PhD-FSTC-2018-72
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 20/11/2018 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Malte HEROLD

Born on 11 January 1988 in Mertesdorf, (Germany)

INTEGRATION OF OMICS DATA FOR BIOTECHNOLOGY-RELEVANT MICROBIAL COMMUNITIES

Dissertation defence committee

Dr Paul Wilmes, dissertation supervisor
Associate Professor, Université du Luxembourg

Dr Florence Abram
National University of Ireland, Galway

Dr Phillip B. Pope
Associate Professor, Norwegian University of Life Sciences

Dr Patrick May, Vice Chairman
Université du Luxembourg

Integration of omics data for biotechnology-relevant microbial communities

A dissertation
by
Malte Herold

Completed in the
Eco-Systems Biology Group, Luxembourg Centre for Systems Biomedicine

To obtain the degree of
DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN *BIOLOGIE*

Dissertation Defence Committee:

Supervisor:	Assoc. Prof. Dr. Paul Wilmes
Comittee members:	Assoc. Prof. Dr. Ines Thiele
	Dr. Patrick May
	Dr. Florence Abram
	Assoc. Prof. Dr. Phillip B. Pope

2018

Declaration

I hereby declare that this dissertation has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this thesis other than those indicated herein.

Malte Herold,
Esch-sur-Alzette, Luxembourg
December 5, 2018

ACKNOWLEDGEMENTS

I would like to thank foremost my supervisor, Prof. Dr. Paul Wilmes, for his support, encouragement, and guidance during these four years.
I am especially grateful to Prof. Dr. Ines Thiele and Dr. Patrick May for encouraging me and their invaluable advice throughout my PhD.
Dr. Florence Abram and Dr. Philipp Pope thank you for taking the time and agreeing to take part in my thesis committee.
A heartfelt thank you goes to Dr. Cedric Laczny, your support and advice throughout has been invaluable.
My utmost gratitude to Dr. Emilie Mueller and Dr. Anna Heintz-Buschart for their guidance and support.
Thank you to all the collaborators of the SysMetEx project, foremost Prof. Dr. Mark Dopson.
I am especially grateful to all present former and current members of the Eco-Systems Biology group for their enthusiasm, support, and making this an unforgettable journey.
Last but not least, I would like to thank my friends and family, and especially my girlfriend Nicole.
Thank you, without you this would not have been possible.

ABSTRACT

Naturally occurring and artificial bacterial communities play an import role in many biotechnological processes. To elucidate bacterial interactions that are important for potential optimized biotechnological applications, high-throughput measurements of biomolecules, metagenomics, metratranscriptomics, metaproteomics, and meta-metabolomics provide a detailed snapshot of mixed microbial consortia.

Integration of multiple layers of omics data allows to reconstruct structure and function of complex microbial communities and is demonstrated for two different model systems. The first chapter focuses on synthetic communities consisting of strains representing key species found in biomining operations and acid mine drainage and that are of economical interest for copper production. A high-quality closed reference genome for *L. ferriphilum* was obtained by DNA sequencing and was subsequently used to integrate functional omics data, i.e. transcriptomic and proteomic profiling. The combination of genomics, genome annotation, and functional omics data allowed an in-depth characterization of *L. ferriphilum* in culture medium and in the presence of the iron sulfide mineral chalcopyrite, an economically relevant copper ore. Subsequently, analyses were performed for co-cultures of up to three organisms highlighting specific interaction mechanisms. The cultures without *L. ferriphilum* showed higher copper solubilisation rates, as the highly efficient iron oxidiser might raise the redox potential above the optimal range.

For *in situ* studies, reference-based analyses are of limited use, e.g. due to a lack in reference genomes of culturable isolates. Hence, the second chapter focuses on an approach to study mixed microbial communities independent of prior knowledge and available reference genomes. A time-series of oleaginous floating sludge samples that spans over one and a half years was analysed by integrating metagenomic, metatranscriptomic, metaproteomic, and meta-metabolomic data. This allowed the reconstruction of population level genomes and the characterization of the niches of the respective populations. The functional potential was assessed, as well as expression profiles over time, yielding a detailed view on lifestyle strategies and the potential impact of abiotic factors. Understanding the niche ecology of the predominant lipid accumulators in the system could lead towards optimized biofuel production.

SCIENTIFIC OUTPUT

Major parts of this thesis are based upon work that has either been published or is in preparation for submission with the candidate as first author. In addition, the candidate has co-authored several publications of which minor parts are incorporated in the thesis. The full list of scientific outputs is listed below and the original manuscripts are provided in **Appendix C**.

Publications in peer-review journals

- Stephan Christel[†], **Malte Herold**[†], Sören Bellenberg, Mohamed El Hajjami, Antoine Buetti-Dinh, Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, Mark Dopson (2017). Multi-omics reveals the lifestyle of the acidophilic, mineral-oxidizing model species *Leptospirillum ferriphilum*^{T}.
  Applied and Environmental Microbiology **84**: e02091-17. doi: 10.1128/AEM.02091-17 [**Appendix C.1**]

- Stephan Christel, Mark Dopson, Mario Vera, Wolfgang Sand, **Malte Herold**, Paul Wilmes, Antoine Buetti-Dinh, Igor Pivking, Christian Trötschel, Ansgar Poetsch, Jan Nygren, Mikael Kubista (2015). Systems Biology of Acidophile Biofilms for Efficient Metal Extraction
  Advanced Materials Research **1130**: 312–315. doi: 10.4028/www.scientific.net/AMR.1130.312. [**Appendix C.2**]

- Sören Bellenberg[†], Antoine Buetti-Dinh[†], Vanni Galli, Olga Ilie, **Malte Herold**, Stephan Christel, Mariia Boretska, Igor Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, Mark Dopson (2018). Automated microscopical analysis of metal sulfide colonization by acidophilic microorganisms
  Applied and Environmental Microbiology **84**: e01835-18. doi: 10.1128/AEM.01835-18 [**Appendix C.4**]

[†] Co-first author[†]

- Shaman Narayanasamy[†], Yohan Jarosz[†], Emilie E.L. Muller, Anna Heintz-Buschart, **Malte Herold**, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses.
  Genome Biology **17**: 260. doi: 10.1186/s13059-016-1116-8 [**Appendix C.5**]

- Cedric C. Laczny, Emilie E.L. Muller, Anna Heintz-Buschart, **Malte Herold**, Laura A. Lebrun, Angela Hogan, Patrick May, Carine de Beaufort, Paul Wilmes (2016). Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract.
  Frontiers in Microbiology **7**. doi: 10.3389/fmicb.2016.00884 [**Appendix C.7**]

- Emilie E.L. Muller[†], Shaman Narayanasamy[†], Myriam Zeimes, Cedric C. Laczny, Laura A. Lebrun, **Malte Herold**, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul Keim, Paul Wilmes (2017). First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*.
  Standards in Genomic Sciences **12**: 64. doi: 10.1186/s40793-017-0274-y [**Appendix C.8**]

- Emilie E.L. Muller, Karoline Faust, Stefanie Widder, **Malte Herold**, Susana Martinez Arbas, Paul Wilmes (2018). Using metabolic networks to resolve ecological properties of microbiomes.
  Current Opinion in Systems Biology **8**: 73-80. doi: 10.1016/j.coisb.2017.12.004 [**Appendix C.6**]

- Linda Wampach [†], Anna Heintz-Buschart [†] , Joëlle V. Fritz [†], Javier Ramiro-Garcia, Janine Habier, **Malte Herold**, Shaman Narayanasamy, Anne Kaysen, Angela H. Hogan, Lutz Bindl, Jean Bottu, Rashi Halder, Conny Sjöqvist, Patrick May, Anders F. Andersson, Carine de Beaufort, Paul Wilmes (2018). Birth mode determines earliest strain-conferred gut microbiome functions and immunostimulatory potential.
  Nature Communications **9**: 5091. [**Appendix C.9**]

Submissions in peer-review journals

- Stephan Christel, **Malte Herold**, Sören Bellenberg, Antoine Buetti-Dinh, Mohamed El Hajjami, Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, and Mark Dopson (2018). Weak Iron Oxidation by *Sulfobacillus thermosulfidooxidans* maintains a favorable redox potential for chalcopyrite bioleaching.
  Frontiers in Microbiology - **in review**. [**Appendix C.3**]

[†] Co-first author[†]

Manuscripts in preparation

- **Herold** *et al.* - Defining fundamental and realized microbial niches using integrated time-resolved multi-omics [**Chapter 3**]
- Kleine-Borgmann *et al.* - Lipid accumulating bacteria from biological wastewater treatment plants: from isolation to *in situ* population dynamics and activity [**Chapter 3**]
- Martinez Arbas [†], Narayanasamy [†] *et al.* - An integrated-omic view of invasive mobile genetic elements and their linked host population dynamics within a microbial community
- **Herold** *et al.* - The SysMetEx Data Collection
- **Herold**[†], Buetti-Dinh[†] *et al.*, - Co-expression networks to determine interactions for defined acidophile communities in chalcopyrite bioleaching.

Oral presentations in scientific conferences, symposia and workshops

- Leptospirillum ferriphilum - Genome, Transcriptome, and Proteome of a Biomining Model Species (2017). *International Biohydrometallurgy Symposium.* Freiberg, Germany.
- Integrated multi-omics for understanding niche ecology of distinct populations in microbial consortia (2016). *16th International Symposium on Microbial Ecology*. Montreal, Canada.
- Integrated multi-omics for understanding niche ecology of distinct populations in microbial consortia (2016) *Life Science PhD Days 2016.* Belval, Luxembourg.

Poster presentations in scientific conferences, symposia and workshops

- SysMetEx. Systems Biology of Acidophile Biofilms for Efficient Metal Extraction (2015). *ErasysApp meeting*. Berlin, Germany.
- Integrated time-resolved multi-omics for understanding microbial niche ecology (2018). *17th International Symposium on Microbial Ecology*. Leipzig, Germany.

[†] Co-first author

CONTENTS

Abstract ii

Scientific output iii

Contents vi

List of Figures x

List of Tables xii

1 Introduction 1

- 1.1 Mixed microbial communities 2
  - 1.1.1 Factors shaping microbial communities 3
- 1.2 Biotechnological processes driven by microbial communities 6
  - 1.2.1 Bioprospecting 6
  - 1.2.2 Towards understanding community interactions and interventions 9
- 1.3 Multi-omics measurements and integration of large-scale datasets 11
  - 1.3.1 Metagenomics 11
  - 1.3.2 Metatranscriptomics 12
  - 1.3.3 Metaproteomics 13
  - 1.3.4 Metabolomics 14
  - 1.3.5 Considerations for integrating multi-omic datasets from microbiomes . . . 14
  - 1.3.6 Computational methods for omics analyses 16
- 1.4 Objectives of this work 19
  - 1.4.1 Genomics, transcriptomics, and proteomics of synthetic acidophile communities 19

1.4.2 Integrated metaomics of an *in situ* time-series of wastewater sludge samples for understanding microbial niche ecology 19

2 Reference-based omics analysis in the context of chalcopyrite bioleaching 21

2.1 Abstract 22
2.2 Background 22
2.2.1 Biomining 22
2.2.2 Acidophile communities for bioleaching and model strains 24
2.2.3 Omics approaches to understand acidophiles 25
2.2.4 The SysMetEx consortium project 26
2.3 Methods 26
2.3.1 Culturing of individual acidophiles and low-complexity communities . . . 26
2.3.2 Imaging analyses and experiments with diffusible signalling factor addition 27
2.3.3 Biomolecular extractions 28
2.3.4 Nucleic acid sequencing 29
2.3.5 Reference genomes and functional annotation 29
2.3.6 Proteomic analyses for *L. ferriphilum* 30
2.3.7 Data analysis 30
2.3.8 Data availability 31
2.4 Results 31
2.4.1 Sequencing and characterisation of the *L. ferriphilum*^{T} isolate genome . . . 31
2.4.2 Detailed characterisation of functional capabilities of *L. ferriphilum*^{T} . . . 35
2.4.3 Comparison of *L. ferriphilum*^{T} continuous and bioleaching cultures 41
2.4.4 Efficient chalcopyrite leaching with favourable redox potential assessed in mixed cultures 43
2.4.5 Biofilm dispersal signals in mixed cultures 48
2.5 Discussion and outlook 52
2.5.1 Functional omics provide an in-depth understanding of acidophile lifestyle in defined conditions 52
2.5.2 From defined conditions and consortia to *in situ* analyses 53

3 Identifying and characterizing fundamental and realized ecological niches of microbial populations in oleaginous floating sludge 54

3.1 Abstract 55
3.2 Background 55
3.2.1 Wastewater treatment plant as model system for microbial ecology 55
3.2.2 Engineering microbial niches for efficient lipid accumulation 57
3.3 Methods 58

3.3.1 Recovery of isolate genomes 58
3.3.2 Sampling, experimental setup, and biomolecular extractions 58
3.3.3 Nucleic acid sequencing, data preprocessing, and assembly 59
3.3.4 Meta-metabolomics 60
3.3.5 Metaproteomics 61
3.3.6 Binning 61
3.3.7 Genome annotation and taxonomic assignments 64
3.3.8 Analysis of functional potential 64
3.3.9 Analysis of gene expression levels 65
3.3.10 Linking abiotic factors to population abundance and expression patterns . . 65
3.4 Results 66
3.4.1 Reconstruction of representative genomes to determine the community structure over time 66
3.4.2 Fundamental niche and resource associations 70
3.4.3 Characterizing realized niches by assessment of gene expression patterns . 77
3.4.4 Expression levels in the *in situ* time-series and short-term *in vitro* experiments 81
3.5 Discussion and outlook 87

4 General conclusions and future perspectives **91**
4.1 General perspectives 92
4.2 Recovery of representative genomes for distinct populations 93
4.2.1 Improving the reconstruction of population-level genomes 94
4.3 Genome annotation 94
4.4 Data management 95
4.5 Integration of omics data in community models 96

Bibliography **98**

Appendices **131**

Appendix A Additional files and tables **132**
A.1 Additional file 2.1 133
A.2 Additional file 2.2 133
A.3 Additional file 3.1 133

Appendix B Additional figures **134**
B.1 Additional figures for **Chapter 3** 134

Appendix C Article manuscripts 144

C.1 Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species *Leptospirillum ferriphilum*^{T}. 145

C.2 Systems Biology of Acidophile Biofilms for Efficient Metal Extraction. 163

C.3 Weak Iron Oxidation by *Sulfobacillus thermosulfidooxidans* maintains a favorable redox potential for chalcopyrite bioleaching. 168

C.4 Automated microscopical analysis of metal sulfide colonization by acidophilic microorganisms. 194

C.5 IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. 210

C.6 Using metabolic networks to resolve ecological properties of microbiomes. 232

C.7 Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract. 241

C.8 First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*. 259

C.9 Birth mode determines earliest strain-conferred gut microbiome functions and immunostimulatory potential. 268

LIST OF FIGURES

1.1 Illustration of niche breadth 5
1.2 Bioprospecting with culture-dependent and -independent methods 8
1.3 The potential societal impact of research on microbial consortia 10
1.4 Biomolecular extraction from a single sample 16
1.5 The IMP pipeline workflow schematic 18

2.1 Illustration of heap bioleaching 23
2.2 Summary equations for chalcopyrite dissolution 24
2.3 Comparison of the newly sequenced genome to the previous draft genome of the *L. ferriphilum* type-strain 33
2.4 Venn diagram of orthologous gene clusters across *L. ferriphilum* reference genomes 34
2.5 Circular representation of the *L. ferriphilumT* genome 36
2.6 Model of energy conservation in *L. ferriphilum* 38
2.7 Model of carbon and nitrogen fixation in *L. ferriphilum*^{T} 39
2.8 Model of biofilm formation and quorum sensing in *L. ferriphilum*^{T} 40
2.9 Differential abundance of transcripts and proteins in L. ferriphilum 42
2.10 Chalcopyrite bioleaching and redox potentials in different combinations 44
2.11 Proposed model of *S. thermosulfidooxidans* transcript regulation of genes related to energy conservation 47
2.12 Expression of DSF system genes in different conditions and combinations 49
2.13 DSF molecules and their effect on attached cells 51

3.1 Conceptual schematic of niche engineering in a BWWTP 58
3.2 Assembly and automated binning workflow 63
3.3 Relative abundance of reconstructed populations over time 67

3.4 Constrained ordination of population abundances 69
3.5 Assignment of cluster according to functional potential 71
3.6 Upset plot of shared KEGG orthologue assignments per FunC 72
3.7 Relative abundance of ReGes related to FunC assignment 73
3.8 Measurements of physico-chemical parameters over the time-series 75
3.9 Rank correlations of abiotic factors and relative abundances 76
3.10 MDS of binarized expression profiles D04_G2.5 78
3.11 Influence of abiotic factors on expression profiles of representative genomes 79
3.12 KO expression profiles, comparison across all ReGes 80
3.13 Ratios of metatranscriptomic and metagenomic depth of coverage in the defined short-term conditions and the long-term time-series 82
3.14 ReGe expression levels for genes related to beta-oxidation 84
3.15 ReGe expression levels for genes related to triacylglycerol metabolism 85
3.16 ReGe expression levels for genes related to nitrogen metabolism 86

B.1 Ribotagger derived community structure 135
B.2 Constrained ordination based on OTU community structure 135
B.3 Correlation of abiotic factors to relative population abundances 136
B.4 Ratios of active genes associated to fatty acid degradation 137
B.5 Metabolite intensities for LCFAs in extra- and intracellular fractions 138
B.6 Metabolite intensities for glycerol in extra- and intracellular fractions 139
B.7 Metabolite intensities for carbohydrates in extra- and intracellular fractions 140
B.8 Metabolite intensities for amino acids in extra- and intracellular fractions 141
B.9 Metabolite intensities for amines in extra- and intracellular fractions 142
B.10 Correlation of abiotic factors to relative population abundances 143

LIST OF TABLES

1.1 Species richness of microbial consortia across different environments 3

2.1 Available reference genomes for *L. ferriphilum* 31
2.2 General genome statistics for the *L. ferriphilum*^{T} genome 32
2.3 Overview of analysed samples for *L. ferriphilum*^{T} 35
2.4 Transcriptomic changes in iron and sulfur metabolism related genes 45
2.5 Presence of DSF family quorum sensing system-encoding genes 48

CHAPTER 1

INTRODUCTION

1.1 Mixed microbial communities

In recent years, studying microbial communities with new molecular methods has solidified the view that microbial communities play integral roles in many parts of the Earth's ecosystems, from biogeochemical cycles [Rousk and Bengtson, 2014] to microbiomes associated with host organisms such as humans [Greenhalgh et al., 2016]. While the more we are learning about these microbial consortia, the more we come to realize that the diversity that we observe is only the tip of the iceberg. Recent estimates suggest that there could be as many as 10^{11}-10^{12} bacterial species on earth of which only ca. 10^4 have been cultured [Locey and Lennon, 2016]. Additionally, microbial communities often include archaea, fungi, viruses, and eukaryotes adding to the complexity.

The classical microbial approach of studying isolate cultures of individual species has improved our understanding of how these organisms function and it is still a prerequisite for accurate characterization of organismal physiology. Yet, classical methods fall short of characterising bacterial communities as a whole and genome sequencing has revealed an unfathomed diversity in uncultured microorganisms [Hug et al., 2016]. Many bacterial species are not readily culturable due to lack of specific nutrients, growth factors or other inter-species communication [Vartoukian et al., 2010], and also physical conditions such as hydraulics [Niederdorfer et al., 2016] or cell-to-surface contact can be a requirement for culturability. While new methods have been developed for culturing microorganisms from complex mixtures [Vartoukian et al., 2010], methods to study microbial communities in a holistic manner are required to understand emergent properties of microbial systems, i.e. properties that cannot be attributed solely to additive effects of the system's components [Konopka, 2009].

Microbial communities are shaped by their biotope and biotic interactions, such as competition for resources, cross-feeding, or horizontal gene transfer [Konopka, 2009]. The notion of the environment as the primary shaping force for microbial communities has been formulated concisely in a famous quote by the Dutch biologist Lourens Baas-Becking in the early 20th century: "Every-thing is everywhere, but, the environment selects" [Baas-Becking, 1934], which can be seen as a precursor for niche assembly theories [De Wit and Bouvier, 2006] that will be discussed in greater detail subsequently (**Section 1.1.1**). Various mechanisms exist a for wide distribution of bacterial cells, e.g. by attachment to aerosols [Joung et al., 2017]. Distribution and persistence across different environments is also facilitated by the ability of many bacterial species to form spores, the ability to resist extreme conditions in a dormant form. Furthermore, microorganisms exhibit diverse resistance mechanisms that guarantees their survival in extreme conditions, e.g. metal rich, acidic conditions [Dopson and Holmes, 2014]. As a result of the various mechanisms that evolved in bacteria, they can be found in nearly all environments on earth.

The complexity or diversity, i.e. species richness and distribution, of an ecosystem's microbiome depends on the environmental conditions and nutrient availability (**Table 1.1**) [Torsvik, 2002]. Mi-

Table 1.1: Species richness of microbial consortia across different environments. The numbers shown are based on different methods for estimation or extrapolation. The definition of a taxa can vary, e.g. for 16S amplicon sequencing it is defined as distinguishable based on 16S RNA gene sequence identity and depends on the selected identity cut-off (Adapted from [Wilmes et al., 2015]).

DNA source	Estimated number of taxa	Basis for estimate	Reference
Acid mine drainage biofilm	159	Total RNA sequencing	[Goltsman et al., 2015]
Activated sludge	>1000	16S amplicon sequencing	[Zhang et al., 2012]
Ocean Water	160	various	[Curtis et al., 2002]
Soil	6,300	various	[Curtis et al., 2002]
Surface freshwater	20,000	various	[Palmer, 1997]
Soil	50,000	16S amplicon sequencing	[Roesch et al., 2007]
Soil	8,000,000	DNA reassociation	[Gans et al., 2005]
Human saliva	>5400	16S amplicon sequencing	[Huse et al., 2012]
Human feces	>21,000	16S amplicon sequencing	[Huse et al., 2012]

crobial organisms can even be found in extreme environments, such as hypersaline lakes [Benlloch et al., 2002], hot deserts [Makhalanyane et al., 2015], or permafrost soils [Hultman et al., 2015] and also in highly acidic, metal rich environments. Communities in these extreme environments often only consist of a limited number of species compared to other environments [Baker and Banfield, 2003; Denef et al., 2010]. Acidophile communities in acid mine drainage (AMD) ecosystems, in which microorganisms mediate the oxidative solubilization of sulfide minerals, have been a model system for applying and developing molecular methods early on [Tyson et al., 2004; Allen and Banfield, 2005; Wilmes and Bond, 2009] also due to the limited number of species [Baker and Banfield, 2003]. Determining ecological interactions of microbial communities in the AMD ecosystem [Denef et al., 2010] is not only important for environmental considerations, but is also of economical relevance (**Section 2.2.1**).

Communities found in sediments and soils typically exhibit high species richness [Torsvik, 2002] (**Table 1.1**). Soil-borne microorganisms are of great importance e.g. for nutrient cycling. A related important field of study are interactions between plants and their associated microbiome, as microorganisms form complex interactions with their plant hosts [Lakshmanan et al., 2014] that influence plant health and productivity. Host-microbiome interactions also are crucial factor in studies about the human microbiome, a field that has received increasing attention in the last years. Here, especially the human gut microbiome has been shown to be implicated in several diseases, e.g. type-2-diabetes [Qin et al., 2012] or inflammatory bowel disease [Morgan et al., 2012]. An in-depth understanding of microbiomes enables us not only to potentially derive treatments beneficial for health, but also provides the background for the optimization of biotechnological processes (**Section 1.2**).

1.1.1 Factors shaping microbial communities

The factors shaping mixed microbial communities can be quite complex and their elucidation is at the centre of microbial ecology. In host-linked microbial communities, naturally the environment provided by the host plays an important role in the formation and persistence of the microbiome. However, active control mechanisms in host-microbiota interactions can also play an important role in shaping the co-evolution of host and microorganisms [Foster et al., 2017]. In the human gut, the microbiome could be shaped for example by immune regulation or by exposing specific mucus layer glycans at the epithelium-interface favouring certain organisms [Schluter and Foster, 2012].

The notion that ecosystems are primarily shaped deterministically, by conditions and accessible resources has found wide resonance in niche assembly theories. According to the niche definition by Hutchinson, a species' fundamental niche is defined by an n-dimensional hypervolume of ecological parameters allowing the species to persist [Hutchinson, 1957]. As two species that are limited by the same resource cannot coexist in an environment, they will utilize a reduced set of resources in the presence of each other, the realised niche [Hutchinson, 1957] (**Figure 1.1**). In microbial ecology the fundamental niche can be seen as reflected by the genomic complement, i.e. functional potential of encoded genes, while the realised niche can be seen as the actually expressed gene functions [Muller et al., 2018]. The fitness of a genotype can vary across the range of different gradients of resources (**Figure 1.1**). A specialist population will have a narrower niche breadth than a generalist, which is adapted to a wider range of resources, however this can occur at a cost reflected in fitness trade-offs [Kassen, 2002].

The evolution of individual microbial genotypes can be tracked experimentally in response to a resource gradient, e.g. for thermal adaptation [Bennett and Lenski, 1993] or exposure to light [Kassen, 2002]. However, testing ecological concepts in *in situ* systems can be challenging due to methodological constraints, as measuring fitness or resource usage of individual species is not trivial. The co-existence and interactions of a multitude of different species and their (co)-evolution in dynamic and spatially heterogeneous environments poses additional challenges. Molecular methods such as DNA sequencing (**Section 1.3.1**) can be utilized to determine the species present in a mixed community and their abundance in the sample. Frequently, concepts of niche theory are then applied implicitly in studies aiming to link environmental parameters to the community structure. Integrating DNA, RNA, protein, and metabolite data (**Section 1.3**) can provide expanded information on community functioning and can be applied for testing ecological concepts in microbial consortia, for example, the concept of niche breadth and the activity of generalist and specialist populations within a microbial community described by Muller et al. [2014a].

Contrasting deterministic niche theories, neutral theory [Hubbell, 2006] models have emerged that model community structure by stochastic birth, death, and immigration events and have also widely been applied for microbial communities [Sloan et al., 2006]. Microbial communities also are characterized by a considerable functional redundancy, i.e. the capability of distinct phylotypes to perform the same ecological function [Prosser, 2012]. While functional redundancy has been seen as a sign for co-existence between competitors in neutral processes, it has been suggested that functionally similar microorganisms could differentiate ecologically in much more complex ways than we are able to observe [Louca et al., 2018]. It has recently been shown that niche specialisation can be reflected by species containing diverse isoforms of enzymes [Rubino et al., 2017] and it was suggested that fine-tuned gene expression of isoforms could convey competitive advantages in fluctuating resource availabilities to a generalist population [Muller et al., 2014a]. Furthermore, it has been suggested that niche segregation takes place on a transcriptional level in the human gut microbiome, as adaptations in gene expression could be linked to a reduction in functional redundancy [Plichta et al., 2016].

Studies have suggested combining models for niche and stochastic effects to characterise species abundance patterns [Ofiteru et al., 2010]. Niche effects can be seen as stabilizing processes that cause intraspecies effects to be more negative than interspecies effects, slowing down growth of abundant populations and thus limiting competitive exclusion [Adler et al., 2007]. As a consequence, co-existence of two species is determined in balancing fitness inequalities and stabilizing processes [Adler et al., 2007]. Unfortunately, quantifying stabilizing effects or fitness for microbial organisms is a complex task. To elucidate how ecosystems shape and are shaped by microbial communities, a framework has been suggested distinguishing between different levels of processes, such as microbial membership, community properties, microbial processes, and ecosystem processes [Hall et al., 2018]. Understanding the various processes involved in microbial systems is a pre-requisite for targetted manipulations with the aim of boosting a desired community phenotype (**Section 1.2.2** and **Section 3.2.2**).

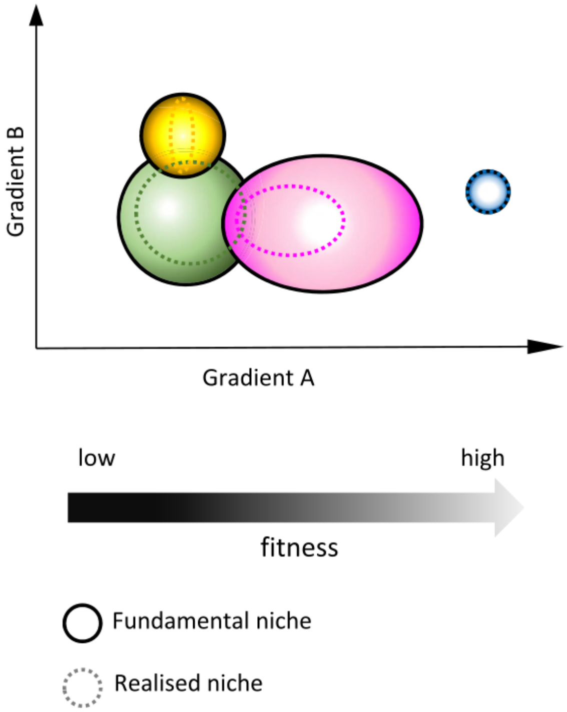


Figure 1.1: Illustration of niche breadth. Visualisation of niche space for a community with 4 organisms along 2 resource gradients A and B. Niche breadth is indicated by circle size and while generalist only realize a smaller portion of their fundamental niche, the realized niche of specialist (blue) could resemble their fundamental niche more closely (from Muller et al. [2018]).

1.2 Biotechnological processes driven by microbial communities

Historically processes driven by microbial activity have been used in early stages of human civilization, as for example evidence of fermented beverages could be traced back as far as 9,000 years ago [McGovern et al., 2004]. Similarly other fermented foods were also already produced in ancient times. Most likely the microbial activity was used by chance and processes were refined through experience. Systematic study of industrial fermentation processes began in the 19th century which is considered the basis for many modern industrial biotechnological processes. Modern biotechnological processes, e.g. recombinant insulin production, or bulk chemical production, e.g. for amino acid production, primarily use pure cultures as effective expression systems [Baeshen et al., 2014]. However, many important processes in modern society such as wastewater treatment with activated sludge (**Section 3.2.1**), anaerobic biogas production [Bremges et al., 2015], or biomining (**Section 2.2.1**) are driven by microbial communities with many yet uncultivable organisms.

1.2.1 Bioprospecting

The vast diversity in microbial communities is also seen as a large untapped resource for discovery of new enzymes or other bioactive compounds from various environments [Keller and Zengler, 2004]. The unique capabilities that microorganisms have evolved to survive in (extreme) environments could provide means to facilitate or improve biotechnological processes. Frequently, environments are selected and sampled based on their physico-chemical properties favouring evolution of a characteristic phenotype, e.g. sampling of environments with high salinity to recover halophilic organisms. Halophilic acidophiles, for example, are of great interest to carry out heap bioleaching (**Section 2.2.1**) with seawater [Watling, 2016] instead of valuable freshwater in dry regions in Chile. Not only is the discovery of specialised organisms commonly pursued, but also individual enzymes can be of value in biotechnological processes. A famous example is the recovery of the heat-stable *Taq*-polymerase of the thermophile organism *Thermus aquaticus* [Chien et al., 1976] commonly found in hot springs and the resulting application in DNA amplification with polymerase-chain reaction (PCR) [Saiki et al., 1988].

While extreme environments are often targeted for the retrieval of potentially interesting organisms, enzymes, or compounds, environments with a higher diversity of prokaryotes are frequently targeted for the retrieval of antimicrobials. Several microorganisms, for example *Streptomyces*, produce antibiotics likely conferring a competitive advantage in nutrient rich micro-environments [Williams and Vickers, 1986]. As commonly culturable soil bacteria have historically been utilized as a source for antimicrobial compounds, discovery of new compounds has been challenging [Lewis, 2012]. New approaches for isolation and culturing [Nichols et al., 2010] have shown successes, exemplified by the recent discovery of a novel candidate class of antibiotics not triggering resistance development in gram-positive pathogens [Ling et al., 2015]. Also, culture-independent approaches have shown success for the discovery of previously unknown classes of antibiotics in soil microbiomes [Hover et al., 2018]. Marine environments also represent important reservoirs of antimicrobial peptides and other bioactive compounds [Reen et al., 2015].

Culture-dependent and -independent screening methods (**Figure 1.2**) have successfully been applied for the discovery of novel bioactive compounds or organisms with biotechnologically relevant properties from various environments. Especially, sequencing-based methods play an important role in accessing the yet uncultivable microbial diversity from mixed communities.

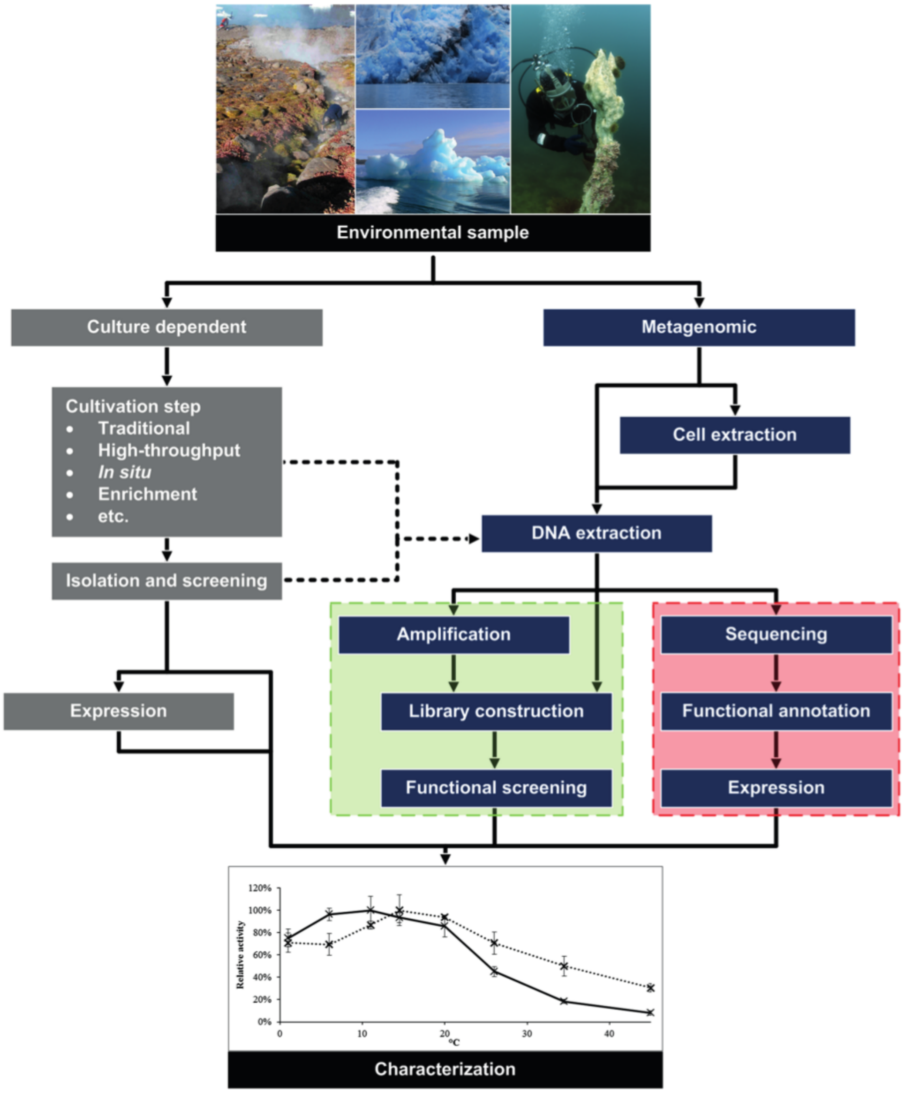


Figure 1.2: Bioprospecting with culture-dependent and -independent methods. Overview of culture-dependent (grey boxes) and culture-independent (metagenomics, blue boxes) screening workflows for bioprospecting of environmental samples. Metagenomic approaches can either be targeted (green) or untargeted (red) (from Vester et al. [2015]).

1.2.2 Towards understanding community interactions and interventions

As microbial communities play fundamental roles in biogeochemical cycles on Earth, elucidating these processes is expected to enable human society to apply biological systems as solutions to environmental needs and problems [Madsen, 2011] (**Figure 1.3**). Still, several challenges to obtain a predictive understanding of microbial interactions within ecosystems remain, such as gaps in decoding microbial gene functions, as well as quantitative measurements across different spatial or temporal scales [Blaser et al., 2016].

The interactions occurring between the constituent populations of microbial consortia can be characterized in different ways, depending on the interaction. These interaction types include among others cooperation, mutual benefits, or competition, for example, by exploiting shared resources, or interfering directly by toxin production. Types of social interactions and their evolution in microorganisms have extensively been reviewed by Mitri and Foster [2013]. In defined co-cultures these interactions can be studied in a targeted way [Grosskopf and Soyer, 2014]. Utilizing synthetic microbial communities also has tremendous potential in biotechnological processes [Shong et al., 2012]. Processes with mixed cultures could be optimized so that product yields could be increased. This could be achieved, for example, by the combined production of several distinct products efficiently utilizing available substrates, by the conversion of toxic by-products, or conversion of mixed and cheaper substrates by combining species with distinct pathways [Sabra et al., 2010]. Additionally microbial communities are considered to be more robust to external perturbations e.g. contamination and could be handled under non-sterile conditions further reducing processing costs [Sabra et al., 2010]. Ultimately engineered microbial consortia could be used as co-operative and stable production systems [Cavaliere et al., 2017]. Designing optimal community composition can be facilitated by computational modelling approaches (**Section 4.5**), e.g. as has been demonstrated for biogas production [Koch et al., 2016].

Understanding interactions within artificial consortia might however not directly be transferable to more complex natural systems. Here, an in-depth understanding of the underlying ecology is required for rationally motivated interventions that could steer a community phenotype towards a desired goal. One example where interventions in microbial communities are commonly pursued can be found in host associated systems, such as the treatment of livestock with pre- and probiotics to increase animal health and replace widespread antibiotics use [Uyeno et al., 2015]. Interventions with probiotics have also shown positive effects when applied for gastrointestinal diseases in humans, yet could be improved by knowledge of which strains best to use or optimization of the dosage [Ritchie and Romanuk, 2012]. Also, the introduction of a foreign microbiome of a healthy donor is applied for treatment, through faecal transplants [Gupta et al., 2016].

Overall, biotechnological processes depending on microbial consortia and microbiome interventions are currently still constrained by the limited knowledge of the of the complex interactions and

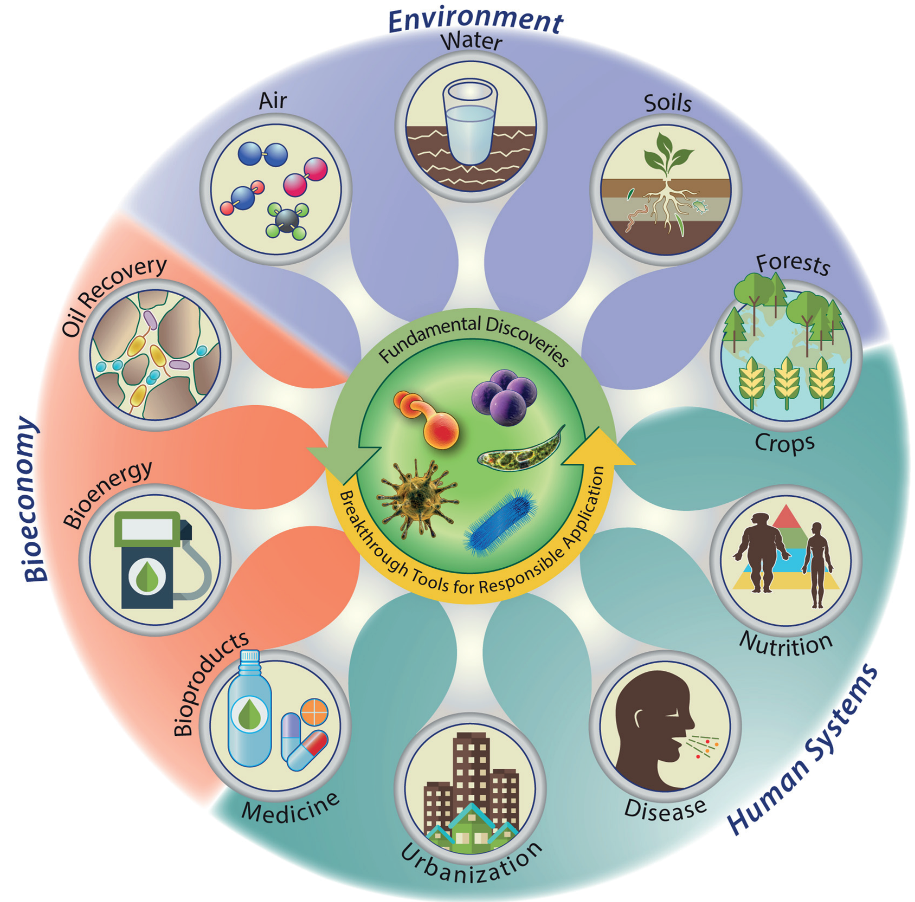


Figure 1.3: The potential societal impact of research on microbial consortia. A unified framework to move from fundamental discoveries in microbial communities in different environments holds the potential to benefit society in many areas from biotechnological applications to human health (from Blaser et al. [2016]).

dynamics within these ecosystems. A detailed and predictive understanding of microbial systems could potentially lead to improvements in many applications, e.g. in bioenergy or human health (**Figure 1.3**) thus representing an important area of research.

1.3 Multi-omics measurements and integration of large-scale datasets

In recent years, so called "-omics techniques" have been developed and improved, allowing a detailed characterisation of the building blocks of cellular systems. The term summarises the various techniques for high-throughput measurements for generating biomolecular data sets, for example genomic, transcriptomic, proteomic, or metabolomic data. While these approaches have been developed and are primarily used for the study of isolates, they can also be applied for mixed microbial communities with metaomics or environmental omics techniques. Metagenomic (MG) sequencing refers to the sequencing of libraries generated from DNA extracted from the entire microbial communities. Metatranscriptomic (MT) sequencing respectively means the sequencing of the RNA complement, mostly indirectly by sequencing reverse transcribed cDNA libraries. Metaproteomic(s) (MP) refers to the study of the whole protein complement of microbial systems and meta-metabolomic(s) (referred to as metabolomics or MM) describes the analysis of the metabolites in an environmental sample. Applied independently or in combination (multi-omics) these techniques enable the study of mixed microbial communities in unprecedented detail and have led to numerous advances in microbial systems ecology [Gutleben et al., 2018].

1.3.1 Metagenomics

With the advent of next-generation sequencing (NGS) techniques, DNA sequencing has seen a tremendous decrease in cost. DNA is extracted and purified, fragmented into templates, which are then amplified (second-generation sequencing techniques) or directly sequenced as single molecules (third-generation methods), with specifics depending on the applied platform [Metzker, 2010]. The different sequencing techniques vary in throughput, and error rates and types, as well as the length of resulting sequencing reads [Scholz et al., 2012]. Second-generation sequencing methods often generate reads with a length around 100 bp, while third-generation methods produce reads with average length above 1 kbp, which allows resolution of extended repetitive genomic regions [Metzker, 2010; van Dijk et al., 2018].

As the DNA extracted from mixed microbial community samples depends on the presence and the quantity of the present microbial populations, MG sequencing has become standard practice to infer the community composition. Frequently, a targeted approach is applied in which a known sequence, often a marker sequence,commonly the 16S ribosomal RNA (rRNA) gene for bacteria, is amplified with primers prior to library preparation. The resulting sequencing reads are then processed and often grouped into operational taxonomic units (OTU) according to the sequence similarity, reflecting a high-resolution taxonomic profile of the original bacterial populations [Hugerth and Andersson, 2017].

In addition to the resolution of taxonomic composition, MG approaches can also be used to elucidate the functional potential of constituent microbial populations of a consortium. Even though functional information can be extrapolated from the amplicon sequencing derived community structure, this approach is dependent on prior knowledge in the availability of reference genomes that is variable across different environments [Langille et al., 2013]. As a result it can only recapitulate the functional potential of already known strains, but not new functional potential of similar, but genetically different strains. For environmental samples suitable reference genomes are not always available [Langille et al., 2013], as even for well-studied system as the human gut microbiome only 43 % of sequences could be associated to existing references [Sunagawa et al., 2013].

Whole genome shotgun (WGS) sequencing, i.e. sequencing of the randomly fragmented whole DNA complement, allows taxonomic and functional profiling either on individual read level or on consensus level, i.e., after an assembly (**Section 3.3.3**) of the reads into longer contiguous sequences (contigs) [Scholz et al., 2012]. Reconstruction of whole genomes can be achieved by binning contigs derived from the same population based on characteristic genomic properties such as GC-content, nucleotide signatures, or depth of coverage distributions (**Section 3.3.6**), an approach that was pioneered in low-complexity AMD consortia allowing inferences of evolutionary origin and metabolic traits of the constituent bacteria and archaea [Tyson et al., 2004]. Reconstructed genomes, also referred to as metagenome-assembled genomes (MAGs), are often incomplete for example due to unresolved repetitive regions or highly-similar genomic regions between different populations. The quality of MAGs is generally assessed by assembly quality, completeness, and contamination estimates [Bowers et al., 2017]. In general MAGs represent an average of a population of cells [Zengler, 2009] and strain heterogeneity can be challenging to resolve [Imelfort et al., 2014]. In recent years single cell genomics has become feasible and potentially allows characterization of individual genotypes [Stepanauskas, 2012]. Overall, MG sequencing enables a detailed view of the microbial community structure and functional potential. However, it does not provide information on the activity of identified functions.

1.3.2 Metatranscriptomics

As not all genes are constitutively expressed in a cell, metatranscriptomics allows the profiling of community-wide gene expression patterns in a sample. Analogously to MG data, MT approaches have tremendously increased in their popularity in the verge of the NGS revolution. By now microarray-based transcriptome profiling [Schena et al., 1995] has largely been replaced by RNAseq due to the wider dynamic range, less bias and background noise, and the possibility to discover novel transcripts or sequence variations [Wang et al., 2009]. RNA can directly be sequenced or first be reverse-transcribed to cDNA, with the latter being the more common application in metatranscriptomics. Profiling of messenger RNA (mRNA) usually implies a depletion of ribosomal RNA (rRNA), as mRNA makes up only a small fraction of the total RNA complement in a cell [Petrova et al., 2017]. Depletion of rRNA can therefore be applied to microbial community samples to massively increase the sequencing depth obtained for mRNA, but the removal rates can vary between different organisms [Petrova et al., 2017]. On the other hand, deriving information on community structure or population dynamics similarly to MG approaches can also be derived from sequencing of rRNA [Goltsman et al., 2015].

MT approaches could show that the most abundant populations within a consortium are not necessarily the most active populations. MT approaches have been applied to microbial consortia across many different environments, such as a thermophilic biogas plant consortium [Maus et al., 2016], the microbiome of the human intestinal tract [Plichta et al., 2016], or arctic peat soil communities [Tveit et al., 2014], among many others. Similarly to MG reads, MT reads are commonly aligned to a reference sequence and matching reads are quantified to characterise the expression e.g. of an individual gene. Computational tools for the quantification of transcripts have recently been reviewed by Conesa et al. [2016] and are described in further detail in **Section 1.3.6**. However, tools for the de novo assembly of transcripts [Schulz et al., 2012] can also be applied to MT data to reconstruct transcripts in the absence of reference genomes [Davids et al., 2016].

1.3.3 Metaproteomics

Metaproteomics, i.e. the measurement of all expressed proteins within an ecosystem, allows the direct assessment of microbial enzymatic activity [Wilmes and Bond, 2004; Hettich et al., 2013]. Profiling the community proteome is commonly done by extraction of proteins from the sample followed by tandem mass spectrometry. A separation of proteins and reduction of sample complexity can be achieved by gel-based separation (e.g. 2D-PAGE) or directly by liquid chromatography (LC). Proteins are cleaved to peptides, e.g. by trypsination. Following separation, mass spectrometry (MS) allows for the measurement of the mass to charge ratios of ionized peptides. An additional MS step after fragmentation of the peptide ions can be used to identify the sequence of amino acids of the original peptides based on the mass to charge ratio of their fragments. This approach is also referred to as shotgun proteomics analogously to WGS. Peptide mass fingerprinting is followed by an protein identification step where identified peptide spectra can be compared to a database of protein sequences. Spectral matches of peptides to the reference protein databases can then be quantified and corrected for the set of reference proteomes [Penzlin et al., 2014]. However, quantification can also be achieved by factoring in peptide peak intensities [Heyer et al., 2017]. In general, the quantification capability of shotgun proteomics approaches is not comparable to that of targeted proteomics, or select-reaction monitoring methods, which however typically lack throughput to resolve complex metaproteomes [Heyer et al., 2017].

Applications of metaproteomics in different environments, such as activated sludge, AMD, marine and freshwater systems, soil, and human gut have been reviewed by Wilmes et al. [2015]. These ecosystems are characterized by a differences in the dynamic range of the protein complement, however only around 1 % of the MP complement is resolvable [Wilmes et al., 2015]. A limitation that has to be accounted for is strain heterogeneity that can prevent matching of peptides to protein databases given amino-acid substitutions [Allen and Banfield, 2005]. This can be accounted for by accurate and complete database generation by integrating other omics data types (**Section 1.3.5**). Recent advances in proteomics are foreseen to solve several existing challenges for MP analyses [Wilmes et al., 2015]. An example of a newly developed method with potential application in metaproteomics is SWATH-MS in which a range of fragment ions is stored and used for peptide identification, thus allowing high-throughput and accurate quantification [Gillet et al., 2012].

1.3.4 Metabolomics

Metabolomics reflect the actual phenotypic state in an environment, as chemical compounds are the end products of microbial conversion reactions. Several considerations have to be made for the measurement of metabolites from community samples. Ongoing metabolic processes have to be quenched to avoid alterations of the metabolite profile after sampling for example by flash freezing in liquid nitrogen [Roume, 2013]. This is especially important when volatile metabolites are measured. Metabolites need then, to be isolated and purified from the sample using suitable solvents. Subsequently, metabolites are derivatized, i.e. chemically modified to increase volatility, depending on the measurement technique. Metabolites are separated e.g. by gas chromatography (GC) or liquid-chromatography (LC) to reduce the complexity of the metabolite mixture and analysed by mass spectrometry (MS) approaches. The resulting mass spectra can then be analysed and linked to databases of known spectra for identification and quantification [Hiller et al., 2009]. Due to its advantages in the methodology such as the high separation efficiency and measurement reproducibility GC based approaches are preferentially used when profiling complex metabolite mixtures without prior knowledge. LC based approaches do not require derivatization and are often applied for the quantification of pre-identified metabolites.

Several challenges exist in the application of metabolomics to microbial community samples. The number of metabolites that are measured in complex samples is high and frequently only a fraction can be identified depending on the sampled environment [Tang, 2011]. Another challenge is connecting metabolite levels to microbial activity, as unlike with MG, MT, or MP data the measured compounds cannot directly be linked to individual populations. This can be circumvented by labelling strategies, e.g. stable isotope probing, which allow inference of metabolic fluxes for specific pathways [Abram, 2015; Srivastava et al., 2016]. However, in *in situ* analyses labelling strategies cannot easily be applied, as samples have to be cultured for the duration of incubation with labelled substrate which then allows tracing the incorporation of the respective isotopes [Abram, 2015; Eyice et al., 2015]

1.3.5 Considerations for integrating multi-omic datasets from microbiomes

An important consideration for the integration of different omic datasets is the co-extraction of biomolecules from an undivided sample (**Figure 1.4**) allowing for meaningful data integration [Roume, 2013] while reducing the effects of sample heterogeneity [Muller et al., 2013]. Sample specific comprehensive lysis is a prerequisite for metaomics analyses, as the biases introduced by variations in lysis efficiency for different species should be minimized given the unknown species composition of a typical sample. However, the applied extraction procedure also needs to account for the sample type as the matrix could also have effects on lysis efficiency and subsequently yields of the purified biomolecules [Lever et al., 2015].

The combination of different omics techniques has several advantages for profiling complex microbial communities and shortcomings of individual omic levels can be overcome by integration with other datasets. Amplicon sequencing allows the profiling of community structure at a great depth and the results can be used to estimate the required sequencing depth for a subsequent application of WGS [Abram, 2015]. An advantage of WGS-MG data is the possibility to recover population-level genomes thus to link predicted gene sequences to their population of origin and also to find novel species- or strain specific genes not characterized in existing reference genomes. While MG data cannot be used to measure activity levels it can therefore be applied to generate reference sequences for functional omics approaches such as those using MT or MP data. For community proteomics, database searches can be informed by MG data derived gene predictions [Wilmes and Bond, 2009], also incorporating methods to account for genetic variants [Heintz-Buschart et al., 2017].

MT and MP measurement techniques are characterized by different dynamic ranges, e.g. while lowly abundant transcripts can still be detected, often only highly abundant proteins can be identified. Combining both MT and MP levels can lead to more complete and reliable profiles of activity. Yet, even though most proteins are considered to have an intrinsic metabolic function [Schneider et al., 2012], even protein levels are not necessarily a direct indication for metabolic function as enzymatic activity also depends on factors such as temperature, pH, or presence of specific co-factors. Metabolomic data in turn can provide an overview of the microbial activity in an ecosystem, which can be related to levels of gene expression or protein abundances [Wilmes et al., 2010b].

A crucial aspect in profiling activity is prior knowledge on the structure of genes and function of encoded proteins reflected in the functional annotation of a genome. The assignment of functional information to gene sequences derived from MG and MT data is a major bottleneck in analysis workflows and remains inherently incomplete [De Filippo et al., 2012]. The lack of functional annotation for many genes can however also may be addressed by the incorporation of MT and

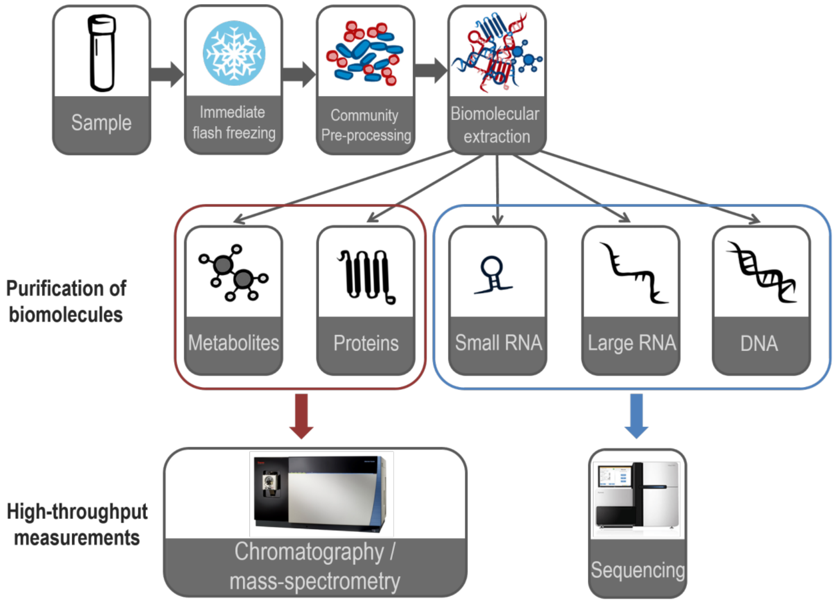


Figure 1.4: Biomolecular extraction from a single sample. After sampling, flash freezing, e.g. in liquid nitrogen, quenches biological and physical processes, such as the conversion of metabolites or the degradation of RNA. An important step in the extraction is the comprehensive lysis of bacterial cells after which the individual biomolecular fractions can be isolated. The data generated either by mass spectrometry or sequencing techniques for the respective biomolecules can then meaningfully be integrated (Courtesy of Linda Wampach and Anne Kaysen).

MP data for example by correcting open reading frame (ORF) annotation by mapping transcripts to the genome sequence [Richardson and Watson, 2013]. Furthermore, transcript levels can also be utilized to improve annotations of genomic context, e.g. operon structures [Fortino et al., 2014]. The activity in different environments or conditions can also be used to assess the potential function of an individual gene. Exact functional annotation of enzymatic activity can be used to incorporate metabolomic data linking gene activity to microbial processes.

Combining various heterogeneous omic datasets is also challenging as the different types biomolecules are generated or processed at different time-scales. While the response to a change in the environment can be quick e.g. on the transcriptomic level, the translation of mRNA to proteins occurs afterwards and a response in protein levels might be observed at a time-point where an increased expression of the respective gene has already ceded. The different time-scales and the different mechanisms of post-transcriptional or post-translational modifications, as well as variable degradation rates of proteins or transcripts explain the perceived lack of correlation between MT and MP levels [Siggins et al., 2012]. Metabolite conversions on the other hand can occur within seconds, often resulting in higher variance between samples [Vemuri and Aristidou, 2005].

A limited number of integrated multi-omic studies has been performed so far, with a comprehensive overview listed in Narayanasamy [2017]. Among others, Heintz-Buschart et al. [2017] highlighted in a MT, MG, MP study of familial type-1 diabetes that describing microbial function on different levels provides insights not achievable by individual omic analyses. Combining multiple omics datasets allows unprecedented insights in microbial activity in diverse ecosystems as has been shown for different permafrost layers [Hultman et al., 2015]. Time-resolved multi-omics approaches can provide a comprehensive picture of dynamics in microbial community functioning while avoiding *a priori* assumptions [Muller et al., 2013, 2014a].

1.3.6 Computational methods for omics analyses

The generation of large-scale datasets is a common feature of all the previously described metaomics techniques. A major bottleneck in multi-omics analyses is the computational analysis of these datasets [Fondi and Liò, 2015]. Therefore, considerable efforts have been devoted to the development of efficient bioinformatic methods for the processing of omics data.

For the analysis of sequencing data efficient alignment methods are of great importance to assign sequencing reads to reference sequences and genomes. Through mapping reads to assembled contigs, the latter can be grouped by their distributions of coverage [Albertsen et al., 2013]. Prediction of the taxonomic origin of DNA sequences also depend on alignment methods and can be performed directly on reads [Gerlach and Stoye, 2011] or on population-level genomes, e.g. by assigning informative marker gene sequences [Wu et al., 2013; Albertsen et al., 2013] to existing databases. Taxonomic classification can also be achieved by hidden Markov model (HMM) searches [Eddy, 2011] of these marker gene sets [Wu and Scott, 2012]. Single copy marker genes can also be used to estimate the completeness and contamination levels of recovered genomes [Parks et al., 2015]. For tracing the evolutionary relationships between different populations, efficient algorithms have been developed for large scale genomic comparisons [Ondov et al., 2015; Brown and Irber, 2016]. For the quantification of MT data, alignment-based methods are also commonly used, which can be a time-consuming step and recently pseudo-alignment or alignment-free heuristic methods have been developed for the quantification of RNAseq reads using kmer-based hashing [Bray et al., 2016].

Another important consideration in computational tool development is reproducibility and automation. The development of automated workflows or computational pipelines can greatly facilitate omics analyses and ensures reproducible results, e.g. for the extraction of genomes from MG data [Karst et al., 2016] or for the processing and assembly of MT and MG reads [Narayanasamy et al., 2016] (**Figure 1.5**). Computational pipelines can be complex and often include a plethora of different software tools. However, results can differ depending on software versions or computational environments. Therefore, achieving reproducibility is a challenging task, especially for complex pipelines. Maintenance and availability of software tools, as well as installation, given different computational platforms and dependencies, can be challenging [Belmann et al., 2015]. Encapsulating these applications in closed environments or containers can provide a solution to greatly enhance reproducibility [Belmann et al., 2015; Narayanasamy et al., 2016]. For the assessment of gene functions, computational predictions are commonly applied in automated pipelines [Seemann, 2014], while recently also machine learning approaches are applied for functional annotation [Farrell et al., 2018].

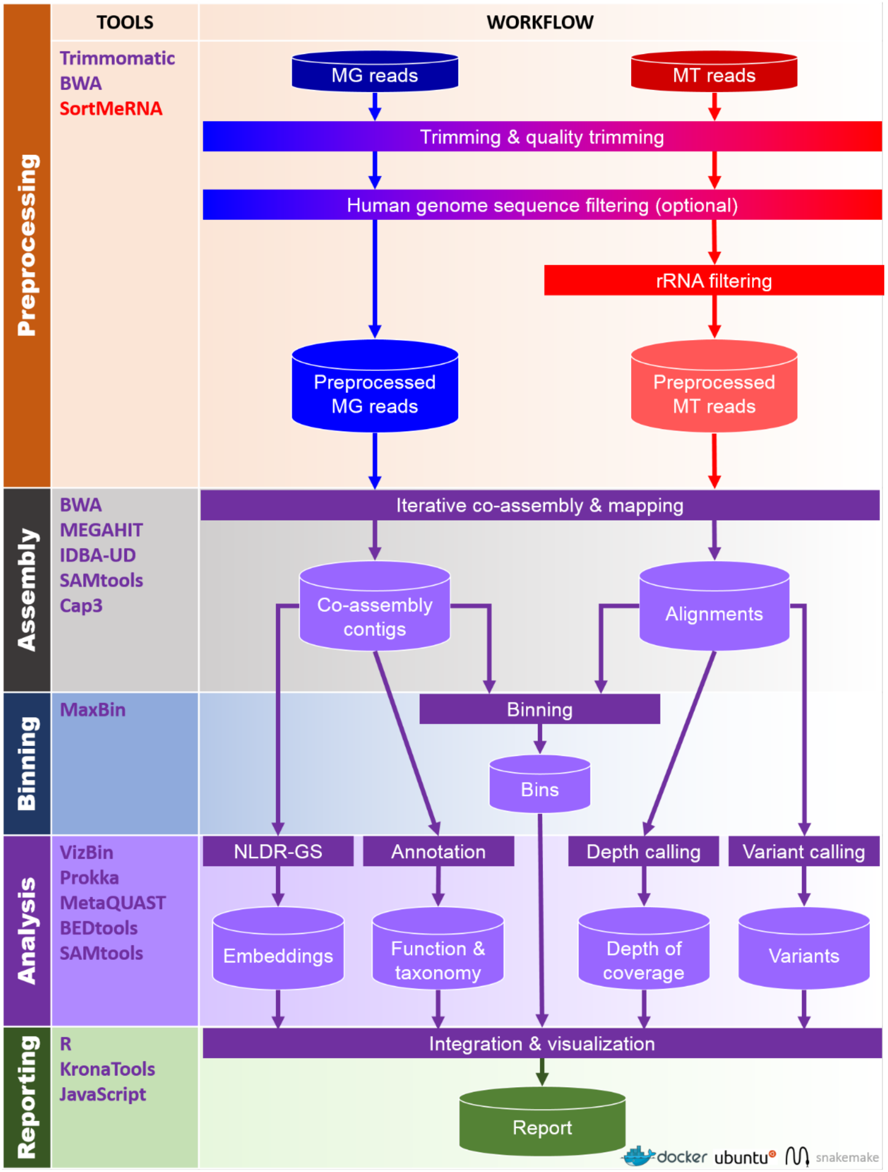


Figure 1.5: The IMP pipeline workflow schematic. The integrated metaomics pipeline is a complex pipeline for the co-assembly of MT and MG reads utilizing various computational tools (left bar) at different processing steps. Cylinders represent input and output while rectangles represent processing steps. NLDR-GS: genomic signature non-linear dimensionality reduction. Individual steps and data derived from or utilizing MG reads are labelled in blue, MT reads in red, and steps combining MT and MG data are coloured in purple respectively. (from Narayanasamy et al. [2016]).

1.4 Objectives of this work

Biogeochemical cycles on a global scale as well as micro-environments are shaped by microbial populations. Yet, methods to understand these complex ecosystems often remain at the level of community structure. Moving to the level of expressed or active functions could accent a more fine-grained picture of the interactions within microbial systems. However, the integration of heterogeneous large-scale datasets is challenging and methods have not been well established. Informed manipulations of microbial ecosystems e.g. for the optimization of an underlying biotechnological process will only be possible with an in-depth understanding of microbial systems ecology. In this context, the work at hand highlights the benefits of multi-omics data integration. The aim was to reconstruct microbial niches of distinct populations by detailing their genomic potential and expression of specific functions of relevance in a biotechnological context. Two model systems were analysed to emphasize the different levels of information that can be obtained by multi-omic data integration.

1.4.1 Genomics, transcriptomics, and proteomics of synthetic acidophile communities

Utilizing defined cultivable isolate strains offers the possibility to sequence complete genomes, while also being able to test a specific hypothesis in a controlled setting. In **Chapter 2** this is demonstrated by the analysis of defined communities of acidophile bacteria in the context of bioleaching of chalcopyrite, a copper mineral. The example of *L. ferriphilum* showcases the recovery and characterisation of an isolate genomes allowing a complete and detailed characterisation of its genomic potential. Changes in the lifestyle of the strain in the presence of the mineral could be observed. Furthermore, the interactions between the different organisms were assessed in defined co-cultures. To determine active processes in combinations exhibiting efficient copper solubilization, transcriptomic data was integrated using reference genomes and incorporating extensive gene annotations of iron- and sulfur-oxidative processes. Finally, a quorum sensing system detected in *L. ferriphilum* and its role in biofilm formation and dispersal was characterized.

1.4.2 Integrated metaomics of an *in situ* time-series of wastewater sludge samples for understanding microbial niche ecology

While synthetic consortia potentially allow a more detailed characterisation of functional traits and responses in defined conditions, they might not necessarily be representative of naturally occurring microbial communities. To characterise microbial function in an *in situ* scenario, a time-series of lipid accumulating organisms from a wastewater treatment plant was analysed in **Chapter 3**. The aim was to *de novo* reconstruct population-level genomes to track and characterise the microbial populations over time. Through the use of multi-omics (MT, MG, MM, MP) data a detailed view on microbial niche ecology in this system was obtained. Distinct lifestyle strategies of populations could be highlighted alongside shifting resource availability. The response to free long-chain fatty acids (LCFAs) in the system was further assessed by incorporation of expression data from microcosm experiments to measure the short-term response of the organisms. Potentially understanding the complex interactions within the floating sludge system, could allow for targeted interventions promoting a lipid accumulating phenotype in order to produce biofuel.

CHAPTER 2

REFERENCE-BASED OMICS ANALYSIS IN THE CONTEXT OF CHALCOPYRITE BIOLEACHING

Parts of this chapter are based on the following peer-reviewed publications:

- Stephan Christel[†], **Malte Herold**[†], Sören Bellenberg, Mohamed El Hajjami, Antoine Buetti-Dinh, Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, Mark Dopson (2017). Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species *Leptospirillum ferriphilum*^{T}.

- *Applied and Environmental Microbiology* **84**: e02091-17. [**Appendix C.1**]

- Sören Bellenberg[†], Antoine Buetti-Dinh[†], Vanni Galli, Olga Ilie, **Malte Herold**, Stephan Christel, Mariia Boretska, Igor Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, Mark Dopson (2018). Automated microscopical analysis of metal sulfide colonization by acidophilic microorganisms
  Applied and Environmental Microbiology **84**: e01835-18. [**Appendix C.4**]

Parts of this chapter are based on the following publication submitted for peer-review:

- Stephan Christel, **Malte Herold**, Sören Bellenberg, Antoine Buetti-Dinh, Mohamed El Hajjami, Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, and Mark Dopson (2018). Weak Iron Oxidation by *Sulfobacillus thermosulfidooxidans* maintains a favorable redox potential for chalcopyrite bioleaching.
  Frontiers in Microbiology **in review**. [**Appendix C.3**].

[†] Co-first author

2.1 Abstract

Heap bioleaching is a an industrial technique for recovering metals from ores by utilizing bacteria that also are relevant in the context of acid mine drainage. For chalcopyrite, a copper iron sulfide mineral of economical interest, heap bioleaching is characterized by a lag-phase of reduced leaching efficiency.

In this work, defined, low-complexity communities of acidophile organisms that are typically dominant members of bioleaching communities were studied. These strains obtain energy by oxidising iron and sulphur in the ore, solubilizing metal cations. A key step in this process is also the formation of biofilms on the mineral surface. To gain insights on iron- and sulfur oxidation pathways, as well as biofilm formation, (meta)transcriptomic and (meta)proteomic data derived from axenic cultures or mixed communities with up to three strains were analysed and supplemented with imaging data and metal solubilization rates.

As high-quality reference genomes were only available for two other strains, the genome of the typestrain of *L. ferriphilum*, a predominant iron oxidiser, was resequenced. Functional capabilities of the strain and their expression in chemostat and bioleaching cultures were assessed in detail. Surprisingly, analysis of mixed cultures revealed that most efficient release of copper was achieved in cultures without *L. ferriphilum* present. In these cultures *S. thermosulfidooxidans* was the primary iron oxidiser, while these cultures did not exhibit elevated redox potentials. Additionally, a quorum sensing system that was initially found to be expressed in *L. ferriphilum* and putatively also in the other strains was analysed. A short effect, i.e., biofilm dispersal could be observed after the addition of a signalling factor targetting the system.

The findings described here, especially those with regards to increased copper solubilization at low redox potentials, could have implications towards chalcopyrite leaching and warrant additional testing in a larger setting closer to the industrial application.

2.2 Background

2.2.1 Biomining

Bioleaching, or in more general terms biomining, refers to the mobilization of metal cations from insoluble ores by exploiting microbial oxidation and complexation processes [Rohwerder et al., 2003]. In recent years, there has been an increasing demand for metals while more and more low-grade ores are being processed in mining operations. Furthermore, environmentally friendly techniques should be developed, as the traditional roasting and smelting of metal ores is high in energy consumption and toxic compounds such as sulphur dioxide are released in the process [Azua-Bustos and González-Silva, 2014]. Biomining is suggested to be low in initial operational investment costs and therefore suitable also for low grade ore processing [Azua-Bustos and González-

Silva, 2014].

An important application of biomining is in copper bioleaching from sulphidic minerals which accounted for 8% of primary copper production world-wide in 2010, but also estimated as high as 20% [Schippers et al., 2011]. Mainly, copper bioleaching is carried out in engineered heaps (**Figure 2.1**) in Chile [Schippers et al., 2011]. Commonly, sulphide minerals are leached in an acidic environment, where the microbial populations catalyse the regeneration of ferric ions by oxidation of ferrous ions and oxidation of reduced sulphur species while further acidifying the environment. The same organisms and processes that drive biomining also occur in acid-mine drainage (AMD) [Baker and Banfield, 2003; Valenzuela et al., 2006], acidic discharge rich in heavy metals that is caused by mining activity when sulfide minerals are exposed to water and atmospheric oxygen. While abiotic corrosion occurs as well, microbial activity plays a significant role in AMD [Baker and Banfield, 2003]. As AMD is a serious environmental concern, confinement of the acidic and metal enriched leachate is a necessity in heap leaching operations.

Bioleaching is also carried out in tanks [Coram and Rawlings, 2002] with the advantage of controlled confined, controlled conditions. Aside from copper also nickel, cobalt, and zinc are commonly leached from sulphide ores [Rohwerder et al., 2003]. Additionally, oxidation of sulphide minerals is applied before cyanide treatment in processes to recover silver or gold [Rohwerder et al., 2003]. Another important application is in the recycling of metal-containing waste [Gabor et al., 2018]. Furthermore, also reductive dissolution techniques exist [Johnson and du Plessis, 2014].

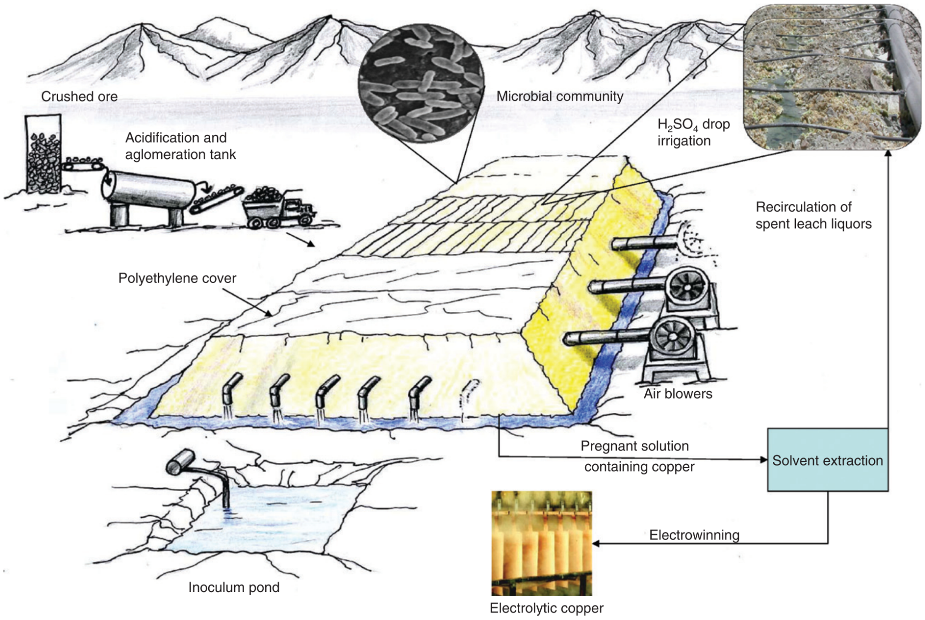

Crushed ore
Microbial community
Acidification and aglomeration tank
H_2SO_4 drop irrigation
Recirculation of spent leach liquors
Polyethylene cover
Air blowers
Pregnant solution containing copper
Solvent extraction
Electrowinning
Inoculum pond
Electrolytic copper

Figure 2.1: Illustration of heap bioleaching. Simplified illustration of an engineered heap for bioleaching. Crushed ore is piled onto a heap that is continuously irrigated with an acidic solution. The microbial communities can either be naturally occurring or inoculated. As processes depend on oxygen levels the heap can be aerated. Flow-through or leachate is enriched in metal cations and captured. (from Jerez [2011]).

2.2.2 Acidophile communities for bioleaching and model strains

Bioleaching of copper is of great economical importance (**Section 2.2.1**). Chalcopyrite ($CuFeS_2$) is the most abundant copper mineral in the world, however efficient chalcopyrite dissolution in low-cost bioheaps is challenging under mesophilic and moderately thermophilic conditions [Watling, 2006]. It has been suggested that chalcopyrite bioleaching is impeded by the formation of a passivation layer on the chalcopyrite surface by sulfur compounds and iron sulphate precipitates [Watling, 2006], however the formation of a passivation layer is still a topic of debate [Khoshkhoo et al., 2014]. In bioleaching applications for chalcopyrite a characteristic lag-time has been described before metal solubilization rates increased [Riekkola-Vanhanen, 2013].

The dissolution of chalcopyrite and other metal sulfides is driven primarily by proton attack and by oxidation by ferric ions with stepwise oxidation of sulfur compounds to sulfate [Rohwerder et al., 2003]. While the process of ore dissolution is driven by ferric ion in an acidic milieu and also occurs as an abiotic process, iron-oxidizing prokaryotes replenish the pool of ferric ions and thus lead to increased dissolution rates [Baker and Banfield, 2003]. Sulfur oxidizing prokaryotes maintain an acidic pH (**Figure 2.2**).

Biofilm formation on the chalcopyrite surface is a crucial step in the leaching process, as contact of mineral-oxidizing microbes with metal sulfides may significantly increase dissolution kinetics. This is at least partially due to glucuronic acid residues in the extracellular polymeric substances

$$\mathrm{CuFeS_2 + 4H^+ + O_2 \rightarrow Cu^{+2} + Fe^{+2} + 2S + 2H_2O} \tag{1}$$

$$\mathrm{4Fe^{+2} + 4H^+ + O_2 \overset{\text{Iron oxidizing Bacteria}}{\rightarrow} 4Fe^{+3} + 2H_2O} \tag{2}$$

$$\mathrm{2S + 3O_2 + 2H_2O \overset{\text{Sulfur oxidizing Bacteria}}{\rightarrow} 2SO_4^- + 4H^+} \tag{3}$$

$$\mathrm{CuFeS_2 + 4Fe^{+3} \rightarrow Cu^{+2} + 2S + 5Fe^{+2}} \tag{4}$$

Figure 2.2: Summary equations for chalcopyrite dissolution. Chalcopyrite is dissolved by protons (1) and Ferric ions (4). Microbial activity is important for replenishing the ferric ion pool by oxidation of ferrous ions (2) and by oxidizing sulfur species while mainting an acidic pH (3) (adapted from Pradhan et al. [2008]).

(EPS) that accumulate the oxidative agent iron(III)-ions, as has been shown for *Acidithiobacillus ferrooxidans* and *Leptospirillum ferrooxidans* [Rohwerder et al., 2003].

Three organisms commonly found in biomining or acid mine drainage environments [Baker and Banfield, 2003; Watling, 2016] and are often applied as model organisms for studying bioleaching were also of primary interest in this work (**Section 2.3.1**): *Acidithiobacillus caldus*, *Leptospirillum ferriphilum*, and *Sulfobacillus thermosulfidooxidans*. *A. caldus* is a moderately thermophilic, Gram-negative bacterium able to oxidise several sulphur compounds, such as tetrathionate, as well as elemental sulfur [Hallberg and Lindstrom, 1994]. *S. thermosulfidooxidans* is a gram-positive, moderately thermophilic organism known to be capable to oxidise ferrous iron and sulfur compounds [Norris et al., 1996]. Contrary to the other two organisms, it is not an obligate autotroph, but also hetero- or mixotrophic growth has been reported [Norris et al., 1996]. *S. thermosulfidooxidans* has a broad metabolic potential and is able to carry out diverse pathways (extensively characterised in Justice et al. [2014]).

L. ferriphilum plays a major role in acidic, metal-rich environments, where it represents one of the most prevalent iron oxidizers [Baker and Banfield, 2003]. It is an obligate aerobe that is capable of gaining energy only via ferrous iron (Fe^{2+}) oxidation [Coram and Rawlings, 2002; Rohwerder et al., 2003]. Although several *Leptospirillum* spp. have been identified and classified in four different groups, current knowledge of how they obtain energy and nutrients for growth is limited. In particular, mechanisms for nitrogen fixation have been under debate, as they has been described in *Leptospirillum* group III [Tyson et al., 2005], but only in some members of group II [Zhang et al., 2018]. Despite the perceived importance of *L. ferriphilum*, no complete genome sequence of the type strain of this model species is available, limiting the possibilities to investigate the strategies and adaptations that applies to survive and compete in its niche. Additionally, the understanding of how members of the leptospirilli survive at acidic pH lags behind that of other acidophiles.

In natural bioleaching microbial consortia, chemolithoautotrophic organisms are prevalent due to the limited accessibility of organic carbon [Baker and Banfield, 2003]. However, also interactions between heterotroph and autotroph organisms are important in these communities [Baker and

Banfield, 2003]. Increased leaching rates were observed in mixed cultures of acidophiles grown on arsenopyrite [Dopson and Lindström, 1999]. *A. caldus* could potentially confer a benefit to *S.thermosulfidooxidans* growth and leaching activity by oxidation of S^0 compounds or by providing benefits for hetero- or mixotrophic growth of *S.thermosulfidooxidans* [Dopson and Lindström, 1999]. Mixed cultures including the three model organisms in bioreactors containing a copper concentrate suggested that in a primary step sulfur oxidisers *A. caldus* and *S.thermosulfidooxidans* increase acidity in early stages of the processes, while in later stages iron-oxidation of primarily *L. ferriphilum* leads to increased metal solubilization [Hedrich et al., 2016]. However, as the analysis was limited to measurements of abundances for the organisms, it could not detail specific interactions.

2.2.3 Omics approaches to understand acidophiles

Microbial communities from AMD ecosystem have been of primary interest in the development and application of metaomic approaches due to relatively the low complexity of these consortia (**Section 1.1**). The recovery of population-level genomes from MG data was first described from samples of an AMD biofilm and allowed characterisation of metabolic relationships between the predominant *Leptospirillum* type II, to which also *L. ferriphilum* belongs, and the heterotroph archeon *Ferroplasma* [Tyson et al., 2004]. Furthermore, pioneering work for metaproteomics approaches has been performed within the same system describing a novel cytochrome central to iron oxidation [Ram et al., 2005]. Analyses based on FISH could resolve spatial and functional organisation of AMD biofilms [Wilmes et al., 2009]. Additionally, correlation patterns between the metabolome and the proteome allowed the characterisation of competition and niche breadth for distinct populations in AMD biofilms [Wilmes et al., 2010a].

Omics techniques with isolates cultures of bioleaching bacteria have contributed to a better understanding of the energy metabolism (i.e. iron and/or sulphur oxidation pathways) as well as their evolutionary relationships for example in *A. caldus* [Mangold et al., 2011]. Proteomics have elucidated functions of *A. ferrooxidans* involved in biofilm formation on pyrite, highlighting the role of carbon metabolism for enhanced EPS production and the role of yet unknown mechanisms in biofilm formation. Furthermore comparison of sessile and planktonic cells revealed increased levels of stress reponse related proteins in biofilm cells [Vera et al., 2013], an observation that has also been made in natural communities [Ram et al., 2005], as well as for other bioleaching organisms [Mangold et al., 2011]. Metal resistance systems in acidophiles have been characterised with omics techniques, e.g., leading the elucidation of copper resistance mechanisms in *A. ferrooxidans* by upregulation of RND-type Cus systems and indicating a role of rusticyanin that is also involved in iron oxidation [Almárcegui et al., 2014]. Metabolomics analyses have indicated novel polyamine-synthesis pathways in acidophiles and their potential role in biofilm formation [Martinez et al., 2015].

2.2.4 The SysMetEx consortium project

Results highlighted in this chapter stem from the participation in a consortium project, Systems Biology of Acidophile Biofilms for Efficient Metal Extraction (SysMetEx). The ERA-NET funded project ran from March 2015 to July 2018 and involved 6 partners from acadaemia and industry: Linneaus University (LNU, Sweden), University of Duisburg-Essen (UDE, Germany), TATAA Biocenter (Sweden), Ruhr University Bochum (Germany), Università della Svizzera italiana (Switzerland), and the University of Luxembourg (UL, Luxembourg). In short, the aim of the project was to improve chalcopyrite bioleaching by understanding dissolution processes and biofilm formation in defined cultures, focusing on omics analyses, leaching kinetics, high-throughput imaging data, and computational modelling. Culturing was performed at LNU with a focus on dissolution rates and UDE for imaging analyses. Biomolecular extractions and omics analyses were performed at UL. The work highlighted within this chapter focuses primarily on results that have been obtained from omic analyses so far. However, analysis of the dataset, generated within the project, is still ongoing.

2.3 Methods

2.3.1 Culturing of individual acidophiles and low-complexity communities

Three bacterial acidophile strains were utilized: *L. ferriphilum* DSM 14647 [Coram and Rawlings, 2002], *S.thermosulfidooxidans* DSM 9293 [Golovacheva and Karavaiko, 1978], and *A. caldus* DSM 8584 [Hallberg and Lindstrom, 1994]. These strains are the typestrains for their respective species. Prior to the bioleaching experiments, cells were maintained in three separate continuous cultures so that the cells were under the same growth state when all experiments were inoculated. The continuous cultures were maintained at 38 °C with MAC medium [Mackintosh, 1978], and electron donor were added in the form of 100 mM ferrous sulfate (*L. ferriphilum*) or 5 mM potassium tetrathionate (*S. thermosulfidooxidans* and *A. caldus*). For the collection of samples for biomolecular extraction, replicate 100-ml samples were taken from the chemostat cultures at least 3 days apart. To minimize RNA degradation, samples were rapidly cooled by mixing with 1 volume of ice-cold sterile MAC medium, and cells were immediately pelleted by centrifugation at 4 °C at 12,000 g for 15 min. The cells were then washed in 40 mL fresh, ice-cold MAC medium before being centrifuged again. Cell pellets were flash-frozen in liquid nitrogen, stored at -80 °C, and shipped on dry-ice.

Bioleaching experiments were conducted in quadruplets in 250 mL Erlenmeyer flasks with different combinations of strains. 100 mL MAC medium was supplemented with 2 % (wt/vol) chalcopyrite concentrate and inoculated with combinations of 10^7 or 10^9 cells per mL of the three bacterial strains, depending on the experiment. Cells were obtained by centrifugation from the continuous cultures (12,500 g, 20 min). Cultures were incubated at 38 °C under slow shaking (120 rpm) for

3, 7, or 14 days, depending on the experiment. Leaching cultures were separated into mineral-attached and planktonic cell sub-populations by centrifugation and resulting samples were handled as described above.

Metal sulfide dissolution was monitored by measurement of the concentration of iron(II) ions, total iron ions, and total copper ions using the spectrophotometric phenanthroline and bicinchoninic acid assays [Anwar et al., 2000], respectively. Precipitation of ferric salts was prevented by the addition of sulfuric acid to maintain the pH in the range 1.6 to 1.8.

Chalcopyrite ore for the bioleaching experiments was provided by Boliden AB (Sweden) and originates from the Aitik copper mine (N 67° 4' 24", E 20° 57' 51"). The flotation concentrate used in this study contained 29.5 % copper. The concentrate was sieved to obtain the size fraction between 50 and 100 µm and subsequently washed in three volumes of 0.1 M EDTA in 0.4 M NaOH for 10 min while stirring. Elemental sulfur was then removed from the surfaces by three iterations of washing with one volume of acetone. Finally, the mineral was dried at 60 °C overnight and then sterilized at 120 °C for 10 h under nitrogen atmosphere.

2.3.2 Imaging analyses and experiments with diffusible signalling factor addition

Experiments were performed as described in [Bellenberg *et al.* 2018 - **Appendix C.4**]. In brief, diffusible signalling family (DSF) compounds (cis-11-methyl-2-dodecenoic acid, DSF) were applied at 5 μM on axenic and mixed leaching cultures (**Section 2.3.1**) for testing their effects on cell growth and soluble substrate oxidation. To quantify the biofilm population attached to chalcopyrite grains, mineral grain particle samples were withdrawn from bioleaching cultures using a flame-sterilized spatula. Particles were incubated in 1 mL MAC medium (pH 1.8) and fixed formaldehyde at room temperature for 1 h. Cells were further incubated for 10 min in 200 μL of an aqueous solution of 0.01 % 4',6-diamidine-2-phenylindole dihydrochloride (DAPI) in 2 % formaldehyde. Prior to and after staining of attached cells, mineral grains were washed with PBS. Automated image acquisition was performed by high-throughput epifluorescence microscopy. Cell counting was carried out computationally by first converting the EFM images into gray-scale images and subsequent counting based on an computer vision technique. The mineral grain area was quantified from corresponding bright-field images with background illumination. Estimation of the chalcopyrite colonization in cells per gram was performed by multiplication of cell numbers of an image set with the specific surface area.

2.3.3 Biomolecular extractions

Isolation of genomic DNA for genome sequencing of *L. ferriphilum*

Cells were grown in continuous culture (**Section 2.3.1**) to late log phase before harvesting by centrifugation at 10,000 g for 10 min. DNA for sequencing was isolated by using the Genomic-tip

100/G extraction kit (Qiagen) according to the manufacturer's instructions, with the exception of a customized purification step recommended by the sequencing facility. Briefly, eluted genomic DNA was precipitated by the addition of isopropanol, immediately spooled by using a sterile pipette tip, and transferred to a microcentrifuge tube containing 70 % (vol/vol) ethanol for 2 min. Spooled DNA was then air dried, finally resuspended in 200 μL 0.1x Tris-EDTA (TE) buffer (pH 8), and allowed to dissolve for 72 h at room temperature.

RNA, and protein extraction from continuous culture samples

RNA and protein fractions were isolated from the planktonic sub-populations from bioleaching experiments. Cell pellets were subjected to biomolecular extractions based on a previously reported protocol [Roume et al., 2013], omitting step for metabolite extraction. In short, cell pellets were lysed by cryo-milling and bead beating followed by the spin column-based isolation of biomolecules with the Qiagen Allprep kit. For isolated RNA, an on-column DNAse digestion step was performed. Protein pellets obtained by precipitation were not dissolved in sample buffer but shipped in dried state for subsequent processing and measurement. Quality control for the isolated RNA (total RNA) fractions was performed on an Agilent bioanalyzer 2100. Samples were stored at -80 °C, and shipped on dry-ice for subsequent measurement.

RNA, and protein extraction from bioleaching culture samples

Isolation of the RNA and protein fractions were performed using a protocol adapted from [Vera et al., 2013]. Samples were washed with sulfuric acid (pH 2) and TE-buffer (pH 8). After removal of remaining liquid, 8 mL of pre-heated (67 °C) extraction buffer (50 mM sodium acetate, 2 mM EDTA, 2 %SDS, pH 5,5) was added. Samples were incubated in a water bath at 67 °C for 10 min, with vortexing for 10 seconds every minute. 8 mL of pre-heated acidic phenol (pH 4) was then added and samples were incubated at 67 °C for an additional 10 min, with vortexing every 2 min. Samples were cooled down and centrifuged for 7 min at 12,000 g to separate the phases. The aqueous phase was recovered for subsequent RNA isolation and the organic phase for protein isolation. The aqueous phase was washed with 8 mL of chloroform followed by resting the samples on ice for 10 minutes with occasional mixing intermittently. After centrifugation (7 min, 12,000 g) and removal of the organic phase, 0.5 vol isopropanol and 0.5 vol 1.2 M sodium chloride, 0.8 M sodium citrate solution were added for RNA precipitation for 1 h on ice. Following a centrifugation for 15 min at 16,000 g the supernatant was discarded and the RNA pellets were resolubilized in lysis buffer, pure ethanol, and water and purified with the RNAeasy kit (Qiagen) including on-column DNAse digestion. The organic phase recoved from lysis was used for protein precipitation. The samples were washed with 1 vol water at 67 °C and cooled on ice for 10 min. After addition of 1.5 vol cold acetone, samples were left at -20 °C overnight for precipitation. After acetone washing of the pellets, dried pellets were stored at -80 °C.

2.3.4 Nucleic acid sequencing

The obtained genomic DNA for *L. ferriphilum* was sent to the Science for Life Laboratory (SciLifeLab, Stockholm, Sweden) and sequenced by using two Pacific Biosciences (PacBio) single molecule real-time sequencing (SMRT) cells. Assembly was conducted with HGAP3 at the sequencing facility, including quiver for consensus corrections [Chin et al., 2013].

Ribosomal RNA was depleted from extracted total RNA samples with the Ribo-Zero rRNA Removal Kit for bacteria (Illumina, USA), except for nine initial continuous culture samples. RNA samples for were adjusted for equimolar concentrations prior to sequencing at SciLifeLab. Library preparation was performed with the Illumina TruSeq Stranded total RNA kit. Paired-end sequencing was performed on an HiSeq2500 instrument.

2.3.5 Reference genomes and functional annotation

The reference genome and functional annotation for *S. thermosulfidooxidans* DSM 9293 were obtained from the Joint-Genome Institute (JGI) Integrated Microbial Genomes (IMG) database [Markowitz et al., 2014] (Accession: 2506210005), while the reference genome and annotations for *A. caldus* DSM 8584 were obtained from NCBI Genbank [Clark et al., 2016] (Accession: GCF_000175575.2).

For *L. ferriphilum* DSM 14647 a newly sequenced genome was obtained (**Section 2.3.4**). The larger of two assembled contigs with overlapping ends could be circularized with `Circlator` [Hunt et al., 2015]. The `-fixstart` option was applied to set the *dnaA* gene as the first gene. The newly sequenced chromosomal contig was annotated with the `Prokka v1.12-beta` pipeline [Seemann, 2014], which included `Prodigal v2.6.3` [Hyatt et al., 2010] for the prediction of protein-encoding sequences. Functional annotation of coding sequences (CDS) was supplemented with a custom genus database with protein sequences of related genomes downloaded from the NCBI RefSeq database or the JGI Genomes Online Database (GOLD) [Mukherjee et al., 2017], consisting of the following genomes: *Leptospirillum* sp. group IV UBA BS (GOLD ID: Ga0053748), *Leptospirillum* sp. group II C75 (GOLD ID: Ga0039193), *L. ferrooxidans* C2-3 (RefSeq Accession.: AP012342), *L. ferriphilum* ML-04 (RefSeq Accession: CP002919), *L. ferriphilum* DSM 14647 (GOLD ID: Ga0059175), *L. ferriphilum* YSK (RefSeq Accession: CP007243).

Within `Prokka` predicted protein sequences were searched with `blastp` [Camacho et al., 2009] against the genus database, and annotations of best-matching hits were transferred. In a following step, the default databases in `Prokka` were searched and an e-value cutoff of $1e^{-9}$ was applied.

Additional functions annotations were transferred from in-house Hidden Markov Model (HMM) databases, including KEGG ortholog groups (KOs), PFAM, TIGRFAM, UniProt-enzymes, and

MetaCyc (additional details on the databases are described in Heintz-Buschart et al. [2017]). Furthermore, the annotation tool `Pannzer` [Koskinen et al., 2015] was applied. Functional categories were assigned based on the KO annotation (COG categories, KEGG pathways). Additionally, genes were assigned functional categories by manual curation, factoring in all of the automatically generated predictions and probability scores (**Appendix A.1** and **Appendix A.2**).

2.3.6 Proteomic analyses for *L. ferriphilum*

Pre-treatment and measurement of protein samples is described in detail in [Christel et al., 2017]. Proteins were identified with `Andromeda` [Cox et al., 2011] and quantified with the label-free quantification (LFQ)-algorithm embedded in `MaxQuant version 1.5.3.175` [Cox et al., 2014]. The FASTA protein database for identification was taken from the output of the functional annotation of the *L. ferriphilum* chromosome assembly or from existing databases (**Section 2.3.5**). After quantification, intensities from the LFQ normalization were filtered and compared with `Perseus v1.5.8.5` [Tyanova et al., 2016] removing rows with fewer than two values under either condition (mineral or continuous). The two conditions were compared with a two-sample Welch's t-test.

2.3.7 Data analysis

A custom pipeline was developed using snakemake [Koster and Rahmann, 2012] for processing and analysis of the transcriptome sequencing (RNA-seq) which involved the following steps. Raw reads for RNA sequencing were preprocessed with `Trimmomatic v0.36` [Bolger et al., 2014]. TrueSeq3-PE adapter sequences were removed using the following parameters: `seed mismatch:2; palindrome clip:30; simple clip:10; leading:20; trailing:20; sliding window: 1:3; minlen: 40; maxinfo: 40:0.5`. Quality control of raw and processed sequencing reads was performed with `FASTQC`. Preprocessed reads were mapped onto a concatenation of the three reference genomes (**Section 2.3.5**) with `Bowtie-2 v2.3.2` ([Langmead and Salzberg, 2012]) using default parameters. Reads mapping to protein coding sequences were counted with the `FeatureCounts, subread package v1.5.1` [Liao et al., 2014] with the `-s 2` parameter accounting for stranded reads. Read counts were then normalized with `DESeq2 v1.16.1` [Love et al., 2014], as well as separately convert to transcripts per million (TPM). Normalization was performed distinctly for the three organisms in a method adapted from Klingenberg and Meinicke [2017]. The pipeline used for the analysis of RNAseq analysis for *L. ferriphilum* is available in the following repository: `https://git-r3lab.uni.lu/malte.herold/LF_omics_analysis`.

The comparison between the previous draft genome of *L. ferriphilum*^{T} [Cardenas et al., 2014] and the newly sequenced genome was generated with `circoletto` ([Darzentas, 2010]), with default settings. The assessment of orthologue gene clusters in the different *L. ferriphilum* strains was done with `OrthoVenn` [Wang et al., 2015]).

2.3.8 Data availability

Sequencing data has been deposited at the European Nucleotide Archive (ENA).
Raw DNA sequencing data and the assembled genome for *L. ferriphilum*^{T} are available in Bioproject PRJEB21703. The annotated chromosome sequence is available under ID: LT966316.1 `https://www.ebi.ac.uk/ena/data/view/LT966316`.
Raw RNA sequencing data is available under the following Bioproject IDs:
PRJEB21842, PRJEB27815, and PRJEB27534.
Processed data and links to raw data are available for *L.ferriphilum* in a structured format under the following link: `https://doi.org/10.15490/fairdomhub.1.investigation.162.1`.

2.4 Results

2.4.1 Sequencing and characterisation of the *L. ferriphilum*^{T} isolate genome

While genome sequences were available for some strains belonging to the *L. ferriphilum* species, no high-quality genome of the *L. ferriphilum* type-strain was available (**Table 2.1**).
To obtain a genome sequence as reference for omics data integration the genome was resequenced. The sequencing and assembly of *L. ferriphilum*^{T} DNA from isolate culture yielded two polished contigs. Contig-1 spanned 2,569,357 bases with 574-fold depth-coverage, while contig-2 consisted of 41,141 bases with 33-fold depth-coverage.
Overall, 2,541 gene feature were predicted for the newly sequenced genome of which 2,486 were protein coding genes (**Table 2.2**). For 1,846 CDS, a functional annotation was transferred.
While contig-1 represented the circular chromosome of the organism (**Section 2.3.5**), initial assumptions that contig-2 could represent a plasmid sequence were not confirmed, as no characteristic plasmid genes could be detected. However, contig-2 could represents a putative phage sequence as predictions with `VIRSorter` [Roux et al., 2015] indicated. Additionally, a region on contig-1

Table 2.1: Available reference genomes for *L. ferriphilum*. Overview of available reference genomes for different strains of *L. ferriphilum*

Strain	Reference	NCBI RefSeq accession no.	State of the genome	No. of genes	Genome size (Mbp)	Coding density (%)
L. ferriphilum^T	[Cardenas et al., 2014]	NZ_JPGK00000000.1	Draft	2,366	2.41	93.1
Sp-Cl	[Issotta et al., 2016]	NZ_LGSH00000000.1	Draft	2,419	2.48	91.7
YSK	[Jiang et al., 2015]	NZ_CP007243.1	Complete	2,273	2.33	90.1
ML-04	[Mi et al., 2011]	NC_018649.1	Complete	2,475	2.41	90.3
DX	[Zhang et al., 2017]	NZ_MPOJ00000000.1	Draft	2,324	2.36	85.8
ZJ	[Zhang et al., 2017]	NZ_MPOK00000000.1	Draft	2,312	2.34	96.4

Table 2.2: General genome statistics for the L. ferriphilum^T genome. Counts of features where derived from the functional annotation.

Attribute	Value	% of total
Genome size (bp)	2,569,357	100.00
DNA coding region (bp)	2,331,855	90.76
DNA G + C content (bp)	1,392,384	54.19
Total no. of genes	2.541	100
No. of protein-encoding genes	2.486	97.84
No. of RNA genes (rRNA/tRNA/tmRNA)	6/48/1	0.24/1.93/0.04
No. of CDSs with functional prediction	1.846	74.25
No. of CDSs with assigned COG category	1.969	79.20
No. of CRISPR repeats	1	

with high similarity to contig-2 putatively could represent a prophage. Due to its low depth of coverage and undetermined origin, contig-2 was excluded from the following analysis, as it also could be a spurious contig originating from assembly errors.

The newly sequenced genome of *L. ferriphilum* DSM 14647 was compared to the previously available draft genome [Cardenas et al., 2014], revealing an additional 163,475 bp, thereby closing gaps in the previous draft (**Figure 2.3**). The most prominent gap with around 100,000 bp most likely had not been previously captured due to the presence of a clustered regularly interspaced short palindromic repeat (CRISPR) stretch. Additional functionalities in the newly identified regions were identified by inspecting the functional annotation (**Section 2.3.5**) and most prominently a cluster of *nif* genes (**Section 2.4.2**) was detected in the stretch that was missing from the previous draft genome ([Cardenas et al., 2014]). While a cluster of *nif* genes is not present in strains ML-04 and DX and the previous draft genome of the type strain, it is also present in strains Sp-Cl, YSK, and ZJ (see **Table 2.1** for an overview of *L. ferriphilum* genomes). Recently, a more detailed comparison of *L. ferriphilum* strains has been published, suggesting that the *nif* cluster is absent in strains ML-04 and DX due to gene loss [Zhang et al., 2018].

The genomes of these six *L. ferriphilum* strains, of which two are considered complete, were additionally used for comparison of orthologe protein sequences**Table 2.1**. All six genomes show a high degree of identity, with 1,769 orthologous gene clusters conserved in all six strains, with the newly sequenced type strain exhibiting the largest number of unique gene clusters (**Figure 2.4**). Many of the genes that are distinct for a particular strain seem to be related to insertions or deletions of mobile elements.

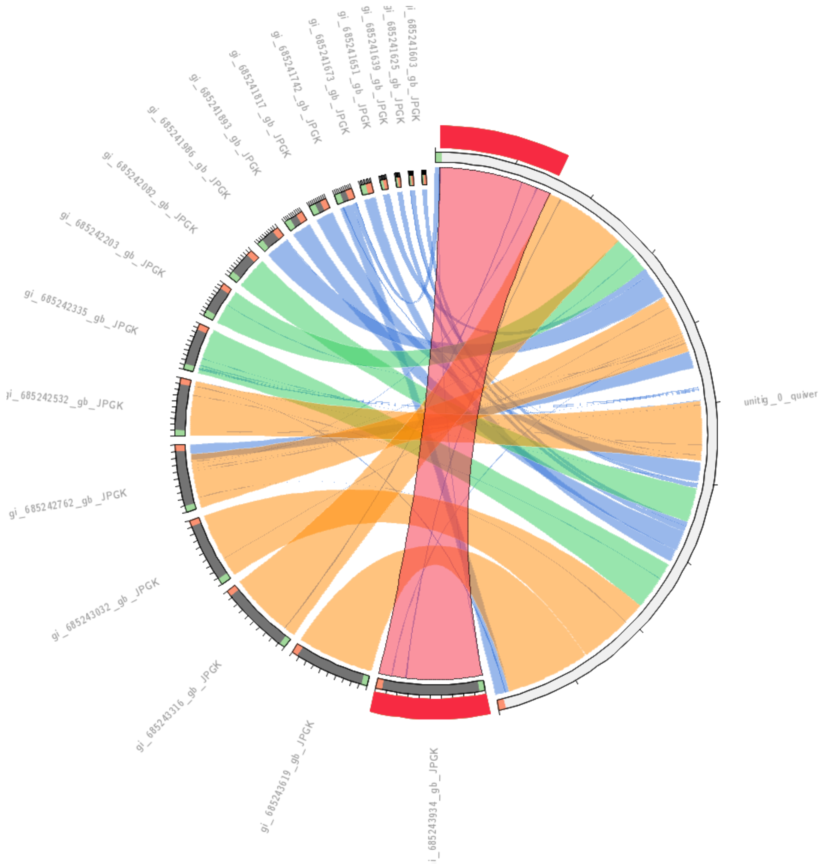


Figure 2.3: Comparison of the newly sequenced genome to the previous draft genome of the *L. ferriphilum* type-strain. A comparison between the new assembly contig-1 (white, right-hand side) and the 18 contigs of the draft genome [Cardenas et al., 2014] (grey, left-hand side). Coloured ribbons indicate ranked alignment scores (BLAST) which primarily reflect sequence length here, due to the high alignment scores overall.

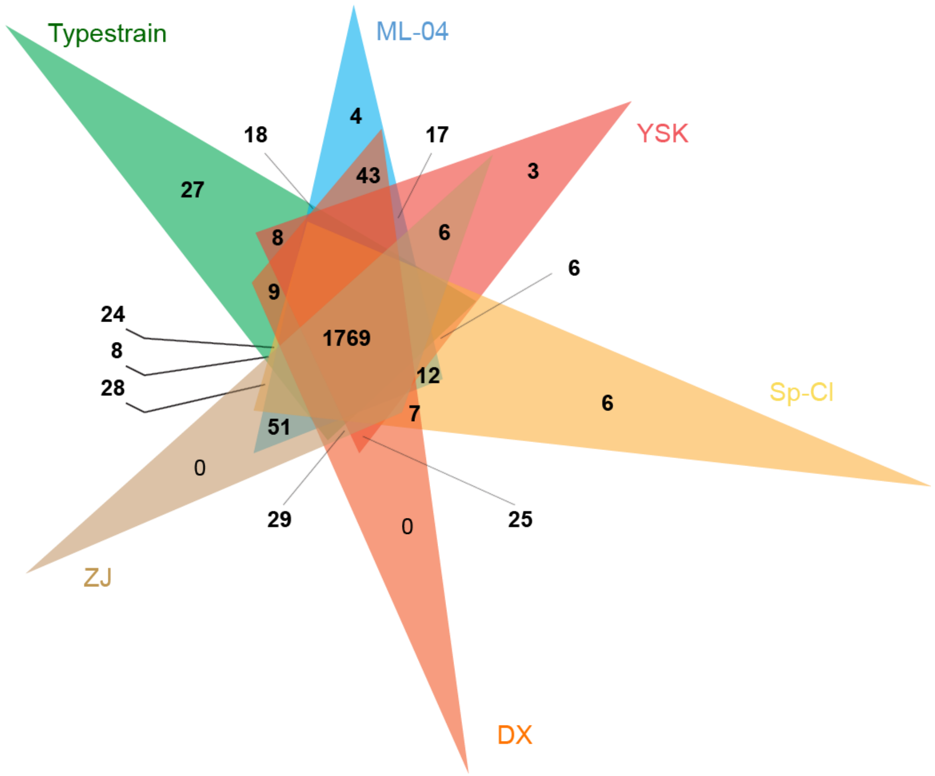


Figure 2.4: Venn diagram of orthologous gene clusters across *L. ferriphilum* reference genomes. Shown are the numbers of gene clusters that are shared between different strains (**Table 2.2**) or uniquely present. Typestrain refers to the newly sequenced genome of *L. ferriphilum* DSM 14647.

2.4.2 Detailed characterisation of functional capabilities of *L. ferriphilum*^{T}

Expressed functions in Fe^{2+} containing medium versus chalcopyrite bioleaching cultures were assessed by transcriptomic and proteomic analyses (**Figure 2.5**). A total of nine samples was analysed, 5 from continuous and 3 from bioleaching cultures (planktonic sub-population). Depending on the quality of extracted biomolecules not all RNA or protein samples could be used for measurements. While RNA was sequenced for 5 samples, proteomes were analysed for 8 samples (**Table 2.3**). RNAseq reads were quantified by mapping to the newly sequenced genome (**Section 2.3.7**), while for protein identification predicted protein coding sequences were used as search database. The resulting values are stated as transcripts per million base pairs (TPM) for RNA and LFQ-intensities for proteins, respectively (**Section 2.3.6**).

Overall, there was considerable overlap between the measured proteome and transcriptome (**Figure 2.5**), even though proteins could not be identified for all coding genes. Especially in the mineral containing samples fewer proteins were identified (**Table 2.1**). For the general characterisation of *L. ferriphilum*, omics data derived from the chemostat culture samples was used. A comparison between expression levels in the two conditions is described in **Section 2.4.3**.

In the following specific functional categories are outlined in greater detail with the full list of assigned genes available in file **Appendix A.1**. TPM and LFQ values listed for all genes alongside the full annotations used for manual assignment of functional categories are available in file **Appendix A.2**.

Table 2.3: Overview of proteomic and transcriptomic data for *L. ferriphilum*^{T} Read counts refer to counts of reads mapped to protein coding sequences of *L. ferriphilum*^{T}.

Sample	Culture type(s)	Total no. of RNA-seq read count	Median no. of RNA-seq counts	No. of proteins identified	No. of proteins with LFQ of > 0	Median LFQ
LNU-LXX9-Si00-CnA-P-B1	Continuous	1,034,434	181	NA	NA	NA
LNU-LXX9-Si00-CnA-P-B2	Continuous	NA	NA	1.698	1.241	160,595,000
LNU-LXX9-Si00-CnA-P-B3	Continuous	NA	NA	1.698	1.509	233,755,000
LNU-LXX9-Si00-CnA-P-B5	Continuous	NA	NA	1.698	1.4092	21,875,000
LNU-LXX9-Si00-CnA-P-B6	Continuous	1,284,834	219	1.698	1.412	212,595,000
LNU-LXX9-Si00-CnA-P-B7	Continuous	1,477,391	256	1.698	1.465	217,165,000
LNU-LXX9-Si00-14B-P	Batch, mineral	10,967,703	1.937	763	432	3,135,800
LNU-LXX9-Si00-14C-P	Batch, mineral	12,842,605	2.099	763	513	3,645,200
LNU-LXX9-Si00-14D-P	Batch, mineral	NA	NA	763	609	3,722,500

[a] NA, not applicable

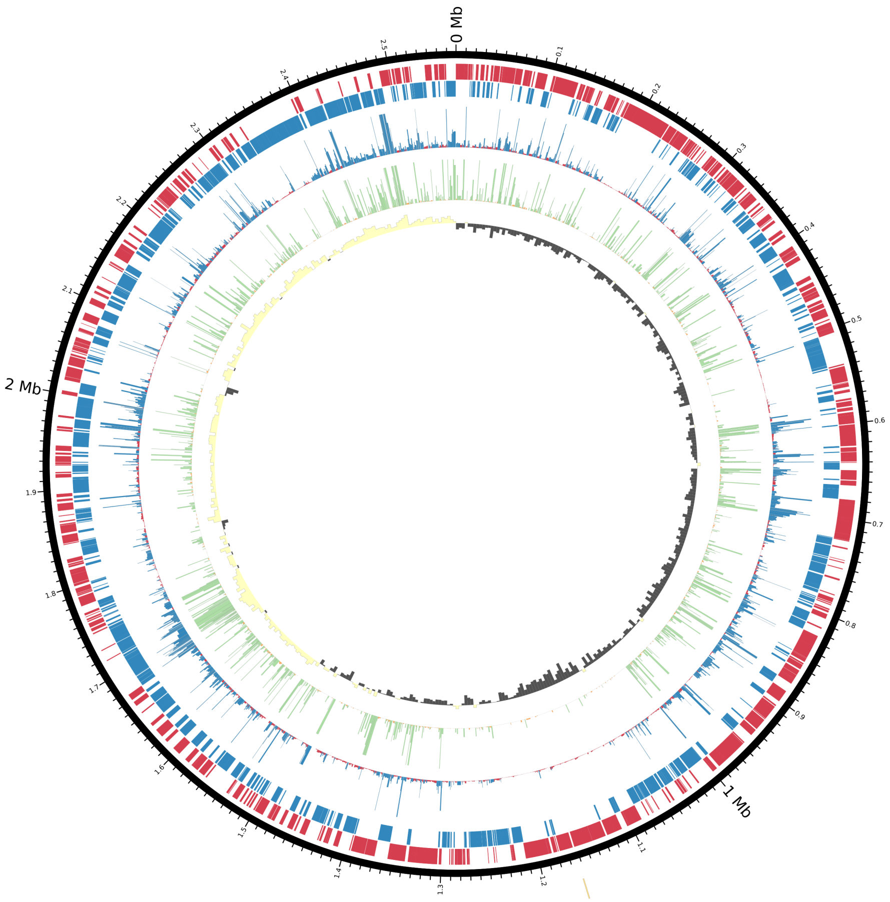


Figure 2.5: Circular representation of the *L. ferriphilum*^{T} genome. From the outside, the bands represent (i) the genome sequence; (ii) protein-encoding sequences on the positive strand (red); (iii) CDSs on the negative strand (blue); (iv) mean transcript expression (TPM), with a maximum of 2,000 TPM (blue indicates TPM values above the median, and red indicates values below the median); (v) mean scaled protein LFQ intensities, with a maximum of 2,000 (green indicates intensity above the median); and (vi) GC-Skew.

Energy conservation

The energy needs of *L. ferriphilum*^{T} are met exclusively by the oxidation of Fe^{2+} (**Figure 2.7**). Analogous to the iron oxidation system reported previously for *L. ferriphilum* ML-04, electrons from *L. ferriphilum*^{T} Fe^{2+} oxidation are transferred to electron carriers [Bonnefoy and Holmes, 2012], which were present in the genome in the form of cytochrome c, cytochrome $c_{551/552}$, cytochrome c_{553}, and cytochrome c_{544}. Thereafter, cytochrome cbb 3 oxidase can be used to directly reduce oxygen as a terminal electron acceptor [Pitcher and Watmough, 2004]. Alternatively, electrons can be used in reverse electron transport from cytochrome c to the quinone pool by the cytochrome b/c_1 complex. The resulting quinols can then be used to generate reducing power in the form of NAD(P)H via the NADH-quinone oxidoreductase (*nuoABCDEFHIJKLMN*) or the NAD(P)H-flavin reductase. Although their functionality is unknown, there are also three copies of subunit 5 of NAD(P)H-quinone oxidoreductase (*ndhF*), with which quinols could be used to produce NAD(P)H [Jünemann, 1997]. Finally, electrons from the quinol pool can be transferred to oxygen by using the cytochrome bd complex [Jünemann, 1997], which was also described for ML-04. Proton motive force generated by iron oxidation can be used for ATP generation by an F_oF_1-type ATP synthase (*atpABCDEFGH*).

RNA transcript counts of the genes involved in energy conservation indicated a preference for cytochrome $c_{551/552}$ (639 ± 26 TPM) compared to other cytochromes. However, this difference was not observed for the protein levels. While several genes of all cytochrome groups were only marginally transcribed and translated, no clear trend in the usage of cytochromes as initial electron carriers was apparent. Further electron transport was likely carried out via cbb 3 cytochromes to oxygen to create a membrane potential for the production of ATP. Although proteins of the competing reverse electron transport chain were expressed, with few exceptions, the pathway utilizing cytochrome cbb 3 had higher transcript counts and protein concentrations than did the pathway utilizing the cytochrome b/c_1 complex and the following quinone pool oxidoreductases.

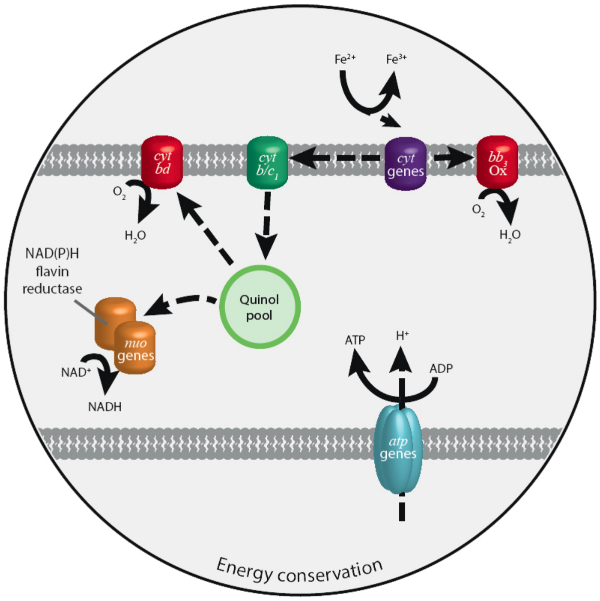


Figure 2.6: Model of energy conservation in *L. ferriphilum*. Solid arrows represent metabolic reactions, while dashed arrows indicate transport, the relocation of electrons or reaction products, and general regulative and metabolic interactions.

Carbon dioxide and nitrogen fixation

A single copy of the large-chain subunit of ribulose bisphosphate carboxylase (RubisCO) was encoded on the *L. ferriphilum*^{T} genome as well as on the genomes of other *L. ferriphilum* strains. However, all *L. ferriphilum* strains are suggested to fix carbon via the reductive tricarboxylic acid (TCA) cycle [Hügler and Sievert, 2011], for which all necessary genes were present on the genome. This was largely confirmed by transcript and proteome data, as gene products of the reductive TCA cycle were expressed and translated to a high extent (**Appendix A.2**). Although RubisCO (276 ± 14 TPM; LFQ $27,738 \pm 258$) exhibited low transcript counts, its protein concentration was comparable to the concentrations of proteins constituting the enzymes of the reductive TCA cycle.

The nitrogen demand of *L. ferriphilum*^{T} can be fulfilled by the fixation of elemental nitrogen by the nitrogenase complex *nifABDEHKNTUXZ* [Hoffman et al., 2014] and accessory protein genes. While having been reported for *L. ferriphilum* strains Sp-Cl and YSK, this gene cluster was not found in the reported *L. ferriphilum*^{T} draft genome [Cardenas et al., 2014] and likewise is lacking in the complete genome sequence of *L. ferriphilum* ML-04 [Mi et al., 2011]. Regulatory capabilities for the gene cluster are suggested to be fulfilled by a *nif*-specific regulatory protein in *L. ferriphilum*^{T}. Additionally, nitrogen can be taken up as nitrite by the nitrate/nitrite transporter *nasA*

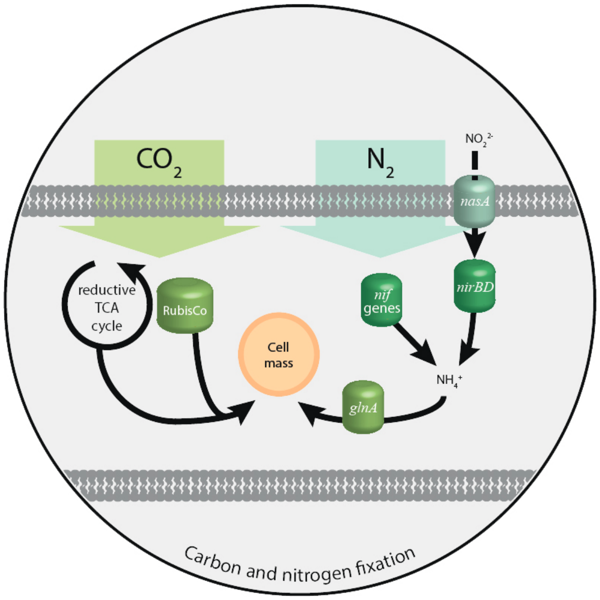


Figure 2.7: Model of carbon and nitrogen fixation in *L. ferriphilum*^{T}.

and assimilated in the form of ammonia by the nitrite reductase *nirBD* [Bykov and Neese, 2015], controlled by regulators of the NtrC and LysR families. RNA transcript analysis of nitrogenase subunits revealed very low counts, and most of the corresponding proteins were also not detected in the proteomic analysis.As the growth medium in this study was rich in ammonium, which can be taken up by the highly expressed glutamine synthetase, this was not surprising and has been reported for *L. ferrooxidans* [Moreno-Paz and Parro, 2006]. The highest transcript count within the nitrogen fixation clusters was that for *nifU* (1,997 ± 268 TPM; LFQ 1,228 ± 58), which is essential for the activation of the nitrogenase complex and is localized together with the cysteine desulfurase gene *nifS* [Agar et al., 2000]. NifS showed the highest protein concentration (661 ± 68 TPM; LFQ; 4,267 ± 175) in the nitrogen fixation clusters despite intermediate transcript counts. In combination with the high expression level of *nifU*, this could indicate an onset of nitrogenase formation due to early-stage ammonium starvation, supported by the intermediate expression of several nitrogen assimilation regulation proteins.

Quorum sensing and c-di-GMP

In Gram-negative bacteria, the regulation of genes encoding proteins for chemotaxis, motility, EPS production, and biofilm formation is often controlled by intracellular levels of the messenger molecule c-di-GMP [Hengge, 2009]. The presented genome sequence provides evidence for complex c-di-GMP metabolism, as is common for many Gram-negative bacteria (**Figure 2.8**). Specifically, the *L. ferriphilum*^{T} genome contains ten genes annotated as encoding putative diguanylate cyclases, thirteen genes encoding both diguanylate cyclase- and c-di-GMP phosphodiesterase -specific GGDEF and EAL protein domains, and two c-di-GMP-specific phosphodiesterases. Furthermore, four genes encoding HD/HDc domain-containing proteins and three genes encoding PilZ domain-containing c-di-GMP effector proteins were found. The latter genes were found in the context of genes annotated as being related to functions such as cellulose and extracellular polysaccharide biosynthesis and export. This suggests that c-di-GMP metabolism in *L. ferriphilum*^{T} also has an important function in the regulation of EPS production and biofilm formation. Several of these genes were expressed at the RNA and protein levels, including a c-di-GMP-specific phosphodiesterase class I-encoding gene, bifunctional diguanylate cyclase/c-di-GMP-specific phosphodiesterase -encoding genes, diguanylate cyclases, and a PilZ domain-containing protein.

Interestingly, the *L. ferriphilum*^{T} genome contains a gene cluster harbouring an *rpf* diffusible signal factor quorum sensing system, which is composed of the diffusible signal factor synthase-encoding gene *rpfF*, two genes encoding *rpfC* homologs annotated as genes encoding the Hpt domain-containing protein and signal transduction kinase, and the respective two-component system response regulator-encoding gene *rpfG*. In addition, further genes related to quorum sensing signalling were identified, such as three *luxR* family transcriptional regulator protein-encoding genes and another autoinducer binding domain-containing gene. The genes encoding the *rpf* quorum sensing system were found to be expressed at enhanced levels, while the orphan LuxR protein-encoding genes were found at very low RNA transcript or protein levels.

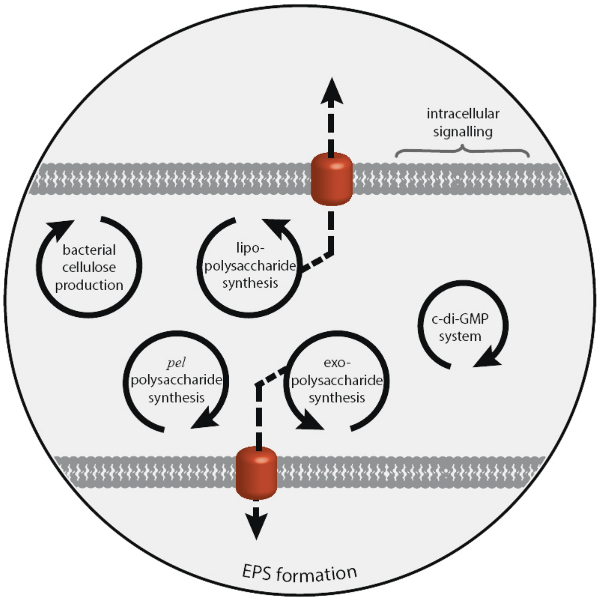


Figure 2.8: Model of biofilm formation and quorum sensing in *L. ferriphilum*^{T}.

2.4.3 Comparison of *L. ferriphilum*^{T} continuous and bioleaching cultures

Bioleaching experiments using pure cultures of *L. ferriphilum*^{T} achieved a significant dissolution of chalcopyrite (see [Christel et al., 2017] and **Figure 2.10**). To investigate important features and adaptation strategies of *L. ferriphilum*^{T}, RNA transcripts and proteins were grouped based on the functional categories established as described in **Section 2.3.5**. Comparison of continuous versus mineral culture samples revealed unexpectedly few differences in expression and translation patterns. In part, this is probably related to the controlled nature of the bioleaching experiments, where, e.g., the initial pH was 1.8 and did not decrease below 1.7, such that pH homeostasis systems seemed unaffected by the presence of chalcopyrite. Longer retention times and the presence of sulfur oxidizers would cause the pH to drop significantly [Watling, 2006]. Despite the remarkable tolerance of *L. ferriphilum*^{T} to high proton concentrations [Kinnunen and Puhakka, 2005], this would likely cause additional stress. Similarly, RNA transcript levels and protein concentrations for genes related to nitrogen fixation were found to be stable under the two conditions, conceivably as the culture medium contained large amounts of biologically available ammonium.

Among the differences observed between continuous and bioleaching cultures were decreased transcript counts related to ATP synthesis in the mineral samples along with bidirectional alterations of protein concentrations in ATP synthesis (**Figure 2.9**) and of specific cytochromes and cytochrome oxidases (**Appendix A.2**). This possibly indicated a shift of electron transport away from proton motive force and ATP generation toward the production of reducing power in the form of NAD(P)H

. However, this was not observable in NADH dehydrogenase RNA transcript counts. In contrast, the protein concentration related to NADH production was decreased in the bioleaching experiments (**Figure 2.9**). Additionally, RNA and protein analysis revealed slight reductions in the levels of proteins involved in both above-mentioned carbon fixation pathways when cells were grown on chalcopyrite (**Figure 2.9**). While the exact reasons for this are unknown, it could indicate a reduced demand for organic carbon, possibly caused by overall slow growth along with a reallocation of efforts for cell maintenance under stress conditions in mineral bioleaching cultures compared to active growth in continuous cultures.

Growth on minerals naturally comes with a heightened exposure of cells to heavy metals. Overall, transcript counts derived from metal resistance genes showed significantly increased levels during growth in chalcopyrite bioleaching cultures, in particular a strong enhancement of counts mapping to copper resistance systems (**Figure 2.9** and **Appendix A.2**). Surprisingly, protein concentrations appeared to be decreased. In-depth analysis revealed increased amounts of proteins belonging to the cus copper efflux system, underlining the strong detrimental effects of copper ions on microbes [Lemire et al., 2013]. Similar to the pH homeostasis response, as metal concentrations increase with time in natural or industrial systems, further upregulation of these systems should be expected.

L. ferriphilum^T was previously reported to rapidly attach to mineral surfaces [Noël et al., 2010], and RNA transcript counts of both chemotaxis and motility systems were revealed to be heavily enhanced during the bioleaching experiments. This was also observed for motility protein concentrations but not chemotaxis protein concentrations (**Figure 2.9** and **Appendix A.1**). The transcription and translation of c-di-GMP and EPS production remained at the same or lower levels in mineral culture samples (**Figure 2.9**). However, this may be explained by the fact that sampling of mineral-grown cells was conducted on the slowly agitated overlying medium and not the biofilm on the mineral grains, where most of the biofilm regulation and EPS production are expected to occur. In contrast, samples taken from the continuous culture were well mixed and likely contained both planktonic and detached biofilm cells.

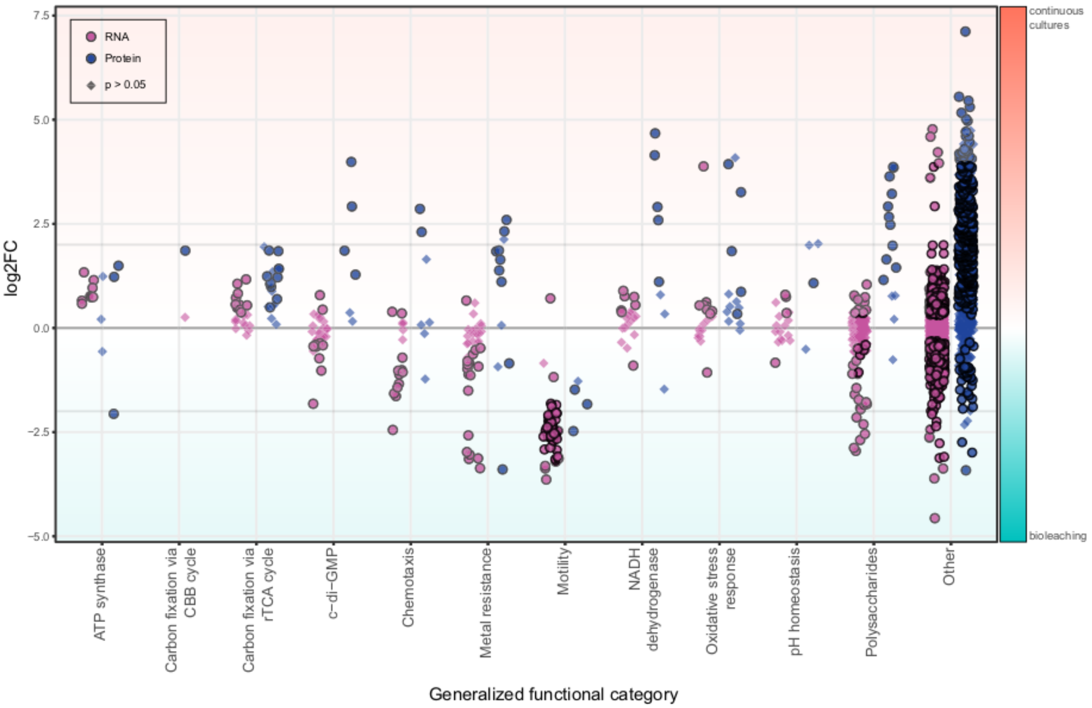


Figure 2.9: Differential abundance of transcripts and proteins in L. ferriphilum. Overview of differential expression (log 2 -fold change) of RNA transcripts and protein concentrations between continuous culture and chalcopyrite-containing bioleaching cultures. Data points represent single transcripts or protein signals. Circular symbols denote statistically significant differences ($P < 0.05$), while diamonds indicate statistically insignificant data. Manually assigned functional categories (**Appendix A.1**) are shown with some categories merged to aid comprehension:
nitrogen metabolism (ammonia and glutamate conversion to glutamine, nitrate/nitrite regulation, nitrite uptake and assimilation to ammonia, and nitrogenase genes), metal resistance (resistance to arsenic, cadmium/cobalt/zinc, copper, copper/silver, and mercury plus general metal tolerance), polysaccharides (cellulose production, extracellular polysaccharide production and export, and lipopolysaccharide synthesis), c-di-GMP (c-di-GMP effector proteins, with the EAL domain, proteins with the GGDEF domain, and proteins with both the EAL and GGDEF domains), and pH homeostasis (proton-consuming reactions, proton transporters, and role of potassium in internal positive membrane potential).

2.4.4 Efficient chalcopyrite leaching with favourable redox potential assessed in mixed cultures

Bioleaching of chalcopyrite was tested with single, binary, and tertiary combinations of the three model species (**Section 2.3.1**) plus uninoculated controls to investigate the effect of species composition on redox potential and copper release (**Figure 2.10**). To aid comprehension, these combinations will be abbreviated in the following using the initial letter of the included species, e.g. ‘ASL’ for the tertiary combination containing all species or ‘LS’ for the binary combination of *L. ferriphilum* and *S. thermosulfidooxidans* etc.

As expected, the single species mobilized less copper than mixed species, but unexpectedly, the tertiary combination ‘ALS’ was also outperformed by all binary combinations (**Figure 2.10**). This indicated that, in contrast to the currently accepted paradigm of inoculation of bioleaching applications with a broad mixture of biomining organisms, a well-chosen and defined mixture of microorganisms could benefit leaching efforts in the early stages of a bioleaching heap. Furthermore, the different combinations showed very distinct oxidation/reduction potential (ORP) profiles that, based on the present iron oxidizer(s), fell into one of two groups. All combinations containing *L. ferriphilum* had redox potentials between 650 and 680 mV compared to combinations in which it was excluded (i.e. ‘AS’ and ‘S’, showing ORPs below 550 mV).

In an attempt to elucidate the biological background for the difference in redox potential, both iron-oxidizing model species’ transcriptomic response towards each other was investigated (i.e. ‘ASL’ vs ‘AL’ for effect of *S. thermosulfidooxidans* on *L. ferriphilum*, and ‘ASL’ vs ‘AS’ for the vice versa effect; **Figure 2.11**).

Emphasis was placed on gene products related to energy metabolism, iron-, and sulfur oxidation **Table 2.4**). In bioleaching co-culture, *L.ferriphilum* remained largely unaffected by the presence of *S. thermosulfidooxidans*. Over its entire genome (2,486 genes), only 36 genes showed significant differential expression in response to *S. thermosulfidooxidans*’ presence. Among the 26 genes attributed to iron oxidation and electron transport, only three cbb 3 -type cytochrome c oxidase subunits (LFTS-01396, LFTS-02094, and LFTS-02276) exhibited significantly increased transcript numbers in the presence of *S. thermosulfidooxidans*, all of which have log2-fold changes below 1.5 **Table 2.4**). No genes involved in iron oxidation or electron transport had significantly higher numbers of RNA transcripts in the absence of *S. thermosulfidooxidans*.

In contrast, *S. thermosulfidooxidans* gene transcript numbers exhibited great variation depending on presence of *L. ferriphilum*. Of its 3,805 identified genes, 828 showed significant differential expression. Among the 83 selected genes involved in iron oxidation, electron transport, and sulfur oxidation, 55 showed significantly greater or lower RNA transcript numbers **Table 2.4**). Large variation was observed in genes related to iron oxidation. In contrast to, e.g., some members of the genus *Acidithiobacillus*, Sulfobacilli genomes lack the common iron oxidation protein rusticyanin [Guo et al., 2014]. Instead, Sulfobacilli are suggested to utilize sulfocyanin, which is also found

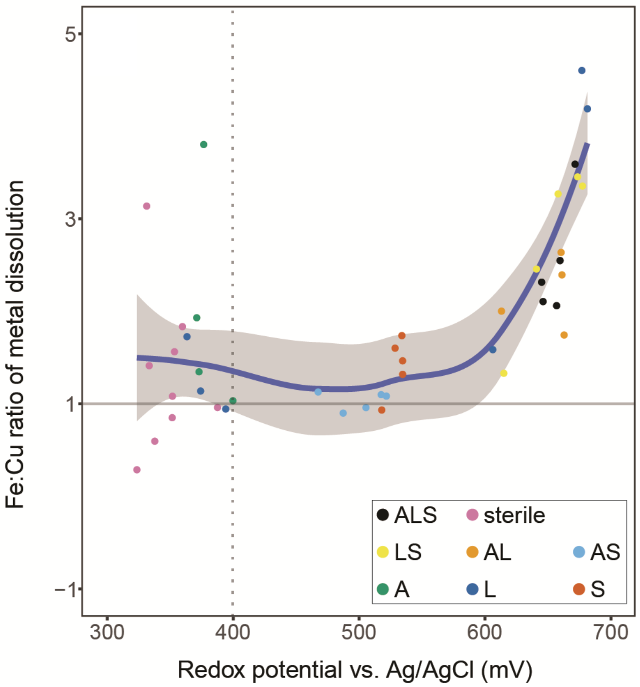


Figure 2.10: Chalcopyrite bioleaching and redox potentials in different combinations Ratio of released iron:copper versus redox potential during bioleaching of chalcopyrite concentrate with various combinations of the three model species. The ratio was calculated by dividing the amounts of the two metals that were released between two consecutive sampling points during the leaching experiment. The regression was calculated using to the LOESS method with 95 % confidence interval marked by the shaded area. The dotted line denotes the onset of microbial iron oxidation indicated by a redox potential above 400 mV. Abbreviations in the legend denote: A = *A. caldus*, L = *L. ferriphilum*, and S = *S. thermosulfidooxidans*. (from Christel et al. [2018] - in review; **Appendix C.3**).

Table 2.4: Transcriptomic changes in iron and sulfur metabolism related genes. Excerpt showing significant ($|log2FC| \geq 1.0, p \leq 0.05$) differential expression of *S. thermosulfidooxidans* genes related to iron and sulfur oxidation as well as electron transport. Negative log2-fold changes indicate higher transcript in presence of *L. ferriphilum* (ASL), positive changes upregulation in its absence (AS). Mean expression values are calculated from three independent experiments (n=3). Abbreviations: std, standard deviation; log2FC, log2-fold change. (from Christel et al. 2018 - in review; **Appendix C.3**)

Gene ID	Product	Deseq normalized expression				log2FC
		AS mean	AS std	ASL mean	ASL std	
Iron oxidation and electron transport chain						
Sulth_0051	Cytochrome *c* assembly protein	1086	91	2412	150	-1.15
Sulth_0119	Cytochrome *c* class I	250	72	105	37	1.25
Sulth_0449	Heme/copper-type cytochrome/quinol oxidase, subunit 3	5850	537	919	85	2.67
Sulth_0450	Cytochrome *c* oxidase subunit I	15675	2453	2857	266	2.46
Sulth_0451	Cytochrome *c* oxidase subunit II	15243	1526	4700	545	1.70
Sulth_0453	Sulfocyanin (SoxE)	7722	884	623	112	3.63
Sulth_0488	Cytochrome *c* oxidase subunit I	17287	3212	533	106	5.02
Sulth_0489	Cytochrome *c* oxidase subunit II	12771	1543	405	58	4.98
Sulth_0494	Cytochrome *d* ubiquinol oxidase, subunit II	161	11	57	38	1.50
Sulth_0495	Cytochrome *bd* ubiquinol oxidase subunit I	355	102	39	10	3.17
Sulth_0840	Cytochrome *c* oxidase, *cbb*$_3$-type, subunit III	557	173	81	30	2.78
Sulth_0843	Heme/copper-type cytochrome/quinol oxidase, subunit 3	154	20	24	6	2.67
Sulth_0844	Cytochrome *c* oxidase subunit I	431	29	35	14	3.62
Sulth_0845	Cytochrome *c* oxidase subunit II	228	4	30	6	2.93
Sulth_1456	Cytochrome *c* oxidase subunit II, periplasmic domain	86	11	43	8	1.01
Sulth_1490	Cytochrome *c* oxidase, cbb3-type, subunit III	76	22	22	6	1.79
Sulth_1513	Cytochrome *c* oxidase subunit II	15255	1578	4733	457	1.69
Sulth_1514	Cytochrome *c* oxidase subunit I	34803	3976	16822	1585	1.05
Sulth_1901	Cytochrome *c* biogenesis protein	442	46	999	136	-1.18
Sulth_1930	Cytochrome *c* oxidase subunit IV	408	92	4081	275	-3.32
Sulth_1931	Cytochrome *c* oxidase subunit III	507	64	4862	339	-3.26
Sulth_1932	Cytochrome *c* oxidase subunit I	1427	269	15277	462	-3.42
Sulth_1933	Cytochrome *c* oxidase subunit II	1764	452	17994	1718	-3.35
Sulth_2044	Cytochrome *c* class I	91	31	18	9	2.36
Sulth_2183	Cytochrome *c* biogenesis protein transmembrane region	291	108	816	311	-1.49
Sulth_2568	Cytochrome *c*-type biogenesis protein CcmE	123	36	47	3	1.37
Sulth_2572	Cytochrome *c*-type biogenesis protein CcmB	68	22	14	3	2.28
Sulth_2573	Cytochrome *c* assembly protein	114	15	12	7	3.33
Sulth_2730	Cytochrome *b*/*b*$_6$ domain	819	19	238	110	1.78
Sulth_2731	Cytochrome *b*/*b*$_6$ domain protein	2148	652	804	64	1.42
Sulth_2749	Sulfocyanin (SoxE)	9756	2642	1112	151	3.13
Sulfur metabolism						
Sulth_0921	Pyrrolo-quinoline quinone repeat-containing protein, tetH	613	101	25972	7210	-5.41
Sulth_0946	FAD-dependent pyridine nucleotide-disulfide oxidoreductase, Sqr_1	207	45	75	11	1.46
Sulth_1024	Hypothetical protein	125	50	57	8	1.13
Sulth_1025	Heterodisulfide reductase, subunit C, *hdrC*	24	0	50	8	-1.08
Sulth_1433	Sulfate adenylyltransferase	369	88	1430	384	-1.95
Sulth_1435	Sulfate adenylyltransferase	252	31	1202	289	-2.26
Sulth_1627	Sulfur oxygenase/reductase, Sor	591	29	1340	319	-1.18
Sulth_1878	Rhodanese-like protein	176	36	381	49	-1.12
Sulth_2076	Rhodanese-like protein	203	23	416	37	-1.03
Sulth_2172	Rhodanese-like protein	3024	1076	12511	2067	-2.05
Sulth_2770	Heterodisulfide reductase, subunit C, *hdrC*	11592	2924	29882	1777	-1.37
Sulth_2771	Heterodisulfide reductase, subunit B, *hdrB*	13422	4935	28681	4989	-1.10
Sulth_2782	DsrE family protein	4123	1767	14140	4441	-1.78
Sulth_3251	Pyrrolo-quinolinequinone repeat-containing protein, *tetH*	163	5	1551	223	-3.25

in the archaeal iron oxidizers of the genus *Ferroplasma* [Dopson, 2005]. In the presence of *L. ferriphilum*, *S. thermosulfidooxidans* showed strongly decreased transcript numbers attributed to two of the five *soxE* genes coding for this protein (Sulth-0453 and Sulth-2749). Additionally, the vast majority of identified cytochromes of all types exhibited decreased transcript counts, along with corresponding biogenesis proteins and quinol oxidases (**Table 2.4**). The strong downregulation of electron chain components that were likely linked to iron oxidation in *S. thermosulfidooxidans* could be explained by the fact that in cultures containing both iron oxidizers, the concentration of available ferrous iron was beyond the detection limit and likely too low for utilization by *S. thermosulfidooxidans*. This may be attributed to *L. ferriphilum* being able to scavenge Fe^{2+} at concentrations far below *S. thermosulfidooxidans*' capabilities and at large Fe^{3+} concentrations that exceed its inhibition limits [Rawlings et al., 1999].

Contrary to this overall trend, one cluster of *S. thermosulfidooxidans* cytochrome c oxidase subunits I-IV showed strongly increased transcript counts in the presence of *L. ferriphilum* (**Table 2.4**; Sulth1930-1933). In addition, two cytochrome c biogenesis proteins (Sulth-1901 and Sulth-2183) and one cytochrome c assembly protein (Sulth-0051) exhibited similarly increased transcript numbers. A direct role of cytochromes in iron oxidation has been suggested in an acid mine drainage biofilm and in *L. ferrooxidans* [Jeans et al., 2008; Blake and Griff, 2012]. Therefore, the strong opposite regulation of cytochrome oxidases in *S. thermosulfidooxidans* raises the question of their potential functional and/or structural differences. It could be possible that the oxidase exhibiting increased transcript counts in the presence of *L. ferriphilum* indirectly facilitates a higher affinity for ferrous iron, or has a lower sensitivity towards oxidative stress induced by accumulating ferric ion. Together with the upregulation of biogenesis and assembly proteins, this could enable *S. thermosulfidooxidans* to gain energy from ferrous ion in the presence of a stronger iron oxidizer. Nevertheless, as only cytochromes and cytochrome oxidases that are upregulated in absence of *L. ferriphilum* correlated with higher copper extraction, they may be of greater interest in the context of this study and should be considered in more detail in the future.

Genes coding for known sulfur oxidation proteins exhibited directionally opposite changes in transcript numbers compared to iron oxidation systems. Conceivably, this was to ensure sufficient supply of energy in a Fe^{2+} deficient environment and the vast majority of *S. thermosulfidooxidans* genes related to sulfur metabolism had significantly higher RNA transcripts in the presence of *L. ferriphilum* (**Table 2.4**). The highest of these log2-fold changes were recorded for two copies of tetrathionate hydrolase gene *tetH* (Sulth-0921 and Sulth-3251) while a third copy (Sulth-1188) exhibited moderately increased transcript counts in the absence of *L. ferriphilum*. *TetH* is responsible for the hydrolysis of tetrathionate, an important intermediate in sulfide mineral dissolution. Furthermore, thiosulfate can be oxidized by thiosulfate quinone oxidoreductase, encoded by a *doxDA* homologue. The two encoded copies of this gene exhibited increased transcript counts when in coculture with *L. ferriphilum*, although the log2-fold changes were low (Sulth-1989 and Sulth-1691).

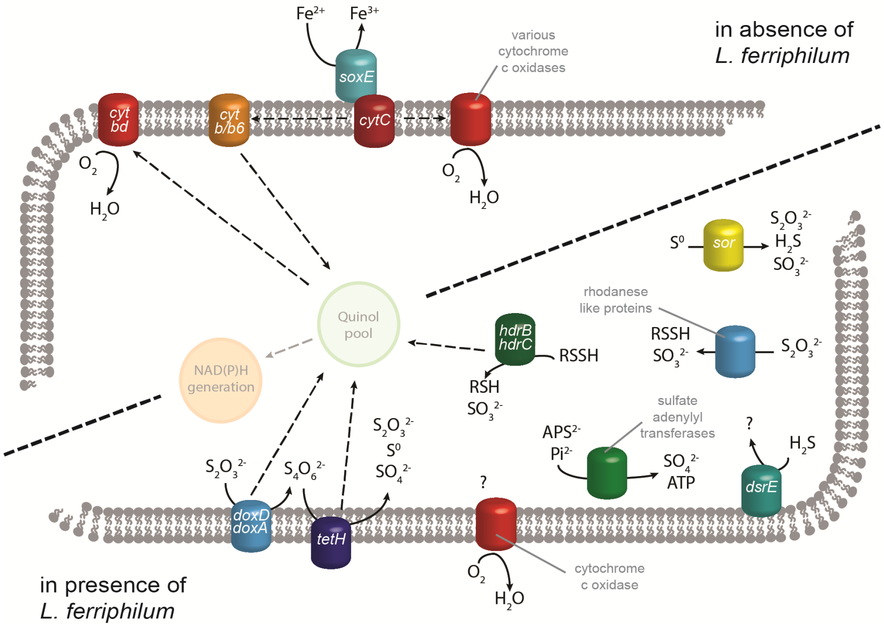


Figure 2.11: Proposed model of *S. thermosulfidooxidans* transcript regulation of genes related to energy conservation. Proposed model of *S. thermosulfidooxidans* transcript regulation of genes related to energy conservation, in cultures with (lower part) or without *L. ferriphilum* (upper part). Iron oxidation systems and electron transport by cytochromes have a greater number of RNA transcripts in the absence of the strong iron oxidizer *L. ferriphilum*. In its presence, *S. thermosulfidooxidans* instead has higher transcript numbers for genes contributing to inorganic sulfur compounds (ISC) oxidation plus one cytochrome c oxidase complex. Quinone pool and NAD(P)H generation are depicted translucently for comprehension, but corresponding genes were not analysed here (from Christel et al. 2018 - in review; **Appendix C.3**).

Additional contributions to sulfur oxidation systems include sulfate adenylyltransferase which is suggested to be involved in sulfite oxidation in *A. ferrooxidans* and *S. thermosulfidooxidans* strain ST [Guo et al., 2014] and is likely to fulfil the same role in *S. thermosulfidooxidans*^{T}. Similarly, DsrE-family protein (Sulth-2782) has been reported to be associated to oxidative sulfite metabolism [Dahl et al., 2005] and was also found to exhibit increased transcript numbers in presence of *L. ferriphilum* (**Table 2.4**).

2.4.5 Biofilm dispersal signals in mixed cultures

As described previously iron- and sulfur-oxidation mechanisms play an important role in chalcopyrite bioleaching. However, also attachment to metal ores can be a crucial factor in the dissolution of metal sulfides [Rohwerder et al., 2003]. To further elucidate interactions between the model strains and their role in biofilm formation, quorum sensing (QS) systems and ci-di-GMP metabolism were assessed. Specifically, the diffusible signalling factor (DSF) system found to be encoded by *rpf* genes in *L. ferriphilum* **Section 2.4.2** was of interest in this regard, as it was shown that DSF signalling molecules directly act on the ci-di-GMP metabolism [Deng et al., 2012]. Signalling molecules are sensed by the two-component system of the sensor kinase *RpfC* and the response regulator *RpfG* that activate a c-di-GMP-hydrolyzing phosphodiesterases encoded by *rpfR*. Low levels of c-di-GMP are typically associated with enhanced motility and decreased expression of biofilm-related genes [Romling et al., 2013].

Genes likely encoding DSF family signal-specific two-component systems or response regulators, suitable for DSF signal perception, were identified in the genomes of *A. caldus, L. ferriphilum*, and *S. thermosulfidooxidans* (**Table 2.5**). However, homologues were only found for *rpfR* and *rpfC* (*A. calddus* and additionally *rpfG*, but not for *rpfF*.

The genes of the DSF QS system were found to be expressed in transcriptome analyses of cells grown in continuous cultures, as well as in chalcopyrite batch cultures (**Figure 2.12**). Expression levels of *L. ferriphilum* strongly exceeded the average expression of gene transcripts of this species in axenic and and especially in binary co-cultures with *S. thermosulfidooxidans*. The DSF synthase LFTS-0514 was especially found to have elevated levels in the planktonic cell sub-opulations. Interestingly, expression of the *rpf* genes in *S. thermosulfidooxidans* showed the opposite trend. Overall, there is a slight trend for lower expression levels in the biofilm sub-population. However, due to the low number of replicates of the biofilm samples this will have to be confirmed in the future.

To further test the activity of the system a DSF signalling molecule was added to the cultures and numbers of attached cells were assessed by high-throughput EFM microscopy in combination with automated cell counting (**Section 2.3.2**). Biofilm dispersal was observed in cultures of *L. ferriphilum*, *S. thermosulfidooxidans*, and their combination in mixed cultures when 5μM DSF was added after 5 days of incubation (**Figure 2.13**). A similar effect was noted in mixed cultures of

Table 2.5: Presence of DSF family quorum sensing system-encoding genes. Locus-tags of genes for three model species that could be assigned to either *rpfF*, *rpfR*, *rpfC*, *rpfG* with blastp. (adapted from Bellenberg et al. [2018])

Species	***rpfF***	***rpfR***	***rpfC***	***rpfG***
A. caldus		ACAty_RS14920, ACAty_RS14615, ACAty_RS02860	ACAty_RS07245, ACAty_RS04080	
L. ferriphilum	LFTS_00514	LFTS_00511	LFTS_00515, LFTS_00516	LFTS_00517
S. thermosulfidooxidans		Sulth_1253, Sulth_1788, Sulth_2384	Sulth_1793	Sulth_2102

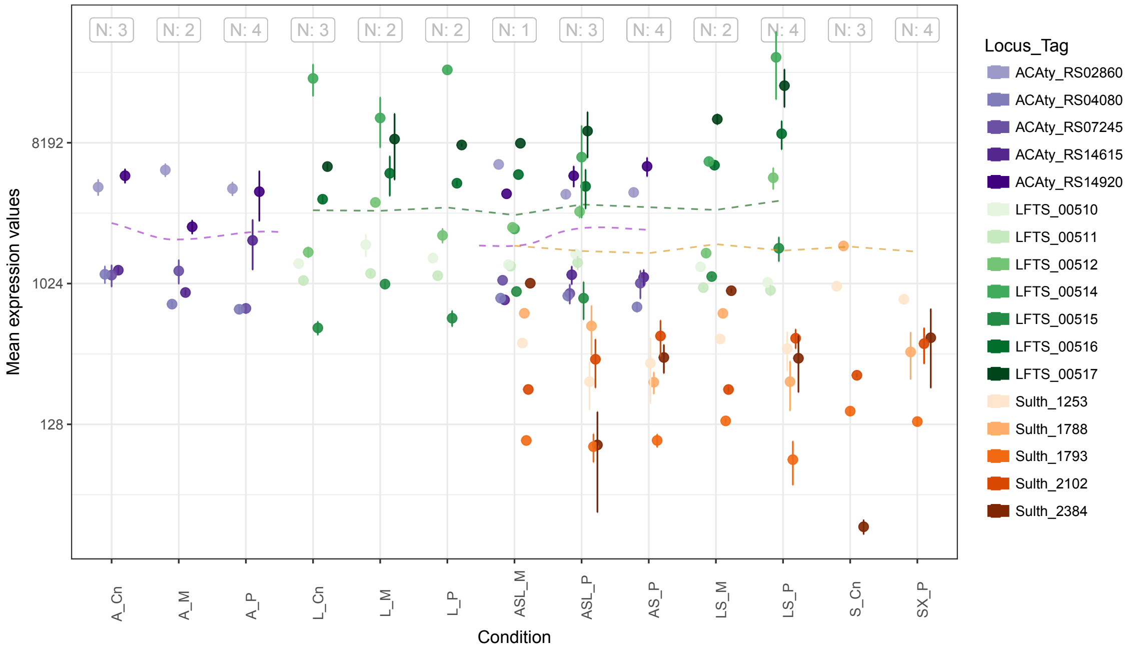


Figure 2.12: Expression of DSF system genes in different conditions and combinations. DESeq2 normalized expression values for genes related to the DSF quorum sensing system in the three model organisms (mean values across replicates). Error bars represent standard deviation across the number of biological replicates which is shown in the top. Dashed lines indicate mean expression levels across all genes per individual organism. Abbreviations used for the conditions (x-axis): A = *A. caldus*, L = *L. ferriphilum*, S = *S. thermosulfidooxidans*, Cn = continuous culture, P = planktonic sub-population, M = mineral attached sub-population

all three species, however no biofilm dispersal was observed in cultures of *A. caldus* ([Bellenberg et al., 2018]). Biofilm dispersal effects were short-lived, and recolonization of the chalcopyrite occurred in the batch experiment assays within 24 h after DSF addition. The addition of DSF to mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* caused a marked difference in the development of the sessile cell population, which was similar to the one observed in pure cultures of *L. ferriphilum* (**Figure 2.13A**).

The results highlight the effects of DSF family signal compounds in cultures of *L. ferriphilum* and S. thermosulfidooxidansand suggest an important role of these signal compounds in colonization of metal sulfides.

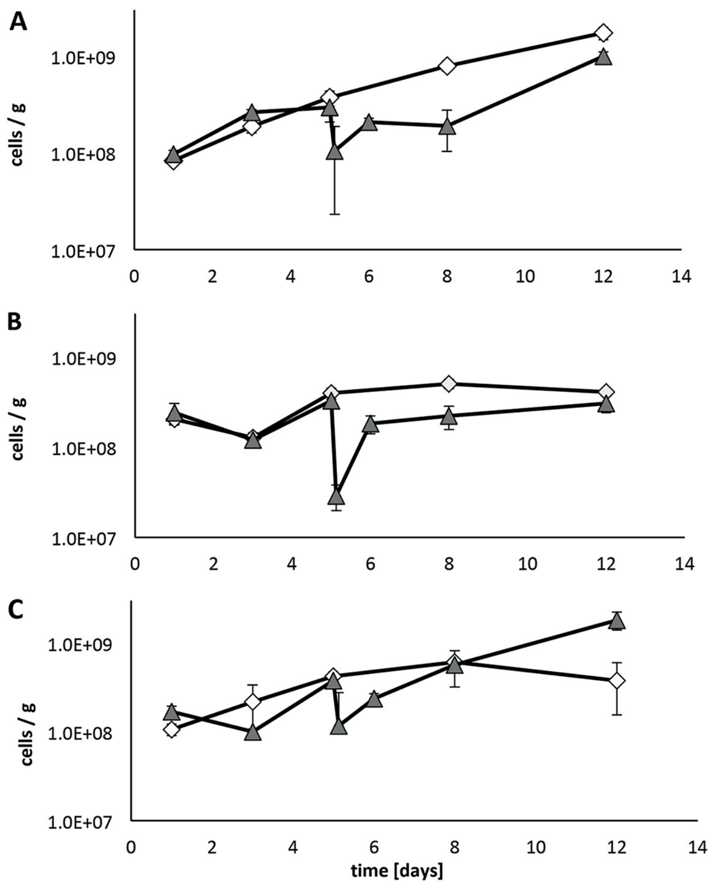


Figure 2.13: DSF molecules and their effect on attached cells. DSF molecules stimulate biofilm dispersal in *L. ferriphilum* and *S. thermosulfidooxidans*. (A to C) Axenic cultures of *L. ferriphilum* (A), *S. thermosulfidooxidans* (B), and mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (C) were cultivated with 2 % chalcopyrite. DSF (5 μM) was added after 5 days of incubation (gray triangles), and the mineral-attached cell population was compared to control experiments without DSF (white diamonds). Cells counts where determined by an automatic counting method and high-throughout microscopic analysis. (from Bellenberg et al. [2018])

2.5 Discussion and outlook

2.5.1 Functional omics provide an in-depth understanding of acidophile lifestyle in defined conditions

The newly sequenced genome of *L. ferriphilum*^{T} aided in the in-depth characterization of this organism's metabolic potential and provided the possibility to interpret its expression and translation behaviour in continuous and batch culture. PacBio long-read sequencing directly allowed the assembly of a circular chromosome and revealed key features of the adaptation of *L. ferriphilum*^{T} to acidic, metal-rich environments, associated with sulfidic minerals, in the environment as well as in industrial applications. Additionally, RNA transcript sequencing and protein identification elucidated stressing factors during chalcopyrite biomining and shed light on resistance systems deployed by *L. ferriphilum*^{T}. The data described in **Section 2.4.2** and **Section 2.4.3** pose a valuable resource for future experiments investigating the role of*L. ferriphilum*^{T} in acid mine and rock drainage as well as bioleaching processes.

However, the isolation of nucleic acids and proteins proved to be challenging, and only two RNA extracts and three protein samples of mineral origins were of sufficient quality for differential expression and translation analyses (**Section 2.4.3** and **Table 2.3**). Extraction of RNA and proteins from the biofilm sub-populations proved to be even more challenging, delayed data generation, and were largely excluded here. Owing to the lower sensitivity and dynamic range of the Orbitrap Elite instrument that was used for analysis of the bioleaching samples, fewer low-abundance proteins were quantified in these samples than in the continuous-culture samples that were analysed with an Q-Exactive HF mass spectrometer (**Section 2.3.6**). This manifested as an apparently higher expression level of such gene products in continuous cultures that could not be correct for with normalization.

During bioleaching of chalcopyrite concentrate in co-cultures, *S. thermosulfidooxidans* but not *L. ferriphilum* maintained a low redox potential that is favorable for the extraction of copper. It can be hypothesized that this was due to differences in affinity and/or effectivity of the species' respective iron oxidation systems, as well as the attachment rate of the microorganisms to the mineral grains. This finding could potentially contribute to overcoming passivation and improving dissolution rates in large-scale chalcopyrite bioleaching. Expression of iron and sulphur oxidation systems in *S. thermosulfidooxidans* were investigated during bioleaching experiments in presence and absence of *L. ferriphilum*. Presence of the strong iron oxidizer induced greatly decreased transcript counts attributed to iron oxidation and increased counts for sulphur oxidation. Analysis of this data revealed gene products potentially responsible for the difference in oxidation/reduction potential, which should be studied in this regard in the future.

While the notion of higher chalcopyrite dissolution rates at lower redox potentials by itself is not new [Watling, 2006], the role that the tested organisms play in the this regard has not been described. The concept of controlling *L. ferriphilum* levels thus achieving higher solubilization rates and lower redox potentials due to less efficient iron oxidation can be challenging to achieve in systems open to the natural environment (**Section 2.5.2**). It would have to be seen if inoculation strategies omitting efficient iron oxidisers could be sustainable.

The diffusible factor system detected in *L. ferriphilum* **Section 2.4.2** might be an interesting target for further studies elucidating if it can be applied as a biofilm dispersal agent. The presence and expression of DSF family genes in mixed cultures points to an important role in biofilm formation The impact of DSF signal compounds was highlighted for *L. ferriphilum* and *S. thermosulfidooxidans*, in mixed or axenic cultures, highlighting its role in biofilm formation. Potentially, inter-species signalling could play a role in maintaining a competitive advantage if attachment sites on ore surfaces, and thus electron donors required for energy conservation, are inaccessible. From the initial results described in **Section 2.4.5**, we can assume that a DSF-system is present and functional in *L. ferriphilum* and *S. thermosulfidooxidans* and that signalling molecules cause a transient dispersal effect. If exploiting DSF signalling was a means to control *L. ferriphilum* growth and biofilm formation this could have potential implications for biomining as well as AMD.

2.5.2 From defined conditions and consortia to *in situ* analyses

A key issue in translating findings from synthetic communities to improved biotechnological processes is being able to assess the applicability of the results in real world systems. Fitness and performance readouts for model organisms have been shown to vary substantially depending on whether they occurred in a synthetic or natural community [Yu et al., 2016].

From the results presented in this chapter, the hypothesis can be formulated that chalcopyrite bioleaching at low redox potentials, i.e., by weak iron oxidizers, would not be hampered by a lag-time for metal release onset or by a passivation effect preventing efficient solubilization. However, controlled conditions often do not reflect microbial activity *in situ*. For example controlled experiments using chalcopyrite concentrates might not be directly transferable to the complex gradients and interactions occurring in heap bioleaching [Watling, 2006]. The conditions within a bioleaching heap can vary tremendously in terms of oxygen and carbon dioxide gradients, as well as for pH and temperature [Pradhan et al., 2008]. Furthermore, the heterogeneity and impurity of the ore itself might have a great impact as well.

A scenario for testing the aforementioned hypothesis could be using a pilot-scale heap with defined inocula omitting efficient iron oxidisers, such as *L. ferriphilum*. However, it has to be seen if *L. ferriphilum* or other efficient iron-oxidisers would not soon dominate in a system open to the environment as they are ubiquitously found in bioleaching or AMD systems. To resolve spatial and temporal scales and to capture microbial community dynamics in a natural ecosystem reference-independent methods could potentially be applied.

CHAPTER 3

IDENTIFYING AND CHARACTERIZING FUNDAMENTAL AND REALIZED ECOLOGICAL NICHES OF MICROBIAL POPULATIONS IN OLEAGINOUS FLOATING SLUDGE

This chapter is based on the following manuscripts in preparation:

- **Herold** *et al.* - Defining fundamental and realized microbial niches using integrated time-resolved multi-omics
- Kleine-Borgmann *et al.* - Lipid accumulating bacteria from biological wastewater treatment plants: from isolation to *in situ* population dynamics and activity

3.1 Abstract

Contrary to an *in vitro* culture, establishing and maintaining a microbial community-based biotechnological processes with a desired phenotype can be challenging and is often attempted by tuning environmental parameters. However, conditions that favour microbial populations contributing to the desired community phenotype could be created by engineering niches. An in-depth understanding of niche ecology is therefore necessary to provide the basis for optimizing such biotechnological processes. Wastewater treatment with activated sludge is one of the most important biotechnological applications, however several avenues are currently being explored to increase sustainability in wastewater treatment operations. One aspect is the efficient accumulation of lipids by microorganisms found in oleaginous floating sludge that could potentially be harnessed for biofuel production. In this chapter, the use of multi-omics data to resolve ecological niches of distinct *de novo* reconstructed populations in this system is highlighted.

To characterize niches of distinct microbial populations, multi-omic (DNA, RNA, protein, metabolites) datasets from weekly samples of activated sludge floating islets over 14 months (long-term *in situ*) were analysed, as well as controlled bioreactor experiments performed (short-term *in vitro*). Using nucleic acid sequencing data, the community structure was determined and population-level genomes were reconstructed for several populations of interest with regards to lipid accumulation. Based on the populations' functional potentials, four groups were defined reflecting distinct types of fundamental niches. Realised niches were characterized over the time-series by linking the MP and MT data to the reconstructed genomes. Abiotic factors, e.g., free fatty acid levels, temperature, or amino acid levels were significantly associated with gene expression, particularly in relation to genes involved in lipid metabolism, thus highlighting the importance of these factors for future niche engineering. The results described in this chapter further deepen our understanding of microbial niche ecology within a biotechnological process, with potential applications beyond wastewater treatment.

3.2 Background

3.2.1 Wastewater treatment plant as model system for microbial ecology

Treatment of wastewater is an essential process in modern civilisation as it is vital for public health and the environment. Biological wastewater treatment plants (BWWTPs) utilize the activated sludge process that was first described over a century ago [Ardern and Lockett, 1914]. In the process, microorganisms remove pollutants such as excess carbon, phosphorous, or ammonium from the wastewater. These compounds are assimilated in microbial biomass or are converted to carbon dioxide, methane, or nitrogen gas. Energy requirements for the operation of BWWTPs are substantial, e.g. they make up to 3 % of global electricity consumption for treatment of domestic wastewater alone and also 5 % of non-carbon dioxide greenhouse-gas emissions [Li et al., 2015]. The chemical energy contained in wastewater is high, however, in its current form, the process is not used optimally, as only a fraction of the chemical energy may be recovered by microbial fuel cells or anaerobic digestion of sludge [Sheik et al., 2014].

Even though the microbial processes in BWWTPs are vital for efficient operation, the microbial diversity in activated sludge has not been well captured by traditional culturing or microscopy-based methods [Wagner and Loy, 2002]. BWWTPs have historically been used as model systems to study microbial ecology and recently significant advances have been made with contemporary molecular methods [Daims et al., 2006]. BWWTP functioning highly depends on the complex interplay of the microbial populations, for example between nitrifiers, ammonia and nitrite oxidiers, and denitrifiers that respire nitrate, or interactions between phosphate and glycogen accumulating bacteria [Wagner and Loy, 2002]. Especially nitrogen cycling in BWWTPs has been the focus of several studies, leading to, e.g., the discovery of organisms capable of performing anaerobic ammonium oxidation (anammox) [Mulder, 1995; Strous et al., 2006]. More recently, also the conversion of ammonium to nitrate within a single organism (comammox) has been shown, challenging the prior assumed division of labour between ammonia and nitrate oxidizing microorganisms in nitrification [van Kessel et al., 2015; Daims et al., 2016].

While frequently findings are based on genetic analysis and in-depth characterization of enrichment cultures, *in situ* methods to measure nutrient uptake have provided extensive advances in determining resource usage. Microautoradiography has been used to show *in situ* metabolite uptake [Nielsen et al., 2002] and has been applied in combination with fluorescent *in situ* hybridization (FISH) approaches to be able to connect metabolite uptake to individual populations [Wagner and Loy, 2002]. Additionally, uptake of labelled compounds can be quantified for individiual cells with nano-scale secondary-ion mass spectrometry (nanoSIMS) in combination with FISH showing phenotypic heterogeneity in populations in a stratified lake [Zimmermann et al., 2015], as well as wastewater [Sheik et al., 2015].

Multiple studies have used gene-amplicon sequencing for community structure profiling within BWWTPs, for example in relation to seasonal sludge bulking and nutrient removal [Xu et al., 2018; Wang et al., 2016; Cydzik-Kwiatkowska and Zielińska, 2016]. In contrast, multi-omics have been applied to characterize microbial functioning and niche ecology in floating sludge [Muller et al., 2014a] or in the context of microbial fuel cells [Ishii et al., 2013].

Studying microbial communities in BWWTPs holds several advantages over other natural or naturally occurring environments. Environmental parameters such as temperature, oxygen, or pH levels in BWWTPs are routinely recorded and the environment is relatively homogeneous in relation to defined physico-chemical boundaries [Narayanasamy et al., 2015]. The microbial communities within activated sludge exhibit medium to high diversity often with few dominant populations and as an intermediate between lower complexity environments, such as AMD, and highly complex environments, such as soil, important properties of both extremes can be assessed [Narayanasamy et al., 2015]. The fact that the ecological interactions of mixed microbial communities in BWWTPs have a profound impact on the biotechnological process, while at the same time, complex interactions can be studied in a controlled and monitored system, makes it an ideal model system for the study of microbial ecology.

3.2.2 Engineering microbial niches for efficient lipid accumulation

Filamentous bacteria are often related to bulking or foaming sludge problems in BWWTPs [Xu et al., 2018], while individual filamentous types can be associated with low food-to-biomass (F/M) ratios and shifts in nitrate and nitrite levels [Musvoto et al., 1999]. The capability to store compounds such as polyphosphates [Martín et al., 2006], glycogen [Crocetti et al., 2002], polyhydroxyalkanoates (PHA) [Yang et al., 2011], or lipids [Muller et al., 2014b] gives these organisms a competitive advantage in fluctuating and/or sparse nutrient conditions [Rossetti et al., 2005].

For lipid accumulation, this has been well described for Candidatus *Microthrix parvicella* (*M. parvicella* in the following text) [Nielsen et al., 2002] as it has been characterised as the often dominant organism in floating sludge and capable of efficiently accumulating lipids (extensively reviewed in Rossetti et al. [2005]). Particularly, growth of *M. parvicella* seems to benefit from low dissolved oxygen concentrations and it is characterised as a microaerophile [Rossetti et al., 2005]. Additionally, in alternating anaerobic-aerobic nutrient-removal plants a competitive advantage might be conferred to *M. parvicella* [Nielsen et al., 2002]. Under anaerobic conditions the storage of lipids putatively predominates, while in the presence of higher levels of electron acceptors (i.e. oxygen or nitrate) levels, stored lipids are metabolised for growth [McIlroy et al., 2013]. Thus *M. parvicella* may be considered as a metabolic generalist being able to cope with a wide range of resource gradients, which in turn has been linked to its genetic complement and the ability to fine-tune its gene expression [Muller et al., 2014a], as well as phenotypic heterogeneity of sub populations [Sheik et al., 2015].

The example of *M. parvicella* highlights the complexity of growth strategies that individual bacteria can pursue within a larger community and the dependence on environmental gradients, particularly oxygen. While *M. parvicella* is often a dominant population among lipid accumulators, other genera exhibit lipid uptake including *Aeromonas*, *Uruburuella*, and *Acinetobacter* [Kleine-Borgmann *et al.* - in preparation]. The metabolic capabilities of these and other lipid accumulating organisms (LAOs), as well as potential interactions could be assessed by screening their genomic potential [Muller et al., 2014b]. The possibility to delineate community interactions and strategies of resource usage within microbial communities can potentially allow to controlling and steering communities towards a desired phenotype. In the context of BWWTPs improved sustainability

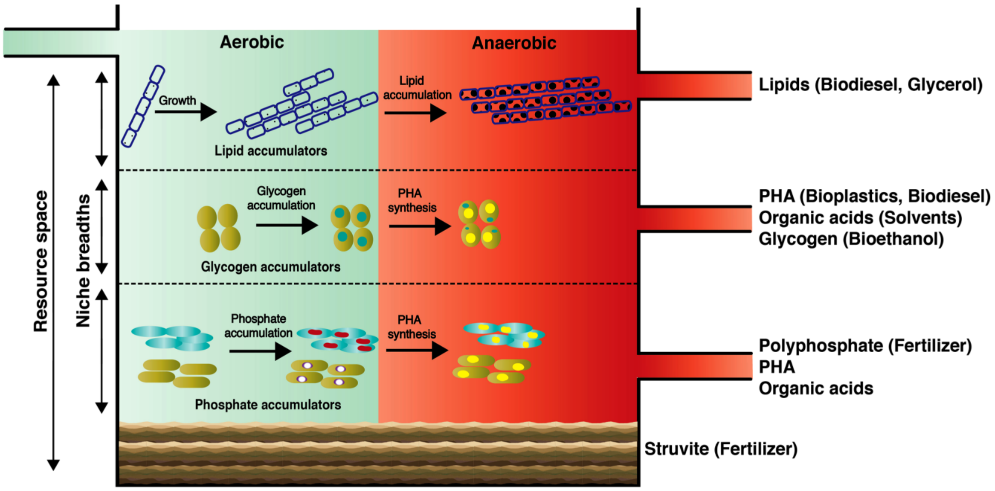


Figure 3.1: Conceptual schematic of niche engineering in a BWWTP. Scheme for the concept of a wastewater biorefinery column, i.e., engineering the activated sludge tank in way so that accumulation products of the various bacteria can be utilized as resources, considering the niches that the different populations occupy (from Sheik et al. [2014]).

could be achieved by engineering niches to enrich for microbial storage compounds, e.g. PHA, glycogen, lipids that could be the basis for products such as biofuels or bioplastics [Sheik et al., 2014] (**Figure 3.3**). This could be realized by physical separation, e.g., by introducing resource gradients or exploiting different settling properties [Sheik et al., 2014].

3.3 Methods

3.3.1 Recovery of isolate genomes

Floating activated sludge samples were collected from the surface of an anoxic tank in a communal wastewater treatment plant in Schifflange, Luxembourg, on October 12th 2011. 688 bacterial strains were isolated on varying media, temperature and oxic conditions, to cover a wide range of cultivation conditions. After describing colony morphology and cell shape, isolates were screened for lipid accumulation using the fluorescence of the lipophilic stain Nile red as readout for intracellular lipid droplets. Whole genome sequencing (Illumina) was performed on the 73 selected nile-red positive isolates. The isolation, screening, and genome sequencing protocols are described in Roume et al. [2015]. The sequencing data was preprocessed, assembled and annotated to produce draft genomes using the protocol described in Muller et al. [2017].

3.3.2 Sampling, experimental setup, and biomolecular extractions

Time-dependent sampling for multi-omics data generation was carried out by collecting floating sludge islets from the surface of the anoxic part of an activated sludge tank of the Schifflange BWWTP as described previously [Muller et al., 2014a]. Overall, the time-resolved sampling included two initial sampling dates (04.10.2011 and 25.01.2011) previously reported by Roume et al. [2015] and Muller et al. [2014a], followed by a higher frequency sampling phase, as a time-series from 23.03.2011 to 03.05.2012. During this period, 51 samples were collected in weekly intervals (mean: 8 days; sd: 4 days), including three longer gaps without sampling due to the absence of floating sludge islets: 08.07.2011 to 25.08.2011 (28 days), 12.10.2011 to 02.011.2011 (21 days), and 29.11.2011 – 28.12.2011 (29 days).

At the time of sample collection, physico-chemical parameters, including conductivity, pH, oxygen-levels, and temperature were measured in the anoxic tank (referred to as on-site measurements). Additionally, measurements were recorded by the BWWTP operators for nitrate-, phosphate-, ammonium-, dry-matter and dissolved oxygen-levels at the outflow of the activated sludge tank as well as conductivity and pH at the inlet, and pH and temperature of the activated sludge tank (referred to as operational measurements). Six missing values in the on-site measurements for pH were imputed from the available measurements with the R-package `imputeTS` [Moritz and Bartz-Beielstein, 2017] with the method `stine`.

Biomolecular fractions of DNA, RNA, proteins, and metabolites were obtained for each time-series sample as described previously in Roume [2013]. In brief, extracellular metabolites were extracted with chloroform and methanol-water, and separated into polar- and non-polar fractions. After lysis by milling, intracellular metabolites were isolated in the same way, followed by sequential spin-column-based purification of RNA, DNA, and proteins.

Additional experiments were carried out in bioreactors seeded with sludge samples. Short-term time-series experiments were set up as described by Sheik et al. [2015] as "alternating aerobic-anoxic phase experiment". In short, sludge samples diluted with artificial wastewater were aliquoted and treated to aerobic, anoxic, or alternating conditions after a 2h preconditioning period. After the preconditioning (time-point 0h) octadecenoic acid was supplemented alongside additional nutrients. Samples for subsequent DNA and RNA extractions were taken at 1h, 5h, and 8h and extracted according to the method in Roume [2013], resulting in twelve samples, i.e., three time-points for four conditions (aerobic, anoxic, aerobically preconditioned and alternating, anoxically preconditioned and alternating).

3.3.3 Nucleic acid sequencing, data preprocessing, and assembly

DNA and RNA biomolecular fractions for the 51 time-series samples and the two preliminary samples were sequenced as described by Muller et al. (2014), resulting in MG and MT reads for

53 samples. On average 2.8 x 10^7 MG reads and 3.3 x 10^7 MT reads were obtained per sample. Large-scale integrated MG and MT data analyses were performed on all the samples using `IMP ver. 1.3` [Narayanasamy et al., 2016] (**Figure 1.5**). Illumina Truseq2 adapters were trimmed, and the step for filtering reads of human origin was omitted for the preprocessing. The `MEGAHIT` *de novo* assembler [Li et al., 2015] was selected for co-assembly of MG and MT data. All other parameters of IMP were left at their default value (**Figure 3.2**).

`Nonpareil2` [Rodriguez-R and Konstantinidis, 2014] was executed using the k-mer based option on each time point by providing the IMP-preprocessed forward reads (R1) in FASTQ format as input sampling one million reads and using default parameters.

DNA for 12 samples of the short-term experiments described above were sequenced on four lanes of an Illumina Hiseq 2500 instrument with a fragment length of 250 bps as paired-end reads. Isolated RNA was reverse transcribed to cDNA and sequenced on five lanes of an Illumina Hiseq 2500 instrument with 100 bp paired-end reads. Resulting MT and MG reads were preprocessed with `IMP` [Narayanasamy et al., 2016]) as described above.

3.3.4 Meta-metabolomics

Four distinct measurements for the metabolite extracts were performed: i) non-polar extracellular (SNP), ii) polar extracellular (SP), iii) non-polar intracellular (BNP), and iv) polar intracellular (BP). Metabolite extracts were derivatized using a multi-purpose sampler (GERSTEL). Dried polar samples were dissolved in 15 µl pyridine, containing 20 mg/mL methoxyamine hydrochloride (Sigma-Aldrich), and incubated under shaking for 60 min at 40 °C. After adding 15 µl N-methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA; Macherey-Nagel), samples were incubated for additional 30 min at 40 °C under continuous shaking. Dried non-polar samples were dissolved in 30 µl MSTFA and incubated under shaking for 60 min at 40 °C.

GC-MS analysis was performed by using an Agilent 7890A GC coupled to an Agilent 5975C inert XL Mass Selective Detector (Agilent Technologies). A sample volume of 1 µL was injected into a Split/Splitless inlet, operating in splitless mode (polar fraction "biomass and supernatant") and split mode (10:1, non-polar fraction "biomass") at 270 °C. The gas chromatograph was equipped with a 30 m (I.D. 250 µm, film 0.25 µm) DB-5MS capillary column (Agilent J & W GC Column). Helium was used as carrier gas with a constant flow rate of 1.2 mL/min.

The GC oven temperature was held at 80 °C for 1 min and increased to 320 °C at 15 °C/min. Then, the temperature was held for 8 min. The total run time was 25 min. The transfer line temperature was set constantly to 280 °C. The mass selective detector (MSD) was operating under electron ionization at 70 eV. The MS source was held at 230 °C and the quadrupole at 150 °C. Full scan mass spectra were acquired from m/z 70 to 700.

All GC-MS chromatograms were processed using the `MetaboliteDetector` software [Hiller et al., 2009]. The software package supports automatic deconvolution of all mass spectra. The following deconvolution settings were applied: Peak threshold: 6; Minimum peak height: 6; Bins per scan: 10; Deconvolution width: 2 scans; No baseline adjustment; Minimum 15 peaks per spectrum; No minimum required base peak intensity. Compounds were annotated by retention time and mass spectrum using an in-house mass spectral library.

Metabolites detected in blanks at a mean intensity level of more than 75% of the mean level in samples were removed as contaminants. Metabolites that were not detected in all pool samples were also removed from subsequent analysis as well as metabolites not detected in at least 90% of measured samples. Metabolite intensities were normalized in respect to pool samples to account for instrument drift as described [Roume, 2013] dividing the intensity values by the mean of up to two preceding and subsequent pools samples according to the measurement sequence. Metabolite derivate names of identified metabolites were manually assigned to KEGG compound identifiers and CHEBI IDs.

3.3.5 Metaproteomics

Protein samples of 51 time-series samples were measured as previously described [Muller et al., 2014a] by LC-MS/MS.

The database searching process for all of the converted mass spectrometry mgf files utilized the `Graph2Pro` pipeline [Tang et al., 2016], which integrated MG, MT, and MP data. In detail, for each time point, `Graph2Pep` first predicted the peptides from one or more short edges in the assembly graph of sequencing data. `FragGeneScan` [Rho et al., 2010] was employed to predict the protein/peptides from contigs consisting of long edges. The combined set of tryptic peptides, including those predicted from long edges and those extracted from one or more short edges in the graph (by `Graph2Pep`), were used as the target database for peptide identification in the MP data by using the `MS-GF+` search engine [Kim and Pevzner, 2014]. `Graph2Pro` then further predicted protein sequences from the graph of the MG/MT assembly, using identified peptides as constraints. The MP data was searched against the database output from `Graph2Pro` to produce final identification results.

`MS-GF+` was used for peptide identification from a given protein sequence database using the following parameters: 1) instrument type: high-resolution LTQ; 2) precursor mass tolerance: 15ppm; 3) isotope error range: -1,2; 4) modifications: oxidation as variable and carboamidomethyl as fixed; 5) maximum charge: 7; and 6) minimum charge: 1. The false discovery rate (FDR) was estimated by using a target-decoy search approach. For full-length proteins predicted from `FragGeneScan` or `Graph2Pro`, the reverse protein sequences were used as decoy. For peptides predicted from Graph2Pep, the decoy peptides were then generated by reversing the peptide sequences while preserving the C-terminal residues (K/R).

3.3.6 Binning

`IMP`-based co-assembled contigs from each time point were binned using a previously described method [Heintz-Buschart et al., 2017; Kaysen et al., 2017], which is based on i) nucleotide signatures, ii) contig-level average depth-of-coverage, and iii) essential gene calls of `IMP`-based co-assembly contigs with length above 1kbp (**Figure 3.2**). Bins from each time point with a completeness above 28% and contamination below 20%, i.e., labelled as “P”, “G”, “O” and “L” [Heintz-Buschart et al., 2017], were retained for downstream selection of representative genomes. In order to de-replicate the collection of high-quality bins from different time-points and obtain representative reconstructed genomes (ReGes) over-time, `dRep` [Olm et al., 2017] was applied with the following parameters: i) completeness threshold of 0.6, ii) strain heterogeneity threshold of 101, removing this threshold for selection iii) primary cluster nucleotide identity of 0.6, and iv) secondary cluster nucleotide identity of 0.965, while other parameters remained as default. In a following step, a subset of ReGes was selected based on `CheckM` [Parks et al., 2015] completeness estimates, requiring at least 50 % in the difference of completeness and contamination estimates. Isolate genomes that had been added to the dereplication process were discarded, due to not representing *de novo* assembled bins and low coverage in the MG and MT data. High-quality ReGes were subsequently used as references for integrating omics data.

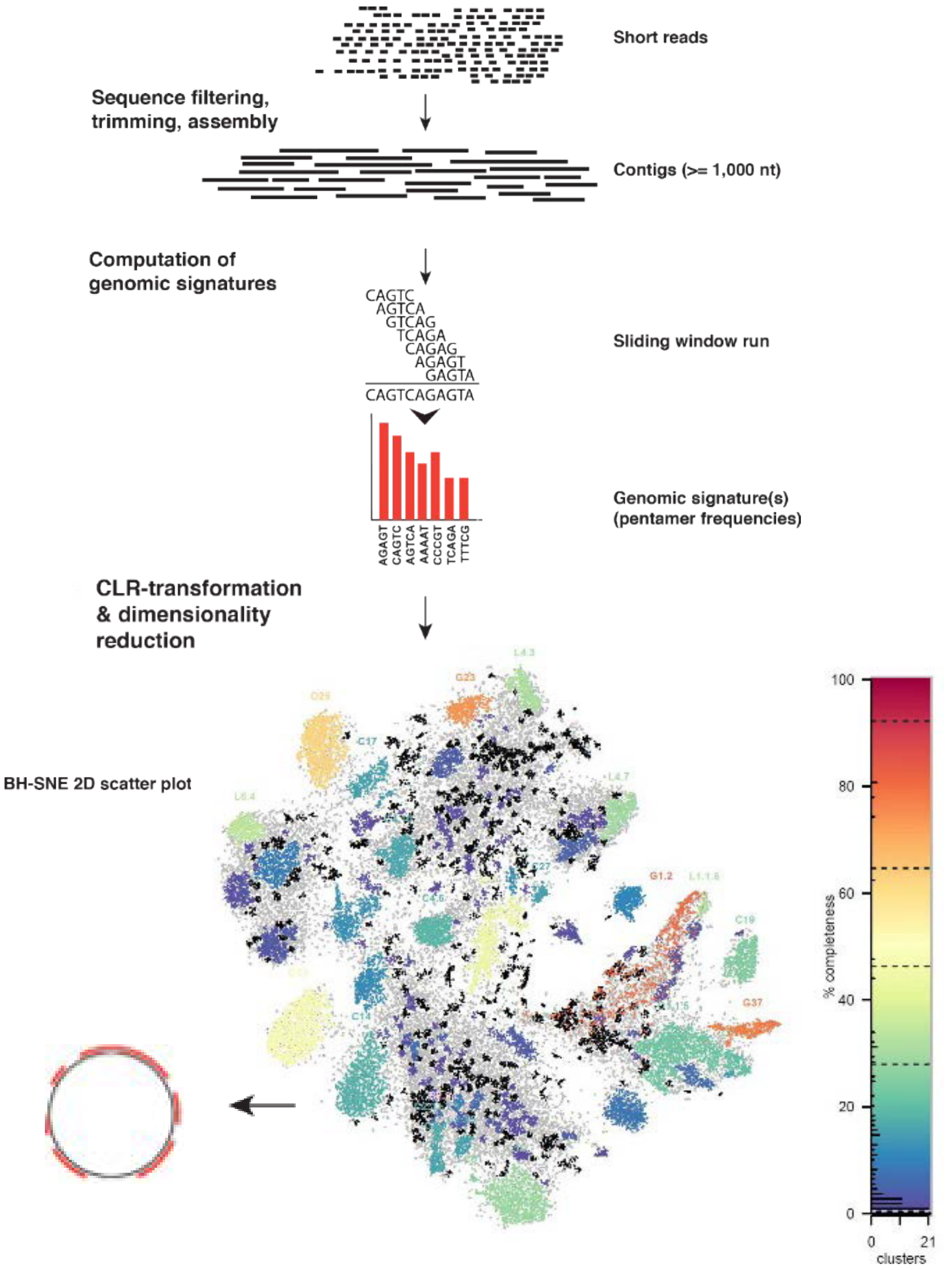


Figure 3.2: Assembly and automated binning workflow. IMP [Narayanasamy et al., 2016] was run on each sample of the time-series, processing MT and MG short reads and performing *de novo* assembly of contigs. Pentamer nucleotide frequencies were calculated and transformed to two-dimensional space [Laczny et al., 2015]. An automated clustering method was applied and selected grouped contigs and refined the recovered genomic bins by assessing marker gene content and coverage distributions [Heintz-Buschart et al., 2017].

3.3.7 Genome annotation and taxonomic assignments

Assembled contigs were annotated with `Prokka v1.11` [Seemann, 2014] including prediction of CDS with `prodigal v2.6` [Hyatt et al., 2010]. Predicted CDS were also searched with an in-house HMM database of the KEGG ortholog groups (KO) with `HMMer v1.12b` (Eddy, 2011) as previously described in Heintz-Buschart et al. [2017]. All annotations and assigned probability scores were stored in a mongoDB database for storage and access [Heintz-Buschart et al., 2017]. Reactions associated with predicted enzymes were added through combining EC-number-to-KEGG-reaction (RN) links and KO-to-RN links. From the combined list of reactions, links to KEGG compounds (CMPs) were inferred by linking gene identifiers to unique products or substrates of putatively catalysed reactions.

Reconstructed genomic bins were analysed with `AMPHORA2` [Wu and Scott, 2012] with a customized version previously applied by Laczny et al. [2016]. Additionally, taxonomic classification was performed with `Sourmash 2.0.0a1` [Brown et al., 2016] lca-version with `kmer-length: 21` and `threshold: 4`, with an existing database including approximately 87,000 microbial genomes (downloaded on 2017-11-09 from `https://osf.io/s3jx8/download`).

`AMPHORA2-based` predictions for individual marker genes were combined by summation of the associated assignment probabilities. If the summed probability scores for the highest-scoring taxonomic level constituted not more than 1/3 of the total probability scores, the assignment was discarded as “low confidence assignment”. Taxonomic assignments of `AMPHORA2` and `sourmash-lca` were merged giving priority to predictions from `sourmash-lca` due to higher expected specificity and an updated database.

3.3.8 Analysis of functional potential

Annotated KOs for the individual ReGes were summarized in a binary matrix combining all ReGes: 0 indicated absence and 1 indicated presence of at least 1 gene annotated with the respective KO. Pairwise binary Jaccard-distances between the ReGes based on their KO profiles were calculated and projected in two-dimensional space by multi-dimensional scaling (MDS). Four clusters of functional potential (FunCs) were selected by applying the `k-means` function (`kmeans` in R, k=4) after inspection of the `k-means` clustering results for a range of centroids settings from one to nine.

Enrichment of individual KOs in the assigned clusters by functional potential (FunCs) was tested individually for every KO with Fisher’s exact test, comparing the number of bins with KO present against the number of assignments in different groups and the number of all KOs within and outside of the FunC assignment. Resulting p-values were adjusted by FDR correction.

To test the validity of the FunC assignment, pairwise correlations (`cor` function in `R, method= "pearson"`) between relative abundances of the ReGes across time were computed. Correlation coefficients (ρ) were transformed to distances with the following formula: $1 - ((\rho + 1)/2)$. Dispersion of the distances was assessed with `betadisper` function of the vegan package in R. Association of the FunC assignment to the distances was tested with `adonis` of the vegan package in R.

3.3.9 Analysis of gene expression levels

Preprocessed MG and MT paired- and single-end reads from all samples were mapped to the genome sequences of the ReGes with `bwa` [Li and Durbin, 2009]. Reads mapping to CDS were extracted with `featureCounts` [Liao et al., 2014] for the long-term time-series and the short-term experiments.

Counts mapped to individual genes were normalized to TPM counts within an individual ReGe for comparison of relative expression levels with a single population [Klingenberg and Meinicke, 2017]. For the analysis of individual expression levels pathway assignments defined in McIlroy et al. [2013] formed the basis for selecting genes based on matching gene name, product name, or EC-number.

MG and MT depth of coverage were computed on gene and contig level by dividing the summed depth per base by the length of the respective sequence.

For the long-term time-series MP data was analysed for 51 time-points. Identified peptides by the `Graph2Pro` pipeline were assigned to the original prodigal-based predictions for coding protein sequences of the ReGes with `peptidematch` [Chen et al., 2013].

3.3.10 Linking abiotic factors to population abundance and expression patterns

Relative abundances of ReGes were associated to abiotic factors by computing the correlation (Spearman rank, `cor.test` function in R) between abundances and z-score transformed metabolite intensities or physico-chemical parameter levels or concentrations.

Analogously to the profiles of potential functions, we generated binarized gene expression profiles per time-point using the MT and MP data. MT and MG depth of coverage for individual genes was normalized by dividing the gene-wise depth (MT or MG) by the total count of mapped reads in the sample and multiplied by the mean of all mapped reads over all samples as a scaling factor. Gene expression profiles were assessed by dividing gene MT depth by gene MG depth, where the MG depth was set to 1 if it was below 1. If the resulting ratio was above 1 the gene was considered expressed. Additionally, the profiles were augmented with MP data and a gene was considered expressed if at least 2 peptides could be matched to its protein sequence for a given time-point.

For the resulting expression profiles based on MT/MG ratios and MP data binary Jaccard-distances were computed. To assign potential driving factors, abiotic factor levels (z-score transformed levels of physico-chemical parameters and metabolites) were used as input in the vegan function `adonis` to test whether the parameter explained the variance in the expression profile distances.

3.4 Results

3.4.1 Reconstruction of representative genomes to determine the community structure over time

Individual foaming sludge islets were sampled in weekly intervals from 23.03.2011 to 03.05.2012. We obtained and analysed the biomolecular fractions of DNA, RNA, proteins, and metabolites for each time-series sample, as well as the operator-recorded abiotic parameters of at the sampling dates (**Section 3.3.2**).

We recovered 73 genomes by sequential isolation and selection of lipid-accumulating populations. Despite being able to recover high quality genomes (**Appendix A.3**) and several novel populations and interesting pathways [Roume et al., 2015; Muller et al., 2017] obtaining a representative set of genomes was not possible. This was mainly due to the general low abundance of these isolated organisms within the *in situ* samples. Therefore, we proceeded to reconstruct genomes directly from the MG and MT data.

In order to identify important players in the system we aimed to reconstruct population-level genomes from the nucleotide sequencing data to later integrate other functional omics data. On average a 2.84×10^7 (sd: 4.0×10^6) MG and 3.30×10^7 (sd: 5.4×10^6) MT read pairs per sample were processed using IMP [Narayanasamy et al., 2016] resulting in an average 4.15×10^5 co-assembly assembled contigs per time-point [Narayanasamy, 2017]. A time point-wise binning procedure of the assembly contigs yielded a total of 1,364 metagenomic-assembled genomes (MAGs) justifying an initial quality cut-off (**Section 3.3.6**). However, to link MAGs originating from similar bacterial populations over the time-series, all MAGs were grouped based on genome similarity resulting in 170 representative genomes (ReGes) (**Appendix A.3**).

After another filtering step using genome completeness estimates as well as taxonomic consistency we obtained 92 ReGes. 14 of the previously completeness-filtered ReGe set of 92 were discarded due to low confidence assignments at the phylum level, resulting in a reduced set of 78 ReGes. ReGes reflect the highest quality genomic bins over time representing clusters at 97.5 % genome similarity. Subsequently, MT and MG depth of coverage, and peptide assignments were associated with contigs and predicted coding sequences (CDS) of the ReGes.

To determine community structure, the MG-depth of the reconstructed ReGes was analysed across the time-series. The most abundant populations at the genus-level (**Section 3.3.7**) included *Acinetobacter* (Gammaproteobacteria), *Albidiferax* and *Dechloromonas* (Betaproteobacteria) of the Proteobacteria phylum, as well as *Intrasporangium* and Candidatus *Microthrix* (referred to as *Microthrix* in the remaining text) of the Actinobacterial phylum, *Haliscomenobacter* and *Chitinophaga* of the Sphingobacteriales, and *Leptospira* of the Spirochaetes phylum (**Figure 3.3**). Several of the recovered ReGes belonged to filamentous taxa according to the MiDAS field guide database for organisms in activated sludge [McIlroy et al., 2015], typically found in foaming sludge, such as the

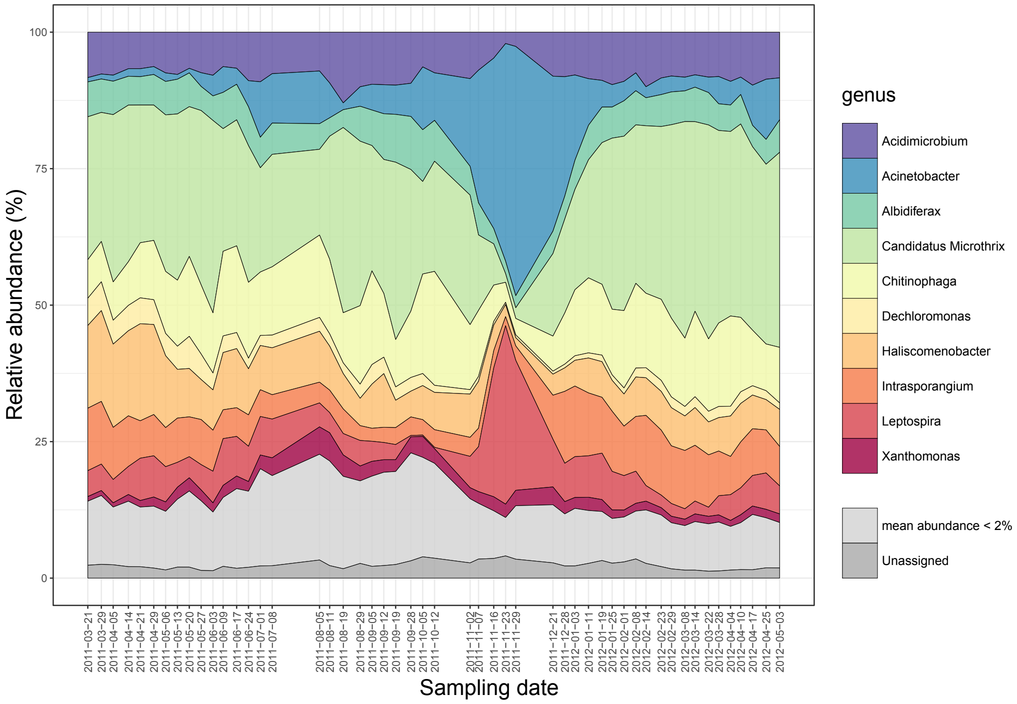


Figure 3.3: Relative abundance of reconstructed populations over time. Relative abundance of representative genomes (ReGes) determined by metagenomic depth of coverage. Relative abundance of individual bins was grouped based on genus-level taxonomic assignment. Genera below mean abundance of 2% were grouped (light grey), as well as ReGes without genus level assignment (dark grey).

highly abundant *Microthrix*, and *Haliscomenobacter*, as well as the less abundant *Gordonia*, and *Anaerolinea*.

The abundances of the individual organisms gradually changed with the seasons (**Figure 3.4**). While the community structure remained relatively stable in the beginning of the time-series (Spring 2011: 21.03.2011 - 03.06.2011), a small shift occurred towards another grouped set of samples during summer 2011 (09.06.2011 - 26.09.2011). In October 2011, community composition began to shift, leading to a notable change of the community structure in late November 2011. This change is marked by spikes in relative abundance of *Leptospira* (peak at 23.11.2011) and *Acinetobacter* (peak at 29.11.2011) (**Figure 3.3**). In the following winter time-points, the community transitioned back to a state more similar, yet separate, to the initial one, with increased dominance of *Microthrix* (11.01.2012 - 03.05.2012). The observation of the shift in community structure in autumn, as well as the trends in overall community structure are corroborated by the results of 16S rRNA gene sequencing in a previous study of the same system showing an increase in Gammaproteobacteria in autumn and an increase in Actinobacteria and Bacteroidetes in winter time-points [Muller et al., 2014a]. While ReGes are not covered by all reads, matching all MG-reads to 16S rRNA gene regions with the tool RiboTagger [Xie et al., 2016] showed similar trends (**Figure B.1**). *Microthrix* was the most dominant individual genus, but genera assigned to the Gammaproteobacteria and Spirochaetes also dramatically increased in relative abundance in late November 2011, and the overall community structure following a similar seasonal pattern (**Figure B.2**). This indicates that the overall community structure and dynamics are reflected by the selected subset of ReGes, even though a complete picture of the community is limited by sequencing depth. An assessment with `Nonpareil2` [Rodriguez-R and Konstantinidis, 2014] that estimates community coverage based on redundant sequencing reads predicted an average of 42 % community coverage by MG sequencing, likely resulting in the undersampling of rare taxa.

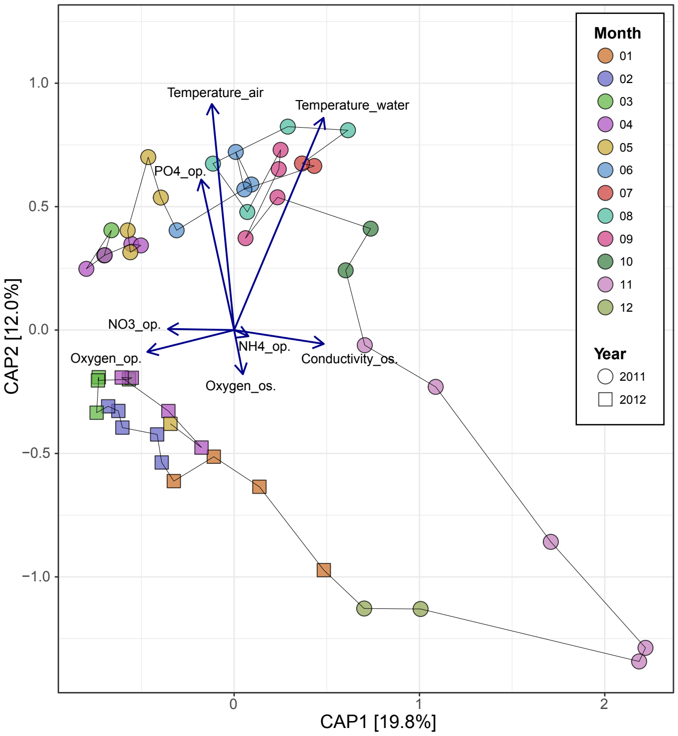


Figure 3.4: Constrained ordination of population abundances. Ordination of species composition based on the Bray-Curtis dissimilarity of relative abundances of individual ReGes constrained by selected abiotic factors (dark blue arrows and labelled arrowheads). Points are coloured by month of sampling and point-shape reflects the year of sampling. Arrow length indicates the environmental score of each factor. Thin black lines connecting the points visualizing the time-course of sampling. Parameters measured during sampling are marked by "os." (on-site sampling), while parameters recorded by the BWWTP operators are marked by "op.".

3.4.2 Fundamental niche and resource associations

The community dynamics, i.e., changes in population structure over time, can partly be explained by environmental parameters (**Figure 3.4**). Strong associations between the abundance profiles and temperature were observed with summer samples also being characterized by higher phosphate levels. Winter samples were characterized by higher oxygen and nitrate levels. Ammonium generally did not show a strong impact on the community structure (**Figure 3.4**).

As the gene pool of a population generally determines the environmental gradient it can survive in (**Section 1.1.1**), the functional potential of the individual populations represented by ReGes was assessed. KEGG orthologue groups (KOs) were assigned to the predicted coding sequences (CDS) as basis for comparison of the functional potential. Projecting pairwise Jaccard distances of KO presence between the ReGes resulted in four larger groups (**Figure 3.5**). These functional potential-based clusters (FunCs) indicate differences in terms of overall lifestyle and metabolic capabilities of the ReGes, and serve as proxies for their respective fundamental niches.

The ReGes grouped into FunCs primarily according to their assigned taxonomy, with FunC-1 including the Actinobacteria phylum, and FunC-2 is comprised primarily of Bacteroidetes (mainly of class Sphingobacteriia). FunC-3 contained Gamma- and Betaproteobacteria, while FunC-4 contained Spirochaetia as a sub-cluster, Deltaproteobacteria, singleton assignments, and unassigned ReGes. The relationship between phylogenetic and metabolic distance has been shown previously for reference genomes, especially for Actinobacteria and Bacteroidetes [Bauer et al., 2015].

To determine the functional capabilities separating the four groups, an enrichment analysis was performed (one sided Fisher's exact test, adjusted p-value <0.05, see **Section 2.3**). In total FunC-1, FunC-3, and FunC-4 show a similar amount of KO assignments with 4,276, 4,177, and 4,129 respectively. Slightly less KOs were assigned to FunC-2 with 3,550 and fewest KOs are shared between FunC-2 and the other groups. However, nearly half 1,857 KOs are shared among all the FunCs (**Figure 3.6**).

FunC-1 showed enrichment for several KOs in starch and sucrose metabolism, as well as distinct KOs involved in lipid metabolism, e.g., diacylglycerol O-acyltransferase [EC:2.3.1.20], short/branched chain acyl-CoA dehydrogenase [EC:1.3.99.12], or glycerate 2-kinase [EC:2.7.1.165]. 10 of the 24 ReGes in FunC-1 had KO assignments for important enzymes in the ethylmalonyl-CoA pathway (crotonyl-CoA reductase [EC:1.3.1.86]), enoyl ACP reductase [EC:1.3.1.9]). Similarly, for 12 ReGes aminobutyraldehyde deydrogenase [EC:1.2.1.19] and related KOs were only present in FunC-1.

FunC-2 showed distinct KOs, especially for (amino)glycan degradation or glycosphingolipid metabolism (hexosaminidase [EC:3.2.1.52], glucosamine kinase [EC:2.7.1.8], glucosylceramidase [EC:3.2.1.45]), for amino acid synthesis (chorismate mutase [EC:5.4.99.5], anthranilate synthase component II [EC:4.1.3.27]), as well as putrescine aminotransferase [EC:2.6.1.82]. Additionally, KOs related to mevalonate metabolism and nitrous-oxide reductase [EC:1.7.2.4] were found to be enriched in

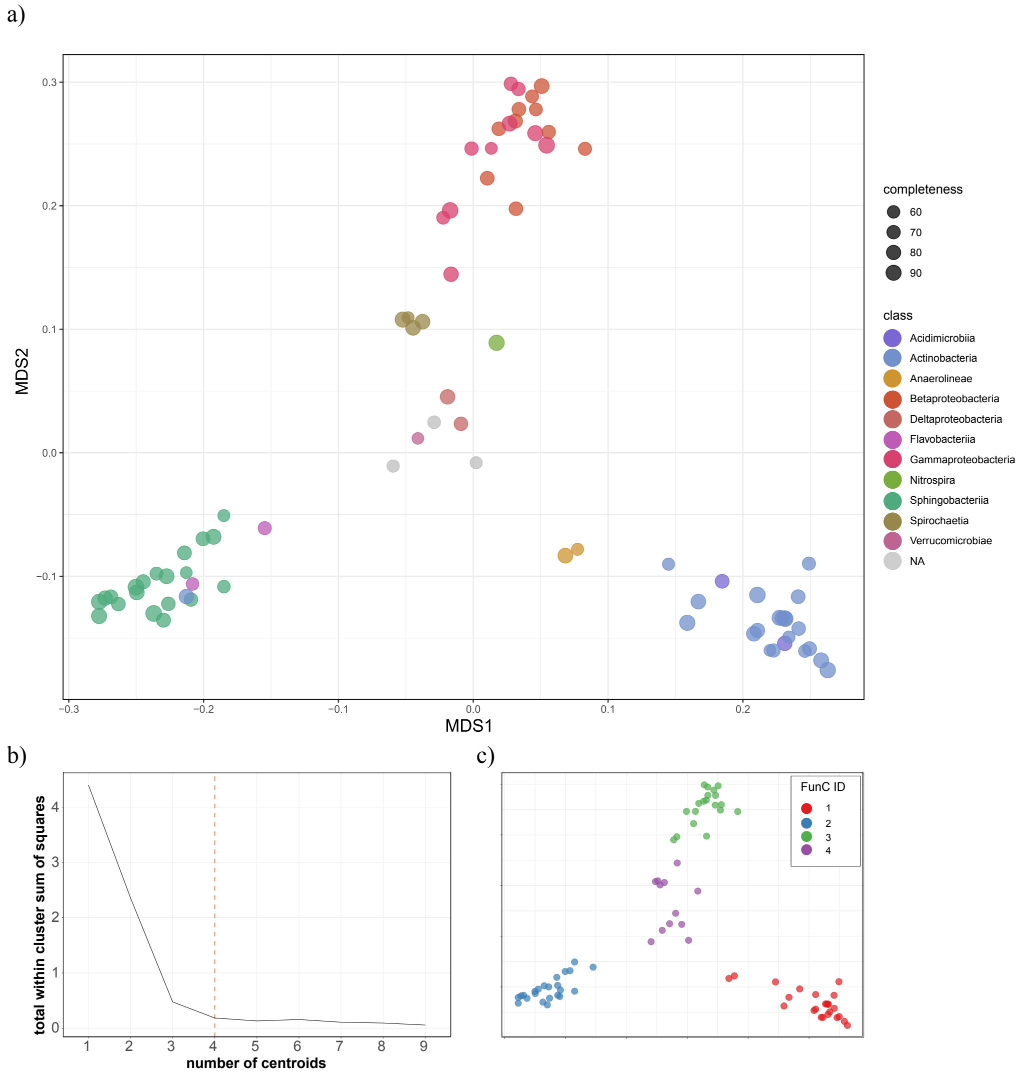


Figure 3.5: Assignment of cluster according to functional potential.
a) Multidimensional scaling (MDS) of Jaccard-distances between profiles of KO presence/absence per representative bin (ReGe). Each point represents a single ReGe with colors reflecting class-level taxonomic assignment. NA (grey) represents ReGes without taxonomic assignment on class-level.
b) K-means clustering of the MDS coordinates with different number of centroids (x-axis) and corresponding number of total within-cluster sum of squares (y-axis).
c) Ordination of a) with colors indicating k-means (centroids=4) cluster assignments.

this cluster. 14 of 23 ReGes in FunC-2 also showed KO assignments for heterodisulfide reductase subunit C [EC:1.8.98.1] only present in this group.

FunC-3 was found to contain more KOs related to chemotaxis and motility, especially KOs related to twitching motility, glutathione synthase, and KOs belonging to the two-component system. KOs involved in the pentose phosphate pathway (phosphomannomutase / phosphoglucomutase

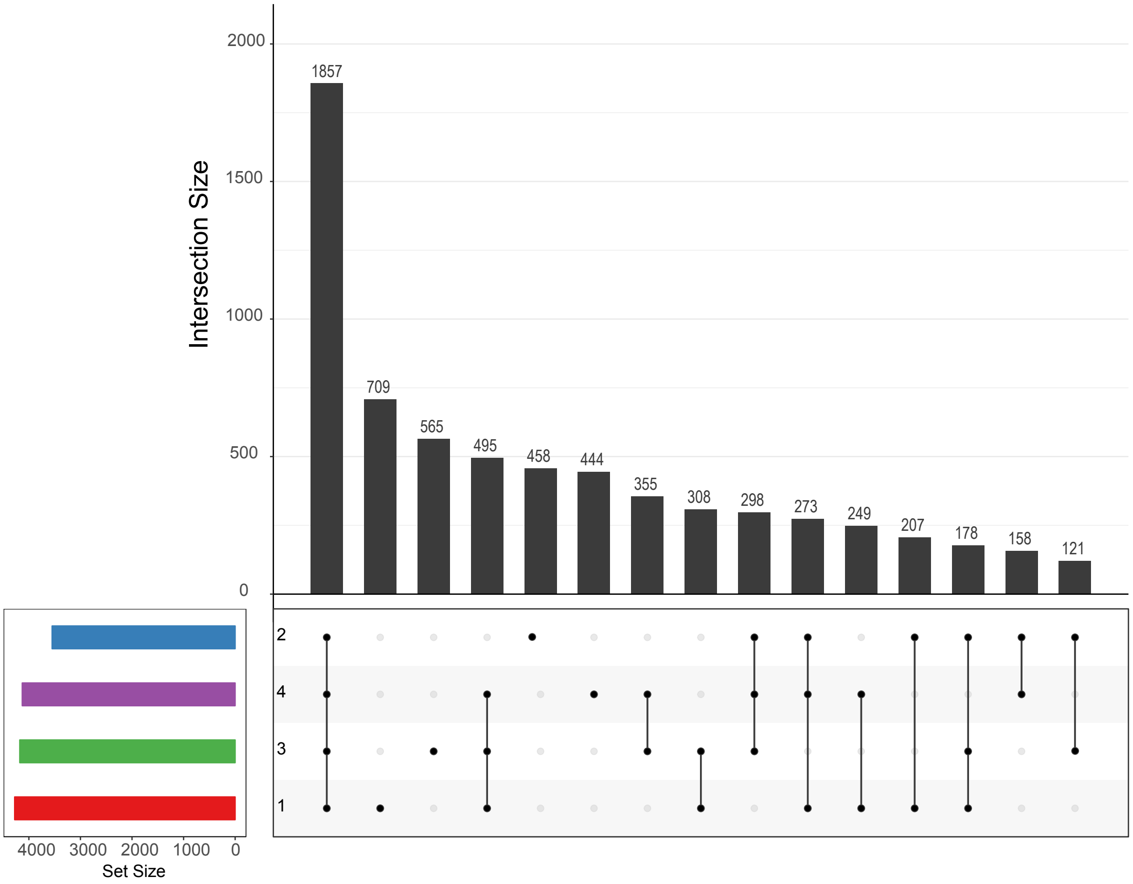


Figure 3.6: Upset plot of shared KEGG orthologue assignments per FunC. Overlap in KOs assigned uniquely to one of the four clusters of functional potential (FunCs). The lower-left panel indicates total number of unique KOs per FunC, coloured according to FunC assignment. Bars (central panel) indicate the number of intersecting KOs between the respective groups (connected dots in lower panel) or KOs unique to a single group (single dot respective for group in lower panel).

[EC:5.4.2.8 5.4.2.2] ribose 5-phosphate isomerase A [EC:5.3.1.6]) were present in many members of FunC-3. Assignments in FunC-4 showed fewest enriched KO assignments. While some KOs were found to be enriched in this group such as tyrosine aminotransferase [EC:2.6.1.5], nitrate reductase, or KOs for flagellar assembly, these were only assigned to half of the members of FunC-4 at maximum.

Overall, many functions are shared between the different FunCs but distinct KOs were assigned to subtypes of similar enzymes, e.g. for fructose-1,6-bisphosphatase (FunC-1: class type 2; FunC-2 mainly class 1; FunC-3: FBPase class 1/SBPase; FunC-4: FBPase class 2/SBPase). KOs related to fatty acid degradation, such as acyl-CoA dehydrogenase [EC:1.3.8.7] and long-chain-acyl-CoA dehydrogenase [EC:1.3.8.8] were mainly present in FunC-1 and FunC-3. While genes associated to general fatty acid metabolism and lipid synthesis such as long-chain acyl-CoA synthetase [EC:6.2.1.3] were present in all FunCs.

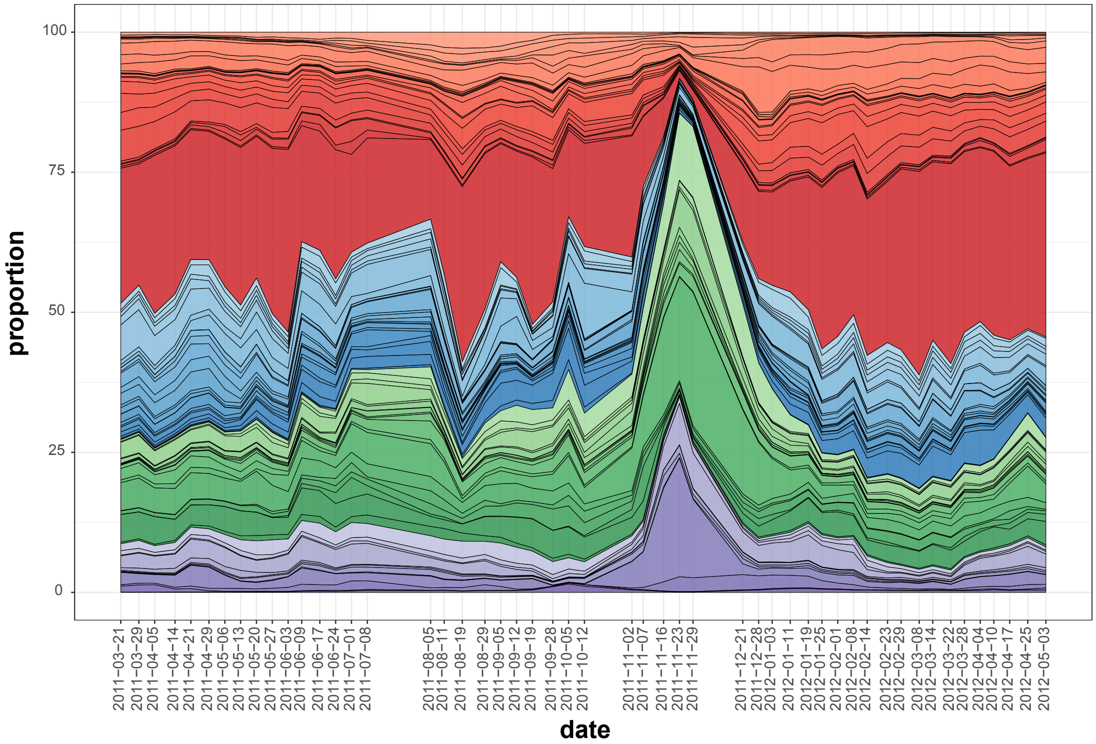


Figure 3.7: Relative abundance of ReGes related to FunC assignment. Relative abundance of ReGes estimated by MG-depth, coloured by FunC assignment:
Reds: FunC-1, Blues: FunC-2, Greens: FunC-3, Purples:FunC-4

Interestingly, the aforementioned changes in relative abundance over time could be related to the FunC assignment of the ReGes, as mainly ReGes within FunC-3 and FunC-4 showed an increase in relative abundance in November 2011 (**Figure 3.7**). To confirm that population abundance is linked to FunC assignment, we computed pairwise correlations between the relative abundances of the ReGes and transformed these to distances. While the dispersion of the distributions of the distances was not significantly different, the distances can partly be attributed to the FunC assignment (`vegan adonis` $R^2 = 0.12, Pr > F = 0.002$).

The environmental conditions, i.e., the resource space (see **Section 1.1.1**), is expected to shape the community structure as environmental factors can explain biological variation across different patches [Ramette and Tiedje, 2007]. However, to link resource parameters to individual populations in an *in situ* system is a complex task, for example as unlabelled metabolites cannot be directly traced back to individual populations. In order to describe connections between individual ReGes and resource parameters, physico-chemical parameters (**Figure 3.8**) and metabolite levels were analysed for the time-series samples. Filtering of the non-targeted metabolite measurements yielded intensities several derivates that could be reliably identified throughout the time-series, 56 for the polar (intra- and extracellular), 6 extracellular non-polar (SNP), and 17 for the intracellular non-polar (BNP) measurement, respectively.

The measurements of the non-polar metabolites showed similar patterns within the extracellular and intracellular fraction. In the extracellular fraction, long-chain fatty acids (LCFAs) were markedly increased in the November 2011 samples, especially unsaturated octadecadienoic and octadecenoic acid (**Figure B.5**). A similar trend could be observed for polar metabolites such as glycerol, glycerol-2-phosphate (**Figure B.6**) lactose, mannose, glucose (**Figure B.7**), putrescine, and ethanolamine (**Figure B.9**). A notable exception were the identified amino acids. Here, the intensities reached a maximum in March and April 2011 where nearly all amino acids showed a reduction in intensity in the intracellular fraction in November 2011 (**Figure B.8**).

In order to link resource parameters to individual populations we computed Spearman rank correlations between the time-courses of relative abundance and the z-score transformed levels of the physico-chemical parameters, the metabolite intensities, as well as ratios between intra- and extracellular metabolites intensities to estimate uptake (**Figure 3.9**). Grouping ReGes according to the correlation of their abundance to abiotic factors revealed a pattern of subgroups distinct from the original FunC assignment (**Figure B.3**). Spirochaete ReGes grouped mainly with Gammaproteobacteria with their abundance positively correlated with intra- and extracellular long-chain fatty acid (LCFA) and glycerol levels. Another mixed group notably containing *Anaerolinae* assigned ReGes, Betaproteobacteria, and Sphingobacteriia were negatively correlated with individual LCFA levels, but positively to LCFA ratios and showed a strong connection to temperature. A large group of mainly FunC-1 and FunC-2 assigned ReGes exhibited positive correlations to amino acid levels and ratios for sugar derivates, as well as for intra- or extracellular fructose, glucose, mannose, or lactose levels.

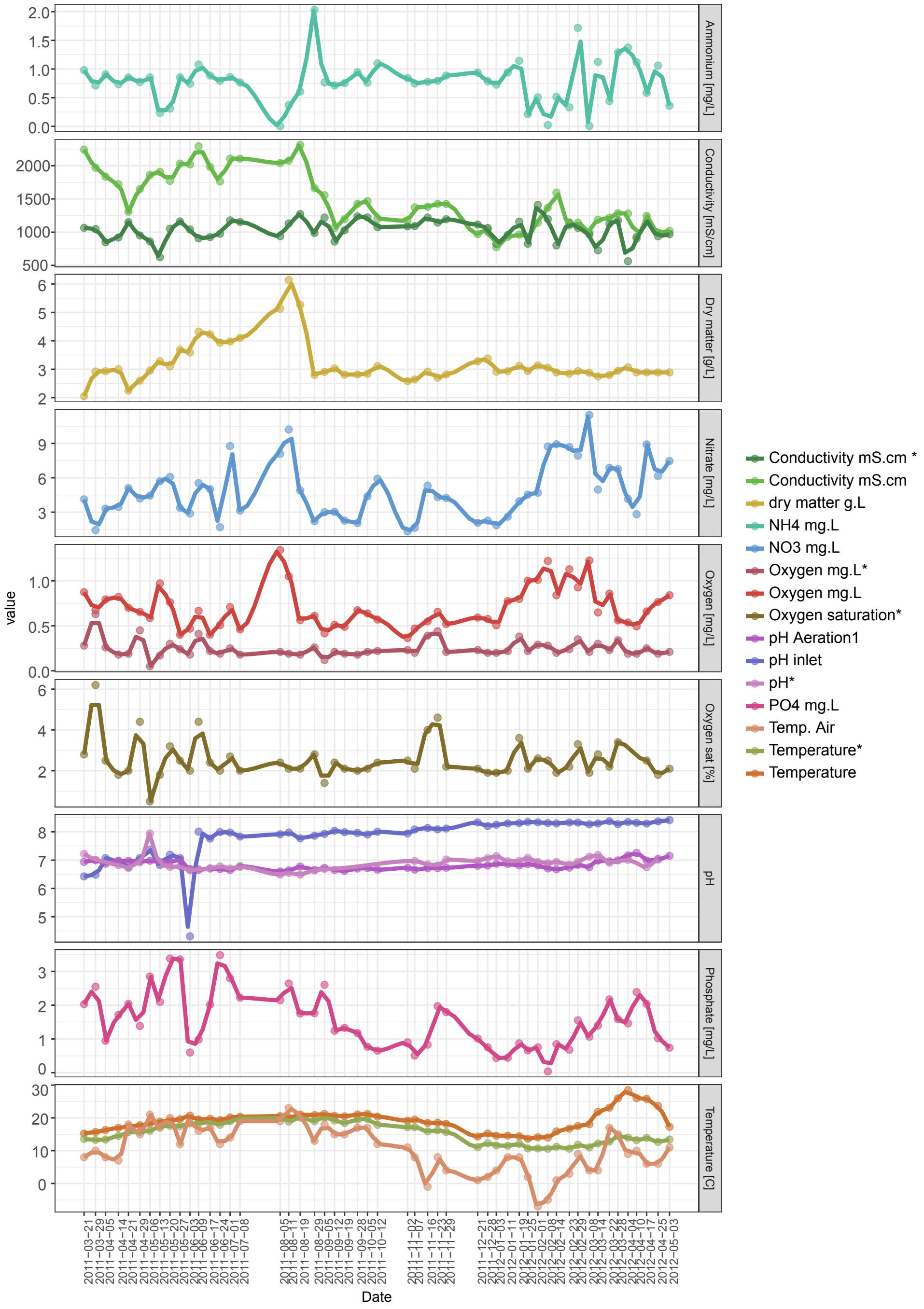


Figure 3.8: Physico-chemical parameters over the time-series. Physico-chemical parameters measured during sampling on the surface of the anoxic zone of the activated sludge tank (on-site measurements taken during sampling are marked by *) and by the WWTP operation in the outflow (oxygen, phosphate, nitrate, ammonium, dry-matter), inflow (conductivity, inlet-pH), or bulk (temperature, pH) of the activated sludge tank ("operational"). Points represent individual measurements for the on-site values or daily averages for sampling days (operational), lines show a locally weighted smoothing (loess) of the values per measured parameter. Panels are separated by type or unit.

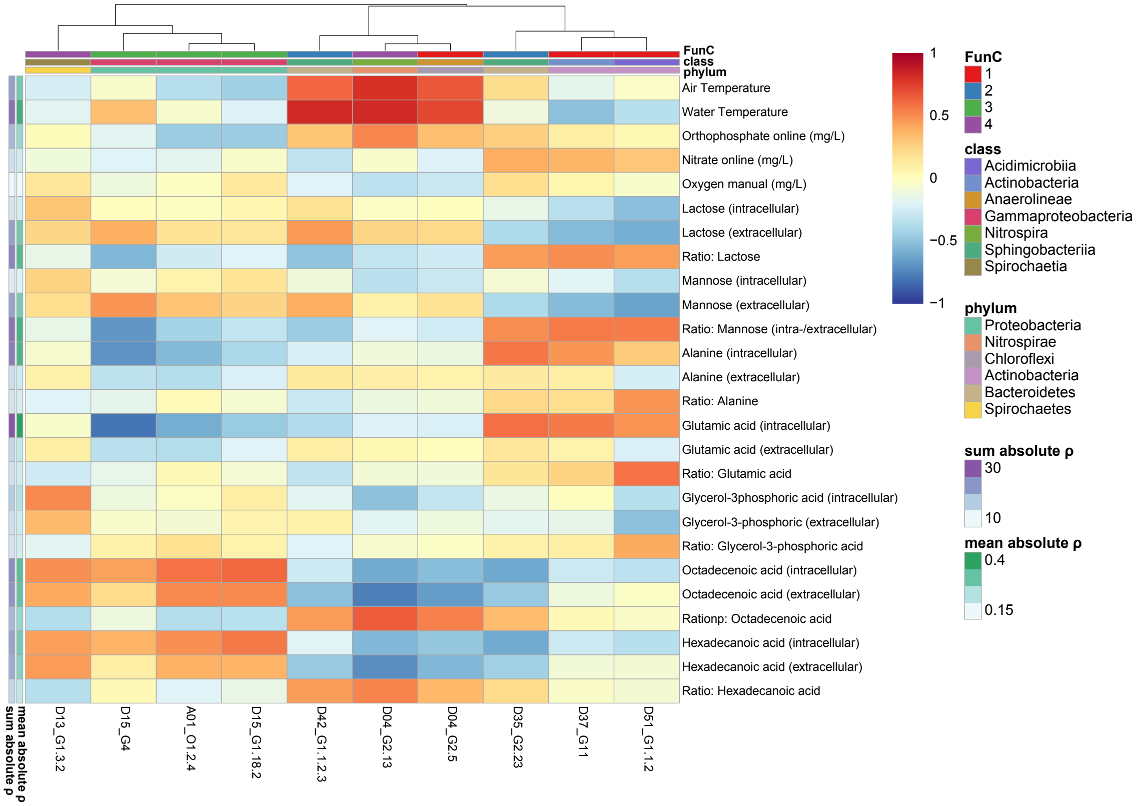


Figure 3.9: Rank correlations of abiotic factors and relative abundances. Correlation coefficients (Spearman rank rho) of abiotic factors (z-score transformed levels) to the relative of abundance (MG-depth) of selected ReGes. A subset of tested parameters and ReGes is shown. Row annotations show the sum and mean of the absolute correlation coefficients derived the full overview, i.e., the filtered set of 78 ReGes (**Figure B.3**).

3.4.3 Characterizing realized niches by assessment of gene expression patterns

In a fluctuating resource space, a competitive advantage is conferred to microbial populations that can adapt to the changing conditions. These adaptations are likely accompanied by changes in gene expression patterns [Gifford et al., 2013]. To trace expression patterns of individual ReGes over time, an index to characterize the activity of individual genes was defined. Genes were considered expressed at a given time-point if the ratio MT-depth over MG-depth ratio was above 1 or if at least two peptides in the MP-data could be assigned to the gene (**Section 3.3.10**). Using this activity index Jaccard distances between the time-point specific binary gene activity profiles of each ReGe were computed.

Subsequently, individual resource parameters (**Section 3.4.2**) were associated to expression-profile distances to determine the influence on gene activity. In a comparison of 78 ReGes, individual ReGes grouped similarly to abundance based correlations (**Figure B.10**).

While metabolite intensity ratios were less frequently found to be significantly associated, temperature was determined as a significant factor for most ReGes, as well as Glycerol and Glycerol-phosphate, LCFA levels, and carbohydrates. Overall, an association was often detected for resource parameters that showed altered levels in the November 2011 time-points, which often coincided with altered gene expression around these time-points (a visualization exemplary for a single ReGe to abiotic factor association is shown in **Figure 3.10**). However, the effect size (R^2) of the individually tested parameters remained relatively low.

Due to the frequent association of gene expression profiles with LCFA levels, the ratio of expressed genes assigned to fatty acid degradation divided by all expressed genes per time-point were determined. The *M. parvicella* population's (D51_G1.1.2) comparably high proportions decreased in late November 2011 coinciding with a dramatic increase in free LCFAs (**Figure B.5**). Additionally, *Acinetobacter* spp. and *Leptospira* spp. ReGes (D13_G1.3.2, D15_G4, and A01_O1.2.4) showed increased proportions of fatty acid degradation related gene activity around these time-points (**Figure B.4**). This could indicate the preferred use of fatty acids for the FunC-3 and FunC-4 assigned ReGes as an energy source during these time-points. Notably, a spike in the overall proportion of active genes associated to fatty acid degradation in sample 2011-05-08 coincides with relatively low levels of extracellular LCFAs (**Figure B.5**).

While associations of individual abiotic factors with gene expression profiles could indicate a response of the population to particular resources, comparing expressed functions between all ReGes indicates different lifestyle strategies. In order to compare expression profiles across different ReGes, the concept of the gene-wise activity index was extended to KOs. Numbers of expressed genes with the same KO assignment in each ReGe were summed per time-point. This allowed the combination of ReGe and time-point specific KO profiles. Distances between these profiles based on the respective KO activity were computed. While the overarching structure remained similar to the comparison of functional repertoires, potentially due to the different number of shared KOs

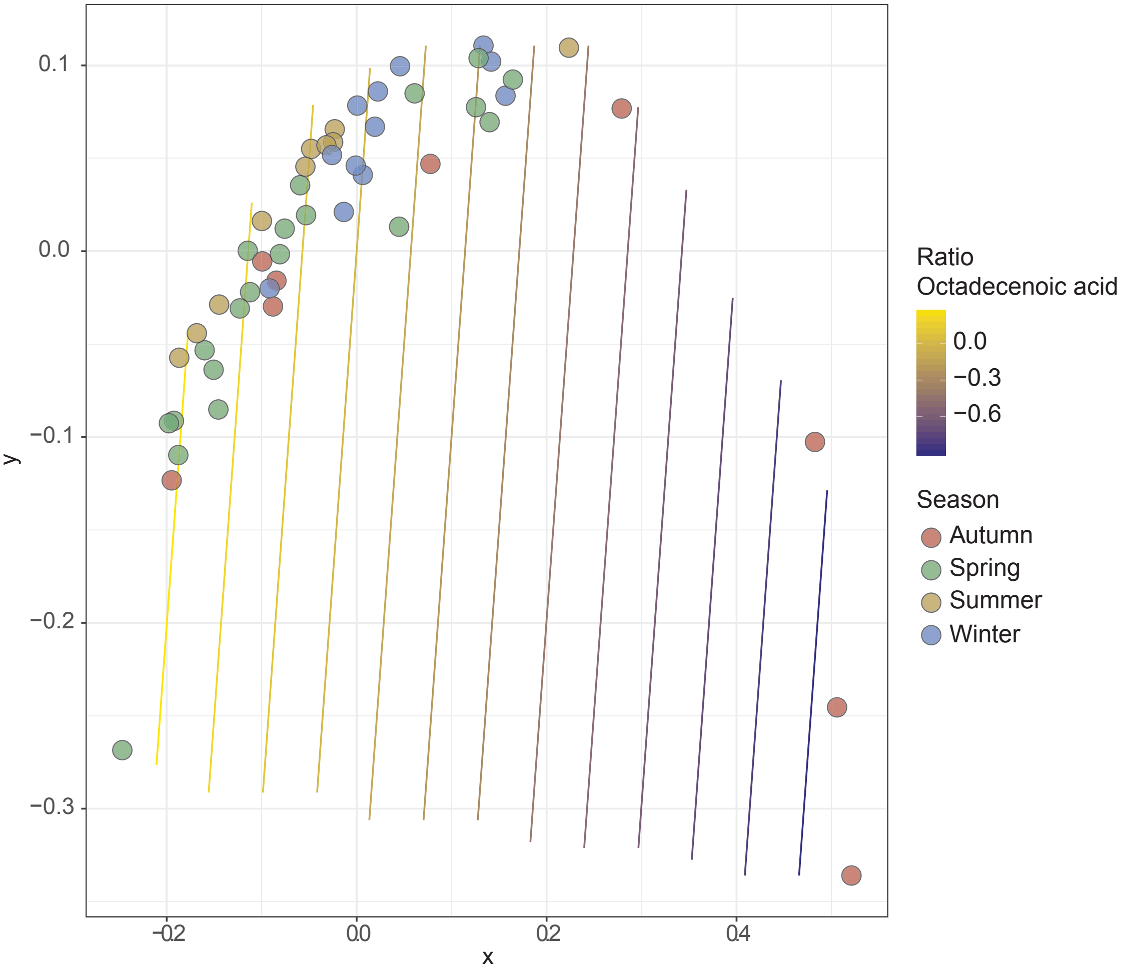


Figure 3.10: MDS of binarized expression profiles D04_G2.5. Multidimensional scaling of distances between binarized expression profiles for ReGe D04_G2.5. Each point represents an expression profile at one of the 51 time-points coloured according to season. Contours mark the levels of the ratio between intra- and extracellular intensities for octadecenoic acid, with higher ratios in yellow and lower ratios marked in purple.

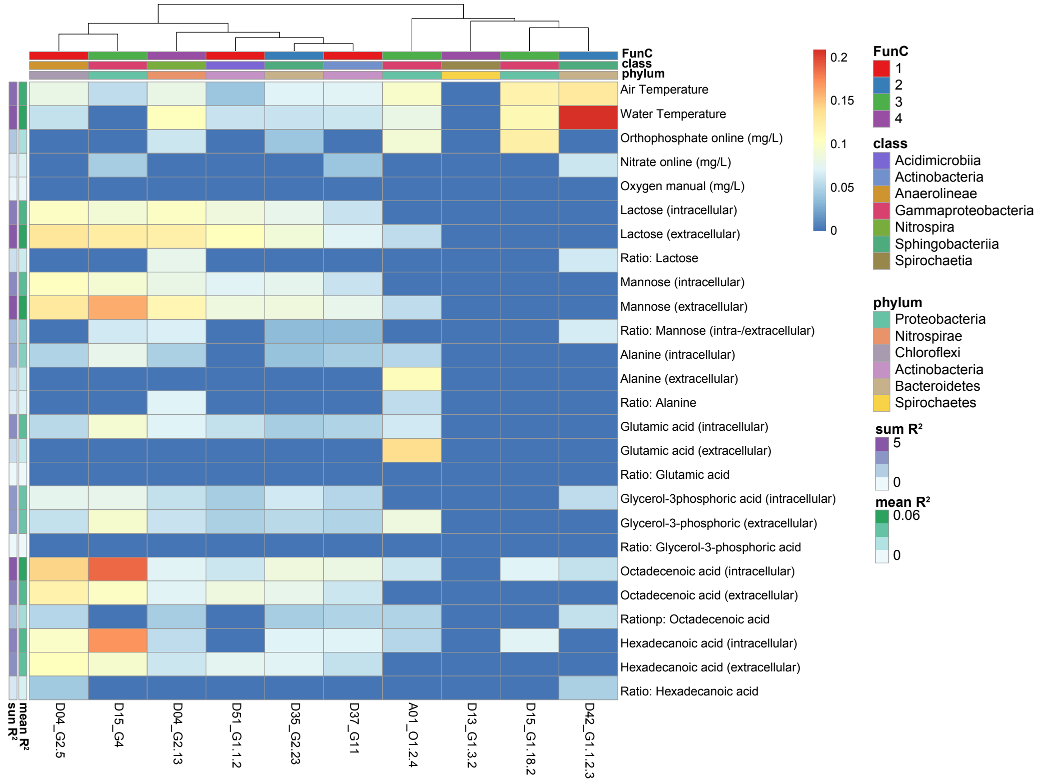


Figure 3.11: Influence of abiotic factors on expression profiles of representative genomes. Selected abiotic factors and their estimated influence on time-point specific gene expression profiles of individual ReGes. Column annotation show taxonomic and FunC assignments. The colour gradient indicates R^2 values of `vegan adonis` analysis for significant (Pr>F < 0.05) parameters, non-significant parameters are shown with $R^2 = 0$. A subset of ReGes and parameters are shown. Row annotations show the sum and mean of the R^2 values derived the full overview, i.e., the filtered set of 78 ReGes (**Figure B.10**).

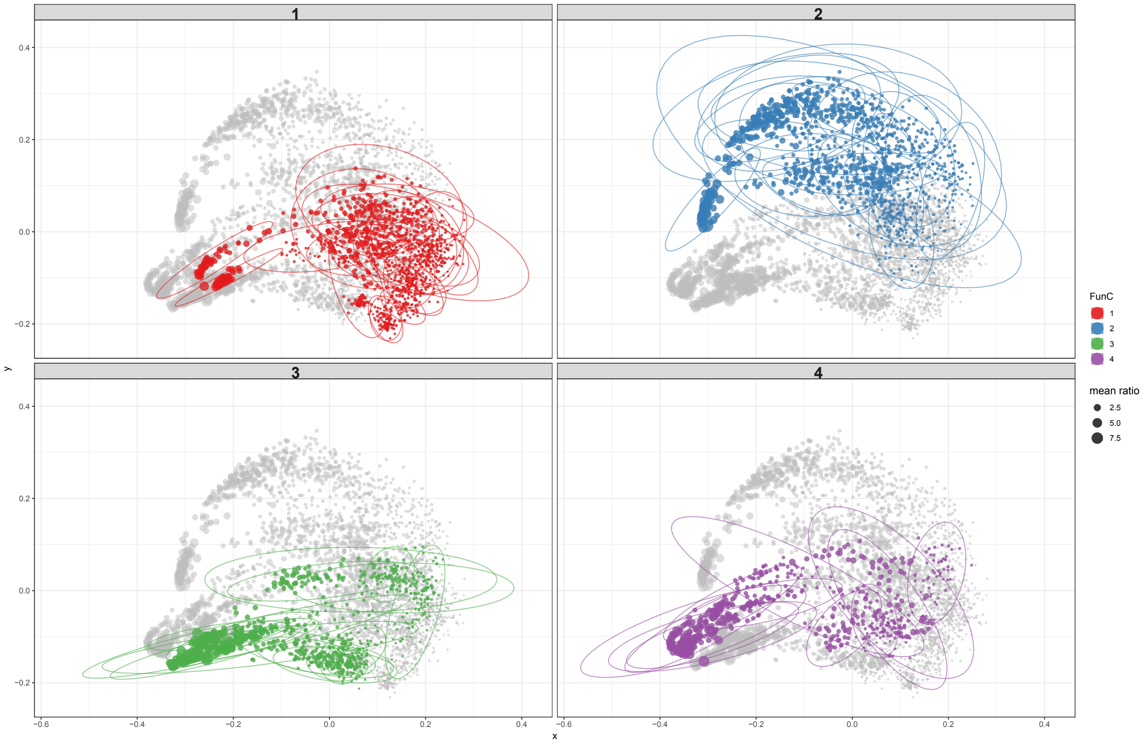


Figure 3.12: KO expression profiles, comparison across all ReGes. Multidimensional scaling of Jaccard-distances between KO profiles of ReGes filtered by completeness (set of 78 ReGes). Each point represents a gene-expression profile (KO level) of a ReGe for a given time-point. For each KO (present in any ReGe) the number of active genes with the respective KO assignment was summed. Ellipses show a 0.95 confidence interval assuming normal distribution. Different panels and colours highlight the FunC association of the ReGes. Point sizes reflect the average gene-wise MT/MG depth ratios for a ReGe at a time-point.

(see **Section 3.4.2**), extensive overlap between the ReGe-specific KO activity profiles could be observed (**Figure 3.12**). This means that even though some populations showed distinct functional repertoires, similar functions are expressed at similar time-points.

Individual ReGes exhibited different levels of variation across their KO activity profiles over time, as highlighted by tightly grouped or widely spread profiles (**Figure 3.12**). Additionally, a grouping according to average MT-depth/MG-depth across all genes could be observed with expression profiles of FunC-1, FunC-3, and FunC-4 associated ReGes showing high average MT/MG ratios throughout the time-series (**Figure 3.12**). The high expression ratios appear to be characteristic traits of the respective organisms and also seem to be relatively robust against fluctuations in population abundance. Overall, comparing expression patterns on a large scale confirmed the assumption that gene expression and thus realized niche space seems to vary to different degrees depending on the individual organisms which could be attributed to differences in niche breadth.

3.4.4 Expression levels in the *in situ* time-series and short-term *in vitro* experiments

The weekly sampling scheme across the *in situ* time-series allowed for the characterisation of long-term trends. To measure short-term responses, *in vitro* experiments were performed in which sampled activated sludge was diluted, incubated, and treated to defined aerobic, anaerobic, or shifting conditions alongside an influx of defined resources (**Section 3.3.2**). Samples were taken 0, 5, and 8 hours after addition of octadecenoic acid, which showed high association to gene expression profiles(**Section 3.4.3**), as well as nutrients (phosphate and nitrate). MT and MG reads from subsequent sequencing of the samples were mapped to the ReGes reconstructed from the *in situ* time-series data. To track expression levels on a more general level, MT over MG ratios of selected ReGes were compared in the defined conditions of the short-term experiments and to the time-series. The ReGes showed distinct patterns of MT/MG ratios in the defined conditions (**Figure 3.13**). ReGes assigned to FunC-3 (A01_O1.2.4, D15_G1.18.2, D15_G4) showed a consistent pattern, with relatively high and more variable MT/MG ratios in the time-series and increasing activity levels in aerobic conditions and lowest in the anoxic conditions. *Anaerolinae* (D04_G2.5) and *Nitrospira* (D04_G2.13) exhibited highest activity in the anoxic condition. D35_G2.23 assigned to FunC-2 and the Sphingobacteriales order, as well as D37_G11, D51_G1.1.2 (FunC-1, Acidomicrobiales order) showed low activity ratios throughout, slightly elevated in the aerobic conditions for the Acidomicrobiales. Overall the MT/MG ratios remained in a comparable range in the short-term experiments and the *in situ* time-series samples for individual ReGes.

To see how genes related to lipid metabolism are regulated MT, expression levels were tracked and compared in the long-term time-series and the short-term experiments. The responses we observed varied distinctly for the individual populations. In aerobic conditions genes related to beta-oxidation, especially acyl-CoA dehydrogenases, were strongly upregulated in the *Microthrix* (D51_G12) and to a lesser extent in closely related ReGe_D37_G11, as well as in D35_G2.23, while the *Acinetobacter*-assigned ReGes (A01_O1.2.4, D15_G1.18.2) and *Anaerolinae* (D04_G2.5) mostly showed rapid upregulation (0-5h) in all the conditions (**Figure 3.14**).

At the same time, genes related to triacylglycerol metabolism are mostly downregulated in aerobic and anaerobic conditions after the addition of the labelled octadecenoic acid (**Figure 3.15**). Notably in D37_G11 upregulation in anaerobic conditions can be observed for diacylglycerol o-acyltransferases. In the long-term time-series clear trends could not readily be identified as the many genes with a function in lipid metabolism, especially for ReGes D04_G2.5, D13_G1.3.2, D37_G11 D51_G1.1.2, highly fluctuated. Some patterns can be observed during the community shift of November 2011, especially for D37_G11 all expression levels of lipid related genes are strongly decreasing during these time-points.

In November 2011 a signal could also be observed in genes related to nitrogen metabolism. In the *Acinetobacter*-related ReGes, as well as in *Microthrix* (D51_G1.1.2) and the related ReGe D37_G11, and in the FunC-4 assigned *Leptospira* ReGe (D13_G1.3.2) ammonium transporters

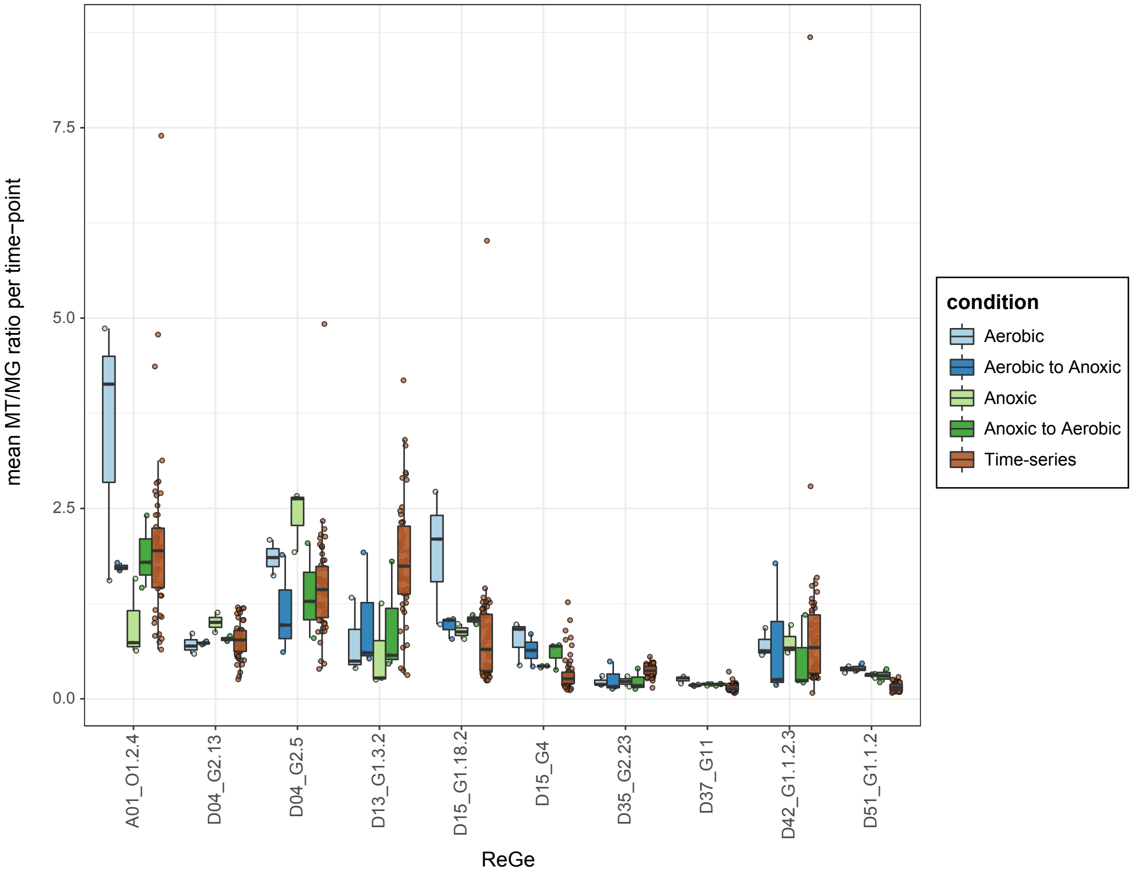


Figure 3.13: Ratios of metatranscriptomic and metagenomic depth of coverage in the defined short-term conditions and the long-term time-series. Metatranscriptomic depth divided by metagenomic depth, MT/MG ratios for a subset of ReGes and 5 different conditions, the long term time-series and the 4 varied aerobic conditions of the short-term experiments. Mean ratios per sample are shown on the y-axis. Colours indicate condition with lightblue: aerobic; darkblue: aerobic preconditioned, shifting; lightgreen: anoxic; darkgreen: anoxic preconditioned, shifting; brown: time-series. Depth of coverage was normalized by dividing by the total read counts in a sample and multiplied by the number of mean total read counts of all samples. To account for the influence of low MG depth, MG depths <1 were set to 1.

and glutamine synthetases are heavily upregulated during November 2011 and the following timepoints (**Figure 3.16**). In the short-term experiments the same genes are correspondingly upregulated in the aerobic condition for the respective ReGes.

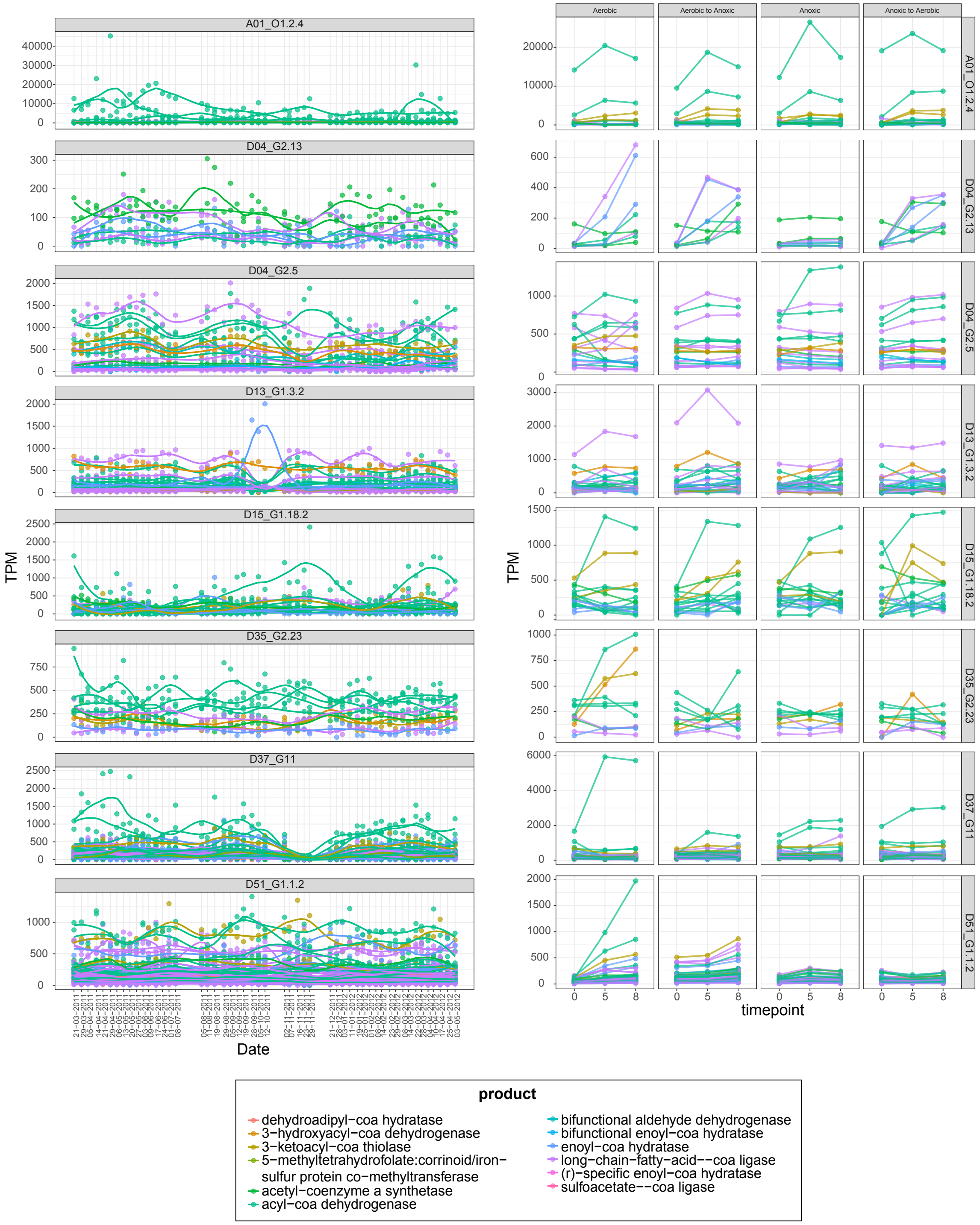


Figure 3.14: ReGe expression levels for genes related to beta-oxidation. Expression levels as transcripts per million (TPM) relative per ReGe. Genes associated to beta-oxidation (links inferred through gene or product-name and EC-number) are shown for the long-term time-series (left-hand side) and the short-term experimental conditions (with conditions separated in different panels; right-hand side). Genes are coloured according to modified product name.

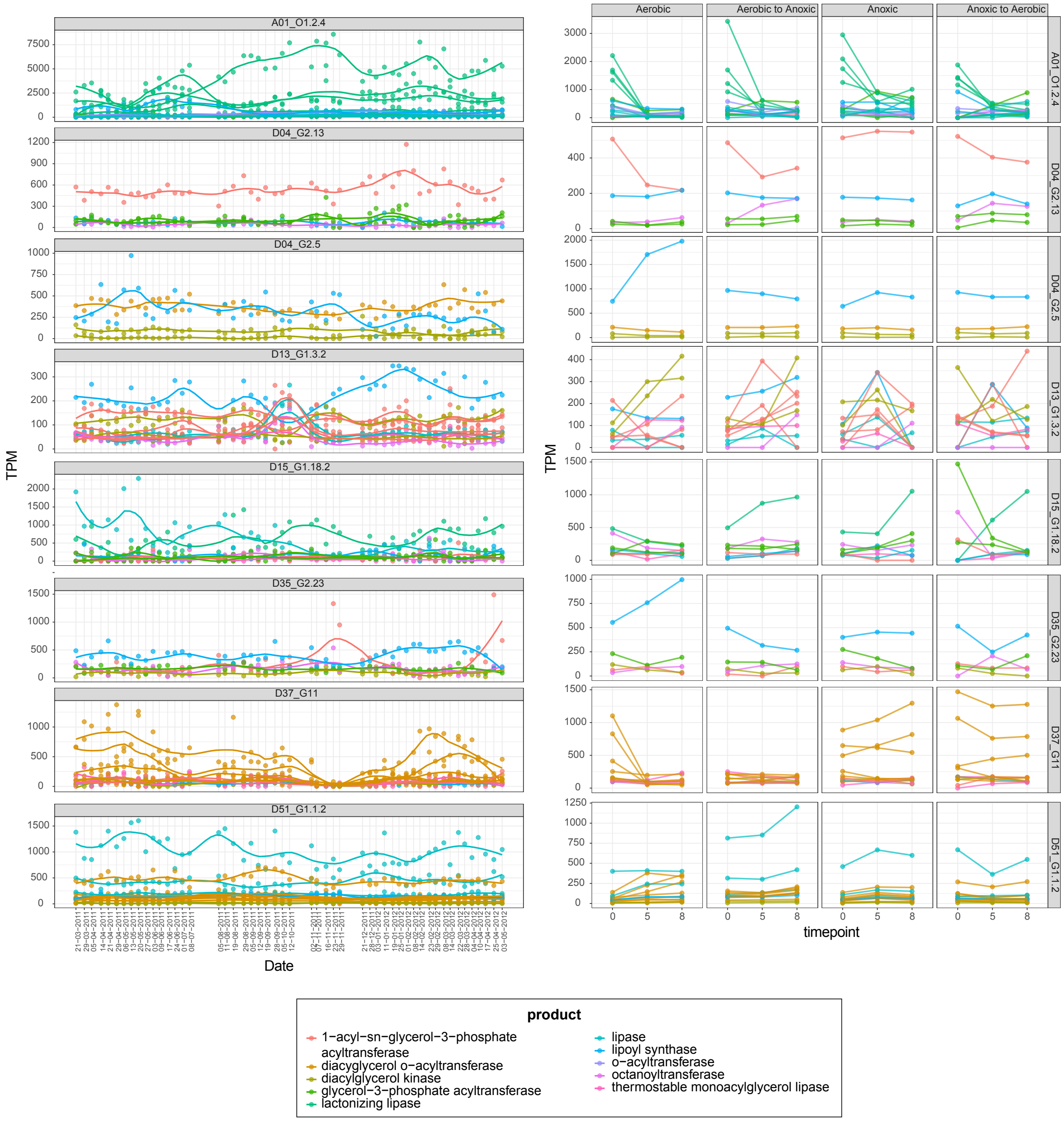


Figure 3.15: ReGe expression levels for genes related to triacylglycerol metabolism. Expression levels as transcripts per million (TPM) relative per ReGe. Genes associated to triacylglycerol metabolism (links inferred through gene or product-name and EC-number) are shown for the long-term time-series (left-hand side) and the short-term experimental conditions (with conditions separated in different panels; right-hand side). Genes are coloured according to modified product name.

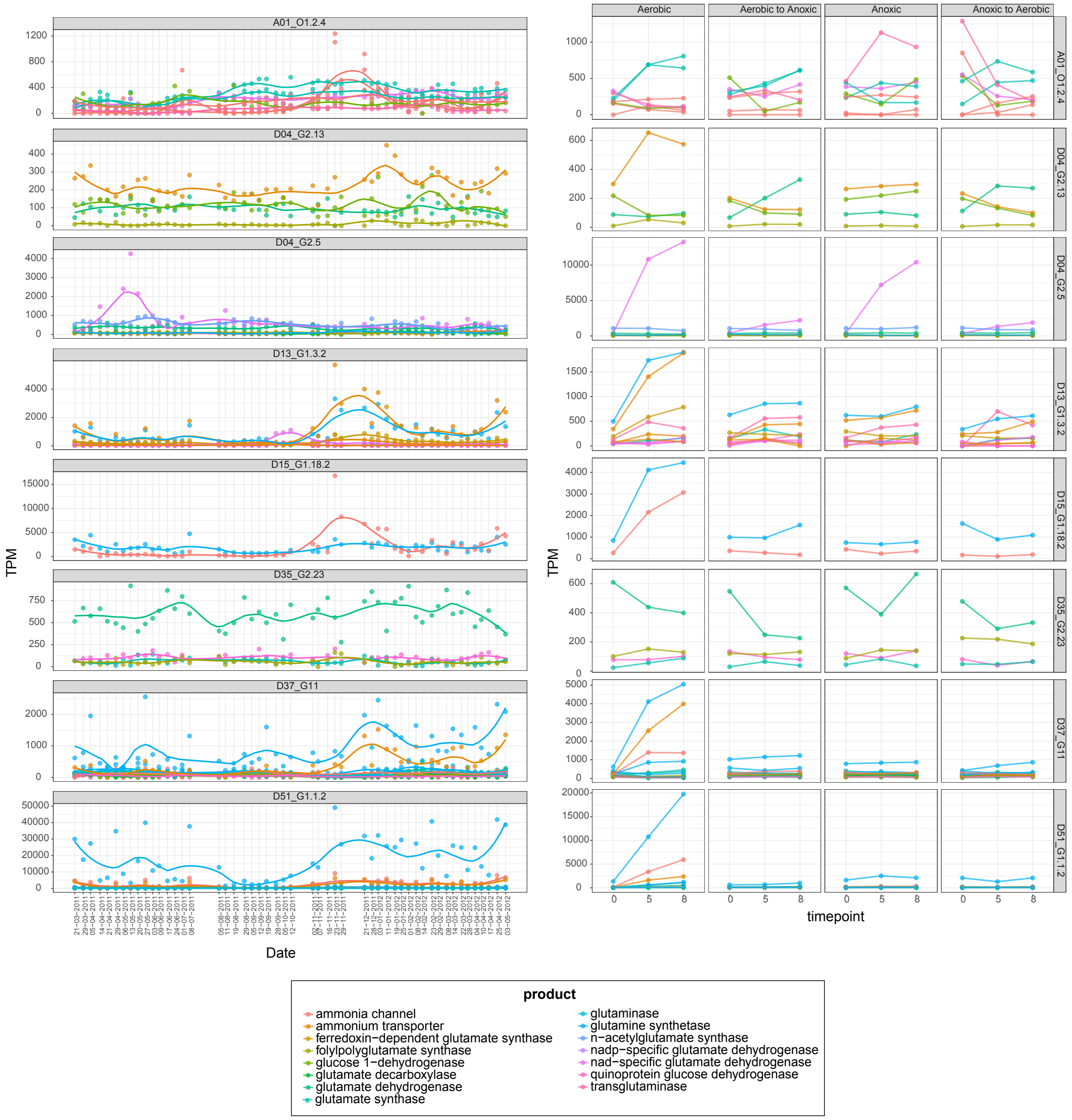


Figure 3.16: ReGe expression levels for genes related to nitrogen metabolism. Expression levels as transcripts per million (TPM) relative per ReGe. Genes associated to glutamine, glutamate, or ammonium metabolism (links inferred through gene or product-name and EC-number) are shown for the long-term time-series (left-hand side) and the short-term experimental conditions (with conditions separated in different panels; right-hand side). Genes are coloured according to modified product name.

3.5 Discussion and outlook

The reconstruction of population-level genomes is a key consideration when integrating functional omics data, i.e., MT, MP and MM, as it allows linking the semi-quantitative readouts to distinct populations. The co-assembly approach of MT and MG data that has been pursued in this work improves contiguity and data usage of the *de novo* assembly over single MG-based assemblies, due to the fact that the addition of MT-assembled contigs faithfully incorporate transcribed regions resulting in additional predictions of full length gene sequences [Narayanasamy et al., 2016]. High-quality genomic bins were recovered from the *in situ* time-series samples, which yield a better representation of the broad community dynamics compared to genomes obtained from isolates, as these reflect rare taxa not present throughout the entire time-series.

The functional potential of the constituent populations of the community is encoded by their respective genetic complement. In this work, we utilized the functional annotation, in the form of KOs, predicted from co-assembled contigs. The resulting FunCs represent types of similar fundamental niches. The groups could be characterized by enriched individual KOs indicative of pathways or functions, such as the ethylmalonyl-CoA pathway in FunC-1 or KOs related to motility in FunC-3. Overall, the functional profiles are not as well distinguishable as expected and not classifiable by characteristic metabolic traits. The fact that only few functional assignments are specific to a FunC (**Figure 3.6**) may indicate a wide and/or shared range of substrates that could be utilized by the individual populations. Interestingly, the populations increasing in their relative abundance in autumn 2011 are related in their functional potential, belonging to FunC-3 and FunC-4, while the populations decreasing in abundance generally belong to two other clusters of functional potential.

In order to delineate if fluctuating levels of available resources can be associated with the shift in population abundances, the available resource space was estimated by utilizing extensive measurements of physico-chemical parameters and untargeted metabolomics. Seasonal patterns in general community profiles corresponded mainly to temperature and phosphate levels (**Figure 3.4**). Interestingly, ammonium levels did not seem to have a substantial effect on shifts in community profiles. Correlation of the levels of abiotic factors and individual abundance profiles revealed broader patterns for ReGes that, on the hand, reflected FunC assignments, but on the other hand also revealed a more fine-grained picture of potentially metabolically related groups. The differences in correlations between metabolite level ratios (intra-/extracellular levels) compared to the individual levels could be indicative of differences in processing of the respective resources, i.e., uptake or storage (higher ratios). However, without the application of higher resolution methods, e.g., directly measuring uptake of metabolites in a population-resolved way, the interpretation remains speculative. Even though, it is not possible to delineate production and consumption of measured metabolites directly, results indicate that temperature, glycerol, carbohydrates, and amino acids are important factors in shaping population structures over time and could play a role in the observed community shift in autumn 2011. Different subsets of ReGes can be associated with these factors, for instance a group consisting primarily of FunC-3, and FunC-4 assigned ReGes showed positive correlations of abundances to LCFA levels and negative correlations to temperature.

ReGe-specific expression profiles were used to estimate shifts in their realized niches, as it has been suggested that niche segregation in microbiota is achieved by transcriptional adaptations [Plichta et al., 2016]. An activity index, even though reflecting a rather simplistic measure, could be used to explore expression profiles over time, thereby integrating MT and MP data. Intra-ReGe comparisons of the resulting expression profiles over time revealed shifting patterns, to which individual abiotic parameters could be associated. While the associated factors depend to a higher degree on individual ReGes and to a lesser extend on the FunC assignment, similar parameters as before exhibited the highest frequency of association, including LCFA levels. Overall, the effect size with that individual parameters could explain variances across time-resolved expression profiles was relatively low and only few metabolite ratios showed significant association. However, microbial populations can respond to a multitude of changing resource parameters and combinations of abiotic factors should be considered for future models.

Furthermore, a inter-ReGe comparison for profiles of expression functions per time-point revealed general trends that could be related to different lifestyles. The degree of variance between expression profiles could point towards differences in niche breadth. Additionally, the trend that the profiles of populations with high MT/MG ratios group suggests that these compete for resources by constitutively high gene expression. This has been indicated to be a feature of metabolic specialists, while metabolic generalists can rely on fine-tuning of gene expression with lower expression levels overall [Muller et al., 2014a]. MT and MG ratios have previously been utilized to assess general activity of individual populations [Hultman et al., 2015]. The gene-wise average of these ratios remains relatively stable for most bins throughout the time-series. For several ReGes, such as several Gammaprotebacteria-assigned ReGes, the MT/MG ratios remain high throughout, while for others such as *M. parvicella*, MT/MG ratios remain low in all time-points. This suggests that MT/MG ratios could be characteristic traits of populations, varying within a certain range depending on the condition and resources. In the *in vitro* short-term experiments characteristic patterns in MT/MG ratio response could be observed depending on the condition. It could be assumed that populations activate metabolic pathways for processing the supplemented nutrients in favourable conditions and thus shower higher expression levels. The fact that for some populations, MT/MG ratios remained low in any of the defined conditions compared to the time-series suggests that the supplemented nutrients did not fall in the spectrum of required resources for these populations. In the MT readouts for the short-term experiments it can be observed that a ReGe assigned to *Acinetobacter* upregulates beta-oxidation related genes regardless of the aerobic condition, while a ReGe assigned to *M. parvicella* primarily upregulates these genes in the aerobic condition, which is in agreement with existing metabolic models for this organism [McIlroy et al., 2013].

Extending the observations of lifestyle strategies of distinct populations points towards the following putative scenario explaining the community shift in autumn 2011. Changes in the environment, as can be observed in increased metabolite levels, decreasing temperatures, and slightly elevated oxygen levels could reflect saturated resource conditions that favour specialist organisms with high expression rates and higher growth rates. These organisms can outcompete the slow-growing generalist populations by quicker adaptation to the high availability of resources, resulting in a shift in observed abundances. With competition among the initially successful populations and fluctuating conditions in the long-term this advantage cannot be maintained triggering a return to a state of the community reflecting the success of generalist species. Populations exhibiting high levels of gene expression could face drawbacks in sparse nutrient condition due to high investment in unneeded transcripts and proteins, while an organism with fine-tuned and overall lower expression levels could benefit [Muller et al., 2014a].

Another factor potentially related to the community shift could be the limitation of nitrogen. Compounds like ethanolamine or putrescine related to catabolism of amino acids are highly increased in the late November 2011 time-points, coinciding with reduced intracellular amino acid levels. Correspondingly, an upregulation of genes involved in nitrogen metabolism, e.g. nitrogen regulatory proteins or glutamine synthetase, as well as ammonium transporters is observed. This could indicate an increased requirement for nitrogen for the bacterial populations potentially associated to a phase of higher growth rates. Measured ammonium levels do not exhibit fluctuations during this period, however the levels measured in the aeration tank might not be reflective of the ammonium levels in the floating sludge islands at the surface.

The balance between maintaining growth rates and the accumulation phenotype to maintain a competitive advantage in the fluctuating environment is of great importance for potential downstream biotechnological applications harnessing the accumulated lipids. As sparse nutrient conditions seem to be favourable for LAOs, a spatial separation of organisms that pursue these different lifestyle strategies along resource gradients could prevent unintended community perturbations [Sheik et al., 2014]. While implementing these aspects will have to be addressed by engineering, the methods presented in this work could provide the foundation for approaches to a required delineation of niche ecology by omics data integration.

As a future perspective, a more accurate description of community dynamics and trophic relationships shaping the system could be achieved by incorporation additional quantitative MM measurements, ,i.e., the compounds for which potential associations were predicted could be measured by targeted approaches or incorporation of labelled metabolites could be tracked. Future work will have to confirm the seasonal effects observed in this work, by extending the time-series. Observed results could be affirmed by incorporating biological replicates, however replicates and the extent of the time-series and will have to be balanced due to the costs of extensive omics profiling. Another important avenue, is considering additional operational parameters in the BWWTP, which could cause perturbations with the microbial communities. E.g., sludge retention time was reduced in August 2011 coinciding with changes in the measured phyisoc-chemical parameters, especially for dry-weight and conductivity (**Figure 3.8**).

The results described here only reflect a small fraction of inferences that could be made by further analysing the rich multi-omics dataset (**Chapter 4**). While in principle the WGS sequencing based approach is also well suited for the assessment of genetic variation over time, the main interest in this work was to characterize populations on a general functional and trophic level. Genetic variation has also not been of a primary interest, as it was expected to be relatively low compared to other systems and for the dominant organism *M. parvicella* slow growth rates, i.e., a doubling time of around eight days, and highly clonal populations have previously been observed [Muller et al., 2014a]. However, evolutionary aspects are of course also expected to shape the analysed microbial community and genetic variation and strain-level analyses could be performed (**Section 4.2.1**). Furthermore, the expression or transference of anti-microbial resistance within this system could be assessed in future studies. On the one hand, anti-microbial resistance genes or toxin/antitoxin systems are important factors in shaping community structure with various mechanisms [Riley and Wertz, 2002; Harms et al., 2018], while on the other hand BWWTPs have been implied in the spread and emergence of antibiotic resistance relevant to human health [Rizzo et al., 2013].

Additionally, community dynamics are also shaped by predator-prey relationships between bacterial hosts and phages, for instance the success of a dominant population could be limited in a "kill the winner"-scenario [Thingstad, 2000] that could also explain drastic shifts in a community as was observed here. The rich multi-omics dataset described herein is ideally suited to characterize the dynamics of the population-resolved CRISPR complement in relation to mobile elements and phages [Martinez Arbas and Narayanasamy *et al.* - in preparation].

CHAPTER 4

GENERAL CONCLUSIONS AND FUTURE PERSPECTIVES

Parts of this chapter are based on the following peer-reviewed publication:

- Emilie E.L. Muller, Karoline Faust, Stefanie Widder, **Malte Herold**, Susana Martinez Arbas, Paul Wilmes (2018). Using metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in Systems Biology* **8**: 73-80.

  [**Appendix C.6**]

4.1 General perspectives

Microbial communities are complex and dynamic systems shaped by environmental parameters (**Section 1.1**). Recent high-throughput measurement techniques for characterising the biomolecular components of microbiomes hold the potential to assess their composition, functional potential, and activity at an unprecedented level of detail (**Section 1.3**). Yet, the integration of the heterogeneous datasets that MG, MT, MP, and MM data represent, is a challenging task and requires development of efficient bioinformatic approaches, especially in combination with time-series analysis (**Chapter 3**).

Herein, two model systems were analysed with the aim of characterising microbial niches of distinct populations by detailing their genomic potential and expression of specific functions of relevance in a biotechnological context. In **Chapter 2** the metabolic capabilities of an isolate strain were analysed by sequencing the complete genome and subsequent gene and function predictions. Functional omics data (MT, MP) was used to characterise the lifestyle of the strain in culture medium and growing on chalcopyrite. Furthermore, mixed cultures of other acidophiles were assessed in the context of chalcopyrite bioleaching. The dataset analysed in **Chapter 3** represents a time-series of lipid accumulating organisms in wastewater. Relationships between different populations, their functional potential, and gene expression in association to shifting abiotic factor levels was detailed by integrating multi-omic data.

In this work, emphasis was directed towards population-level analyses, i.e., the characterisation of constituent members of microbial communities. Specifically, a pronounced difference between the herein studied systems lies in their diversity. The bioleaching environment represents rather extreme conditions and is thus associated with a lower diversity [Baker and Banfield, 2003] compared to the BWWTP environment. Accordingly, while the analyses share many commonalities, not least the integration of multi-omic data, the aim of obtaining an in-depth understanding of microbial niches was pursued from two different angles: a) reconstruction of a model system based on known populations b) reconstruction of populations from an *in situ* system.

These approaches have also been described as bottom-up and top-down approaches [Zengler, 2009]. Naturally, they provide a different level of resolution on the biological mechanisms studied, and should ideally be applied in a complementary way to characterise microbial communities [Zengler, 2009]. Using cultivable strains in defined settings allowed a detailed characterisation of functional potential and expressed functions, which led to inferences that potentially could be scaled up and tested in a setting closer the foreseen application of biomining (**Section 2.5.2**). On the other hand, the patterns observed in large-scale analyses can determine broader functional profiles of constituent populations in microbial communities. Due to the multitude of potential interactions within complexer consortia, a more mechanistic understanding of metabolic processes can be obtained by transferring environmental samples to defined conditions. These microcosm experiments resemble a bridge between top-down and bottom-up approaches. Here, community samples are cultured in conditions resembling the *in situ* environment, usually for a certain period of time. Similarly to the approach described in **Section 3.4.4** this allows targeted perturbations of the system e.g. by varying conditions, such as oxygen or temperature or addition of defined or labelled nutrients.

Integrating data derived from different experimental setups or different temporal and spatial scales, is a key challenge in defining a framework capable of predicting community phenotypes in complex consortia. The ability to characterize microbial niches over time and in changing environmental conditions will enable us to systematically define and alter the realised niches of constituent populations in situ and manage community conferred traits, leading to exciting prospects for biotechnology [Muller et al., 2018].

4.2 Recovery of representative genomes for distinct populations

For *in vitro* experiments with synthetic microbial communities as described in **Chapter 2**, commonly reference genomes derived from isolate culture of the utilized strains are available from databases such as NCBI GenBank [Clark et al., 2016] or the JGI IMG/M database [Markowitz et al., 2014]. However, non-model organisms are typically underepresented or completely missing from the reference databases. With a decrease in cost and improvements in DNA sequencing methods in recent years (**Section 1.3.1**) obtaining high-quality reference genome sequences from isolate culture has become feasible, as was demonstrated in **Section 2.4.1** for *L. ferriphilum* for which a circular chromosomal contig of 2.5 Mbp could directly be assembled from third-generation sequencing reads. Furthermore, long-read sequencing methods have become a cost-efficient method to assemble isolate microbial genomes with high accuracy [Liao et al., 2015] and can be extended by methylation profiles [Wibberg et al., 2016; Fomenkov et al., 2017]. Assembly methods have to be further refined for being applicable in MG sequencing, but have also been successfully utilized to recover complete genomes from low-diversity microbial communities [Driscoll et al., 2017]. However, given the higher requirements in terms of quantity and quality of DNA [van Dijk et al., 2018], they are often limited to cultured isolates.

An advantage of *de novo* assembly-based approaches in general is the possibility to discover previously uncharacterised organisms [Laczny et al., 2016; Tully et al., 2018; Delmont et al., 2018], thus truly assessing microbial diversity [Keller and Zengler, 2004]. As demonstrated herein, this allows the analysis of the respective organisms' functional potential as well activity without the requirement of a priori assumptions. However, MAGs derived from the environmental samples are in general of lower quality in terms of completeness and contamination compared to genomes derived from isolate sequencing. While several high-quality MAGs were recovered herein (**Chapter 3**), various MAGs remained incomplete or potentially contaminated, requiring stringent filtering criteria.

Tracking populations over time posed an additional challenge for recovering representative MAGs. Initial attempts with multiple sample binning that relied on abundance profiles [Nielsen et al., 2014; Alneberg et al., 2014] were unable to resolve population-level genomes, as was expected due to their requirement of independent samples which was not satisfied by the time-series dataset. Linking MAGs from different time-points or connected samples, was demonstrated by connecting MAGs through essential marker gene content [Wampach *et al.* 2018 - **Appendix C.9**] or through genome-wide signature comparison [Olm et al., 2017] yielding representative genomes as shown in **Chapter 3**.

4.2.1 Improving the reconstruction of population-level genomes

State-of-the-art assembly and binning strategies have shown to enhance MAG quality [Parks et al., 2017]. Recently, an ensemble approach that combines several binning tools has shown to outperform the individual methods in terms of number, completeness, and contamination of MAGs recovered from microbial communities of different environments and varying complexity [Sieber et al., 2018]. This approach could be implemented in two settings, either by optimizing the quality of bins derived from a single samples and/or by combining bin sets across multiple samples.

For the analysis presented in **Chapter 3**), more complete genomic reconstructions could allow estimation of growth rates based on MG coverage [Brown et al., 2016]. As the results point towards differences in growth strategies of distinct populations across different season, this could be an important direction for future work. Additionally, combining isolate culture and MG/MT derived genomes is expected to provide additional information, e.g., by using the isolate references for a guided assembly to recover MAGs of lowly abundant taxa [Cepeda et al., 2017]. Additionally, metabolic reconstructions obtained from MAGs can also be applied to identify required conditions for isolate cultures [Pope et al., 2011].

Furthermore, time-series analysis could be augmented by assessing strain-level dynamics, e.g., by profiling strain-specific single nucleotide variants in combination with species-specific marker genes [Truong et al., 2017]. Strain-level analyses would provide a more fine-grained picture, and could highlight phenotypic differences due to intra-species genomic variation [Mallick et al., 2017]. An approach to track genomic variation, while maintaining an population-level association could also be realized by pursuing a pangenome-based approach [Delmont and Eren, 2018] instead of selecting representative MAGs from individual time-points (**Chapter 3**). However, strain-level based analysis are typically restricted to genomic regions that are found within the species, i.e, single nucleotide variants in shared genes [Truong et al., 2017]. Strains can differ also in the genes or plasmids they contain, which remains largely unexplored to date.

4.3 Genome annotation

Genome annotation is an important subsequent step to the recovery of population-level genomes and is required to characterise the functional complement of the sequenced population. In general, the position of genetic features, such as CDS, tRNAs, rRNAs or other coding and non-coding regions of interest are determined within the genome sequence, commonly by automated computational tools. Reliable annotation, also of protein functions, is a prerequisite for any inferences on functional repertoire and gene expression. While automatic tools can be applied for the accurate annotation of large-scale datasets, computational predictions can also introduce erroneous annotations, e.g., by propagation of misannotations from existing databases [Richardson and Watson, 2013]. In this work, multiple sources or databases were used to generate the respective functional annotations. As the combination and weighting of multiple annotation sources poses unsolved challenges, manual curation by experts is often required, yet only feasible for a limited number of genomes. As described in **Section 2.3.5**, results of multiple automated prediction tools were utilized to manually augment existing annotations and infer protein functions and assignments to functional categories or phenotypic traits. However, large-scale datasets (**Chapter 3**) preclude comprehensive manual curation of functional annotations. Here, functional annotations were weighted by probability of correct assignment and the best-scoring protein annotation was selected (**Section 3.3.7**). The inference of functional categories and protein-to-reaction links was based on the of union several annotations favouring sensitivity over specificity. Similarly, the characterisation of the expression for specific gene functions was based on multiple annotated features (**Section 3.4.4**).

Methods for consolidating annotations of multiple sources could be used to improve functional predictions [Cozzetto et al., 2013]. Furthermore, also the incorporation of omics data can be applied to improve annotations (**Section 1.3.5**). Omic data aided driven approaches could could be incorporated to determine putatively false annotations. Or combined in an accessible framework omic data could be used to visually inspect ambiguous cases, e.g. with divergent automated predictions.

4.4 Data management

Omics measurements and subsequent processing with bioinformatic tools and pipelines can generate massive amounts of data. Therefore, in addition to computationally efficient processing of these datasets, strategies for data management are of key concern. Public databases for storage and access of omics data have been established, such as the NCBI short read archive [Leinonen et al., 2011] for sequencing data or the PRIDE repository for proteomics data [Jones, 2006]. These however, are often not well suited for combining heterogeneous data types and frameworks for more flexible solutions are required [Bauch et al., 2011]. Organism specific databases exist that offer warehousing and web-service for analysis Kalderimis et al. [2014]. In addition, platforms for collaborative data management in systems biology projects have emerged [Wolstencroft et al., 2017].

Specifically for omics data integration, data management is a key concern. Databases capable of storing and associating data of different omic levels can greatly facilitate analyses. Non-relational databases are suitable to store omics data associations to due their flexibility in data structures. In this work, a database was constructed to link contigs to, e.g., gene associations and bin assignments, as well as functional annotations for which it was primarily used in the analysis workflow (**Section 3.3.7**). This approach could be extended by associating functional readouts, such as MT read-counts or MP spectral counts per sample. A flexible database scheme could be a powerful tool for omics data integration in time-series analyses and could be combined with existing platforms that provide methods for statistical multi-omics analyses and visualisation such as anvio'o [Eren et al., 2015] or mixOmics [Rohart et al., 2017].

4.5 Integration of omics data in community models

Building on a detailed characterisation of microbial communities with omic data, computational modelling approaches are envisaged to provide a solution for disentangling the complex dynamics in microbial systems and for obtaining a predictive understanding required for controlling community phenotypes [Widder et al., 2016]. Yet, a missing link between empirical data and theoretical models needs to be overcome and the integration of omic data in models is an area of great interest [Widder et al., 2016].

Metabolic modelling methods can broadly be grouped into two different groups, stoichiometric approaches that model metabolism quantitatively [Bordbar et al., 2014] and topological or network-based methods more suitable for qualitative modelling [Faust et al., 2011]. Stoichiometric models can be based on metabolic reconstructions derived from genomic data, i.,e., gene-to-reaction relationships and curated with data derived from physiological experiments [Feist et al., 2009] and frameworks for modelling microbial communities have been developed [Khandelwal et al., 2013]. While the generation of comprehensive genome scale models requires manual curation [Thiele and Palsson, 2010], which makes the application for complex communities challenging, recent efforts to automate curation steps showed that reconstructing a large set of genome scale models can be feasible [Magnúsdóttir et al., 2016]. Theoretical approaches have been used to elucidate ecological interactions by potential exchanges of metabolites [Zelezniak et al., 2015]. Algorithms for the integration of omic datasets for stoichiometric models exist for transcriptomics, proteomics, or metabolics data. However, a framework for an integrated multi-omics driven approach is still lacking.

Network-based approaches in microbial ecology have primarily focused on the integration population abundances to formulate co-occurence networks Berry and Widder [2014]. These methods are well established for characterising time dynamics in ecology [Faust and Raes, 2012]. Limitations of these approaches include for example that the interaction strengths between different species is often assumed to be constant [Faust and Raes, 2012]. Additionally, networks based on co-expression data can be difficult to interpret, as observed correlations might arise from indirect interactions. Similarly, network-based approaches can also be pursued by constructing models directly from various omics data types. For example, co-expression networks can be generated from proteomic or transcriptomic data [Manzoni et al., 2018]. Techniques to integrate networks derived from transcriptomics and proteomics have been described [Walley et al., 2016] and could be pursued in the context of microbial communities.

Both datasets described within this work, could be well suited to be applied for omics integration in computational modelling approaches. The synthetic bioleaching communities could be used to characterise metabolic interactions based on detailed metabolic reconstructions inferred from the in-depth characterisation of functional potential. Also previous metabolic reconstructions could be utilized as reference points, as a stoichiometric model for *Leptospirillum ferroxidans* has been formulated [Merino et al., 2010]. Also, models for multiple bioleaching strains could be combined, highlighting the interactions of hetero- and autotroph species [Merino et al., 2014]. Similar approaches could be extended with functional omics data, as well as the physiological leaching data for the system consisting of three organisms. However, also resolving interactions on a spatial scale, i.e., on the biofilm level, is important in the system due to the relevance of attachment in bioleaching [Rohwerder et al., 2003]. This could be achieved with particle-based modelling. Classifying interactions between bioleaching organisms could also be realized by co-expression analyses. Here, the reduced complexity of the system and the possibility to formulate data-driven hypotheses on interactions from the omics data (**Chapter 2**) could be beneficial for validating network-derived predictions.

It remains to be seen if stoichiometric modelling approaching could also be applied for the LAO time-series data. The quality of the recovered population-level genomes could pose a challenge, even if it could be improved (**Section 4.2.1**). However, also gapfilling methods that can be applied on an individual and/or on a community level [Henry et al., 2016] could be tested. A comprehensive set of reference genome scale models, as exists for the human gut [Magnúsdóttir et al., 2016], is lacking for the wastewater system. Nonetheless, it would be interesting to see, if community scale models can reproduce different life-style strategies inferred from the reconstruction of metabolic niches. The different growth strategies of the individual populations could pose an additional challenge as some community modelling frameworks can only account for this in a well-defined system [Gottstein et al., 2016]. Due to the temporal nature data, non-model organisms, and the complexity of the system, topological or network-based approaches seem more feasible, at least at an initial stage. Data-driven metabolic network reconstruction on the LAO system has been applied for detecting keystone genes at individual time-points [Roume et al., 2015] and it could be promising exploring a similar concept in a time-resolved manner.

A critical consideration for multi-omics integration especially for computational modelling would be the signal to noise ratio in the data. Strategies to classify and filter noise that have been described for dynamic co-abundance models [Faust et al., 2018] could also allow to asses noise in multi-omic data derived networks.

Future augmented community-level metabolic models need to account for trophic interactions, changing environmental conditions, and spatial scales ideally by integrating dynamic community models with genome-scale metabolic models [Muller et al., 2018].

BIBLIOGRAPHY

Florence Abram. Systems-based approaches to unravel multi-species microbial community functioning. *Computational and Structural Biotechnology Journal*, 13:24–32, dec 2015. ISSN 20010370. doi: 10.1016/j.csbj.2014.11.009.

Peter B. Adler, Janneke HilleRisLambers, and Jonathan M. Levine. A niche for neutrality. *Ecology Letters*, 10(2):95–104, feb 2007. ISSN 1461-023X. doi: 10.1111/j.1461-0248.2006.00996.x.

Jeffrey N. Agar, Pramvadee Yuvaniyama, Richard F. Jack, Valerie L. Cash, Archer D. Smith, Dennis R. Dean, and Michael K. Johnson. Modular organization and identification of a mononuclear iron-binding site within the NifU protein. *Journal of Biological Inorganic Chemistry*, 5(2):167–177, 2000. ISSN 09498257. doi: 10.1007/s007750050361.

Mads Albertsen, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538, 2013. ISSN 1087-0156. doi: 10.1038/nbt.2579.

Eric E Allen and Jillian F Banfield. Community genomics in microbial ecology and evolution. *Nature reviews. Microbiology*, 3(6):489–498, 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1157.

Rodrigo J. Almárcegui, Claudio A. Navarro, Alberto Paradela, Juan Pablo Albar, Diego Von Bernath, and Carlos A. Jerez. New copper resistance determinants in the extremophile acidithiobacillus ferrooxidans: A quantitative proteomic analysis. *Journal of Proteome Research*, 13(2):946–960, 2014. ISSN 15353893. doi: 10.1021/pr4009833.

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3103.

Munir A. Anwar, Mazhar Iqbal, Muhammad A. Qamar, Moazur Rehman, and Ahmad M. Khalid. Technical communication: Determination of cuprous ions in bacterial leachates and for environmental monitoring. *World Journal of Microbiology and Biotechnology*, 16(2):135–138, 2000. ISSN 09593993. doi: 10.1023/A:1008978501177.

Edward Ardern and William T. Lockett. Experiments on the oxidation of sewage without the aid of filters. *Journal of the Society of Chemical Industry*, 33(10):523–539, may 1914. ISSN 03684075. doi: 10.1002/jctb.5000331005.

Armando Azua-Bustos and Carlos González-Silva. Biotechnological Applications Derived from Microorganisms of the Atacama Desert. *BioMed Research International*, 2014, 2014. ISSN 23146141. doi: 10.1155/2014/909312.

Lourens Gerhard Marinus Baas-Becking. *Geobiologie of inleiding tot de milieukunde*. WP Van Stockum & Zoon NV, 1934.

Nabih Baeshen, Mohammed N Baeshen, Abdullah Sheikh, Roop S Bora, Mohamed Morsi M Ahmed, Hassan a I Ramadan, Kulvinder Singh Saini, and Elrashdy M Redwan. Cell factories for insulin production. *Microbial cell factories*, 13:141, 2014. ISSN 1475-2859. doi: 10.1186/s12934-014-0141-0.

Brett J. Baker and Jillian F. Banfield. Microbial communities in acid mine drainage. *FEMS Microbiology Ecology*, 44(2):139–152, may 2003. ISSN 01686496. doi: 10.1016/S0168-6496(03)00028-X.

Angela Bauch, Izabela Adamczyk, Piotr Buczek, Franz-Josef Elmer, Kaloyan Enimanev, Pawel Glyzewski, Manuel Kohler, Tomasz Pylak, Andreas Quandt, Chandrasekhar Ramakrishnan, Christian Beisel, Lars Malmstrom, Ruedi Aebersold, and Bernd Rinn. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12 (1):468, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-468.

Eugen Bauer, Cedric Christian Laczny, Stefania Magnusdottir, Paul Wilmes, and Ines Thiele. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome*, 3:55, 2015. ISSN 2049-2618. doi: 10.1186/s40168-015-0121-6.

Sören Bellenberg, Antoine Buetti-Dinh, Vanni Galli, Olga Ilie, Malte Herold, Stephan Christel, Mariia Boretska, Igor V Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, and Mark Dopson. Automated Microscopic Analysis of Metal Sulfide Colonization by Acidophilic Microorganisms. *Applied and Environmental Microbiology*, 84(20):AEM.01835–18, aug 2018. ISSN 0099-2240. doi: 10.1128/AEM.01835-18.

Peter Belmann, Johannes Dröge, Andreas Bremges, Alice C. McHardy, Alexander Sczyrba, and Michael D. Barton. Bioboxes: standardised containers for interchangeable bioinformatics software. *GigaScience*, 4(1):47, dec 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0087-0.

Susana Benlloch, Arantxa López-López, Emilio O. Casamayor, Lise Øvreås, Victoria Goddard, Frida Lise Daae, Gary Smerdon, Ramón Massana, Ian Joint, Frede Thingstad, Carlos Pedrós-Alió, and Francisco Rodríguez-Valera. Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environmental Microbiology*, 4(6):349–360, 2002. ISSN 14622912. doi: 10.1046/j.1462-2920.2002.00306.x.

Albert F Bennett and Richard E Lenski. EVOLUTIONARY ADAPTATION TO TEMPERATURE II. THERMAL NICHES OF EXPERIMENTAL LINES OF ESCHERICHIA COLI. *Evolution*, 47(1):1–12, feb 1993. ISSN 00143820. doi: 10.1111/j.1558-5646.1993.tb01194.x.

David Berry and Stefanie Widder. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5(MAY):1–14, 2014. ISSN 1664302X. doi: 10.3389/fmicb.2014.00219.

Robert C. Blake and Megan N. Griff. In situ spectroscopy on intact Leptospirillum ferrooxidans reveals that reduced cytochrome 579 is an obligatory intermediate in the aerobic iron respiratory chain. *Frontiers in Microbiology*, 3(APR):1–10, 2012. ISSN 1664302X. doi: 10.3389/fmicb.2012.00136.

Martin J Blaser, Zoe G Cardon, Mildred K Cho, Jeffrey L Dangl, Timothy J Donohue, Jessica L Green, Rob Knight, Mary E Maxon, Trent R Northen, Katherine S Pollard, and Eoin L Brodie. Toward a Predictive Understanding of Earth ' s Microbiomes to Address 21st Century Challenges. *American Society for Microbiology*, 7(3):1–16, 2016. ISSN 2150-7511. doi: 10.1128/mBio.00714-16.

Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu170.

Violaine Bonnefoy and David S. Holmes. Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. *Environmental Microbiology*, 14(7):1597–1611, 2012. ISSN 14622912. doi: 10.1111/j.1462-2920.2011.02626.x.

Aarash Bordbar, Jonathan M Monk, Zachary a King, and Bernhard O Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature reviews. Genetics*, 15(2): 107–20, 2014. ISSN 1471-0064. doi: 10.1038/nrg3643.

Robert M. Bowers, Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T B K Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, Susannah G. Tringe, Natalia N. Ivanova, Alex Copeland, Alicia Clum, Eric D. Becraft, Rex R. Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M. Weinstock, George M. Garrity, Jeremy A. Dodsworth, Shibu Yooseph, Granger Sutton, Frank O. Glöckner, Jack A. Gilbert, William C. Nelson, Steven J. Hallam, Sean P. Jungbluth, Thijs J G Ettema, Scott Tighe, Konstantinos T. Konstantinidis, Wen-Tso Liu, Brett J. Baker, Thomas Rattei, Jonathan A. Eisen, Brian Hedlund, Katherine D McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W. Tyson, Christian Rinke, Nikos C. Kyrpides, Lynn Schriml, George M. Garrity, Philip Hugenholtz, Granger Sutton, Pelin Yilmaz, Folker Meyer, Frank O. Glöckner, Jack A. Gilbert, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H. Parks, A. M. Eren, Lynn Schriml, Jillian F. Banfield, Philip Hugenholtz, and Tanja Woyke. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, 35(8):725–731, aug 2017. ISSN 1087-0156. doi: 10.1038/nbt.3893.

Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, may 2016. ISSN 1087-0156. doi: 10.1038/nbt.3519.

Andreas Bremges, Irena Maus, Peter Belmann, Felix Eikmeyer, Anika Winkler, Andreas Albersmeier, Alfred Pühler, Andreas Schlüter, and Alexander Sczyrba. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *GigaScience*, 4(1):33, 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0073-6.

C. Titus Brown and Luiz Irber. sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software*, 1(5):27, sep 2016. ISSN 2475-9066. doi: 10.21105/joss.00027.

Christopher T Brown, Matthew R Olm, Brian C Thomas, and Jillian F Banfield. In situ replication rates for uncultivated bacteria in microbial communities. *bioRxiv*, (November):057992, 2016. ISSN 1087-0156. doi: 10.1101/057992.

Dmytro Bykov and Frank Neese. Six-Electron Reduction of Nitrite to Ammonia by Cytochrome c Nitrite Reductase: Insights from Density Functional Theory Studies. *Inorganic Chemistry*, 54 (19):9303–9316, 2015. ISSN 1520510X. doi: 10.1021/acs.inorgchem.5b01506.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.

J. P. Cardenas, M. Lazcano, Francisco J Ossandon, Melissa Corbett, David S Holmes, and E. Watkin. Draft Genome Sequence of the Iron-Oxidizing Acidophile Leptospirillum ferriphilum Type Strain DSM 14647. *Genome Announcements*, 2(6):e01153–14–e01153–14, nov 2014. ISSN 2169-8287. doi: 10.1128/genomeA.01153-14.

Matteo Cavaliere, Song Feng, Orkun S. Soyer, and José I. Jiménez. Cooperation in microbial communities and their biotechnological applications. *Environmental Microbiology*, 19(8):2949–2963, 2017. ISSN 14622920. doi: 10.1111/1462-2920.13767.

Victoria Cepeda, Bo Liu, Mathieu Almeida, Christopher M. Hill, Sergey Koren, Todd J. Treangen, and Mihai Pop. MetaCompass: Reference-guided Assembly of Metagenomes. *bioRxiv*, 2017. doi: 10.1101/212506.

Chuming Chen, Zhiwen Li, Hongzhan Huang, Baris E. Suzek, and Cathy H. Wu. A fast peptide match service for UniProt knowledgebase. *Bioinformatics*, 29(21):2808–2809, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/btt484.

A. Chien, D. B. Edgar, and J. M. Trela. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *Journal of Bacteriology*, 127(3):1550–1557, 1976. ISSN 00219193.

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, jun 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2474.

Stephan Christel, Malte Herold, Sören Bellenberg, Mohamed El Hajjami, Antoine Buetti-Dinh, Igor V. Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, and Mark Dopson. Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species Leptospirillum ferriphilum T. *Applied and Environmental Microbiology*, 84(3):e02091–17, nov 2017. ISSN 0099-2240. doi: 10.1128/AEM.02091-17.

Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 44(D1):D67–D72, jan 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1276.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1): 13, dec 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8.

N. J. Coram and Douglas E. Rawlings. Molecular Relationship between Two Groups of the Genus Leptospirillum and the Finding that Leptospirillum ferriphilum sp. nov. Dominates South African Commercial Biooxidation Tanks That Operate at 40 C. *Applied and Environmental Microbiology*, 68(2):838–845, feb 2002. ISSN 0099-2240. doi: 10.1128/AEM.68.2.838-845.2002.

Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10(4):1794–1805, 2011. ISSN 15353893. doi: 10.1021/pr101065j.

Jürgen Cox, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9): 2513–2526, 2014. ISSN 1535-9476. doi: 10.1074/mcp.M113.031591.

Domenico Cozzetto, Daniel W a Buchan, Kevin Bryson, and David T Jones. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC bioinformatics*, 14 Suppl 3(Suppl 3):S1, jan 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S3-S1.

Gregory R. Crocetti, Philip L. Bond, Jillian F. Banfield, Linda L. Blackall, and Jürg Keller. Glycogen-accumulating organisms in laboratory-scale and full-scale wastewater treatment processes b. *Microbiology*, 148(11):3353–3364, nov 2002. ISSN 1350-0872. doi: 10.1099/00221287-148-11-3353.

T. P. Curtis, W. T. Sloan, and J. W. Scannell. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, 99(16):10494–10499, aug 2002. ISSN 0027-8424. doi: 10.1073/pnas.142680199.

Agnieszka Cydzik-Kwiatkowska and Magdalena Zielińska. Bacterial communities in full-scale wastewater treatment systems. *World Journal of Microbiology and Biotechnology*, 32(4):66, apr 2016. ISSN 0959-3993. doi: 10.1007/s11274-016-2012-9.

Christiane Dahl, Sabine Engels, A. S. Pott-Sperling, Andrea Schulte, Johannes Sander, Y. Lubbe, Oliver Deuster, and Daniel C Brune. Novel Genes of the dsr Gene Cluster and Evidence for Close Interaction of Dsr Proteins during Sulfur Oxidation in the Phototrophic Sulfur Bacterium Allochromatium vinosum. *Journal of Bacteriology*, 187(4):1392–1404, feb 2005. ISSN 0021-9193. doi: 10.1128/JB.187.4.1392-1404.2005.

Holger Daims, Michael W. Taylor, and Michael Wagner. Wastewater treatment: a model system for microbial ecology. *Trends in Biotechnology*, 24(11):483–489, nov 2006. ISSN 01677799. doi: 10.1016/j.tibtech.2006.09.002.

Holger Daims, Sebastian Lücker, and Michael Wagner. A New Perspective on Microbes Formerly Known as Nitrite-Oxidizing Bacteria. *Trends in Microbiology*, 24(9):699–712, 2016. ISSN 18784380. doi: 10.1016/j.tim.2016.05.004.

Nikos Darzentas. Circoletto: Visualizing sequence similarity with Circos. *Bioinformatics*, 26(20): 2620–2621, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq484.

Mark Davids, Floor Hugenholtz, Vitor Martins dos Santos, Hauke Smidt, Michiel Kleerebezem, and Peter J. Schaap. Functional Profiling of Unfamiliar Microbial Communities Using a Validated De Novo Assembly Metatranscriptome Pipeline. *PLOS ONE*, 11(1):e0146423, jan 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0146423.

Carlotta De Filippo, Matteo Ramazzotti, Paolo Fontana, and Duccio Cavalieri. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics*, 13(6):696–710, nov 2012. ISSN 1467-5463. doi: 10.1093/bib/bbs070.

Rutger De Wit and Thierry Bouvier. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, 8(4):755–758, 2006. ISSN 14622912. doi: 10.1111/j.1462-2920.2006.01017.x.

Tom O. Delmont and A. Murat Eren. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, 6:e4320, jan 2018. ISSN 2167-8359. doi: 10.7717/peerj.4320.

Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. Author Correction: Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(8):963–963, aug 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0209-4.

Vincent J. Denef, Ryan S. Mueller, and Jillian F. Banfield. AMD biofilms: Using model communities to study microbial evolution and ecological complexity in nature. *ISME Journal*, 4(5): 599–610, 2010. ISSN 17517362. doi: 10.1038/ismej.2009.158.

Y. Deng, N. Schmid, C. Wang, J. Wang, G. Pessi, D. Wu, J. Lee, C. Aguilar, C. H. Ahrens, C. Chang, H. Song, L. Eberl, and L.-H. Zhang. Cis-2-dodecenoic acid receptor RpfR links quorum-sensing signal perception with regulation of virulence through cyclic dimeric guanosine monophosphate turnover. *Proceedings of the National Academy of Sciences*, 109(38):15479–15484, sep 2012. ISSN 0027-8424. doi: 10.1073/pnas.1205037109.

Mark Dopson. Analysis of differential protein expression during growth states of Ferroplasma strains and insights into electron transport for iron oxidation. *Microbiology*, 151(12):4127–4137, dec 2005. ISSN 1350-0872. doi: 10.1099/mic.0.28362-0.

Mark Dopson and David S. Holmes. Metal resistance in acidophilic microorganisms and its significance for biotechnologies. *Applied Microbiology and Biotechnology*, pages 8133–8144, 2014. ISSN 01757598. doi: 10.1007/s00253-014-5982-2.

Mark Dopson and E Börje Lindström. Potential Role of Thiobacillus caldus in Arsenopyrite Bioleaching. *Applied and Environmental Microbiology*, 65(1):36–40, 1999. ISSN 0099-2240.

Connor B. Driscoll, Timothy G. Otten, Nathan M. Brown, and Theo W. Dreher. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in Genomic Sciences*, 12(1):9, dec 2017. ISSN 1944-3277. doi: 10.1186/s40793-017-0224-8.

Sean R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10), 2011. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002195.

A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, oct 2015. ISSN 2167-8359. doi: 10.7717/peerj.1319.

Özge Eyice, Motonobu Namura, Yin Chen, Andrew Mead, Siva Samavedam, and Hendrik Schäfer. SIP metagenomics identifies uncultivated Methylophilaceae as dimethylsulphide degrading bacteria in soil and lake sediment. *The ISME Journal*, 9(11):2336–2348, nov 2015. ISSN 1751-7362. doi: 10.1038/ismej.2015.37.

Fred Farrell, Orkun S Soyer, and Christopher Quince. Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. pages 1–22, 2018.

Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012. ISSN 1740-1526. doi: 10.1038/nrmicro2832.

Karoline Faust, Didier Croes, and Jacques van Helden. Prediction of metabolic pathways from genome-scale metabolic networks. *BioSystems*, 105(2):109–121, 2011. ISSN 03032647. doi: 10.1016/j.biosystems.2011.05.004.

Karoline Faust, Franziska Bauchinger, Béatrice Laroche, Sophie de Buyl, Leo Lahti, Alex D. Washburne, Didier Gonze, and Stefanie Widder. Signatures of ecological processes in microbial community time series. *Microbiome*, 6(1):1–13, 2018. ISSN 20492618. doi: 10.1186/s40168-018-0496-2.

Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7 (FEBRuARy):129–143, 2009. ISSN 1740-1526. doi: 10.1038/nrmicro1949.

Alexey Fomenkov, Tamas Vincze, Sergey K. Degtyarev, and Richard J. Roberts. Complete Genome Sequence and Methylome Analysis of Acinetobacter calcoaceticus 65. *Genome Announcements*, 5(12):1–2, mar 2017. ISSN 2169-8287. doi: 10.1128/genomeA.00060-17.

Marco Fondi and Pietro Liò. Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. *Microbiological Research*, 171:52–64, 2015. ISSN 09445013. doi: 10.1016/j.micres.2015.01.003.

Vittorio Fortino, Olli-Pekka Smolander, Petri Auvinen, Roberto Tagliaferri, and Dario Greco. Transcriptome dynamics-based operon prediction in prokaryotes. *BMC bioinformatics*, 15(1):145, 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-145.

Kevin R. Foster, Jonas Schluter, Katharine Z. Coyte, and Seth Rakoff-Nahoum. The evolution of the host microbiome as an ecosystem on a leash. *Nature*, 548(7665):43–51, 2017. ISSN 0028-0836. doi: 10.1038/nature23292.

Esther M Gabor, Benedikt Hoffmann, Thomas Deichmann, and Brain Ag. BRAIN BioXtractor: Biobased Metal Extraction for the Circular Economy. pages 68–73, 2018. doi: 10.1089/ind.2018.29122.emg.

Jason Gans, Murray Wolinsky, and John Dunbar. Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. *Science*, 309(5739):1387–1390, aug 2005. ISSN 0036-8075. doi: 10.1126/science.1112665.

Wolfgang Gerlach and Jens Stoye. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids research*, 39(14):e91, aug 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr225.

Scott M Gifford, Shalabh Sharma, Melissa Booth, and Mary Ann Moran. Expression patterns reveal niche diversification in a marine microbial assemblage. *The ISME Journal*, 7(2):281–298, 2013. ISSN 1751-7370. doi: 10.1038/ismej.2012.96.

Ludovic C. Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics*, 11(6):O111.016717, 2012. ISSN 1535-9476. doi: 10.1074/mcp.O111.016717.

R S Golovacheva and G I Karavaiko. [Sulfobacillus, a new genus of thermophilic sporulating bacteria]. *Mikrobiologiia*, 47(5):815–822, 1978. ISSN 0026-3656 (Print).

Daniela S Aliaga Goltsman, Luis R Comolli, Brian C Thomas, and Jillian F Banfield. Community transcriptomics reveals unexpected high microbial diversity in acidophilic biofilm communities. *The ISME Journal*, 9(4):1014–1023, apr 2015. ISSN 1751-7362. doi: 10.1038/ismej.2014.200.

Willi Gottstein, Brett G. Olivier, Frank J. Bruggeman, and Bas Teusink. Constraint-based stoichiometric modelling from single organisms to microbial communities. *Journal of the Royal Society Interface*, 13(124), 2016. ISSN 17425662. doi: 10.1098/rsif.2016.0627.

Kacy Greenhalgh, Kristen M. Meyer, Kjersti M. Aagaard, and Paul Wilmes. The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental Microbiology*, 18(7):2103–2116, jul 2016. ISSN 14622912. doi: 10.1111/1462-2920.13318.

Tobias Grosskopf and Orkun S Soyer. Synthetic microbial communities. *Current opinion in microbiology*, 18:72–7, apr 2014. ISSN 1879-0364. doi: 10.1016/j.mib.2014.02.002.

Xue Guo, Huaqun Yin, Yili Liang, Qi Hu, Xishu Zhou, Yunhua Xiao, Liyuan Ma, Xian Zhang, Guanzhou Qiu, and Xueduan Liu. Comparative genome analysis reveals metabolic versatility and environmental adaptations of Sulfobacillus thermosulfidooxidans strain ST. *PLoS ONE*, 9 (6), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0099417.

Shaan Gupta, Emma Allen-Vercoe, and Elaine O. Petrof. Fecal microbiota transplantation: in perspective. *Therapeutic Advances in Gastroenterology*, 9(2):229–239, mar 2016. ISSN 1756-283X. doi: 10.1177/1756283X15607414.

Johanna Gutleben, Maryam Chaib De Mares, Jan Dirk van Elsas, Hauke Smidt, Jörg Overmann, and Detmer Sipkema. The multi-omics promise in context: from sequence to microbial isolate. *Critical Reviews in Microbiology*, 44(2):212–229, mar 2018. ISSN 1040-841X. doi: 10.1080/1040841X.2017.1332003.

Ed K. Hall, Emily S. Bernhardt, Raven L. Bier, Mark A. Bradford, Claudia M. Boot, James B. Cotner, Paul A. del Giorgio, Sarah E. Evans, Emily B. Graham, Stuart E. Jones, Jay T. Lennon, Kenneth J. Locey, Diana Nemergut, Brooke B. Osborne, Jennifer D. Rocca, Joshua P. Schimel, Mark P. Waldrop, and Matthew D. Wallenstein. Understanding how microbiomes influence the systems they inhabit. *Nature Microbiology*, 3(9):977–982, sep 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0201-z.

K. B. Hallberg and E. B. Lindstrom. Characterization of Thiobacillus caldus sp. nov., a moderately thermophilic acidophile. *Microbiology*, 140(12):3451–3456, dec 1994. ISSN 1350-0872. doi: 10.1099/13500872-140-12-3451.

Alexander Harms, Ditlev Egeskov Brodersen, Namiko Mitarai, and Kenn Gerdes. Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Molecular Cell*, 70(5):768–784, 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.01.003.

Sabrina Hedrich, Anne Gwenaëlle Guézennec, Mickaël Charron, Axel Schippers, and Catherine Joulian. Quantitative monitoring of microbial species during bioleaching of a copper concentrate. *Frontiers in Microbiology*, 7(DEC):1–11, 2016. ISSN 1664302X. doi: 10.3389/fmicb.2016.02044.

Anna Heintz-Buschart, Patrick May, Cédric C. Laczny, Laura A. Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, Jochen G. Schneider, Angela Hogan, Carine de Beaufort, and Paul Wilmes. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, 2(1):16180, jan 2017. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.180.

Regine Hengge. Principles of c-di-GMP signalling in bacteria. *Nature Reviews Microbiology*, 7 (4):263–273, 2009. ISSN 17401526. doi: 10.1038/nrmicro2109.

Christopher S. Henry, Hans C. Bernstein, Pamela Weisenhorn, Ronald C. Taylor, Joon Yong Lee, Jeremy Zucker, and Hyun Seob Song. Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction. *Journal of Cellular Physiology*, 231(11):2339–2345, 2016. ISSN 10974652. doi: 10.1002/jcp.25428.

Robert L. Hettich, Chongle Pan, Karuna Chourey, and Richard J. Giannone. Metaproteomics : Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities. *Analytical Chemistry*, 85(9): 4203–4214, may 2013. ISSN 0003-2700. doi: 10.1021/ac303053e.

Robert Heyer, Kay Schallert, Roman Zoun, Beatrice Becher, Gunter Saake, and Dirk Benndorf. Challenges and perspectives of metaproteomic data analysis. *Journal of Biotechnology*, 261: 24–36, nov 2017. ISSN 01681656. doi: 10.1016/j.jbiotec.2017.06.1201.

Karsten Hiller, Jasper Hangebrauk, Christian Jäger, Jana Spura, Kerstin Schreiber, and Dietmar Schomburg. Metabolite detector: Comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Analytical Chemistry*, 81(9):3429–3439, 2009. ISSN 00032700. doi: 10.1021/ac802689c.

Brian M. Hoffman, Dmitriy Lukoyanov, Zhi Yong Yang, Dennis R. Dean, and Lance C. Seefeldt. Mechanism of nitrogen fixation by nitrogenase: The next stage. *Chemical Reviews*, 114(8): 4041–4062, 2014. ISSN 15206890. doi: 10.1021/cr400641x.

Bradley M. Hover, Seong Hwan Kim, Micah Katz, Zachary Charlop-Powers, Jeremy G. Owen, Melinda A. Ternei, Jeffrey Maniko, Andreia B. Estrela, Henrik Molina, Steven Park, David S. Perlin, and Sean F. Brady. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature Microbiology*, 3(4):415–422, 2018. ISSN 20585276. doi: 10.1038/s41564-018-0110-1.

Stephen P. Hubbell. Neutral theory and the evolution of ecological equivalence. *Ecology*, 87(6): 1387–1398, 2006. ISSN 00129658. doi: 10.1890/0012-9658(2006)87[1387:NTATEO]2.0.CO; 2.

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Ise Kotaro, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. A new view of the tree and life's diversity. (April):Manuscript submitted for publication, 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.48.

Luisa W. Hugerth and Anders F. Andersson. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology*, 8(SEP): 1–22, 2017. ISSN 1664302X. doi: 10.3389/fmicb.2017.01561.

Michael Hügler and Stefan M. Sievert. Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Annual Review of Marine Science*, 3(1):261–289, 2011. ISSN 1941-1405. doi: 10.1146/annurev-marine-120709-142712.

Jenni Hultman, Mark P. Waldrop, Rachel Mackelprang, Maude M. David, Jack McFarland, Steven J. Blazewicz, Jennifer Harden, Merritt R. Turetsky, A. David McGuire, Manesh B. Shah, Nathan C. VerBerkmoes, Lang Ho Lee, Kostas Mavrommatis, and Janet K. Jansson. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, 521(7551):208–212, 2015. ISSN 14764687. doi: 10.1038/nature14238.

Martin Hunt, Nishadi De Silva, Thomas D. Otto, Julian Parkhill, Jacqueline A. Keane, and Simon R. Harris. Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biology*, 16(1):1–10, 2015. ISSN 1474760X. doi: 10.1186/s13059-015-0849-0.

Susan M. Huse, Yuzhen Ye, Yanjiao Zhou, and Anthony A. Fodor. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS ONE*, 7(6):1–12, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0034242.

G. E. Hutchinson. Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22(0):415–427, 1957. ISSN 0091-7451. doi: 10.1101/SQB.1957.022.01.039.

Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.

Michael Imelfort, Donovan Parks, Ben J Woodcroft, Paul Dennis, Philip Hugenholtz, and Gene W Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, sep 2014. ISSN 2167-8359. doi: 10.7717/peerj.603.

Shun'ichi Ishii, Shino Suzuki, Trina M Norden-Krichmar, Aaron Tenney, Patrick S G Chain, Matthew B Scholz, Kenneth H Nealson, and Orianna Bretschger. A novel metatranscriptomic approach to identify gene expression dynamics during extracellular electron transfer. *Nature communications*, 4:1601, 2013. ISSN 2041-1723. doi: 10.1038/ncomms2615.

Francisco Issotta, Pedro A. Galleguillos, Ana Moya-Beltrán, Carol S. Davis-Belmar, George Rautenbach, Paulo C. Covarrubias, Mauricio Acosta, Francisco J. Ossandon, Yasna Contador, David S. Holmes, Sabrina Marín-Eliantonio, Raquel Quatrini, and Cecilia Demergasso. Draft genome sequence of chloride-tolerant Leptospirillum ferriphilum Sp-Cl from industrial bioleaching operations in northern Chile. *Standards in Genomic Sciences*, 11(1):1–7, 2016. ISSN 19443277. doi: 10.1186/s40793-016-0142-1.

Chris Jeans, Steven W Singer, Clara S Chan, Nathan C Verberkmoes, Manesh Shah, Robert L Hettich, Jillian F Banfield, and Michael P Thelen. Cytochrome 572 is a conspicuous membrane protein with iron oxidation activity purified directly from a natural acidophilic microbial community. *The ISME journal*, 2(5):542–50, may 2008. ISSN 1751-7362. doi: 10.1038/ismej.2008.17.

C.A. Jerez. Bioleaching and Biomining for the Industrial Recovery of Metals. In *Comprehensive Biotechnology*, volume 3, pages 717–729. Elsevier, second edi edition, 2011. ISBN 9780080885049. doi: 10.1016/B978-0-08-088504-9.00234-8.

Huidan Jiang, Yili Liang, Huaqun Yin, Yunhua Xiao, Xue Guo, Ying Xu, Qi Hu, Hongwei Liu, and Xueduan Liu. Effects of Arsenite Resistance on the Growth and Functional Gene Expression of <i>Leptospirillum ferriphilum</i> and <i>Acidithiobacillus thiooxidans</i> in Pure Culture and Coculture. *BioMed Research International*, 2015:1–13, 2015. ISSN 2314-6133. doi: 10.1155/2015/203197.

D. Barrie Johnson and Chris a. du Plessis. Biomining in reverse gear: Using bacteria to extract metals from oxidised ores. *Minerals Engineering*, 75:2–5, 2014. ISSN 08926875. doi: 10.1016/j.mineng.2014.09.024.

P. Jones. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Research*, 34(90001):D659–D663, jan 2006. ISSN 0305-1048. doi: 10.1093/nar/gkj138.

Young Soo Joung, Zhifei Ge, and Cullen R. Buie. Bioaerosol generation by raindrops on soil. *Nature Communications*, 8:1–10, 2017. ISSN 20411723. doi: 10.1038/ncomms14668.

Susanne Jünemann. Cytochrome bd terminal oxidase1All amino acid numbering refers to the E. coli enzyme.1. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1321(2):107–127, aug 1997. ISSN 00052728. doi: 10.1016/S0005-2728(97)00046-7.

Nicholas B. Justice, Anders Norman, Christopher T. Brown, Andrea Singh, Brian C. Thomas, and Jillian F. Banfield. Comparison of environmental and isolate Sulfobacillus genomes reveals diverse carbon, sulfur, nitrogen, and hydrogen metabolisms. *BMC Genomics*, 15(1):1–17, 2014. ISSN 14712164. doi: 10.1186/1471-2164-15-1107.

Alex Kalderimis, Rachel Lyne, Daniela Butano, Sergio Contrino, Mike Lyne, Joshua Heimbach, Fengyuan Hu, Richard Smith, Radek Štěpán, Julie Sullivan, and Gos Micklem. InterMine: Extensive web services for modern biology. *Nucleic Acids Research*, 42(W1):468–472, 2014. ISSN 13624962. doi: 10.1093/nar/gku301.

Soeren M Karst, Rasmus H Kirkegaard, and Mads Albertsen. mmgenome: a toolbox for reproducible genome extraction from metagenomes. *bioRxiv*, 2016. doi: 10.1101/059121.

Rees Kassen. The experimental evolution of specialists, generalists, and the maintenance of diversity. *Journal of Evolutionary Biology*, 15(2):173–190, mar 2002. ISSN 1010061X. doi: 10.1046/j.1420-9101.2002.00377.x.

Anne Kaysen, Anna Heintz-Buschart, Emilie E L Muller, Shaman Narayanasamy, Linda Wampach, Cédric C Laczny, Norbert Graf, Arne Simon, Katharina Franke, Jörg Bittenbring, Paul Wilmes, and Jochen G Schneider. Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic hematopoietic stem cell transplantation. *Translational research : the journal of laboratory and clinical medicine*, 186:79–94.e1, aug 2017. ISSN 1878-1810. doi: 10.1016/j.trsl.2017.06.008.

Martin Keller and Karsten Zengler. Tapping into microbial diversity. *Nature Reviews Microbiology*, 2(2):141–150, feb 2004. ISSN 1740-1526. doi: 10.1038/nrmicro819.

Ruchir A. Khandelwal, Brett G. Olivier, Wilfred F M Röling, Bas Teusink, and Frank J. Bruggeman. Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE*, 8(5), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0064567.

Mohammad Khoshkhoo, Mark Dopson, Andrey Shchukarev, and Åke Sandström. Chalcopyrite leaching and bioleaching: An X-ray photoelectron spectroscopic (XPS) investigation on the nature of hindered dissolution. *Hydrometallurgy*, 149:220–227, 2014. ISSN 0304386X. doi: 10.1016/j.hydromet.2014.08.012.

Sangtae Kim and Pavel A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5(1):5277, dec 2014. ISSN 2041-1723. doi: 10.1038/ncomms6277.

P. H M Kinnunen and Jaakko A. Puhakka. High-rate iron oxidation at below pH 1 and at elevated iron and copper concentrations by a Leptospirillum ferriphilum dominated biofilm. *Process Biochemistry*, 40(11):3536–3541, 2005. ISSN 13595113. doi: 10.1016/j.procbio.2005.03.050.

Heiner Klingenberg and Peter Meinicke. How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ*, 5:e3859, 2017. ISSN 2167-8359. doi: 10.7717/peerj.3859.

Sabine Koch, Dirk Benndorf, Karen Fronk, Udo Reichl, and Steffen Klamt. Predicting compositions of microbial communities from stoichiometric models with applications for the biogas process. *Biotechnology for Biofuels*, 9(1):17, dec 2016. ISSN 1754-6834. doi: 10.1186/s13068-016-0429-x.

Allan Konopka. What is microbial community ecology. *ISME Journal*, 3(11):1223–1230, 2009. ISSN 17517362. doi: 10.1038/ismej.2009.88.

P. Koskinen, P. Toronen, J. Nokso-Koivisto, and L. Holm. PANNZER - High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, jan 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu851.

J. Koster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, oct 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480.

Cedric C Laczny, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Margossian, Sergio Coronado, Laurens der Maaten, Nikos Vlassis, and Paul Wilmes. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3(1):1, 2015. ISSN 2049-2618. doi: 10.1186/s40168-014-0066-1.

Cedric C Laczny, Emilie E. L. Muller, Anna Heintz-Buschart, Malte Herold, Laura A Lebrun, Angela Hogan, Patrick May, Carine de Beaufort, and Paul Wilmes. Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract. *Frontiers in Microbiology*, 7, jun 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00884.

V. Lakshmanan, G. Selvaraj, and H. P. Bais. Functional Soil Microbiome: Belowground Solutions to an Aboveground Problem. *Plant Physiology*, 166(2):689–700, 2014. ISSN 0032-0889. doi: 10.1104/pp.114.245811.

Morgan G.I. Langille, Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, Deron E. Burkepile, Rebecca L. Vega Thurber, Rob Knight, Robert G. Beiko, and Curtis Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9):814–821, 2013. ISSN 10870156. doi: 10.1038/nbt.2676.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, apr 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923.

Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database):D19–D21, jan 2011. ISSN 0305-1048. doi: 10.1093/nar/gkq1019.

Joseph A. Lemire, Joe J. Harrison, and Raymond J. Turner. Antimicrobial activity of metals: Mechanisms, molecular targets and applications. *Nature Reviews Microbiology*, 11(6):371–384, 2013. ISSN 17401526. doi: 10.1038/nrmicro3028.

Mark A. Lever, Andrea Torti, Philip Eickenbusch, Alexander B. Michaud, Tina Santl-Temkiv, and Bo Barker Jorgensen. A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Frontiers in Microbiology*, 6(MAY), 2015. ISSN 1664302X. doi: 10.3389/fmicb.2015.00476.

Kim Lewis. Antibiotics: Recover the lost art of drug discovery. *Nature*, 485(7399):439–440, 2012. ISSN 00280836. doi: 10.1038/485439a.

Dinghua Li, Chi-man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, may 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv033.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324.

Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btt656.

Yu-Chieh Liao, Shu-Hung Lin, and Hsin-Hung Lin. Completing bacterial genome assemblies: strategy and performance comparisons. *Scientific Reports*, 5(1):8747, aug 2015. ISSN 2045-2322. doi: 10.1038/srep08747.

Losee L Ling, Tanja Schneider, Aaron J Peoples, Amy L Spoering, Ina Engels, Brian P Conlon, Anna Mueller, Dallas E Hughes, Slava Epstein, Michael Jones, Linos Lazarides, Victoria a Steadman, Douglas R Cohen, Cintia R Felix, K Ashley Fetterman, William P Millett, Anthony G Nitti, Ashley M Zullo, Chao Chen, and Kim Lewis. A new antibiotic kills pathogens without detectable resistance. *Nature*, 517(7535):455–459, 2015. ISSN 1476-4687. doi: 10.1038/nature14098.

Kenneth J. Locey and Jay T. Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1521291113.

Stilianos Louca, Martin F Polz, Florent Mazel, Michaeline B N Albright, Julie A Huber, Mary I. O'Connor, Martin Ackermann, Aria S Hahn, Diane S Srivastava, Sean A Crowe, Michael Doebeli, and Laura Wegener Parfrey. Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, 2(6):936–943, jun 2018. ISSN 2397-334X. doi: 10.1038/s41559-018-0519-1.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.

M. E. Mackintosh. Nitrogen Fixation by Thiobacillus ferrooxidans. *Journal of General Microbiology*, 105(2):215–218, apr 1978. ISSN 0022-1287. doi: 10.1099/00221287-105-2-215.

Eugene L. Madsen. Microorganisms and their roles in fundamental biogeochemical cycles. *Current Opinion in Biotechnology*, 22(3):456–464, jun 2011. ISSN 09581669. doi: 10.1016/j.copbio.2011.01.008.

S. Magnúsdóttir, A. Heinken, L. Kutt, D.A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, R.M.T. Fleming, and I. Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, (November), 2016. ISSN 1087-0156. doi: 10.1038/nbt.3703.

Thulani P. Makhalanyane, Angel Valverde, Eoin Gunnigle, Aline Frossard, Jean Baptiste Ramond, and Don A. Cowan. Microbial ecology of hot desert edaphic systems. *FEMS Microbiology Reviews*, 39(2):203–221, 2015. ISSN 15746976. doi: 10.1093/femsre/fuu011.

Himel Mallick, Siyuan Ma, Eric A. Franzosa, Tommi Vatanen, Xochitl C. Morgan, and Curtis Huttenhower. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*, 18(1):228, dec 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1359-z.

Stefanie Mangold, Jorge Valdés, David S. Holmes, and Mark Dopson. Sulfur metabolism in the extreme acidophile Acidithiobacillus caldus. *Frontiers in Microbiology*, 2(FEB):1–18, 2011. ISSN 1664302X. doi: 10.3389/fmicb.2011.00017.

Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302, mar 2018. ISSN 1467-5463. doi: 10.1093/bib/bbw114.

Victor M Markowitz, I-Min a Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Ioanna Pagani, Susannah Tringe, Marcel Huntemann, Konstantinos Billis, Neha Varghese, Kristin Tennessen, Konstantinos Mavromatis, Amrita Pati, Natalia N Ivanova, and Nikos C Kyrpides. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, 42(D1):D568–D573, jan 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt919.

Héctor García Martín, Natalia Ivanova, Victor Kunin, Falk Warnecke, Kerrie W Barry, Alice C McHardy, Christine Yeates, Shaomei He, Asaf A Salamov, Ernest Szeto, Eileen Dalin, Nik H Putnam, Harris J Shapiro, Jasmyn L Pangilinan, Isidore Rigoutsos, Nikos C Kyrpides, Linda Louise Blackall, Katherine D McMahon, and Philip Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*, 24(10):1263–1269, 2006. ISSN 1087-0156. doi: 10.1038/nbt1247.

Patricio Martinez, Mario Vera, and Roberto A. Bobadilla-Fazzini. Omics on bioleaching: current and future impacts. *Applied Microbiology and Biotechnology*, 99(20):8337–8350, 2015. ISSN 14320614. doi: 10.1007/s00253-015-6903-8.

Irena Maus, Daniela E. Koeck, Katharina G. Cibis, Sarah Hahnke, Yong S. Kim, Thomas Langer, Jana Kreubel, Marcel Erhard, Andreas Bremges, Sandra Off, Yvonne Stolze, Sebastian Jaenicke, Alexander Goesmann, Alexander Sczyrba, Paul Scherer, Helmut König, Wolfgang H. Schwarz, Vladimir V. Zverlov, Wolfgang Liebl, Alfred Pühler, Andreas Schlüter, and Michael Klocke. Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates. *Biotechnology for Biofuels*, 9(1):171, dec 2016. ISSN 1754-6834. doi: 10.1186/s13068-016-0581-3.

P. E. McGovern, J. Zhang, J. Tang, Z. Zhang, G. R. Hall, R. A. Moreau, A. Nunez, E. D. Butrym, M. P. Richards, C.-s. Wang, G. Cheng, Z. Zhao, and C. Wang. Fermented beverages of pre- and proto-historic China. *Proceedings of the National Academy of Sciences*, 101(51):17593–17598, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0407921102.

Simon Jon McIlroy, Rikke Kristiansen, Mads Albertsen, Søren Michael Karst, Simona Rossetti, Jeppe Lund Nielsen, Valter Tandoi, Robert James Seviour, and Per Halkjær Nielsen. Metabolic model for the filamentous 'Candidatus Microthrix parvicella' based on genomic and metagenomic analyses. *The ISME Journal*, 7(6):1161–1172, jun 2013. ISSN 1751-7362. doi: 10.1038/ismej.2013.6.

Simon Jon McIlroy, Aaron Marc Saunders, Mads Albertsen, Marta Nierychlo, Bianca McIlroy, Aviaja Anna Hansen, Søren Michael Karst, Jeppe Lund Nielsen, and Per Halkjær Nielsen. MiDAS: the field guide to the microbes of activated sludge. *Database*, 2015(September):bav062, jun 2015. ISSN 1758-0463. doi: 10.1093/database/bav062.

M. P. Merino, B. a. Andrews, and J. a. Asenjo. Stoichiometric model and metabolic flux analysis for Leptospirillum ferrooxidans. *Biotechnology and Bioengineering*, 107(4):696–706, 2010. ISSN 00063592. doi: 10.1002/bit.22851.

M. P. Merino, B. a. Andrews, and J. a. Asenjo. Stoichiometric model and flux balance analysis for a mixed culture of Leptospirillum ferriphilum and Ferroplasma acidiphilum. *Biotechnology Progress*, pages n/a–n/a, 2014. ISSN 87567938. doi: 10.1002/btpr.2028.

Michael L. Metzker. Sequencing technologies the next generation. *Nature Reviews Genetics*, 11 (1):31–46, 2010. ISSN 14710056. doi: 10.1038/nrg2626.

Shuang Mi, Jian Song, Jianqun Lin, Yuanyuan Che, Huajun Zheng, and Jianqiang Lin. Complete genome of Leptospirillum ferriphilum ML-04 provides insight into its physiology and environmental adaptation. *Journal of Microbiology*, 49(6):890–901, 2011. ISSN 12258873. doi: 10.1007/s12275-011-1099-9.

Sara Mitri and Kevin Richard Foster. The Genotypic View of Social Interactions in Microbial Communities. *Annual Review of Genetics*, 47(1):247–273, 2013. ISSN 0066-4197. doi: 10.1146/annurev-genet-111212-133307.

Mercedes Moreno-Paz and Víctor Parro. Amplification of low quantity bacterial RNA for microarray studies: Time-course analysis of Leptospirillum ferrooxidans under nitrogen-fixing conditions. *Environmental Microbiology*, 8(6):1064–1073, 2006. ISSN 14622912. doi: 10.1111/j.1462-2920.2006.00998.x.

Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E Sands, Ramnik J Xavier, and Curtis Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012. ISSN 1465-6906. doi: 10.1186/gb-2012-13-9-r79.

Steffen Moritz and Thomas Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218, 2017. ISSN 2073-4859.

Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Olena Verezemska, Michelle Isbandi, Alex D. Thomas, Rida Ali, Kaushal Sharma, Nikos C. Kyrpides, and T. B. K. Reddy. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Research*, 45(D1):D446–D456, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw992.

A. Mulder. Anaerobic ammonium oxidation discovered in a denitrifying fluidized bed reactor. *FEMS Microbiology Ecology*, 16(3):177–183, mar 1995. ISSN 01686496. doi: 10.1016/0168-6496(94)00081-7.

Emilie E L Muller, Enrico Glaab, Patrick May, Nikos Vlassis, and Paul Wilmes. Condensing the omics fog of microbial communities. *Trends in microbiology*, 21(7):325–33, jul 2013. ISSN 1878-4380. doi: 10.1016/j.tim.2013.04.009.

Emilie E L Muller, Nicolás Pinel, Cédric C Laczny, Michael R Hoopmann, Shaman Narayanasamy, Laura a Lebrun, Hugo Roume, Jake Lin, Patrick May, Nathan D Hicks, Anna Heintz-Buschart, Linda Wampach, Cindy M Liu, Lance B Price, John D Gillece, Cédric Guignard, James M Schupp, Nikos Vlassis, Nitin S Baliga, Robert L Moritz, Paul S Keim, and Paul Wilmes. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications*, 5(1):5603, dec 2014a. ISSN 2041-1723. doi: 10.1038/ncomms6603.

Emilie E. L. Muller, Shaman Narayanasamy, Myriam Zeimes, Cédric C. Laczny, Laura A. Lebrun, Malte Herold, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul Keim, and Paul Wilmes. First draft genome sequence of a strain belonging to the Zoogloea genus and its gene expression in situ. *Standards in Genomic Sciences*, 12(1):64, dec 2017. ISSN 1944-3277. doi: 10.1186/s40793-017-0274-y.

Emilie El Muller, Abdul R Sheik, and Paul Wilmes. Lipid-based biofuel production from wastewater. *Current Opinion in Biotechnology*, 30:9–16, dec 2014b. ISSN 09581669. doi: 10.1016/j.copbio.2014.03.007.

Emilie E.L. Muller, Karoline Faust, Stefanie Widder, Malte Herold, Susana Martínez Arbas, and Paul Wilmes. Using metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in Systems Biology*, 8:73–80, apr 2018. ISSN 24523100. doi: 10.1016/j.coisb.2017.12.004.

E. V. Musvoto, M. T. Lakay, T. G. Casey, M. C. Wentzel, and G. A. Ekama. Filamentous organism bulking in nutrient removal activated sludge systems. Paper 8: The effect of nitrate and nitrite. *Water SA*, 25(4):397–407, 1999. ISSN 03784738.

Shaman Narayanasamy. *Development of an integrated omics in silico workflow and its application for studying bacteria-phage interactions in a model microbial community*. Phd thesis, University of Luxembourg, 2017.

Shaman Narayanasamy, Emilie E. L. Muller, Abdul R. Sheik, and Paul Wilmes. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microbial Biotechnology*, 8(3):363–368, may 2015. ISSN 17517915. doi: 10.1111/1751-7915.12255.

Shaman Narayanasamy, Yohan Jarosz, Emilie E L Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolás Pinel, Patrick May, and Paul Wilmes. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology*, 17(1):260, dec 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1116-8.

D. Nichols, N. Cahoon, E. M. Trakhtenberg, L. Pham, a. Mehta, a. Belanger, T. Kanigan, K. Lewis, and S. S. Epstein. Use of Ichip for High-Throughput In Situ Cultivation of "Uncultivable" Microbial Species. *Applied and Environmental Microbiology*, 76(8):2445–2450, 2010. ISSN 0099-2240. doi: 10.1128/AEM.01754-09.

Robert Niederdorfer, Hannes Peter, and Tom J. Battin. Attached biofilms and suspended aggregates are distinct microbial lifestyles emanating from differing hydraulics. *Nature Microbiology*, 1 (October), 2016. ISSN 20585276. doi: 10.1038/nmicrobiol.2016.178.

H Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R Plichta, Laurent Gautier, Anders G Pedersen, Emmanuelle Le Chatelier, Eric Pelletier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B Quintanilha Dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S Kaas, Sean Kennedy, Karsten Kristiansen, Jens Roat Kultima, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W Ussery, Takuji Yamada, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, and S Dusko Ehrlich. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, aug 2014. ISSN 1546-1696. doi: 10.1038/nbt.2939.

P.H. Nielsen, P. Roslev, T.E. Dueholm, and J.L. Nielsen. Microthrix parvicella, a specialized lipid consumer in anaerobic–aerobic activated sludge plants. *Water Science and Technology*, 46(1-2): 73–80, jul 2002. ISSN 0273-1223. doi: 10.2166/wst.2002.0459.

N. Noël, B. Florian, and W. Sand. AFM & EFM study on attachment of acidophilic leaching organisms. *Hydrometallurgy*, 104(3-4):370–375, 2010. ISSN 0304386X. doi: 10.1016/j.hydromet.2010.02.021.

Paul R Norris, Darren A Clark, Jonathan P Owen, and Sara Waterhouse. Characteristics of Sulfobacillus acidophilus sp. nov. and other moderately thermophilic mineral-sulphide-oxidizing bacteria. *Microbiology*, 142(4):775–783, apr 1996. ISSN 1350-0872. doi: 10.1099/00221287-142-4-775.

I. D. Ofiteru, Mary Lunn, Thomas P Curtis, George F Wells, Craig S Criddle, Christopher A Francis, and William T Sloan. Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences*, 107(35):15345–15350, aug 2010. ISSN 0027-8424. doi: 10.1073/pnas.1000604107.

Matthew R Olm, Christopher T Brown, Brandon Brooks, and Jillian F Banfield. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, 11(12):2864–2868, dec 2017. ISSN 1751-7362. doi: 10.1038/ismej.2017.126.

Brian D. Ondov, Todd J. Treangen, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Fast genome and metagenome distance estimation using MinHash. *bioRxiv*, page 029827, 2015. doi: 10.1101/029827.

Margaret A Palmer. Biodiversity and Ecosystem Processes in Freshwater Sediments. *Ambio*, 26 (8):571–577, 1997. ISSN 00447447 (ISSN).

Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, jul 2015. ISSN 1088-9051. doi: 10.1101/gr.186072.114.

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542, 2017. ISSN 20585276. doi: 10.1038/s41564-017-0012-7.

Anke Penzlin, Martin S. Lindner, Joerg Doellinger, Piotr Wojtek Dabrowski, Andreas Nitsche, and Bernhard Y. Renard. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics*, 30(12):i149–i156, jun 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu267.

Olga E. Petrova, Fernando Garcia-Alcalde, Claudia Zampaloni, and Karin Sauer. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Scientific Reports*, 7(1):41114, dec 2017. ISSN 2045-2322. doi: 10.1038/srep41114.

Robert S. Pitcher and Nicholas J. Watmough. The bacterial cytochrome cbb3 oxidases. *Biochimica et Biophysica Acta - Bioenergetics*, 1655(1-3):388–399, 2004. ISSN 00052728. doi: 10.1016/j.bbabio.2003.09.017.

Damian Rafal Plichta, Agnieszka Sierakowska Juncker, Marcelo Bertalan, Elizabeth Rettedal, Laurent Gautier, Encarna Varela, Chaysavanh Manichanh, Charlène Fouqueray, Florence Levenez, Trine Nielsen, Joël Doré, Ana Manuel Dantas Machado, Mari Cristina Rodriguez de Evgrafov, Torben Hansen, Torben Jørgensen, Peer Bork, Francisco Guarner, Oluf Pedersen, Morten O. A. Sommer, S. Dusko Ehrlich, Thomas Sicheritz-Pontén, Søren Brunak, and H. Bjørn Nielsen. Transcriptional interactions suggest niche segregation among microorganisms in the human gut. *Nature Microbiology*, 1(11):16152, 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.152.

P. B. Pope, W. Smith, S. E. Denman, S. G. Tringe, K. Barry, P. Hugenholtz, C. S. McSweeney, A. C. McHardy, and M. Morrison. Isolation of Succinivibrionaceae Implicated in Low Methane Emissions from Tammar Wallabies. *Science*, 333(6042):646–648, jul 2011. ISSN 0036-8075. doi: 10.1126/science.1205760.

N. Pradhan, K. C. Nathsarma, K. Srinivasa Rao, L. B. Sukla, and B. K. Mishra. Heap bioleaching of chalcopyrite: A review. *Minerals Engineering*, 21:355–365, 2008. ISSN 08926875. doi: 10.1016/j.mineng.2007.10.018.

James I. Prosser. Ecosystem processes and interactions in a morass of diversity. *FEMS Microbiology Ecology*, 81(3):507–519, 2012. ISSN 01686496. doi: 10.1111/j.1574-6941.2012.01435.x.

Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian Wu, Yali Dai, Xiaojuan Sun, Zesong Li, Aifa Tang, Shilong Zhong, Xiaoping Li, Weineng Chen, Ran Xu, Mingbang Wang, Qiang Feng, Meihua Gong, Jing Yu, Yanyan Zhang, Ming Zhang, Torben Hansen, Gaston Sanchez, Jeroen Raes, Gwen Falony, Shujiro Okuda, Mathieu Almeida, Emmanuelle LeChatelier, Pierre Renault, Nicolas Pons, Jean-Michel Batto, Zhaoxi Zhang, Hua Chen, Ruifu Yang, Weimou Zheng, Songgang Li, Huanming Yang, Jian Wang, S. Dusko Ehrlich, Rasmus Nielsen, Oluf Pedersen, Karsten Kristiansen, and Jun Wang. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, sep 2012. ISSN 0028-0836. doi: 10.1038/nature11450.

Rachna J. Ram, Nathan C. VerBerkmoes, Michael P. Thelen, Gene W. Tyson, Brett J. Baker, Robert C. Blake, Manesh Shah, Robert L. Hettich, and Jillian F. Banfield. Community proteomics of a Natural Microbial Biofilm. *Science*, 308(2005):1915—-1920, 2005. ISSN 0036-8075. doi: 10.1126/science.1109070.

A. Ramette and J. M. Tiedje. Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proceedings of the National Academy of Sciences*, 104(8):2761–2766, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0610671104.

D. E. Rawlings, H. Tributsch, and G. S. Hansford. Reasons why 'Leptospirillum'-like species rather than Thiobacillus ferrooxidans are the dominant iron-oxidizing bacteria in many commercial processes for the biooxidation of pyrite and related ores. *Microbiology*, 145(1):5–13, 1999. ISSN 13500872. doi: 10.1099/13500872-145-1-5.

F. Reen, José Gutiérrez-Barranquero, Alan Dobson, Claire Adams, and Fergal O'Gara. *Emerging Concepts Promising New Horizons for Marine Biodiscovery and Synthetic Biology*, volume 13. 2015. ISBN 3532142759. doi: 10.3390/md13052924.

Mina Rho, Haixu Tang, and Yuzhen Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191–e191, nov 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq747.

Emily J. Richardson and Mick Watson. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*, 14:1–12, 2013. ISSN 14675463. doi: 10.1093/bib/bbs007.

Marja Riekkola-Vanhanen. Talvivaara mining company – From a project to a mine. *Minerals Engineering*, 48:2–9, jul 2013. ISSN 08926875. doi: 10.1016/j.mineng.2013.04.018.

Margaret A. Riley and John E. Wertz. Bacteriocins: Evolution, Ecology, and Application. *Annual Review of Microbiology*, 56(1):117–137, 2002. ISSN 0066-4227. doi: 10.1146/annurev.micro.56.012302.161024.

Marina L. Ritchie and Tamara N. Romanuk. A meta-analysis of probiotic efficacy for gastrointestinal diseases. *PLoS ONE*, 7(4), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0034938.

L. Rizzo, C. Manaia, C. Merlin, T. Schwartz, C. Dagot, M.C. Ploy, I. Michael, and D. Fatta-Kassinos. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: A review. *Science of The Total Environment*, 447:345–360, mar 2013. ISSN 00489697. doi: 10.1016/j.scitotenv.2013.01.032.

Luis M. Rodriguez-R and Konstantinos T. Konstantinidis. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*, 30(5):629–635, mar 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt584.

Luiz F W Roesch, Roberta R Fulthorpe, Alberto Riva, George Casella, Alison K M Hadwin, Angela D Kent, Samira H Daroub, Flavio A O Camargo, William G Farmerie, and Eric W Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, 1(4):283–290, aug 2007. ISSN 1751-7362. doi: 10.1038/ismej.2007.53.

Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11): e1005752, nov 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005752.

T. Rohwerder, T. Gehrke, K. Kinzler, and W. Sand. Bioleaching review part A: Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation. *Applied Microbiology and Biotechnology*, 63(3):239–248, dec 2003. ISSN 0175-7598. doi: 10.1007/s00253-003-1448-7.

U. Romling, M. Y. Galperin, and M. Gomelsky. Cyclic di-GMP: the First 25 Years of a Universal Bacterial Second Messenger. *Microbiology and Molecular Biology Reviews*, 77(1):1–52, mar 2013. ISSN 1092-2172. doi: 10.1128/MMBR.00043-12.

Simona Rossetti, Maria C. Tomei, Per H. Nielsen, and Valter Tandoi. “ Microthrix parvicella ”, a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiology Reviews*, 29(1):49–64, jan 2005. ISSN 1574-6976. doi: 10.1016/j.femsre.2004.09.005.

Hugo Roume. Molecular Eco-Systems Biology of Lipid Accumulating Microbial Communities in Biological Wastewater Treatment Plants. 2013.

Hugo Roume, Emilie EL Muller, Thekla Cordes, Jenny Renaut, Karsten Hiller, and Paul Wilmes. A biomolecular isolation framework for eco-systems biology. *The ISME Journal*, 7(1):110–121, jan 2013. ISSN 1751-7362. doi: 10.1038/ismej.2012.72.

Hugo Roume, Anna Heintz-Buschart, Emilie E L Muller, Patrick May, Venkata P Satagopam, Cédric C Laczny, Shaman Narayanasamy, Laura A Lebrun, Michael R Hoopmann, James M Schupp, John D Gillece, Nathan D Hicks, David M Engelthaler, Thomas Sauter, Paul S Keim, Robert L Moritz, and Paul Wilmes. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms and Microbiomes*, 1 (1):15007, dec 2015. ISSN 2055-5008. doi: 10.1038/npjbiofilms.2015.7.

Johannes Rousk and Per Bengtson. Microbial regulation of global biogeochemical cycles. *Frontiers in Microbiology*, 5(7441):305–307, mar 2014. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00103.

Simon Roux, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, may 2015. ISSN 2167-8359. doi: 10.7717/peerj.985.

Francesco Rubino, Ciara Carberry, Sinéad M Waters, David Kenny, Matthew S. McCabe, and Christopher J. Creevey. Divergent functional isoforms drive niche specialisation for nutrient acquisition and use in rumen microbiome. *ISME Journal*, 11(4):932–944, 2017. ISSN 17517370. doi: 10.1038/ismej.2016.172.

Wael Sabra, David Dietz, Donna Tjahjasari, and An-Ping Zeng. Biosystems analysis and engineering of microbial consortia for industrial biotechnology. *Engineering in Life Sciences*, 10(5): 407–421, oct 2010. ISSN 16180240. doi: 10.1002/elsc.201000111.

R. Saiki, D. Gelfand, S Stoffel, S. Scharf, R Higuchi, G. Horn, K. Mullis, and H. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239 (4839):487–491, jan 1988. ISSN 0036-8075. doi: 10.1126/science.2448875.

Mark Schena, Dari Shalon, R W Davis, and P O Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, oct 1995. ISSN 0036-8075. doi: 10.1126/science.270.5235.467.

Axel Schippers, Jürgen Vasters, and Malte Drobe. Biomining - Entwicklung der Metallgewinnung mittels Mikroorganismen im Bergbau. *Commodity Top News*, 39:1–10, 2011.

Jonas Schluter and Kevin R. Foster. The Evolution of Mutualism in Gut Microbiota Via Host Epithelial Selection. *PLoS Biology*, 10(11), 2012. ISSN 15449173. doi: 10.1371/journal.pbio.1001424.

Thomas Schneider, Katharina M. Keiblinger, Emanuel Schmid, Katja Sterflinger-Gleixner, Günther Ellersdorfer, Bernd Roschitzki, Andreas Richter, Leo Eberl, Sophie Zechmeister-Boltenstern, and Kathrin Riedel. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. *The ISME Journal*, 6(9):1749–1762, sep 2012. ISSN 1751-7362. doi: 10.1038/ismej.2012.11.

Matthew B Scholz, Chien-Chi Lo, and Patrick S G Chain. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current opinion in biotechnology*, 23(1):9–15, feb 2012. ISSN 1879-0429. doi: 10.1016/j.copbio.2011.11.013.

Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts094.

Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(4):2068–2069, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu153.

Abdul Sheik, Emilie Muller, Jean-Nicolas Audinot, Laura Lebrun, Patrick Grysan, and Paul Wilmes. In situ phenotypic heterogeneity among single cells of the filamentous bacterium Candidatus Microthrix parvicella. *The ISME Journal*, under revi:1–6, 2015. ISSN 1751-7362. doi: 10.1038/ismej.2015.181.

Abdul R Sheik, Emilie E L Muller, and Paul Wilmes. A hundred years of activated sludge: time for a rethink. *Frontiers in Microbiology*, 5(March):1–7, 2014. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00047.

Jasmine Shong, Manuel Rafael Jimenez Diaz, and Cynthia H. Collins. Towards synthetic microbial consortia for bioprocessing. *Current Opinion in Biotechnology*, 23(5):798–802, oct 2012. ISSN 09581669. doi: 10.1016/j.copbio.2012.02.001.

Christian M. K. Sieber, Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3(7):836–843, jul 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0171-1.

Alma Siggins, Eoin Gunnigle, and Florence Abram. Exploring mixed microbial community functioning: recent advances in metaproteomics. *FEMS Microbiology Ecology*, 80(2):265–280, apr 2012. ISSN 01686496. doi: 10.1111/j.1574-6941.2011.01284.x.

William T. Sloan, Mary Lunn, Stephen Woodcock, Ian M. Head, Sean Nee, and Thomas P. Curtis. Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, 8(4):732–740, 2006. ISSN 14622912. doi: 10.1111/j.1462-2920.2005.00956.x.

Anubhav Srivastava, Greg Kowalski, Damien Callahan, Peter Meikle, and Darren Creek. Strategies for Extending Metabolomics Studies with Stable Isotope Labelling and Fluxomics. *Metabolites*, 6(4):32, oct 2016. ISSN 2218-1989. doi: 10.3390/metabo6040032.

Ramunas Stepanauskas. Single cell genomics: An individual look at microbes. *Current Opinion in Microbiology*, 15(5):613–620, 2012. ISSN 13695274. doi: 10.1016/j.mib.2012.09.001.

Marc Strous, Eric Pelletier, Sophie Mangenot, Thomas Rattei, Angelika Lehner, Michael W. Taylor, Matthias Horn, Holger Daims, Delphine Bartol-Mavel, Patrick Wincker, Valérie Barbe, Nuria Fonknechten, David Vallenet, Béatrice Segurens, Chantal Schenowitz-Truong, Claudine Médigue, Astrid Collingro, Berend Snel, Bas E. Dutilh, Huub J.M. Op Den Camp, Chris Van Der Drift, Irina Cirpus, Katinka T. Van De Pas-Schoonen, Harry R. Harhangi, Laura Van Niftrik, Markus Schmid, Jan Keltjens, Jack Van De Vossenberg, Boran Kartal, Harald Meier, Dmitrij Frishman, Martijn A. Huynen, Hans Werner Mewes, Jean Weissenbach, Mike S.M. Jetten, Michael Wagner, and Denis Le Paslier. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, 440(7085):790–794, 2006. ISSN 14764687. doi: 10.1038/nature04647.

Shinichi Sunagawa, Daniel R Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon a Berger, Jens Roat Kultima, Luis Pedro Coelho, Manimozhiyan Arumugam, Julien Tap, Henrik Bjørn Nielsen, Simon Rasmussen, Søren Brunak, Oluf Pedersen, Francisco Guarner, Willem M de Vos, Jun Wang, Junhua Li, Joël Doré, S Dusko Ehrlich, Alexandros Stamatakis, and Peer Bork. Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12):1196–1199, dec 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2693.

Haixu Tang, Sujun Li, and Yuzhen Ye. A Graph-Centric Approach for Metagenome-Guided Peptide and Protein Identification in Metaproteomics. *PLoS Computational Biology*, 12(12):1–16, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1005224.

Jane Tang. Microbial Metabolomics. *Current Genomics*, 12(6):391–403, 2011. ISSN 13892029. doi: 10.2174/138920211797248619.

Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010. ISSN 1754-2189. doi: 10.1038/nprot.2009.203.

T Frede Thingstad. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, 45(6):1320–1328, sep 2000. ISSN 00243590. doi: 10.4319/lo.2000.45.6.1320.

Vigdis Torsvik. Prokaryotic Diversity–Magnitude, Dynamics, and Controlling Factors. *Science*, 296(5570):1064–1066, may 2002. ISSN 00368075. doi: 10.1126/science.1071698.

Duy Tin Truong, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, 27 (4):626–638, apr 2017. ISSN 1088-9051. doi: 10.1101/gr.216242.116.

Benjamin J. Tully, Elaina D. Graham, and John F. Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5:170203, jan 2018. ISSN 2052-4463. doi: 10.1038/sdata.2017.203.

Alexander Tveit, Tim Urich, and Mette M Svenning. Metatranscriptomic analysis of Arctic peat soil microbiota. *Applied and environmental microbiology*, 80(July):5761–5772, 2014. ISSN 1098-5336. doi: 10.1128/AEM.01030-14.

Stefka Tyanova, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y. Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9):731–740, 2016. ISSN 15487105. doi: 10.1038/nmeth.3901.

Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004. ISSN 0028-0836. doi: 10.1038/nature02340.

Gene W Tyson, Ian Lo, Brett J. Baker, Eric E Allen, and Phillip Hugenholtz. Genome-Directed Isolation of the Key Nitrogen Fixer. *Society*, 71(10):6319–6324, 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.10.6319.

Yutaka Uyeno, Suguru Shigemori, and Takeshi Shimosato. Effect of Probiotics/Prebiotics on Cattle Health and Productivity. *Microbes and environments*, 30(2):126–132, 2015. ISSN 1342-6311. doi: 10.1264/jsme2.ME14176.

Lissette Valenzuela, An Chi, Simon Beard, Alvaro Orell, Nicolas Guiliani, Jeff Shabanowitz, Donald F Hunt, and Carlos a Jerez. Genomics, metagenomics and proteomics in biomining microorganisms. *Biotechnology advances*, 24(2):197–211, 2006. ISSN 0734-9750. doi: 10.1016/j.biotechadv.2005.09.004.

Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34(9):666–681, 2018. ISSN 13624555. doi: 10.1016/j.tig.2018.05.008.

Maartje A. H. J. van Kessel, Daan R. Speth, Mads Albertsen, Per H. Nielsen, Huub J. M. Op den Camp, Boran Kartal, Mike S. M. Jetten, and Sebastian Lücker. Complete nitrification by a single microorganism. *Nature*, 528(7583):555–559, dec 2015. ISSN 0028-0836. doi: 10.1038/nature16459.

Sonia R. Vartoukian, Richard M. Palmer, and William G. Wade. Strategies for culture of 'unculturable' bacteria. *FEMS Microbiology Letters*, 309(1):1–7, 2010. ISSN 15746968. doi: 10.1111/j.1574-6968.2010.02000.x.

Goutham N Vemuri and Aristos a Aristidou. Metabolic Engineering in the -omics Era: Elucidating and Modulating Regulatory Networks. *Microbiology and Molecular Biology Reviews*, 69(2): 197–216, jun 2005. ISSN 1092-2172. doi: 10.1128/MMBR.69.2.197-216.2005.

Mario Vera, Beate Krok, Sören Bellenberg, Wolfgang Sand, and Ansgar Poetsch. Shotgun proteomics study of early biofilm formation process of Acidithiobacillus ferrooxidans ATCC 23270 on pyrite. *PROTEOMICS*, 13(7):1133–1144, apr 2013. ISSN 16159853. doi: 10.1002/pmic.201200386.

Jan Kjølhede Vester, Mikkel Andreas Glaring, and Peter Stougaard. Improved cultivation and metagenomics as new tools for bioprospecting in cold environments. *Extremophiles*, 19(1):17–29, 2015. ISSN 1431-0651. doi: 10.1007/s00792-014-0704-3.

Michael Wagner and Alexander Loy. Bacterial community composition and function in sewage treatment systems. *Current Opinion in Biotechnology*, 13(3):218–227, jun 2002. ISSN 09581669. doi: 10.1016/S0958-1669(02)00315-4.

Justin W Walley, Ryan C Sartor, Zhouxin Shen, Robert J Schmitz, Kevin J Wu, Mark A Urich, Joseph R Nery, Laurie G Smith, James C Schnable, Joseph R Ecker, and Steven P Briggs. Integration of omic networks in a developmental atlas of maize. *Science*, 353(6301):814–818, aug 2016. ISSN 0036-8075. doi: 10.1126/science.aag1125.

Yi Wang, Devin Coleman-Derr, Guoping Chen, and Yong Q. Gu. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research*, 43(W1):W78–W84, 2015. ISSN 13624962. doi: 10.1093/nar/gkv487.

Zhi-Bin Wang, Ming-Sheng Miao, Qiang Kong, and Shou-Qing Ni. Evaluation of microbial diversity of activated sludge in a municipal wastewater treatment plant of northern China by high-throughput sequencing technology. *Desalination and Water Treatment*, 57(50):23516–23521, 2016. ISSN 1944-3994. doi: 10.1080/19443994.2015.1137232.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, jan 2009. ISSN 1471-0056. doi: 10.1038/nrg2484.

Helen Watling. Microbiological Advances in Biohydrometallurgy. *Minerals*, 6(2):49, 2016. ISSN 2075-163X. doi: 10.3390/min6020049.

H.R. Watling. The bioleaching of sulphide minerals with emphasis on copper sulphides — A review. *Hydrometallurgy*, 84(1-2):81–108, oct 2006. ISSN 0304386X. doi: 10.1016/j.hydromet.2006.05.001.

Daniel Wibberg, Andreas Bremges, Tanja Dammann-Kalinowski, Irena Maus, Mª Isabel Igeño, Ralph Vogelsang, Christoph König, Víctor M. Luque-Almagro, Mª Dolores Roldán, Alexander Sczyrba, Conrado Moreno-Vivián, Rafael Blasco, Alfred Pühler, and Andreas Schlüter. Finished genome sequence and methylome of the cyanide-degrading Pseudomonas pseudoalcaligenes strain CECT5344 as resolved by single-molecule real-time sequencing. *Journal of Biotechnology*, 232:61–68, aug 2016. ISSN 01681656. doi: 10.1016/j.jbiotec.2016.04.008.

Stefanie Widder, Rosalind J Allen, Thomas Pfeiffer, Thomas P Curtis, Carsten Wiuf, William T Sloan, Otto X Cordero, Sam P Brown, Babak Momeni, Wenying Shou, Helen Kettle, Harry J Flint, Andreas F Haas, Béatrice Laroche, Jan-ulrich Kreft, Paul B Rainey, Shiri Freilich, Stefan Schuster, Kim Milferstedt, Jan R van der Meer, Tobias Groβkopf, Jef Huisman, Andrew Free, Cristian Picioreanu, Christopher Quince, Isaac Klapper, Simon Labarthe, Barth F Smets, Harris Wang, and Orkun S Soyer. Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME Journal*, 10(11):2557–2568, nov 2016. ISSN 1751-7362. doi: 10.1038/ismej.2016.45.

S. T. Williams and J. C. Vickers. The ecology of antibiotic production. *Microbial Ecology*, 12(1): 43–52, mar 1986. ISSN 0095-3628. doi: 10.1007/BF02153221.

P. Wilmes, B. P. Bowen, B. C. Thomas, R. S. Mueller, V. J. Denef, N. C. VerBerkmoes, R. L. Hettich, T. R. Northen, and J. F. Banfield. Metabolome-Proteome Differentiation Coupled to Microbial Divergence. *mBio*, 1(5):459–464, oct 2010a. ISSN 2150-7511. doi: 10.1128/mBio.00246-10.

Paul Wilmes and Philip L. Bond. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, 6(9):911–920, 2004. ISSN 14622912. doi: 10.1111/j.1462-2920.2004.00687.x.

Paul Wilmes and Philip L. Bond. Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. *Current Opinion in Microbiology*, 12(3):310–317, 2009. ISSN 13695274. doi: 10.1016/j.mib.2009.03.004.

Paul Wilmes, Jonathan P Remis, Mona Hwang, Manfred Auer, Michael P Thelen, and Jillian F Banfield. Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *The ISME journal*, 3(2):266–70, feb 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.90.

Paul Wilmes, Benjamin P. Bowen, Brian C. Thomas, Ryan S. Mueller, Vincent J. Denef, Nathan C. VerBerkmoes, Robert L. Hettich, Trent R. Northen, and J. F. Banfield. Metabolome-Proteome Differentiation Coupled to Microbial Divergence. *mBio*, 1(5):3–7, oct 2010b. ISSN 2150-7511. doi: 10.1128/mBio.00246-10.

Paul Wilmes, Anna Heintz-Buschart, and Philip L. Bond. A decade of metaproteomics: Where we stand and what the future holds. *Proteomics*, 15(20):3409–3417, 2015. ISSN 16159861. doi: 10.1002/pmic.201500183.

Katherine Wolstencroft, Olga Krebs, Jacky L. Snoep, Natalie J. Stanford, Finn Bacall, Martin Golebiewski, Rostyk Kuzyakiv, Quyen Nguyen, Stuart Owen, Stian Soiland-Reyes, Jakub Straszewski, David D. van Niekerk, Alan R. Williams, Lars Malmström, Bernd Rinn, Wolfgang Müller, and Carole Goble. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, 45(D1):D404–D407, jan 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1032.

Dongying Wu, Guillaume Jospin, and Jonathan A. Eisen. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS ONE*, 8(10), 2013. ISSN 19326203. doi: 10.1371/journal.pone.0077033.

Martin Wu and Alexandra J. Scott. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, 28(7):1033–1034, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts079.

Chao Xie, Chin Lui Wesley Goi, Daniel H. Huson, Peter F. R. Little, and Rohan B. H. Williams. RiboTagger: fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or metatranscriptome surveys. *BMC Bioinformatics*, 17(S19):508, dec 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1378-x.

Shuang Xu, Junqin Yao, Meihaguli Ainiwaer, Ying Hong, and Yanjiang Zhang. Analysis of Bacterial Community Structure of Activated Sludge from Wastewater Treatment Plants in Winter. 2018, 2018. ISSN 23146141. doi: 10.1155/2018/8278970.

Chao Yang, Wei Zhang, Ruihua Liu, Qiang Li, Baobin Li, Shufang Wang, Cunjiang Song, Chuanling Qiao, and Ashok Mulchandani. Phylogenetic diversity and metabolic potential of activated sludge microbial communities in full-scale wastewater treatment plants. *Environmental Science and Technology*, 45(17):7408–7415, 2011. ISSN 0013936X. doi: 10.1021/es2010545.

Zheng Yu, Sascha M.B. Krause, David A.C. Beck, and Ludmila Chistoserdova. A synthetic ecology perspective: How well does behavior of model organisms in the laboratory predict microbial activities in natural habitats? *Frontiers in Microbiology*, 7(JUN):1–7, 2016. ISSN 1664302X. doi: 10.3389/fmicb.2016.00946.

Aleksej Zelezniak, Sergej Andrejev, Olga Ponomarova, Daniel R Mende, Peer Bork, and Kiran Raosaheb Patil. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences*, 112(20):6449–6454, may 2015. ISSN 0027-8424. doi: 10.1073/pnas.1421834112.

K. Zengler. Central Role of the Cell in Microbial Ecology. *Microbiology and Molecular Biology Reviews*, 73(4):712–729, 2009. ISSN 1092-2172. doi: 10.1128/MMBR.00027-09.

Tong Zhang, Ming Fei Shao, and Lin Ye. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *ISME Journal*, 6(6):1137–1147, 2012. ISSN 17517362. doi: 10.1038/ismej.2011.188.

Xian Zhang, Xueduan Liu, Yili Liang, Yunhua Xiao, Liyuan Ma, Xue Guo, Bo Miao, Hongwei Liu, Deliang Peng, Wenkun Huang, and Huaqun Yin. Comparative genomics unravels the functional roles of co-occurring acidophilic bacteria in bioleaching heaps. *Frontiers in Microbiology*, 8 (MAY):1–15, 2017. ISSN 1664302X. doi: 10.3389/fmicb.2017.00790.

Xian Zhang, Xueduan Liu, Fei Yang, and Lv Chen. Pan-Genome Analysis Links the Hereditary Variation of Leptospirillum ferriphilum With Its Evolutionary Adaptation. *Frontiers in Microbiology*, 9(MAR):1–12, mar 2018. ISSN 1664-302X. doi: 10.3389/fmicb.2018.00577.

Matthias Zimmermann, Stéphane Escrig, Thomas Hübschmann, Mathias K. Kirf, Andreas Brand, R. Fredrik Inglis, Niculina Musat, Susann Müller, Anders Meibom, Martin Ackermann, and Frank Schreiber. Phenotypic heterogeneity in metabolic traits among single cells of a rare bacterial species in its natural environment quantified with a combination of flow cell sorting and NanoSIMS. *Frontiers in Microbiology*, 06(MAR):1–11, apr 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00243.

Appendices

APPENDIX A

ADDITIONAL FILES AND TABLES

A.1 Additional file 2.1: *L. ferriphilum* genelists with functional categories

Lists of functional categories and gene functions to protein coding genes (Tables S1 - S8) The file is available in the original publication [**Appendix C.1**] and the following link.
`https://aem.asm.org/highwire/filestream/15501/field_highwire_adjunct_files/0/zam003188291s1.pdf`.

A.2 Additional file 2.2: *L. ferriphilum* combined omics data and additional functional annotations

The file is available via the original publication [**Appendix C.1**] and available through the following link:
`https://doi.org/10.15490/fairdomhub.1.datafile.1807.5`.

A.3 Additional File 3.1: Overview of recovered ReGes and isolates

An overview of recovered representative genomes and sequenced isolate genomes recovered from oleaginous floating sludge. The file is available under the following link:
`https://dropit.uni.lu/invitations?share=3414a4385a3912d3b729&dl=0`

APPENDIX B

ADDITIONAL FIGURES

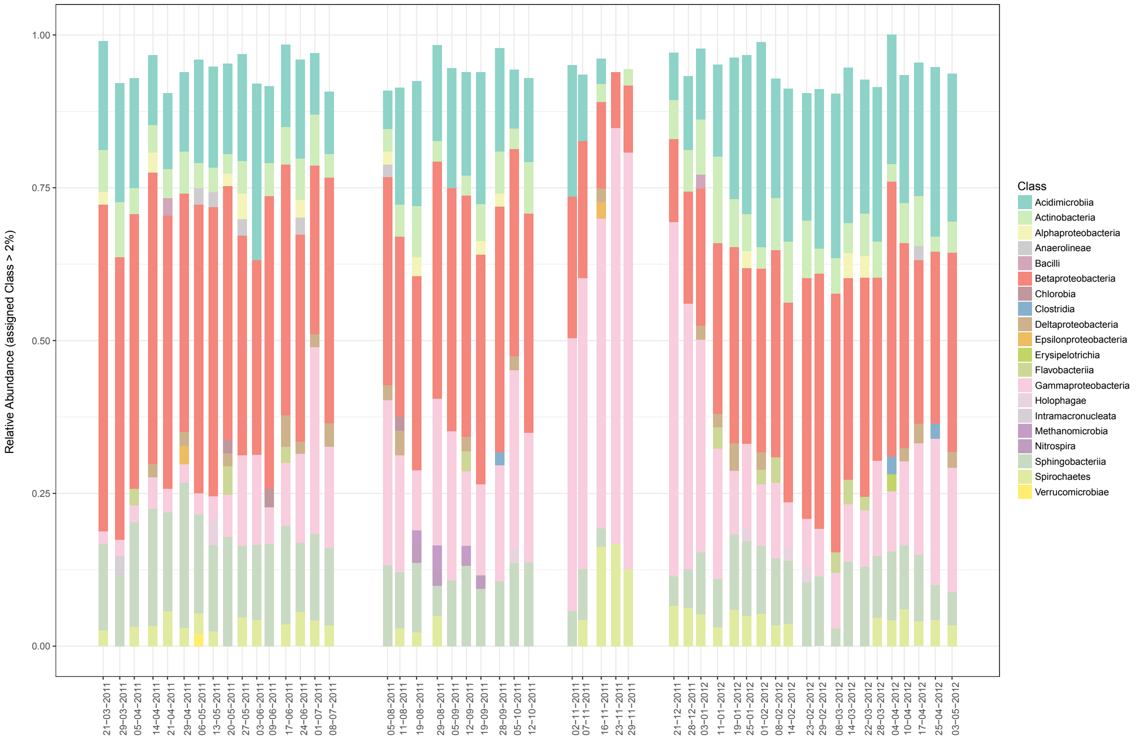


Figure B.1: Relative abundance based on OTU-counts based on extracted ribotags (region v4) from the metagenomic (MG) reads over the time-series. Relative OTU-counts summed on class-level taxonomic assignment, including groups only above 2%. Only OTUs with class-level assignment are shown.

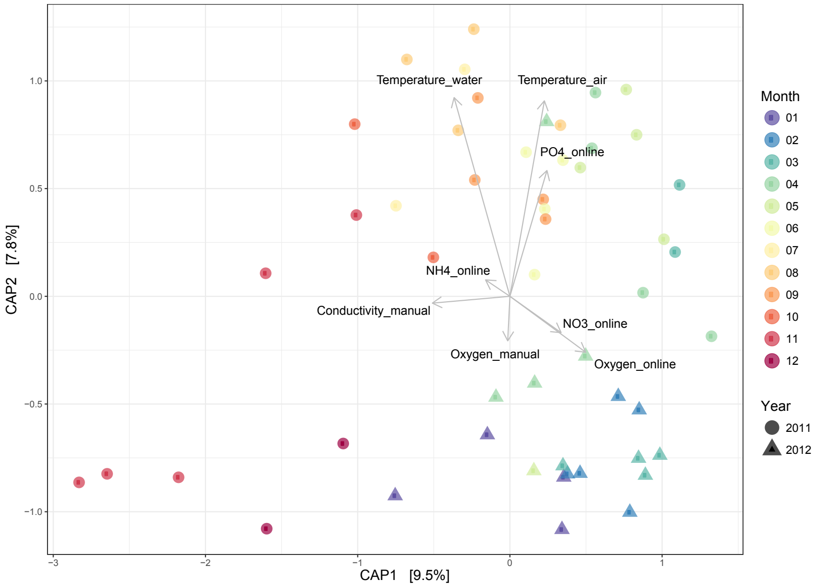


Figure B.2: Constrained ordination of species composition based on Bray-Curtis dissimilarity of relative OTU-counts (based on ribotag assignment) constrained by selected abiotic factors (labelled arrowheads). The function ordinate of the R-package physloseq (method="CAP") was used. Points are coloured by month of sampling and point-shape reflects the year of sampling. Arrow length indicates environmental scores

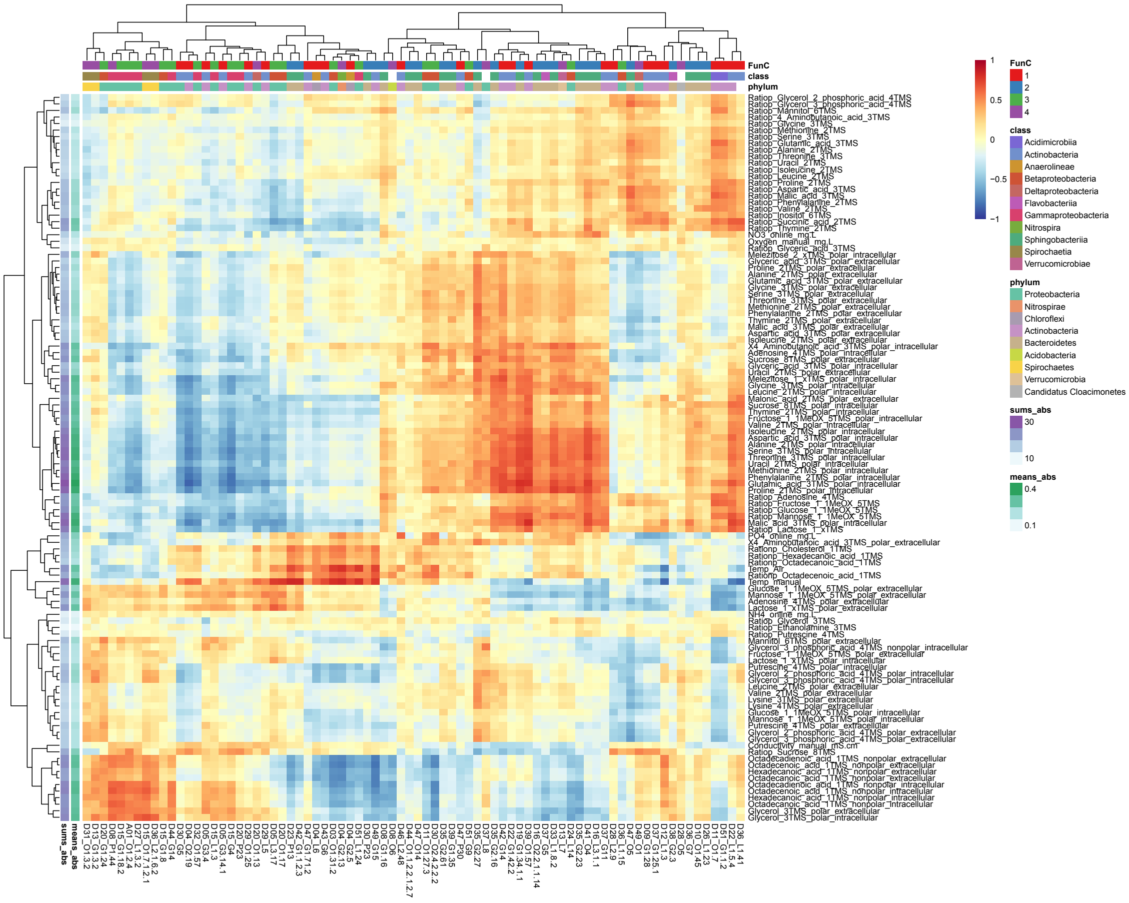


Figure B.3: Correlations between ReGes (completeness-filtered set of 78 ReGes) relative abundance time-courses and time-courses of pool and z-score normalized metabolite intensities (median of measurement replicates) and physico-chemical parameter levels (Spearman rank correlation, cor.test in R). Red indicates a positive correlation coefficient, respectively blue shows negative coefficients. Column annotation tracks indicate class- and phylum-level taxonomic assignments and FunC assignments. Row annotation tracks show mean and sum of absolute correlation coefficients of the respective parameter over all ReGe abundances.

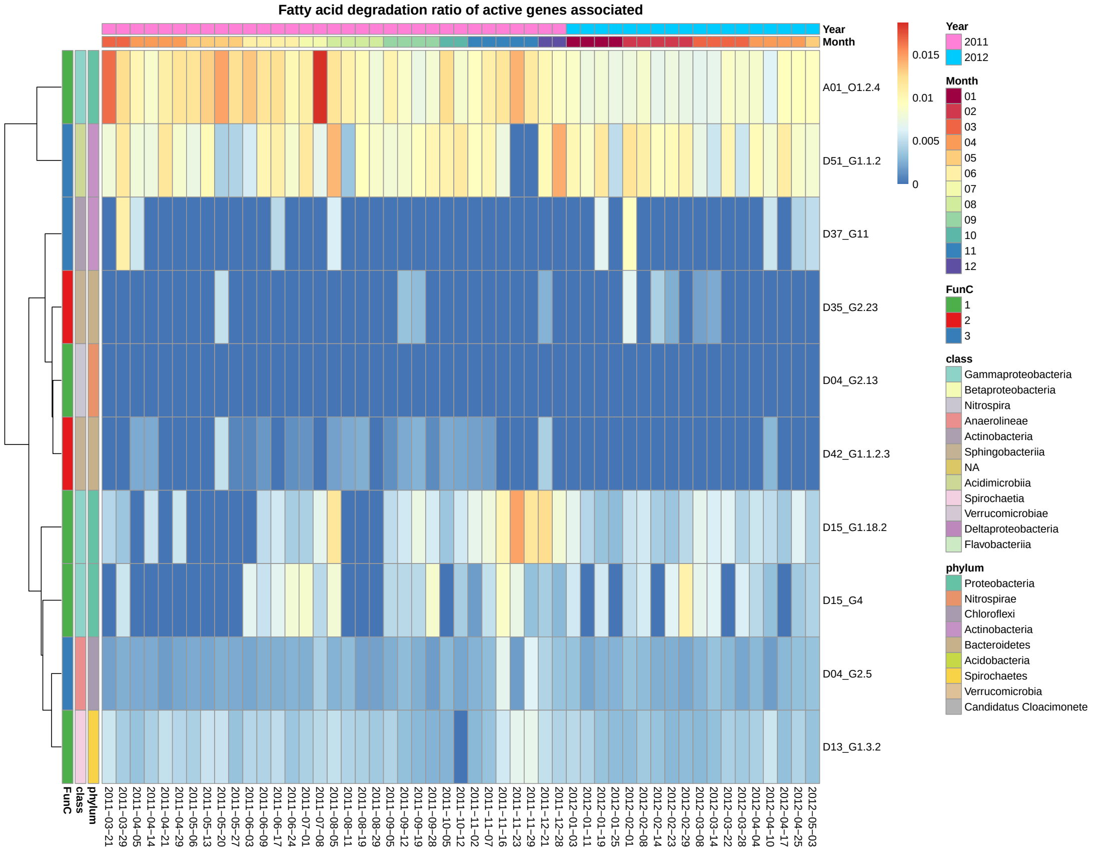


Figure B.4: Ratios of active genes associated to fatty acid degradation (KEGG orthologue assignment to KEGG pathway link) divided by all active genes per time-point. Columns show the values of the ratios for a subset of selected ReGes (rows).

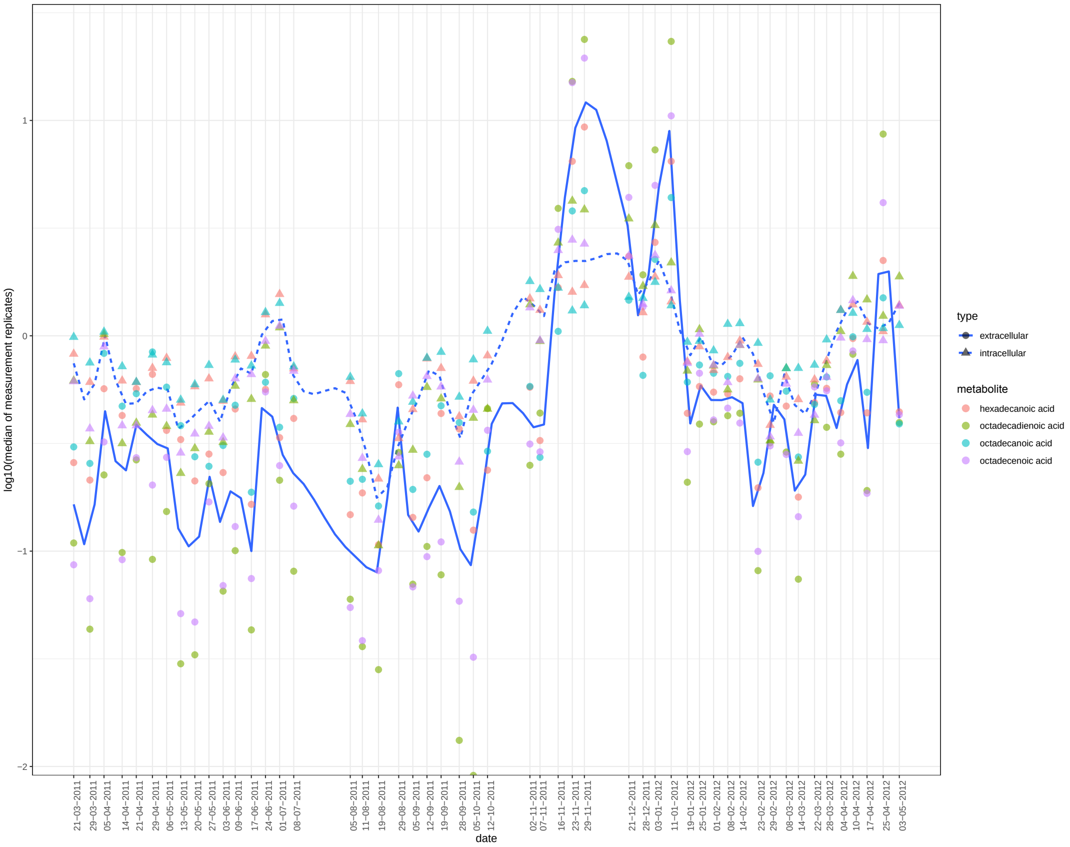


Figure B.5: Intensities of pool-normalized metabolite intensities over time. Points represent log10-scaled intensity values (median of measurement replicates) with triangles for intracellular measurements and circles for extracellular measurements. Dashed lines represent a loess smoothing of the intracellular metabolite intensities, while solid lines show a loess smoothing of extracellular metabolite intensities. Smoothing lines over the whole set of grouped metabolite intensities are shown. Only confidently identifiable metabolites present in the both measurements for intra- and extracellular metabolites are shown and represent the most abundant derivate. Derivate names have been replaced with associated CHEBI-nomenclature names.

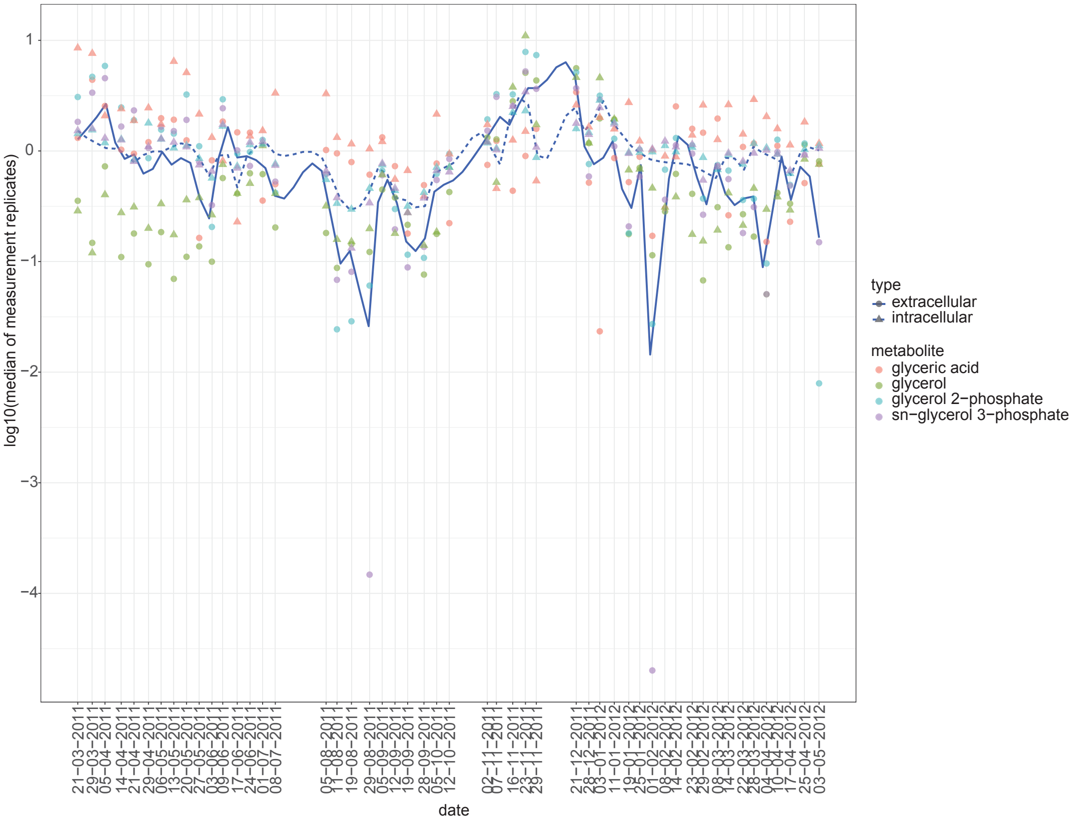


Figure B.6: Intensities of pool-normalized metabolite intensities over time. Points represent log10-scaled intensity values (median of measurement replicates) with triangles for intracellular measurements and circles for extracellular measurements. Dashed lines represent a loess smoothing of the intracellular metabolite intensities, while solid lines show a loess smoothing of extracellular metabolite intensities. Smoothing lines over the whole set of grouped metabolite intensities are shown. Only confidently identifiable metabolites present in the both measurements for intra- and extracellular metabolites are shown and represent the most abundant derivate. Derivate names have been replaced with associated CHEBI-nomenclature names.

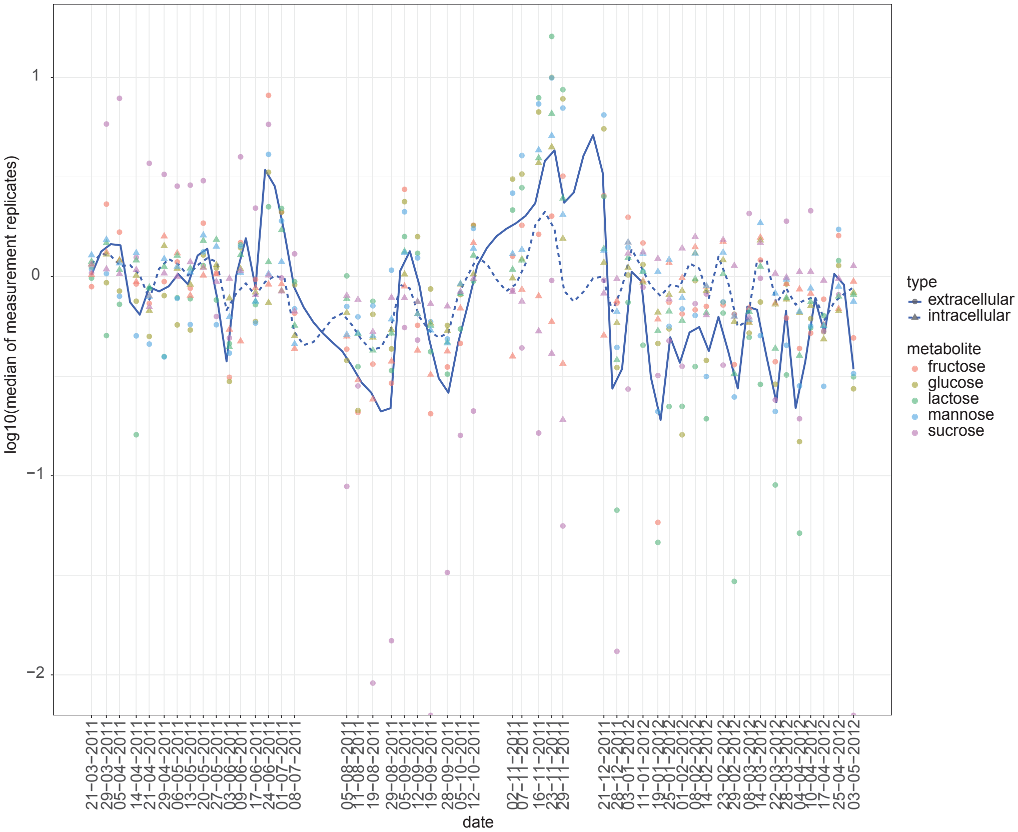


Figure B.7: Intensities of pool-normalized metabolite intensities over time. Points represent log10-scaled intensity values (median of measurement replicates) with triangles for intracellular measurements and circles for extracellular measurements. Dashed lines represent a loess smoothing of the intracellular metabolite intensities, while solid lines show a loess smoothing of extracellular metabolite intensities. Smoothing lines over the whole set of grouped metabolite intensities are shown. Only confidently identifiable metabolites present in the both measurements for intra- and extracellular metabolites are shown and represent the most abundant derivate. Derivate names have been replaced with associated CHEBI-nomenclature names.

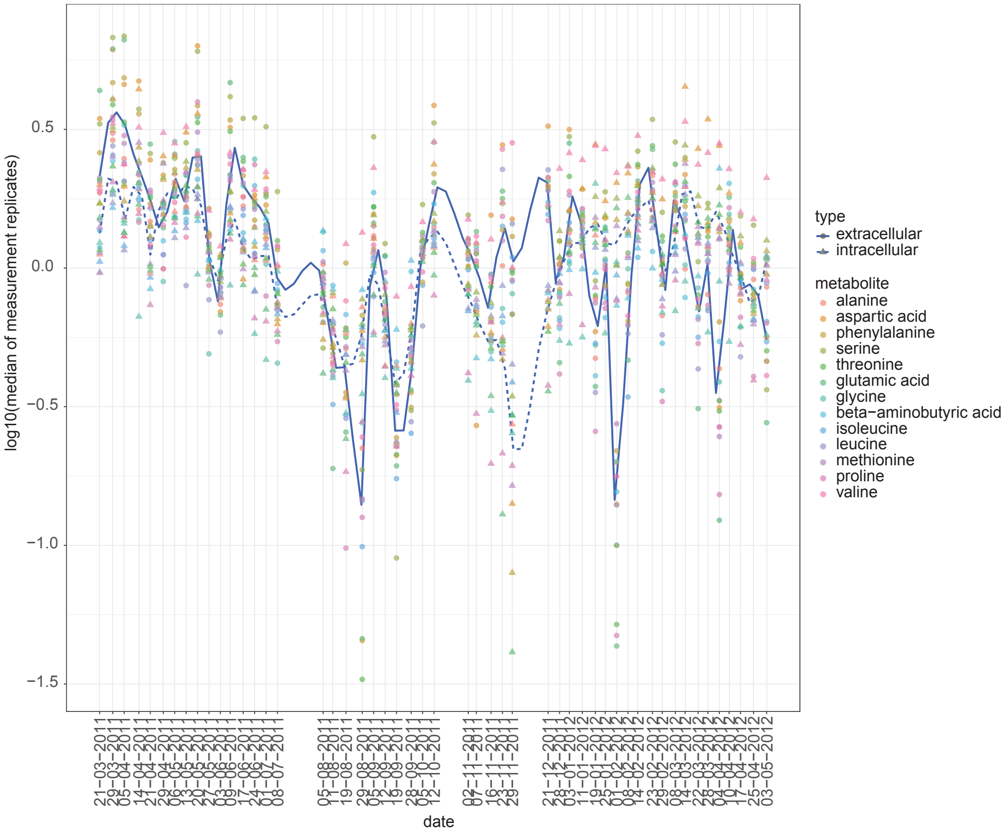


Figure B.8: Intensities of pool-normalized metabolite intensities over time. Points represent log10-scaled intensity values (median of measurement replicates) with triangles for intracellular measurements and circles for extracellular measurements. Dashed lines represent a loess smoothing of the intracellular metabolite intensities, while solid lines show a loess smoothing of extracellular metabolite intensities. Smoothing lines over the whole set of grouped metabolite intensities are shown. Only confidently identifiable metabolites present in the both measurements for intra- and extracellular metabolites are shown and represent the most abundant derivate. Derivate names have been replaced with associated CHEBI-nomenclature names.

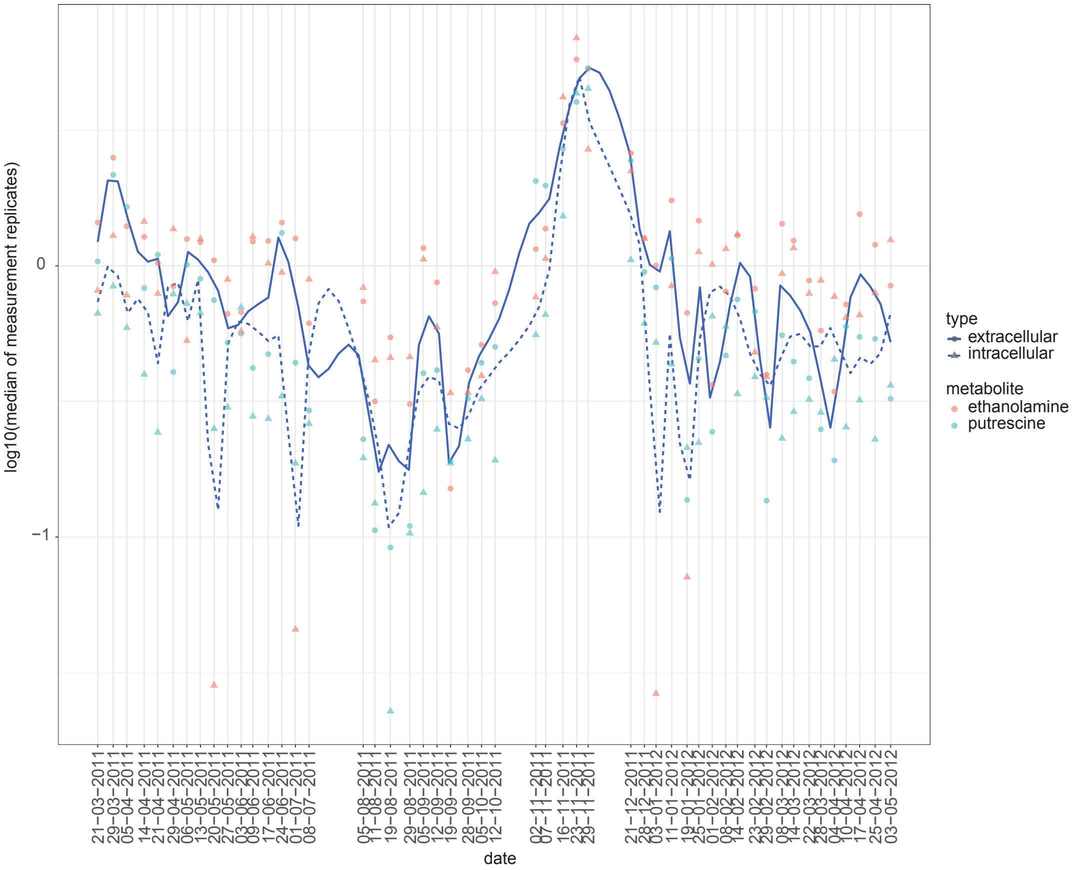


Figure B.9: Intensities of pool-normalized metabolite intensities over time. Points represent log10-scaled intensity values (median of measurement replicates) with triangles for intracellular measurements and circles for extracellular measurements. Dashed lines represent a loess smoothing of the intracellular metabolite intensities, while solid lines show a loess smoothing of extracellular metabolite intensities. Smoothing lines over the whole set of grouped metabolite intensities are shown. Only confidently identifiable metabolites present in the both measurements for intra- and extracellular metabolites are shown and represent the most abundant derivate. Derivate names have been replaced with associated CHEBI-nomenclature names.

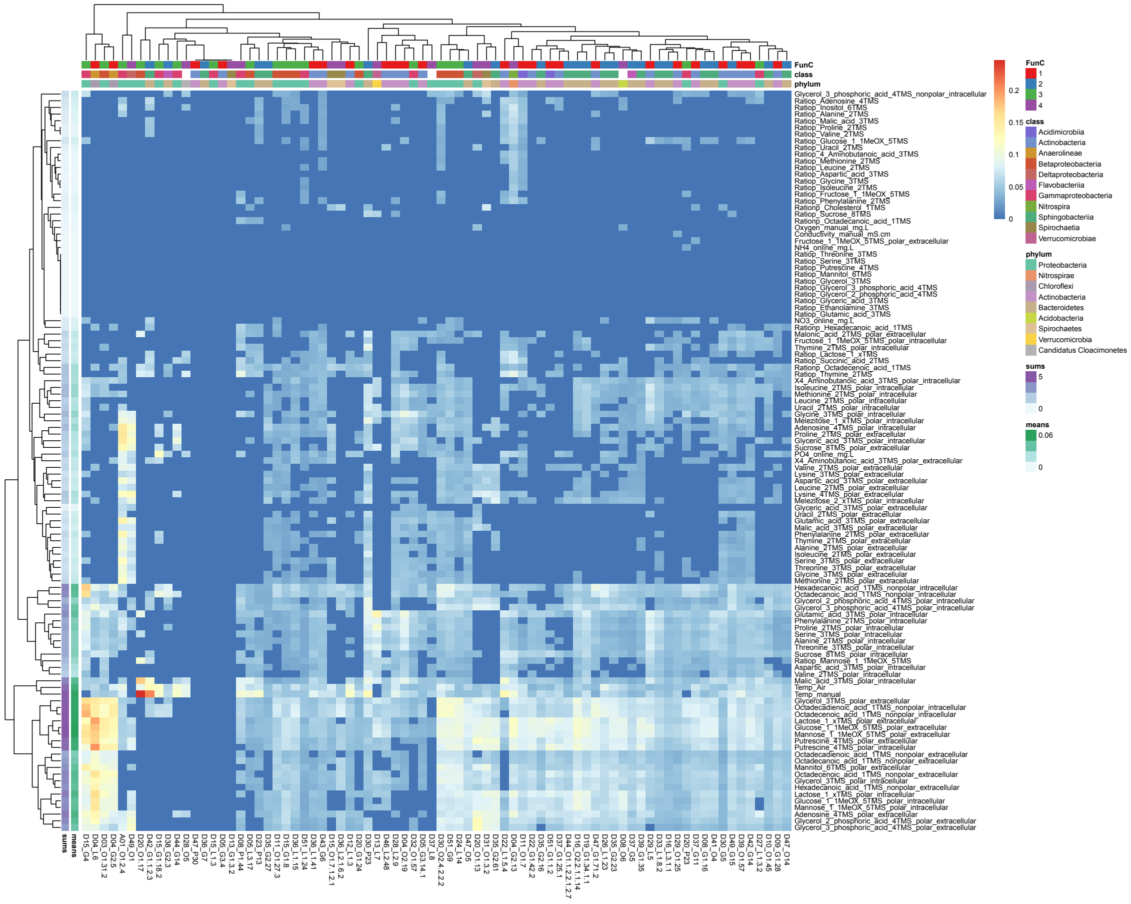


Figure B.10: Correlations between ReGes (completeness-filtered set of 78 ReGes) relative abundance time-courses and time-courses of pool and z-score normalized metabolite intensities (median of measurement replicates) and physico-chemical parameter levels (Spearman rank correlation, cor.test in R). Red indicates a positive correlation coefficient, respectively blue shows negative coefficients. Column annotation tracks indicate class- and phylum-level taxonomic assignments and FunC assignments. Row annotation tracks show mean and sum of absolute correlation coefficients of the respective parameter over all ReGe abundances.

APPENDIX C

ARTICLE MANUSCRIPTS

The appendix contains all manuscripts authored as a first author or co-author. Journal formatted articles are provided for published manuscripts. Manuscripts currently under review are provided as the submitted versions.

C.1 Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species *Leptospirillum ferriphilum*^{T}.

Stephan Christel[†], **Malte Herold**[†], Sören Bellenberg, Mohamed El Hajjami, Antoine Buetti-Dinh,
Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, Mark Dopson

2017
Applied and Environmental Microbiology **84**: e02091-17
DOI: 10.1128/AEM.02091-17

Contributions of author include:

- Coordination
- Analytical research design
- Software development
- Data analysis and visualization
- Writing and revision of manuscript

[†] Co-first author

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

EVOLUTIONARY AND GENOMIC MICROBIOLOGY

Multi-omics Reveals the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species *Leptospirillum ferriphilum*^{T}

Stephan Christel,[a] Malte Herold,[b] Sören Bellenberg,[c] Mohamed El Hajjami,[d] Antoine Buetti-Dinh,[e,f] Igor V. Pivkin,[e,f] Wolfgang Sand,[c,g,h] Paul Wilmes,[b] Ansgar Poetsch,[d,i] Mark Dopson[a]

[a]Centre for Ecology and Evolution in Microbial Model Systems, Linnaeus University, Kalmar, Sweden
[b]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg
[c]Aquatic Biotechnology, Universität Duisburg-Essen, Essen, Germany
[d]Plant Biochemistry, Ruhr Universität Bochum, Bochum, Germany
[e]Institute of Computational Science, Faculty of Informatics, Università della Svizzera Italiana, Lugano, Switzerland
[f]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[g]College of Environmental Science and Engineering, Donghua University, Shanghai, People's Republic of China
[h]Mining Academy and Technical University Freiberg, Freiberg, Germany
[i]School of Biomedical and Healthcare Sciences, Plymouth University, Plymouth, United Kingdom

ABSTRACT *Leptospirillum ferriphilum* plays a major role in acidic, metal-rich environments, where it represents one of the most prevalent iron oxidizers. These milieus include acid rock and mine drainage as well as biomining operations. Despite its perceived importance, no complete genome sequence of the type strain of this model species is available, limiting the possibilities to investigate the strategies and adaptations that *Leptospirillum ferriphilum* DSM 14647^T (here referred to as *Leptospirillum ferriphilum*^{T}) applies to survive and compete in its niche. This study presents a complete, circular genome of *Leptospirillum ferriphilum*^{T} obtained by PacBio single-molecule real-time (SMRT) long-read sequencing for use as a high-quality reference. Analysis of the functionally annotated genome, mRNA transcripts, and protein concentrations revealed a previously undiscovered nitrogenase cluster for atmospheric nitrogen fixation and elucidated metabolic systems taking part in energy conservation, carbon fixation, pH homeostasis, heavy metal tolerance, the oxidative stress response, chemotaxis and motility, quorum sensing, and biofilm formation. Additionally, mRNA transcript counts and protein concentrations were compared between cells grown in continuous culture using ferrous iron as the substrate and those grown in bioleaching cultures containing chalcopyrite ($CuFeS_2$). Adaptations of *Leptospirillum ferriphilum*^{T} to growth on chalcopyrite included the possibly enhanced production of reducing power, reduced carbon dioxide fixation, as well as elevated levels of RNA transcripts and proteins involved in heavy metal resistance, with special emphasis on copper efflux systems. Finally, the expression and translation of genes responsible for chemotaxis and motility were enhanced.

IMPORTANCE *Leptospirillum ferriphilum* is one of the most important iron oxidizers in the context of acidic and metal-rich environments during moderately thermophilic biomining. A high-quality circular genome of *Leptospirillum ferriphilum*^{T} coupled with functional omics data provides new insights into its metabolic properties, such as the novel identification of genes for atmospheric nitrogen fixation, and represents an essential step for further accurate proteomic and transcriptomic investigation of this acidophile model species in the future. Additionally, light is shed on adaptation strategies of *Leptospirillum ferriphilum*^{T} for growth on the copper mineral chalcopyrite. These data can be applied to deepen our understanding and optimization of bi-

Received 22 September 2017 **Accepted** 9 November 2017
Accepted manuscript posted online 17 November 2017
Citation Christel S, Herold M, Bellenberg S, El Hajjami M, Buetti-Dinh A, Pivkin IV, Sand W, Wilmes P, Poetsch A, Dopson M. 2018. Multi-omics reveals the lifestyle of the acidophilic, mineral-oxidizing model species *Leptospirillum ferriphilum*^{T}. Appl Environ Microbiol 84:e02091-17. https://doi.org/10.1128/AEM.02091-17.
Editor M. Julia Pettinari, University of Buenos Aires

Copyright © 2018 Christel et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Mark Dopson, mark.dopson@lnu.se.

S.C. and M.H. contributed equally to this work.

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

oleaching and biooxidation, techniques that present sustainable and environmentally friendly alternatives to many traditional methods for metal extraction.

KEYWORDS *Leptospirillum*, acidophile, biomining, genome, metabolism, omics, proteome, transcriptome

The *Leptospirillum* genus comprises four described species of Gram-negative, chemolithoautotrophic, and acidophilic bacteria: *Leptospirillum ferrooxidans* (group I) (1), *Leptospirillum rubarum* (group II) (2), and *Leptospirillum ferriphilum* and "*Leptospirillum ferrodiazotrophum*" (group III) (2, 3). In addition, community genomics has identified a further candidate species, "*Leptospirillum* sp. group IV UBA BS" (2, 4). The original description of the *L. ferriphilum* type strain gives a temperature optimum of 30°C to 37°C, although many isolated strains are defined as being moderately thermophilic (reviewed in reference 5). *Leptospirillum ferriphilum* DSM 14647^T (here referred to as *Leptospirillum ferriphilum*^{T}) is an obligate aerobe that is capable of gaining energy only via ferrous iron (Fe^{2+}) oxidation (3). Finally, it has a pH optimum of 1.4 to 1.8, which requires the cells to maintain an internal, cytoplasmic pH close to neutral in the face of an $\sim 10^4$-fold proton gradient across the cytoplasmic membrane. As a result, acidophiles have several pH homeostasis mechanisms, including primary (1°) and secondary (2°) proton pumps, an inside positive membrane potential that hinders the influx of protons, proton-consuming reactions, and a cytoplasmic buffering capacity (reviewed in reference 6). Although several *Leptospirillum* spp. have been identified, current knowledge of how they obtain energy and nutrients for growth is limited. In particular, mechanisms for nitrogen fixation have been under debate. Additionally, the understanding of how members of the leptospirilli survive at acidic pH lags behind that of other acidophiles, such as those from the *Acidithiobacillus* genus (reviewed in references 5 and 7–9).

Leptospirillum spp. are often identified in sulfide mineral-containing environments, where they catalyze the cleavage of the metal sulfide bond by oxidizing ferrous iron (Fe^{2+}) back to ferric iron (Fe^{3+}) (10). The result of metal sulfide oxidation is an acidic solution typically containing high metal concentrations (reviewed in references 11 and 12). This requires acidophiles to have multiple chemical and biological metal resistance strategies, such as efflux pumps, metal sequestration methods, and the ability to reduce metal uptake via the inside positive membrane potential (reviewed in references 12 and 13). A second consequence of high iron concentrations is the need to mitigate oxidative stress (14), as acidophiles generate intracellular reactive oxygen species (ROS) as well as being exposed to extracellular ROS sources generated by reactions between water or molecular oxygen, dissolved metal ions, and/or surface-bound metal ions on metal sulfides (15, 16). As mentioned above, knowledge of how the leptospirilli survive high metal concentrations and ROS is limited.

The ability to catalyze mineral dissolution has been exploited in the industrial process of "biomining" (reviewed in reference 17), where *L. ferriphilum* dominates biooxidation tanks for the recovery of gold (3) and has been identified in bioleaching heaps for the recovery of copper from chalcopyrite ($CuFeS_2$) (e.g., see reference 18). However, efficient chalcopyrite dissolution in low-cost bioheaps is challenging under mesophilic and moderately thermophilic conditions (19). A critical stage in biomining is cell attachment and biofilm formation on the ore surface (11). Consequently, understanding the genetic basis for cell attachment on metal sulfides may help in the design of strategies to stimulate bioleaching rates, speed up the initiation of bioleaching operations, and improve the persistence of active cells in heap bioleaching operations.

The identification of the genes responsible for biological processes in acidophiles has been hindered until the very recent development of gene knockout systems (e.g., see reference 20), and these methods are still lacking for the leptospirilli. A method to circumvent this limitation is the identification of gene homologs in genome sequences and "metagenome-assembled genomes" (MAGs) that have been used to construct models of individual acidophile strains (reviewed in reference 21) through to commu-

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

TABLE 1 Overview of previously available *L. ferriphilum* genomes

Strain	Reference or source	NCBI RefSeq accession no.	State of the genome	No. of genes	Genome size (Mbp)	Coding density (%)
L. ferriphilum^T	24	NZ_JPGK00000000.1	Draft	2,366	2.41	93.1
Sp-Cl	89	NZ_LGSH00000000.1	Draft	2,419	2.48	91.7
YSK	86	NZ_CP007243.1	Complete	2,273	2.33	90.1
ML-04	90	NC_018649.1	Complete	2,475	2.41	90.3
DX	91	NZ_MPOJ00000000.1	Draft	2,324	2.36	85.8
ZJ	91	NZ_MPOK00000000.1	Draft	2,312	2.34	96.4

nity interactions (22). Although several genomes and MAGs from *L. ferriphilum* strains have been reported (Table 1), the fact that only a draft genome of the *L. ferriphilum* type strain is available has hindered efforts to elucidate its metabolic properties and evolutionary relationships with the other leptospirilli.

The present study provides the complete, closed genome of *L. ferriphilum*^{T} that allows metabolic insights and reveals evolutionary relationships to the leptospirilli and other acidophiles. In addition, we have used RNA transcript sequencing and proteomics to identify the genes used for growth on Fe^{2+} and during biomining of chalcopyrite.

RESULTS AND DISCUSSION

General genome data. The sequencing and assembly of *L. ferriphilum*^{T} DNA gave two polished contiguous sequences (contigs) (Table 2; see also Report S1 in the supplemental material). Contig 1 was 2,569,357 bases with a depth of coverage of 574-fold, while contig 2 was 41,141 bases with a depth of coverage of 33-fold. Contig 2 was predicted to be a putative phage with VIRSorter (23), and a region on contig 1 with high similarity to contig 2 putatively represents a prophage. Although further analysis is required to determine its origin, contig 2 was excluded due to its low-depth coverage and absence of typical plasmid genes. Circular contig 1 represents the closed chromosome sequence of *L. ferriphilum*^{T} (Fig. 1). A comparison with the previously available draft genome (24) revealed an additional 163,475 bp, closing gaps in the previous draft (Fig. S1). The most prominent gap with around 100,000 bp most likely had not been previously captured due to the presence of a clustered regularly interspaced short palindromic repeat (CRISPR) stretch (Report S2). Further functionalities in the additional sequences were identified, such as a cluster of *nif* genes. Additional functional capabilities encoded in the *L. ferriphilum*^{T} genome (Fig. 2) are detailed below, while expressed functions in Fe^{2+}-containing medium versus chalcopyrite bioleaching cultures were assessed by transcriptomic and proteomic analyses (Table 3 and Fig. 3). The resulting values are given as transcripts per million base pairs (TPM) for RNA and label-free quantification (LFQ) intensities (25) for proteins, respectively.

Comparison with other *Leptospirillum ferriphilum* genomes. The genomes of six *L. ferriphilum* strains are available (Table 1), two of which are considered complete. All six genomes show a high degree of identity, with 1,769 orthologous gene clusters conserved in all 6 strains (see Fig. S2 in the supplemental material). The type strain exhibited the largest number of unique gene clusters, which is reflected in the

TABLE 2 General *L. ferriphilum*^{T} genome statistics

Attribute	Value	% of total
Genome size (bp)	2,569,357	100.00
DNA coding region (bp)	2,331,855	90.76
DNA G+C content (bp)	1,392,384	54.19
Total no. of genes	2,541	100
No. of protein-encoding genes	2,486	97.84
No. of RNA genes (rRNA/tRNA/tmRNA)	6/48/1	0.24/1.93/0.04
No. of CDSs with functional prediction	1,846	74.25
No. of CDSs with assigned COG category	1,969	79.20
No. of CRISPR repeats	1	

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

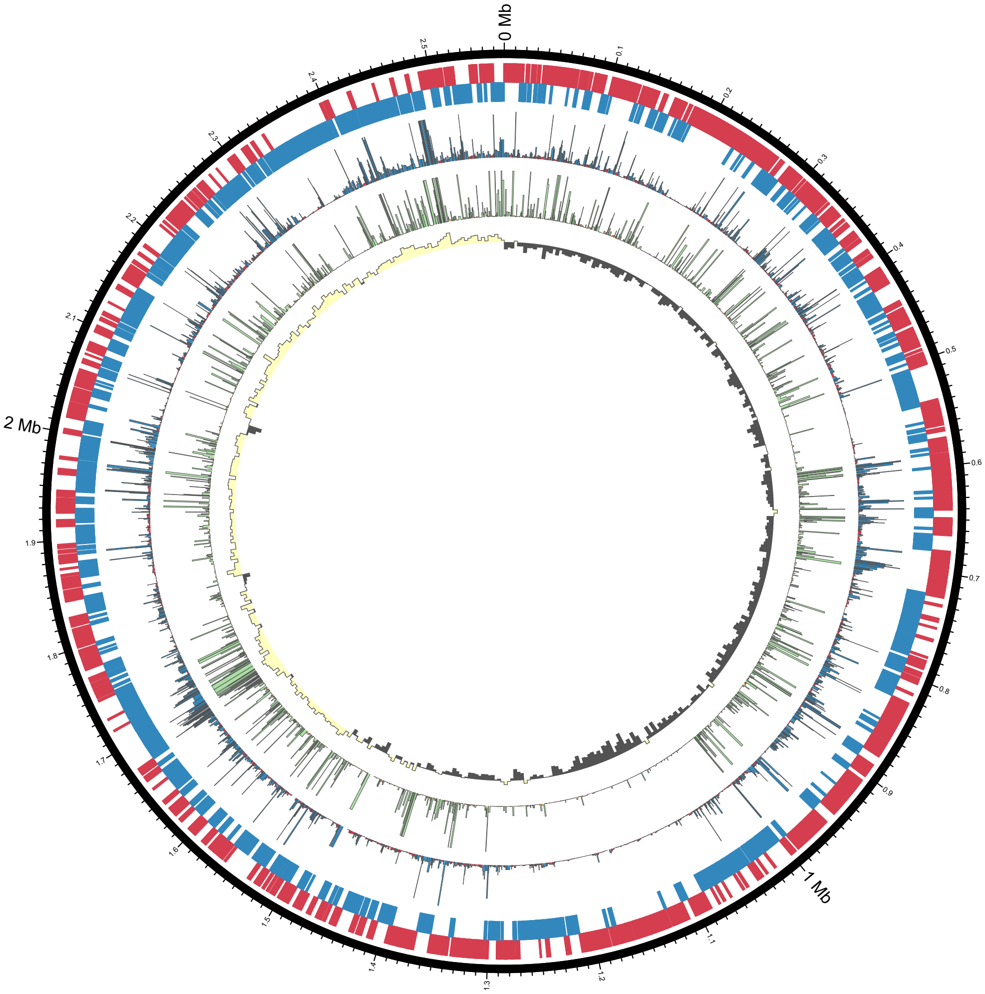


FIG 1 Circular representation (87) of the genome sequence of *L. ferriphilum*^{T} (configuration in parts based on data in reference 88). From the outside, the bands represent (i) the genome sequence; (ii) protein-encoding sequences on the positive strand (red); (iii) CDSs on the negative strand (blue); (iv) mean transcript expression (TPM), with a maximum of 2,000 TPM (blue indicates TPM values above the median, and red indicates values below the median); (v) mean scaled protein LFQ intensity, with a maximum of 2,000 (green indicates intensity above the median); and (vi) GC-Skew {as calculated by the equation $[(C - G)/(C + G)] \times 100$} in windows of 5,000 nucleotides (yellow indicates values above zero, and gray indicates values below zero).

outgrouping of the two type strain sequences in the phylogenetic tree (Fig. S2A). Many of the genes that are distinct for a particular strain seem to be related to insertions or deletions of mobile elements. A cluster of *nif* genes is harbored in the newly sequenced type strain genome and in strains Sp-Cl, YSK, and ZJ. This gene cluster is not present in strains ML-04 and DX and the previous draft of the type strain. The longer stretch (1.0 and 1.3 Mbp) surrounding the *nif* cluster is unique to the type strain, with large parts being present only in the newly sequenced contig (data not shown).

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

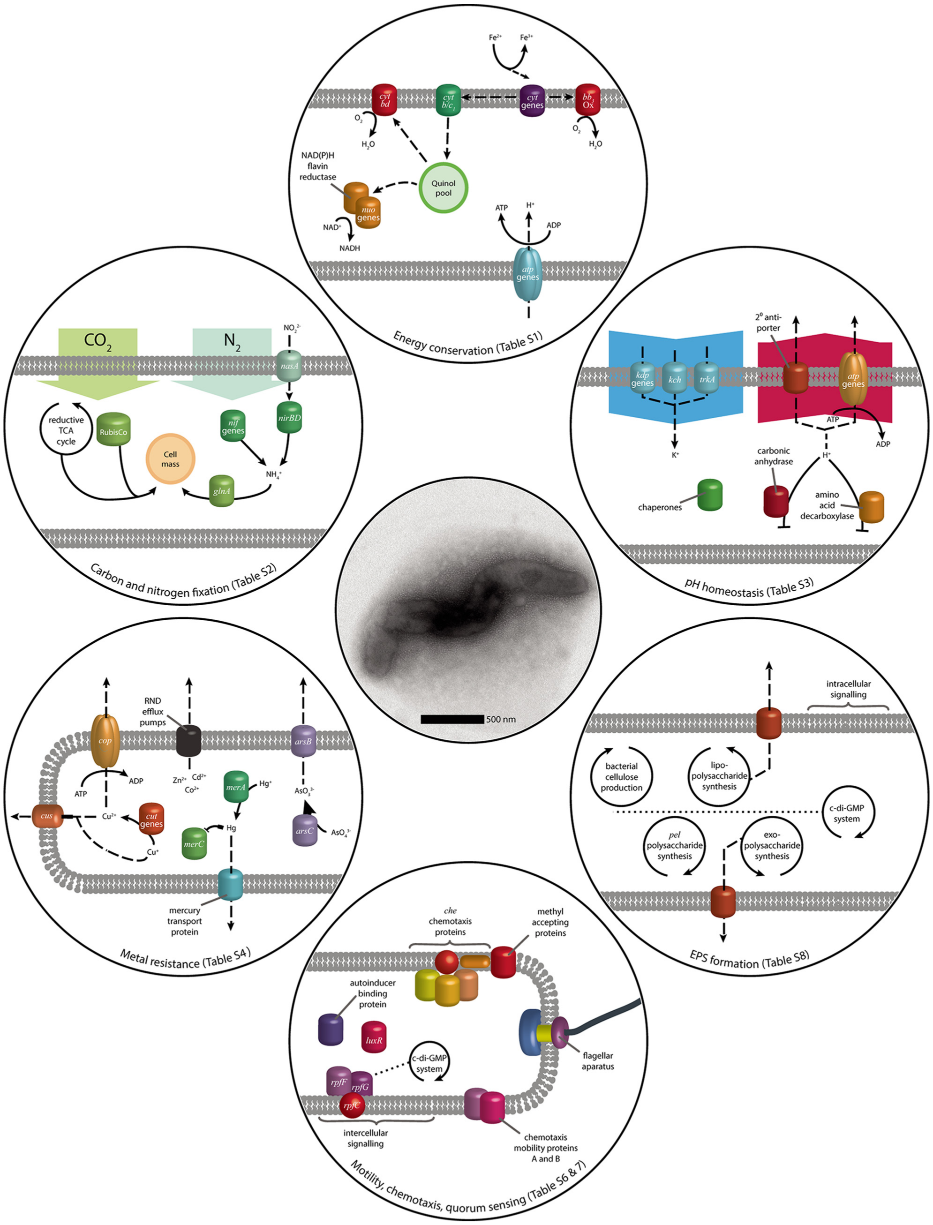


FIG 2 Model of the genomic potential observed in the *L. ferriphilum*^{T} genome, focusing on functions relevant in acidic environments and its application in biomining (see Tables S1 to S8 in the supplemental material). Solid arrows represent metabolic reactions, while dashed arrows indicate transport, the relocation of electrons or reaction products, and general regulative and metabolic interactions.

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

TABLE 3 Overview of samples and corresponding transcriptomics and proteomics data[a]

Sample	Culture type(s)	Total no. of RNA-seq read counts	Median no. of RNA-seq counts	No. of proteins identified	No. of proteins with LFQ of >0	Median LFQ
LNU-LXX9-Si00-CnA-P-B1	Continuous	1,034,434	181	NA	NA	NA
LNU-LXX9-Si00-CnA-P-B2	Continuous	NA	NA	1,698	1,241	160,595,000
LNU-LXX9-Si00-CnA-P-B3	Continuous	NA	NA	1,698	1,509	233,755,000
LNU-LXX9-Si00-CnA-P-B5	Continuous	NA	NA	1,698	1,409	221,875,000
LNU-LXX9-Si00-CnA-P-B6	Continuous	1,284,834	219	1,698	1,412	212,595,000
LNU-LXX9-Si00-CnA-P-B7	Continuous	1,477,391	256	1,698	1,465	217,165,000
LNU-LXX9-Si00-14B-P	Batch, mineral	10,967,703	1,937	763	432	3,135,800
LNU-LXX9-Si00-14C-P	Batch, mineral	12,842,605	2,099	763	513	3,645,200
LNU-LXX9-Si00-14D-P	Batch, mineral	NA	NA	763	609	3,722,500

[a] NA, not applicable.

Energy conservation. As established above, the energy needs of *L. ferriphilum* are met exclusively by the oxidation of Fe^{2+} (Fig. 2; see also Table S1 in the supplemental material). Analogous to the iron oxidation system reported previously for *L. ferriphilum* ML-04, electrons from *L. ferriphilum*$^{\mathrm{T}}$ Fe^{2+} oxidation are transferred to electron carriers (26), which were present on the genome in the form of cytochrome *c*, cytochrome $c_{551/552}$, cytochrome c_{553}, and cytochrome c_{544}. Thereafter, cytochrome cbb_3 oxidase can be used to directly reduce oxygen as a terminal electron acceptor (27). Alternatively, electrons can be used in reverse electron transport from cytochrome *c* to the quinone pool by the cytochrome b/c_1 complex. The resulting quinols can then be used to generate reducing power in the form of NAD(P)H via the NADH-quinone oxidoreductase (*nuoABCDEFHIJKLMN*) (Table S1) or the NAD(P)H-flavin reductase. Although their functionality is unknown, there are also three copies of subunit 5 of NAD(P)H-quinone oxidoreductase (*ndhF*), with which quinols could be used to produce NAD(P)H (28). Finally, electrons from the quinol pool can be transferred to oxygen by using the cytochrome *bd* complex (28), which was also described for ML-04. Proton motive force generated by iron oxidation can be used for ATP generation by an F_oF_1-type ATP synthase (*atpABCDEFGH*) (Table S1). RNA transcript counts of the genes involved in energy conservation indicated a preference for cytochrome $c_{551/552}$ (639 ± 26 TPM). However, this difference was not observed for the protein concentration. While several genes of all cytochrome groups were only marginally transcribed and translated, no clear trend in the usage of cytochromes as initial electron carriers was apparent (Data Set S1). Further electron transport was likely carried out via cbb_3 cytochromes to oxygen to create a membrane potential for the production of ATP. Although proteins of the competing reverse electron transport chain were expressed, with few exceptions, the pathway utilizing cytochrome cbb_3 had higher transcript counts and protein concentrations than did the pathway utilizing the cytochrome b/c_1 complex and the following quinone pool oxidoreductases (Data Set S1).

Carbon dioxide fixation. A single copy of the large-chain subunit of ribulose bisphosphate carboxylase (RubisCO) was encoded on the *L. ferriphilum*$^{\mathrm{T}}$ genome as well as on the genomes of other *L. ferriphilum* strains. However, all *L. ferriphilum* strains are suggested to fix carbon via the reductive tricarboxylic acid (TCA) cycle (29), for which all necessary genes were present on the genome (see Table S2 in the supplemental material). This was largely confirmed by transcript and proteome data, as gene products of the reductive TCA cycle were expressed and translated to a high extent (Data Set S1). Although RubisCO (276 ± 14 TPM; LFQ, 27,738 ± 258) exhibited low transcript counts, its protein concentration was comparable to the concentrations of proteins constituting the enzymes of the reductive TCA cycle. However, any role of RubisCO in *L. ferriphilum*$^{\mathrm{T}}$ is unknown.

Nitrogen fixation. The nitrogen demand of *L. ferriphilum*$^{\mathrm{T}}$ can be fulfilled by the fixation of elemental nitrogen by the nitrogenase complex *nifABDEHKNTUXZ* (30) and accessory protein genes (see Table S2 in the supplemental material). While present in *L. ferrooxidans* C2-3 (31) and *L. ferriphilum* strains Sp-Cl and YSK, this gene cluster was not found in the reported *L. ferriphilum*$^{\mathrm{T}}$ draft genome and likewise is lacking in the

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

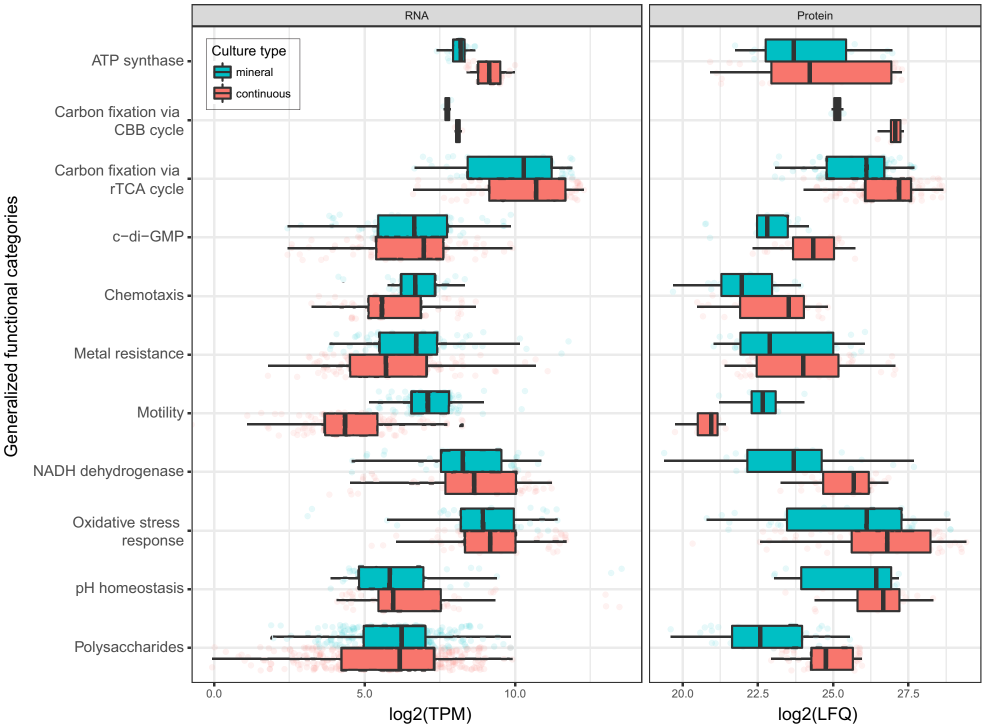


FIG 3 Overview of gene expression values for RNA-seq (TPM) (left) and proteomics (LFQ) (right) with samples grouped according to culture type. The data represent averages of expression values for genes assigned to selected functional categories (based on data in Data Sets S1 to S3 in the supplemental material), with some categories merged to aid comprehension: nitrogen metabolism (ammonia and glutamate conversion to glutamine, nitrate/nitrite regulation, nitrite uptake and assimilation to ammonia, and nitrogenase genes), metal resistance (resistance to arsenic, cadmium/cobalt/zinc, copper, copper/silver, and mercury plus general metal tolerance), polysaccharides (cellulose production, extracellular polysaccharide production and export, and lipopolysaccharide synthesis), c-di-GMP (c-di-GMP effector proteins, with the EAL domain, proteins with the GGDEF domain, and proteins with both the EAL and GGDEF domains), and pH homeostasis (proton-consuming reactions, proton transporters, and role of potassium in internal positive membrane potential). Note that the average translation values depicted are based on various proteins identified in the corresponding samples. Therefore, assumptions about the up- and downregulation of functional categories of proteins have to be made with caution and only in combination with data from differential translation analysis (Fig. 4 and Data Sets S2 and S3). Abbreviation: CBB, Calvin-Benson-Bassham.

complete genome sequence of *L. ferriphilum* ML-04. Regulatory capabilities for the gene cluster are suggested to be fulfilled by a *nif*-specific regulatory protein in *L. ferriphilum*^{T}. Additionally, nitrogen can be taken up as nitrite by the nitrate/nitrite transporter *nasA* and assimilated in the form of ammonia by the nitrite reductase *nirBD* (32), controlled by regulators of the NtrC and LysR families. RNA transcript analysis of nitrogenase subunits revealed negligible counts, and most of the corresponding proteins were also not detected in the proteomic analysis (Data Set S1). As the growth medium in this study was rich in ammonium, which can be taken up by the highly expressed glutamine synthetase, this was not surprising and has been reported for *L. ferrooxidans* (33). The highest transcript count within the nitrogen fixation clusters was that for *nifU* (1,997 ± 268 TPM; LFQ, 1,228 ± 58), which is essential for the activation of the nitrogenase complex and is localized together with the cysteine desulfurase gene *nifS* (34). NifS showed the highest protein concentration (661 ± 68 TPM; LFQ, 4,267 ± 175) in the nitrogen fixation clusters despite intermediate transcript counts. In combination with the high expression level of *nifU*, this could indicate an onset of

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

nitrogenase formation due to early-stage ammonium starvation, supported by the intermediate expression of several nitrogen assimilation regulation proteins (Data Set S1).

pH homeostasis mechanisms. Acidophiles maintain a near-neutral cytoplasmic pH by several methods, including proton efflux via 1° transport pumps in the electron transport chain (35), and this is discussed in "Energy conservation," above (see Table S3 in the supplemental material). A second method to maintain pH homeostasis is the inside positive membrane potential that repels the influx of protons (reviewed in reference 6). The internal positive membrane potential is suggested to be formed by K^+ ions, and this is supported by K^+-deficient medium inducing acid shock in *Sulfolobus acidocaldarius* (36). The *L. ferriphilum*^{T} genome has two copies of the *kdpD*-encoded sensor protein along with the *kdpABC*-encoded K^+-transporting system as well as two voltage-gated potassium channel genes (*kch* and *trkA*). The Kdp system and the TrkA voltage-gated potassium channel, but not the *kch* gene, were identified on the *L. ferriphilum* ML-04 genome. In addition, the *L. ferriphilum*^{T} genome has several 2° proton pumps, such as cation/H^+ antiporters, and similar antiporter systems were also present on the ML-04 genome. Protons can also be consumed in chemical reactions, such as amino acid decarboxylases in both neutrophiles (37) and acidophiles (35). Three amino acid decarboxylases were identified on the *L. ferriphilum*^{T} genome, while only glutamate and arginine decarboxylases were present on the *L. ferriphilum* ML-04 genome. A further proton-consuming reaction encoded in the *L. ferriphilum*^{T} genome is that of carbonic anhydrase, which has been demonstrated to aid in pH homeostasis (38). A fourth method of pH homeostasis is the production of spermidines that, among other functions (e.g., see "Oxidative stress management," below), reduce membrane permeability to protons (39), and three genes related to spermidine production were present on the *L. ferriphilum*^{T} genome. Finally, members of the general stress response protect against acid stress (40), including GroEL, ClpBC, Clp protease, and DnaK. Several of these chaperones were previously identified on the *L. ferriphilum* ML-04 as well as the *Leptospirillum* sp. group II strain CF-1 (41) genomes. With the exception of those with additional functions, few of the predicted pH homeostasis genes or proteins had high TPM values or protein levels, respectively (Data Set S1). For instance, the *kdpABC* potassium-transporting genes had TPM values of $\leq 62 \pm 13$ and LFQ values of $\leq 6 \pm 2$, while the general stress proteins DnaK and GroEL had LFQ values of $6{,}149 \pm 153$ and $68{,}468 \pm 6{,}961$, respectively. This suggested that the growth pH of 1.4 did not impose a high level of acid stress on *L. ferriphilum*^{T}.

Metal resistance systems. *L. ferriphilum*^{T} is often exposed to high metal concentrations, and the genome contains the *arsRBC* genes coding for the negative regulator, the arsenite efflux pump, and arsenate reductase, respectively. These genes are present in many acidophiles (reviewed in reference 13), such as *L. ferriphilum* ML-04 and *L. ferriphilum* strain Fairview (42). Separate Cu^{2+} and Cu^+ resistance systems were harbored on the *L. ferriphilum*^{T} genome, including the copper resistance gene *cop*. This gene can be divided into two functional groups: multicopper oxidases and P-type ATPases used to export copper ions (43). The *L. ferriphilum*^{T} *cop* gene aligned most closely with sequences of species with confirmed ATPase *cop* activity (data not shown), indicating a similar function in *L. ferriphilum*^{T}. In addition, *cut* is present in the genome for Cu^+ oxidase as well as the *cusABCF* Cu/Ag system. Cus-like metal resistance systems are part of the *Acidithiobacillus ferrivorans* SS3 mobilome that is thought to reflect selective pressure by the presence of heavy metals (44). A total of 13 open reading frames (ORFs) were determined to encode the RND family $Cd^{2+}/Co^{2+}/Zn^{2+}$ CzcABCD resistance complex and associated proteins that were also present on the *L. ferriphilum* ML-04 genome (see Table S4 in the supplemental material). Finally, the mercury resistance *merRAC* genes and a mercury transport protein were identified on the *L. ferriphilum*^{T} and *L. ferriphilum* ML-04 genomes. In continuous culture, *L. ferriphilum*^{T} was grown with only trace concentrations of metals required for cellular metabolism, and consequently, their metal resistance systems were not highly expressed (Data Set S1).

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

In several cases, low metal resistance TPM values were not reflected in concurrent protein production, and this may be due to *L. ferriphilum*^{T} maintaining a readiness to protect cells against heavy metals. For instance, the expression of the *arsR* negative regulator inhibits the expression of the arsenic resistance operon in many species, including the acidophilic archaeon "*Ferroplasma acidarmanus*" Fer1 (45).

Oxidative stress management. Oxidative stress management is crucial for all aerobic organisms. ROS are generated (i) intracellularly via molecular oxygen reactions with metal ions (46); (ii) in the extracellular, acidic, and metal-rich environment of aerobic mineral-oxidizing acidophiles (47); and (iii) on surfaces of metal sulfide minerals (15, 16). Several genes associated with oxidative stress management and ROS degradation were identified in the presented genome sequence (although homologs of catalases and superoxide dismutases were not found [see Table S5 in the supplemental material]). Among these genes, several peroxiredoxins were identified, including genes encoding alkyl hydroperoxide reductase subunit C (*ahpC*), peroxiredoxins (*ccmG* and *dsbE*), and a putative iron-dependent peroxidase (*efeB*). Several thioredoxins, a thioredoxin reductase (*trxB*), and glutaredoxins were also identified. Further genes encoding proteins involved in peroxide degradation are those encoding rubrerythrin and the periplasmic cytochrome *c* peroxidase (48). Furthermore, in the context of an *ahpC* gene and the latter gene, a gene homologous to the transcriptional regulator *perR* was found. Genes involved in the repair or degradation of oxidatively damaged proteins were also identified. ROS degradation and oxidative stress management are also complemented by protective mechanisms such as the production of antioxidants, including spermidine, ectoine, and cobalamin (49, 50). The regulation of metal homeostasis is also involved in the oxidative stress response, and the Fe^{3+} uptake regulator (Fur) family transcriptional regulator and peroxide stress response regulator proteins were identified. Genes for these functions are highly expressed in transcriptomes and proteomes of iron-grown cells (Data Set S1). Expression levels of *ahpC* and the genes encoding further peroxiredoxins, thioredoxin, glutaredoxins, cytochrome *c* peroxidase, and rubrerythrin were especially prominent (Data Set S1). Furthermore, it was also suggested that biofilm formation plus diverse mechanisms of extracellular polysaccharide production and secretion are also part of the *L. ferriphilum*^{T} ROS defense strategy in a manner similar to that of the *Acidithiobacillus ferrooxidans* type strain (51), which may be especially relevant during growth on metal sulfide minerals such as chalcopyrite and pyrite.

Chemotaxis and motility. Among the mineral dissolution-catalyzing bacteria, *Leptospirillum* spp. colonize metal sulfide surfaces more efficiently than do the *Acidithiobacilli* (52, 53), and they often comprise a substantial fraction of the community in acid mine drainage streamer biofilms (54, 55). Attached cells on solid metal sulfides are considered to enhance the oxidation of the mineral that serves as an energy source and substratum for mineral-oxidizing bacteria (56, 57). The regulation of biofilm formation involves chemotaxis and motility (see Table S6 in the supplemental material), intracellular signaling via c-di-GMP and intercellular quorum sensing (Table S7), and the production of extracellular polymeric substances (EPSs) (Table S8).

All genes involved in the assembly of a functional flagellar apparatus and its controlling chemotaxis system were identified in the presented genome sequence (Table S6). The *L. ferriphilum*^{T} chemotaxis system is composed of seven methyl-accepting chemotaxis proteins (MCPs) involved in sensing environmental signals. These genes are scattered across the chromosome, except for LFTS_00227, which was found in the context of the chemotaxis gene cluster. Of the MCPs, only LFTS_01731 was significantly expressed at both the RNA and protein levels in iron-grown chemostat cells. ORFs encoding the flagellar motor switch proteins FliN and FliM were found in different regions on the chromosome. Except for the *fliM* gene, all genes relevant for a functional flagellar motility system were found in two large gene clusters (Table S6). All these genes were found at very low expression levels by using RNA transcript analysis, and either the corresponding protein levels were low or proteins were not detected

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

(Data Set S1), indicating that motility and chemotaxis are of no relevance in a well-mixed, homogenous environment such as a chemostat reactor.

Quorum sensing and c-di-GMP. In Gram-negative bacteria, the regulation of genes encoding proteins for chemotaxis, motility, EPS production, and biofilm formation is often controlled by intracellular levels of the messenger molecule c-di-GMP (58). The presented genome sequence provides evidence for complex c-di-GMP metabolism, as is common for many Gram-negative bacteria, including acidophilic mineral-oxidizing *Acidithiobacillus* spp. (59, 60). The *L. ferriphilum*^{T} genome contains 10 genes annotated as encoding putative diguanylate cyclases, 13 genes encoding both diguanylate cyclase- and c-di-GMP phosphodiesterase-specific GGDEF and EAL protein domains, and two c-di-GMP-specific phosphodiesterases (see Table S7 in the supplemental material). Furthermore, four genes encoding HD/HDc domain-containing proteins and three genes encoding PilZ domain-containing c-di-GMP effector proteins were found. The latter genes were found in the context of genes annotated as being related to functions such as cellulose and extracellular polysaccharide biosynthesis and export. This suggests that c-di-GMP metabolism in *L. ferriphilum*^{T} also has an important function in the regulation of EPS production and biofilm formation. Several of these genes were expressed at the RNA and protein levels, including a c-di-GMP-specific phosphodiesterase class I-encoding gene, bifunctional diguanylate cyclase/c-di-GMP-specific phosphodiesterase-encoding genes, diguanylate cyclases, and a PilZ domain-containing protein.

Interestingly, the *L. ferriphilum*^{T} genome contains a gene cluster harboring an *rpf* diffusible signal factor quorum sensing system, which is composed of the diffusible signal factor synthase-encoding gene *rpfF*, two genes encoding *rpfC* homologs annotated as genes encoding the Hpt domain-containing protein and signal transduction kinase, and the respective two-component system response regulator-encoding gene *rpfG*. In addition, further genes related to quorum sensing signaling were identified, such as three *luxR* family transcriptional regulator protein-encoding genes and another autoinducer binding domain-containing gene. The genes encoding the *rpf* quorum sensing system were found to be expressed at enhanced levels, while the orphan LuxR protein-encoding genes were found at very low RNA transcript or protein levels (Data Set S1).

Biofilm formation. A total of 103 genes were annotated with functions related to sugar processing, polysaccharide biosynthesis, and export and may be involved in the synthesis of polysaccharides as a constituent of EPSs (see Table S8 in the supplemental material). Among these genes, 47 represent or were in the context of genes primarily associated with lipopolysaccharide synthesis. Several of these genes were found to be expressed in iron-grown chemostat cells (Data Set S1). Interestingly, two gene clusters contain bacterial cellulose synthesis genes, and one of them also contains a gene encoding a cellulase family 8 protein, suggesting that aside from a structural component of EPSs and/or cell walls, cellulose may be used as an intracellular sugar storage compound in *L. ferriphilum*^{T}. Furthermore, a gene cluster highly similar to the *Pseudomonas aeruginosa pel* operon was found. This cluster is responsible for Pel polysaccharide production as part of its EPS constituents. Further genes associated with extracellular polysaccharide export and biosynthesis were found in a large cluster. Among these ORFs, poly-β-1,6-*N*-acetyl-D-glucosamine (PGA) synthesis and export protein-encoding genes were found directly next to a *wza* polysaccharide outer membrane export protein-encoding gene plus further genes with functions associated with polysaccharide assembly and export. In addition, 12 of the ORFs in this cluster were determined to encode glycosyltransferases. However, the majority of these genes were found at very low transcript levels, while the corresponding proteins were not detected in the chemostat (Data Set S1). An exception to the low RNA transcript levels but high protein levels was observed for a UTP-glucose-1-phosphate uridylyltransferase. The *rfbBAC* genes were found in one gene cluster close to an *algK* homolog that is a recently described outer membrane secretin that differs from canonical bacterial capsular polysac-

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

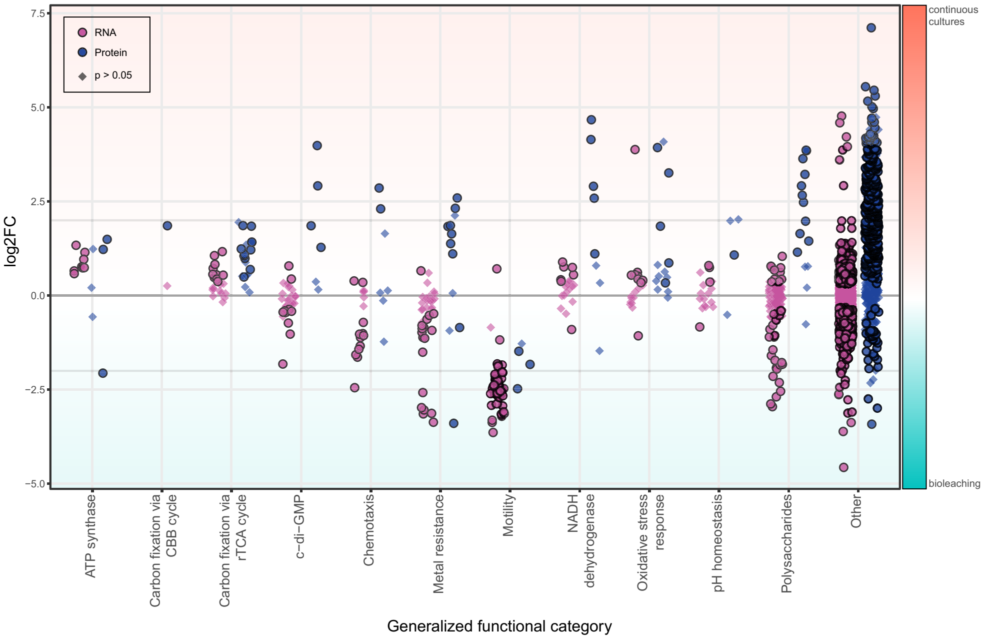


FIG 4 Overview of differential expression ($\log_2$-fold change) of RNA transcripts and protein concentrations between continuous culture and chalcopyrite-containing bioleaching cultures. Data points represent single transcripts or protein signals, categorized as explained in the legend to Fig. 3. Circular symbols denote statistically significant differences ($P < 0.05$), while diamonds indicate statistically insignificant data.

charide secretion systems (61) and a *wzzE* polysaccharide chain length modulation protein. Furthermore, ORFs were determined to encode undecaprenyl (UDP)-galactose-4-epimerases, UDP-galactopyranose mutase, UDP-glucuronate-4-epimerase, and UDP-*N*-acetyl-D-glucosamine dehydrogenase (*wbpA*).

Biomining lifestyle. Bioleaching experiments using pure cultures of *L. ferriphilum*^{T} achieved a significant dissolution of chalcopyrite (see Fig. S3 in the supplemental material). The isolation of nucleic acids and proteins proved to be challenging, and only two RNA extracts and three protein samples of mineral origins were of sufficient quality for differential expression and translation analyses (Fig. 4 and Data Sets S2 and S3). Owing to the lower sensitivity and dynamic range of the Orbitrap Elite instrument, fewer low-abundance proteins were quantified in the bioleaching samples than in the continuous-culture samples. This manifested as an apparently higher expression level of such gene products in continuous cultures. Therefore, more studies will have to be conducted to confirm the data presented below. To investigate important features and adaptation strategies of *L. ferriphilum*^{T}, RNA transcripts and proteins were grouped based on the functional categories established as described above (Fig. 2 and 3). Comparison of continuous versus mineral culture samples revealed unexpectedly few differences in expression and translation patterns. In part, this is probably related to the controlled nature of the bioleaching experiments, where, e.g., the initial pH was 1.8 and did not decrease below 1.7 (data not shown), such that pH homeostasis systems seemed unaffected by the presence of chalcopyrite. Longer retention times and the presence of sulfur oxidizers would cause the pH to drop significantly (19, 62). Despite the remarkable tolerance of *L. ferriphilum*^{T} to high proton concentrations (63), this would likely cause additional stress. Similarly, RNA transcript levels and protein con-

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

centrations for genes related to nitrogen fixation were found to be stable under the two conditions, conceivably as the culture medium contained large amounts of biologically available ammonium.

Among the differences observed between continuous and bioleaching cultures were decreased transcript counts related to ATP synthesis in the mineral samples along with bidirectional alterations of protein concentrations in ATP synthesis (Fig. 4) and of specific cytochromes and cytochrome oxidases (Data Set S1). This possibly indicated a shift of electron transport away from proton motive force and ATP generation toward the production of reducing power in the form of NAD(P)H (Fig. 2). However, this was not observable in NADH dehydrogenase RNA transcript counts. In contrast, the protein concentration related to NADH production was decreased in the bioleaching experiments (Fig. 4). Additionally, RNA and protein analysis revealed slight reductions in the levels of proteins involved in both above-mentioned carbon fixation pathways when cells were grown on chalcopyrite (Fig. 4). While the exact reasons for this are unknown, it could indicate a reduced demand for organic carbon, possibly caused by overall slow growth along with a reallocation of efforts for cell maintenance under stress conditions in mineral batch cultures compared to active growth in continuous cultures.

Growth on minerals naturally comes with a heightened exposure of cells to heavy metals. Overall, transcript counts derived from metal resistance genes showed significantly increased levels during growth in chalcopyrite bioleaching cultures, in particular a strong enhancement of counts mapping to copper resistance systems (Fig. 4 and Data Set S2). Surprisingly, protein concentrations appeared to be decreased, with two exceptions. In-depth analysis revealed severely elevated amounts of proteins belonging to the *cus* copper efflux system (Fig. 4 and Data Set S3), underlining the strong detrimental effects of copper ions on microbes (64). Similar to the pH homeostasis response, as metal concentrations increase with time in natural or industrial systems, further upregulation of these systems should be expected.

Damage caused by heavy metal ions can often be mitigated by oxidative damage repair systems (65). The majority of these genes were found to exhibit the same or even fewer transcript counts and protein levels in mineral-grown cells (Fig. 4 and Data Set S3). This was surprising, as they have been suggested to combat the heightened oxidative damage caused by ROS produced at the mineral surface (15, 16). An explanation for this behavior could be that the combined effect of high Fe^{3+} concentrations and excessive sparging with air in continuous culture induced more oxidative damage than ROS produced on mineral surfaces.

L. ferriphilum^T was previously reported to rapidly attach to mineral surfaces (52), and RNA transcript counts of both chemotaxis and motility systems were revealed to be heavily enhanced during the bioleaching experiments. This was also observed for motility protein concentrations but not chemotaxis protein concentrations (Fig. 4 and Data Set S2). The transcription and translation of c-di-GMP and EPS production remained at the same or lower levels in mineral culture samples (Fig. 4). However, this may be explained by the fact that sampling of mineral-grown cells was conducted on the slowly agitated overlying medium and not the biofilm on the mineral grains, where most of the biofilm regulation and EPS production are expected to occur (56, 66). In contrast, samples taken from the continuous culture were well mixed and likely contained both planktonic and detached biofilm cells.

Conclusions. The newly sequenced genome of *L. ferriphilum*^{T} allows an in-depth and complete characterization of this organism's metabolic potential as well as its expression and translation behaviors in continuous culture and batch bioleaching experiments. PacBio single-molecule real-time (SMRT) long-read sequencing allowed the assembly of a circular chromosome and revealed key features of the adaptation of *L. ferriphilum*^{T} to acidic, metal-rich environments associated with sulfidic minerals, in the environment as well as in industrial applications. Additionally, RNA transcript sequencing and protein identification elucidated stressing factors during chalcopyrite biomining and shed light on resistance systems deployed by *L. ferriphilum*^{T}. The data

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

provided by this study pose a valuable resource for future experiments investigating the role of *L. ferriphilum*^{T} in acid mine and rock drainage as well as bioleaching processes.

MATERIALS AND METHODS

Batch cell culture and DNA extraction. The *Leptospirillum ferriphilum* type strain (ATCC 49881 and DSM 14647) was cultured aerobically at 37°C at pH 1.5 to 1.6 in MAC medium (67) containing 100 mM sterile-filtered (0.2-μm filter) Fe^{2+} as the electron donor. Cells were grown to late log phase before harvesting at 10,000 × *g* for 10 min. DNA for sequencing was isolated by using the Genomic-tip 100/G extraction kit (Qiagen) according to the manufacturer's instructions, with the exception of a customized purification step recommended by the sequencing facility. Briefly, eluted genomic DNA was precipitated by the addition of isopropanol, immediately spooled by using a sterile pipette tip, and transferred to a microcentrifuge tube containing 70% (vol/vol) ethanol for 2 min. Spooled DNA was then air dried, finally resuspended in 200 μl 0.1× Tris-EDTA (TE) buffer (pH 8), and allowed to dissolve for 72 h at room temperature.

Continuous cultivation, bioleaching experiments, and RNA and protein isolation. *L. ferriphilum*^{T} was grown in a substrate-limited, 1-liter-working-volume, continuous-culture vessel at 37°C. The electron donor was provided in the form of MAC medium containing sterile-filtered 100 mM Fe^{2+} (dilution rate [*D*] = 0.3 liters/day). The pH of the medium was adjusted to pH 1.1 by the addition of sulfuric acid that maintained a constant pH of 1.4 within the culture. For the collection of RNA and protein, replicate 100-ml samples were taken from the cultures at least 3 days apart. To minimize RNA degradation, samples were rapidly cooled by mixing with 1 volume of ice-cold sterile MAC medium, and cells were immediately pelleted by centrifugation at 4°C at 12,000 × *g* for 15 min. The cells were then washed in 40 ml fresh, ice-cold MAC medium before being centrifuged again. Finally, pellets were flash-frozen in liquid nitrogen and stored at −80°C for the extraction step.

Additionally, *L. ferriphilum*^{T} was cultured in four bioleaching flasks containing 100 ml MAC medium (pH 1.8) supplemented with 2% (wt/vol) copper mineral chalcopyrite ($CuFeS_2$) as the only energy source. Mixtures for the bioleaching experiments were incubated for 14 days under slow shaking (120 rpm). Seventy-five milliliters of the overlying medium was taken as a sample and processed as described above.

Cell pellets were subjected to biomolecular extractions based on a previously reported protocol (68), skipping the metabolite extraction step. In short, cell pellets were lysed by cryo-milling and bead beating followed by the column-based isolation of biomolecules with the Qiagen Allprep kit. Quality control was performed with SDS-PAGE (protein) and measurements on an Agilent bioanalyzer (total RNA).

DNA sequencing and genomic analysis. The obtained genomic DNA was sent to the Science for Life Laboratory (Stockholm, Sweden) and sequenced by using two PacBio SMRT cells. Assembly was conducted with HGAP3 at the sequencing facility, including quiver for consensus corrections. The large contig was circularized with Circlator (69) after inspection of dot plots produced with Gepard (70). The −fixstart option was applied to set the *dnaA* gene as the first gene. Prokka v1.12-beta (71) was used for genome annotation, which included Prodigal v2.6.3 (72) for the prediction of protein-encoding sequences. Functional annotation of coding sequences (CDSs) was performed with a custom genus database of related genomes downloaded from the Integrated Microbial Genomes (IMG) system (73): *Leptospirillum* sp. group IV UBA BS (GOLD identification Ga0053748 [https://gold.jgi.doe.gov/analysis_projects?id=Ga0053748]), "*Leptospirillum* sp. group II C75" (GOLD identification Ga0039193 [https://gold.jgi.doe.gov/analysis_projects?id=Ga0039193]), *L. ferrooxidans* C2-3 (NCBI RefSeq accession number AP012342), *L. ferriphilum* ML-04 (NCBI RefSeq accession number CP002919), *L. ferriphilum* DSM 14647 (GOLD identification Ga0059175 [https://gold.jgi.doe.gov/analysis_projects?id=Ga0059175]), and *L. ferriphilum* YSK (NCBI RefSeq accession number CP007243). Protein sequences were searched (blastp) in Prokka against this genus database, and annotations of best-matching hits were transferred with an E value cutoff of 1e−9. Additionally, the standard databases in Prokka were searched. For additional information, in-house hidden Markov model (HMM) databases were searched, including KEGG orthologous groups (KOs), PFAM, TIGRFAM, UniProt-enzymes, and MetaCyc (additional details are available in reference 74). Furthermore, the annotation tool Pannzer (75) was applied. Additional annotations are listed in Data Set S1 in the supplemental material. Functional categories were assigned based on the KO annotation (COG, KEGG class). Additionally, genes were grouped into functional categories by manual assignment.

RNA sequencing and transcript analysis. RNA samples were adjusted for equimolar concentrations and sent to the Science for Life Laboratory (Stockholm, Sweden). Library preparation was performed with the Illumina TruSeq Stranded total RNA kit. Paired-end sequencing was performed on one HiSeq2500 lane for a total of five *L. ferriphilum*^{T} samples, three from continuous cultures and two successful extracts from batch cultures with chalcopyrite. Batch culture samples were depleted of rRNA with the bacterial Ribo-Zero rRNA removal kit (Illumina) prior to library preparation.

A custom pipeline was written in snakemake (76) for processing and analysis of the transcriptome sequencing (RNA-seq) data (the source code is available at https://git-r3lab.uni.lu/malte.herold/LF_omics_analysis). Raw reads for RNA sequencing were preprocessed with Trimmomatic v0.36 (77) with the file TruSeq3-PE.fa for adapters. Preprocessed reads were mapped onto a concatenation of reference genomes of three acidophiles, including the newly assembled chromosome of *L. ferriphilum*^{T}, with bowtie2 v2.3.2 with default settings. Read mappings to CDSs were counted with the software featureCounts from subread package v1.5.2 (78), and the −s 2 option was used to include only reads on the correct strand. Raw read counts were normalized to the gene length and the sum of total counted reads. Normalized

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

counts were represented as transcripts per million base pairs (TPM). Raw counts for the CDS features of continuous and batch culture samples were subjected to differential analysis with DeSeq2 v1.16.1 (79).

Proteomics and protein identification. Five separate protein extracts from a continuous culture and three batch cultures were precipitated in acetone, dried, and then dissolved in 20 μl of 6 M urea–2 M thiourea by vortexing. The reduction of cysteine was done by incubation with 1 μl 1 M dithiothreitol for 30 min at room temperature. Cysteines were alkylated with 1 μl 550 mM iodoacetamide for 20 min in the dark. Proteins were then digested with lysyl endopeptidase (Wako) at a 1:100 protease/protein ratio at room temperature for 3 h. Upon the dilution of urea to 2 M with 50 mM ammonium bicarbonate, further digestion occurred with trypsin (sequencing grade; Promega) at a protease/protein ratio of 1:100 at room temperature for 12 h. Peptides were extracted from the gel pieces with acetonitrile, loaded onto stop-and-go extraction (STAGE) tips for storage, and eluted from the tips shortly before mass spectrometry (MS) analysis (80).

Mass spectrometry for continuous-culture samples was carried out by using an EASY-nLC 1000 liquid chromatography (LC) system (Thermo Scientific) and a Q-Exactive HF mass spectrometer (Thermo Scientific), as described previously (81). Mass spectra were recorded with Xcalibur software 3.1.66.10 (Thermo Scientific). Mass spectrometry for mineral culture samples was carried out by using a nanoACQUITY gradient ultraperformance liquid chromatography (UPLC) pump system (Waters, Milford, MA, USA) coupled to an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). An UPLC HSS T3 M-class column (1.8 μm, 75 μm by 150 mm; Waters, Milford, MA, USA) and an UPLC Symmetry C_{18} trapping column (5 μm, 180 μm by 20 mm; Waters, Milford, MA, USA) were used for LC in combination with a PicoTip emitter (SilicaTip, 10-μm internal diameter [i.d.]; New Objective, Woburn, MA, USA). For elution of the peptides, a linear gradient with increasing concentrations of buffer B (0.1% formic acid in acetonitrile [ULC/MS grade]; Biosolve, Netherlands) from 1% to 95% within 166.5 min was applied, followed by a linear gradient from 1% acetonitrile within 13.5 min (1% buffer B from 0 to 10 min, 5% buffer B from 10 to 161 min, 40% buffer B from 161 to 161.5 min, 85% buffer B from 161.5 to 166.5 min, 95% buffer B from 166.5 to 167.1 min, and 1% buffer B from 167.1 to 180 min) at a flow rate of 400 nl min^{-1} and a spray voltage of 1.5 to 1.8 kV. The column was reequilibrated with 2% buffer B within 15 min. The analytical column oven was set to 55°C, and the heated desolvation capillary was set to 275°C. The LTQ Orbitrap Elite instrument was operated by using instrument method files of Xcalibur (Rev.2.1.0) in the positive-ion mode. The linear ion trap and Orbitrap instruments were operated in parallel; i.e., during a full MS scan on the Orbitrap instrument in the range of 150 to 2,000 *m/z* at a resolution of 60,000, tandem MS (MS/MS) spectra of the 10 most intense precursors, from the most intense to the least intense, were detected in the ion trap. The relative collision energy for rapid collision-induced dissociation (rCID) was set to 35%. Dynamic exclusion was enabled with a repeat count of 1 and a 45-s exclusion duration window. Singly charged ions and ions of an unknown charge state were rejected for MS/MS. Mass spectra were recorded with Xcalibur software 2.2 SP1.48 (Thermo Scientific).

Proteins under both culture conditions were identified with Andromeda (82) and quantified with the LFQ algorithm (25) embedded in MaxQuant version 1.5.3.175 (81). The FASTA protein database for identification was taken from the output of the functional annotation of the chromosome and contained 2,486 entries. After quantification, intensities from the LFQ normalization were filtered and compared with Perseus (v1.5.8.5) (83), removing rows with fewer than two values under either condition (mineral or continuous). The two conditions were compared with two-sample Welch's *t* test.

Data availability. DNA raw sequencing data and the resulting assembly are available under BioProject accession number PRJEB21703 and Assembly accession number GCA_900198525.1. Raw reads for transcriptome sequencing are available under BioProject accession number PRJEB21842. Links to raw data repositories, processed data files, and repositories containing the respective computational workflows are available through the fairdomhub platform (84) in a structured format (see reference 85 [https://doi.org/10.15490/fairdomhub.1.investigation.162.1]).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.02091-17.

SUPPLEMENTAL FILE 1, PDF file, 1.7 MB.
SUPPLEMENTAL FILE 2, XLSX file, 1.1 MB.
SUPPLEMENTAL FILE 3, XLSX file, 0.3 MB.
SUPPLEMENTAL FILE 4, XLSX file, 0.2 MB.

ACKNOWLEDGMENTS

This project was supported by the Vetenskapsrådet (contract 2014-6545), the Luxembourg National Research Fund (FNR) (INTER/SYSAPP/14/05), the Bundesministerium für Bildung und Forschung (BMBF) (031A600A and -B), and the Swiss Initiative in Systems Biology (SystemsX.ch, SysMetEx) under the frame of ERASysAPP. We greatly acknowledge the Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

computational infrastructure (project b2013127). Parts of the analysis presented in this paper were carried out by using the HPC facilities of the University of Luxembourg.

We thank Emilie Mueller and Anna Heintz-Buschart for helpful comments for the manuscript, Patrick May and Venkata Satagopam for providing HMM databases for annotation, and Susana Martinez-Arbas and Shaman Narayanasamy for aiding in the analysis.

REFERENCES

1. Hippe H. 2000. *Leptospirillium* gen. nov (ex Markoysan 1972), nom. rev., including *Leptospirillium ferrooxidans* sp. nov. (ex Markoysan 1972), nom. rev. and *Leptospirillium thermoferrooxidans* sp. nov. (Golovacheva et al. 1992). Int J Syst Evol Microbiol 50:501–503.
2. Goltsman DS, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS, Dick GJ, Sun CL, Wheeler KE, Zemla A, Baker BJ, Hauser L, Land M, Shah MB, Thelen MP, Hettich RL, Banfield JF. 2009. Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "*Leptospirillum rubarum*" (group II) and "*Leptospirillum ferrodiazotrophum*" (group III) bacteria in acid mine drainage biofilms. Appl Environ Microbiol 75:4599–4615. https://doi.org/10.1128/AEM.02943-08.
3. Coram NJ, Rawlings DE. 2002. Molecular relationship between two groups of the genus Leptospirillum and the finding that *Leptospirillum ferriphilum* sp. nov. dominates South African commercial biooxidation tanks that operate at 40°C. Appl Environ Microbiol 68:838–845. https://doi.org/10.1128/AEM.68.2.838-845.2002.
4. Aliaga Goltsman DS, Dasari M, Thomas BC, Shah MB, VerBerkmoes NC, Hettich RL, Banfield JF. 2013. New group in the *Leptospirillum* clade: cultivation-independent community genomics, proteomics, and transcriptomics of the new species "*Leptospirillum* group IV UBA BS." Appl Environ Microbiol 79:5384–5393. https://doi.org/10.1128/AEM.00202-13.
5. Dopson M, Johnson DB. 2012. Biodiversity, metabolism and applications of acidophilic sulfur-metabolizing microorganisms. Environ Microbiol 14:2620–2631. https://doi.org/10.1111/j.1462-2920.2012.02749.x.
6. Slonczewski JL, Fujisawa M, Dopson M, Krulwich TA. 2009. Cytoplasmic pH measurement and homeostasis in bacteria and archaea. Adv Microb Physiol 55:1–79. https://doi.org/10.1016/S0065-2911(09)05501-5.
7. Dopson M. 2012. Physiological adaptations and biotechnological applications of acidophiles. *In* Anitori RP (ed), Extremophiles: microbiology and biotechnology. Caister Academic Press, Norfolk, United Kingdom.
8. Bird LJ, Bonnefoy V, Newman DK. 2011. Bioenergetic challenges of microbial iron metabolisms. Trends Microbiol 19:330–340. https://doi.org/10.1016/j.tim.2011.05.001.
9. Esparza M, Cardenas JP, Bowien B, Jedlicki E, Holmes DS. 2010. Genes and pathways for CO_2 fixation in the obligate, chemolithoautotrophic acidophile, *Acidithiobacillus ferrooxidans*. BMC Microbiol 10:229. https://doi.org/10.1186/1471-2180-10-229.
10. Sand W, Gerke T, Hallmann R, Schippers A. 1995. Sulfur chemistry, biofilm, and the (in)direct attack mechanism—a critical evaluation of bacterial leaching. Appl Microbiol Biotechnol 43:961–966. https://doi.org/10.1007/BF00166909.
11. Vera M, Schippers A, Sand W. 2013. Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation—part A. Appl Microbiol Biotechnol 97:7529–7541. https://doi.org/10.1007/s00253-013-4954-2.
12. Dopson M, Ossandon F, Lövgren L, Holmes DS. 2014. Metal resistance or tolerance? Acidophiles confront high metal loads via both abiotic and biotic mechanisms. Front Microbiol 5:157. https://doi.org/10.3389/fmicb.2014.00157.
13. Dopson M, Holmes DS. 2014. Metal resistance in acidophilic microorganisms and its significance for biotechnologies. Appl Microbiol Biotechnol 19:8133–8144. https://doi.org/10.1007/s00253-014-5982-2.
14. Frawley ER, Fang FC. 2014. The ins and outs of bacterial iron metabolism. Mol Microbiol 93:609–616. https://doi.org/10.1111/mmi.12709.
15. Schoonen MAA, Cohn CA, Roemer E, Laffers R, Simon SR, O'Riordan T. 2006. Mineral-induced formation of reactive oxygen species. Rev Miner Geochem 64:179–221. https://doi.org/10.2138/rmg.2006.64.7.
16. Borda MJ, Elsetinow AR, Strongin DR, Schoonen MA. 2003. A mechanism for the production of hydroxyl radical at surface defect sites on pyrite. Geochim Cosmochim Acta 67:935–939. https://doi.org/10.1016/S0016-7037(02)01222-X.
17. Brierley CL, Brierley JA. 2013. Progress in bioleaching. Part B: applications of microbial processes by the minerals industries. Appl Microbiol Biotechnol 97:7543–7552. https://doi.org/10.1007/s00253-013-5095-3.
18. Acosta M, Galleguillos P, Ghorbani Y, Tapia P, Contador Y, Velásquez A, Espoz C, Pinilla C, Demergasso C. 2014. Variation in microbial community from predominantly mesophilic to thermotolerant and moderately thermophilic species in an industrial copper heap bioleaching operation. Hydrometallurgy 150:281–289. https://doi.org/10.1016/j.hydromet.2014.09.010.
19. Watling HR. 2006. The bioleaching of sulphide minerals with emphasis on copper sulphides—a review. Hydrometallurgy 84:81–108. https://doi.org/10.1016/j.hydromet.2006.05.001.
20. Yu YY, Liu XM, Wang HY, Li XT, Lin JQ. 2014. Construction and characterization of *tetH* overexpression and knockout strains of *Acidithiobacillus ferrooxidans*. J Bacteriol 196:2255–2264. https://doi.org/10.1128/JB.01472-13.
21. Cardenas JP, Quatrini R, Holmes DS. 2016. Progress in acidophile genomics, p 179–198. *In* Quatrini R, Johnson DB (ed), Acidophiles: life in extremely acidic environments. Caister Academic Press, Norfolk, United Kingdom.
22. Yelton AP, Comolli LR, Justice NB, Castelle C, Denef VJ, Thomas BC, Banfield JF. 2013. Comparative genomics in acid mine drainage biofilm communities reveals metabolic and structural differentiation of co-occurring archaea. BMC Genomics 14:485. https://doi.org/10.1186/1471-2164-14-485.
23. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:985. https://doi.org/10.7717/peerj.985.
24. Cardenas JP, Lazcano M, Ossandon FJ, Corbett M, Holmes DS, Watkin E. 2014. Draft genome sequence of the iron-oxidizing acidophile *Leptospirillum ferriphilum* type strain DSM 14647. Genome Announc 2:e01153-14. https://doi.org/10.1128/genomeA.01153-14.
25. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 13:2513–2526. https://doi.org/10.1074/mcp.M113.031591.
26. Bonnefoy V, Holmes DS. 2012. Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. Environ Microbiol 14:1597–1611. https://doi.org/10.1111/j.1462-2920.2011.02626.x.
27. Pitcher RS, Watmough NJ. 2004. The bacterial cytochrome cbb_3 oxidases. Biochim Biophys Acta 1655:388–399. https://doi.org/10.1016/j.bbabio.2003.09.017.
28. Jünemann S. 1997. Cytochrome *bd* terminal oxidase. Biochim Biophys Acta 1321:107–127. https://doi.org/10.1016/S0005-2728(97)00046-7.
29. Hugler M, Sievert SM. 2011. Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Annu Rev Mar Sci 3:261–289. https://doi.org/10.1146/annurev-marine-120709-142712.
30. Hoffman BM, Lukoyanov D, Yang ZY, Dean DR, Seefeldt LC. 2014. Mechanism of nitrogen fixation by nitrogenase: the next stage. Chem Rev 114:4041–4062. https://doi.org/10.1021/cr400641x.
31. Fujimura R, Sato Y, Nishizawa T, Oshima K, Kim SW, Hattori M, Kamijo T, Ohta H. 2012. Complete genome sequence of *Leptospirillum ferrooxidans* strain C2-3, isolated from a fresh volcanic ash deposit on the island of Miyake, Japan. J Bacteriol 194:4122–4123. https://doi.org/10.1128/JB.00696-12.
32. Bykov D, Neese F. 2015. Six-electron reduction of nitrite to ammonia by cytochrome *c* nitrite reductase: insights from density functional theory studies. Inorg Chem 54:9303–9316. https://doi.org/10.1021/acs.inorgchem.5b01506.
33. Moreno-Paz M, Parro V. 2006. Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum fer-*

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

rooxidans under nitrogen-fixing conditions. Environ Microbiol 8:1064–1073. https://doi.org/10.1111/j.1462-2920.2006.00998.x.

34. Agar JN, Yuvaniyama P, Jack RF, Cash VL, Smith AD, Dean DR, Johnson MK. 2000. Modular organization and identification of a mononuclear iron-binding site within the NifU protein. J Biol Inorg Chem 5:167–177. https://doi.org/10.1007/s007750050361.
35. Mangold S, Rao Jonna V, Dopson M. 2013. Response of *Acidithiobacillus caldus* toward suboptimal pH conditions. Extremophiles 17:689–696. https://doi.org/10.1007/s00792-013-0553-5.
36. Buetti-Dinh A, Dethlefsen O, Friedman R, Dopson M. 2016. Transcriptomic analysis reveals how potassium ions affect *Sulfolobus acidocaldarius* pH homeostasis. Microbiology 162:1422–1434. https://doi.org/10.1099/mic.0.000314.
37. Richard H, Foster JW. 2004. *Escherichia coli* glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. J Bacteriol 186:6032–6041. https://doi.org/10.1128/JB.186.18.6032-6041.2004.
38. Bury-Moné S, Mendz GL, Ball GE, Thibonnier M, Stingl K, Ecobichon C, Avé P, Huerre M, Labigne A, Thiberge JM, De Reuse H. 2008. Roles of α and β carbonic anhydrases of *Helicobacter pylori* in the urease-dependent response to acidity and in colonization of the murine gastric mucosa. Infect Immun 76:497–509. https://doi.org/10.1128/IAI.00993-07.
39. Samartzidou H, Mehrazin M, Xu Z, Benedik MJ, Delcour AH. 2003. Cadaverine inhibition of porin plays a role in cell survival at acidic pH. J Bacteriol 185:13–19. https://doi.org/10.1128/JB.185.1.13-19.2003.
40. Leverrier P, Vissers JPC, Rouault A, Boyaval P, Jan G. 2004. Mass spectrometry proteomic analysis of stress adaptation reveals both common and distinct response pathways in *Propionibacterium freudenreichii*. Arch Microbiol 181:215–230. https://doi.org/10.1007/s00203-003-0646-0.
41. Ferrer A, Bunk B, Spröer C, Biedendieck R, Valdés N, Jahn M, Jahn D, Orellana O, Levicán G. 2016. Complete genome sequence of the bioleaching bacterium *Leptospirillum* sp. group II strain CF-1. J Biotechnol 222:21–22. https://doi.org/10.1016/j.jbiotec.2016.02.008.
42. Kotze AA, Tuffin IM, Deane SM, Rawlings DE. 2006. Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacillus caldus* and enhanced arsenic resistance on conjugal transfer of *ars* genes located on transposon *TnAtcArs*. Microbiology 152:3551–3560. https://doi.org/10.1099/mic.0.29247-0.
43. Li X, Zhu Y-G, Shaban B, Bruxner TJC, Bond PL, Huang L. 2015. Assessing the genetic diversity of Cu resistance in mine tailings through high-throughput recovery of full-length *copA* genes. Sci Rep 5:13258. https://doi.org/10.1038/srep13258.
44. González C, Yanquepe M, Cardenas JP, Valdes J, Quatrini R, Holmes DS, Dopson M, Gonz C, Cardenas JP, Valdes J, Quatrini R, Holmes DS, Dopson M, González C, Yanquepe M, Cardenas JP, Valdes J, Quatrini R, Holmes DS, Dopson M. 2014. Genetic variability of psychrotolerant *Acidithiobacillus ferrivorans* revealed by (meta)genomic analysis. Res Microbiol 165:726–734. https://doi.org/10.1016/j.resmic.2014.08.005.
45. Baker-Austin C, Dopson M, Wexler M, Sawers RG, Stemmler A, Rosen BP, Bond PL. 2007. Extreme arsenic resistance by the acidophilic archaeon "*Ferroplasma acidarmanus*" Fer1. Extremophiles 11:425–434. https://doi.org/10.1007/s00792-006-0052-z.
46. Imlay JA. 2003. Pathways of oxidative damage. Annu Rev Microbiol 57:395–418. https://doi.org/10.1146/annurev.micro.57.030502.090938.
47. Cohn CA, Mueller S, Wimmer E, Leifer N, Greenbaum S, Strongin DR, Schoonen MA. 2006. Pyrite-induced hydroxyl radical formation and its effect on nucleic acids. Geochem Trans 7:3. https://doi.org/10.1186/1467-4866-7-3.
48. Zapata C, Paillavil B, Chávez R, Álamos P, Levicán G. 2017. Cytochrome *c* peroxidase (CcP) is a molecular determinant of the oxidative stress response in the extreme acidophilic *Leptospirillum* sp. CF-1. FEMS Microbiol Ecol 93:fix001. https://doi.org/10.1093/femsec/fix001.
49. Rider JE, Hacker A, Mackintosh CA, Pegg AE, Woster PM, Casero RA. 2007. Spermine and spermidine mediate protection against oxidative damage caused by hydrogen peroxide. Amino Acids 33:231–240. https://doi.org/10.1007/s00726-007-0513-4.
50. Ferrer A, Rivera J, Zapata C, Norambuena J, Sandoval Á Chávez R, Orellana O, Levicán G. 2016. Cobalamin protection against oxidative stress in the acidophilic iron-oxidizing bacterium *Leptospirillum* group II CF-1. Front Microbiol 7:748. https://doi.org/10.3389/fmicb.2016.00748.
51. Bellenberg S, Barthen R, Boretska M, Zhang R, Sand W, Vera M. 2014. Manipulation of pyrite colonization and leaching by iron-oxidizing *Acidithiobacillus* species. Appl Microbiol Biotechnol 99:1435–1449. https://doi.org/10.1007/s00253-014-6180-y.
52. Noël N, Florian B, Sand W. 2010. AFM & EFM study on attachment of acidophilic leaching organisms. Hydrometallurgy 104:370–375. https://doi.org/10.1016/j.hydromet.2010.02.021.
53. Bellenberg S, Díaz M, Noël N, Sand W, Poetsch A, Guiliani N, Vera M. 2014. Biofilm formation, communication and interactions of leaching bacteria during colonization of pyrite and sulfur surfaces. Res Microbiol 165:773–781. https://doi.org/10.1016/j.resmic.2014.08.006.
54. Bond PL, Smriga SP, Banfield JF. 2000. Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. Appl Environ Microbiol 66:3842–3849. https://doi.org/10.1128/AEM.66.9.3842-3849.2000.
55. Wilmes P, Remis JP, Hwang M, Auer M, Thelen MP, Banfield JF. 2009. Natural acidophilic biofilm communities reflect distinct organismal and functional organization. ISME J 3:266–270. https://doi.org/10.1038/ismej.2008.90.
56. Harneit K, Goksel A, Kock D, Klock JH, Gehrke T, Sand W. 2006. Adhesion to metal sulfide surfaces by cells of *Acidithiobacillus ferrooxidans*, *Acidithiobacillus thiooxidans* and *Leptospirillum ferrooxidans*. Hydrometallurgy 83:245–254. https://doi.org/10.1016/j.hydromet.2006.03.044.
57. Rohwerder T, Sand W. 2007. Mechanisms and biochemical fundamentals of bacterial metal sulfide oxidation, p 35–58. *In* Donati ER, Sand W (ed), Microbial processing of metal sulfides. Springer, Dordrecht, Netherlands.
58. Hengge R. 2009. Principles of c-di-GMP signalling in bacteria. Nat Rev Microbiol 7:263–273. https://doi.org/10.1038/nrmicro2109.
59. Ruiz LM, Castro M, Barriga A, Jerez CA, Guiliani N. 2012. The extremophile *Acidithiobacillus ferrooxidans* possesses a c-di-GMP signalling pathway that could play a significant role during bioleaching of minerals. Lett Appl Microbiol 54:133–139. https://doi.org/10.1111/j.1472-765X.2011.03180.x.
60. Castro M, Deane SM, Ruiz L, Rawlings DE, Guiliani N. 2015. Diguanylate cyclase null mutant reveals that c-di-GMP pathway regulates the motility and adherence of the extremophile bacterium *Acidithiobacillus caldus*. PLoS One 10:e0116399. https://doi.org/10.1371/journal.pone.0116399.
61. Keiski CL, Harwich M, Jain S, Neculai AM, Yip P, Robinson H, Whitney JC, Riley L, Burrows LL, Ohman DE, Howell PL. 2010. AlgK is a TPR-containing protein and the periplasmic component of a novel exopolysaccharide secretin. Structure 18:265–273. https://doi.org/10.1016/j.str.2009.11.015.
62. Rohwerder T, Gehrke T, Kinzler K, Sand W. 2003. Bioleaching review part A. Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation. Appl Microbiol Biotechnol 63:239–248. https://doi.org/10.1007/s00253-003-1448-7.
63. Kinnunen PHM, Puhakka JA. 2005. High-rate iron oxidation at below pH 1 and at elevated iron and copper concentrations by a *Leptospirillum ferriphilum* dominated biofilm. Process Biochem 40:3536–3541. https://doi.org/10.1016/j.procbio.2005.03.050.
64. Lemire JA, Harrison JJ, Turner RJ. 2013. Antimicrobial activity of metals: mechanisms, molecular targets and applications. Nat Rev Microbiol 11:371–384. https://doi.org/10.1038/nrmicro3028.
65. Quatrini R, Johnson DB (ed). 2016. Acidophiles: life in extremely acidic environments. Caister Academic Press, Norfolk, United Kingdom.
66. Gehrke T, Hallmann R, Kinzler K, Sand W. 2001. The EPS of *Acidithiobacillus ferrooxidans*—a model for structure-function relationships of attached bacteria and their physiology. Water Sci Technol 43(6):159–167.
67. Mackintosh ME, Down P, Ojg SSP, Mackintosh ME. 1978. Nitrogen fixation by *Thiobacillus ferrooxidans*. Biotechnol Bioeng 105:215–218.
68. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P. 2013. A biomolecular isolation framework for eco-systems biology. ISME J 7:110–121. https://doi.org/10.1038/ismej.2012.72.
69. Hunt M, De Silva N, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16:294. https://doi.org/10.1186/s13059-015-0849-0.
70. Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028. https://doi.org/10.1093/bioinformatics/btm039.
71. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
72. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
73. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122. https://doi.org/10.1093/nar/gkr1044.

74. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. 2016. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol 2:16180. https://doi.org/10.1038/nmicrobiol.2016.180.
75. Koskinen P, Toronen P, Nokso-Koivisto J, Holm L. 2015. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics 31:1544–1552. https://doi.org/10.1093/bioinformatics/btu851.
76. Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480.
77. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
78. Liao Y, Smyth GK, Shi W. 2014. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656.
79. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.
80. Rappsilber J, Mann M, Ishihama Y. 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2:1896–1906. https://doi.org/10.1038/nprot.2007.261.
81. Kaur H, Takefuji M, Ngai CY, Carvalho J, Bayer J, Wietelmann A, Poetsch A, Hoelper S, Conway SJ, Möllmann H, Looso M, Troidl C, Offermanns S, Wettschureck N. 2016. Targeted ablation of periostin-expressing activated fibroblasts prevents adverse cardiac remodeling in mice. Circ Res 118:1906–1917. https://doi.org/10.1161/CIRCRESAHA.116.308643.
82. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794–1805. https://doi.org/10.1021/pr101065j.
83. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13:731–740. https://doi.org/10.1038/nmeth.3901.
84. Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, Kuzyakiv R, Nguyen Q, Owen S, Soiland-Reyes S, Straszewski J, Van Niekerk DD, Williams AR, Malmstrom L, Rinn B, Muller W, Goble C. 2017. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. Nucleic Acids Res 45:D404–D407. https://doi.org/10.1093/nar/gkw1032.
85. Herold M. 2017. Multi-omics reveal lifestyle of acidophile, mineral-oxidizing model species *Leptospirillum ferriphilum*^{T}. fairdomhub https://doi.org/10.15490/fairdomhub.1.investigation.162.1.
86. Jiang H, Liang Y, Yin H, Xiao Y, Guo X, Xu Y, Hu Q, Liu H, Liu X. 2015. Effects of arsenite resistance on the growth and functional gene expression of *Leptospirillum ferriphilum* and *Acidithiobacillus thiooxidans* in pure culture and coculture. Biomed Res Int 2015:203197. https://doi.org/10.1155/2015/203197.
87. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109.
88. Wibberg D, Bremges A, Dammann-Kalinowski T, Maus I, Igeno MI, Vogelsang R, Konig C, Luque-Almagro VM, Roldan MD, Sczyrba A, Moreno-Vivian C, Blasco R, Pohler A, Schluter A. 2016. Finished genome sequence and methylome of the cyanide-degrading *Pseudomonas pseudoalcaligenes* strain CECT5344 as resolved by single-molecule real-time sequencing. J Biotechnol 232:61–68. https://doi.org/10.1016/j.jbiotec.2016.04.008.
89. Issotta F, Galleguillos PA, Moya-Beltrán A, Davis-Belmar CS, Rautenbach G, Covarrubias PC, Acosta M, Ossandon FJ, Contador Y, Holmes DS, Marín-Eliantonio S, Quatrini R, Demergasso C. 2016. Draft genome sequence of chloride-tolerant *Leptospirillum ferriphilum* Sp-Cl from industrial bioleaching operations in northern Chile. Stand Genomic Sci 11:19. https://doi.org/10.1186/s40793-016-0142-1.
90. Mi S, Song J, Lin JJ, Che Y, Zheng H, Lin JJ. 2011. Complete genome of *Leptospirillum ferriphilum* ML-04 provides insight into its physiology and environmental adaptation. J Microbiol 49:890–901. https://doi.org/10.1007/s12275-011-1099-9.
91. Zhang X, Liu X, Liang Y, Xiao Y, Ma L, Guo X, Miao B, Liu H, Peng D, Huang W, Yin H. 2017. Comparative genomics unravels the functional roles of co-occurring acidophilic bacteria in bioleaching heaps. Front Microbiol 8:790. https://doi.org/10.3389/fmicb.2017.00790.

Downloaded from http://aem.asm.org/ on June 11, 2018 by guest

C.2 Systems Biology of Acidophile Biofilms for Efficient Metal Extraction.

Stephan Christel, Mark Dopson, Mario Vera, Wolfgang Sand, **Malte Herold**, Paul Wilmes,
Antoine Buetti-Dinh, Igor Pivkin, Christian Trötschel, Ansgar Poetsch, Jan Nygren, Mikael
Kubista

Advanced Materials Research **1130**: 312–315
2015
DOI: 10.4028/www.scientific.net/AMR.1130.312

Contributions of author include:

- Writing and revision of manuscript

Advanced Materials Research Vol. 1130 (2015) pp 312-315
© (2015) Trans Tech Publications, Switzerland
doi:10.4028/www.scientific.net/AMR.1130.312

Submitted: 2015-06-15
Revised: 2015-07-13
Accepted: 2015-08-02

Systems Biology of Acidophile Biofilms for Efficient Metal Extraction

Stephan Christel[1,*], Mark Dopson[1], Mario Vera[2], Wolfgang Sand[2],
Malte Herold[3], Paul Wilmes[3], Antoine Buetti-Dinh[4], Igor Pivkin[4],
Christian Trötschel[5], Ansgar Poetsch[5], Jan Nygren[6], Mikael Kubista[6]

[1]Linnaeus University, Sweden
[2]Universität Duisburg-Essen, Germany
[3]University of Luxembourg, Luxembourg
[4]Università della Svizzera italiana, Switzerland
[5]Ruhruniversität Bochum, Germany
[6]TATAA Biocenter AB, Sweden

*Author to whom correspondence should be addressed; E-Mail: stephan.christel@lnu.se;
Tel.: +46-480-446156; Fax: +46-480-447305

Keywords: Biofilm, Acidophiles, Chalcopyrite, Heap Bioleaching, Copper, Lag Phase, Omics

Abstract. The objectives of a recently funded European Union (ERASysApp) are to understand and provide solutions to the problem of the long lag period typically encountered in new mineral heap bioleaching operations of the copper containing mineral chalcopyrite. In practice, this lag phase can be up to three years and the long time period adds to the operating expenses of bioheaps for chalcopyrite dissolution. One of the major time determining factors in bioleaching heaps is suggested to be the speed of mineral colonization by the acidophilic microorganisms present. By applying confocal microscopy, metatranscriptomics, metaproteomics, bioinformatics, and computer modeling the study aims to investigate the processes leading up to, and influencing the attachment of three moderately thermophilic sulfur- and/or iron-oxidizing model species: *Acidithiobacillus caldus*, *Leptospirillum ferriphilum*, and *Sulfobacillus thermosulfidooxidans*. Stirred tank reactors containing chalcopyrite concentrate, inoculated with these species, allows investigation of the effects of various inoculation orders and proportions on the lag phase and rates of metal release. Meanwhile, confocal microscopy studies of cell attachment to chalcopyrite mineral particles, as well as metatranscriptomics and metaproteomics of the formed biofilms further increases the so far limited understanding of the attachment process and help develop a model thereof. By fulfilling the projects goal to decrease the length of the lag phase in chalcopyrite bioleaching operations it is hoped to increase their economic feasibility and thereby, raise industrial interest in bioleaching as a suitable technology to extract copper from chalcopyrite mineral.

Introduction

Presently there is an increase in the European demand for metals. Consequently, preferably environmentally friendly techniques must be developed to meet this demand. Despite the need to dig the ore and crush it for both ‘biomining’ and conventional metal extraction processes, biomining is suggested to partially uphold this criteria as it exploits acidophilic microorganisms for metal solubilisation from sulphide ores in tanks, heaps and dumps [1, 2]. Bioleaching of copper sulphide minerals such as chalcopyrite ($CuFeS_2$; the largest copper resource in the world) is usually conducted in engineered heaps and accounts for approximately 15% of the present world copper production. In comparison with the traditional roasting and smelting processes used for metal extraction of sulphide ores, bioleaching reduces the release of toxic compounds, such as sulphur dioxide, associated with these techniques. The major role of microbial populations in biomining is to catalyse the regeneration of ferric iron from ferrous iron and generate protons by oxidation of

All rights reserved. No part of contents of this paper may be reproduced or transmitted in any form or by any means without the written permission of Trans Tech Publications, www.ttp.net. (ID: 194.47.169.28-14/10/15,11:41:52)

sulphuric acid species. Although heap bioleaching is a low cost method, it is a slow process that can take up to several years to achieve economic metal recoveries.

One factor that strongly influences the economics of an industrial bioheap plant (Fig. 1) is the lag time between addition of acid to the top of the heap and metal recoveries. An example is a test heap constructed for metal recovery from a black schist in Talvivaara, Finland where nickel and zinc recoveries were >80% after 480 days while copper recovery from chalcopyrite did not proceed until after 500 days [3]. The duration of this lag time is suggested to be determined predominantly by the speed of microbial colonization of the mineral and the mineralogy of the ore. However, to date no optimisation has been achieved in terms of biofilm formation for enhanced bioleaching of industrially relevant metal sulphides. Microorganisms in biofilm communities thrive when attached to substrates by 'extracellular polymeric substances' (EPS). A fundamental question in microbial ecology is to understand the interaction(s) among the different species present in natural and/or industrial or engineered biofilm systems. This is especially the case in bioleaching, since successional processes, such as attachment of microbes to the mineral, play important roles in biofilm formation [4].

Consequently, fundamental knowledge on the mechanisms of biofilm formation is central for the design of heap inoculation strategies to increase the efficiency of ore processing, in particular by reducing the lag time between heap initiation and metal recovery.

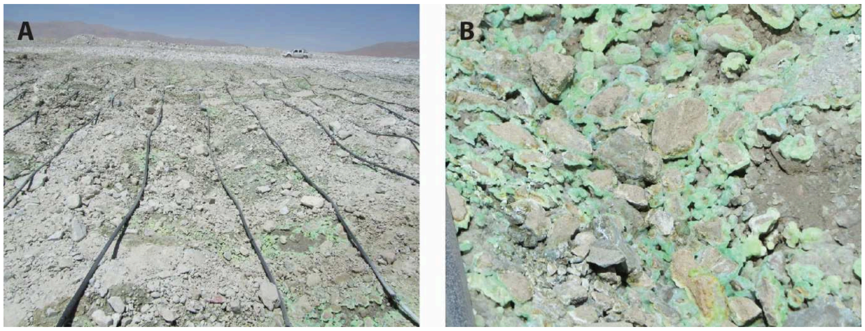


Figure 1. The top of a bioheap showing irrigation of the mineral (A) and copper precipitates formed as a result of chalcopyrite dissolution (B).

Study of acidophile biofilm by 'omics'

Examination of biofilm formation by "omics" and microscopy serve as a way to model the rate and influence of microbial species on the biofilm formation, as well as the time required for copper to be released from the mineral matrix. However, one of the issues of multi-species microbial biofilms is that they are inherently heterogeneous and community-wide networks are expected to be more than merely sums of their respective parts. In order to meaningfully integrate "omics" datasets, obtained biomolecular fractions have to be representative of the sampled microbial assemblages before high-resolution molecular analyses. For this reason, appropriate laboratory methods have been developed which allow the isolation of concomitant RNA and protein from single unique microbial community samples. The resulting 'omics' data fulfil the premise of standardised systematic measurements and can be meaningfully integrated. In particular, the RNA sequencing and protein data can be mapped onto the genomes of constituent community members with high accuracy which in turn, forms the basis for formulating community-wide multi-scale models.

A second issue is that the application of 'omics' techniques for the investigation of acidophilic biofilms during bioleaching is highly limited. However, many studies have been carried out for other types of acidophile biofilms [5, 6]. Although metatranscriptomics and metaproteomics (i.e.

RNA transcript and protein complement analysis of microbial communities, respectively) is rapidly evolving, the field is still in its infancy and for many tasks no established, robust tools are available. In this study, some of the most important moderately thermophilic bioleaching bacteria on chalcopyrite surfaces, *Acidithiobacillus caldus*, *Leptospirillum ferriphilum* and *Sulfobacillus thermosulfidooxidans* are used to investigate the biofilm formation [7, 8]. Therefore, a unique feature of the experimental setup is the use of well-defined microbial communities of limited diversity and known cultivation conditions. Consequentially, despite the described challenges of acidophile 'omics' techniques, establishing and testing novel approaches for measuring and modelling a mixed-microorganism biofilm formation process is possible. In particular, the following analyses assist in meeting this goal: i. integrative analysis using microscopy and high throughput omics of temporal and spatial biofilm development on the species and molecular network level; ii. exploitation of large data sets for extensive bioinformatic analysis; and iii. processing of the resulting data for the formulation of multi-species biological network models (based on meta–omics data) alongside multi-species particle-based models (based on imaging data) to define the key factors affecting biofilm formation.

Modelling of 'omics' data

Modelling biological processes at the molecular level is usually based on idealised network models that represent interacting components in a simplified fashion. Efficient tools are available to deduce causal links between biological components from 'omics' experiments. This allows the assimilation of metatranscriptomic and metaproteomic data into network models which can subsequently be analysed with appropriate mathematical tools. Bayesian networks, regression and simulated annealing allow inference of network topology from quantitative, omics data [9]. Network analysis tools consist of a diversified set of methods such as bifurcation and sensitivity analysis which have a long tradition in engineering and more recently, have been applied to biology. These tools allow the identification of important control points and parameters for the studied system. Further, models describing cells at the molecular level can be integrated into particle-based methods and consequently account for the interactions between different cells, and allow the study of bacterial communities.

Planned outcomes

The models will valorise systems biology knowledge and will be used to predict and manipulate biofilm development to reduce the lag time between heap initiation and onset of copper solubilisation. These 'improved' biofilms will be iteratively evaluated to connect the model development with mineral oxidation rates, as well as their tolerance to changes in environmental conditions. Decisively, the project will transfer systems biology knowledge into an application by including end user companies actively carrying out biomining and other biotechnological applications. Thereby, we are maximising the industrial application of the data and hopefully make this environmentally friendly technique more attractive to mining companies.

References

[1] N. Pradhan, K.C. Nathsarma, K. Srinivasa Rao, L.B. Sukla, B.K. Mishra, Heap bioleaching of chalcopyrite: A review, Minerals Engineering 21 (2008) 355-365.

[2] H.R. Watling, The bioleaching of sulphide minerals with emphasis on copper sulphides – A review, Hydrometallurgy 84 (2006) 81-108.

[3] M. Riekkola-Vanhanen, Talvivaara mining company – From a project to a mine, Minerals Engineering 48 (2013) 2-9.

[4] P. Wilmes, J.P. Remis, M. Hwang, M. Auer, M.P. Thelen, J.F. Banfield, Natural acidophilic biofilm communities reflect distinct organismal and functional organization, ISME J 3 (2009) 266-270.

[5] M. Vera, B. Krok, S. Bellenberg, W. Sand, A. Poetsch, Shotgun proteomics study of early biofilm formation process of *Acidithiobacillus ferrooxidans* ATCC 23270 on pyrite, Proteomics 13 (2013) 1133-1144.

[6] C. Baker-Austin, J. Potrykus, M. Wexler, P.L. Bond, M. Dopson, Biofilm development in the extremely acidophilic archaeon '*Ferroplasma acidarmanus*' Fer1, Extremophiles 14 (2010) 485-491.

[7] J.J. Plumb, N.J. McSweeney, P.D. Franzmann, Growth and activity of pure and mixed bioleaching strains on low grade chalcopyrite ore, Minerals Engineering 21 (2008) 93-99.

[8] M. Dopson, E.B. Lindstrom, Analysis of commuinity composition during moderately thermophilic bioleaching of pyrite, arsenical pyrite, and chalcopyrite, Microbial Ecology 48 (2004) 19-28.

[9] D. Hurley, H. Araki, Y. Tamada, B. Dunmore, D. Sanders, S. Humphreys, M. Affara, S. Imoto, K. Yasuda, Y. Tomiyasu, K. Tashiro, C. Savoie, V. Cho, S. Smith, S. Kuhara, S. Miyano, D.S. Charnock-Jones, E.J. Crampin, C.G. Print, Gene network inference and visualization tools for biologists: application to new human transcriptome datasets, Nucleic Acids Res 40 (2012) 2377-2398.

C.3 Weak Iron Oxidation by *Sulfobacillus thermosulfidooxidans* maintains a favorable redox potential for chalcopyrite bioleaching.

Stephan Christel, **Malte Herold**, Sören Bellenberg, Antoine Buetti-Dinh, Mohamed El Hajjami,
Igor Pivkin, Wolfgang Sand, Paul Wilmes, Ansgar Poetsch, and Mark Dopson
2018
Frontiers in Microbiology **in review**

Contributions of author include:

- Data analysis
- Writing and revision of manuscript

Weak Iron Oxidation by Sulfobacillus thermosulfidooxidans Maintains a Favorable Redox Potential for Chalcopyrite Bioleaching

Stephan Christel[1*], Malte Herold[2], Soeren Bellenberg[3], Antoine Buetti-Dinh[4, 5], Mohamed El Hajjami[6], Igor Pivkin[4, 5], Wolfgang Sand[3, 7, 8], Paul Wilmes[2], Ansgar Poetsch[6, 9], Mark Dopson[1]

[1]Centre for Ecology and Evolution in Microbial Model Systems, Linnaeus University, Sweden, [2]University of Luxembourg, Luxembourg, [3]Universität Duisburg-Essen, Germany, [4]Università della Svizzera italiana, Switzerland, [5]Swiss Institute of Bioinformatics (SIB), Switzerland, [6]Ruhr-Universität Bochum, Germany, [7]Donghua University, China, [8]Freiberg University of Mining and Technology, Germany, [9]Plymouth University, United Kingdom

In review

Submitted to Journal:
Frontiers in Microbiology

Specialty Section:
Extreme Microbiology

Article type:
Original Research Article

Manuscript ID:
421883

Received on:
30 Aug 2018

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

SC conducted the laboratory experiments. MH performed bioinformatic analysis of transcript data. SB provided microscopic imaging. SC, SB, AB, ME, IP, WS, PW, AP, MD were involved in data analysis and biological interpretation of the results. SC drafted the manuscript and all authors contributed to its preparation.

Keywords

Redox control, Microbial, Chalcopyrite, Iron oxidation, bioleaching, Sulfobacillus, Leptospirillum

Abstract

Word count: 274

Bioleaching is an emerging technology, describing the microbially assisted dissolution of sulfidic ores that provides a more environmentally friendly alternative to many traditional metal extraction methods, such as roasting or smelting. Industrial interest increases steadily and today, circa 15-20% of the world's copper production can be traced back to this method. However, bioleaching of the world's most abundant copper mineral chalcopyrite suffers from low dissolution rates, often attributed to passivating layers, which need to be overcome to use this technology to its full potential. To prevent these passivating layers from forming, leaching needs to occur at a low oxidation/reduction potential (ORP), but chemical redox control in bioleaching heaps is difficult and costly. As an alternative, selected weak iron-oxidizers could be employed that are incapable of scavenging exceedingly low concentrations of iron and therefore, raise the ORP just above the onset of bioleaching, but not high enough to allow for the occurrence of passivation. In this study, we report that microbial iron oxidation by Sulfobacillus thermosulfidooxidans meets these specifications. Chalcopyrite concentrate bioleaching experiments with S. thermosulfidooxidans as the sole iron oxidizer exhibited significantly lower redox potentials and higher release of copper compared to communities containing the strong iron oxidizer Leptospirillum ferriphilum. Transcriptomic response to single and co-culture of these two iron oxidizers was studied and revealed a greatly decreased number of mRNA transcripts ascribed to iron oxidation in S. thermosulfidooxidans when cultured in the presence of L. ferriphilum. This allowed for the identification of genes potentially responsible for S. thermosulfidooxidans' weaker iron oxidation to be studied in the future, as well as underlined the need for mechanisms to control the microbial population in bioleaching heaps.

In review

Ethics statements

(Authors are required to state the ethical considerations of their study in the manuscript, including for cases where the study was exempt from ethical approval procedures)

Does the study presented in the manuscript involve human or animal subjects: No

Weak Iron Oxidation by *Sulfobacillus thermosulfidooxidans* Maintains a Favorable Redox Potential for Chalcopyrite Bioleaching

Stephan Christel[1*], Malte Herold[2], Sören Bellenberg[3], Antoine Buetti-Dinh[4,5], Mohamed El Hajjami[6], Igor V. Pivkin[4,5], Wolfgang Sand[3,7,8] , Paul Wilmes[2], Ansgar Poetsch[6,9], and Mark Dopson[1]

[1]Centre for Ecology and Evolution in Microbial Model Systems, Linnaeus University, Kalmar, Sweden

[2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg

[3]Aquatic Biotechnology, Universität Duisburg-Essen, Essen, Germany

[4]Institute of Computational Science, Faculty of Informatics, Università della Svizzera Italiana, Lugano, Switzerland

[5]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[6]Plant Biochemistry, Ruhr Universität Bochum, Germany

[7]College of Environmental Science and Engineering, Donghua University, Shanghai, PR China

[8]Mining Academy and Technical University Freiberg, Freiberg, Germany

[9]School of Biomedical and Healthcare Sciences, Plymouth University, UK

***Correspondence:**
Stephan Christel
stephan.christel@lnu.se

Keywords: redox control, microbial, bioleaching, chalcopyrite, iron oxidation, sulfobacillus, leptospirillum.

In review

Abstract

Bioleaching is an emerging technology, describing the microbially assisted dissolution of sulfidic ores that provides a more environmentally friendly alternative to many traditional metal extraction methods, such as roasting or smelting. Industrial interest increases steadily and today, circa 15-20% of the world's copper production can be traced back to this method. However, bioleaching of the world's most abundant copper mineral chalcopyrite suffers from low dissolution rates, often attributed to passivating layers, which need to be overcome to use this technology to its full potential. To prevent these passivating layers from forming, leaching needs to occur at a low oxidation/reduction potential (ORP), but chemical redox control in bioleaching heaps is difficult and costly. As an alternative, selected weak iron-oxidizers could be employed that are incapable of scavenging exceedingly low concentrations of iron and therefore, raise the ORP just above the onset of bioleaching, but not high enough to allow for the occurrence of passivation. In this study, we report that microbial iron oxidation by *Sulfobacillus thermosulfidooxidans* meets these specifications. Chalcopyrite concentrate bioleaching experiments with *S. thermosulfidooxidans* as the sole iron oxidizer exhibited significantly lower redox potentials and higher release of copper compared to communities containing the strong iron oxidizer *Leptospirillum ferriphilum*. Transcriptomic response to single and co-culture of these two iron oxidizers was studied and revealed a greatly decreased number of mRNA transcripts ascribed to iron oxidation in *S. thermosulfidooxidans* when cultured in the presence of *L. ferriphilum*. This allowed for the identification of genes potentially responsible for *S. thermosulfidooxidans*' weaker iron oxidation to be studied in the future, as well as underlined the need for mechanisms to control the microbial population in bioleaching heaps.

In review

1 Introduction

Biomining is a sustainable process for metal extraction from sulfidic ores that has been studied by researchers around the globe since its emergence in the early 1950s (Temple and Colmer, 1951; Bryner and Jameson, 1958). In the recent decades and with its industrial application in mind, understanding of this natural process has significantly improved (Rohwerder et al., 2003; Watling, 2006; Vera et al., 2013; Jerez, 2017) and today, biomining is defined to be the microbial promoted oxidation of insoluble metal sulfides to acid soluble sulfates. In a technique termed bioleaching, this is undertaken to solubilize and recover metals of interest that form part of the metal sulfide mineral matrix. Ferrous iron (Fe^{2+})-oxidizing acidophilic microorganisms are responsible for the regeneration of the chemical oxidant ferric iron (Fe^{3+}), which in turn attacks the sulfidic mineral, and breaks its covalent bonds. This releases ferrous iron plus any other contained metals and completes the catalytic cycle (Vera et al., 2013). While having initial economic disadvantages, mainly attributed to the long lag phase after construction of a bioleaching heap, biomining technologies are commonly considered more environmentally friendly than most conventional methods (Johnson, 2014). Today, increasing amounts of metals are extracted or processed by biomining technologies in many countries that include Chile, Australia, and South Africa, with the bioleaching of secondary copper sulfides accounting for an estimated 15-20% of the world wide copper production (Brierley and Brierley, 2013).

Bioleaching of primary copper minerals, such as the world's most abundant copper mineral chalcopyrite ($CuFeS_2$), remains challenging and suffers from slow dissolution rates. This is often attributed to the formation of passivation layers on the mineral surface (Cordoba et al., 2009; Wang et al., 2016), but it has also been argued that the semiconductor properties of chalcopyrite itself could be responsible (Crundwell, 2015). To date, extensive efforts to elucidate the exact nature of

This is a provisional file, not the final typeset article

chalcopyrite passivation have not been successful (Khoshkhoo et al., 2014a; Khoshkhoo et al., 2017). Despite this, strategies have been discovered to diminish the passivating effect, including bioleaching at high temperatures and low redox potentials (Li et al., 2013; Panda et al., 2015). Due to their large concentrations, the oxidation/reduction potential in bioleaching systems is predominantly determined by the Fe^{3+}/Fe^{2+} redox couple, whereby high concentrations of Fe^{3+} indicate high potentials. At low redox potentials and in the presence of millimolar concentrations of Fe^{2+} and Cu^{2+}, chalcopyrite is suggested to be transformed into the secondary copper sulfide chalcocite (Cu_2S) which is more readily oxidized by the Fe^{3+} provided by microbial action (Hiroyoshi et al., 2013). Methods to control the redox potential of the leaching solution include the addition of chemical reductants (Zhao et al., 2017) or limitation of oxygen (Third et al., 2002). However, the technical realization of such methods in a large industrial bioheap with gradients of e.g. temperature, oxygen concentration, and substrates has not been accomplished. Many studies have investigated the optimal microbial consortia in bioleaching operations (Rawlings and Johnson, 2007), usually focusing on the need to efficiently oxidize Fe^{2+} that drives the redox potential above that optimal for chalcopyrite dissolution (Hiroyoshi et al., 2013). In contrast, little attention has been paid to the possibility of controlling the redox potential of a bioleaching system by influencing the ratio of ferric to ferrous iron via suitable iron-oxidizing microbes (Masaki et al., 2018).

A large range of acidophile microbes have the capability to oxidize Fe^{2+} to gain energy under acidic conditions (Hedrich et al., 2011) and are therefore applicable in biomining operations. Among those are members of the *Acidithiobacillus*, *Acidimicrobium*, *Acidiferrobacter*, *Sulfobacillus*, and *Ferroplasma* genera (reviewed in Quatrini and Johnson, 2016). In bioleaching systems, one of the most abundant iron oxidizers is the moderately thermophilic, autotroph *Leptospirillum ferriphilum* (Penev and Karamanev, 2010; Christel et al., 2017). This moderate thermophile solely derives its energy from the oxidation of ferrous iron (Coram and Rawlings, 2002) and is capable of doing so at very low Fe^{2+} ion concentrations and redox potentials as high as 700 mV vs. Ag/AgCl (Rawlings et al., 1999), giving it a significant advantage over other species. Another iron-oxidizer commonly found in acidic, sulfur rich environments is the moderately thermophilic *Sulfobacillus thermosulfidooxidans* (Karavaiko et al., 2005) that in contrast to *L. ferriphilum*, is unable to scavenge exceedingly scarce ferrous iron and is therefore considered a “weak” iron oxidizer in this study. In addition to Fe^{2+}, *S. thermosulfidooxidans* is capable of oxidizing inorganic sulfur compounds (ISCs) and can utilize organic molecules to meet its carbon demands (Tsaplina et al., 2000). ISC oxidation is an important process in bioleaching heaps to remove excess sulfur compounds (Dopson and Lindstrom, 1999) and generate the necessary acidity, which is otherwise consumed by gangue minerals in low grade ores (Baldi et al., 1991). Often, this role is fulfilled by obligate ISC-oxidizing species, such as the mesophile *Acidithiobacillus thiooxidans* or moderately thermophile *A. caldus* (Hallberg and Lindstrom, 1994).

In this study, we hypothesized that by inoculation of chalcopyrite ore with suitable iron-oxidizing bacteria the redox potential of the leachate in the initial phase of bioleaching experiments can be controlled. By these means, the redox potential can be maintained close to the optimum range. The initial rate of chalcopyrite dissolution is enhanced and thereby increases the amount of released copper. The applicability of this approach to industrial bioleaching operations is discussed.

2 Materials and Methods

2.1 Mineral

Chalcopyrite was provided by Boliden AB (Sweden) and originates from the Aitik copper mine (N 67° 4' 24", E 20° 57' 51"). The flotation concentrate used in this study contained 29.5 % copper (Supplementary File 1). For bioleaching experiments, the concentrate was sieved to obtain the size fraction between 50 and 100 µm and subsequently washed in three volumes of 0.1 M EDTA in 0.4 M NaOH for 10 min under stirring. Elemental sulfur was then removed from the surfaces by three iterations of washing with one volume of acetone. Finally, the mineral was dried at 60 °C overnight and then sterilized at 120 °C for 10 h under nitrogen.

2.2 Bacterial strains and growth conditions

Three bacterial acidophile species were used in this study, *L. ferriphilum* DSM 14647, *S. thermosulfidooxidans* DSM 9293, and *A. caldus* DSM 8584. Prior to the bioleaching experiments, cells were maintained in three separate continuous cultures so that the cells were under the same growth state when all experiments were inoculated. The continuous cultures were maintained at 38 °C, fed with MAC medium (Mackintosh, 1978), and electron donor added in the form of 100 mM ferrous sulfate (*L. ferriphilum*) or 5 mM potassium tetrathionate (*S. thermosulfidooxidans* and *A. caldus*). The continuous culture vessels, all tubing, plus MAC medium were autoclaved while the ferrous sulfate and potassium tetrathionate were sterile filtered (0.2 µm pore size, cellulose acetate filter, PALL).

In review

2.3 Bioleaching experiments

Bioleaching experiments were conducted in quadruplets in 250 mL Erlenmeyer flasks. 100 mL MAC medium was supplemented with 2% (wt/vol) chalcopyrite concentrate and inoculated with combinations of 10^7 cells per mL of the three bacterial species obtained by centrifugation from the continuous cultures (12 500 × g, 20 min) including three single, three binary, and one tertiary combination, plus one sterile control. Cultures were incubated at 38 ± 2 °C under slow shaking (120 rpm) for 14 days after the redox potential reached 400 mV versus Ag/AgCl for the first time.

Experiments were analyzed for pH (pHenomenal® 221, VWR), redox potential (Ag/AgCl with 3 M KCl; InLab® Redox-L, Mettler-Toledo), ferrous iron, total dissolved sulfur, elemental sulfur, as well as total iron and copper concentration in the leach liquor. Ferrous iron concentration was assessed by titration of its 1,10 phenanthroline complex (Walden et al., 1933; Dopson and Lindstrom, 1999). In short, 200 µL of bioleaching sample was centrifuged for 5 min at 16 000 × g. Supernatant was mixed with the same volume of 15 mM 1,10-phenantroline in 5 mM $FeSO_4$, added to 1 mL of 1 M H_2SO_4 and subsequently titrated from orange to blue with 1 mM $CeSO_4$. Total soluble and elemental sulfur were measured by photospectrometric measurement of thiocyanate complexes obtained by cyanolysis (Kelly et al., 1969) from the supernatant and pellet of a bioleaching sample respectively. For total soluble sulfur, 500 µL sample were centrifuged for 5 min at 16 000 × g and the supernatant mixed with 100 µL 0.5 M NaCN. After 10 min of incubation at room temperature, 500 µL of phosphate buffer (pH 7.2) and 100 µl 50 mM $CuSO_4$ was added, followed by 30 min of incubation at room temperature. Then, complexes were formed by addition of 400 µL 1.5 M $FeNO_3$ in 4 M $HClO_4$ and distilled H_2O to a volume of 2 mL. The reaction was then measured at 460 nm against a calibration curve of thiocyanate treated in the same way. For measurement of elemental sulfur, the mineral pellet of the same sample was dissolved in 2 mL of absolute acetone. 200 µL of this solution was then processed as described in the total soluble sulfur analysis, except that no $CuSO_4$ was added, and the

This is a provisional file, not the final typeset article

calibration curve consisted of elemental sulfur dissolved in acetone. Total metal concentration was obtained by adding 1.8 mL of 5 M HCl to 200 µL of unaltered bioleaching sample, followed by incubation at 65 °C for 30 min. Then, samples for both total and soluble metals were diluted in 0.1 M HCl appropriately for measurement by atomic absorption spectroscopy (AAS) using a Perkin Elmer AAnalyst 400.

2.4 Extraction and analysis of nucleic acids

After 14 days of active bioleaching time (defined by a redox potential above 400 mV vs. Ag/AgCl), experiments were sampled for nucleic acid extraction. The flasks were left to settle for 5 min before removing 75 mL supernatant to be immediately mixed with an equal volume of sterile, ice-cold MAC medium. Then, the sample was centrifuged at 12 500 × g for 20 minutes at 4 °C. The resulting cell pellet was washed twice by resuspending in 10 and 2 mL of sterile, ice-cold MAC, respectively and then frozen in liquid nitrogen. Cell pellets were subjected to biomolecular extractions according to a previously published method (Roume et al., 2013), skipping the metabolite extraction step. In short, cell pellets were lysed by cryo-milling and bead-beating followed by spin column based isolation of biomolecules with the Allprep kit (Qiagen, Belgium). Purified total RNA was stored at -80 ºC and shipped on dry ice. Ribosomal RNA was depleted with the Ribo-Zero rRNA Removal Kit for bacteria (Illumina, USA). rRNA-depleted RNA for nine samples was then sent to Science for Life Laboratory (Stockholm, Sweden) for sequencing.

2.5 Sequence analysis

Library preparation was performed with the Illumina TruSeq Stranded mRNA kit. Paired-end sequencing was performed on two HiSeq 2500 lanes resulting in on average 94 million reads per sample with length of 126bp and GC% of 54% (Supplementary Table 1). Raw reads were filtered with Trimmomatic v0.32 (Bolger et al., 2014), TrueSeq3-PE adapter sequences were removed using the following parameters: seed mismatch:2; palindrome clip:30; simple clip:10; leading:20; trailing:20; sliding window: 1:3; minlen: 40; maxinfo: 40:0.5. Filtered reads were mapped onto a concatenation of the three reference genomes (*A. caldus* DSM 8584: GCF_000175575.2; *S. thermosulfidooxidans* DSM 9293: GCF_900176145.1; *L. ferriphilum* DSM 14647: GCF_900198525.1) with Bowtie-2 v2.3.2 (Langmead and Salzberg, 2012) with default parameters. Reads mapping to protein coding sequences were counted with the FeatureCounts program of the subread package v1.5.1 (Liao et al., 2014) with the –s 2 parameter accounting for strandedness. Read counts were then normalized and compared per organism with a custom R-script using the DESeq2 package v1.16.1 (Love et al., 2014) in R v3.4.4. Normalization was adapted from scripts provided in a previous publication (Klingenberg and Meinicke, 2017).

2.6 Data availability

Raw sequencing reads are available from ENA SRA under study accession PRJEB27534. Scripts used in the analysis of the sequencing data can be accessed under the following link: https://git-r3lab.uni.lu/malte.herold/RNAseq_LF_ST_redox. Lists of genes relevant for analysis were generated by manual curation of the reference genome annotations and from previous publications (Janosch et al., 2015; Christel et al., 2017).

In review

3 Results and Discussion

3.1 Bioleaching of chalcopyrite concentrate

Bioleaching of chalcopyrite was tested with single, binary, and tertiary combinations of the three model species (*A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans*) plus uninoculated controls to investigate the effect of species composition on redox potential and copper release (Figure 1, Supplementary Figure 1). To aid comprehension, these combinations will be abbreviated using the initial letter of the included species (e.g. 'ASL' for the tertiary combination containing all species or 'LS' for the binary combination of *L. ferriphilum* and *S. thermosulfidooxidans* etc.).

Physical and chemical analysis (Figure 1, Supplementary Figure 1) of the uninoculated controls showed a redox potential of circa 310-330 mV (vs. Ag/AgCl) after stabilization, while the Fe^{2+} concentration steadily increased until plateauing at 5.7 ± 0.2 mM in later stages of the experiment (i.e. day 32, late stage data not shown). The abiotic leaching released a small amount of metal (4.2 ± 0.3 mM Fe and 2.4 ± 0.3 mM Cu after 15 days) from the chalcopyrite by proton attack and/or a small concentration of Fe^{3+} present on the mineral or in the medium. The same behavior was observed in the experiment inoculated exclusively with *A. caldus*, where the redox potential remained at circa 370 mV (i.e. well below the 400 mV perceived to mark the onset of bioleaching) until day 12 when likely environmental bacteria that survived the sterilization process on the mineral became active and commenced iron oxidation. Metal release was only marginally higher than from uninoculated controls, but showed a slight acceleration after the redox increase and reached 6.0 ± 0.7 mM Fe and 4.4 ± 0.4 mM Cu after 14 days. As sulfur compounds released from the mineral matrix are debated to be involved in formation of passivating layers (Khoshkhoo et al., 2014a; b), it is of importance to note that experiments containing *A. caldus* exhibited examples of ISC degradation (Supplementary Figure 1). However, in the timeframe investigated during the leaching experiments, significant accumulation of soluble ISCs and elemental sulfur was also not observed in the combinations excluding *A. caldus* or in the sterile controls (Supplementary Figure 1). Experiments including inoculation with iron-oxidizing bacteria allowed for transformation of Fe^{2+} to Fe^{3+} and therefore the redox potential rose up to e.g. 682 ± 8 mV in the case of 'L'. Accordingly, iron and copper release from all such experiments was significantly higher than from uninoculated controls and 'A'; e.g. reaching the highest Fe concentration of 21.6 ± 1.7 mM in 'L' and highest copper concentration of 14.2 ± 0.3 mM in 'AS' (Figure 1).

As expected, the single species mobilized less copper than mixed species but unexpectedly, the tertiary combination 'ALS' was also outperformed by all binary combinations (Figure 1). This indicated that, in contrast to the currently accepted paradigm of inoculation of bioleaching applications with a broad mixture of biomining organisms, a well-chosen and defined mixture of microorganisms could benefit leaching efforts in the early stages of a bioleaching heap. Furthermore, the different combinations showed very distinct oxidation/reduction potential (ORP) profiles that, based on the present iron oxidizer(s), fell into one of two groups. All combinations containing *L. ferriphilum* had redox potentials between 650 and 680 mV compared to combinations in which it was excluded (i.e. 'AS' and 'S', showing ORPs below 550 mV). To confirm the lower redox potential in bioleaching cultures without *L. ferriphilum*, the 'AS' combination was repeated seven times with the redox potential reaching a maximum of 593 ± 15 mV (Figure 2A). Previous studies report that low redox potentials are favorable for chalcopyrite bioleaching (Third et al., 2002; Hiroyoshi et al., 2013). In accordance with that, in this study the redox potential of the leaching experiments correlated positively with the ratio of released iron/copper (Figure 2B). As dissolution of pure chalcopyrite is theoretically characterized by a 1:1 ratio of released iron to copper, this confirms the preferential oxidation of this copper mineral, or the transiently produced chalcocite, over associated copper-deficient minerals, such as pyrite, at low redox potentials. Our data independently confirms a study by Masaki et al. (2018), in which microbial redox control was also attempted, likewise by members of the Sulfobacilli, i.e. *Sb. sibiricus* and *Sb. acidophilus*. Using iron-oxidizing bacteria in bioleaching processes, which raise the ORP only minimally over the threshold for the onset of leaching is therefore possible and could benefit the performance of chalcopyrite bioleaching processes. Multiple reasons for the induction of different redox potentials by different species are conceivable. First and foremost, the effect could be explained by the effectivity and/or affinity of the respective species' iron oxidation system. Species with a low affinity to ferrous iron, or inferior capability to oxidize it, should in theory maintain lower redox potentials. Additionally, high concentrations of Fe^{3+} ions are known to inhibit iron oxidation differently in different species (Rawlings et al., 1999), which could also contribute to this effect. However, fluorescence microscopy examination of chalcopyrite grains in our bioleaching experiments revealed another difference that could contribute to the observed effect. *L. ferriphilum* showed significantly higher rates of colonization of mineral grains compared to *S. thermosulfidooxidans* (Bellenberg et al., 2018; Supplementary Figure 2). Attachment to metal sulfides is considered important for bioleaching, since in the so called 'contact mechanism', mineral-attached microbes concentrate Fe^{3+} in their EPS, effectively locally increasing the ORP at the microbe-mineral interface compared to the rest of the medium. Likewise, low levels of cell attachment on chalcopyrite mineral grains support the idea that a non-contact mechanism is likely observed for *S. thermosulfidooxidans*. Consequently, ferric ions diluted in the bulk medium maintain a more homogeneous redox environment. Unfortunately, Masaki et al. (2018) did not report on attachment rates and this hypothesis remains to be tested.

Owing to its heterogeneity and complexity, control of both chemical and biological parameters in a bioleaching heap is challenging (Petersen, 2016). Biologically, a major cause of this challenge is that due to implied costs, the mineral cannot be sterilized, and environmental bacteria will be present and thrive in the heap, competing with the inoculated 'strategic' microorganisms. Procedures will have to be developed in order to fully exploit the potential of selected microorganisms, i.e. preventing undesired environmental bacteria from raising the redox potential above critical levels for chalcopyrite dissolution. Conceivably, the heap could be treated with compounds inhibiting their growth, while not impacting the strategic microorganisms. Another approach could be to continuously inoculate the heap with the strategic organisms through the irrigation system. This could also benefit the bioleaching in longer terms, as with the onset of the bioleaching activity the heap temperature rises (Petersen, 2016) and the strategically added microorganisms would have to be adapted to the different temperature zones. In future studies, efforts should therefore be made to identify species with both low Fe^{2+} scavenging capabilities and increasingly high optimal growth temperatures.

In any case, manipulation of leachate, mineral, or other components of a bioleaching heap will naturally increase running costs. Further studies and ultimately large scale testing are needed to validate the viability of such approaches.

3.2 Transcriptomic analysis of Fe^{2+} oxidation and electron transport

In an attempt to elucidate the biological background for the difference in redox potential, both iron-oxidizing model species' transcriptomic response towards each other was investigated (i.e. 'ASL' vs 'AL' for effect of *S. thermosulfidooxidans* on *L. ferriphilum*, and 'ASL' vs 'AS' for the vice versa effect; Figure 3). The entirety of the ecological interactions between the two species are beyond the scope of this study and instead, this section concentrates on gene products related to energy metabolism, iron-, and sulfur oxidation (Supplementary tables 1 and 2). In bioleaching co-culture, *L. ferriphilum* remained remarkably unaffected by the presence of *S. thermosulfidooxidans*. Over its

In review

entire genome (2486 genes), only 36 genes showed significant differential expression in response to *S. thermosulfidooxidans* ($p \leq 0.05$; data not shown). Among the 26 genes attributed to iron oxidation and electron transport, merely three *cbb*$_3$-type cytochrome *c* oxidase subunits (LFTS_01396, _02094, and _02276) exhibited significantly increased transcript numbers in the presence of *S. thermosulfidooxidans*, all of which have log2-fold changes below 1.5 (Supplementary Table 2). No genes involved in iron oxidation or electron transport had significantly higher numbers of RNA transcripts in the absence of *S. thermosulfidooxidans*.

In contrast, *S. thermosulfidooxidans* gene transcript numbers exhibited great variation depending on presence of *L. ferriphilum*. Of its 3805 identified genes, 828 showed significant differential expression. Among the 83 selected genes involved in iron oxidation, electron transport, and sulfur oxidation, 55 had significantly greater or lower RNA transcripts (Table 1, Supplementary Table 3). Large variation was observed in genes related to iron oxidation. In contrast to e.g. some members of the genus *Acidithiobacillus*, Sulfobacilli genomes lack the common iron oxidation protein rusticyanin (Guo et al., 2014). Instead, Sulfobacilli are suggested to utilize sulfocyanin, which is also found in the archaeal iron oxidizers of the genus Ferroplasma (Dopson et al., 2005). In the presence of *L. ferriphilum*, *S. thermosulfidooxidans* strongly decreased transcript numbers attributed to two of the five *soxE* genes coding for this protein (Sulth_0453 and _2749). Additionally, the vast majority of identified cytochromes of all types exhibited decreased transcript counts, along with corresponding biogenesis proteins and quinol oxidases (Table 1, Supplementary Table 3). The strong downregulation of electron chain components that were likely linked to iron oxidation in *S. thermosulfidooxidans* could be explained by the chemical data reported in the previous section. In cultures containing both iron oxidizers, the concentration of available ferrous iron was beyond the detection limit and likely too low for utilization by *S. thermosulfidooxidans*. This may be attributed to *L. ferriphilum* being able to scavenge Fe^{2+} at concentrations far below *S. thermosulfidooxidans*' capabilities and at large Fe^{3+} concentrations that exceed its inhibition limits (Rawlings et al., 1999).

Contrary to this overall trend, one cluster of *S. thermosulfidooxidans* cytochrome *c* oxidase subunits I-IV showed strongly increased transcript counts in the presence of *L. ferriphilum* (Table 1, Sulth_1930-1933). In addition, two cytochrome *c* biogenesis proteins (Sulth_1901 and _2183) and one cytochrome *c* assembly protein (Sulth_0051) exhibited similarly increased transcript numbers. A direct role of cytochromes in iron oxidation has been suggested in an acid mine drainage biofilm and in *L. ferrooxidans* (Jeans et al., 2008; Blake and Griff, 2012). Therefore, the strong opposite regulation of cytochrome oxidases in *S. thermosulfidooxidans* raises the question of their potential functional and/or structural differences. It could be possible that the oxidase exhibiting increased transcript counts in the presence of *L. ferriphilum* indirectly facilitates a higher affinity for Fe^{2+}, or has a lower sensitivity towards oxidative stress induced by accumulating Fe^{3+}. Together with the upregulation of biogenesis and assembly proteins, this could enable *S. thermosulfidooxidans* to at least gain some energy from Fe^{2+} in the presence of a stronger iron oxidizer. Alternatively, the cytochrome *c* oxidase complex upregulated in the presence of *L. ferriphilum* could be part of the strong upregulation of genes described in the following section, i.e. reducing oxygen in the final step of sulfur oxidation systems. Nevertheless, as only cytochromes and cytochrome oxidases that are upregulated in absence of *L. ferriphilum* correlated with higher copper extraction, they may be of greater interest in the context of this study and should be considered in more detail in the future.

3.3 Transcriptomic analysis of ISC oxidation genes

Genes coding for known sulfur oxidation proteins exhibited directionally opposite changes in transcript numbers compared to iron oxidation systems. Conceivably, this was to ensure sufficient supply of energy in a Fe^{2+} deficient environment and the vast majority of *S. thermosulfidooxidans*

This is a provisional file, not the final typeset article

genes related to sulfur metabolism had significantly higher RNA transcripts in the presence of *L. ferriphilum* (Table 1, Supplementary Table 3). The highest of these log2-fold changes were recorded for two copies of tetrathionate hydrolase gene *tetH* (Sulth_0921 and _3251) while a third copy (Sulth_1188) exhibited moderately increased transcript counts in the absence of *L. ferriphilum*. TetH is responsible for the hydrolysis of tetrathionate, an important intermediate in sulfide mineral dissolution. Additionally, thiosulfate can be oxidized by thiosulfate quinone oxidoreductase, encoded by a *doxDA* homologue. The two encoded copies of this gene exhibited increased transcript counts when in co-culture with *L. ferriphilum*, although the log2-fold changes were low (Sulth_1989 and _1691). A similar function is suggested for rhodanese-like proteins, which in humans are sulfur transferases involved in the detoxification of cyanide by transformation to thiocyanate in the liver (Nakajima, 2015). Proteins sharing their active domain are also suggested to play a role in microbial sulfur oxidation, in particular of thiosulfate (Valdes et al., 2008). In *S. thermosulfidooxidans*, seven genes coding for such proteins were present (Table 1, Supplementary Table 3) and four of them exhibited significantly increased transcript numbers in the presence of *L. ferriphilum* (Sulth_1878, _2076, _2172, and _3294). The product of their enzymatic reaction can be further oxidized by heterodisulfide reductase, which is encoded in *S. thermosulfidooxidans* by three sets of genes for its respective subunits HdrBC (Supplementary Table 3). Four Hdr subunit gene loci exhibited significant slightly increased transcript counts in presence of *L. ferriphilum* (Sulth_1025, _1026, _2770, and _2771). Elemental sulfur oxidation is also relevant in the context of sulfide mineral dissolution, and is conducted by the product of two copies of sulfur oxygenase/reductase gene *sor* (Janosch et al., 2015). While one copy was only minimally expressed in both conditions (Sulth_1798; indicating a possibly defunct gene), the second exhibited greater transcript numbers (Sulth_1627) and appeared to be enhanced by *L. ferriphilum*.

Additional contributions to sulfur oxidation systems include sulfate adenylyltransferase which is suggested to be involved in sulfite oxidation in *A. ferrooxidans* and *S. thermosulfidooxidans* strain ST (Guo et al., 2014) and is likely to fulfill the same role in *S. thermosulfidooxidans*^{T}. Similarly, DsrE-family protein (Sulth_2782) has been reported to be associated to oxidative sulfite metabolism (Dahl et al., 2005) and was also found to exhibit increased transcript numbers in presence of *L. ferriphilum* (Table 1).

In the absence of *L. ferriphilum*, only a few genes related to sulfur oxidation were significantly enhanced. Of note in this regard is one of three present copies of sulfide quinone reductase gene *sqr1* (Sulth_0946), which is responsible for the oxidation of sulfide to elemental sulfur a hypothetical protein within the *hdr* gene cluster (Sulth_1024), as well as the upper mentioned copy of *tetH* (Sulth_1188).

4 Conclusions

During bioleaching of chalcopyrite concentrate, *S. thermosulfidooxidans* but not *L. ferriphilum* maintained a low redox potential that is favorable for the extraction of copper. We hypothesize that this was due to differences in affinity and/or effectivity of the species' respective iron oxidation systems, as well as the attachment rate of the microorganisms to the mineral grains. This finding could potentially contribute to overcoming passivation and improving dissolution rates in large-scale chalcopyrite bioleaching. Expression of iron and sulfur oxidation systems in *S. thermosulfidooxidans* were investigated during bioleaching experiments in presence and absence of *L. ferriphilum*. Presence of the strong iron oxidizer induced greatly decreased transcript counts attributed to iron oxidation and increased counts for sulfur oxidation. Analysis of this data revealed genes products potentially responsible for the difference in ORP, which should be studied in this regard in the future.

In review

Additionally, this study underlines the importance of developing methods to control microbial populations in a bioleaching heap in order to exploit desired properties of selected microorganisms.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

SC conducted the laboratory experiments. MH performed bioinformatic analysis of transcript data. SB provided microscopic imaging. SC, SB, AB, ME, IP, WS, PW, AP, MD were involved in data analysis and biological interpretation of the results. SC drafted the manuscript and all authors contributed to its preparation.

In review

Funding

This project was supported by Vetenskapsrådet (contract 2014-6545), the Luxembourg National Research Fund (FNR, INTER/SYSAPP/14/05), Bundesministerium für Bildung und Forschung (BMBF, 031A600A and B), and the Swiss Initiative in Systems Biology (SystemsX.ch, SysMetEx) under the frame of ERASysAPP.

Acknowledgments

The authors acknowledge support from Science for Life Laboratory and the National Genomics Infrastructure for providing assistance in massive parallel sequencing and computational infrastructure. Uppsala Multidisciplinary Center for Advanced Computational Science is greatly acknowledged for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. Parts of the sequencing analysis presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

This is a provisional file, not the final typeset article

References

Baldi, F., Bralia, A., Riccobono, F., and Sabatini, G. (1991). Bioleaching of Cobalt and Zinc from Pyrite Ore in Relation to Calcitic Gangue Content. *World Journal for Microbiology and Biotechnology* 7(3), 298-308. doi: 10.1007/BF00329395.

Bellenberg, S., Buetti-Dinh, A., Galli, V., Ilie, O., Herold, M., Christel, S., et al. (2018). Automated Microscopical Analysis of Metal Sulfide Colonization by Acidophilic Microorganisms. *Applied and Environmental Microbiology*, AEM.01835-01818. doi: 10.1128/AEM.01835-18.

Blake, R.C., 2nd, and Griff, M.N. (2012). In Situ Spectroscopy on Intact *Leptospirillum Ferrooxidans* Reveals That Reduced Cytochrome 579 Is an Obligatory Intermediate in the Aerobic Iron Respiratory Chain. *Frontiers in Microbiology* 3, 136. doi: 10.3389/fmicb.2012.00136.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170.

Brierley, C.L., and Brierley, J.A. (2013). Progress in Bioleaching: Part B: Applications of Microbial Processes by the Minerals Industries. *Applied Microbiology and Biotechnology* 97(17), 7543-7552. doi: 10.1007/s00253-013-5095-3.

Bryner, L.C., and Jameson, A.K. (1958). Microorganisms in Leaching Sulfide Minerals. *Applied Microbiology* 6(4), 281-287.

Christel, S., Herold, M., Bellenberg, S., El Hajjami, M., Buetti-Dinh, A., Pivkin, I.V., et al. (2017). Multi-Omics Reveal the Lifestyle of the Acidophilic, Mineral-Oxidizing Model Species *Leptospirillum Ferriphilum*^{t}. *Applied and Environmental Microbiology*. doi: 10.1128/AEM.02091-17.

Coram, N.J., and Rawlings, D.E. (2002). Molecular Relationship between Two Groups of the Genus *Leptospirillum* and the Finding That *Leptosphillum Ferriphilum* Sp Nov Dominates South African Commercial Biooxidation Tanks That Operate at 40 Degrees C. *Applied and Environmental Microbiology* 68(2), 838-845. doi: Doi 10.1128/Aem.68.2.838-845.2002.

Cordoba, E.M., Munoz, J.A., Blazquez, M.L., Gonzalez, F., and Ballester, A. (2009). Passivation of Chalcopyrite During Its Chemical Leaching with Ferric Ion at 68 Degrees C. *Minerals Engineering* 22(3), 229-235. doi: 10.1016/j.mineng.2008.07.004.

Crundwell, F.K. (2015). The Semiconductor Mechanism of Dissolution and the Pseudo-Passivation of Chalcopyrite. *Canadian Metallurgical Quarterly* 54(3), 279-288. doi: 10.1179/1879139515y.0000000007.

Dahl, C., Engels, S., Pott-Sperling, A.S., Schulte, A., Sander, J., Lubbe, Y., et al. (2005). Novel Genes of the Dsr Gene Cluster and Evidence for Close Interaction of Dsr Proteins During Sulfur Oxidation in the Phototrophic Sulfur Bacterium Allochromatium Vinosum. *Journal of Bacteriology* 187(4), 1392-1404. doi: 10.1128/JB.187.4.1392-1404.2005.

Dopson, M., Baker-Austin, C., and Bond, P.L. (2005). Analysis of Differential Protein Expression During Growth States of *Ferroplasma* Strains and Insights into Electron Transport for Iron Oxidation. *Microbiology* 151(Pt 12), 4127-4137. doi: 10.1099/mic.0.28362-0.

Dopson, M., and Lindstrom, E.B. (1999). Potential Role of *Thiobacillus Caldus* in Arsenopyrite Bioleaching. *Applied and Environmental Microbiology* 65(1), 36-40.

In review

Guo, X., Yin, H., Liang, Y., Hu, Q., Zhou, X., Xiao, Y., et al. (2014). Comparative Genome Analysis Reveals Metabolic Versatility and Environmental Adaptations of *Sulfobacillus Thermosulfidooxidans* Strain St. *PLoS One* 9(6)**,** e99417. doi: 10.1371/journal.pone.0099417.

Hallberg, K.B., and Lindstrom, E.B. (1994). Characterization of *Thiobacillus Caldus* Sp. Nov., a Moderately Thermophilic Acidophile. *Microbiology* 140 (Pt 12)**,** 3451-3456. doi: 10.1099/13500872-140-12-3451.

Hedrich, S., Schlomann, M., and Johnson, D.B. (2011). The Iron-Oxidizing Proteobacteria. *Microbiology* 157(Pt 6)**,** 1551-1564. doi: 10.1099/mic.0.045344-0.

Hiroyoshi, N., Tsunekawa, M., Okamoto, H., Nakayama, R., and Kuroiwa, S. (2013). Improved Chalcopyrite Leaching through Optimization of Redox Potential. *Canadian Metallurgical Quarterly* 47(3)**,** 253-258. doi: 10.1179/cmq.2008.47.3.253.

Janosch, C., Remonsellez, F., Sand, W., and Vera, M. (2015). Sulfur Oxygenase Reductase (Sor) in the Moderately Thermoacidophilic Leaching Bacteria: Studies in *Sulfobacillus Thermosulfidooxidans* and *Acidithiobacillus Caldus*. *Microorganisms* 3(4)**,** 707-724. doi: 10.3390/microorganisms3040707.

Jeans, C., Singer, S.W., Chan, C.S., Verberkmoes, N.C., Shah, M., Hettich, R.L., et al. (2008). Cytochrome 572 Is a Conspicuous Membrane Protein with Iron Oxidation Activity Purified Directly from a Natural Acidophilic Microbial Community. *ISME Journal* 2(5)**,** 542-550. doi: 10.1038/ismej.2008.17.

Jerez, C.A. (2017). Biomining of Metals: How to Access and Exploit Natural Resource Sustainably. *Microbial Biotechnology* 10(5)**,** 1191-1193. doi: 10.1111/1751-7915.12792.

Johnson, D.B. (2014). Biomining-Biotechnologies for Extracting and Recovering Metals from Ores and Waste Materials. *Current Opinions in Biotechnology* 30**,** 24-31. doi: 10.1016/j.copbio.2014.04.008.

Karavaiko, G.I., Bogdanova, T.I., Tourova, T.P., Kondrat'eva, T.F., Tsaplina, I.A., Egorova, M.A., et al. (2005). Reclassification of '*Sulfobacillus Thermosulfidooxidans* Subsp. *Thermotolerans*' Strain K1 as *Alicyclobacillus Tolerans* Sp. Nov. And *Sulfobacillus Disulfidooxidans* Dufresne Et Al. 1996 as *Alicyclobacillus Disulfidooxidans* Comb. Nov., and Emended Description of the Genus *Alicyclobacillus*. *International Journal of Systematic and Evolutionary Microbiology* 55(Pt 2)**,** 941-947. doi: 10.1099/ijs.0.63300-0.

Kelly, D.P., Chambers, L.A., and Trudinge.Pa (1969). Cyanolysis and Spectrophotometric Estimation of Trithionate in Mixture with Thiosulfate and Tetrathionate. *Analytical Chemistry* 41(7)**,** 898-&. doi: DOI 10.1021/ac60276a029.

Khoshkhoo, M., Dopson, M., Engstrom, F., and Sandstrom, A. (2017). New Insights into the Influence of Redox Potential on Chalcopyrite Leaching Behaviour. *Minerals Engineering* 100**,** 9-16. doi: 10.1016/j.mineng.2016.10.003.

Khoshkhoo, M., Dopson, M., Shchukarev, A., and Sandstrom, A. (2014a). Chalcopyrite Leaching and Bioleaching: An X-Ray Photoelectron Spectroscopic (Xps) Investigation on the Nature of Hindered Dissolution. *Hydrometallurgy* 149**,** 220-227. doi: 10.1016/j.hydromet.2014.08.012.

Khoshkhoo, M., Dopson, M., Shchukarev, A., and Sandstrom, A. (2014b). Electrochemical Simulation of Redox Potential Development in Bioleaching of a Pyritic Chalcopyrite Concentrate. *Hydrometallurgy* 144**,** 7-14. doi: 10.1016/j.hydromet.2013.12.003.

In review

This is a provisional file, not the final typeset article

Klingenberg, H., and Meinicke, P. (2017). How to Normalize Metatranscriptomic Count Data for Differential Expression Analysis. *PeerJ* 5**,** e3859. doi: 10.7717/peerj.3859.

Langmead, B., and Salzberg, S.L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat Methods* 9(4)**,** 357-359. doi: 10.1038/nmeth.1923.

Li, Y., Kawashima, N., Li, J., Chandra, A.P., and Gerson, A.R. (2013). A Review of the Structure, and Fundamental Mechanisms and Kinetics of the Leaching of Chalcopyrite. *Advances in Colloid and Interface Science* 197-198**,** 1-32. doi: 10.1016/j.cis.2013.03.004.

Liao, Y., Smyth, G.K., and Shi, W. (2014). Featurecounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics* 30(7)**,** 923-930. doi: 10.1093/bioinformatics/btt656.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2. *Genome Biology* 15(12)**,** 550. doi: 10.1186/s13059-014-0550-8.

Mackintosh, M.E. (1978). Nitrogen-Fixation by *Thiobacillus Ferrooxidans*. *Journal of General Microbiology* 105(Apr)**,** 215-218. doi: Doi 10.1099/00221287-105-2-215.

Masaki, Y., Hirajima, T., Sasaki, K., Miki, H., and Okibe, N. (2018). Microbiological Redox Potential Control to Improve the Efficiency of Chalcopyrite Bioleaching. *Geomicrobiology Journal* 35(8)**,** 648-656. doi: 10.1080/01490451.2018.1443170.

Nakajima, T. (2015). Roles of Sulfur Metabolism and Rhodanese in Detoxification and Anti-Oxidative Stress Functions in the Liver: Responses to Radiation Exposure. *Medical Science Monitor* 21**,** 1721-1725. doi: 10.12659/MSM.893234.

Panda, S., Akcil, A., Pradhan, N., and Deveci, H. (2015). Current Scenario of Chalcopyrite Bioleaching: A Review on the Recent Advances to Its Heap-Leach Technology. *Bioresource Technology* 196**,** 694-706. doi: 10.1016/j.biortech.2015.08.064.

Penev, K., and Karamanev, D. (2010). Batch Kinetics of Ferrous Iron Oxidation by *Leptospirillum Ferriphilum* at Moderate to High Total Iron Concentration. *Biochemical Engineering Journal* 50(1-2)**,** 54-62. doi: 10.1016/j.bej.2010.03.004.

Petersen, J. (2016). Heap Leaching as a Key Technology for Recovery of Values from Low-Grade Ores – a Brief Overview. *Hydrometallurgy* 165**,** 206-212. doi: 10.1016/j.hydromet.2015.09.001.

Quatrini, R., and Johnson, D.B. (eds.). (2016). *Acidophiles: Life in Extremely Acidic Environments.* Caister Academic Press.

Rawlings, D.E., and Johnson, D.B. (2007). The Microbiology of Biomining: Development and Optimization of Mineral-Oxidizing Microbial Consortia. *Microbiology* 153(Pt 2)**,** 315-324. doi: 10.1099/mic.0.2006/001206-0.

Rawlings, D.E., Tributsch, H., and Hansford, G.S. (1999). Reasons Why '*Leptospirillum*'-Like Species Rather Than *Thiobacillus Ferrooxidans* Are the Dominant Iron-Oxidizing Bacteria in Many Commercial Processes for the Biooxidation of Pyrite and Related Ores. *Microbiology* 145 (Pt 1)**,** 5-13. doi: 10.1099/13500872-145-1-5.

Rohwerder, T., Gehrke, T., Kinzler, K., and Sand, W. (2003). Bioleaching Review Part A: Progress in Bioleaching: Fundamentals and Mechanisms of Bacterial Metal Sulfide Oxidation. *Applied Microbiology and Biotechnology* 63(3)**,** 239-248. doi: 10.1007/s00253-003-1448-7.

In review

Roume, H., Muller, E.E., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013). A Biomolecular Isolation Framework for Eco-Systems Biology. *ISME Journal* 7(1)**,** 110-121. doi: 10.1038/ismej.2012.72.

Temple, K.L., and Colmer, A.R. (1951). The Autotrophic Oxidation of Iron by a New Bacterium, *Thiobacillus Ferrooxidans*. *Journal of Bacteriology* 62(5)**,** 605-611.

Third, K.A., Cord-Ruwisch, R., and Watling, H.R. (2002). Control of the Redox Potential by Oxygen Limitation Improves Bacterial Leaching of Chalcopyrite. *Biotechnology and Bioengineering* 78(4)**,** 433-441.

Tsaplina, I.A., Krasil'nikova, E.N., Zakharchuk, L.M., Egorova, M.A., Bogdanova, T.I., and Karavaiko, G.I. (2000). Carbon Metabolism in *Sulfobacillus Thermosulfidooxidans* Subsp. *Asporogenes,* Strain 41. *Mikrobiologiia* 69(3)**,** 334-340.

Walden, G.H., Hammett, L.P., and Chapman, R.P. (1933). Phenanthroline-Ferrous Ion: A Reversible Oxidation—Reduction Indicator of High Potential and Its Use in Oxidimetric Titrations. *Journal of the American Chemical Society* 55(7)**,** 2649-2654. doi: 10.1021/ja01334a005.

Valdes, J., Pedroso, I., Quatrini, R., Dodson, R.J., Tettelin, H., Blake, R., 2nd, et al. (2008). *Acidithiobacillus Ferrooxidans* Metabolism: From Genome Sequence to Industrial Applications. *BMC Genomics* 9**,** 597. doi: 10.1186/1471-2164-9-597.

Wang, J., Gan, X.W., Zhao, H.B., Hu, M.H., Li, K.Y., Qin, W.Q., et al. (2016). Dissolution and Passivation Mechanisms of Chalcopyrite During Bioleaching: Dft Calculation, Xps and Electrochemistry Analysis. *Minerals Engineering* 98**,** 264-278. doi: 10.1016/j.mineng.2016.09.008.

Watling, H.R. (2006). The Bioleaching of Sulphide Minerals with Emphasis on Copper Sulphides - a Review. *Hydrometallurgy* 84(1-2)**,** 81-108. doi: 10.1016/j.hydromet.2006.05.001.

Vera, M., Schippers, A., and Sand, W. (2013). Progress in Bioleaching: Fundamentals and Mechanisms of Bacterial Metal Sulfide Oxidation--Part A. *Applied Microbiology and Biotechnology* 97(17)**,** 7529-7541. doi: 10.1007/s00253-013-4954-2.

Zhao, H.B., Wang, J., Tao, L., Cao, P., Yang, C.R., Qin, W.Q., et al. (2017). Roles of Oxidants and Reductants in Bioleaching System of Chalcopyrite at Normal Atmospheric Pressure and 45 Degrees C. *International Journal of Mineral Processing* 162**,** 81-91. doi: 10.1016/j.minpro.2017.04.002.

In review

This is a provisional file, not the final typeset article

Tables

Table 1 Excerpt of Supplementary Table 1 showing significant ($|\log2FC| \geq 1.0$, $p \leq 0.05$) differential expression of *S. thermosulfidooxidans* genes related to iron and sulfur oxidation as well as electron transport. Negative log2-fold changes indicate higher transcript in presence of *L. ferriphilum* (ASL), positive changes upregulation in its absence (AS). Mean expression values are calculated from three independent experiments (n=3). Abbreviations: std, standard deviation; log2FC, log2-fold change.

Gene ID	Product	Deseq normalized expression				log2FC
		AS mean	AS std	ASL mean	ASL std	
Iron oxidation and electron transport chain						
Sulth_0051	Cytochrome *c* assembly protein	1086	91	2412	150	-1.15
Sulth_0119	Cytochrome *c* class I	250	72	105	37	1.25
Sulth_0449	Heme/copper-type cytochrome/quinol oxidase, subunit 3	5850	537	919	85	2.67
Sulth_0450	Cytochrome *c* oxidase subunit I	15675	2453	2857	266	2.46
Sulth_0451	Cytochrome *c* oxidase subunit II	15243	1526	4700	545	1.70
Sulth_0453	Sulfocyanin (SoxE)	7722	884	623	112	3.63
Sulth_0488	Cytochrome *c* oxidase subunit I	17287	3212	533	106	5.02
Sulth_0489	Cytochrome *c* oxidase subunit II	12771	1543	405	58	4.98
Sulth_0494	Cytochrome *d* ubiquinol oxidase, subunit II	161	11	57	38	1.50
Sulth_0495	Cytochrome *bd* ubiquinol oxidase subunit I	355	102	39	10	3.17
Sulth_0840	Cytochrome *c* oxidase, cbb_3-type, subunit III	557	173	81	30	2.78
Sulth_0843	Heme/copper-type cytochrome/quinol oxidase, subunit 3	154	20	24	6	2.67
Sulth_0844	Cytochrome *c* oxidase subunit I	431	29	35	14	3.62
Sulth_0845	Cytochrome *c* oxidase subunit II	228	4	30	6	2.93
Sulth_1456	Cytochrome *c* oxidase subunit II, periplasmic domain	86	11	43	8	1.01
Sulth_1490	Cytochrome *c* oxidase, cbb3-type, subunit III	76	22	22	6	1.79
Sulth_1513	Cytochrome *c* oxidase subunit II	15255	1578	4733	457	1.69
Sulth_1514	Cytochrome *c* oxidase subunit I	34803	3976	16822	1585	1.05
Sulth_1901	Cytochrome *c* biogenesis protein	442	46	999	136	-1.18
Sulth_1930	Cytochrome *c* oxidase subunit IV	408	92	4081	275	-3.32
Sulth_1931	Cytochrome *c* oxidase subunit III	507	64	4862	339	-3.26
Sulth_1932	Cytochrome *c* oxidase subunit I	1427	269	15277	462	-3.42
Sulth_1933	Cytochrome *c* oxidase subunit II	1764	452	17994	1718	-3.35
Sulth_2044	Cytochrome *c* class I	91	31	18	9	2.36
Sulth_2183	Cytochrome *c* biogenesis protein transmembrane region	291	108	816	311	-1.49
Sulth_2568	Cytochrome *c*-type biogenesis protein CcmE	123	36	47	3	1.37
Sulth_2572	Cytochrome *c*-type biogenesis protein CcmB	68	22	14	3	2.28
Sulth_2573	Cytochrome *c* assembly protein	114	15	12	7	3.33
Sulth_2730	Cytochrome b/b_6 domain	819	19	238	110	1.78
Sulth_2731	Cytochrome b/b_6 domain protein	2148	652	804	64	1.42
Sulth_2749	Sulfocyanin (SoxE)	9756	2642	1112	151	3.13
Sulfur metabolism						

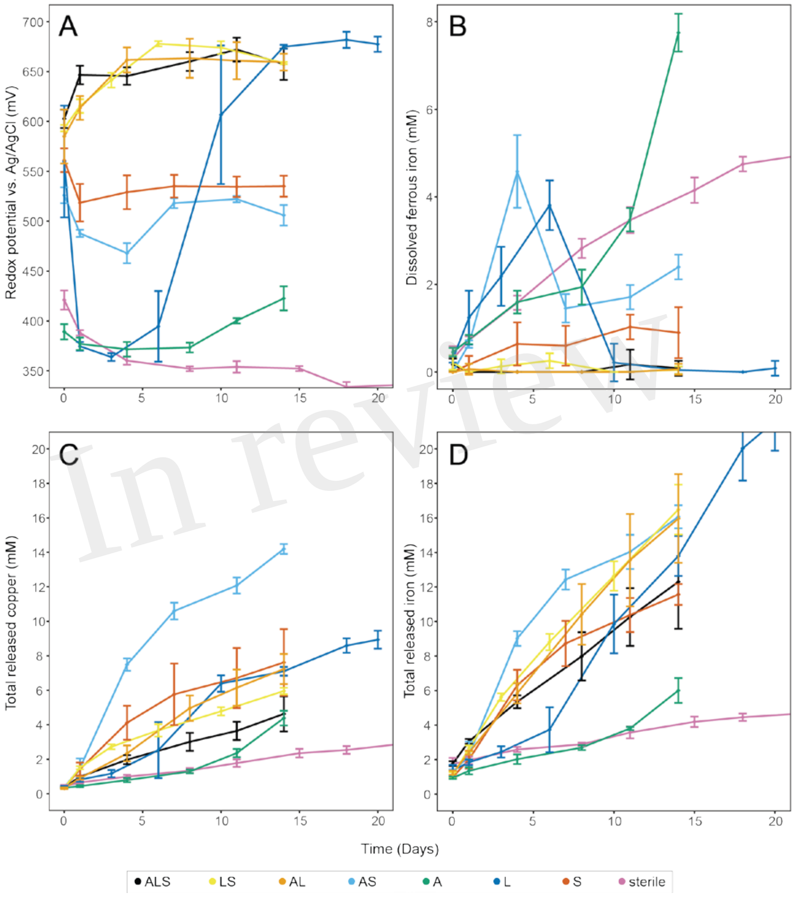
In review

Sulth_0921	Pyrrolo-quinoline quinone repeat-containing protein, tetH	613	101	25972	7210	-5.41
Sulth_0946	FAD-dependent pyridine nucleotide-disulfide oxidoreductase, Sqr_1	207	45	75	11	1.46
Sulth_1024	Hypothetical protein	125	50	57	8	1.13
Sulth_1025	Heterodisulfide reductase, subunit C, *hdrC*	24	0	50	8	-1.08
Sulth_1433	Sulfate adenylyltransferase	369	88	1430	384	-1.95
Sulth_1435	Sulfate adenylyltransferase	252	31	1202	289	-2.26
Sulth_1627	Sulfur oxygenase/reductase, Sor	591	29	1340	319	-1.18
Sulth_1878	Rhodanese-like protein	176	36	381	49	-1.12
Sulth_2076	Rhodanese-like protein	203	23	416	37	-1.03
Sulth_2172	Rhodanese-like protein	3024	1076	12511	2067	-2.05
Sulth_2770	Heterodisulfide reductase, subunit C, *hdrC*	11592	2924	29882	1777	-1.37
Sulth_2771	Heterodisulfide reductase, subunit B, *hdrB*	13422	4935	28681	4989	-1.10
Sulth_2782	DsrE family protein	4123	1767	14140	4441	-1.78
Sulth_3251	Pyrrolo-quinolinequinone repeat-containing protein, *tetH*	163	5	1551	223	-3.25

In review

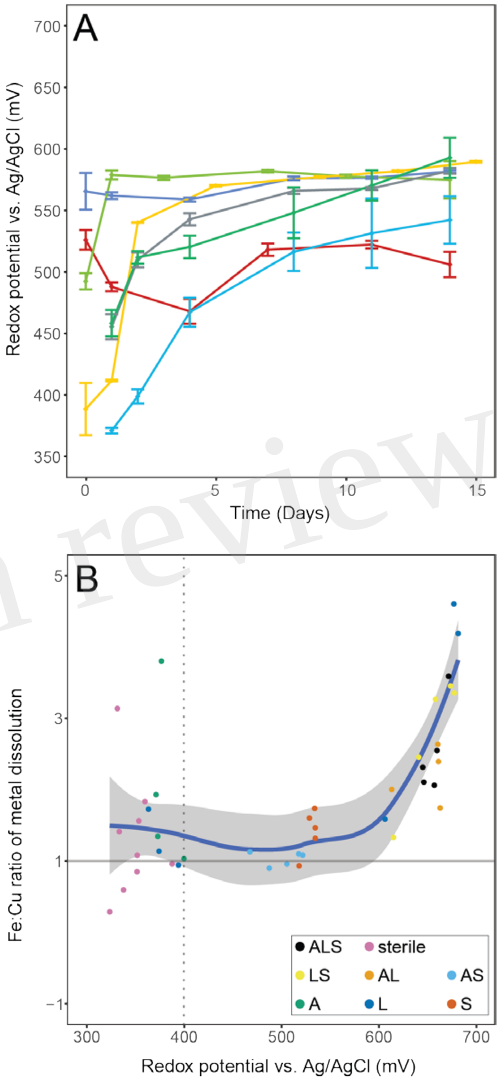
This is a provisional file, not the final typeset article

Figures



In review

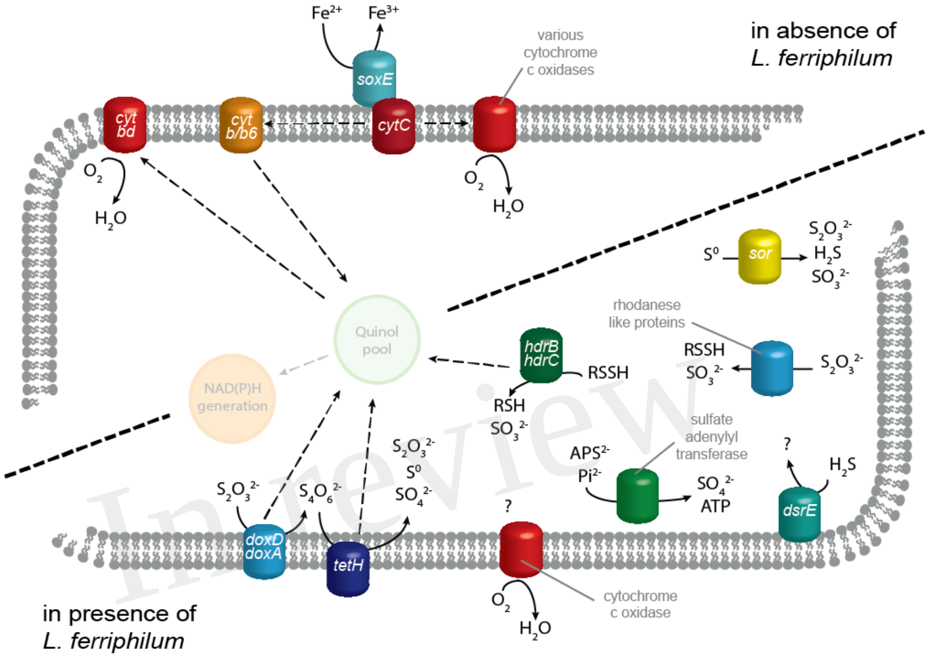
Figure 1. Bioleaching of chalcopyrite concentrate with single species, binary, and tertiary combinations of the three studied model species, plus unioculated control. The panels show redox potential (A), dissolved Fe^{2+} (B), and total released copper and iron (C and D, respectively). Data points represent means ± standard deviations (*n* = 4). Abbreviations in the legend denote: A = *A. caldus*, L = *L. ferriphilum*, and S = *S. thermosulfidooxidans*.



In review

Figure 2. Panel A shows redox potentials remaining below 600 mV during seven independent experiments containing only *S. thermosulfidooxidans* for microbial iron oxidation. Data points represent means ± standard deviations (n = 4). Panel B illustrates the correlation of the ratio of released iron:copper versus redox potential during bioleaching of chalcopyrite concentrate with various combinations of the three model species. The ratio was calculated by dividing the amounts of the two metals that were released between two consecutive sampling points during the leaching experiment. The regression was calculated using to the LOESS method with 95% confidence interval marked by the shaded area. The dotted line denotes the onset of microbial iron oxidation indicated by a redox potential above 400 mV. Abbreviations in the legend denote: A = *A. caldus*, L = *L. ferriphilum*, and S = *S. thermosulfidooxidans*.

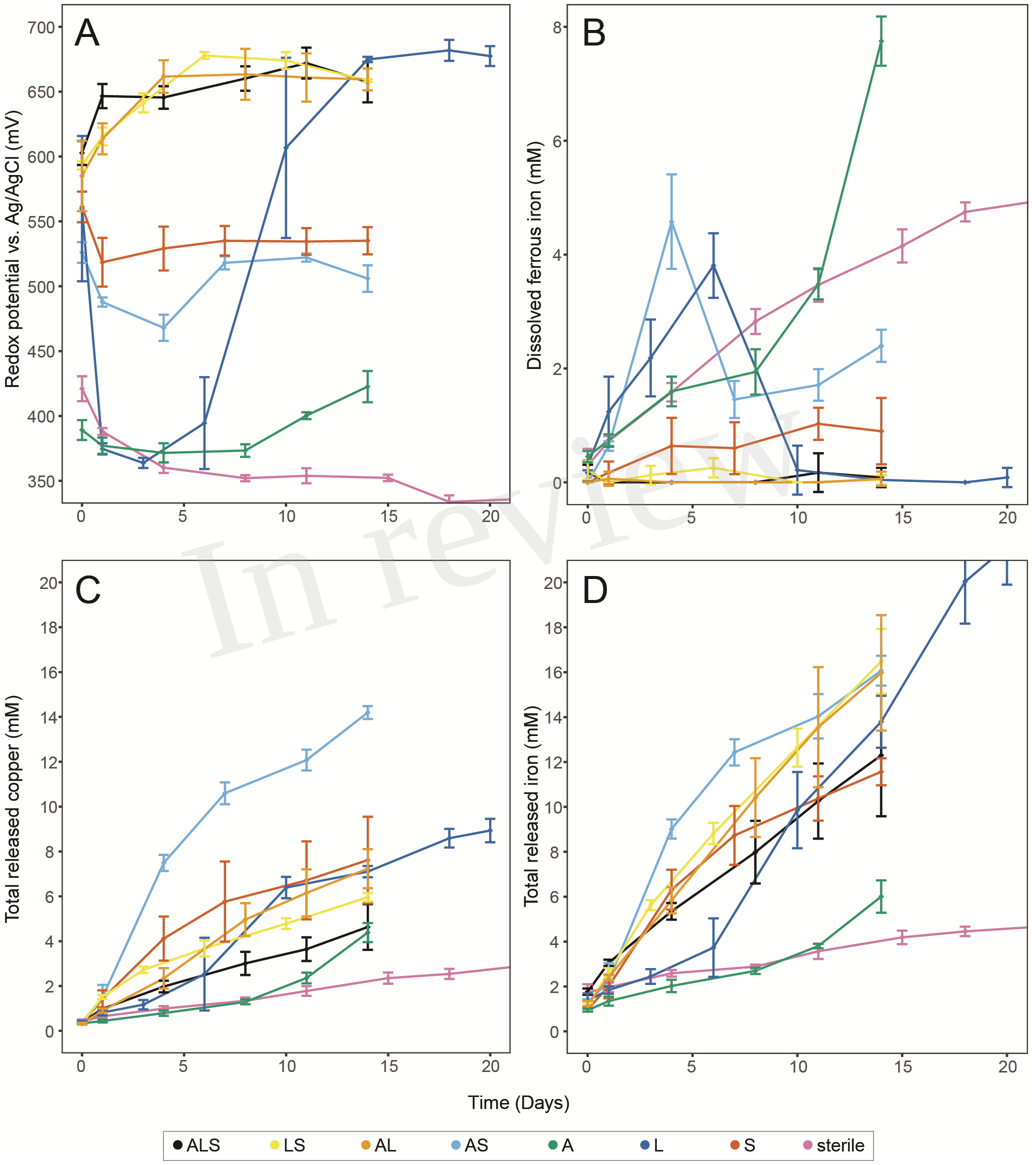
In review



In review

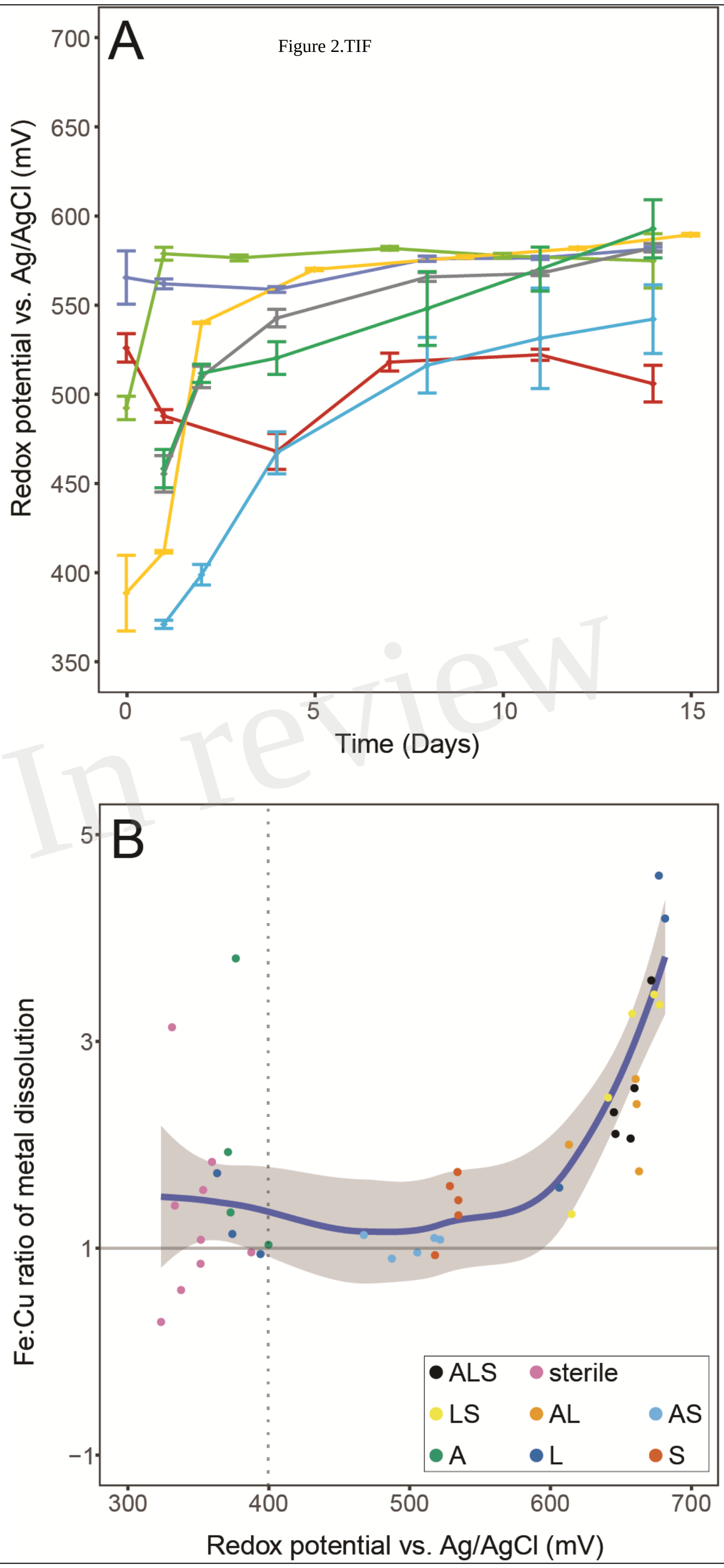
Figure 3 Proposed model of *S. thermosulfidooxidans* transcript regulation of genes related to energy generation. Iron oxidation systems and electron transport by cytochromes has a greater number of RNA transcripts in the absence of the strong iron oxidizer *L. ferriphilum*. In its presence, *S. thermosulfidooxidans* instead has higher transcript numbers for genes contributing to ISC oxidation plus one cytochrome *c* oxidase complex. Quinone pool and NAD(P)H generation are depicted translucently for comprehension, but corresponding genes were not analyzed in this study. Solid arrows represent metabolic reactions while dashed arrows indicate the relocation of electrons.

Figure 1.TIF



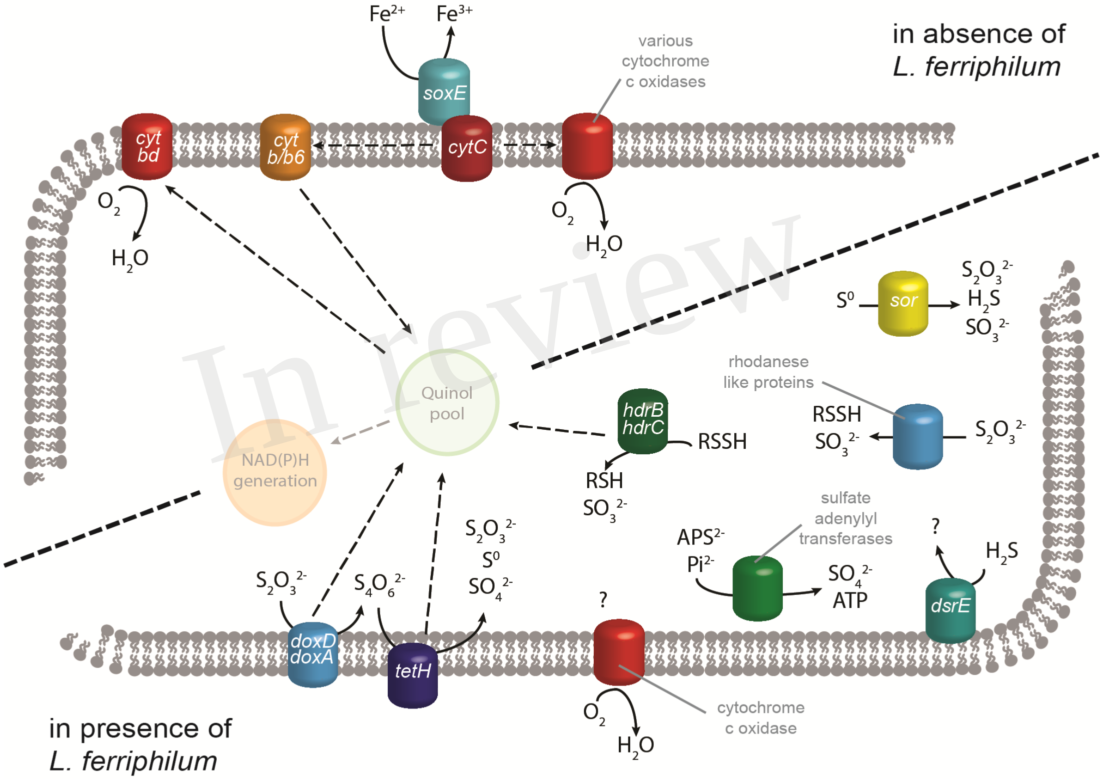
In review

Figure 2.TIF



In review

Figure 3.TIF



In review

C.4 Automated microscopical analysis of metal sulfide colonization by acidophilic microorganisms.

Sören Bellenberg[†], Antoine Buetti-Dinh[†], Vanni Galli, Olga Ilie, **Malte Herold**, Stephan Christel,
Mariia boretska, Igor Pivkin, Paul Wilmes, Wolfgang Sand, Mario Vera, Mark Dopson

2018
Applied and Environmental Microbiology **84**: e01835.
DOI: 10.1128/AEM.01835-18

Contributions of author include:

- Data analysis
- Writing and revision of manuscript

[†] Co-first author

AMERICAN SOCIETY FOR MICROBIOLOGY Applied and Environmental Microbiology®

METHODS

Automated Microscopic Analysis of Metal Sulfide Colonization by Acidophilic Microorganisms

Sören Bellenberg,[a] Antoine Buetti-Dinh,[b,c] Vanni Galli,[d] Olga Ilie,[b,c] Malte Herold,[e] Stephan Christel,[f] Mariia Boretska,[a] Igor V. Pivkin,[b,c] Paul Wilmes,[e] Wolfgang Sand,[a,g,h] Mario Vera,[i] Mark Dopson[f]

[a]Fakultät für Chemie, Biofilm Centre, Universität Duisburg-Essen, Essen, Germany
[b]Institute of Computational Science, Faculty of Informatics, Università della Svizzera Italiana, Lugano, Switzerland
[c]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[d]Institute for Information Systems and Networking, University of Applied Sciences of Southern Switzerland, Manno, Switzerland
[e]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg
[f]Centre for Ecology and Evolution in Microbial Model Systems, Linnaeus University, Kalmar, Sweden
[g]Donghua University, Shanghai, People's Republic of China
[h]Technische Universität Bergakademie Freiberg, Freiberg, Germany
[i]Institute for Biological and Medical Engineering, Schools of Engineering, Medicine & Biological Sciences, Department of Hydraulic & Environmental Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

ABSTRACT Industrial biomining processes are currently focused on metal sulfides and their dissolution, which is catalyzed by acidophilic iron(II)- and/or sulfur-oxidizing microorganisms. Cell attachment on metal sulfides is important for this process. Biofilm formation is necessary for seeding and persistence of the active microbial community in industrial biomining heaps and tank reactors, and it enhances metal release. In this study, we used a method for direct quantification of the mineral-attached cell population on pyrite or chalcopyrite particles in bioleaching experiments by coupling high-throughput, automated epifluorescence microscopy imaging of mineral particles with algorithms for image analysis and cell quantification, thus avoiding human bias in cell counting. The method was validated by quantifying cell attachment on pyrite and chalcopyrite surfaces with axenic cultures of *Acidithiobacillus caldus*, *Leptospirillum ferriphilum*, and *Sulfobacillus thermosulfidooxidans*. The method confirmed the high affinity of *L. ferriphilum* cells to colonize pyrite and chalcopyrite surfaces and indicated that biofilm dispersal occurs in mature pyrite batch cultures of this species. Deep neural networks were also applied to analyze biofilms of different microbial consortia. Recent analysis of the *L. ferriphilum* genome revealed the presence of a diffusible soluble factor (DSF) family quorum sensing system. The respective signal compounds are known as biofilm dispersal agents. Biofilm dispersal was confirmed to occur in batch cultures of *L. ferriphilum* and *S. thermosulfidooxidans* upon the addition of DSF family signal compounds.

IMPORTANCE The presented method for the assessment of mineral colonization allows accurate relative comparisons of the microbial colonization of metal sulfide concentrate particles in a time-resolved manner. Quantitative assessment of the mineral colonization development is important for the compilation of improved mathematical models for metal sulfide dissolution. In addition, deep-learning algorithms proved that axenic or mixed cultures of the three species exhibited characteristic biofilm patterns and predicted the biofilm species composition. The method may be extended to the assessment of microbial colonization on other solid particles and may serve in the optimization of bioleaching processes in laboratory scale experi-

Received 26 July 2018 **Accepted** 31 July 2018
Accepted manuscript posted online 3 August 2018
Citation Bellenberg S, Buetti-Dinh A, Galli V, Ilie O, Herold M, Christel S, Boretska M, Pivkin IV, Wilmes P, Sand W, Vera M, Dopson M. 2018. Automated microscopic analysis of metal sulfide colonization by acidophilic microorganisms. Appl Environ Microbiol 84:e01835-18. https://doi.org/10.1128/AEM.01835-18.
Editor Frank E. Löffler, University of Tennessee and Oak Ridge National Laboratory

Copyright © 2018 American Society for Microbiology. All Rights Reserved.

Address correspondence to Sören Bellenberg, soeren.bellenberg@uni-due.de, or Antoine Buetti-Dinh, antoine.buetti.dinh@usi.ch.
S.B. and A.B.-D. contributed equally to this work.

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

ments with industrially relevant metal sulfide concentrates. Furthermore, the method was used to demonstrate that DSF quorum sensing signals directly influence colonization and dissolution of metal sulfides by mineral-oxidizing bacteria, such as *L. ferriphilum* and *S. thermosulfidooxidans*.

KEYWORDS bioleaching, biofilm formation, biofilm dispersal, image analysis, microbe-mineral interaction, quorum sensing, diffusible soluble factor, biofilms, fluorescent image analysis, microbe-mineral interactions

The dissolution of metal sulfides is a chemical process catalyzed by the microbial oxidation of iron(II) ions and inorganic sulfur compounds (ISCs). It leads to the generation of acidic, sulfate, and heavy-metal laden acid mine drainage (AMD) waters. Mineral-attached microorganisms are crucial for the mineral breakdown (1) and are industrially exploited for the recovery of valuable metals from sulfide ores in biomining processes (2, 3). Although the mechanism of metal sulfide oxidation is an indirect chemical process (4, 5), contact of mineral-oxidizing microbes with metal sulfides may significantly increase dissolution kinetics. This is at least partially due to glucuronic acid residues in the extracellular polymeric substances (EPS) of *Acidithiobacillus ferrooxidans* and *Leptospirillum ferrooxidans* that accumulate the oxidative agent iron(III) ions (6, 7). The presence of biofilms is especially important for the persistence of active bioleaching microorganisms in commercial heap bioleaching operations (2, 8, 9). In addition, mineral-attached cells are particularly important for initiation of the metal sulfide dissolution. For instance, at dissolved iron ion concentrations of <200 mg/liter, mineral-attached cells of iron(II)- and ISC-oxidizing *Acidithiobacillus ferrooxidans* or *Acidithiobacillus ferrivorans* on pyrite surfaces are exclusively responsible for catalyzing its dissolution (10). Consequently, cell attachment to metal sulfides has been extensively studied (10–16). We compare metal sulfide colonization by the ISC oxidizer *Acidithiobacillus caldus*, the iron(II) oxidizer *Leptospirillum ferriphilum*, and the ISC- and iron(II)-oxidizing species *Sulfobacillus thermosulfidooxidans*.

Several methods for the assessment of mineral-attached cells on pyrite or chalcopyrite have been developed. Indirect microscopic cell counts rely on the decrease of planktonic cells during the initial contact with metal sulfides. However, this method cannot assess the temporal development of the mineral-attached cell population for prolonged cultivation periods. Other methods involve a cell detachment step, although quantitative separation of cells from the mineral is not possible and is prone to biases due to the release of fine mineral particles when samples are rigorously mixed. Molecular methods and microcalorimetry are alternative options for quantification and characterization of cells on mineral surfaces and were compared in a recent study (17). Quantitative PCR assays are currently the most reliable and common method for absolute quantification of attached cells in a species-specific manner, although DNA extraction from mineral samples has specific biases, such as differential susceptibility of different cell types to cell lysis, as well as interferences of iron ions with the remaining nucleic acids. In addition, intact cells have been found on chalcopyrite and pyrite mineral grains after aggressive chemical extraction methods, such as hot cell lysis and phenol treatment (see Fig. S1 in the supplemental material) (18). Epifluorescence microscopy (EFM) can be used to study the number of attached cells, as well as the structure of the biofilm (19). However, model systems that employ polished mineral coupons are not comparable with fine-ground mineral particles. Microbial metal sulfide colonization is generally heterogeneous, as particles devoid of bacterial colonization coexist with well-colonized surfaces. This requires the analysis of a sufficiently large number of particles to take into account random variation in mineral grain colonization.

In addition to quantitative information on mineral colonization, biofilm structures can also be investigated by using computational methods. Deep neural networks are the algorithms underlying "deep learning," a method broadly used in areas of computer vision, for instance, to analyze and classify images. Popular examples are object recognition with smartphones and self-driving cars. Several tools exist for processing

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

microscopy images and extracting relevant biological features that include cell or nucleus counting and eukaryotic phenotype analysis. These tools are based on open-source (R, EBImage [20]; and Java, ImageJ [21, 22]) or proprietary (MatLab, CellProfiler [23] and CellClassifier [24]) programming languages. The automated image analysis allows processing of many images with large numbers of mineral grains to be analyzed, reducing time for replication and therefore allowing testing and comparisons of multiple experimental conditions in a relative manner.

Automated image analyses could provide insights into aspects of biofilm development that are not yet fully understood. As such, the temporal dynamics of mineral colonization in acidophilic bacteria are largely unknown. However, it is known that the initial colonization of metal sulfide surfaces by *A. ferrooxidans*, *A. ferrivorans*, *L. ferrooxidans*, and *Acidiferrobacter* sp. strain SPIII/3 is influenced by quorum sensing (QS) signal compounds, such as *N*-acyl-homoserine lactones (25, 26). Those compounds are not produced by the three strains used in this study. However, genes encoding a diffusible soluble factor (DSF) QS system have recently been described for the *L. ferriphilum*^{T} (27). DSF family QS signal compounds synchronize virulence and biofilm dispersal in *Xanthomonas campestris* (*cis*-11-methyl-dodecenoic acid, termed DSF) and *Burkholderia cenocepacia* (*cis*-2-dodecenoic acid, termed BDSF). These compounds are also known to disperse biofilms. Pronounced interspecies biofilm dispersal effects are associated with DSF family signaling (28, 29). DSF QS systems are encoded by the *rpfCFGR* genes in those species where RpfF is the signaling compound synthase, while the corresponding two-component signal recognition system consists of the sensor kinase RpfC and the response regulator RpfG, which act directly on cyclic diguanylate (c-di-GMP) metabolism. In addition, the DSF signal receptor proteins homologous to the RpfR protein are known to become active c-di-GMP-hydrolyzing phosphodiesterases upon binding of DSF family signals. Lowered levels of c-di-GMP are typically associated with enhanced motility and decreased expression of biofilm-related genes (30).

In this study, we used a motorized EFM for automated image acquisition (Fig. 1) coupled to automated image analysis using algorithms that allowed quantification of mineral-attached cells (Fig. 2). In addition, we used deep neural network algorithms for classification of images based on species-specific biofilm patterns in samples with low microbial diversity. This methodology provides the possibility to assess directly microbial mineral colonization laboratory bioleaching assays of metal sulfide concentrate ores. We demonstrate that the method is suitable to follow the temporal development of biofilms in model cultures of *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans* in chalcopyrite and pyrite bioleaching assays. Furthermore, biofilm dispersal upon the addition of DSF molecules to biofilms formed by *L. ferriphilum* and *S. thermosulfidooxidans* is suggested to occur.

RESULTS

Automated image analysis for monitoring biofilm on mineral grains. Specimens with mineral grains for microscopy were prepared in a manner to achieve images with a minimum mineral coverage of 70%. For the assessment of the mineral grain colonization, images were grouped into four arbitrarily chosen equally large groups of one, two, nine, 18, 36, or 72 images. This was done in order to average the naturally nonhomogeneous mineral colonization in single microscopy images over a more representative mineral surface area in multiple images. The variation of the amount of images per group showed that the coefficient of variation of the mean values of each of the four groups decreased in a linear manner from 25% ± 10% to 8% ± 2.5% when two images per group (eight images in total) or 18 images per group (72 images in total) were considered from the same biological sample (Fig. 3).

Mineral colonization data of every sample of mineral grains were derived from analysis of at least 36 images, corresponding to a coefficient of variation not larger than 16% ± 8%. For a hypothetical average mineral coverage of 75% of each image, this consideration corresponded to an analyzed top-view mineral surface area of 4.6 mm^2. This can be deduced to be the minimum mineral surface area that should be analyzed for

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

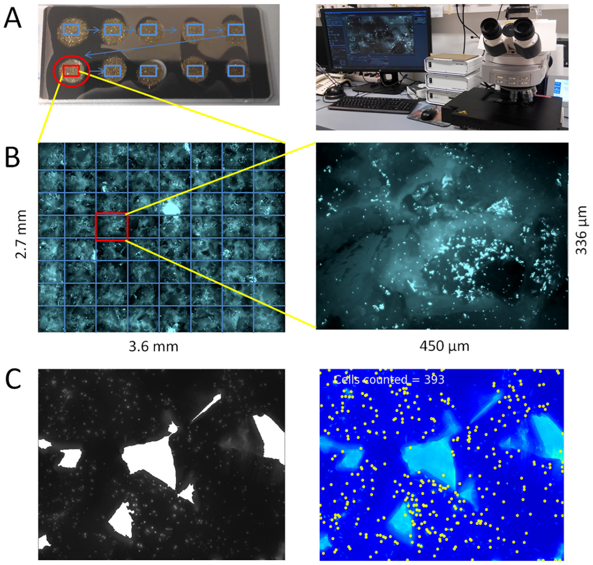


FIG 1 Experimental setup for automated imaging and mounting of mineral grain samples. (A) Ten-well diagnostic glass slides were used for spotting mineral samples in mounting medium. (B) Stack images were recorded using a motorized epifluorescence microscope, for calculation of extended depth of focus image projections. (C) Determination of the mineral grain area (left) and cell counting (right) is illustrated. Detected cell counts are indicated by a yellow circle for generation of a report file.

colonization assessment of pyrite or chalcopyrite concentrate particles in the size range of 50 to 100 μm in order to achieve a coefficient of variation not larger than 16% ± 8%. Figure 3 shows that a higher accuracy of the method was achieved with more analyzed images, as the coefficient of variation fell below 10% with >80 analyzed images per sample.

***L. ferriphilum* efficiently colonizes pyrite and chalcopyrite surfaces.** Axenic cultures of *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans* were compared regarding their ability to colonize pyrite and chalcopyrite (Fig. 4). The inocula were not previously adapted by growth on pyrite or chalcopyrite or to the presence of copper ions. *L. ferriphilum* significantly outperformed *A. caldus* and *S. thermosulfidooxidans* in its capacity to attach on the minerals and was estimated to have $1.5 \times 10^{-9} \pm 6.2 \times 10^{-7}$ or $1.2 \times 10^{-9} \pm 5.0 \times 10^{-7}$ cells · g^{-1} in chalcopyrite or pyrite cultures, respectively, averaging the highest levels of mineral colonization on days 14 and 21. The corresponding values for *A. caldus* and *S. thermosulfidooxidans* were $4.4 \times 10^{-8} \pm 7.3 \times 10^{-7}$ or $4.8 \times 10^{-8} \pm 10^{-8}$ and $3.1 \times 10^{-8} \pm 4.8 \times 10^{-7}$ or $3.1 \times 10^{-8} \pm 4.5 \times 10^{-7}$ cells · g^{-1} in chalcopyrite or pyrite cultures, respectively. Student's *t* tests showed that the difference is statistically significant ($P < 10^{-4}$) between groups made of colonization data from 72 individual images (36 images from day 14 and 36 images from day 21 samples) of each of the mineral cultures of *L. ferriphilum*, *A. caldus*, and *S. thermosulfidooxidans*. In addition, *L. ferriphilum* was most effective in dissolution of pyrite or chalcopyrite in axenic batch experiments. This was reflected by the release of iron and for chalcopyrite copper ions (Fig. 5). In the case of chalcopyrite cultures, this difference was also represented by the development of the total cell numbers (Fig. 4A). For the pyrite cultures, *S. thermosulfidooxidans* showed the highest total cell numbers, likely due to its ability to utilize ISCs and iron(II) ions (31). On the one hand, ISCs are not used by the obligate iron(II) oxidizer *L. ferriphilum*, and due to its inability to oxidize iron(II)

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

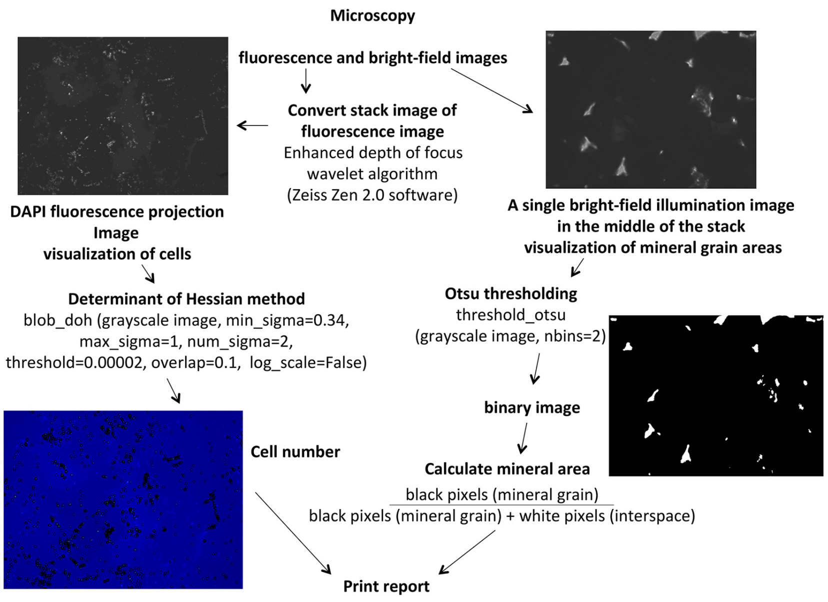

Microscopy
fluorescence and bright-field images
Convert stack image of
fluorescence image
Enhanced depth of focus
wavelet algorithm
(Zeiss Zen 2.0 software)
DAPI fluorescence projection
Image
visualization of cells
A single bright-field illumination image
in the middle of the stack
visualization of mineral grain areas
Determinant of Hessian method
blob_doh (grayscale image, min_sigma=0.34,
max_sigma=1, num_sigma=2,
threshold=0.00002, overlap=0.1, log_scale=False)
Otsu thresholding
threshold_otsu
(grayscale image, nbins=2)
binary image
Cell number
Calculate mineral area
black pixels (mineral grain)
black pixels (mineral grain) + white pixels (interspace)
Print report

FIG 2 Illustration of the Python image analysis algorithm for quantification of attached cells on mineral grains.

ions, *A. caldus* was unable to grow on pyrite. *A. caldus* and *S. thermosulfidooxidans* formed a lesser but detectable biofilm on both minerals (Fig. 4B). However, the initial colonization of chalcopyrite by cells of *S. thermosulfidooxidans* was significantly lower than that by *A. caldus* (Fig. 4B1).

In general, the development of the mineral-attached cell fractions in axenic cultures of all three strains clearly showed mineral-dependent differences. In chalcopyrite cultures, an initial peak of 45 to 78% attached cells was followed by a rapid decline within the first 10 days of cultivation to a level of 25 to 40% for all strains (Fig. 4C1). Interestingly, this peak in the percentage fraction of mineral-attached cells was the

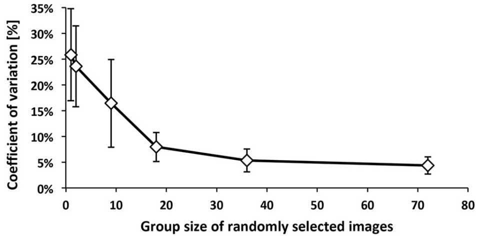


FIG 3 Development of the coefficient of variation with the amount of analyzed images. The coefficient of variation was calculated using mean values of mineral colonization data. Colonization data of the individual images were randomly binned into four arbitrarily chosen groups. The group size was varied from colonization values derived from 1, 2, 9, 18, 36, or 72 images from a data set of 300 images from the same mineral sample condition (mixed culture of *A. caldus* and *S. thermosulfidooxidans* after 12 days of cultivation on chalcopyrite). The colonized mineral stems from a biological triplicate experiment. The coefficient of variation among the groups was calculated repetitively with randomized selection of the colonization data 25 times in order to calculate the standard deviation of the coefficient of variation.

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

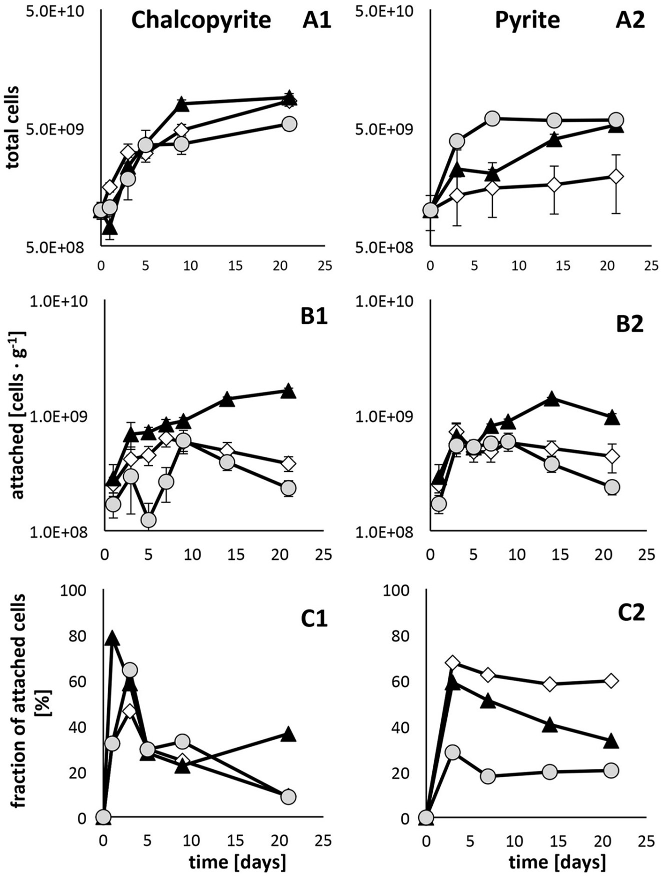


FIG 4 *L. ferriphilum* efficiently colonizes chalcopyrite and pyrite surfaces. (A to C) The temporal development of total cell numbers (A), mineral-attached cell per gram of metal sulfide mineral (B), and the fraction of the mineral-attached cell population of the total cell population (C) were compared in 150-ml cultures of axenic cultures of *A. caldus* (white diamonds), *L. ferriphilum* (black triangles), or *S. thermosulfidooxidans* (gray circles) containing 2% chalcopyrite grains (1) or 2% pyrite grains (2) of 50- to 100-μm grain size.

result of the initial mineral colonization, followed by growth of the planktonic cell population rather than detachment of biofilm cells. This finding is supported by the fact that the amount of mineral-attached cells did not decrease significantly during the respective time period (Fig. 4B1). Also, the total cell numbers per assay kept rising steadily from 10^9 cells to at least 5×10^{-9} cells during the duration of the experiment for cultures of all three strains (Fig. 4A1). In the case of pyrite, the initial peak in the fraction of attached cells was less pronounced.

The fraction of attached *L. ferriphilum* cells in chalcopyrite cultures remained stable at 25 to 35% after the first 5 days of incubation, even though the amount of mineral-attached cells increased from $8.3 \times 10^{-8} \pm 8.1 \times 10^{-7}$ on day seven to $1.6 \times 10^{-9} \pm 7.2 \times 10^{-7}$ cells · g^{-1} chalcopyrite on day 21. *A. caldus* and *S. thermosulfidooxidans* showed a different behavior, as after the first 10 days of incubation, the amount of

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

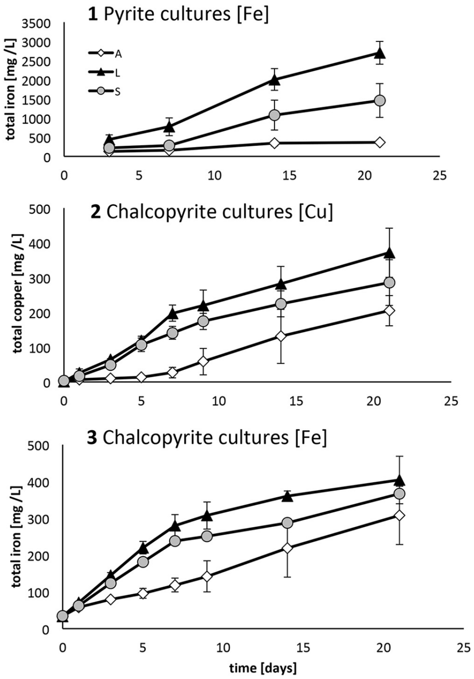


FIG 5 Dissolution of pyrite or chalcopyrite indicates microbial growth in bioleaching assays. Axenic cultures of *A. caldus* (A, white diamonds), *L. ferriphilum* (L, black triangles), and *S. thermosulfidooxidans* (S, gray circles) were cultivated with pyrite (1) or chalcopyrite (2 and 3), as described. The development of the total iron and copper ion concentrations is shown.

attached cells decreased slightly to $4.4 \times 10^{-8} \pm 1.2 \times 10^{-8}$ and $2.4 \times 10^{-8} \pm 3.3 \times 10^{-7}$ cells · g^{-1} on day 21, respectively. During this time, their percent fractions of attached cells gradually decreased to circa 10% (Fig. 4C1).

In the case of pyrite bioleaching, the fraction of mineral-attached cells in cultures of *L. ferriphilum* averaged over the time from day five until the end of the experiment on day 21 (Fig. 4C2) was ~40 to 60% enhanced in comparison to the levels observed in chalcopyrite assays (25 to 35%). A similar observation was made for *A. caldus* (60 to 70% attached cells in pyrite assays compared to circa 10 to 30% in chalcopyrite assays), while *S. thermosulfidooxidans* showed the lowest pyrite colonization efficiency with a fraction of 20 to 30% attached cells (10 to 30% in chalcopyrite assays).

Deep neural networks can identify characteristic biofilm patterns on chalcopyrite in axenic and mixed cultures. Deep neural networks trained on 600 microscopy images per experimental category were used to test their performance in recognizing cell attachment patterns on chalcopyrite grains. Samples from cultures with different inoculum compositions of *A. caldus* (A), *L. ferriphilum* (L), and *S. thermosulfidooxidans* (S) were used as pure or mixed cultures, resulting in the following categories: A, L, S, AS, LS, and ASL. These categories represent the biofilms formed on chalcopyrite grains after 5 days of incubation. A set of 100 test images per category not included in the training set were used to test the ability of the deep neural network to assign test images to one of the training set categories. Under the restrictions that only low-species-abundance samples are considered and individual training sets are available for

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

TABLE 1 Deep learning prediction of species composition of mineral-attached cell populations

	Probability (%) by predicted class[a]					
Actual class	**A**	**L**	**S**	**AS**	**LS**	**ALS**
A	96	0	3	1	0	0
L	0	94	0	1	0	5
S	2	0	93	3	0	2
AS	0	1	2	78	14	5
LS	1	0	0	11	84	4
ALS	0	0	3	0	1	96

[a] Probabilities (%) were assigned by the deep learning analysis for the similarity of the 100 test set images to the convolutional neural network (CNN) class prediction. CNNs were trained with 600 images from five-day-old mineral cultures with different inoculum compositions of *A. caldus* (A), *L. ferriphilum* (L), and *S. thermosulfidooxidans* (S) that were used as pure or mixed cultures, resulting in the following categories: A, L, S, AS, LS, and ASL.

each of the three species in axenic and mixed cultures, the technique allows the prediction of the microbial species present within a mixed-species biofilm on chalcopyrite samples (Table 1).

Expression of the DSF family quorum sensing system in *L. ferriphilum*. A DSF synthase was found encoded in the *L. ferriphilum* genome (Table 2) (27). Genes likely encoding DSF family signal-specific two-component systems or response regulators, suitable for DSF signal perception, were identified in the genomes of *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans* (Table 2). The genes of the *L. ferriphilum* DSF QS system were found to be expressed in transcriptome analyses of cells grown in continuous cultures, as well as in chalcopyrite batch cultures. The DSF synthase LFTS_0514 was especially found to have high expression levels in the planktonic cell subpopulations. Those levels strongly exceeded the average expression of gene transcripts of this species in axenic, but also in mixed, cultures with *S. thermosulfidooxidans* (Fig. S2).

DSF and BDSF signal compounds inhibit iron(II) oxidation and chalcopyrite dissolution. A strong inhibitory effect on the metabolic activity of bioleaching bacteria was observed after the external addition of DSF or BDSF. These compounds prevented oxidation of the soluble substrates iron(II) ions and tetrathionate (Fig. S3) or the insoluble substrate chalcopyrite during a cultivation period of 32 days (Fig. S4), when 5 μM DSF or BDSF signal molecules were added simultaneously with the inoculum into cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (Table 3). No effect of DSF or BDSF addition on soluble substrate oxidation was observed in tetrathionate cultures of *A. caldus*, while growth with chalcopyrite and its dissolution were inhibited by the addition of 5 μM DSF (Table 3).

Computational image analysis detects biofilm dispersal upon addition of DSF family signaling compounds. Biofilm dispersal was observed in cultures of *L. ferriphilum*, *S. thermosulfidooxidans*, and their combination in mixed cultures when 5 μM DSF was added after 5 days of incubation (Fig. 6). A similar effect was noted in mixed cultures of all three species (Fig. S5). In contrast, no biofilm dispersal was observed in

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

TABLE 2 Presence of DSF family QS system-encoding genes in *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans* genomes identified using BLASTP[a]

	DSF quorum sensing system genes			
Species (reference and/or accession no.)	***rpfF***	***rpfR***	***rpfC***	***rpfG***
Acidithiobacillus caldus^T (59) (GCA_000175575.2)		ACAty_RS14920, ACAty_RS14615, ACAty_RS02860	ACAty_RS07245, ACAty_RS04080	
Leptospirillum ferriphilum^T (27) (GCA_900198525.1)	LFTS_00514	LFTS_00511	LFTS_00515, LFTS_00516	LFTS_00517
Sulfobacillus thermosulfidooxidans^T (GCA_900176145.1)		Sulth_1253, Sulth_1788, Sulth_2384	Sulth_1793	Sulth_2102

[a] E value ($<10^{-30}$) (48).

TABLE 3 Inhibitory effect of 5 μM DSF or BDSF addition on oxidation of soluble and insoluble energy sources in cultures of *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans*

	Inhibition by species[a]		
Energy source (reference figure)	***A. caldus***	***L. ferriphilum***	***S. thermosulfidooxidans***
Soluble [tetrathionate/iron(II) ions] (Fig. S3)	−	+	+
Insoluble (chalcopyrite) (Fig. S4)	+	+	+

[a] +, no biological oxidation of soluble substrates occurred within 32 days of incubation; chalcopyrite dissolution in assays with DSF or BDSF were significantly lower than in the control assays without DSF or BDSF addition; −, no inhibition, and substrate oxidation similar to assays without DSF or BDSF addition. Tetrathionate was used for *A. caldus*, iron(II) ions were used for *L. ferriphilum*, and tetrathionate or iron(II) ions were used for *S. thermosulfidooxidans*. Fig. S3 and S4 substantiate the summary represented by the indicators (±) shown here by providing quantitative measurements of iron(II) ions, pH values, and planktonic cell counts [Fig. S3, soluble energy sources of iron(II)ions or tetrathionate] and total copper ions (Fig. S4, insoluble energy source of chalcopyrite).

cultures of *A. caldus* (Fig. S5). However, biofilm dispersal effects were short-lived, and recolonization of the chalcopyrite occurred in the batch experiment assays within 24 h after DSF addition. The addition of DSF to mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (Fig. 6C) caused a marked difference in the development of the sessile cell population, which was similar to the one observed in pure cultures of *L. ferriphilum* (Fig. 6A). Deep-learning analysis of this mixed-species biofilm under the

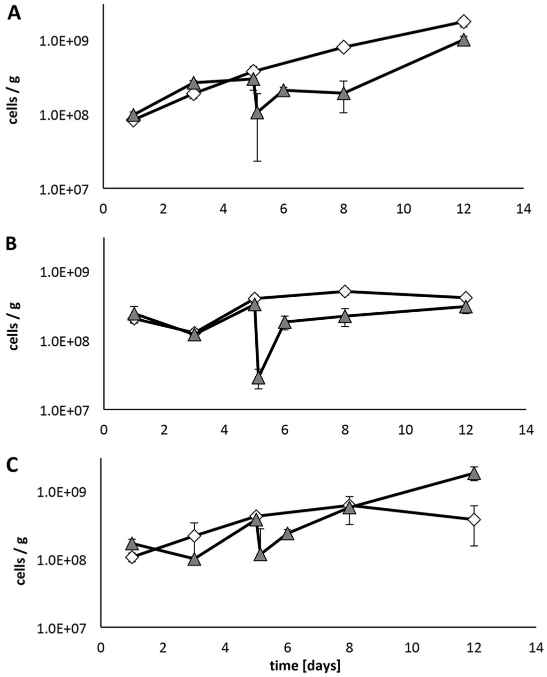


FIG 6 DSF molecules stimulate biofilm dispersal in *L. ferriphilum* and *S. thermosulfidooxidans*. (A to C) Axenic cultures of *L. ferriphilum* (A), *S. thermosulfidooxidans* (B), and mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (C) were cultivated with 2% chalcopyrite. DSF (5 μM) was added after 5 days of incubation (gray triangles), and the mineral-attached cell population was compared to control experiments without DSF (white diamonds).

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

TABLE 4 Deep-learning classification of biofilm patterns on chalcopyrite after 12 days of incubation and addition of 5 μM DSF

	Predicted class (+ DSF/control)[a]					
Actual class	**A**	**L**	**S**	**AS**	**LS**	**ALS**
A	89/89	6/0	3/3	NA	NA	8/3
L	6/3	86/92	3/0	NA	6/NA	3/3
S	6/3	0/3	83/89	NA	6/NA	6/6
LS	10/8	33/3	4/0	11/NA	38/83	4/6
ALS	3/3	3/3	6/3	NA	NA/3	89/89

[a] Probabilities (%) were assigned by the deep-learning analysis for the similarity of 36 images to the CNN class prediction. CNNs were trained with 600 images from five-day-old mineral cultures with different inoculum compositions of *A. caldus* (A), *L. ferriphilum* (L), and *S. thermosulfidooxidans* (S) that were used as pure or mixed cultures, resulting in the following categories: A, L, S, AS, LS, and ASL. NA, not analyzed.

influence of the DSF molecule (Table 4) confirmed a relatively high similarity with the biofilm pattern of axenic *L. ferriphilum* cultures (33%) and the one of mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (38%). However, DSF molecules had no influence on the biofilm pattern classification in all the other mixed or axenic cultures. In general, biofilm patterns on chalcopyrite grains after 12 days of incubation matched well to the true species composition in axenic or mixed cultures and are therefore similar to those observed in the training set images from day five of the experiment (Table 4).

DISCUSSION

The presented method allows the direct assessment of the relative amount of mineral-attached bacterial cells in laboratory bioleaching cultures. It avoids laborious biochemical or molecular biology sample pretreatment procedures, such as nucleic acid extraction, and their biases. The method has its main strength in performing relative comparisons rather than accurate absolute quantification of the amounts of attached cells and was tested for mineral concentrates as a proof of concept. However, we propose that the method is extendable, with some specific *ad hoc* parameterization for the analysis of other industrially relevant concentrates and low-grade ore preparations. These requirements include, but are not restricted to, the mineral particle size of the ore sample, which has to be sufficiently small and homogenous for enabling the visible deliberation of metal sulfide phases and gangue mineral phases using standard microscopy equipment. Adapted image analysis algorithms may have to include manual or automated differentiation of mineral phases and exclusion of gangue and autofluorescent mineral phases. Consequently, we suggest that it will be possible to employ similar techniques for assessment of the microbial colonization of metal sulfides in complex and low-grade mineral samples.

For the species-specific attachment behavior on metal sulfides, similar findings have been published (32–34), supporting the validity of our approach. The reliable, relative, and quantitative evaluation of biofilm populations is an innovative and powerful avenue for industrial and academic efforts to improve biomining operations and devise inoculation strategies of bioleaching operations.

L. ferriphilum cells have a high capacity to form biofilms on chalcopyrite and pyrite ores, and our method proved this directly in time-series studies (Fig. 4). In contrast, *A. caldus* cells that are unable to oxidize pyrite exhibited a low affinity to its surface in short-duration studies (13, 33), which are based on an indirect assessment of the attached cells by counting planktonic cell numbers and following their decline during initial contact with metal sulfides. However, the ostensibly high affinity of this ISC-oxidizing strain to pyrite surfaces in longer-duration axenic culture experiments presented here (Fig. 4C2) may be explained since biofilm formation is a common microbial starvation survival strategy (35). *S. thermosulfidooxidans* showed fewer attached cells than *A. caldus* within the first week of cultivation in chalcopyrite cultures (Fig. 4B1), and this may explain difficulties encountered in RNA and protein extraction from biofilm cells of chalcopyrite cultures of this species. It may indicate that the attached *S.*

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

thermosulfidooxidans population on chalcopyrite did not multiply. The poor initial attachment of *S. thermosulfidooxidans* alongside the slow increase of the number of attached *A. caldus* cells on chalcopyrite compared to pyrite grains (Fig. 4B1 and B2) is possibly related to the physiological effect of inhibitory levels of copper ions. Those reached concentrations of approximately 100 mg/liter after 5 days of incubation in cultures of *S. thermosulfidooxidans* (Fig. 5), and even lower copper concentrations are known to inhibit biofilm formation by iron-oxidizing acidithiobacilli (10). This is supported by the characteristic difference in the development of the fraction of biofilm cells in chalcopyrite (Fig. 4C1) and pyrite (Fig. 4C2) cultures.

The strong decrease in the mineral-attached cell population in *L. ferriphilum* pyrite cultures measured on day 21 (Fig. 4B2) may indicate a pronounced biofilm dispersal event, since *A. caldus* and *S. thermosulfidooxidans* cells exhibited a slower and more gradual decrease in attached cells than did *L. ferriphilum*. This dispersal may be related to multiple factors, including the toxicity of exudates, a lowered pH, and enhanced ionic strength that are known limiting and inhibitory factors for pyrite colonization (10, 19). However, an additional explanation for the dispersal may involve a QS-related effect. Christel and coworkers (27) revealed the presence of a DSF family QS system in *L. ferriphilum*. Even though DSF family signaling compounds of this species are not chemically identified, bioinformatic analyses suggest a possible function of these signal molecules in AMD and bioleaching microbial communities (Table 2). The high relative expression levels of the DSF synthase (LFTS_0514) support this suggestion (Fig. S2). Fatty acids were identified in extracts of pyrite cultures of *L. ferriphilum* and several other leptospirilli (36). Unknown compounds in those extracts inhibited iron oxidation in several acidophilic iron oxidizers, including *S. thermosulfidooxidans* and *A. ferrooxidans*^{T} (36). A similar observation was made in this study with DSF family signal compounds from *Xanthomonas campestris* (*cis*-11-methyl-dodecenoic acid, DSF) and *B. cenocepacia* (*cis*-2-dodecenoic acid [BDSF]) in *S. thermosulfidooxidans* and *L. ferriphilum* (Table 3 and Fig. S3 and S4). Furthermore, DSF family molecules are known biofilm dispersal agents with pronounced interspecies effects (28, 29, 37, 38). Consequently, it is not surprising that these compounds also caused biofilm dispersal in *L. ferriphilum* and *S. thermosulfidooxidans* (Fig. 6 and S4). Even though the biofilm dispersal effects were of short duration under batch culture conditions, the implications of this observation are of great importance under environmental conditions. Here, biofilm dispersal may be ensued by a succession in attachment by other microorganisms and detachment of bioleaching microorganisms may impact the performance of heap or stirred-tank bioleaching reactors. Furthermore, if DSF molecules are produced by mineral-oxidizing bacteria, cell-cell signaling mechanisms exerting strong inhibitory and presumably also biofilm dispersal effects on competing species may provide strategies to manipulate leaching activities in target strains.

Deep learning was used to classify biofilm images from experimental conditions that were not represented in the training image sets. Based on the visual features learned during the training, the deep learning correctly inferred the bacterial composition of the biofilms composed of combinations of the three species used in this study. The high accuracy achieved in classification of biofilm images after training with convolutional neural networks (CNNs) with a reduced number of images, compared to recent successful deep-learning applications (39–41), demonstrates that deep learning represents a valid imaging-based method for the analysis of low-diversity mixed-biofilm populations (Table 1). In combination with molecular validation, we anticipate that this method may be extended as an alternative to classical molecular methods for specific applications with characteristic and low-species-abundance microbial consortia.

Deep learning applied to images from chalcopyrite grains from mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* after the addition of 5 μM DSF molecules suggested an intermediate situation between biofilms from axenic *L. ferriphilum* cultures (probability, 33%) and mixed *L. ferriphilum* and *S. thermosulfidooxidans* cultures (38%, Table 1). Further indications suggested a dominance of *L. ferriphilum* cells in those cultures after the addition of DSF. Phase-contrast microscopy indicated mainly small,

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

curved, motile, and rod-shaped cells characteristic of *L. ferriphilum* in the planktonic cell population on day 12 of this experiment. Furthermore, a similar increase in the amount of biofilm cells, as shown in Fig. 6C, was observed in axenic cultures of *L. ferriphilum* with or without the addition of DSF molecules (Fig. 6A). Taken together, these results suggest that DSF molecules facilitated and accelerated *L. ferriphilum* to dominate the mixed culture with *S. thermosulfidooxidans*. The presence of DSF family QS genes in both species (Table 2) suggests that a complex signal molecule interaction of *L. ferriphilum* and *S. thermosulfidooxidans* may exist in mixed cultures. In general, competition for dissolved iron(II) ions and attachment sites on metal sulfides may be directly mediated by the DSF signal compounds, which trigger degradation of the second messenger c-di-GMP (42, 43). Low levels of c-di-GMP are primarily associated with upregulation of bacterial motility genes and downregulation of genes related to bacterial biofilm formation and exopolysaccharide (EPS) production (30, 44). However, the mechanism that explains the inhibition of iron(II) oxidation by DSF family signaling compounds is not yet understood. Likewise, it remains to be demonstrated if inhibition of iron(II) oxidation in *L. ferriphilum* is valid also for the DSF family compounds that are hypothesized to be produced by *L. ferriphilum*.

Conclusion. The presented study is a proof of concept for a direct method for relative quantification of attached cells on metal sulfides using automated image acquisition and analysis. The results highlight the effects of DSF family signal compounds in cultures of *L. ferriphilum* and *S. thermosulfidooxidans* and suggest an important role of these signal compounds in colonization of metal sulfides, microbial interactions, and niche defense among chemolithotrophic mineral-oxidizing bacteria that compete for electron donors originating from interfacial processes that determine metal sulfide dissolution.

MATERIALS AND METHODS

Microorganisms, cultivation media, and mineral cultures. The type strains *Acidithiobacillus caldus* DSM 8584 (45), *Leptospirillum ferriphilum* DSM 14647 (46), and *Sulfobacillus thermosulfidooxidans* DSM 9293 (31) were cultured with Mackintosh basal salt medium (MAC) (47). The medium was autoclaved at 121°C for 20 min. Cells were grown with soluble electron donors for inoculation of mineral cultures. This approach is a realistic scenario for the production of industrial bioleaching inoculum cells. In the case of *L. ferriphilum*, 4 g/liter iron(II) ions (provided as $FeSO_4 \cdot 7H_2O$) was used. Precipitation of ferric salts was prevented by the addition of sulfuric acid to maintain the pH in the range 1.6 to 1.8. *A. caldus* and *S. thermosulfidooxidans* precultures were grown using 0.9 g/liter potassium tetrathionate ($K_2S_4O_6$), and for *S. thermosulfidooxidans*, the medium was amended with 0.02% yeast extract (YE) and 0.1 g/liter iron(II) ions. Cells were harvested by centrifugation at 11,270 × *g* for 10 min and washed with 100 ml MAC medium. Subsequently, cells were inoculated at an initial cell density of 10^7 cells/ml to mineral cultures in 300-ml Erlenmeyer flasks with 150 ml MAC medium and 2% (wt/vol) pyrite or chalcopyrite grains (50- to 100-μm grain size). Equal proportions of cells of each species were used in mixed cultures. All strains were cultivated on a rotary shaker at 37°C and 150 rpm. For transcriptomic analyses, *L. ferriphilum* was additionally grown in continuous cultures, as described previously (27). Nucleic acid and protein extractions from free-swimming planktonic cells from batch mineral cultures, mineral-attached cells, and continuous-culture iron(II)-grown planktonic cells were done using a hot phenol protocol, as previously described (18, 27). Basic local alignment search tool (BLASTP) (48) was used to identify homologous proteins of known DSF family QS systems in the genome sequences of the three species.

For testing the effects of DSF family signal compounds, *cis*-11-methyl-dodecenoic acid (DSF; CAS 677354-23-3; Sigma) or *cis*-2-dodecenoic acid (BDSF; CAS 55928-65-9; Sigma) was used. *A. caldus*, *L. ferriphilum*, and *S. thermosulfidooxidans* were grown as described above, with the exception that YE was omitted in *S. thermosulfidooxidans* cultures. DSF family signal compounds were applied at 5 μM for testing their effects on cell growth and soluble substrate oxidation. Growth was evaluated by monitoring the planktonic cell number using a Thoma counting chamber and a phase-contrast microscope, spectrophotometric measurement of iron(II) ions (49), and following the development of pH for the tetrathionate cultures. DSF was also spiked into chalcopyrite cultures at a concentration of 5 μM for testing their effects on metal sulfide colonization and oxidation in axenic and mixed cultures of *A. caldus*, *L. ferriphilum*, and *S. thermosulfidoxidan*s. Metal sulfide dissolution was monitored by measurement of the concentration of iron(II) ions, total iron ions, and total copper ions using the spectrophotometric phenanthroline and bicinchoninic acid assays, respectively (49, 50). All experiments were done in triplicate.

Mineral preparation. Pure mineral samples were used in this study. Museum-grade pyrite grains (Navajun, Spain) used in leaching and attachment assays were from cube crystals crushed with a disc swing-mill (HSM 100M; Herzog). Chalcopyrite grains were obtained from a flotation concentrate provided by Boliden AB (Sweden). Mineral grains were wet sieved (Retsch, Germany) in order to use the particle fraction between 50 and 100 μm. Pyrite grains were boiled for 30 min in approximately 10 volumes of 6 M HCl, washed with deionized water until the pH was neutral, and stirred twice in approximately 5 volumes of acetone for 30 min in order to remove soluble sulfur compounds (51). Chalcopyrite grains

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

were washed twice for 30 min in 10 volumes of washing solution (0.1 M EDTA, 0.4 M NaOH), followed by treatment with acetone, as described for pyrite grains. For sterilization of mineral preparations, aliquots were sealed under a nitrogen atmosphere and incubated for 10 h at 125°C.

Microscopy sample preparation. Mineral grain particle samples were withdrawn from mineral cultures (~25 mg) using a flame-sterilized spatula. These particles were incubated in 1 ml MAC medium (pH 1.8) with 4% formaldehyde at room temperature for 1 h for fixation of mineral-attached cells, followed by two washing steps with water and subsequently with 1 ml phosphate-buffered saline (PBS). Samples were stored at −20°C in 50% ethanol in PBS. Mineral particles were incubated for 10 min in 200 μl of an aqueous solution of 0.01% 4′,6-diamidine-2′-phenylindole dihydrochloride (DAPI) in 2% formaldehyde. Prior to and after staining of attached cells, mineral grains were washed with 1 ml PBS. Finally, mineral particles were mounted on 10-well diagnostic glass slides (10-well, 6.7 mm; Thermo Scientific) using a glycerol-based mounting medium (CitiFluor AF2) and 22- by 50-mm cover glasses (Fig. 1A).

High-throughput epifluorescence microscopy. Automated image acquisition was performed as illustrated in Fig. 1A and B using an AxioImager M2m (Zeiss) fluorescence microscope equipped with a motorized microscopy stage (IM SCAN 130 × 85, DC 1 mm; Märzhäuser Wetzlar) and a AxioCam MRm camera. Image acquisition used a Zeiss filter set 09 for DAPI-stained samples or bright-field mode with background illumination for visualization of the localization of opaque mineral grains and transparent regions between. Images were recorded using a Zeiss Plan-Neofluar 20×/0.50 objective. Images were recorded as stack images with 2-μm step size, covering the entire maximum grain depth of 100 μm (50 layers). The extended-focus module of the Zen 2 software (blue edition, 2011; Carl Zeiss GmbH) was used to calculate projection images using the Wavelet option. Projections were exported as JPEG files. At least 36 images were analyzed for assessment of the amount of mineral-attached cells for every mineral sample and time point.

Image analysis. (i) Cell counting and mineral grain area determination. Cell counting was carried out computationally as illustrated in Fig. 2 by first converting the EFM images into gray-scale images and subsequently using the "Determinant of the Hessian" method ("blob_doh" function of Python's scikit-image package) with the following parameters: min_sigma = 0.34, max_sigma = 1, num_sigma = 2, threshold = 2×10^{-5}, overlap = 0.1, log_scale = false. The parameters were adjusted such that the analysis was accurate for a set of test images. A full description of the parameters is found on the "Determinant of the Hessian" Python scikit-image package. The analysis was subsequently applied to the entire image set. The mineral grain area was quantified from corresponding bright-field images with background illumination that were converted into gray-scale images and by setting a threshold in the color distribution using Otsu's method ("threshold_otsu" function of Python's scikit-image package assuming a bimodal pixel distribution in color intensity [nbins = 2]).

(ii) Calculation of mineral colonization and total cell numbers. The evaluation of the method's statistical accuracy depended on the number of images considered. Cell counts were related to the two-dimensional mineral grain area depicted in microscopy images and expressed as cells per mm^{-2}. After manual removal of extreme values, representing the top and bottom deciles of images with extremely low or high cell counts, metal sulfide colonization values [cells per mm^{-2}] of at least 36 images were normalized for the representation of 100% mineral grain area (i.e., the true percentage mineral area of each image and the corresponding cell count value were extrapolated to a theoretical image with 100% mineral coverage). Then, the values were randomly sorted using Microsoft Excel's random function and grouped in four arbitrarily chosen classes in order to calculate the mean of each class. These four classes can be understood as four sets of equal mineral areas used for averaging of the naturally nonhomogeneous mineral colonization over a larger area than that represented in a single microscopy image. The mean of the four mean values from each group and its coefficient of variation were calculated. For estimation of the metal sulfide colonization in cells per gram, the values in cells per mm^{-2} were multiplied with the specific surface area in $\text{mm}^2 \cdot \text{g}^{-1}$ of the mineral preparations (4.2×10^4 and 4.8×10^4 $\text{mm}^2 \cdot \text{g}^{-1}$) for the used pyrite and chalcopyrite concentrates, as determined by gas adsorption according to the BET (Brunauer Emmet and Taylor) theory. In order to take into account the fact that the mineral grains were viewed only from the top, the resulting values were doubled in order to account for the unobserved bottom side, while no correction factor was used for extrapolation from two-dimensional areas to the true three-dimensional mineral objects. Total cell numbers were estimated by calculation from direct counts of planktonic cells using phase-contrast microscopy with a Thoma chamber in cells per milliliter multiplied by the medium volume in milliliters plus the estimated amount of mineral-attached cells, which were determined using the image analysis method presented in this study in cells per gram multiplied by the mass of mineral in the bacterial culture in grams.

Deep learning. CNNs are a class of neural networks used in applications known as deep learning. They have shown high efficacy in areas of computer vision, such as image recognition and classification (52–54). The open-source program CAFFE was used to perform the deep-learning analysis (55). CNNs were used to perform deep-learning analysis of EFM images, where >600 images were used for model training and 100 images for model testing. In order to train our CNNs, images from mineral cultures with different inoculum compositions of *A. caldus* (A), *L. ferriphilum* (L), and *S. thermosulfidooxidans* (S) were used as pure or mixed cultures, resulting in the following categories: A, L, S, AS, LS, and ASL. These categories represent the biofilms formed on chalcopyrite grains after 5 days of incubation. Then, a network model for the CAFFE framework was defined and used along with the classified data to train the CNNs. Finally, the neural network analysis was validated by processing 100 images of each test category that were not used during the neural network training phase. It was also used to classify 36 images per species composition in chalcopyrite cultures after 12 days of incubation with or without addition of 5 μM DSF on day five.

RNA isolation, sequencing, and data analysis. Leaching cultures were separated into mineral-attached and planktonic cell subpopulations. RNA was extracted from continuous culture samples and

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

planktonic fractions according to Christel et al. (27), while RNA from mineral-attached cells was obtained as described previously (18). The RNA was purified with the RNeasy kit (Qiagen), including DNase treatment. RNA with sufficient quality was sequenced as described previously (27). Suitable RNA samples from chalcopyrite cultures of axenic *L. ferriphilum* (2 samples of mineral-attached cell subpopulation), mixed cultures of *L. ferriphilum* and *S. thermosulfidooxidans* (2 samples from the attached cell population and 4 samples from the planktonic cell subpopulation) were obtained. However, the success rate using this protocol was below 50% for chalcopyrite culture mineral samples. Raw reads for those samples are available under the accession no. PRJEB27815. Previously sequenced samples [3 *L. ferriphilum* continuous iron(II) culture samples and 2 samples from planktonic cells from chalcopyrite cultures] can be accessed under the accession no. PRJEB21842. Transcriptomic data were processed as described previously (27). In short, the resulting sequencing reads were mapped to the *L. ferriphilum* (27) reference genome with bowtie2 (56) after a quality filtering step. The resulting read counts for annotated coding sequences were normalized with DESeq2 (57) using a method introduced by Klingenberg and Meinicke (58).

Accession number(s). Raw reads for the RNA samples are available under the accession no. PRJEB27815.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.01835-18.

SUPPLEMENTAL FILE 1, PDF file, 9.9 MB.

ACKNOWLEDGMENTS

Image processing was done at the Swiss National Supercomputer Centre under projects u4, s653, and s747.

This project was supported by Bundesministerium für Bildung und Forschung (BMBF, grants 031A600A and 031A600B), Vetenskapsrådet (contract 2014-6545), the Luxembourg National Research Fund (FNR) (grant INTER/SYSAPP/14/05), and the Swiss Initiative in Systems Biology (SystemsX.ch, SysMetEx) under the frame of ERASysAPP. M.V. acknowledges support from Fondecyt 1161007 grant.

We acknowledge the effort of the reviewers who substantially improved our manuscript.

We declare no conflicts of interest.

REFERENCES

1. Rohwerder T, Sand W. 2007. Mechanisms and biochemical fundamentals of bacterial metal sulfide oxidation, p 35–58. *In* Donati ER, Sand W (ed), Microbial processing of metal sulfides. Springer, Dordrecht, The Netherlands.
2. Vera M, Schippers A, Sand W. 2013. Progress in bioleaching: fundamentals and mechanisms of bacterial metal sulfide oxidation–part A. Appl Microbiol Biotechnol 97:7529–7541. https://doi.org/10.1007/s00253-013-4954-2.
3. Brierley CL, Brierley JA. 2013. Progress in bioleaching: part B: applications of microbial processes by the minerals industries. Appl Microbiol Biotechnol 97:7543–7552. https://doi.org/10.1007/s00253-013-5095-3.
4. Schippers A, Sand W. 1999. Bacterial leaching of metal sulfides proceeds by two indirect mechanisms via thiosulfate or via polysulfides and sulfur. Appl Environ Microbiol 65:319–321.
5. Sand WGT, Hallmann R, Schippers A. 1995. Sulfur chemistry, biofilm, and the (in)direct attack mechanism—a critical evaluation of bacterial leaching. Appl Microbiol Biotechnol 43:961–966. https://doi.org/10.1007/BF00166909.
6. Gehrke T, Telegdi J, Thierry D, Sand W. 1998. Importance of extracellular polymeric substances from *Thiobacillus ferrooxidans* for bioleaching. Appl Environ Microbiol 64:2743–2747.
7. Gehrke T, Hallmann R, Kinzler K, Sand W. 2001. The EPS of *Acidithiobacillus ferrooxidans*–a model for structure-function relationships of attached bacteria and their physiology. Water Sci Technol 43:159–167. https://doi.org/10.2166/wst.2001.0365.
8. Sand W, Gehrke T, Jozsa PG, Schippers A. 2001. (Bio)chemistry of bacterial leaching–direct vs. indirect bioleaching. Hydrometallurgy 59:159–175. https://doi.org/10.1016/S0304-386X(00)00180-8.
9. Zhang R, Bellenberg S, Neu TR, Sand W, Vera M. 2016. The biofilm lifestyle of acidophilic metal/sulfur-oxidizing microorganisms, p 177–213. *In* Rampelotto PH (ed), Biotechnology of extremophiles: grand challenges in biology and biotechnology, vol 1, Springer International Publishing, Cham, Switzerland.
10. Bellenberg S, Barthen R, Boretska M, Zhang R, Sand W, Vera M. 2015. Manipulation of pyrite colonization and leaching by iron-oxidizing *Acidithiobacillus* species. Appl Microbiol Biotechnol 99:1435–1449. https://doi.org/10.1007/s00253-014-6180-y.
11. Rojas-Chapana JA, Tributsch H. 2004. Interfacial activity and leaching patterns of *Leptospirillum ferrooxidans* on pyrite. FEMS Microbiol Ecol 47:19–29. https://doi.org/10.1016/S0168-6496(03)00221-6.
12. Rodriguez-Leiva M, Tributsch H. 1988. Morphology of bacterial leaching patterns by *Thiobacillus ferrooxidans* on synthetic pyrite. Arch Microbiol 149:401–405. https://doi.org/10.1007/BF00425578.
13. Noël N, Florian B, Sand W. 2010. AFM & EFM study on attachment of acidophilic leaching organisms. Hydrometallurgy 104:370–375. https://doi.org/10.1016/j.hydromet.2010.02.021.
14. Bellenberg S, Diaz M, Noël N, Sand W, Poetsch A, Guiliani N, Vera M. 2014. Biofilm formation, communication and interactions of leaching bacteria during colonization of pyrite and sulfur surfaces. Res Microbiol 165:773–781. https://doi.org/10.1016/j.resmic.2014.08.006.
15. Africa CJ, van Hille RP, Sand W, Harrison STL. 2013. Investigation and *in situ* visualisation of interfacial interactions of thermophilic microorganisms with metal-sulphides in a simulated heap environment. Miner Eng 48:100–107. https://doi.org/10.1016/j.mineng.2012.09.011.
16. Africa CJ, Harrison STL, Becker M, Hille RP. 2010. In *situ* investigation and visualisation of microbial attachment and colonisation in a heap bioleach environment: the novel biofilm reactor. Miner Eng 23:486–491. https://doi.org/10.1016/j.mineng.2009.12.011.
17. Hedrich S, Guézennec A-G, Charron M, Schippers A, Joulian C. 2016. Quantitative monitoring of microbial species during bioleaching of a copper concentrate. Front Microbiol 7:2044. https://doi.org/10.3389/fmicb.2016.02044.
18. Vera M, Krok B, Bellenberg S, Sand W, Poetsch A. 2013. Shotgun proteomics study of early biofilm formation process of *Acidithiobacillus ferrooxidans* ATCC 23270 on pyrite. Proteomics 13:1133–1144. https://doi.org/10.1002/pmic.201200386.
19. Echeverría-Vega A, Demergasso C. 2015. Copper resistance, motility and

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

the mineral dissolution behavior were assessed as novel factors involved in bacterial adhesion in bioleaching. Hydrometallurgy 157:107–115. https://doi.org/10.1016/j.hydromet.2015.07.018.

20. Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. 2010. EBImage–an R package for image processing with applications to cellular phenotypes. Bioinformatics 26:979–981. https://doi.org/10.1093/bioinformatics/btq046.
21. Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9:671. https://doi.org/10.1038/nmeth.2089.
22. Schindelin J, Rueden CT, Hiner MC, Eliceiri KW. 2015. The ImageJ ecosystem: an open platform for biomedical image analysis. Mol Reprod Dev 82:518–529. https://doi.org/10.1002/mrd.22489.
23. Lamprecht MR, Sabatini DM, Carpenter AE. 2007. CellProfiler: free, versatile software for automated biological image analysis. Biotechniques 42:71–75. https://doi.org/10.2144/000112257.
24. Rämö P, Sacher R, Snijder B, Begemann B, Pelkmans L. 2009. CellClassifier: supervised learning of cellular phenotypes. Bioinformatics 25:3028–3030. https://doi.org/10.1093/bioinformatics/btp524.
25. Ruiz LM, Valenzuela S, Castro M, Gonzalez A, Frezza M, Soulère L, Rohwerder T, Queneau Y, Doutheau A, Sand W, Jerez CA, Guiliani N. 2008. AHL communication is a widespread phenomenon in biomining bacteria and seems to be involved in mineral-adhesion efficiency. Hydrometallurgy 94: 133–137. https://doi.org/10.1016/j.hydromet.2008.05.028.
26. González A, Bellenberg S, Mamani S, Ruiz L, Echeverria A, Soulère L, Doutheau A, Demergasso C, Sand W, Queneau Y, Vera M, Guiliani N. 2013. AHL signaling molecules with a large acyl chain enhance biofilm formation on sulfur and metal sulfides by the bioleaching bacterium *Acidithiobacillus ferrooxidans*. Appl Microbiol Biotechnol 97:3729–3737. https://doi.org/10.1007/s00253-012-4229-3.
27. Christel S, Herold M, Bellenberg S, El Hajjami M, Buetti-Dinh A, Pivkin IV, Sand W, Wilmes P, Poetsch A, Dopson M. 2017. Multi-omics reveal the lifestyle of the acidophilic, mineral-oxidizing model species *Leptospirillum ferriphilum*^{T}. Appl Environ Microbiol https://doi.org/10.1128/aem.02091-17.
28. Ryan RP, McCarthy Y, Watt SA, Niehaus K, Dow JM. 2009. Intraspecies signaling involving the diffusible signal factor BDSF (cis-2-dodecenoic acid) influences virulence in *Burkholderia cenocepacia*. J Bacteriol 191: 5013–5019. https://doi.org/10.1128/JB.00473-09.
29. Ryan RP, Dow JM. 2011. Communication with a growing family: diffusible signal factor (DSF) signaling in bacteria. Trends Microbiol 19: 145–152. https://doi.org/10.1016/j.tim.2010.12.003.
30. Römling U, Galperin MY, Gomelsky M. 2013. Cyclic di-GMP: the first 25 years of a universal bacterial second messenger. Microbiol Mol Biol Rev 77:1–52. https://doi.org/10.1128/MMBR.00043-12.
31. Golovacheva RS, Karavaiko GI. 1978. *Sulfobacillus*, a new genus of thermophilic sporulating bacteria. Mikrobiologiia 47:815–822. (In Russian.)
32. Foucher S, Battaglia-Brunet F, d'Hugues P, Clarens M, Godon JJ, Morin D. 2003. Evolution of the bacterial population during the batch bioleaching of a cobaltiferous pyrite in a suspended-solids bubble column and comparison with a mechanically agitated reactor. Hydrometallurgy 71: 5–12. https://doi.org/10.1016/S0304-386X(03)00142-7.
33. Florian B, Noël N, Thyssen C, Felschau I, Sand W. 2011. Some quantitative data on bacterial attachment to pyrite. Miner Eng 24:1132–1138. https://doi.org/10.1016/j.mineng.2011.03.008.
34. Sampson MI, Phillips CV, Blake RC, II. 2000. Influence of the attachment of acidophilic bacteria during the oxidation of mineral sulfides. Miner Eng 13:373–389. https://doi.org/10.1016/S0892-6875(00)00020-0.
35. Flemming H-C, Wingender J, Szewzyk U, Steinberg P, Rice SA, Kjelleberg S. 2016. Biofilms: an emergent form of bacterial life. Nat Rev Micro 14:563–575. https://doi.org/10.1038/nrmicro.2016.94.
36. Noël N. 2013. Attachment of acidophilic bacteria to solid substrata. Ph.D. thesis. Universität Duisburg-Essen, Duisburg, Germany.
37. Dow JM, Crossman L, Findlay K, He YQ, Feng JX, Tang JL. 2003. Biofilm dispersal in *Xanthomonas campestris* is controlled by cell-cell signaling and is required for full virulence to plants. Proc Natl Acad Sci U S A 100:10995–11000. https://doi.org/10.1073/pnas.1833360100.
38. Dean SN, Chung M-C, van Hoek ML. 2015. *Burkholderia* diffusible signal factor signals to *Francisella novicida* to disperse biofilm and increase siderophore production. Appl Environ Microbiol 81:7057–7066. https://doi.org/10.1128/AEM.02165-15.
39. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115. https://doi.org/10.1038/nature21056.
40. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316 22:2402–2410. https://doi.org/10.1001/jama.2016.17216.
41. Buggenthin F, Buettner F, Hoppe PS, Endele M, Kroiss M, Strasser M, Schwarzfischer M, Loeffler D, Kokkaliaris KD, Hilsenbeck O, Schroeder T, Theis FJ, Marr C. 2017. Prospective identification of hematopoietic lineage choice by deep learning. Nat Methods 14:403–406. https://doi.org/10.1038/nmeth.4182.
42. Ryan RP, Fouhy Y, Lucey JF, Crossman LC, Spiro S, He YW, Zhang LH, Heeb S, Cámara M, Williams P, Dow JM. 2006. Cell-cell signaling in *Xanthomonas campestris* involves an HD-GYP domain protein that functions in cyclic di-GMP turnover. Proc Natl Acad Sci U S A 103:6712. https://doi.org/10.1073/pnas.0600345103.
43. Deng Y, Schmid N, Wang C, Wang J, Pessi G, Wu D, Lee J, Aguilar C, Ahrens CH, Chang C, Song H, Eberl L, Zhang LH. 2012. *cis*-2-dodecenoic acid receptor RpfR links quorum-sensing signal perception with regulation of virulence through cyclic dimeric guanosine monophosphate turnover. Proc Natl Acad Sci U S A 109:15479–15484. https://doi.org/10.1073/pnas.1205037109.
44. Hengge R. 2009. Principles of c-di-GMP signalling in bacteria. Nat Rev Microbiol 7:263–273. https://doi.org/10.1038/nrmicro2109.
45. Hallberg KB, Lindström EB. 1994. Characterization of *Thiobacillus caldus* sp. nov., a moderately thermophilic acidophile. Microbiology 140: 3451–3456. https://doi.org/10.1099/13500872-140-12-3451.
46. Coram NJ, Rawlings DE. 2002. Molecular relationship between two groups of the genus *Leptospirillum* and the finding that *Leptospirillum ferriphilum* sp. nov. dominates South African commercial biooxidation tanks that operate at 40°C. Appl Environ Microbiol 68:838–845. https://doi.org/10.1128/AEM.68.2.838-845.2002.
47. Mackintosh M. 1978. Nitrogen fixation by *Thiobacillus ferrooxidans*. J Gen Microbiol 105:215–218. https://doi.org/10.1099/00221287-105-2-215.
48. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389.
49. Harvey AE, Jr, Smart JA, Amis E. 1955. Simultaneous spectrophotometric determination of iron(II) and total iron with 1,10-phenanthroline. Anal Chem 27:26–29. https://doi.org/10.1021/ac60097a009.
50. Anwar MA, Iqbal M, Qamar MA, Rehman M, Khalid AM. 2000. Technical communication: determination of cuprous ions in bacterial leachates and for environmental monitoring. World J Microbiol Biotechnol 16: 135–138. https://doi.org/10.1023/A:1008978501177.
51. Moses CO, Nordstrom DK, Herman JS, Mills AL. 1987. Aqueous pyrite oxidation by dissolved oxygen and by ferric iron. Geochim Cosmochim Acta 51:1561–1571. https://doi.org/10.1016/0016-7037(87)90337-1.
52. Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J, Zieba K. 2016. End to end learning for self-driving cars. arXiv arXiv:1604.07316v1.
53. Parkhi OM, Vedaldi A, Zisserman A. 2015. Deep face recognition. University of Oxford, Oxford, United Kingdom.
54. Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90. https://doi.org/10.1145/3065386.
55. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. 2014. Caffe: convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on multimedia, Orlando, FL. https://doi.org/10.1145/2647868.2654889.
56. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357. https://doi.org/10.1038/nmeth.1923.
57. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8.
58. Klingenberg H, Meinicke P. 2017. How to normalize metatranscriptomic count data for differential expression analysis. PeerJ 5:e3859. https://doi.org/10.7717/peerj.3859.
59. Valdes J, Quatrini R, Hallberg K, Dopson M, Valenzuela PD, Holmes DS. 2009. Draft genome sequence of the extremely acidophilic bacterium *Acidithiobacillus caldus* ATCC 51756 reveals metabolic versatility in the genus *Acidithiobacillus*. J Bacteriol 191:5877–5878. https://doi.org/10.1128/JB.00843-09.

Downloaded from http://aem.asm.org/ on October 3, 2018 by guest

C.5 IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses.

Shaman Narayanasamy[†], Yohan Jarosz[†], Emilie E.L. Muller, Anna Heintz-Buschart, **Malte Herold**, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes
2016
Genome Biology **17**: 260
DOI: 10.1186/s13059-016-1116-8

Contributions of author include:

- Software development and testing
- Data analysis and visualization
- Writing and revision of manuscript

[†] Co-first author

Narayanasamy *et al. Genome Biology* (2016) 17:260
DOI 10.1186/s13059-016-1116-8

Genome Biology

SOFTWARE Open Access

IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses

Shaman Narayanasamy[1†], Yohan Jarosz[1†], Emilie E. L. Muller[1,2], Anna Heintz-Buschart[1], Malte Herold[1], Anne Kaysen[1], Cédric C. Laczny[1,3], Nicolás Pinel[4,5], Patrick May[1] and Paul Wilmes[1*]

Abstract

Existing workflows for the analysis of multi-omic microbiome datasets are lab-specific and often result in sub-optimal data usage. Here we present IMP, a reproducible and modular pipeline for the integrated and reference-independent analysis of coupled metagenomic and metatranscriptomic data. IMP incorporates robust read preprocessing, iterative co-assembly, analyses of microbial community structure and function, automated binning, as well as genomic signature-based visualizations. The IMP-based data integration strategy enhances data usage, output volume, and output quality as demonstrated using relevant use-cases. Finally, IMP is encapsulated within a user-friendly implementation using Python and Docker. IMP is available at http://r3lab.uni.lu/web/imp/ (MIT license).

Keywords: Multi-omics data integration, Metagenomics, Metatranscriptomics, Microbial ecology, Microbiome, Reproducibility

Background

Microbial communities are ubiquitous in nature and govern important processes related to human health and biotechnology [1, 2]. A significant fraction of naturally occurring microorganisms elude detection and investigation using classic microbiological methods due to their unculturability under standard laboratory conditions [3]. The issue of unculturability is largely circumvented through the direct application of high-resolution and high-throughput molecular measurements to samples collected in situ [4–6]. In particular, the application of high-throughput next-generation sequencing (NGS) of DNA extracted from microbial consortia yields metagenomic (MG) data which allow the study of microbial communities from the perspective of community structure and functional potential [4–6]. Beyond metagenomics, there is also a clear need to obtain functional readouts in the form of other omics data. The sequencing of reverse transcribed RNA (cDNA) yields metatranscriptomic (MT) data, which provides information about gene expression and therefore allows a more faithful assessment of community function [4–6]. Although both MG and MT data allow unprecedented insights into microbial consortia, the integration of such multi-omic data is necessary to more conclusively link genetic potential to actual phenotype in situ [4, 6]. Given the characteristics of microbial communities and the resulting omic data types, specialized workflows are required. For example, the common practice of subsampling collected samples prior to dedicated biomolecular extractions of DNA, RNA, etc. has been shown to inflate variation, thereby hampering the subsequent integration of the individual omic datasets [7, 8]. For this purpose, specialized wet-lab methods which allow the extraction of concomitant DNA, RNA, proteins, and metabolites from single, unique samples were developed to ensure that the generated data could be directly compared across the individual omic levels [7, 8]. Although standardized and reproducible wet-lab methods have been developed for integrated omics of microbial communities, corresponding bioinformatic analysis workflows have yet to be formalized.

---

* Correspondence: paul.wilmes@uni.lu
†Equal contributors
[1] Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur-Alzette L-4362, Luxembourg
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2016 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Bioinformatic analysis methods for MG and MT NGS data can be broadly classified into reference-dependent or reference-independent (de novo) methods [5]. Reference-dependent methods are based on the alignment/mapping of sequencing reads onto isolate genomes, gene catalogs, or existing MG data. A major drawback of such methods is the large number of sequencing reads from uncultured species and/or divergent strains which are discarded during data analysis, thereby resulting in the loss of potentially useful information. For example, based on analyses of MG data from the human gut microbiome (arguably the best characterized microbial community in terms of culture-derived isolate genomes), approximately 43% of the data are typically not mappable to the available isolate genomes [9]. Conversely, reference-independent methodologies, such as approaches based on de novo assemblies, enable the retrieval of the actual genomes and/or potentially novel genes present in samples, thereby allowing more of the data to be mapped and exploited for analysis [4, 5, 10]. Furthermore, it has been demonstrated that the assembly of sequencing reads into longer contiguous sequences (contigs) greatly improves the taxonomic assignments and prediction of genes as opposed to their direct identification from short sequencing reads [11, 12]. Finally, de novo MG assemblies may be further leveraged by binning the data to resolve and retrieve population-level genomes, including those from hitherto undescribed taxa [13–21].

Given the advantages of reference-independent methods, a wide array of MG-specific assemblers such as IDBA-UD [22] and MEGAHIT [23] have been developed. Most MT data analyses involve reference-based [24–26] or MG-dependent analysis workflows [27–29]. A comparative study by Celaj et al. [12] demonstrated that reference-independent approaches for MT data analyses are also applicable using either specialized MT assemblers (e.g., IDBA-MT [12, 30]), MG assemblers (e.g., IDBA-UD [22, 30, 31] and MetaVelvet [12, 32]) or single-species transcriptome assemblers (e.g., Trinity [12, 33]). In all cases, the available assemblers are able to handle the uneven sequencing depths of MG and MT data. Although dedicated assembly methods have been developed for MG and MT data, formalized pipelines allowing the integrated use of both data types are not available yet.

Automated bioinformatic pipelines have so far been mainly developed for MG data. These include MOCAT [34] and MetAMOS [10], which incorporate the entire process of MG data analysis, ranging from preprocessing of sequencing reads, de novo assembly, and post-assembly analysis (read alignment, taxonomic classification, gene annotation, etc.). MOCAT has been used in large-scale studies such as those within the MetaHIT Consortium [35, 36], while MetAMOS is a flexible pipeline which allows customizable workflows [10]. Both pipelines use SOAPdenovo [37] as the default de novo assembler, performing single-length *k*mer-based assemblies which usually result in fragmented (low contiguity) assemblies with low gene coverage values [38].

Multi-omic analyses have already provided new insights into microbial community structure and function in various ecosystems. These include studies of the human gut microbiome [28, 39], aquatic microbial communities from the Amazon river [27], soil microbial communities [40, 41], production-scale biogas plants [29], hydrothermal vents [42], and microbial communities from biological wastewater treatment plants [43, 44]. These studies employed differing ways for analyzing the data, including reference-based approaches [27, 28, 42], MG assembly-based approaches [29, 40], MT assembly-based approaches [42], and integrated analyses of the meta-omic data [39, 42–44]. Although these studies clearly demonstrate the power of multi-omic analyses by providing deep insights into community structure and function, standardized and reproducible computational workflows for integrating and analyzing the multi-omic data have so far been unavailable. Importantly, such approaches are, however, required to compare results between different studies and systems of study.

Due to the absence of established tools/workflows to handle multi-omic datasets, most of the aforementioned studies utilized non-standardized, ad hoc analyses, mostly consisting of custom workflows, thereby creating a challenge in reproducing the analyses [10, 45–47]. Given that the lack of reproducible bioinformatic workflows is not limited to those used for the multi-omic analysis of microbial consortia [10, 45–47], several approaches have recently been developed with the explicit aim of enhancing software reproducibility. These include a wide range of tools for constructing bioinformatic workflows [48–50] as well as containerizing bioinformatic tools/pipelines using Docker [29, 46–48].

Here, we present IMP, the Integrated Meta-omic Pipeline, the first open source de novo assembly-based pipeline which performs standardized, automated, flexible, and reproducible large-scale integrated analysis of combined multi-omic (MG and MT) datasets. IMP incorporates robust read preprocessing, iterative co-assembly of metagenomic and metatranscriptomic data, analyses of microbial community structure and function, automated binning, as well as genomic signature-based visualizations. We demonstrate the functionalities of IMP by presenting the results obtained on an exemplary data set. IMP was evaluated using datasets from ten different microbial communities derived from three distinct environments as well as a simulated mock microbial community dataset. We compare the assembly and data integration measures of IMP against standard MG analysis strategies (reference-based and reference-independent) to demonstrate that IMP vastly improves overall data usage. Additionally, we benchmark our assembly procedure against available MG analysis pipelines to show that IMP consistently produces high-quality assemblies across all the processed datasets. Finally, we describe a number of particular use cases which highlight biological applications of the IMP workflow.

Results

Overview of the IMP implementation and workflow

IMP leverages Docker for reproducibility and deployment. The interfacing with Docker is facilitated through a user-friendly Python wrapper script (see the "Details of the IMP implementation and workflow" section). As such, Python and Docker are the only prerequisites for the pipeline, enabling an easy installation and execution process. Workflow implementation and automation is achieved using Snakemake [49, 51]. The IMP workflow can be broadly divided into five major parts: i) preprocessing, ii) assembly, iii) automated binning, iv) analysis, and v) reporting (Fig. 1).

The preprocessing and filtering of sequencing reads is essential for the removal of low quality bases/reads, and potentially unwanted sequences, prior to assembly and analysis. The input to IMP consists of MG and MT (the latter preferably depleted of ribosomal RNA prior to sequencing) paired-end reads in FASTQ format (section "Input data"). MG and MT reads are preprocessed independently of each other. This involves an initial quality control step (Fig. 1 and section "Trimming and quality filtering") [52] followed by an optional screening for host/contaminant sequences, whereby the default screening is performed against the human genome while other host genome/contaminant sequences may also be used (Fig. 1 and section "Screening host or contaminant sequences"). In silico rRNA sequence depletion is exclusively applied to MT data (Fig. 1 and section "Ribosomal RNA filtering").

The customized assembly procedure of IMP starts with an initial assembly of preprocessed MT reads to generate an initial set of MT contigs (Additional file 1: Figure S1). MT reads unmappable to the initial set of MT contigs undergo a second round of assembly. The process of assembling unused reads, i.e., MG or MT reads unmappable to the previously assembled contigs, is henceforth referred to as "iterative assembly". The assembly of MT reads is performed, first as transcribed regions are covered much more deeply and evenly in MT data. The resulting MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data, overall leading to enhanced assemblies [43]. Therefore, the combined set of MT contigs from the initial and iterative MT assemblies are used to enhance the subsequent assembly with the MG data. MT data are assembled using the MEGAHIT de novo assembler using the appropriate option to prevent the merging of bubbles within the de Bruijn assembly graph [23, 36]. Subsequently, all preprocessed MT and MG reads, together with the generated MT contigs, are used as input to perform a first co-assembly, producing a first set of co-assembled contigs. The MG and MT reads unmappable to this first set of co-assembled contigs then undergo an additional iterative co-assembly step. IMP implements two assembler options for the de novo co-assembly step, namely IDBA-UD or MEGAHIT. The contigs resulting from the co-assembly procedure undergo a subsequent assembly refinement step by a contig-level assembly using the cap3 [53] de novo assembler. This aligns highly similar contigs against each other, thus reducing overall redundancy by collapsing shorter contigs into longer contigs and/or improving contiguity by extending contigs via overlapping contig ends (Additional file 1: Figure S1). This step produces the final set of contigs. Preprocessed MG and MT reads are then mapped back against the final contig set and the resulting alignment information is used in the various downstream analysis procedures (Fig. 1). In summary, IMP employs four measures for the de novo assembly of preprocessed MG and MT reads, including: i) iterative assemblies of unmappable reads, ii) use of MT contigs to scaffold the downstream assembly of MG data, iii) co-assembly of MG and MT data, and iv) assembly refinement by contig-level assembly. The entire de novo assembly procedure of IMP is henceforth referred to as the "IMP-based iterative co-assembly" (Additional file 1: Figure S1).

Contigs from the IMP-based iterative co-assembly undergo quality assessment as well as taxonomic annotation [54] followed by gene prediction and functional annotation [55] (Fig. 1 and section "Annotation and assembly quality assessment"). MaxBin 2.0 [20], an automated binning procedure (Fig. 1 and section "Automated binning") which performs automated binning on assemblies produced from single datasets, was chosen as the de facto binning procedure in IMP. Experimental designs involving single coupled MG and MT datasets are currently the norm. However, IMP's flexibility does not forego the implementation of multi-sample binning algorithms such as CONCOCT [16], MetaBAT [18], and canopy clustering [15] as experimental designs evolve in the future.

Non-linear dimensionality reduction of the contigs' genomic signatures (Fig. 1 and section "Non-linear dimensionality reduction of genomic signatures") is performed using the Barnes-Hut Stochastic Neighborhood Embedding (BH-SNE) algorithm allowing visualization of the data as two-dimensional scatter plots (henceforth referred to as VizBin maps [13, 56]). Further analysis steps include, but are not limited to, calculations of the contig- and gene-level depths of coverage (section

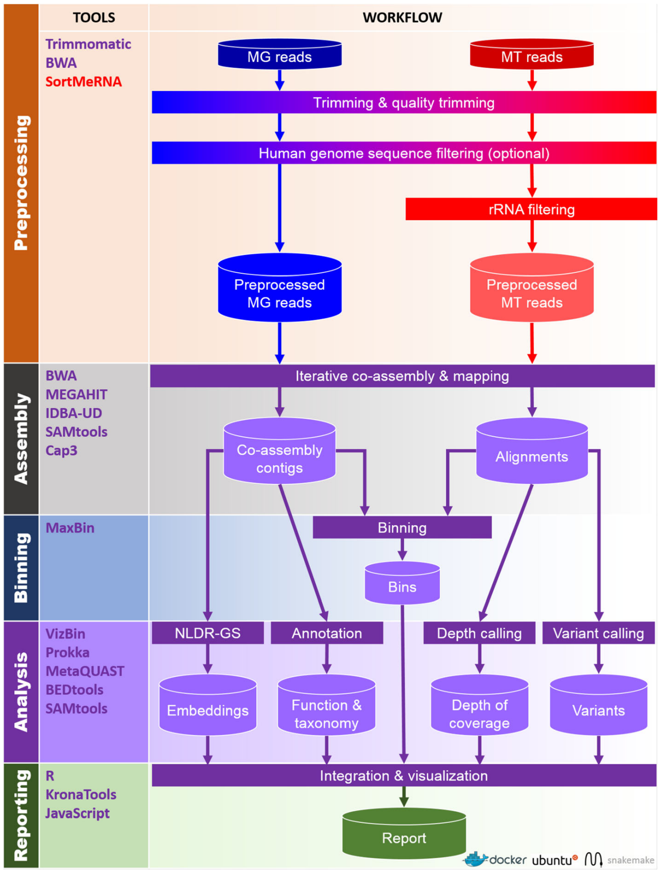


Fig. 1 Schematic overview of the IMP pipeline. *Cylinders* represent input and output while *rectangles* represent processes. *Arrows* indicate the flow between input, processes, and output. *MG* — Metagenomic data, *MT* — Metatranscriptomic data, *rRNA* — ribosomal RNA, *NLDR-GS* — genomic signature non-linear dimensionality reduction. Processes, input, and output specific to MG and MT data are labeled in *blue* and *red*, respectively. Processes and output that involve usage of both MG and MT data are represented in *purple*. A detailed illustration of the "iterative co-assembly" is available in Additional file 1: Figure S1

"Depth of coverage") as well as the calling of genomic variants (variant calling is performed using two distinct variant callers; section "Variant calling"). The information from these analyses are condensed and integrated into the generated VizBin maps to produce augmented visualizations (sections "Visualization and reporting"). These visualizations and various summaries of the output are compiled into a HTML report (examples of the HTML reports available via Zenodo [57]).

Exemplary output of IMP (using the default IDBA-UD assembler) based on a human fecal microbiome dataset is summarized in Fig. 2. The IMP output includes taxonomic (Fig. 2a) and functional (Fig. 2b, c) overviews. The representation of gene abundances at the MG and MT levels enables comparison of potential (Fig. 2b) and actual expression (Fig 2c) for specific functional gene categories (see Krona charts within HTML S1 [57]). IMP provides augmented VizBin maps [13, 56], including, for

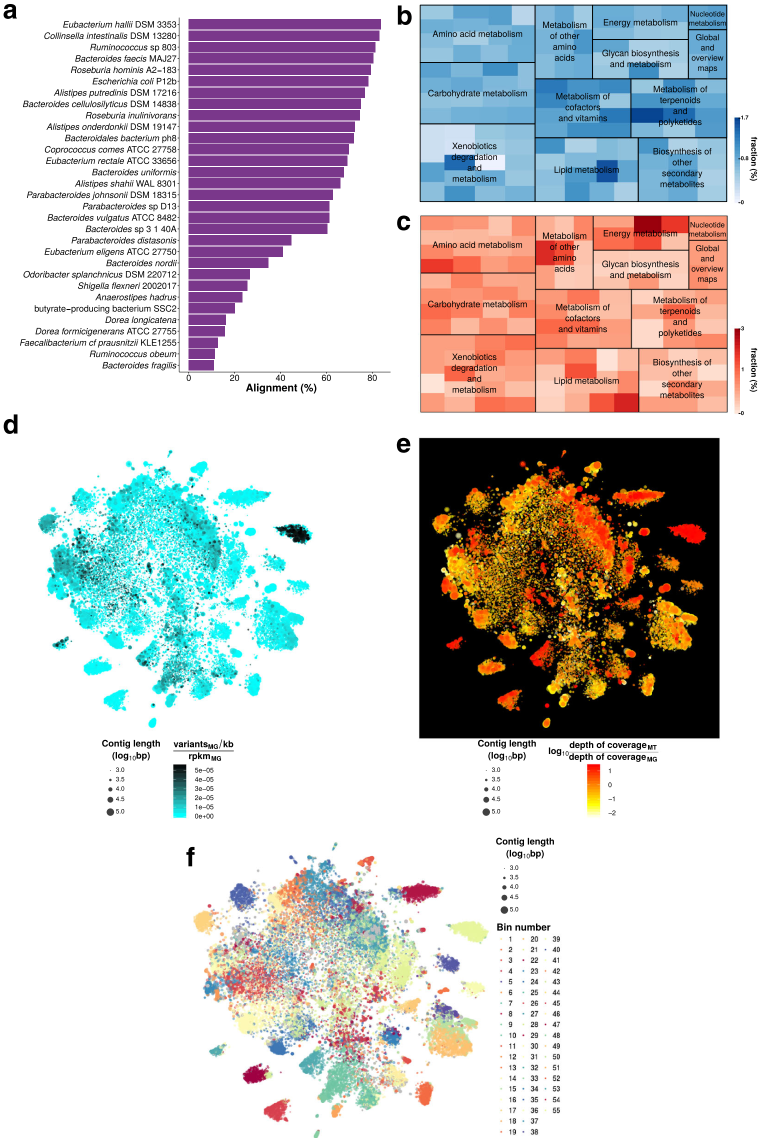


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Example output from the IMP analysis of a human microbiome dataset (HF1). **a** Taxonomic overview based on the alignment of contigs to the most closely related genomes present in the NCBI genome database (see also HTML report S1 [57]). **a**, **b** Abundances of predicted genes (based on average depths of coverage) of various KEGG Ontology categories represented both at the MG (**b**) and MT (**c**) levels (see also Krona charts within HTML report S1). **d**–**f** Augmented VizBin maps of contigs ≥1 kb, representing contig-level MG variant densities (**d**), contig-level ratios of MT to MG average depth of coverage (**e**), and bins generated by the automated binning procedure (**f**). Please refer to the HTML reports [57] for additional examples

example, variant densities (Fig. 2d) as well as MT to MG depth of coverage ratios (Fig. 2e). These visualizations may aid users in highlighting subsets of contigs based on certain characteristics of interest, i.e., population heterogeneity/homogeneity, low/high transcriptional activity, etc. Although an automated binning method [20] is incorporated within IMP (Fig. 2f), the output is also compatible with and may be exported to other manual/interactive binning tools such as VizBin [56] and Anvi'o [17] for additional manual curation. Please refer to the HTML reports for additional examples [57].

The modular design (section "Automation and modularity") and open source nature of IMP allow for customization of the pipeline to suit specific user-defined analysis requirements (section "Customization and further development"). As an additional feature, IMP also allows single-omic MG or MT analyses (section "Details of the IMP implementation and workflow"). Detailed parameters for the processes implemented in IMP are described in the section "Details of the IMP implementation and workflow" and examples of detailed workflow schematics are provided within the HTML reports [57].

Assessment and benchmarking

IMP was applied to ten published coupled MG and MT datasets, derived from three types of microbial systems, including five human fecal microbiome samples (HF1, HF2, HF3, HF4, HF5) [28], four wastewater sludge microbial communities (WW1, WW2, WW3, WW4) [43, 44], and one microbial community from a production-scale biogas (BG) plant [29]. In addition, a simulated mock (SM) community dataset based on 73 bacterial genomes [12], comprising both MG and MT data was generated to serve as a means for ground truth-based assessment of IMP (details in section "Coupled metagenomic and metatranscriptomic datasets"). The SM dataset was devised given the absence of a standardized benchmarking dataset for coupled MG and MT data (this does solely exist for MG data as part of the CAMI initiative (http://www.cami-challenge.org)).

Analysis with IMP was carried out with the two available de novo assembler options for the co-assembly step (Fig. 1; Additional file 1: Figure S1), namely the default IDBA-UD assembler [22] (hereafter referred to as IMP) and the optional MEGAHIT assembler [23] (henceforth referred to as IMP-megahit). IMP was quantitatively assessed based on resource requirement and analytical capabilities. The analytical capabilities of IMP were evaluated based on data usage, output volume, and output quality. Accordingly, we assessed the advantages of the iterative assembly procedure as well as the overall data integration strategy.

Resource requirement and runtimes

IMP is an extensive pipeline that utilizes both MG and MT data within a reference-independent (assembly-based) analysis framework which renders it resource- and time-intensive. Therefore, we aimed to assess the required computational resource and runtimes of IMP.

All IMP-based runs on all datasets were performed on eight compute cores with 32 GB RAM per core and 1024 GB of total memory (section "Computational platforms"). IMP runtimes ranged from approximately 23 h (HF1) to 234 h (BG) and the IMP-megahit runtimes ranged from approximately 21 h (HF1) up to 281 h (BG). IMP was also executed on the Amazon cloud computing (AWS) infrastructure, using the HF1 dataset on a machine with 16 cores (section "Computational platforms") whereby the run lasted approximately 13 h (refer to Additional file 1: Note S1 for more details). The analysis of IMP resulted in an increase in additional data of around 1.2–3.6 times the original input (Additional file 2: Table S1). Therefore, users should account for the disc space for both the final output and intermediate (temporary) files generated during an IMP run. Detailed runtimes and data generated for all the processed data sets are reported in Additional file 2: Table S1.

We further evaluated the effect of increasing resources using a small scale test dataset (section "Test dataset for runtime assessment"). The tests demonstrated that reduced runtimes are possible by allocating more threads to IMP-megahit (Additional file 2: Table S2). However, no apparent speed-up is achieved beyond allocation of eight threads, suggesting that this would be the optimal number of threads for this particular test dataset. Contrastingly, no speed-up was observed with additional memory allocation (Additional file 2: Table S3). Apart from the resources, runtime may also be affected by the input size, the underlying complexity of the dataset and/or behavior of individual tools within IMP.

Data usage: iterative assembly

De novo assemblies of MG data alone usually result in a large fraction of reads that are unmappable to the assembled contigs and therefore remain unused, thereby leading to suboptimal data usage [43, 58–60]. Previous studies have assembled sets of unmappable reads iteratively to successfully obtain additional contigs, leading to an overall increase in the number of predicted genes, which in turn results in improved data usage [43, 58–60]. Therefore, IMP uses an iterative assembly strategy to maximize NGS read usage. In order to evaluate the best iterative assembly approach for application within the IMP-based iterative co-assembly strategy, we attempted to determine the opportune number of assembly iterations in relation to assembly quality metrics and computational resources/runtimes.

The evaluation of the iterative assembly strategy was applied to MG and MT datasets. For both omic data types, it involved an "initial assembly" which is defined as the de novo assembly of all preprocessed reads. Additional iterations of assembly were then conducted using the reads that remained unmappable to the generated set of contigs (see section "Iterative single-omic assemblies" for details and parameters). The evaluation of the iterative assembly procedure was carried out based on the gain of additional contigs, cumulative contig length (bp), numbers of genes, and numbers of reads mappable to contigs. Table 1 shows the evaluation results of four representative data sets and Additional file 2: Table S4 shows the detailed results of the application of the approach to 11 datasets. In all the datasets evaluated, all iterations (1 to 3) after the initial assembly lead to an increase in total length of the assembly and numbers of mappable reads (Table 1; Additional file 2: Table S4). However, there was a notable decline in the number of additional contigs and predicted genes beyond the first iteration. Specifically, the first iteration of the MG assembly yielded up to 1.6% additional predicted genes while the equivalent on the MT data yielded up to 9% additional predicted genes (Additional file 2: Table S4). Considering the small increase (<1%) in the number of additional contigs and predicted genes beyond the first assembly iteration on one hand and the extended runtimes required to perform additional assembly iterations on the other hand, a generalized single iteration assembly approach was retained and implemented within the IMP-based iterative co-assembly (Fig. 1; Additional file 1: Figure S1). This approach aims to maximize data usage without drastically extending runtimes.

Despite being developed specifically for the analysis of coupled MG and MT datasets, the iterative assembly can also be used for single omic datasets. To assess IMP's performance on MG datasets, it was applied to the simulated MG datasets from the CAMI challenge (http://www.cami-challenge.org) and the results are shown in Additional file 1: Figure S2. IMP-based MG assembly using the MEGAHIT assembler on the CAMI dataset outperforms well-established MG pipelines such

Table 1 Statistics of iterative assemblies performed on MG and MT datasets

		MG iterative assembly				MT iterative assembly			
Dataset	Iteration	Number of contigs (≥1 kb)	Cumulative length of assembled contigs (bp)	Number of predicted genes	Number of mapped reads	Number of contigs (all)	Cumulative length of assembled contigs (bp)	Number of predicted genes	Number of mapped reads
SM	Initial assembly	29063	182673343	186939	18977716	13436	8994518	13946	822718
	1	16	483336	329	9515	1286	502535	1272	16038
	2	6	213094	126	3425	48	18460	49	656
	3	1	86711	47	1536	0	0	0	0
HF1	Initial assembly	27028	145938650	154760	20715368	40989	45300233	66249	17525586
	1	15	966872	274	39839	2471	969614	2238	329400
	2	−1	26822	5	1276	26	10315	24	45642
	3	0	4855	0	172	3	1640	6	54788
WW1	Initial assembly	14815	77059275	81060	6513708	45118	22525759	49859	8423603
	1	28	3146390	1136	73511	2115	723904	1589	529441
	2	2	175634	114	4031	250	82048	201	13335
	3	1	30032	16	572	31	10280	18	65866
BG	Initial assembly	105282	545494441	593688	109949931	47628	27493690	60566	3754432
	1	417	10998269	3902	456821	3956	1397409	3061	130131
	2	5	335313	219	21647	717	250223	754	12766
	3	7	79022	20	2511	24	9060	22	5827

Results for all datasets available in Additional file 2: Table S2 as MOCAT in all measures. In addition, IMP-based iterative assemblies also exhibit comparable performance to the gold standard assembly with regards to contigs ≥1 kb and number of predicted genes (http://www.cami-challenge.org). Detailed results of the CAMI assemblies are available in Additional file 2: Table S5. However, as no MT and/or coupled MG and MT datasets so far exist for the CAMI challenge, the full capabilities of IMP could not be assessed in relation to this initiative.

Data usage: multi-omic iterative co-assembly

In order to assess the advantages of integrated multi-omic co-assemblies of MG and MT data, IMP-based iterative co-assemblies (IMP and IMP-megahit) were compared against MG-only-based assemblies which include single-omic iterative MG assemblies generated using IMP (referred to as IMP_MG) and standard MG assemblies by MOCAT (hereafter referred to as MOCAT_MG) and MetAMOS (hereafter referred to as MetAMOS_MG). Furthermore, the available reads from the human fecal microbiome dataset (preprocessed with IMP) were mapped to the MetaHIT Integrated Gene Catalog (IGC) reference database [35] to compare the data usage of the different assembly procedures against a reference-dependent approach.

IMP-based iterative co-assemblies consistently recruited larger fractions of properly paired MG (Fig. 3a) and/or MT (Fig. 3b) reads compared to single-omic

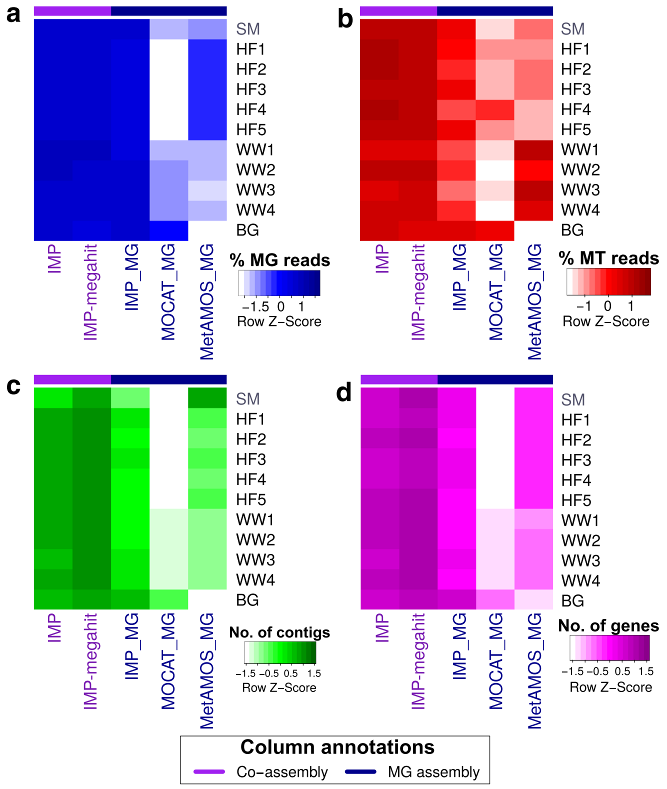


Fig. 3 Assessment of data usage and output generated from co-assemblies compared to single-omic assemblies. Heat maps show (**a**) fractions of properly mapped MG read pairs, (**b**) fractions of properly mapped MT read pairs, (**c**) numbers of contigs ≥1 kb, and (**d**) numbers of unique predicted genes. IMP and IMP-megahit represent integrated multi-omic MG and MT iterative co-assemblies while IMP_MG, MOCAT_MG, and MetAMOS_MG represent single-omic MG assemblies. All numbers were row Z-score normalized for visualization. Detailed results available in Additional file 2: Table S5

assemblies. The resulting assemblies also produced larger numbers of contigs ≥1 kb (Fig. 3c), predicted non-redundant unique genes (Fig. 3d), and, even more important, complete genes as predicted with start and stop codon by Prodigal [61] (Additional file 2: Table S5). Using the reference genomes from the SM data as ground truth, IMP-based iterative co-assemblies resulted in up to 25.7% additional recovery of the reference genomes compared to the single-omic MG assemblies (Additional file 2: Table S5).

IMP-based iterative co-assemblies of the human fecal microbiome datasets (HF1–5) allowed recruitment of comparable fractions of properly paired MG reads and an overall larger fraction of properly paired MT reads compared to those mapping to the IGC reference database (Table 2). The total fraction (union) of MG or MT reads mapping to either IMP-based iterative co-assemblies and/or the IGC reference database was higher than 90%, thus demonstrating that the IMP-based iterative co-assemblies allow at least 10% of additional data to be mapped when using these assemblies in addition to the IGC reference database. In summary, the complementary use of de novo co-assembly of MG and MT datasets in combination with iterative assemblies enhances overall MG and MT data usage and thereby significantly increases the yield of useable information, especially when combined with comprehensive reference catalogs such as the IGC reference database.

Assembly quality: multi-omic iterative co-assembly

In order to compare the quality of the IMP-based iterative co-assembly procedure to simple co-assemblies, we compared the IMP-based iterative co-assemblies against co-assemblies generated using MetAMOS [10] (henceforth referred to as MetAMOS_MGMT) and MOCAT [34] (henceforth referred to as MOCAT_MGMT). Although MetAMOS and MOCAT were developed for MG data analysis, we extended their use for obtaining MG and MT co-assemblies by including both MG and MT read libraries as input (section “Execution of pipelines”). The assemblies were assessed based on contiguity (N50 length), data usage (MG and MT reads mapped), and output volume (number of contigs above 1 kb and number of genes; Additional file 2: Table S5). Only the SM dataset allowed for ground truth-based assessment by means of aligning the generated de novo assembly contigs to the original 73 bacterial genomes used to simulate the data set (section “Simulated coupled metagenomic and metatranscriptomic dataset”) [12, 54]. This allowed the comparison of two additional quality metrics, i.e., the recovered genome fraction and the composite performance metric (CPM) proposed by Deng et al. [62].

Table 2 Mapping statistics for human microbiome samples

Reference	Average MG pairs mapping (%)	Average MT pairs mapping (%)
IGC	70.91	53.57
IMP	70.25	86.21
IMP-megahit	70.62	83.33
IMP_MG	68.08	58.54
MetAMOS_MG	57.31	37.34
MOCAT_MG	36.73	36.68
IMP + IGC	92.66	95.77
IMP-megahit + IGC	92.80	93.24

Average fractions (%) of properly paired reads from the human microbiome datasets (HF1–5) mapping to various references, including IMP-based iterative co-assemblies (IMP and IMP-megahit) and single-omic co-assemblies (IMP_MG, MetAMOS_MG, and MOCAT_MG) as well as the IGC reference database. IMP + IGC and IMP-megahit + IGC reports the total number of properly paired reads mapping to IMP-based iterative co-assemblies and/or the IGC reference database. Refer to Additional file 2: Table S3 for detailed information

Assessments based on real datasets demonstrate comparable performance between IMP and IMP-megahit while both outperform MetAMOS_MGMT and MOCAT_MGMT in all measures (Fig. 4a–c). The ground truth assessment using the SM dataset shows that IMP-based iterative co-assemblies are effective in recovering the largest fraction of the original reference genomes while achieving a higher CPM score compared to co-assemblies from the other pipelines. Misassembled (chimeric) contigs are a legitimate concern within extensive de novo assembly procedures such as the IMP-based iterative co-assembly. It has been previously demonstrated that highly contiguous assemblies (represented by high N50 lengths) tend to contain higher absolute numbers of misassembled contigs compared to highly fragmented assemblies, thereby misrepresenting the actual quality of assemblies [38, 62, 63]. Therefore, the CPM score was devised as it represents a normalized measure reflecting both contiguity and accuracy for a given assembly [62]. Based on the CPM score, both IMP and IMP-megahit yield assemblies that balance high contiguity with accuracy and thereby outperform the other methods (Fig. 4c, d). In summary, cumulative measures of numbers of contigs ≥1 kb, N50 lengths, numbers of unique genes, recovered genome fractions (%), and CPM scores (the latter two were only calculated for the SM dataset), as well as the mean fractions (%) of mappable MG and MT reads, show that the IMP-based iterative co-assemblies (IMP and IMP-megahit) clearly outperform all other available methods (Fig. 4e; Additional file 2: Table S5).

Use-cases of integrated metagenomic and metatranscriptomic analyses in IMP

The integration of MG and MT data provides unique opportunities for uncovering community- or population-specific traits, which cannot be resolved from MG or MT data alone. Here we provide two examples of

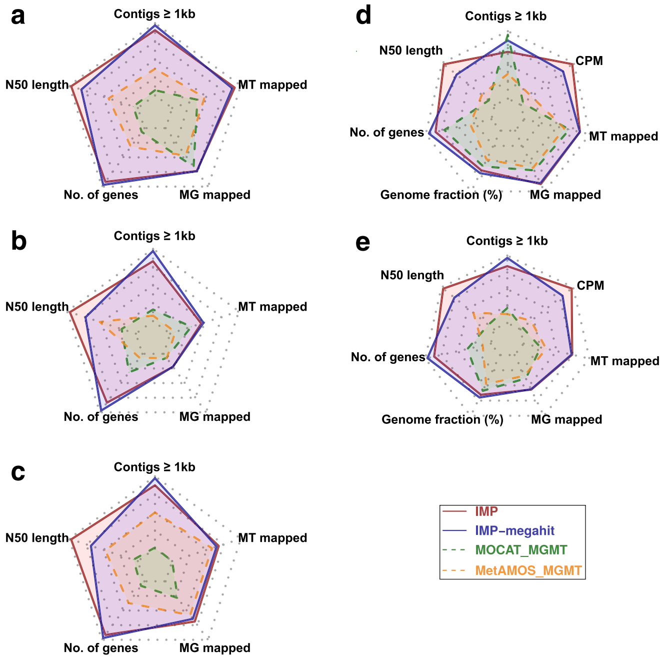


Fig. 4 Assessment of the IMP-based iterative co-assemblies in comparison to MOCAT- and MetAMOS-based co-assemblies. Radar charts summarizing the characteristics of the co-assemblies generated using IMP, MetAMOS, and MOCAT pipelines on: **a** human fecal microbiome, **b** wastewater sludge community, **c** biogas reactor, **d** simulated mock community. IMP co-assemblies were performed with two de novo assembler options, IDBA_UD and MEGAHIT, whereas MetAMOS and MOCAT were executed using default settings. Assessment metrics within the radar charts include number of contigs ≥1 kb, N50 length (contiguity, cutoff 500 bp), number of predicted genes (unique), and fraction of properly mapped MG and MT read pairs. N50 statistics are reported using a 500-bp cutoff. Additional ground truth assessments for simulated mock dataset included recovered genome fractions (%) and the composite performance metric (CPM) score with a cutoff of 500 bp [62]. **e** Summary radar chart reflecting the cumulative measures and mean fraction of properly mapped MG and MT read pairs from all analyzed 11 datasets while incorporating ground truth-based measures from the simulated mock dataset. Higher values within the radar charts (furthest from center) represent better performance. Detailed information on the assembly assessments is available in Additional file 2: Table S5

insights gained through the direct inspection of results provided by IMP.

Tailored preprocessing and filtering of MG and MT data

The preprocessing of the datasets HF1–5 included filtering of human-derived sequences, while the same step was not necessary for the non-human-derived datasets, WW1–4 and BG. MT data analyzed within this article included RNA extracts which were not subjected to wet-lab rRNA depletion, i.e., BG [29], and samples which were treated with wet-lab rRNA removal kits (namely HF1–5 [28] and WW1–4 [43]). Overall, the removal of rRNA pairs from the MT data showed a large variation, ranging from as low as 0.51% (HF5) to 60.91% (BG), demonstrating that wet-lab methods vary in terms of effectiveness and highlighting the need for such MT-specific filtering procedures (Additional file 1: Note S2; Additional file 2: Table S6).

Identification of RNA viruses

To identify differences in the information content of MG and MT complements, the contigs generated using IMP were inspected with respect to coverage by MG and MT reads (Additional file 2: Table S7). In two exemplary datasets HF1 and WW1, a small fraction of the contigs resulted exclusively from MT data (Additional file 2: Table S7). Longer contigs (≥1 kb) composed exclusively of MT reads and annotated with known viral/bacteriophage genes were retained for further inspection (Table 3; complete list contigs in Additional file 2: Table S8

Table 3 Contigs with a likely viral/bacteriophage origin/function reconstructed from the metatranscriptomic data

Sample	Contig ID*	Contig length	Average contig depth of coverage	Gene product	Average gene depth of coverage
HF1	Contig_34	6468	20927	Virus coat protein (TMV like)	30668
				Viral movement protein (MP)	26043
				RNA-dependent RNA polymerase	22578
				Viral methyltransferase	18817
	Contig_13948	2074	46	RNA-dependent RNA polymerase	41
				Viral movement protein (MP)	56
WW2	Contig_6405	4062	46	Tombusvirus p33	43
				Viral RNA-dependent RNA polymerase	42
				Viral coat protein (S domain)	36
	Contig_7409	3217	21	Viral RNA-dependent RNA polymerase	18
				Viral coat protein (S domain)	21
	Contig_7872	2955	77	Hypothetical protein	112
				Phage maturation protein	103

*Contigs of ≥1 kb and average depth of coverage ≥20 were selected

and S9). A subsequent sequence similarity search against the NCBI NR nucleotide database [64] of these candidate contigs revealed that the longer contigs represent almost complete genomes of RNA viruses (Additional file 2: Table S10 and S11). This demonstrates that the incorporation of MT data and their contrasting to the MG data allow the identification and recovery of nearly complete RNA viral genomes, thereby allowing their detailed future study in a range of microbial ecosystems.

Identification of populations with apparent high transcriptional activity

To further demonstrate the unique analytical capabilities of IMP, we aimed to identify microbial populations with a high transcriptional activity in the HF1 human fecal microbiome sample. Average depth of coverage at the contig- and gene-level is a common measure used to evaluate the abundance of microbial populations within communities [14, 16, 43]. The IMP-based integrative analysis of MG and MT data further extends this measure by calculation of average MT to MG depth of coverage ratios, which provide information on transcriptional activity and which can be visualized using augmented VizBin maps [56].

In our example, one particular cluster of contigs within the augmented VizBin maps exhibited high MT to MG depth of coverage ratios (Additional file 1: Figure S3). The subset of contigs within this cluster aligned to the genome of the *Escherichia coli* P12B strain (henceforth referred to as *E. coli*). For comparison, we also identified a subset, which was highly abundant at the MG level (lower MT to MG ratio), which aligned to the genome of *Collinsella intestinalis* DSM 13280 strain (henceforth referred to as *C. intestinalis*). Based on these observations, we highlighted the subsets of these contigs in an augmented VizBin map (Fig. 5a). The *C. intestinalis* and *E. coli* subsets are mainly represented by clear peripheral clusters which exhibit consistent intra-cluster MT to MG depth of coverage ratios (Fig. 5a). The subsets were manually inspected in terms of their distribution of average MG and MT depths of coverage and were compared against the corresponding distributions for all contigs. The MG-based average depths of coverage of the contigs from the entire community exhibited a bell-shape like distribution, with a clear peak (Fig. 5b). In contrast, MT depths of coverage exhibited more spread, with a relatively low mean (compared to MG distribution) and no clear peak (Fig. 5b). The *C. intestinalis* subset displays similar distributions to that of the entire community, whereas the *E. coli* subset clearly exhibits unusually high MT-based and low MG-based depths of coverage (Fig. 5b). Further inspection of the individual omic datasets revealed that the *E. coli* subset was not covered by the MG contigs, while approximately 80% of the *E. coli* genome was recoverable from a single-omic MT assembly (Fig. 5c). In contrast, the *C. intestinalis* subset demonstrated genomic recovery in all co-assemblies (IMP, IMP-megahit, MOCAT_MGMT, MetAMOS_MGMT) and the single-omic MG assemblies (IMP_MG, MOCAT_MG, MetAMOS_MG; Fig. 5c).

As noted by the authors of the original study by Franzosa et al. [28], the cDNA conversion protocol used to produce the MT data is known to introduce approximately 1–2% of *E. coli* genomic DNA into the cDNA as contamination which is then reflected in the MT data. According to our analyses, 0.12% of MG reads and

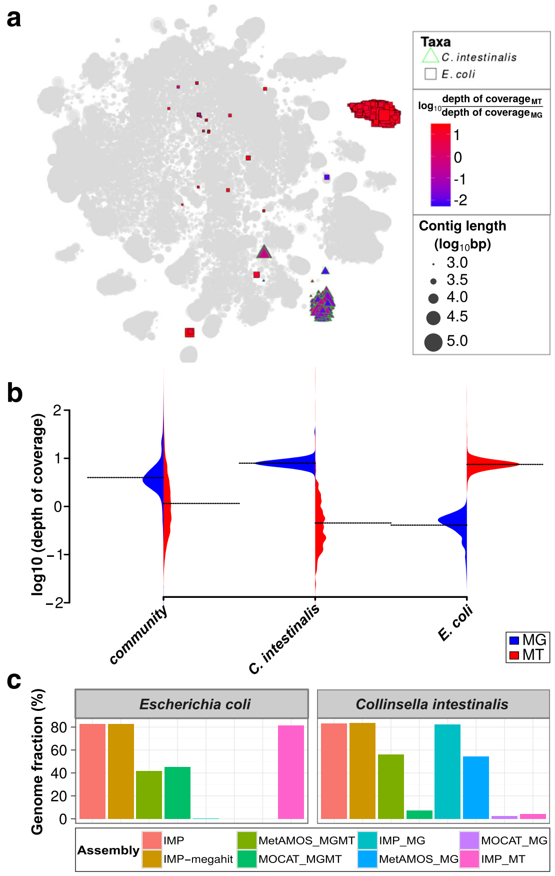


Fig. 5 Metagenomic and metatranscriptomic data integration of a human fecal microbiome. **a** Augmented VizBin map highlighting contig subsets with sequences that are most similar to *Escherichia coli* P12b and *Collinsella intestinalis* DSM 13280 genomes. **b** Beanplots representing the densities of metagenomic (*MG*) and metatranscriptomic (*MT*) average contig-level depth of coverage for the entire microbial community and two subsets (population-level genomes) of interest. The *dotted lines* represent the mean. **c** Recovered portion of genomes of the aforementioned taxa based on different single-omic assemblies and multi-omic co-assemblies (Additional file 2: Table S5)

1.95% of MT reads derived from this sample could be mapped onto the *E. coli* contigs, which is consistent with the numbers quoted by Franzosa et al. [28].

Consistent recovery of the *E. coli* genome was also observed across all other assemblies of the human fecal microbiome datasets (HF2–5) which included their respective MT data (Additional file 1: Figure S4; Additional file 2: Table S12). The integrative analyses of MG and MT data within IMP enables users to efficiently highlight notable cases such as this and to further investigate inconsistencies and/or interesting characteristics within these multi-omic datasets.

Discussion

The microbiome analysis workflow of IMP is unique in that it allows the integrated analysis of MG and MT data. To the best of our knowledge, IMP represents the only pipeline that spans the preprocessing of NGS reads to the binning of the assembled contigs, in addition to being the first automated pipeline for reproducible reference-independent metagenomic and metatranscriptomic data analysis. Although existing pipelines such as MetAMOS or MOCAT may be applied to perform co-assemblies of MG and MT data [44], these tools do not include specific steps for the two data types in their pre- and post-assembly procedures, which is important given the disparate nature of these datasets. The use of Docker promotes reproducibility and sharing, thereby allowing researchers to precisely replicate the IMP workflow with relative ease and with minimal impact on overall performance of the employed bioinformatic tools [29, 46–48]. Furthermore, static websites will be created and associated with every new version of IMP (Docker image), such that users will be able to download and launch specific versions of the pipeline to reproduce the work of others. Thereby, IMP enables standardized comparative studies between datasets from different labs, studies, and environments. The open source nature of IMP encourages a community-driven effort to contribute to and further improve the pipeline. Snakemake allows the seamless integration of Python code and shell (bash) commands and the use of *make* scripting style, which are arguably some of the most widely used bioinformatic scripting languages. Snakemake also supports parallel processing and the ability to interoperate with various tools and/or web services [49, 51]. Thus, users will be able to customize and enhance the features of the IMP according to their analysis requirements with minimal training/learning.

Quality control of NGS data prior to de novo assemblies has been shown to increase the quality of downstream assembly and analyses (predicted genes) [63]. In addition to standard preprocessing procedures (i.e., removal of low quality reads, trimming of adapter sequences and removal), IMP incorporates additional tailored and customizable filtering procedures which account for the different sample and/or omic data types. For instance, the removal of host-derived sequences in the context of human microbiomes is required for protecting the privacy of study subjects. The MT-specific in silico rRNA removal procedure yielded varying fractions of rRNA reads between the different MT datasets despite the previous depletion of rRNA (section “Tailored preprocessing and filtering of MG and MT data”), indicating that improvements in wet-lab protocols are necessary. Given that rRNA sequences are known to be highly similar, they are removed in IMP in order to mitigate any possible misassemblies resulting from such reads and/or regions [65, 66]. In summary, IMP is designed to perform stringent and standardized preprocessing of MG and MT data in a data-specific way, thereby enabling efficient data usage and resulting in high-quality output.

It is common practice that MG and MT reads are mapped against a reference (e.g., genes, genomes, and/or MG assemblies) [28, 29, 40] prior to subsequent data interpretation. However, these standard practices lead to suboptimal usage of the original data. IMP enhances overall data usage through its specifically tailored iterative co-assembly procedure, which involves four measures to achieve better data usage and yield overall larger volumes of output (i.e., a larger number of contigs ≥1 kb and predicted unique and complete genes).

First, the iterative assembly procedure leads to increases in data usage and output volume in each additional iterative assembly step (section “Data usage: iterative assembly”). The exclusion of mappable reads in each iteration of the assembly serves as a means of partitioning the data, thereby reducing the complexity of the data and overall, resulting in a higher cumulative volume of output [60, 63, 67].

Second, the initial assembly of MT-based contigs enhances the overall assembly, as transcribed regions are covered much more deeply and evenly in MT data, resulting in better assemblies for these regions [43]. The MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data.

Third, the co-assembly of MG and MT data allows the integration of these two data types while resulting in a larger number of contigs and predicted complete genes against which, in turn, a substantially higher fraction of reads can be mapped (section “Data usage: multi-omic iterative co-assembly”). Furthermore, the analyses of the human fecal microbiome datasets (HF1–5) demonstrate that the numbers of MG reads mapping to the IMP-based iterative co-assemblies for each sample are comparable to the numbers of reads mapping to the comprehensive IGC reference database (Table 2). Previously, only fractions of 74–81% of metagenomic reads mapping to the IGC have been reported [35]. However, such numbers have yet to be reported for MT data, in which case we observe lower mapping rates to the IGC reference database (35.5–70.5%) compared to IMP-based assemblies (Additional file 2: Table S3). This may be attributed to the fact that the IGC reference database was generated from MG-based assemblies only, thus creating a bias [35]. Moreover, an excess of 90% of MG and MT reads from the human fecal datasets (HF1–5) are mappable to either the IGC reference database and/or IMP-based iterative co-assemblies, emphasizing that a combined reference-based and IMP-based integrated-omics approach vastly improves data usage (Table 2). Although large fractions of MG and/or MT reads can be mapped to the IGC, a significant advantage of using a de novo reference-independent approach lies within the fact that reads can be linked to genes within their respective genomic context and microbial populations of origin.

Exploiting the maximal amount of information is especially relevant for microbial communities with small sample sizes and which lack comprehensive references such as the IGC reference database.

Fourth, the assembly refinement step via a contig-level assembly with cap3 improves the quality of the assemblies by reducing redundancy and increasing contiguity by collapsing and merging contigs (section "Assembly quality: multi-omic iterative co-assembly"). Consequently, our results support the described notion that the sequential use of multi-*k*mer-based de Bruijn graph assemblers, such as IDBA-UD and MEGAHIT, with overlap-layout-consensus assemblers, such as cap3, result in improved MG assemblies [38, 62] but importantly also extend this to MG and MT co-assemblies.

When compared to commonly used assembly strategies, the IMP-based iterative co-assemblies consisted of a larger output volume while maintaining a relatively high quality of the generated contigs. High-quality assemblies yield higher quality taxonomic information and gene annotations while longer contigs (≥1 kb) are a prerequisite for unsupervised population-level genome reconstruction [14, 19, 56] and subsequent multi-omics data integration [39, 43, 44]. Throughout all the different comparative analyses which we performed, IMP performed more consistently across all the different datasets when compared to existing methods, thereby emphasizing the overall stability and broad range of applicability of the method (section "Assembly quality: multi-omic iterative co-assembly").

Integrated analyses of MG and MT data with IMP provide the opportunity for analyses that are not possible based on MG data alone, such as the detection of RNA viruses (section "Identification of RNA viruses") and the identification of transcriptionally active populations (section "Identification of populations with apparent high transcriptional activity"). The predicted/annotated genes may be used for further analyses and integration of additional omic datasets, most notably metaproteomic data [39, 43, 44]. Furthermore, the higher number of complete genes improves the downstream functional analysis, because the read counts per gene will be much more accurate when having full length transcript sequences and will increase the probability to identify peptides. More specifically, the large number of predicted genes may enhance the usage of generated metaproteomic data, allowing more peptides, and thus proteins, to be identified.

Conclusions

IMP represents the first self-contained and standardized pipeline developed to leverage the advantages associated with integrating MG and MT data for large-scale analyses of microbial community structure and function in situ [4, 6]. IMP performs all the necessary large-scale bioinformatic analyses, including preprocessing, assembly, binning (automated), and analyses within an automated, reproducible, and user-friendly pipeline. In addition, we demonstrate that IMP vastly enhances data usage to produce high-volume and high-quality output. Finally, the combination of open development and reproducibility should promote the general paradigm of reproducible research within the microbiome research community.

Methods

The details of the IMP workflow, implementation, and customizability are described in further detail. We also describe the additional analyses carried out for assessment and benchmarking of IMP.

Details of the IMP implementation and workflow

A Python (v3) wrapper script was implemented for user-friendly execution of IMP via the command line. The full list of dependencies, parameters (see below), and documentation is available on the IMP website (http://r3lab.uni.lu/web/imp/doc.html). Although IMP was designed specifically for integrated analysis of MG and MT data, it can also be used for single MG or MT analyses as an additional functionality.

Reproducibility

IMP is implemented around a Docker container that runs the Ubuntu 14.04 operating system, with all relevant dependencies. Five mounting points are defined for the Docker container with the -v option: i) input directory, ii) output directory, iii) database directory, iv) code directory, and v) configuration file directory. Environment variables are defined using the -e parameter, including: i) paired MG data, ii) paired MT data, and iii) configuration file. The latest IMP Docker image will be downloaded and installed automatically upon launching the command, but users may also launch specific versions based on tags or use modified/customized versions of their local code base (documentation at http://r3lab.uni.lu/web/imp/doc.html).

Automation and modularity

Automation of the workflow is achieved using Snakemake 3.4.2 [49, 51], a Python-based make language implemented specifically for building reproducible bioinformatic workflows and pipelines. Snakemake is inherently modular and thus allows various features to be implemented within IMP, including the options of i) executing specific/selected steps within the pipeline, ii) check-pointing, i.e., resuming analysis from a point of possible interruption/termination, iii) analysis of single-omic datasets (MG or MT). For more details regarding the functionalities of IMP, please refer to the documentation of IMP (http://r3lab.uni.lu/web/imp/doc.html).

Input data

The input to IMP includes MG and/or MT FASTQ paired files, i.e., pairs-1 and pairs-2 are in individual files. The required arguments for the IMP wrapper script are metagenomic paired-end reads ("-m" options) and/or metatranscriptomic paired-end reads ("-t" option) with the specified output folder ("-o" option). Users may customize the command with the options and flags described in the documentation (http://r3lab.uni.lu/web/imp/doc.html) and in the "Customization and further development" section.

Trimming and quality filtering

Trimmomatic 0.32 [52] is used to perform trimming and quality filtering of MG and MT Illumina paired-end reads, using the following parameters: ILLUMINACLIP: TruSeq3-PE.fa:2:30:10; LEADING:20; TRAILING:20; SLIDINGWINDOW:1:3; MAXINFO:40:0.5; MINLEN:40. The parameters may be tuned via the command line or within the IMP config file. The output from this step includes retained paired-end and single-end reads (mate discarded), which are all used for downstream processes. These parameters are configurable in the IMP config file (section "Customization and further development")

Ribosomal RNA filtering

SortMeRNA 2.0 [68] is used for filtering rRNA from the MT data. The process is applied on FASTQ files for both paired- and single-end reads generated from the trimming and quality filtering step. Paired-end FASTQ files are interleaved prior to running SortMeRNA. If one of the mates within the paired-end read is classified as an rRNA sequence, then the entire pair is filtered out. After running SortMeRNA, the interleaved paired-end output is split into two separate paired-end FASTQ files. The filtered sequences (without rRNA reads) are used for the downstream processes. All available databases provided within SortMeRNA are used for filtering and the maximum memory usage parameter is set to 4 GB (option: "-m 4000"), which can be adjusted in the IMP config file (section "Customization and further development").

Read mapping

The read mapping procedure is performed using the bwa mem aligner [69] with settings: " -v 1" (verbose output level), "-M" (Picard compatibility) introducing an automated samtools header using the "-R" option [69]. Paired- and single-end reads are mapped separately and the resulting alignments are merged (using samtools merge [70]). The output is written as a binary aligment map (BAM) file. Read mapping is performed at various steps in the workflow, including: i) screening for host or contaminant sequences (section "Screening host or contaminant sequences"), ii) recruitment of unmapped reads within the IMP-based iterative co-assembly (section "Extracting unmapped reads"), and iii) mapping of preprocessed MG and MT reads to the final contigs. The memory usage is configurable in the IMP config file (section "Customization and further development").

Extracting unmapped reads

The extraction of unmapped reads (paired- and single-end) begins by mapping reads to a given reference sequence (section "Read mapping"). The resulting BAM file is used as input for the extraction of unmapped reads. A set of paired-end reads are considered unmappable if both or either one of the mates do not map to the given reference. The unmapped reads are converted from BAM to FASTQ format using samtools [70] and BEDtools 2.17.0—bamToFastq utility [71]. Similarly, unmapped single-end reads are also extracted from the alignment information.

Screening host or contaminant sequences

By default, the host/contaminant sequence screening is performed by mapping both paired- and single-end reads (section "Read mapping") onto the human genome version 38 (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/), followed by extraction of unmapped reads (section "Extracting unmapped reads"). Within the IMP command line, users are provided with the option of i) excluding this procedure with the "--no-filtering" flag, ii) using other sequence(s) for screening by providing the FASTA file (or URL) using "--screen" option, or iii) specifying it in the configuration file (section "Customization and further development").

Parameters of the IMP-based iterative co-assembly

The IMP-based iterative co-assembly implements MEGAHIT 1.0.3 [23] as the MT assembler while IDBA-UD 1.1.1 [22] is used as the default co-assembler (MG and MT), with MEGAHIT [23] as an alternative option for the co-assembler (specified by the "-a" option of the IMP command line). All de novo assemblies are performed on *k*mers ranging from 25-mers to 99-mers, with an incremental step of four. Accordingly, the command line parameters for IDBA-UD are "--mink 25 --maxk 99 --step 4 --similar 0.98 --pre-correction" [22]. Similarly, the command line parameters for MEGAHIT are "--k-min 25 --k-max 99 --k-step 4", except for the MT assemblies which are performed with an additional "--no-bubble" option to prevent merging of bubbles within the assembly graph [23]. Furthermore, contigs generated from the MT assembly are used as "long read" input within the "-l" flag of IDBA-UD or "-r" flag of MEGAHIT [22, 23]. *K*mer ranges for the IDBA-UD and MEGAHIT can be adjusted/specified in the configuration file (section "Customization and further development"). Cap3 is used to reduce the redundancy and improve contiguity of the assemblies using a minimum alignment identity of 98% ("-p 0.98") with a minimum overlap of 100 bases ("-o 100"), which are adjustable in the configuration file (section "Customization and further development"). Finally, the extraction of reads that are unmappable to the initial MT assembly and initial co-assembly is described in the "Extracting unmapped reads" section.

Annotation and assembly quality assessment

Prokka 1.11 [55] with the "--metagenome" setting is used to perform functional annotation. The default BLAST and HMM databases of Prokka are used for the functional annotation. Custom databases may be provided by the user (refer to the "Databases" and "Customization and further development" sections for details).

MetaQUAST 3.1 [54] is used to perform taxonomic annotation of contigs with the maximum number of downloadable reference genomes set to 20 ("--max-ref-number 20"). In addition, MetaQUAST provides various assembly statistics. The maximum number of downloadable reference genomes can be changed in the IMP config file (see "Customization and further development" for details).

Depth of coverage

Contig- and gene-wise depth of coverage values are calculated (per base) using BEDtools 2.17.0 [71] and aggregated (by average) using awk, adapted from the CONCOCT code [16] (script: map-bowtie2-markduplicates.sh; https://github.com/BinPro/CONCOCT) and is non-configurable.

Variant calling

The variant calling procedure is performed using Samtools 0.1.19 [70] (mpileup tool) and Platypus 0.8.1 [72], each using their respective default settings and which are non-configurable. The input is the merged paired- and single-end read alignment (BAM) against the final assembly FASTA file (section "Read mapping"). The output files from both the methods are indexed using tabix and compressed using gzip. No filtering is applied to the variant calls, so that users may access all the information and filter it according to their requirements. The output from samtools mpileup is used for the augmented VizBin visualization.

Non-linear dimensionality reduction of genomic signatures

VizBin [56] performs non-linear dimensionality reduction of genomic signatures onto contigs ≥1 kb, using default settings, to obtain two-dimensional embeddings. Parameters can be modified in the IMP config file (section "Customization and further development").

Automated binning

Automated binning of the assembled contigs is performed using MaxBin 2.0. Default setting are applied and paired-end reads are provided as input for abundance estimation [20]. The sequence length cutoff is set to be same as VizBin (section "Non-linear dimensionality reduction of genomic signatures") and is customizable using the config file (section "Customization and further development").

Visualization and reporting

IMP compiles the multiple summaries and visualizations into a HTML report [57]. FASTQC [73] is used to visualize the quality and quantity of reads before and after preprocessing. MetaQUAST [54] is used to report assembly quality and taxonomic associations of contigs. A custom script is used to generate KEGG-based [74] functional Krona plots by running KronaTools [75] (script: genes.to.kronaTable.py, GitHub URL: https://github.com/EnvGen/metagenomics-workshop). Additionally, VizBin output (two-dimensional embeddings) is integrated with the information derived from the IMP analyses, using a custom R script for analysis and visualization of the augmented maps. The R workspace image is saved such that users are able to access it for further analyses. All the steps executed within an IMP run, including parameters and runtimes, are summarized in the form of a workflow diagram and a log-file. The visualization script is not configurable.

Output

The output generated by IMP includes a multitude of large files. Paired- and single-end FASTQ files of preprocessed MG and MT reads are provided such that the user may employ them for additional downstream analyses. The output of the IMP-based iterative co-assembly consists of a FASTA file, while the alignments/mapping of MG and MT preprocessed reads to the final co-assembly are also provided as BAM files, such that users may use these for further processing. Predicted genes and their respective annotations are provided in the various formats produced by Prokka [55]. Assembly quality statistics and taxonomic annotations of contigs are provided as per the output of MetaQUAST [54]. Two-dimensional embeddings from the NLDR-GS are provided such that they can be exported to and further curated using VizBin [56]. Additionally, abundance and expression information is represented by contig- and gene-level average depth of coverage values. MG and MT genomic variant information (VCF format), including both SNPs and INDELs (insertions and deletions), is also provided. The results of the automated binning using MaxBin 2.0 [20] are provided in a folder which contains the default output from the program (i.e., fasta files of bins and summary files).

The HTML reports [57], e.g., HTML S1 and S2, compile various summaries and visualizations, including, i) augmented VizBin maps, ii) MG- and MT-level functional Krona charts [75], iii) detailed schematics of the steps carried out within the IMP run, iv) list of parameters and commands, and v) additional reports (FASTQC report [73], MetaQUAST report [54]). Please refer to the documentation of IMP for a detailed list and description of the output (http://r3lab.uni.lu/web/imp/doc.html).

Databases

The IMP database folder (db) contains required databases required for IMP analysis. The folder contains the following subfolders and files with their specific content:

i. adapters folder — sequencing adapter sequences. Default version contains all sequences provided by Trimmomatic version 0.32 [52]
ii. cm, genus, hmm, and kingdom folders — contains databases provided by Prokka 1.11 [55]. Additional databases may be added into the corresponding folders as per the instructions in the Prokka documentation (https://github.com/tseemann/prokka#databases)
iii. sortmerna folder — contains all the databases provided in SortMeRNA 2.0 [68]. Additional databases may be added into the corresponding folders as per the instructions in the SortMeRNA documentation (http://bioinfo.lifl.fr/RNA/sortmerna/code/SortMeRNA-user-manual-v2.0.pdf)
iv. ec2pathways.txt — enzyme commission (EC) number mapping of amino acid sequences to pathways
v. pathways2hierarchy.txt — pathway hierarchies used to generated for KEGG-based functional Krona plot (section “Visualization and reporting”)

Customization and further development

Additional advanced parameters can be specified via the IMP command line, including specifying a custom configuration file (“-c” option) and/or specifying a custom database folders (“-d” option). Threads (“--threads”) and memory allocation (“--memcore” and “--memtotal”) can be adjusted via the command line and the configuration file. The IMP launcher script provides a flag (“--enter”) to launch the Docker container interactively and the option to specify the path to the customized source code folder (“-s” option). These commands are provided for development and testing purposes (described on the IMP website and documentation: http://r3lab.uni.lu/web/imp/doc.html). Further customization is possible using a custom configuration file (JSON format). The customizable options within the JSON file are specified in individual subsections within the “Details of the IMP implementation and workflow” section. Finally, the open source implementation of IMP allows users to customize the Docker image and source code of IMP according to their requirements.

Iterative single-omic assemblies

In order to determine the opportune number of iterations within the IMP-based iterative co-assembly strategy an initial assembly was performed using IMP preprocessed MG reads with IDBA-UD [22]. Cap3 [53] was used to further collapse the contigs and reduce the redundancy of the assembly. This initial assembly was followed by a total of three assembly iterations, whereby each iteration was made up of four separate steps: i) extraction of reads unmappable to the previous assembly (using the procedure described in the “Extracting unmapped reads” section), ii) assembly of unmapped reads using IDBA-UD [22], iii) merging/collapsing the contigs from the previous assembly using cap3 [53], and iv) evaluation of the merged assembly using MetaQUAST [54]. The assembly was evaluated in terms of the per-iteration increase in mappable reads, assembly length, numbers of contigs ≥1 kb, and numbers of unique genes.

Similar iterative assemblies were also performed for MT data using MEGAHIT [23], except CD-HIT-EST [76] was used to collapse the contigs at ≥95% identity (“-c 0.95”) while MetaGeneMark [77] was used to predict genes. The parameters and settings of the other programs were the same as those defined in the “Details of the IMP implementation and workflow” section.

The aforementioned procedures were applied to all the datasets analyzed within this article. The merged contig sets (non-redundant) from the first iteration of both the MG and MT iterative assemblies were selected to represent the IMP single-omics assemblies (IMP_MG and IMP_MT) and were compared against co-assemblies.

Execution of pipelines

MetAMOS v1.5rc3 was executed using default settings. MG data were provided as input for single-omic assemblies (MetAMOS_MG) while MG and MT data were provided as input for multi-omic co-assemblies (MetAMOS_MGMT). All computations using MetAMOS were set to use eight computing cores (“-p 8”).

MOCAT v1.3 (MOCAT.pl) was executed using default settings. Paired-end MG data were provided as input for single-omic assemblies (MOCAT_MG) while paired-end MG and MT data were provided as input for multi-omic co-assemblies (MOCAT_MGMT). All computations using MOCAT were set to use eight computing cores (“-cpus 8”). Paired-end reads were first preprocessed using the read_trim_filter step of MOCAT (“-rtf”). For the human fecal microbiome datasets (HF1–5), the preprocessed paired- and single-end reads were additionally screened for human genome-derived sequences (“-s hg19”). The resulting reads were afterwards assembled with default parameters (“-gp assembly -r hg19”) using SOAPdenovo.

IMP v1.4 was executed for each dataset using different assemblers for the co-assembly step: i) default setting using IDBA-UD, and ii) MEGAHIT ("-a megahit"). Additionally, the analysis of human fecal microbiome datasets (HF1–5) included the preprocessing step of filtering human genome sequences, which was omitted for the wastewater sludge datasets (WW1–4) and the biogas (BG) reactor dataset. Illumina TruSeq2 adapter trimming was used for wastewater dataset preprocessing since the information was available. Computation was performed using eight computing cores ("- -threads 8"), 32 GB memory per core ("- -memcore 32") and total memory of 256 GB ("- -memtotal 256 GB"). The customized parameters were specified in the IMP configuration file (exact configurations listed in the HTML reports [57]). The analysis of the CAMI datasets were carried using the MEGAHIT assembler option ("-a megahit"), while the other options remained as default settings.

In addition, IMP was also used on a small scale dataset to evaluate performance of increasing the number of threads from 1 to 32 and recording the runtime ("time" command). IMP was launched on the AWS cloud computing platform running the MEGAHIT as the assembler ("-a megahit") with 16 threads ("- -threads 16") and 122 GB of memory ("- -memtotal 122").

Data usage assessment

Preprocessed paired-end and single-end MG and MT reads from IMP were mapped (section Read mapping) onto the IMP-based iterative co-assemblies and IMP_MG assembly. Similarly, preprocessed paired-end and single-end MG and MT reads from MOCAT were mapped onto the MOCAT co-assembly (MOCAT_MGMT) and the MOCAT single-omic MG assembly (MOCAT_MG). MetAMOS does not retain single-end reads; therefore, preprocessed MG and MT paired-end reads from MetAMOS were mapped onto the MetAMOS co-assembly (MetAMOS_MGMT) and MetAMOS single-omic MG assembly (MetAMOS_MG).

Preprocessed MG and MT reads from the human fecal datasets (HF1–5) were mapped using the same parameters described in the "Read mapping" section to the IGC reference database [35] for evaluation of a reference-based approach. Alignment files of MG and MT reads mapping to the IMP-based iterative co-assemblies and the aforementioned alignments to the IGC reference database were used to report the fractions of properly paired reads mapping in either IMP-based iterative co-assembly, IGC reference database, or both. These fractions were then averaged across all the human fecal datasets (HF1–5).

Assembly assessment and comparison

Assemblies were assessed and compared using MetaQUAST by providing contigs (FASTA format) from all different (single- and multi-omic) assemblies of the same dataset as input [54]. The gene calling function ("-f") was utilized to obtain the number of genes which were predicted from the various assemblies. An additional parameter within MetaQUAST was used for ground truth assessment of the simulated mock (SM) community assemblies by providing the list of 73 FASTA format reference genomes ("-R"). The CPM measure was computed based on the information derived from the results of MetaQUAST [54]. In order to be consistent with the reported values (i.e., N50 length), the CPM measures reported within this article are based on alignments of 500 bp and above, unlike the 1-kb cutoff used in the original work [62]. Prodigal was also used for gene prediction to obtain the number of complete and incomplete genes [61].

Analysis of contigs assembled from MT data

A list of contigs with no MG depth of coverage together with additional information on these contigs (contig length, annotation, MT depth of coverage) was retrieved using the R workspace image, which is provided as part IMP output (sections "Visualization and reporting" and "Output"). The sequences of these contigs were extracted and subjected to a BLAST search on NCBI to determine their potential origin. Furthermore, contigs with length ≥1 kb, average depth of coverage ≥20 bases, and containing genes encoding known virus/bacteriophage functions were extracted.

Analysis of subsets of contigs

Subsets of contigs within the HF1 dataset were identified by visual inspection of augmented VizBin maps generated by IMP. Specifically, detailed inspection of contig-level MT to MG depth of coverage ratios was carried out using the R workspace provided as part of IMP output (sections "Visualization and reporting" and "Output"). The alignment information of contigs to isolate genomes provided by MetaQUAST [54] was used to highlight subsets of contigs aligning to genomes of the *Escherichia coli* P12B strain (*E. coli*) and *Collinsella intestinalis* DSM 13280 (*C. intestinalis*).

An additional reference-based analysis of MetaQUAST [54] was carried out for all the human fecal microbiome assemblies (HF1–5) by providing the genomes of *E. coli* P12B and *C. intestinalis* DSM 13280 as reference (flag: "-R") to assess the recovery fraction of the aforementioned genomes within the different assemblies.

Computational platforms

IMP and MetAMOS were executed on a Dell R820 machine with 32 Intel(R) Xeon(R) CPU E5-4640 @ 2.40GHz physical computing cores (64 virtual), 1024 TB of DDR3 RAM (32 GB per core) with Debian 7 Wheezy as the operating system. MOCAT, IMP single-omic assemblies, and additional analyses were performed on the Gaia cluster of the University of Luxembourg HPC platform [78].

IMP was executed on the Amazon Web Services (AWS) cloud computing platform using EC2 R3 type (memory optimized) model r3.4xlarge instance with 16 compute cores, 122 GB memory, and 320 GB of storage space running a virtual Amazon Machine Image (AMI) Ubuntu v16.04 operating system.

Additional files

Additional file 1: Supplementary figures and notes. **Figures S1–S3** and **Notes S1–S2**. Detailed figure legends available within file. (PDF 1047 kb)

Additional file 2: Supplementary tables. **Tables S1–S12.** Detailed table legends available within file. (XLSX 4350 kb)

Abbreviations

AWS: Amazon Web Services; BAM: Binary Alignment Maps; BG: Biogas; bp: Base pair; CAMI: Critical Assessment of Metagenome Interpretation; cDNA: Complementary DNA; Contigs: Contiguous sequence(s); HF: Human fecal; IGC: Integrated Gene Catalog; IMP: Integrated Meta-omic Pipeline; INDELs: Insertions and deletions; kb: Kilo base; KEGG: Kyoto Encyclopedia of Genes and Genomes; MetaHIT: Metagenomics of the Human Intestinal Tract; MG: Metagenomic; MT: Metatranscriptomic; NCBI: National Center for Biotechnology Information; NGS: Next-generation sequencing; rRNA: Ribosomal RNA; SM: Simulated mock; SNPs: Single nucleotide polymorphisms; SRA: Sequence read archive; VCF: Variant call format; WW: Wastewater

Acknowledgements

We would like to acknowledge John Larsson from SciLifeLab (Sweden) for kindly providing the KEGG-based functional Krona plot scripts. Albi Celaj from the University of Toronto is thanked for supplying the in silico simulated metatranscriptomic data and the corresponding reference genomes. The University of Luxembourg High Performance Computing (HPC) facility is duly thanked for providing and maintaining the computing platform. The Reproducible Research Results (R3) team of the Luxembourg Centre for Systems Biomedicine is acknowledged for support of the project and for promoting reproducible research. Finally, we acknowledge the Luxembourg National Research Fund (FNR) for funding the project via the ATTRACT, AFR, CORE, INTER and PoC grants.

Funding

This work was supported by an ATTRACT program grant (A09/03), a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01), a proof-of-concept grant (PoC/13/02), an ERASysAPP grant (INTER/8888488), and CORE programme grant (CORE/15/BM/10404093) to PW, an Aide à la Formation Recherche (AFR) grant to SN (PHD-2014-1/7934898), and a CORE junior (C15/SR/10404839) to EELM, all funded by the Luxembourg National Research Fund (FNR).

Availability and requirements

All the data, software, and source code related to this manuscript are publicly available.

Coupled metagenomic and metatranscriptomic datasets

The published human fecal microbiome datasets (MG and MT) were obtained from NCBI Bioproject PRJNA188481 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA188481). They include samples from individuals X310763260, X311245214, X316192082, X316701492, and X317690558 [28], designated within this article as HF1–5, respectively. Only samples labeled as "Whole" (samples preserved by flash-freezing) were selected for analysis [28]. The published wastewater sludge microbial community datasets (MG and MT) were obtained from NCBI Bioproject with the accession code PRJNA230567 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA230567). These include samples A02, D32, D36, and D49, designated within this article as WW1–4, respectively [43]. The published biogas reactor microbial community data set (MG and MT) was obtained from the European Nucleotide Archive (ENA) project PRJEB8813 (http://www.ebi.ac.uk/ena/data/view/PRJEB8813) and is designated within this article as BG [29].

Simulated coupled metagenomic and metatranscriptomic dataset

The simulated MT data were obtained upon request from the original authors [12]. A complementary metagenome was simulated using the same set of 73 bacterial genomes used for the aforementioned simulated MT [12]. Simulated reads were obtained using the NeSSM MG simulator (default settings) [79]. The simulated mock community is designated as SM within this article [79]. The simulated data along with the corresponding reference genomes used to generate the MG data are made available via LCSB WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/) and is archived on Zenodo [80].

CAMI simulated community metagenomic datasets

The medium complexity CAMI simulated MG data and the corresponding gold standard assembly were obtained from the CAMI website (http://www.cami-challenge.org).

Test dataset for runtime assessment

A subset of ~5% of reads from both the WW1 MG and MT datasets (section "Coupled metagenomic and metatranscriptomic datasets") was selected and used as the data to perform runtime assessments. This dataset could be used to test IMP on regular platforms such as laptops and desktops. It is made available via the LCSB R3 WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/) and is archived on Zenodo [81].

Software and source code

IMP is available under the MIT license on the LCSB R3 website (http://r3lab.uni.lu/web/imp/), which contains necessary information related to IMP. These include links to the Docker images on the LCSB R3 WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/dist/) and is archived on Zenodo [82]. Source code is available on LCSB R3 GitLab (https://git-r3lab.uni.lu/IMP/IMP), GitHub (https://github.com/shaman-narayanasamy/IMP), and is archived on Zenodo [83]. Scripts and commands for additional analyses performed specifically within this manuscript are available on LCSB R3 GitLab (https://git-r3lab.uni.lu/IMP/IMP_manuscript_analysis) and on GitHub (https://github.com/shaman-narayanasamy/IMP_manuscript_analysis). Frozen pages containing all necessary material related to this article are available at http://r3lab.uni.lu/frozen/imp/.

Authors' contributions

SN, NP, EELM, PM, and PW conceived the analysis and designed the workflow. SN, YJ, MH, and CCL developed the software, wrote the documentation and tested the software. YJ ensured reproducibility of the software. SN, PM, and MH performed data analyses. EELM, PM, AHB, AK, NP, and PW participated in discussions and tested the software. SN, EELM, AHB, PM, NP, AK, MH, and PW wrote and edited the manuscript. PW designed and supported the project. All authors read and agreed on the final version of the manuscript.

Authors' information

Current affiliations: CCL—Saarland University, Building E2 1, 66123 Saarbrücken, Germany; NP—Universidad EAFIT, Carrera 49 No 7 sur 50, Medellín, Colombia; EELM—Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA—CNRS, Université de Strasbourg, Strasbourg, France.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur-Alzette L-4362, Luxembourg. [2]Present address: Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA—CNRS, Université de Strasbourg, Strasbourg, France. [3]Present address: Saarland University, Building E2 1, Saarbrücken 66123, Germany. [4]Institute of Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA. [5]Present address: Universidad EAFIT, Carrera 49 No 7 sur 50, Medellín, Colombia.

Received: 18 October 2016 Accepted: 22 November 2016
Published online: 16 December 2016

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007;449:804–10.
2. Rittmann BE. Microbial ecology to manage processes in environmental biotechnology. Trends Biotechnol. 2006;24:261–6.
3. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012;194:4151–60.
4. Narayanasamy S, Muller EEL, Sheik AR, Wilmes P. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. Microb Biotechnol. 2015;8:363–8.
5. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. Mol Syst Biol. 2013;9:666.
6. Muller EEL, Glaab E, May P, Vlassis N, Wilmes P. Condensing the omics fog of microbial communities. Trends Microbiol. 2013;21:325–33.
7. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P. A biomolecular isolation framework for eco-systems biology. ISME J. 2013;7:110–21.
8. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Methods Enzymol. 2013;531:219–36.
9. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10:1196–9.
10. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14:R2.
11. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics. 2011;12:41.
12. Celaj A, Markle J, Danska J, Parkinson J. Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. Microbiome. 2014;2:39.
13. Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. Sci Rep. 2014;4:4516.
14. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.
15. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.
16. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
17. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
18. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
19. Laczny CC, Muller EEL, Heintz-Buschart A, Herold M, Lebrun LA, Hogan A, May P, De Beaufort C, Wilmes P. Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract. Front Microbiol. 2016;7:884.
20. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW, Metzker M, Dick G, Andersson A, Baker B, Simmons S, Thomas B, Yelton A, Banfield J, Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J, Mackelprang R, Waldrop M, DeAngelis K, David M, Chavarria K, Blazewicz S, Rubin E, et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014;2:26.
21. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2:e603.
22. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
23. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
24. Westreich ST, Korf I, Mills DA, Lemay DG, Moran M, Leimena M, Embree M, McGrath K, Dimitrov D, Cho I, Blaser M, Round J, Mazmanian S, Gosalbes M, Giannoukos G, Reck M, Hainzl E, Bolger A, Lohse M, Usadel B, Magoc T, Salzberg S, Meyer F, Tatusova T, Wilke A, Overbeek R, Love M, Huber W, Anders S, Costa V, et al. SAMSA: a comprehensive metatranscriptome analysis pipeline. BMC Bioinformatics. 2016;17:399.
25. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, Azpiroz F, Guarner F, Manichanh C, Li J, Gosalbes MJ, Helbling DE, Ackermann M, Fenner K, Kohler HP, Johnson DR, Tulin S, Aguiar D, Istrail S, Smith J, Leimena MM, He S, Murakami S, Fujishima K, Tomita M, Kanai A, Manichanh C, Li R, McDonald D, Wilke A, et al. MetaTrans: an open-source pipeline for metatranscriptomics. Sci Rep. 2016;6:26447.
26. Leimena MM, Ramiro-Garcia J, Davids M, van den Bogert B, Smidt H, Smid EJ, Boekhorst J, Zoetendal EG, Schaap PJ, Kleerebezem M. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics. 2013;14:530.
27. Satinsky BBM, Fortunato CS, Doherty M, Smith CBC, Sharma S, Ward NDNND, Krusche AAV, Yager PL, Richey JE, Moran MA, Crump BBC, Richey JE, Devol A, Wofsy S, Victoria R, Riberio M, Nebel G, Dragsted J, Vega A, Hedges J, Clark W, Quay P, Richey JE, Devol A, Santos U, Spencer R, Hernes P, Aufdenkampe A, Baker A, Gulliver P, et al. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. Microbiome. 2015;3:39.
28. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111:E2329–38.
29. Bremges A, Maus I, Belmann P, Eikmeyer F, Winkler A, Albersmeier A, Pühler A, Schlüter A, Sczyrba A. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. Gigascience. 2015;4:33.
30. Leung HCM, Yiu S-M, Parkinson J, Chin FYL. IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology. J Comput Biol. 2013;20:540–50.
31. Leung HCM, Yiu SM, Chin FYL. IDBA-MTP: a hybrid metatranscriptomic assembler based on protein information. Res Comput Mol Biol. 2014; 160–172.
32. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40;e155.
33. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Nusbaum C, Lindblad-toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
34. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7:e47656.
35. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41.
36. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
37. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu

S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
38. Lai B, Wang F, Wang X, Duan L, Zhu H. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics. 2015;16:244.
39. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2:16180.
40. Hultman J, Waldrop MP, Mackelprang R, David MM, Mcfarland J, Blazewicz SJ, Harden J, Turetsky MR, Mcguire AD, Shah MB, Verberkmoes NC, Lee LH. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521:208–12.
41. Beulig F, Urich T, Nowak M, Trumbore SE, Gleixner G, Gilfillan GD, Fjelland KE, Küsel K. Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO2 exposure. Nat Microbiol. 2016;1:15025.
42. Urich T, Lanzén A, Stokke R, Pedersen RB, Bayer C, Thorseth IH, Schleper C, Steen IH, Ovreas L. Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. Environ Microbiol. 2014;16:2699–710.
43. Muller EEL, Pinel N, Laczny CC, Hoopman MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, Heintz-Buschart A, Wampach L, Liu CM, Price LB, Gillece JD, Guignard C, Schupp JM, Vlassis N, Baliga NS, Moritz RL, Keim PS, Wilmes P. Community integrated omics links the dominance of a microbial generalist to fine-tuned resource usage. Nat Commun. 2014;5:5603.
44. Roume H, Heintz-Buschart A, Muller EEL, May P, Satagopam VP, Laczny CC, Narayanasamy S, Lebrun LA, Hoopmann MR, Schupp JM, Gillece JD, Hicks ND, Engelthaler DM, Sauter T, Keim PS, Moritz RL, Wilmes P. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. npj Biofilms Microbiomes. 2015;1:15007.
45. Kenall A, Edmunds S, Goodman L, Bal L, Flintoft L, Shanahan DR, Shipley T. Better reporting for better research: a checklist for reproducibility. BMC Neurosci. 2015;16:44.
46. Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. Bioboxes: standardised containers for interchangeable bioinformatics software. Gigascience. 2015;4:47.
47. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.
48. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2016. http://bib.oxfordjournals.org/content/early/2016/03/23/bib.bbw020.full.
49. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
50. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, Kartashov A, Leehr D, Ménager H, Nedeljkovich M, Scales M, Soiland-Reyes S, Stojanovic L. Common Workflow Language, v1.0. 2016. https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156.
51. Koster J. Reproducibility in next-generation sequencing analysis. 2014.
52. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
53. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9:868–77.
54. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2015;32:1088–90.
55. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
56. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, der Maaten L, Vlassis N, Wilmes P. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3:1.
57. IMP HTML reports. October 17, 2016. http://dx.doi.org/10.5281/zenodo.161321.
58. Schürch AC, Schipper D, Bijl MA, Dau J, Beckmen KB, Schapendonk CME, Raj VS, Osterhaus ADME, Haagmans BL, Tryland M, Smits SL. Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. PLoS One. 2014;9:e105227.
59. Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW, Rohwer F, Gordon JI. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A. 2015;112:11941–6.
60. Hitch T, Creevey C. Spherical: an iterative workflow for assembling metagenomic datasets. bioRxiv. 2016. http://biorxiv.org/content/early/2016/08/02/067256.
61. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ, Delcher A, Bratke K, Powers E, Salzberg S, Lukashin A, Borodovsky M, Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E, Larsen T, Krogh A, Zhu H, Hu G, Yang Y, Wang J, She Z, Ou H, Guo F, Zhang C, Tech M, Pfeifer N, Morgenstern B, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
62. Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu Y, Delwart EL. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res. 2015;43:e46.
63. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J, Bork P. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012;7:e31386.
64. Pruitt K, Brown G, Tatusova T, Maglott D. The Reference Sequence (RefSeq) Database. In: NCBI Handbook. 2002. p. 1–24.
65. Salzberg SL, Yorke JA. Beware of mis-assembled genomes. Bioinformatics. 2005;21:4320–1.
66. Mariano DCB, Sousa Tde J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TDL, Dorella FA, de Carvalho AF, Leal CAG, Figueiredo HCP, Silva A, Ramos RTJ, Azevedo VAC, Dorella F, Pacheco LC, Oliveira S, Miyoshi A, Azevedo V, Aleman M, Spier S, Wilson W, Doherr M, Soares S, Silva A, Trost E, Blom J, Ramos R, et al. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. BMC Genomics. 2016;17:315.
67. Hug LA, Thomas BC, Sharon I, Brown CT, Sharma R, Hettich RL, Wilkins MJ, Williams KH, Singh A, Banfield JF. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ Microbiol. 2016;18:159–73.
68. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
69. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:589–95.
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
72. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
73. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7:e30619.
74. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
75. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.
76. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
77. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38:e132.
78. Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an Academic HPC Cluster: the UL Experience. In: Proceedings of the 2014 International Conference on High Performance Computing Simulation. 2014. p. 959–67.
79. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: a next-generation sequencing simulator for metagenomics. PLoS One. 2013;8:e75448.
80. IMP simulated mock community data set. October 12, 2016. http://doi.org/10.5281/zenodo.160261.
81. IMP small scale test dataset. October 14, 2016. http://doi.org/10.5281/zenodo.160708.
82. IMP v1.4 docker image. October 12, 2016. http://doi.org/10.5281/zenodo.160263.
83. IMP v1.4 source code. October 14, 2016. http://doi.org/10.5281/zenodo.160703.

C.6 Using metabolic networks to resolve ecological properties of microbiomes.

Emilie E.L. Muller, Karoline Faust, Stefanie Widder, **Malte Herold**, Susana Martinez Arbas, Paul Wilmes

2018

Current Opinion in Systems Biology **8**: 73-80

DOI: 10.1016/j.coisb.2017.12.004

Contributions of author include:

- Writing and revision of manuscript

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Systems Biology

Using metabolic networks to resolve ecological properties of microbiomes

Emilie E. L. Muller[1,2], Karoline Faust[3], Stefanie Widder[4,5,6], Malte Herold[1], Susana Martínez Arbas[1] and Paul Wilmes[1]

Abstract

The systematic collection, integration and modelling of high-throughput molecular data (multi-omics) allows the detailed characterisation of microbiomes *in situ*. Through metabolic trait inference, metabolic network reconstruction and modelling, we are now able to define ecological interactions based on metabolic exchanges, identify keystone genes, functions and species, and resolve ecological niches of constituent microbial populations. The resulting knowledge provides detailed information on ecosystem functioning. However, as microbial communities are dynamic in nature the field needs to move towards the integration of time- and space-resolved multi-omic data along with detailed environmental information to fully harness the power of community- and population-level metabolic network modelling. Such approaches will be fundamental for future targeted management strategies with wide-ranging applications in biotechnology and biomedicine.

Addresses
[1] Eco-Systems Biology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, 7 Avenue des Hauts-Fourneaux, Esch-sur-Alzette, L-4362, Luxembourg
[2] Equipe Adaptations et Interactions Microbiennes, Université de Strasbourg, UMR 7156 UNISTRA–CNRS Génétique Moléculaire, Génomique, Microbiologie, Strasbourg, France
[3] Department of Microbiology and Immunology, Rega Institute KU Leuven, Leuven, Belgium
[4] CeMM-Reseach Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, 1090 Vienna, Austria
[5] Department of Medicine 1, Research Laboratory of Infection Biology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria
[6] Konrad Lorenz Institute for Evolution and Cognition Research, Martinstr. 12, 4300 Klosterneuburg, Austria

Corresponding author: Wilmes, Paul (paul.wilmes@uni.lu)

Current Opinion in Systems Biology 2018, **8**:73–80

This review comes from a themed issue on **Regulatory and metabolic networks (2018)**

Edited by **Bas Teusink** and **Uwe Sauer**

For a complete overview see the Issue and the Editorial

Available online 21 December 2017

https://doi.org/10.1016/j.coisb.2017.12.004

2452-3100/© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords
Ecological interactions, Keystone gene, Function and species, Metabolic network and model, Microbial systems ecology, Niche breadth.

Microbial systems ecology

Microbial communities (microbiomes) are involved in all biogeochemical cycles by contributing functions which may be common to most ecosystems (underlined words are defined in Box 1), e.g. nitrogen fixation, or by being first-line to very specific ecosystem services, e.g. the degradation of particular xenobiotics. Although the global relevance of microbial activities for ecosystem functioning is now widely accepted, methods to study the ecology of the tremendous richness of the microbial realm are relatively recent. In order to model, predict and understand the behaviour of microbial constituents in their native environments, Microbial Systems Ecology heavily relies on high-throughput, high-fidelity and high-resolution measurements of microbial consortia (Figure 1A) as well as the integration of the resulting data [1]. Thereby, Microbial Systems Ecology relies on specialised wet- and dry-lab approaches to achieve coherent assessments of microbial community structure and function *in situ* [1–5]. In addition to the valuable insights on community structure and functional potential (metagenomics), expressed functions (metatranscriptomics and metaproteomics) and metabolic activity (metabolomics), the integration of the individual omic levels (Figure 1B) allows comprehensive resolution of the emergent properties of ecosystems [1,6]. Furthermore, integrative approaches can significantly reduce the current limitations associated with single omics by enhancing the interpretability of data [1], allowing for example to obtain improved genome reconstructions from constituent populations [7] and to link the expression of phenotype-associated microbial functions to distinct taxa [8].

Natural microbial communities are comprised of constituent, interacting populations. Therefore, to move from descriptive, comparative or statistical studies to ecological inferences [9], in Microbial Systems Ecology, microbial communities must be seen as networks of networks: community members (populations), consisting of collections of interwoven molecular networks, form the interacting units of higher-order ecological systems. Although different types of molecular networks exist (e.g. gene regulatory networks, co-occurrence networks, etc.), we particularly focus our review on metabolic network reconstruction and related modelling approaches as applied to microbial communities in view of resolving specific properties underpinning ecosystem functioning. We also present our opinion on how harnessing this ecological knowledge will facilitate

Box 1. Glossary

- Ecosystem: ecological self-supporting unit constituted of an environment (the biotope) and the living organisms inhabiting it (the biocoenosis). Despite flows of materials, organisms and energy occurring across the boundary of individual units, the two components of an ecosystem interact more strongly between each other than with the neighbouring units.
- Ecosystem functioning: all activities, processes and properties driving biogeochemical activities and leading to the relative ecological stability of an ecosystem.
- Ecological niche (Hutchinson): the hypervolume comprised of n dimensions representing the environmental conditions and resources gradients enabling a species to persist. This definition led to the subsequent description of the fundamental niche (the maximal usable space) and the realised niche (the actual used space).
- Ecological interactions (or biological interactions or symbiosis): long-term relationship between individuals of different species including mutualism (win–win), commensalism (win–neutral), parasitism/predation (win–lose) and amensalism (lose–neutral). Metabolic interactions represent a subset of these relationships when the interaction is mediated through one or multiple metabolite(s), as opposed to non-metabolic relationships.
- Metabolic models: in silico description of the metabolic potential of a biological unit (e.g. community, guild, species), often represented as a bipartite directed network consisting of metabolites and reactions/enzymes/genes [12]. While topological metabolic models represent a qualitative view of metabolism, stoichiometric metabolic models require the specification of each reaction's stoichiometry in a stoichiometric matrix, which forms the basis for quantitative metabolic modelling.
- Microbial Systems Ecology: the holistic study of microbial communities using systems biology approaches.
- Systematic measurement: "the standardised, reproducible, and simultaneous measurement of multiple features from a single sample. Resulting datasets are fully integrable and relate system-wide behaviours" [1].

targeted manipulations of microbial communities in the future. More specifically, space- and time-resolved integrated multi-omic datasets will allow us to define and subsequently alter the realised niches of constituent populations for the management of community–conferred traits.

Using metabolic networks to obtain meaningful ecological insights

Reconstruction, analysis and modelling of metabolic networks

Community-level metabolic modelling approaches are classified according to the unit being modelled (entire community, guilds, species or strains, see Figure 1B and C) [10] and the level of detail. Metabolic modelling approaches may be divided into i) stoichiometric approaches that model the metabolism quantitatively [11], and ii) topological (network-based) approaches, which are more suitable for qualitative metabolic modelling [12].

In any case, a prerequisite to metabolic modelling is metabolic network reconstruction, i.e. the assembly of a metabolic map for the unit of interest. A number of automatic pipelines generate metabolic reconstructions directly from the genome [13–15] or metagenome [16], which can subsequently serve as the starting point for manual curation [17]. Alternatively, a selected subset of pathways relevant in a particular environment can be targeted for metabolic reconstruction [18]. Two major challenges for metabolic reconstruction are i) the large number of genes without functional annotation, which can be partially overcome using gap filling methods [19], and ii) the association of genes to reactions. Semi-curated metabolic models are collected in repositories such as AGORA [20].

Once a metabolic network reconstruction has been obtained, the community's metabolism can be analysed qualitatively or quantitatively. For instance, a topological analysis can serve to identify specific metabolic pathways of interest or to extract the active part of a community's metabolism from metatranscriptomic [21], metaproteomic or (meta-)metabolomic data (Figure 1B). A widespread quantitative metabolic modelling approach is flux balance analysis (FBA), which calculates the metabolite flow through reactions such that a particular objective function, e.g. biomass production, is maximised [11]. While topological metabolic models can integrate omics data via node or edge weights, stoichiometric models can take them into account for instance by modifying flux distributions [22]. FBA, which was originally developed for single species, was recently extended to multiple species [23,24]. However, these approaches only provide a static picture of the community. Dynamic community-level metabolic modelling, which describes the change of species abundances and metabolite concentrations over time, currently is an active field of development [25,26].

In the following paragraphs, we will discuss some applications of metabolic modelling in more detail, namely the prediction of ecological interactions, identification of keystone species and functions as well as metabolic niche inference.

Metabolic interactions

Metabolic models can be exploited to predict ecological interactions between species via metabolic cross-feeding, for instance in the case of mutualistic growth on the toxic end-products of other species, or when two species compete for the same nutrients (Figure 1C and D). Importantly, the extracellular environment, which can be characterised by metabolomics and physicochemical measurements, needs to be taken into account when predicting interactions, since not all potential interactions will be actually realised particularly in nutrient-rich environments [27]. A number of stoichiometric interaction prediction approaches compare growth rates computed in the presence or the absence of an interaction partner [28–30] or under different environmental condition [31] to determine the interaction type. Here, COMETS [26] also takes into account the impact of spatial structure on cross-feeding.

In contrast to analyses based on stoichiometric modelling, topology-based interaction prediction [32–34] first involves the inference of seed metabolites for a given microbial population, which include all metabolites that cannot be produced by the network itself [35]. It then assesses whether some of these seeds can be produced by the metabolic network of another species, which in turn allows quantification of the potential for commensalism or mutualism. The metabolic interaction potential measures the maximum number of essential nutrients that an organism can obtain by interacting with its community [34]. Furthermore, the competitive potential between two species can be determined by computing the overlap between their seed metabolites [36].

An alternative topological approach finds genome segments that maximise the number of consecutive enzyme-coding genes. The enzymes in turn catalyse metabolic transformations which are complementary across species [37]. Metabolic pathway complementarity or overlap can also be exploited to screen metagenomic data for interactions. This form of topological analysis has for instance been applied to explore metabolic strategies in human gut microbiota [38].

Recent work has involved the use of multi-omics to refine or validate model predictions in different environmental conditions [39–42]. Beyond interactions mediated through exchange or competition for metabolites, trophic interactions such as phage predation can also be inferred using omic data (see Box 2 for an example of non-metabolic interactions). Similarly, additional ecological insights such as keystone roles of some species can be inferred when metabolic networks are combined with other layers of knowledge such as co-occurrence of genes/transcripts/proteins/metabolites or to regression- and rule-based network analysis [43].

Keystone functions, genes and species

Ecological keystone species are commonly understood as species that have a pronounced impact on their environment independent of their abundance, i.e. they have a disproportionate deleterious effect on the community upon their removal [44,45]. This concept reflects the dependencies within a community governed by interactions among its members and is clearly context-dependent: the importance of any organism for stabilising the community is conferred by the particular group. Thus being a keystone species is not a Boolean trait, but it is rather a continuous property that emerges in the context of community function and different selection pressures. In order to predict which organism is a functional keystone species, the topological properties of networks derived from metabolic models that represent the community-wide organisation of microbial interactions may be used (Figure 1E) in synergy with co-occurrence networks [46,47]. Measures such as degree, clustering coefficient and closeness centrality reflect the scale of the embeddedness of the constituting organisms (nodes) in the microbial community ranging from direct ecological partners to local and global neighbourhoods, respectively [46].

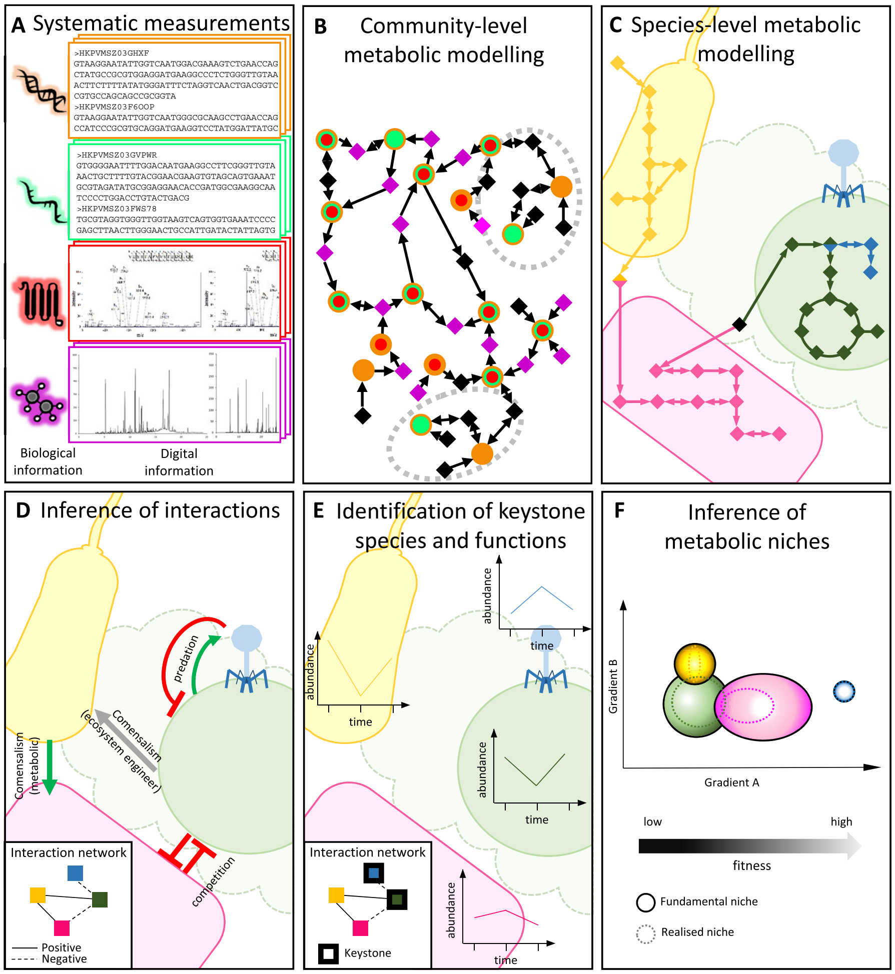
Different categories of keystone species have been proposed including ecosystem engineer (or modifier) keystone species (Figure 1E), trophic (prey or predator) keystone species or resource provider keystone species [48]. In any case, keystone species confer keystone functionalities to the ecosystem [49]. For example, the degradation of dietary fibres in the human gut is the result of a community-driven effort. However, the pivotal step is the breakdown of the complex resistant starches like amylopectin and amylose by primary degraders, which release simple sugar molecules to be fermented by the rest of the microbial consortium. *Ruminococcus bromii* is a keystone species in this context [50]. The organism possesses a highly specific cluster of keystone genes essential for efficient amylolysis [51].

Keystone metabolic genes are predicted to be highly expressed despite typically low gene copy numbers (reflecting the typical relatively low abundance of keystone species) and to catalyse key biochemical transformations (enzymes represent “load points” in the community-wide metabolic networks [52]). Therefore, a framework has been developed for the identification of such genes in reconstructed community-wide metabolic networks [49]. High relative gene expression (extracted from metatranscriptomic and/or metaproteomic data relative to gene abundance information derived from the corresponding metagenomic data) as well as specific network topological features (low relative degree and high betweenness centrality) are taken into account for the identification of such keystone genes which, through genomic linkage to reconstructed population-level genomes, can be linked to specific constituent populations which represent keystone species [49]. This approach has highlighted ammonia monooxygenase as a keystone gene in a biological wastewater treatment plant which is contributed to the community function by a specific keystone strain of *Nitrosomonas* spp [49]. Community-wide reconstructed metabolic networks are thereby particularly informative for the identification of keystone traits conferred by specific keystone species.

Microbial niche ecology

Even though it has been shown that clusters in a co-occurrence network based on 16S rRNA sequencing data reflect overlapping ecological niche preferences and common habitats of populations [53], the inference of niches of distinct bacterial populations in microbial

Figure 1


A Systematic measurements
>HKPVMSZ03GHXF
GTAAGGAATATTGGTCAATGGACGAAAGTCTGAACCAG
CTATGCCGCGTGGAGGATGAAGGCCCTCTGGGTTGTAA
ACTTCTTTTATATGGATTTCTAGGTCAACTGACGGTC
CGTGCCAGCAGCCGCGGTA
>HKPVMSZ03F6OOP
GTAAGGAATATTGGTCAATGGGCGCAAGCCTGAACCAG
CCATCCCGCGTGCAGGATGAAGGTCCTATGGATTATGC
>HKPVMSZ03GVPWR
GTGGGGAATTTTGGACAATGAAGGCCTTCGGGTTGTA
AACTGCTTTTGTACGGAACGAAGTGTAGCAGTGAAAT
GCGTAGATATGCGGAGGAACACCGATGGCGAAGGCAA
TCCCCTGGACCTGTACTGACG
>HKPVMSZ03FWS78
TGCGTAGGTGGGTTGGTAAGTCAGTGGTGAAATCCCC
GAGCTTAACTTGGGAACTGCCATTGATACTATTAGTG
Biological information
Digital information
B Community-level metabolic modelling
C Species-level metabolic modelling
D Inference of interactions
predation
Comensalism (ecosystem engineer)
Comensalism (metabolic)
competition
Interaction network
Positive
Negative
E Identification of keystone species and functions
abundance
time
abundance
time
abundance
time
abundance
time
Interaction network
Keystone
F Inference of metabolic niches
Gradient B
Gradient A
low
high
fitness
Fundamental niche
Realised niche

From metabolic models to ecological insights. (A) Following carefully adapted wet-lab procedures and systematic measurements of the purified biomolecules, (B) metabolic modelling (here resolved to the community level) by stepwise integration and modelling of the metagenomic (blue), metatranscriptomic (green), metaproteomic (red) and (meta-)metabolomic (pink) data, allows to detect, for example, parts of the metabolic network that are inactive (dotted line circle) at the sample collection. (C) Metabolic modelling (here resolved to the species level), often represented as a directed network consisting of metabolites (nodes) and reactions (edges), can be a starting point to determine (D) an ecological interaction network (nodes = species; edges = interactions). Although some non-metabolic interactions, such as commensalism by niche engineering (e.g. the green organism is a biofilm founder, allowing a secondary colonisation by the yellow microbe) or predation (see Box 2) cannot be predicted from inferred metabolic networks, other complimentary analyses, such as co-occurrence networks, will allow to predict such behaviour. (E) Topological analysis of metabolic, interaction and co-occurrence networks allow the detection of metabolic keystone species (highlighted in green; bacterial species) and trophic keystone species (highlighted in blue; phage). (F) The use of population-resolved metagenomic data to describe the fundamental niche is extended by the use of functional omic data to characterise the realised niche of different species. From this information, predictions can be made for example in relation to the fitness gradients of constituent populations.

Box 2. Causal inference of non-metabolic interactions (i.e. phage-host interactions from metagenomic data)

Phages are the most abundant and diverse entities in any environment, greatly influencing microbial community structure and dynamics through affecting the prokaryotic (host) metabolism [68,69], modulating nutrient cycles, and driving long-term host evolution [70]. The unculturability of the vast majority of host and phage strains can be circumvented by integrating meta-omic data [71,72]. Accordingly, computational methods have been developed to identify phages [73,74] and predict links to their putative hosts [75].
In addition, time-resolved datasets enable the inference of phage-host dynamics [76,77], which will result in improved knowledge and, thereby, the formulation of potential phage treatment strategies for biomedical and biotechnological applications [78].

communities remains a challenging task, due to the inherent complexity of trophic interactions and fluctuating environmental conditions. In that sense, integrated multi-omic approaches have been shown to be useful for studying microbial niche ecology. State-of-the-art binning approaches [54], or ensemble methods [55], allow near complete reconstruction of population-level genomes from assembled sequencing reads. By applying the traditional concepts of niche ecology by Hutchinson, the genomic functional potential of a microbial population reflects its fundamental niche [56,57]. Conversely, metatranscriptomic or metaproteomic data can be used to infer a population's realised niche at the time of sampling [57], while intra- and extracellular metabolomic data allows inferences regarding resource usage and the overall resource space available, respectively [57] (Figure 1F). Previous studies have relied on gene expression patterns to assess lifestyle strategies (generalists versus specialists) and the metabolic niche breadth of distinct populations [57,58]. Computational approaches that automatically predict phenotypic traits of reconstructed genomes [59] are an important resource for the in-depth characterisation of niche occupation. In this context, metabolic models can provide a detailed picture on growth conditions, such as available carbon or nitrogen sources and models have indeed been used to predict medium requirements reflecting niche breadths [60].

Apart from resource availability and usage, niche breadth also reflects tolerance ranges to physico-chemical variables, such as pH, temperature or dissolved oxygen, which are generally available only for cultured isolates. Currently, a popular approach involves the linking of inferred organismal abundances to environmental conditions, which can be challenging due to the compositional nature of rRNA amplicon sequencing data. Leveraging integrated multi-omic data and metabolic models may in turn provide a detailed mechanistic understanding of the adaptation to environmental factors for single organismal groups, as demonstrated for pH-dependent metabolic adaptations of *Enterococcus faecalis* [61].

Harnessing the power of data integration in Microbial Systems Ecology

The integration, contextualisation and analysis of multi-omic data using metabolic network approaches (in synergy with other network approaches) offer many exciting opportunities in the context of Microbial Systems Ecology, a few of which are highlighted above. While such tools are commonly used in systems biology [62], their utilisation in (microbial) ecology is still limited.

In order to move beyond associations and hypotheses derived from integrated multi-omic data, model predictions will have to be tested using combinations of detailed field and/or laboratory experiments [1,5,63], as described for example in Ref. [64]. A discovery-driven planning approach, wherein systematic measurements, data integration, model generation, hypothesis testing and new ecological hypotheses follow each other iteratively, should culminate in predictive models [1]. Thus, system-wide data has to be collected in a manner consistent with the subsequent integration and modelling to continuously improve the community models; ultimately we aim for models which allow the systematic and knowledge-guided control of different microbial community functions and/or structures. In this context, keystones functions, genes and species represent primary targets for community management, because of their disproportionate effect on ecosystem functioning. For example, lipid accumulating organisms present in wastewater treatment plants are an abundant source of lipids which may be directly converted into biodiesel [65], but as the community phenotype shows seasonal fluctuations, economical interest remains limited. Biostimulation of endogenous keystone specie(s) or targeted activation of keystone gene(s) would help tune the community towards the desired phenotype robustly around the year [63]. Conversely, a targeted removal of keystone functions may provoke a collapse of the community. In this context, the keystone concept was successfully used for the prediction of drug targets that control the pathological lung microbiome of persons with cystic fibrosis [66].

In the future, by determining the respective ecological niches of the constituent populations, we will be able to move beyond 'basic' ecological classifications of lifestyle strategy for microbes such as generalists and specialists towards more specific classifications such as the Universal Adaptive Strategy Theory (UAST) describing trade-offs between ruderal, stress tolerant and competitor behaviours [67]. This will further enable us to determine the metabolic basis of colonisation/immigration, successional stages and the community response to perturbations. In order to establish such concepts, the field needs to move towards the integration of time- and space-resolved multi-omic data to unravel the functional dynamics of complex microbial communities. In our opinion, the elucidation of networks requires such longitudinal data and corresponding time-series analyses to model the populations' interplay as well as to highlight which parts of these networks are active under specific conditions. Hence, future augmented community-level metabolic models need to account for trophic interactions and changing environmental conditions, ideally by integrating dynamic community models with genome-scale metabolic models. Therefore, within the framework of Microbial Systems Ecology, we will in the future be able to systematically define and alter the realised niches of constituent populations *in situ* and manage community—conferred traits, leading to exciting prospects for biotechnology and biomedicine.

Acknowledgements

This work was supported by CORE grants (C15/SR/10404839 to EELM and PW, and C15/BM/10404093, C16/BM/11333923 and C17/SR/11689322 to PW), a PRIDE doctoral training unit grant to PW (PRIDE15/10907093) and a European Union ERASysAPP grant (INTER/SYSAPP/14/05 to PW), all funded by the Luxembourg National Research Fund (FNR). This work was also carried out under the framework of the IdEx Unistra (MAD project granted to EELM), and benefits from a funding from the state managed by the French Research Agency as part of the Investments for the Future program. KF is supported by KU Leuven. SW was partly supported by the Konrad Lorenz Institute for Evolution and Cognition Research, Klosterneuburg, Austria, and by the Austrian Science Fund (Elise Richter V585-B31).

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Muller EEL, Glaab E, May P, Vlassis N, Wilmes P: **Condensing the omics fog of microbial communities**. *Trends Microbiol* 2013, **21**:325–333.

2. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Trouble R, *et al.*: **Open science resources for the discovery and analysis of Tara Oceans data**. *Sci Data* 2015, **2**:150023.

3. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P: **A biomolecular isolation framework for eco-systems biology**. *ISME J* 2013, **7**:110–121.

4. • Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C: **Sequencing and beyond: integrating molecular 'omics' for microbial community profiling**. *Nat Rev Microbiol* 2015, **13**:360–372.

An overview of current multi-omic approaches in microbial ecology.

5. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, Cordero OX, Brown SP, Momeni B, Shou W, *et al.*: **Challenges in microbial ecology: building predictive understanding of community function and dynamics**. *ISME J* 2016, **10**:2557.

6. Abram F: **Systems-based approaches to unravel multi-species microbial community functioning**. *Comput Struct Biotechnol J* 2015, **13**:24–32.

7. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, Laczny CC, Pinel N, May P, Wilmes P: **IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses**. *Genome Biol* 2016, **17**:260.

8. •• Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, *et al.*: **Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes**. *Nat Microbiol* 2016, **2**:16180.

This article comprehensively describes for the first time disruption of ecosystem services in human disease as evidenced across the different omic levels.

9. Greenblum S, Chiu H-C, Levy R, Carr R, Borenstein E: **Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities**. *Curr Opin Biotechnol* 2013, **24**:810–820.

10. Biggs MB, Medlock GL, Kolling GL, Papin JA: **Metabolic network modeling of microbial communities**. *WIREs Syst Biol Med* 2015, **7**:317–334.

11. Bordbar A, Monk JM, King ZA, Palsson BO: **Constraint-based models predict metabolic and associated cellular functions**. *Nat Rev Genet* 2014, **15**:107–120.

12. Faust K, Croes D, van Helden J: **Prediction of metabolic pathways from genome-scale metabolic networks**. *Biosystems* 2011, **105**:109–121.

13. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J: **The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum***. *PLoS Comput Biol* 2013, **9**:e1002980.

14. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best A, Henry C: **Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED**. *Methods Mol Biol* 2013, **985**:17–45.

15. Dias O, Rocha M, Ferreira EC, Rocha I: **Reconstructing genome-scale metabolic models with merlin**. *Nucleic Acids Res* 2015, **43**:3899–3910.

16. Konwar KM, Hanson NW, Bhatia MP, Kim D, Wu S-J, Hahn AS, Morgan-Lang C, Cheung HK, Hallam SJ: **MetaPathways v2.5: quantitative functional, taxonomic and usability improvements**. *Bioinformatics* 2015, **31**:3345–3347.

17. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93–121.

18. Darzi Y, Falony G, Vieira-Silva S, Raes J: **Towards biome-specific analysis of meta-omics data**. *ISME J* 2016, **10**:1025–1028.

19. Thiele I, Vlassis N, Fleming RMT: **fastGapFill: efficient gap filling in metabolic networks**. *Bioinformatics* 2014, **30**:2529–2531.

20. Magnusdottir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jager C, Baginska J, Wilmes P, *et al.*: **Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota**. *Nat Biotechnol* 2017, **35**:81–89.

21. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, Malmstrom RR, Forest KT, McMahon KD: **Metabolic network analysis and metatranscriptomics reveal auxotrophies and nutrient sources of the cosmopolitan freshwater microbial lineage acI**. *mSystems* 2017, **2**.

22. Kim MK, Lun DS: **Methods for integration of transcriptomic data in genome-scale metabolic models**. *Comput Struct Biotechnol J* 2014, **11**:59–65.

23. Zomorrodi AR, Maranas CD: **OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities**. *PLoS Comput Biol* 2012, **8**:e1002363.

24. Khandelwal RA, Olivier BG, Roeling WFM, Teusink B, Bruggeman FJ: **Community flux balance analysis for microbial consortia at balanced growth**. *PLos One* 2013, **8**:e64567.

25. Zomorrodi AR, Islam MM, Maranas CD: **d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities**. *ACS Synth Biol* 2014, **3**:247–257.

26. • Harcombe RW, Riehl JW, Dukovski I, Granger RB, Betts A, Lang HA, Bonilla G, Kar A, Leiby N, Mehta P, *et al.*: **Metabolic**

resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep* 2014, **7**:1104–1115.

Harcombe and coauthors present a dynamic, multi-species metabolic modelling framework that takes spatial structure into account (COMETS – Computation Of Microbial Ecosystems in Time and Space).

27. Klitgord N, Segrè D: **Environments that induce synthetic microbial ecosystems**. *PLoS Comput Biol* 2010, **6**:e1001002.

28. Chiu H-C, Levy R, Borenstein E: **Emergent biosynthetic capacity in simple microbial communities**. *PLoS Comput Biol* 2014, **10**:e1003695.

29. Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R, Ruppin E: **Competitive and cooperative metabolic interactions in bacterial communities**. *Nat Commun* 2011, **2**:589.

30. Mendes-Soares H, Mundy M, Soares LM, Chia N: **MMinte: an application for predicting metabolic interactions among the microbial species in a community**. *BMC Bioinf* 2016, **17**:343.

31. Heinken A, Thiele I: **Anoxic conditions promote species-specific mutualism between gut microbes in silico**. *Appl Environ Microbiol* 2015, **81**(12):4049–4061.

32. Borenstein E, Feldman MW: **Topological signatures of species interactions in metabolic networks**. *J Comput Biol* 2009, **16**: 191–200.

33. Levy R, Borenstein E: **Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules**. *Proc Natl Acad Sci U S A* 2013, **10**: 12804–12809.

34. • Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR: **Metabolic dependencies drive species co-occurrence in diverse microbial communities**. *Proc Natl Acad Sci U S A* 2015, **112**:6449–6454.

By analysing the composition of more than 800 communities, this article presents a mechanistic understanding for observed co-occurrence patterns among distinct populations through metabolic dependencies.

35. Carr R, Borenstein E: **NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment**. *Bioinformatics* 2012, **28**: 734–735.

36. Kreimer A, Doron-Faigenboim A, Borenstein E, Freilich S: **NetCmpt: a network-based tool for calculating the metabolic competition between bacterial species**. *Bioinformatics* 2012, **28**:2195–2197.

37. Bordron P, Latorre M, Cortés M-P, González M, Thiele S, Siegel A, Maass A, Eveillard D: **Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach**. *Mirobiol Open* 2016, **5**:106–117.

38. • Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Yunta RG, Okuda S, Vandeputte D, Valles-Colomer M, Hildebrand F, Chaffron S, *et al.*: **Species–function relationships shape ecological properties of the human gut microbiome**. *Nat Microbiol* 2016, **1**:16088.

Linkage of specific functional traits to constituent populations which are drivers of ecological networks.

39. Shoaie S, Karlsson F, Mardinoglu A, Nookaew I, Bordel S, Nielsen J: **Understanding the interactions between bacteria in the human gut through metabolic modeling**. *Sci Rep* 2013, **3**: 2532.

40. Lawson CE, Wu S, Bhattacharjee AS, Hamilton JJ, McMahon KD, Goel R, Noguera DR: **Metabolic network analysis reveals microbial community interactions in anammox granules**. *Nat Commun* 2017, **8**:15416.

41. Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, Thomas BC, Handley KM, Mullin SW, Nicora CD, Singh A, *et al.*: **Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer**. *ISME J* 2014, **8**:1452–1463.

42. Si Ishii, Suzuki S, Tenney A, Norden-Krichmar TM, Nealson KH, Bretschger O: **Microbial metabolic networks in a complex electrogenic biofilm recovered from a stimulus-induced metatranscriptomics approach**. *Sci Rep* 2015, **5**:14840.

43. Faust K, Raes J: **Microbial interactions: from networks to models**. *Nat Rev Microbiol* 2012, **10**:538–550.

44. Power ME, Tilman D, Estes JA, Menge BA, Bond WJ, Mills LS, Daily G, Castilla JC, Lubchenco J, Paine RT: **Challenges in the quest for keystones**. *Bioscience* 1996, **46**:609–620.

45. Paine RT: **A conversation on refining the concept of keystone species**. *Conserv Biol* 1995, **9**:962–964.

46. • Berry D, Widder S: **Deciphering microbial interactions and detecting keystone species with co-occurrence networks**. *Front Microbiol* 2014, **5**:219.

In this article, current approaches for identifying keystone species in microbial communities are comprehensively overview and tested on simulated communities with known interactions. It is demonstrated that the interpretability of co-occurrence networks can be lost when the effects of habitat filtering become significant.

47. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**: 751–753.

48. Mills LS, Soulé ME, Doak DF: **The keystone-species concept in ecology and conservation management and policy must explicitly consider the complexity of interactions in natural systems**. *Bioscience* 1993, **43**:219–224.

49. •• Roume H, Heintz-Buschart A, Muller EEL, May P, Satagopam VP, Laczny CC, Narayanasamy S, Lebrun LA, Hoopmann MR, Schupp JM, *et al.*: **Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks**. *NPJ Biofilms Microbiomes* 2015, **1**:15007.

This work takes a function-centric approach based on community-wide metabolic network reconstructions to identify "keystone genes" and links these to constituent keystone species.

50. •• Ze X, Duncan SH, Louis P, Flint HJ: ***Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon**. *ISME J* 2012, **6**:1535–1543.

By describing the first amylosome, a cell surface enzyme complex devoted to starch degradation, Ze and co-authors identified key functional attributes of keystone species within the human gut microbiome.

51. Ze X, Ben David Y, Laverde-Gomez JA, Dassa B, Sheridan PO, Duncan SH, Louis P, Henrissat B, Juge N, Koropatkin NM, *et al.*: **Unique organization of extracellular amylases into amylosomes in the resistant starch-utilizing human colonic Firmicutes bacterium *Ruminococcus bromii***. *mBio* 2015, **6**. e01058–01015.

52. Rahman SA, Schomburg D: **Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks**. *Bioinformatics* 2006, **22**:1767–1774.

53. Chaffron S, Rehrauer H, Pernthaler J, von Mering C: **A global network of coexisting microbes from environmental and whole-genome sequence data**. *Genome Res* 2010, **20**: 947–959.

54. Sedlar K, Kupkova K, Provaznik I: **Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics**. *Comput Struct Biotechnol J* 2017, **15**:48–55.

55. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF: **Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy**. *bioRxiv* 2017, https://doi.org/10.1101/107789.

56. Mou X, Sun S, Edwards RA, Hodson RE, Moran MA: **Bacterial carbon processing by generalist species in the coastal ocean**. *Nature* 2008, **451**:708–711.

57. •• Muller EEL, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, *et al.*: **Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage**. *Nat Commun* 2014, **5**:5603.

The first study to integrate metagenomic, metatranscriptomic, metaproteomic and (meta-)metabolomic data. The integrated data analysis provides unique insights into the lifestyle strategies of constituent populations and links the levels of genetic variation within populations to their ecological niche breadth.

58. Gifford SM, Sharma S, Booth M, Moran MA: **Expression patterns reveal niche diversification in a marine microbial assemblage**. *ISME J* 2013, **7**:281–298.

59. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC: **From genomes to phenotypes: traitar, the microbial trait analyzer**. *mSystems* 2016, **1**.

60. Zarecki R, Oberhardt MA, Reshef L, Gophna U, Ruppin E: **A novel nutritional predictor links microbial fastidiousness with lowered ubiquity, growth rate, and cooperativeness**. *PLoS Comput Biol* 2014, **10**:e1003726.

61. Grosseholz R, Koh CC, Veith N, Fiedler T, Strauss M, Olivier B, Collins BC, Schubert OT, Bergmann F, Kreikemeyer B, *et al.*: **Integrating highly quantitative proteomics and genome-scale metabolic modeling to study pH adaptation in the human pathogen *Enterococcus faecalis***. *NPJ Syst Biol Appl* 2016, **2**: 16017.

62. Gomez-Cabrero D, Tegnér J: **Iterative systems biology for medicine – time for advancing from network signatures to mechanistic equations**. *Curr Opin Syst Biol* 2017, **3**: 111–118.

63. Narayanasamy S, Muller EEL, Sheik AR, Wilmes P: **Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities**. *Microb Biotechnol* 2015, **8**:363–368.

64. Sheik AR, Muller EE, Audinot JN, Lebrun LA, Grysan P, Guignard C, Wilmes P: ***In situ* phenotypic heterogeneity among single cells of the filamentous bacterium *Candidatus* Microthrix parvicella**. *ISME J* 2016, **10**:1274–1279.

65. Sheik AR, Muller EELL, Wilmes P: **A hundred years of activated sludge: time for a rethink**. *Front Microbiol* 2014, **5**:47.

66. Quinn RA, Whiteson K, Lim YW, Zhao J, Conrad D, LiPuma JJ, Rohwer F, Widder S: **Ecological networking of cystic fibrosis lung infections**. *NPJ Biofilms Microbiomes* 2016, **2**:4.

67. Grime JP, Pierce S: **Primary adaptive strategies in organisms other than plants**. In *The evolutionary strategies that shape ecosystems*. John Wiley & Sons, Ltd; 2012:40–104.

68. • De Smet J, Zimmermann M, Kogadeeva M, Ceyssens PJ, Vermaelen W, Blasdel B, Bin Jang H, Sauer U, Lavigne R: **High coverage metabolomics analysis reveals phage-specific alterations to *Pseudomonas aeruginosa* physiology during infection**. *ISME J* 2016, **10**:1823–1835.

In this work, the dynamics of phenotypic features of a bacterial host is tracked following viral infections using untargeted high coverage metabolomics. A clear phage-specific and infection stage-specific metabolic reshuffling was observed, highlighting the importance of the overlooked viral realm in Microbial System Ecology.

69. Zhao X, Shen M, Jiang X, Shen W, Zhong Q, Yang Y, Tan Y, Agnello M, He X, Hu F, *et al.*: **Transcriptomic and metabolomics profiling of phage-host interactions between phage PaP1 and *Pseudomonas aeruginosa***. *Front Microbiol* 2017, **8**:548.

70. Burmeister AR, Lenski RE: **Host coevolution alters the adaptive landscape of a virus**. *Proc Biol Sci* 2016, **283**.

71. Hayes S, Mahony J, Nauta A, van Sinderen D: **Metagenomic approaches to assess bacteriophages in various environmental niches**. *Viruses* 2017, **9**.

72. Motlagh AM, Bhattacharjee AS, Coutinho FH, Dutilh BE, Casjens SR, Goel RK: **Insights of phage-host interaction in hypersaline ecosystem through metagenomics analyses**. *Front Microbiol* 2017, **8**:352.

73. Roux S, Enault F, Hurwitz BL, Sullivan MB: **VirSorter: mining viral signal from microbial genomic data**. *PeerJ* 2015, **3**:e985.

74. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F: **VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data**. *Microbiome* 2017, **5**:69.

75. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE: **Computational approaches to predict bacteriophage-host relationships**. *FEMS Microbiol Rev* 2016, **40**:258–272.

76. Parsons RJ, Breitbart M, Lomas MW, Carlson CA: **Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea**. *ISME J* 2012, **6**: 273–284.

77. Zhang J, Gao Q, Zhang Q, Wang T, Yue H, Wu L, Shi J, Qin Z, Zhou J, Zuo J, *et al.*: **Bacteriophage-prokaryote dynamics and interaction within anaerobic digestion processes across time and space**. *Microbiome* 2017, **5**:57.

78. Jassim SA, Limoges RG, El-Cheikh H: **Bacteriophage biocontrol in wastewater treatment**. *World J Microbiol Biotechnol* 2016, **32**:70.

C.7 Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract.

Cedric C. Laczny, Emilie E.L. Muller, Anna Heintz-Buschart, Malte Herold, Laura A. Lebrun,
Angela Hogan, Patrick May, Carine de Beaufort, Paul Wilmes
2016
Frontiers in Microbiology **7**
DOI: 10.3389/fmicb.2016.00884

Contributions of author include:

- Data analysis and visualisation
- Revision of manuscript

frontiers in Microbiology

ORIGINAL RESEARCH
published: 21 June 2016
doi: 10.3389/fmicb.2016.00884

Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract

Cedric C. Laczny[1]†, Emilie E. L. Muller[1], Anna Heintz-Buschart[1], Malte Herold[1], Laura A. Lebrun[1], Angela Hogan[2], Patrick May[1], Carine de Beaufort[1,3] and Paul Wilmes[1]*

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg, [2] Integrated Biobank of Luxembourg, Luxembourg, Luxembourg, [3] Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg

OPEN ACCESS

Edited by:
Marc Strous,
University of Calgary, Canada

Reviewed by:
James Hemp,
California Institute of Technology, USA
Sebastian Luecker,
Radboud University Nijmegen, Netherlands

***Correspondence:**
Paul Wilmes
paul.wilmes@uni.lu

†Present address:
Cedric C. Laczny,
Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany

Specialty section:
This article was submitted to Microbial Physiology and Metabolism, a section of the journal Frontiers in Microbiology

Received: *29 January 2016*
Accepted: *25 May 2016*
Published: *21 June 2016*

Citation:
Laczny CC, Muller EEL, Heintz-Buschart A, Herold M, Lebrun LA, Hogan A, May P, de Beaufort C and Wilmes P (2016) Identification, Recovery, and Refinement of Hitherto Undescribed Population-Level Genomes from the Human Gastrointestinal Tract. Front. Microbiol. 7:884. doi: 10.3389/fmicb.2016.00884

Linking taxonomic identity and functional potential at the population-level is important for the study of mixed microbial communities and is greatly facilitated by the availability of microbial reference genomes. While the culture-independent recovery of population-level genomes from environmental samples using the binning of metagenomic data has expanded available reference genome catalogs, several microbial lineages remain underrepresented. Here, we present two reference-independent approaches for the identification, recovery, and refinement of hitherto undescribed population-level genomes. The first approach is aimed at genome recovery of varied taxa and involves multi-sample automated binning using CANOPY CLUSTERING complemented by visualization and human-augmented binning using VIZBIN *post hoc*. The second approach is particularly well-suited for the study of specific taxa and employs VIZBIN *de novo*. Using these approaches, we reconstructed a total of six population-level genomes of distinct and divergent representatives of the Alphaproteobacteria class, the Mollicutes class, the Clostridiales order, and the Melainabacteria class from human gastrointestinal tract-derived metagenomic data. Our results demonstrate that, while automated binning approaches provide great potential for large-scale studies of mixed microbial communities, these approaches should be complemented with informative visualizations because expert-driven inspection and refinements are critical for the recovery of high-quality population-level genomes.

Keywords: metagenome, binning, genome recovery, refinement, reference genomes

INTRODUCTION

Substantial efforts have recently been undertaken for the in-depth structural and functional characterization of human-derived microbiota from various body sites (Qin et al., 2010; Huttenhower et al., 2012; Methé et al., 2012). These efforts largely involve the use of microbial isolate genome sequences (Lagier et al., 2012; Rajilić-Stojanović and de Vos, 2014) or population-level genomes recovered following the “binning” of metagenomic data (Di Rienzi et al., 2013; Sharon et al., 2013; Alneberg et al., 2014; Nielsen et al., 2014). Importantly, the term “population-level genome“ indicates that the genomic complements recovered from metagenomic data will typically be derived from a mixture (population) of closely related microorganisms in which individuals are not expected to be clonal (Kunin et al., 2008), thus resulting in composite genomic reconstructions. Obtaining population-level genome resolution, e.g., at the genus- or species-levels, allows linkage of taxonomic identity and function at the level of individual populations (Dick et al., 2009; Albertsen et al., 2013; Muller et al., 2014).

Unsupervised binning approaches, as opposed to their supervised counterparts, are independent of prior information and exploit data-inherent characteristics, e.g., genomic signatures based on oligonucleotide frequencies and/or sequence abundance information (Dick et al., 2009; Albertsen et al., 2013; Alneberg et al., 2014; Laczny et al., 2014; Nielsen et al., 2014; Wu et al., 2014; Kang et al., 2015). These reference-independent binning approaches may further be subdivided into automated approaches (Nielsen et al., 2014; Wu et al., 2014), e.g., CANOPY CLUSTERING, and user-driven approaches (Dick et al., 2009; Laczny et al., 2014, 2015), e.g., VIZBIN. For the former, minimal, if any, input by a human user is needed, whereas for the latter a low-dimensional representation, typically in two dimensions, allows for human input in the cluster definition process. The automated CANOPY CLUSTERING employs sequence abundance covariation across a set of multiple samples and has been used for the large-scale recovery of population-level genomes from the human gastrointestinal tract (GIT; Nielsen et al., 2014). In the user-driven VIZBIN, bins are represented as two-dimensional sequence cluster structures and resolved using human pattern recognition capabilities (Laczny et al., 2014, 2015). While the automation of the clustering process typically leads to an increased throughput in the recovery of population-level genomes from metagenomic data, suboptimal clusters might be created and should be manually refined (Laczny et al., 2015).

Despite previous efforts to further expand reference genome catalogs, individual microbial lineages might have been missed even in deeply studied environments such as the human GIT. Unsupervised binning approaches are particularly pertinent for the recovery of genomes derived from members of hitherto undescribed microbial lineages due to their independence of prior information, especially if no closely related representatives have yet been recovered. An example of such a novel lineage discovered following the binning of metagenomic data are the Melainabacteria which occur in environmental as well as human-derived samples (Di Rienzi et al., 2013; Soo et al., 2014). While this lineage was reported in earlier 16S rRNA gene-based studies of the human GIT (Ley et al., 2005), sets of complete or partial population-level genomes were only recently recovered from metagenomic data derived from human and koala fecal samples (Di Rienzi et al., 2013; Soo et al., 2014). The Melainabacteria-lineage was originally proposed as a sister-phylum to the Cyanobacteria (Di Rienzi et al., 2013) but Soo et al. (2014) subsequently suggested that it represents a non-photosynthetic class within the Cyanobacteria phylum instead.

Here, we present two approaches for the identification, recovery, and refinement of hitherto undescribed population-level genomes without the need for *a priori* knowledge in the form of reference genomes. These approaches involving automated binning using CANOPY CLUSTERING and/or visualization and human-augmented binning using VIZBIN were applied to human GIT-derived metagenomic data from a multiplex family study of type 1 diabetes mellitus (MuSt). VIZBIN was first applied *post hoc* (automatically generated clusters were inspected and manually refined using VIZBIN) and using this approach one automatically identified cluster was found to be a mixture of at least three distinct organisms. Second, given the recent discovery of the Melainabacteria class, VIZBIN was used to explore whether genomic complements of further, hitherto undescribed representatives could be recovered by the *de novo* application of VIZBIN, i.e., without prior automated clustering. Overall, a total of six almost complete or partial population-level genomes from the Alphaproteobacteria class, the Mollicutes class, the Clostridiales order, and the Melainabacteria class were recovered thereby extending existing reference genome catalogs. The reconstructed, high-quality population-level genomes will be valuable for the successful interpretation of additional multi-omic data from the human GIT. Moreover, we would expect our approach to be applicable for the reconstruction of high-quality population-level genomes from metagenomic data derived from other, less well-characterized environments.

MATERIALS AND METHODS

Sample Collection, Processing, and Metagenomic Sequencing

Study Context

The multiplex family study of type 1 diabetes mellitus is a Luxembourg-based, observational study of selected family groups of two or three generations in which there are multiple incidents of type 1 diabetes mellitus (T1DM). Fecal samples were collected at different time points from patients with T1DM and healthy family members. A total of 55 fecal samples were collected from 10 patients with T1DM and 10 healthy family members. The generated metagenomic data is used herein for the identification and recovery of hitherto undescribed microbial population-level genomes, independent of disease burden. The study was approved by the Comité d'Ethique de Recherche (CNER; Reference: 201110/05) and the National Commission for Data Protection in Luxembourg. Written informed consent was obtained from all subjects enrolled in the study.

Stool Sampling

Fecal samples were self-collected and immediately frozen on dry ice at three time points (if bowel movement permitted on day of scheduled sampling) at intervals between 4 and 8 weeks.

Extraction of DNA from Fecal Samples

DNA was extracted from frozen subsamples of 150 mg after pre-treatment of the weighed subsamples with 1.5 ml RNAlater ICE (Life Technologies) at −20°C over night. The faeces-RNAlater ICE mixtures were homogenized by bead-beating (Roume et al., 2013). Differential centrifugation and extraction using All-In-One kit (Norgen Biotek) were performed according to Roume et al. (2013). DNA fractions were further supplemented with DNA extracted from 200 mg subsamples using the MOBIO Power Soil Kit according to the manufacturer's instructions.

Library Preparation and Sequencing

Libraries with an insert size of 350 base pair (bp) were constructed from metagenomic DNA following fragmentation by sonification (Covaris), end-repair, adenylation, adapter ligation, and amplification of adapter-ligated DNA fragments using appropriate enzymes (Enzymatics). Library amplification and cluster generation were performed using TruSeq PE Cluster Kit V3–cBot–HS (Illumina). The resulting flow cells were sequenced on a HiSeq2000 system (Illumina) generating 101 bp paired-end reads. Sequencing was performed by BGI (Hong Kong, China).

Preprocessing of the Metagenomic Data

The per-sample metagenomic paired-end sequencing data in FASTQ format was processed using the MOCAT trimming and quality filtering step (MOCAT.pl -rtf) and the parameters used were as follows: readtrimfilter_length_cutoff = 40 readtrimfilter_qual_cutoff = 20 readtrimfilter_use_sanger_scale = yes readtrimfilter_trim_5_prime = yes readtrimfilter_use_precalc _5prime_trimming = no (Kultima et al., 2012). The preprocessed reads were then mapped onto the human reference genome (hg19) using the MOCAT screening step (MOCAT.pl -s hg19) and using the following parameters: screen_length_cutoff = 30 screen_percent_cutoff = 90 screen_soap_seed_length = 30 screen_soap_max_mm = 10 screen_soap_cmd = −M 4 screen_save_format = sam and SOAPaligner v2.21 (Li R. et al., 2009). The preprocessing resulted in two sets of reads in FASTQ format (human and non-human) consisting each of paired-end and single-end reads. The human reads were discarded. A schematic overview of the individual steps is provided in Supplementary Figure S1.

Assembly of the Metagenomic Data

The preprocessed, non-human, paired-end reads of each sample were assembled separately using IDBA-UD (Peng et al., 2012). More specifically, the reads were converted from FASTQ to FASTA format using the fq2fa script (fq2fa --merge --filter) provided by IDBA-UD. Subsequently, IDBA-UD was applied using its pre-error-correction step for read error correction (idba_ud --pre_correction). The resulting contigs were extended using the paired-end and single-end reads not used by IDBA-UD using the Velvet assembly tool (Zerbino and Birney, 2008). First, paired-end reads were mapped onto the previously assembled contigs using SOAP (-r 2 -M 4 -l 30 -v 10 -p 8 -u unmapped.fa) and "unmapped" reads were identified. Then, the unused single-end reads (IDBA-UD only supports paired-end reads) were combined with the unmapped reads, and cd-hit-dup from the cd-hit software suite was used to remove duplicate reads (Fu et al., 2012). The IDBA-UD-based contigs were provided as long-read input to Velvet v1.2.07 with the following parameters for velveth: -long contig.fa, for velvetg: -conserveLong yes, and over a range of *k*-mer sizes (27, 31, 35, 39, 43, 47, 51, 55, 59, 63). The resulting contigs from the assemblies using different *k*-mer sizes and the IDBA-UD-based initial set of contigs were pooled and clustered using cd-hit-est (parameter: -c 0.99) to remove redundancy. minimus2 (AMOS genome assembly software suite v3.1) was used to join and extend the clustered contigs based on the detection of sequence overlaps (Treangen et al., 2011). Gene prediction was performed on the final set of contigs using Prodigal v2.60 (parameter: -p meta; Hyatt et al., 2012). A schematic overview of the individual steps is provided in Supplementary Figure S1.

Automated Clustering Using Canopy Clustering

Genes from the individual assemblies were pooled and genes with a sequence length <100 nt were discarded. The remaining genes were made non-redundant by applying cd-hit-est (Li and Godzik, 2006; Fu et al., 2012) to collapse sequences with 95% sequence identity over 90% of the shorter sequence (-c 0.95 -aS 0.9). Bowtie2 v2.0.2 was used to map the preprocessed reads for the individual samples to the gene catalog. The resulting SAM files were sorted using SAMtools v0.1.19 and converted to BAM format. BEDTools v2.18.1 genomeCoverageBed and awk were used to compute the per-sample fold-coverage for each catalog-gene. Genes with a fold-coverage <2× in all of the 55 samples were discarded to reduce the data amount and to limit the runtime on the University of Luxembourg's High Performance Computing platform. Put differently, a gene was preserved if its fold-coverage was ≥2× in at least one of the 55 samples. Canopy Clustering was run with the following parameters: --max_canopy_dist 0.1 --max_close_dist 0.4 --max_merge_dist 0.1 --min_step_dist 0.005 --stop_fraction 1 and 30 threads (-n 30). Following the original definition, the resulting co-abundance gene groups (CAGs) are referred to as metagenomic species (MGS) if they contained at least 700 genes (Nielsen et al., 2014).

Completeness and Contamination Estimation of Population-Level Genomes

A set of 107 genes found in single copy in 95% of sequenced bacterial genomes (essential genes) was used to assess the degrees of completeness and contamination of individual population-level genomes (Dupont et al., 2012). Peptide sequences of the *in silico* predicted genes were screened against the hidden Markov models (HMMs) of the essential genes[1] (Albertsen et al., 2013) using HMMER v3.1b1 hmmsearch with parameters: --cut_tc –notextw (Eddy, 2007). High quality population-level genomes are characterized by high levels of completeness (high fraction of essential genes recovered) and low levels of contamination (low number of duplicated essential genes).

Taxonomic Characterization

Complementary approaches for the taxonomic characterization of the bins were used. These included two approaches based on protein-coding genes [PhyloPhlAn (Segata et al., 2013) and AMPHORA2 (Wu and Scott, 2012)] and a whole genome-based approach [BLAST (Altschul et al., 1990, 1997; Zhang et al.,

---

[1] https://github.com/MadsAlbertsen/multi-metagenome/blob/master/R.data.generation/essential.hmm

2000; Wheeler et al., 2007) + MEGAN (Huson et al., 2007)]. For individual genomic complements, PhyloPhlAn provides a consensus taxonomic classification, whereas AMPHORA2 returns taxonomic classifications for each gene separately out of a set of phylogenetic marker genes. While the whole genome-based approach queries a continuously updated database at the NCBI and may thus potentially be more specific, the two protein-coding gene-based approaches are more robust with respect to the taxonomic characterization of organisms without closely related representatives in a database.

PhyloPhlAn

The translated peptide sequences for each gene (see sections "Assembly of the Metagenomic Data" and "Automated Clustering Using Canopy Clustering") were prepared for PhyloPhlAn by ensuring unique peptide sequence identifiers and removing the asterisk symbol (if present) at the ends of the sequences. PhyloPhlAn's impute option (-t) was used for taxonomic assignment of individual populations and 12 threads were used (--nproc 12).

AMPHORA2

The default parameters of AMPHORA2 were used.

BLAST + MEGAN

Online BLAST searches were carried out on the "Nucleotide collection (nr/nt)" database. "Uncultured/environmental sample sequences" were excluded and "bacteria (taxid:2)" was specified as "Organism." The "Max target sequences" were set to 10. All other parameters were left at their default values. The results were downloaded and imported into MEGAN using the lowest common ancestor (LCA)-option to obtain per-sequence taxonomic classifications.

Whole Genome-Based Comparisons

The online average nucleotide identity (ANI) calculator[2] was used for determining the ANI values between genomic complements (Goris et al., 2007). Genome pairs with ANI values >95% were considered to belong to the same species (Goris et al., 2007).

Cyanobacteria-Like Sequence Groups

Identification of Marker Genes

AMPHORA2 (Wu and Scott, 2012) was used to identify and to taxonomically classify phylogenetic marker genes among the genes encoded by the *de novo* assembled contigs of the 55 MuSt metagenomic datasets (see section "Assembly of the Metagenomic Data"). All genes annotated to belong to the Cyanobacteria phylum with an associated confidence value ≥75% were retained. The DNA-directed RNA polymerase subunit beta (*rpoB*) genes were selected for further analyses, as they represented the largest set of marker genes annotated as cyanobacterial. Incomplete *rpoB* gene predictions, i.e., those lacking a start or a stop codon according to the gene prediction, were discarded. The remaining *rpoB* genes were then clustered according to sequence similarity using cd-hit-est at 95% sequence identity over 50% of the shorter sequences (-c 0.95 -aS 0.5; Li and Godzik, 2006; Fu et al., 2012). The resulting sequence clusters are referred to as "Cyanobacteria-like sequence groups (CLSGs)" and the respective representative genes as "CLSG marker genes."

Coverage Computation

The preprocessed reads of each sample were individually mapped to the CLSG genes using Bowtie2 v2.0.2. The resulting SAM files were sorted using SAMtools v0.1.19 and converted to BAM format (Li H. et al., 2009). Per-CLSG gene fold-coverages for each sample were computed using BEDTools v2.23.0 genomeCoverageBed (Quinlan and Hall, 2010) and awk.

Construction of Phylogenetic Trees Using *rpoB*

The *rpoB* gene was used for the construction of phylogenetic trees as it represents an alternative to the ribosomal rRNA genes for phylogenetic analyses (Case et al., 2007; Bondoso et al., 2013). For the MGS (see section "Automated Clustering Using Canopy Clustering"), the NCBI's MOLE-BLAST webservice[3] was used which includes a database search to retrieve the most closely related sequences, thus accounting for the varied MGS taxa, i.e., of diverse phylogenetic origin. In contrast, for the CLSG genomes (see section "Cyanobacteria-Like Sequence Groups"), melainabacterial *rpoB* genes were manually extracted from published melainabacterial genomes.

MGS

The NCBI's MOLE-BLAST webservice[3] was used to query the *rpoB* gene sequences encoded by the VizBin-refined MGS population-level genomes against the "Nucleotide collection (nr/nt)" database. "Uncultured/environmental sample sequences" were excluded and "Bacteria" was specified as an "Entrez Query." The "Number of database sequences" was set to 10. In brief, MOLE-BLAST generally works as follows. In the first step, the query sequences are grouped by locus using BLAST (Altschul et al., 1997). Second, a BLAST database search is performed to identify each query's nearest-neighbor target sequences. A multiple sequence alignment is subsequently computed for each locus, including the query sequences and their nearest neighbors, using MUSCLE (Edgar, 2004). MOLE-BLAST then computes a phylogenetic tree for each locus multiple sequence alignment using Neighbor Joining (Saitou and Nei, 1987) or Fast Minimum Evolution (Desper and Gascuel, 2004).

CLSGs and Melainabacteria

Gut-derived and environment-derived Melainabacteria genomes from Di Rienzi et al. (2013) were downloaded[4]. Gut-derived Melainabacteria genomes from Soo et al. (2014) were downloaded from JGI IMG/ER under the accession numbers 2523533517 (Zag_1), 2531839741 (Zag_111), 2523533519 (Zag_221), 2522572068 (MH_37). The *rpoB* genes were identified using prokka (Seemann, 2014) and their nucleotide sequences were extracted. All *rpoB* gene sequences in the

---

[2] http://enve-omics.ce.gatech.edu/ani/

[3] http://blast.ncbi.nlm.nih.gov/moleblast/moleblast.cgi

[4] http://ggkbase.berkeley.edu/mel/organisms

publicly available Melainabacteria genomes and the CLSG genes identified in this work were submitted to phylogeny.fr using the "One Click" mode[5] (Dereeper et al., 2008). The option "Use the Gblocks program to eliminate poorly aligned positions and divergent regions" was enabled. The alignment of the *rpoB* gene sequences was computed using MUSCLE (Edgar, 2004) and automatically curated using GBLOCKS (min. seq. for flank pos.: 85%; max. contig. Non-conserved pos.: 8; min. block length: 10; gaps in final blocks: no; Castresana, 2000). PHYML including the approximate Likelihood-Ratio Test (aLRT) was used to compute the phylogeny (model: default; statistical test: alrt; number of categories: 4, gamma: estimated; invariable sites: estimated; remove gaps: enabled; Guindon and Gascuel, 2003; Anisimova and Gascuel, 2006; Guindon et al., 2010). The tree was saved in Newick-format and rendered using EVOLVIEW (Zhang et al., 2012). Coloring of the resulting tree was performed manually in ADOBE ILLUSTRATOR.

Visualization and Binning Using VIZBIN

Contigs from the respective samples were used as input for VIZBIN (Laczny et al., 2015). The resulting two-dimensional embeddings were employed to select bins of interest. If not stated otherwise, reconstructed metagenomic sequence fragments <1,000 nt were omitted from the visualization and binning using VIZBIN (Laczny et al., 2015). Remaining sequences longer than 5,000 nt were iteratively cut into segments (chunks) of 5,000 nt as long as the remaining sequence was at least 5,000 nt long. Otherwise the entire (remaining) sequence was used. The creation of sequence chunks helps to mitigate effects of variations in assembly quality on the visualization: well-recovered genomes are assembled in longer and fewer contigs than other genomes. However, long-assembled genomes would, without the creation of chunks, be represented by only a small number of points. This step can thus be considered as a means to normalize sequence cluster size and density for improved cluster identification and delineation. The resulting sequences were used as input for VIZBIN. Additional information was added to the visualizations on a per-case basis. Generally, if coverage information was used, the opaqueness value (alpha) of each point was determined based on the natural logarithm of the corresponding fold-coverage value and provided as the "coverage" annotation option in VIZBIN. Particular sequences of interest were highlighted using either the "label" annotation type (distinct color and shape per individual label) in the case of the manual refinement of automatically generated MGS or the "isMarker" annotation type (star-shape; "beacon contig") in the case of the manually defined CLSG bins.

Re-assembly and Analysis of Recovered Population-Level Genomes

The preprocessed reads (pairs and singletons) from the sample with the highest average contig fold-coverage for each population-level genome were aligned to the contigs of the population-level genome using BOWTIE2 v.2.2.2 with default parameters. Contigs with exceptionally high or low fold-coverage, i.e., outliers, were identified and excluded. More specifically, a contig was considered an outlier if the absolute value of the modified Z-score was greater than 3.5 (Iglewicz and Hoaglin, 1993). The reads (pairs and singletons) from the preserved contigs were recruited using SAMTOFASTQ from PICARD v.1.130[6] and assembled using SPADES v.3.1.0 (Bankevich et al., 2012) using the "careful" option.

RAST-Based Annotation and Accession

Functional annotation of the re-assembled genomes was performed using the RAST webservice[7] (Aziz et al., 2008; Overbeek et al., 2014). The respective annotation results, including the original genomes, are accessible under the following accession IDs via a guest account: MGS00153 – 6666666.163363, MGS00248 – 6666666.163364, MGS00113-CG02 – 6666666.163361, CLSG01 – 6666666.163354, CLSG02 – 6666666.163355, CLSG03 – 6666666.155161. The genome analyses were performed by using automatically computed 'Scenarios' as well as by user-driven search of specific genes in the 'Genome Browser' of the RAST webservice. Gaps in nearly complete pathways or complexes were filled manually using a BLAST search of the missing genes.

Data Accession

The raw non-human metagenomic reads are deposited at the NCBI under the BioProject accession number PRJNA289586.

RESULTS AND DISCUSSION

Here, we present the results of the application of our two approaches, involving CANOPY CLUSTERING and/or VIZBIN, for the identification, recovery, and refinement of population-level genomes from human GIT-derived metagenomic data of the MuSt project (**Figure 1**). The metagenomic data was preprocessed and assembled using a custom pipeline (Supplementary Figure S1). On average, 20,862,561 paired-end reads ± 608,594 (mean ± SD) remained after preprocessing per sample thus resulting in highly similar read library sizes across all samples. On average, 877,215 contigs were assembled (Supplementary Table S1).

Inspection and Refinement of Automatically Generated Bins

CANOPY CLUSTERING requires the construction of a non-redundant gene catalog and exploits the cross-sample correlation of fold-coverages of genes in the catalog for bin (cluster) definition. Here, CANOPY CLUSTERING was first applied to the MuSt metagenomic dataset consisting of 55 samples from 20 individuals in total. A subset of the resulting clusters was subsequently inspected and refined using VIZBIN (**Figure 1**).

---

[5] http://www.phylogeny.fr/simple_phylogeny.cgi

[6] http://broadinstitute.github.io/picard

[7] http://rast.nmpdr.org/

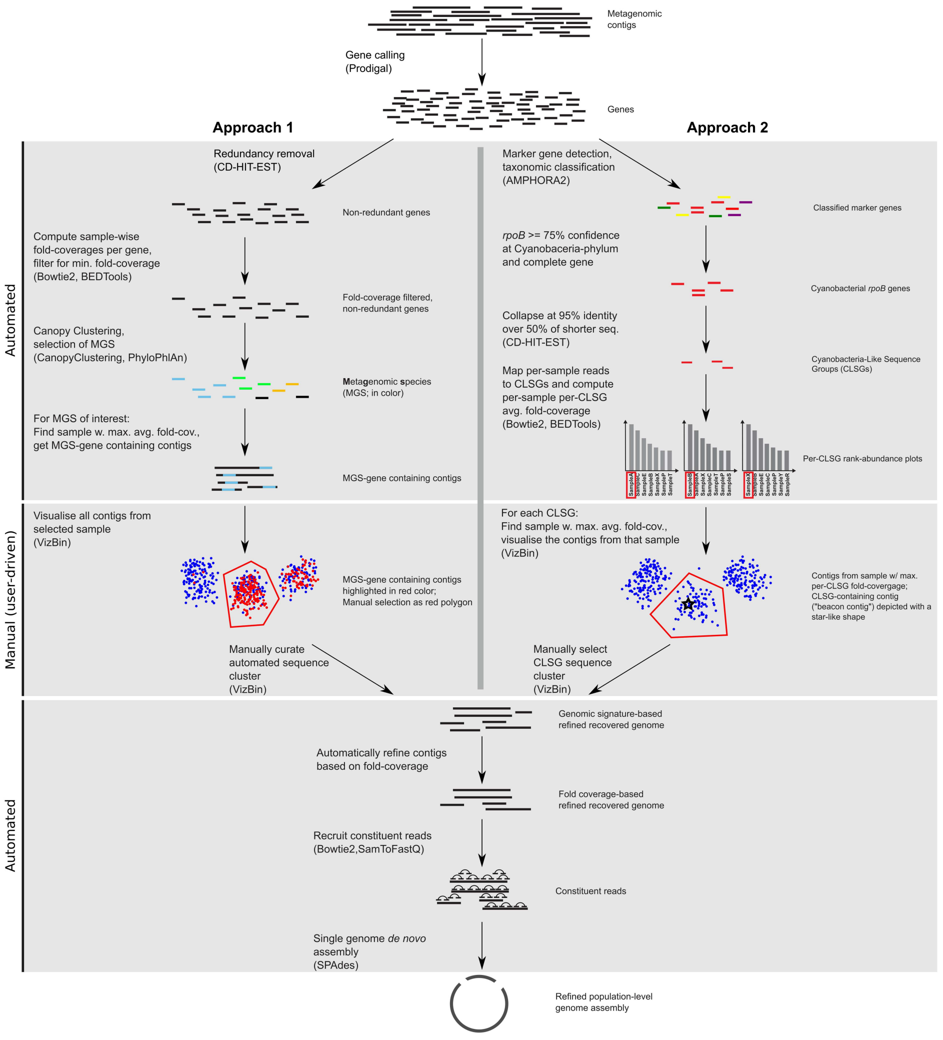

Metagenomic contigs
Gene calling (Prodigal)
Genes
Approach 1
Approach 2
Automated
Redundancy removal (CD-HIT-EST)
Non-redundant genes
Compute sample-wise fold-coverages per gene, filter for min. fold-coverage (Bowtie2, BEDTools)
Fold-coverage filtered, non-redundant genes
Canopy Clustering, selection of MGS (CanopyClustering, PhyloPhlAn)
Metagenomic species (MGS; in color)
For MGS of interest: Find sample w. max. avg. fold-cov., get MGS-gene containing contigs
MGS-gene containing contigs
Marker gene detection, taxonomic classification (AMPHORA2)
Classified marker genes
rpoB >= 75% confidence at Cyanobacteria-phylum and complete gene
Cyanobacterial rpoB genes
Collapse at 95% identity over 50% of shorter seq. (CD-HIT-EST)
Cyanobacteria-Like Sequence Groups (CLSGs)
Map per-sample reads to CLSGs and compute per-sample per-CLSG avg. fold-coverage (Bowtie2, BEDTools)
SampleA SampleC SampleE SampleB SampleX SampleP SampleT
SampleB SampleA SampleX SampleC SampleT SampleP SampleS
SampleX SampleB SampleE SampleC SampleP SampleY SampleR
Per-CLSG rank-abundance plots
Manual (user-driven)
Visualise all contigs from selected sample (VizBin)
MGS-gene containing contigs highlighted in red color; Manual selection as red polygon
For each CLSG: Find sample w. max. avg. fold-cov., visualise the contigs from that sample (VizBin)
Contigs from sample w/ max. per-CLSG fold-covergage; CLSG-containing contig ("beacon contig") depicted with a star-like shape
Manually curate automated sequence cluster (VizBin)
Manually select CLSG sequence cluster (VizBin)
Automated
Genomic signature-based refined recovered genome
Automatically refine contigs based on fold-coverage
Fold coverage-based refined recovered genome
Recruit constituent reads (Bowtie2,SamToFastQ)
Constituent reads
Single genome de novo assembly (SPAdes)
Refined population-level genome assembly

FIGURE 1 | Workflow scheme for the identification, recovery, and refinement of hitherto undescribed population-level genomes from metagenomic data. Approach 1: the Canopy Clustering-based clusters are manually inspected and refined using VizBin. Approach 2: CLSGs are identified and used as beacon contigs to highlight the respective clusters. The main tools used at each step (see "Materials and Methods") are listed in parentheses.

This allowed an assessment whether Canopy Clustering could be used for initial cluster definition/identification and whether it would benefit from a *post hoc* application of VizBin for the recovery of hitherto undescribed population-level genomes from human GIT-derived metagenomic data.

The non-redundant gene catalog comprised 8,576,852 non-redundant genes. After filtering for low-fold-coverage genes (<fold-coverage < 2x per sample in all samples), the per-sample fold-coverage values of the remaining genes were aggregated into a 4,343,293-by-55 matrix serving as the input for Canopy Clustering. In total, 365 clusters containing at least 700 genes were returned by Canopy Clustering. Clusters with a minimum of 700 genes are referred to as "MGS" (Nielsen et al., 2014) and the following results focus on the 365 identified MGS. The degrees of completeness and contamination were assessed using the set of 107 essential genes (see "Materials and Methods"). The distribution of the numbers of different essential genes per MGS (reflecting completeness) exhibited two major modes, one around a total of 10 and one around a total of 100 essential genes per MGS (Supplementary Figure S2; top marginal distribution). The median was 47 essential genes (mean was 49.55), with 75% of the MGS having ≤85 of the 107 essential genes in at least one copy. The 365 MGS generally demonstrated a low degree of contamination (Supplementary Figure S2; right marginal distribution). More specifically, the median was three essential genes in multiple copies (mean was 9.1) and 75% of the MGS had ≤10 essential genes in multiple copies. However, an increase in the degree of contamination, i.e., an increase in the numbers of essential genes in multiple copies per MGS, could be observed with an increase in completeness (Supplementary Figure S2). This may be due to various reasons including suboptimal clustering or covarying microorganisms. In any case, suboptimal completeness and contamination results suggested that it could be promising to use a complementary approach, here VizBin, for the inspection and potential refinement of Canopy Clustering-based clusters.

In order to prioritize the automatically generated clusters for inspection and refinement using VizBin, all 365 MGS were taxonomically classified using PhyloPhlAn. This allowed us to focus our efforts on the recovery of genomic complements derived from hitherto undescribed organisms. Among the 365 MGS, 16 lacked a taxonomic assignment at the order/class-level or lower (**Figure 2**). Three of these 16 MGS were selected for further processing using VizBin: MGS00153 – an Alphaproteobacteria-like MGS, MGS00248 – a Mollicutes-like MGS, and MGS113 – a Cyanobacteria-like MGS. For each of the three MGS, the sample that exhibited the highest average fold-coverage of the respective MGS was chosen and the contigs of that sample were used as input for VizBin-based visualization. Using VizBin, contigs were highlighted, which contained genes of the MGS of interest, and the contig bins were refined by manual selection (**Figure 3**).

Contigs encoding genes of MGS00153 and MGS00248 were mostly found within a single cluster each with few outlying contigs in the respective VizBin maps (**Figures 3A,B**). Moreover,

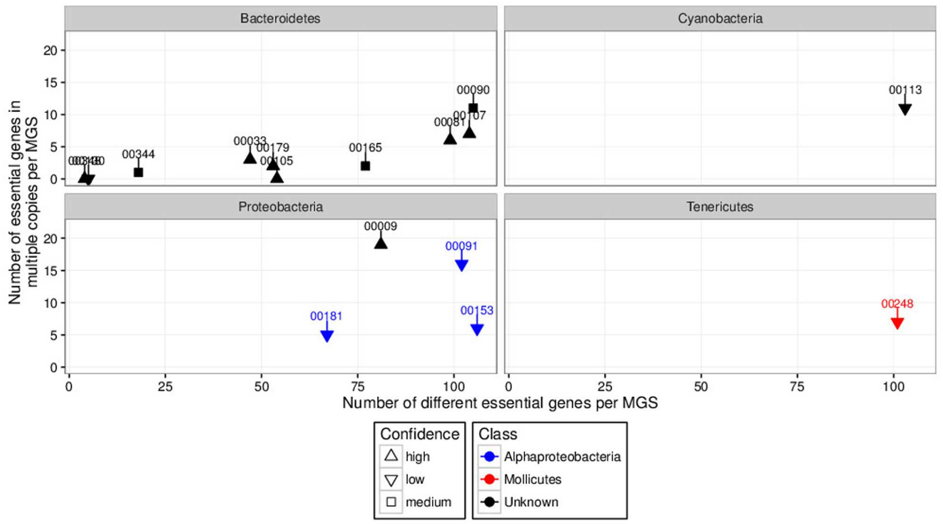


FIGURE 2 | Degrees of completeness (number of different essential genes per MGS; *x*-axis) and contamination (number of essential genes in multiple copies per MGS; *y*-axis) analyses results of automatically generated MGS with unknown order-level assignment grouped by phyla of interest. Some MGS were unclassified at the class-level (black). Numbers above and connected to points indicate respective MGS identifiers. MGS00153, MGS00248, and MGS00113 were selected for inspection and manual refinement using VizBin. The "confidence levels" (high, medium, or low) of the individual taxonomic assignments were assigned by PhyloPhlAn.

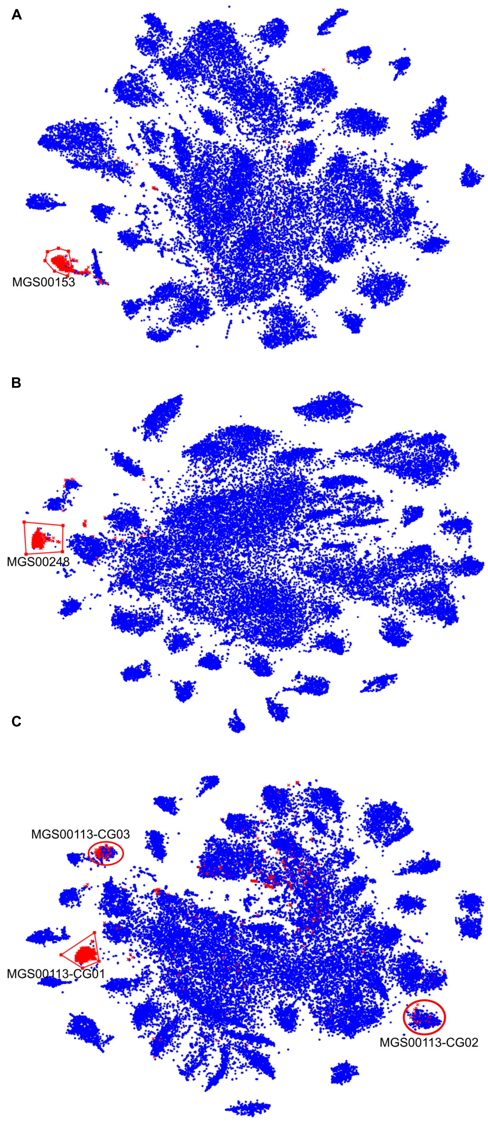


FIGURE 3 | VizBin-based inspection and refinement of three automatically generated MGS. (A–C) Contigs ≥1,000 nt and cut into fragments of 5,000 nt in length (see "Materials and Methods") are shown. Symbols highlighted in red represent fragments coding for at least one gene included in the respective MGS, blue symbols represent fragments not encoding any gene in the respective MGS. Red polygons indicate the selection boundaries manually defined in VizBin.
(A) MGS00153 – Alphaproteobacteria-like. **(B)** MGS00248 – Mollicutes-like. **(C)** MGS00113 – Cyanobacteria-like. Red circles indicate two additional clusters represented by the MGS that were chosen for further inspection.

the gene-wise %GC distributions of the automatically generated or manually defined clusters were highly similar for these two MGS (**Figures 4A,B**). In contrast to MGS00153 and MGS00248, the genes of the Cyanobacteria-like MGS00113 were present in contigs forming part of multiple contig clusters in the VizBin map (**Figure 3C**). The gene-wise %GC distribution of MGS00113 exhibited a large spread as well as relatively high frequencies of genes with divergent %GC, in particular for high %GC values (**Figure 4C**). This further supported that MGS00113 represented a mixture of population-level genomic complements. The three most prominent clusters (composite genomes – CGs) representing genes of MGS00113 were chosen in the VizBin map for further inspection and are referred to as MGS00113-CG01, MGS00113-CG02, and MGS00113-CG03 in the following (**Figure 3C**). All CGs were found to be almost complete with only MGS00113-CG03 containing several essential genes in multiple copies (65/107). An online BLAST search of the CGs' *rpoB* genes revealed limited sequence similarity for MGS00113-CG01 (78% identity over 11% of the query sequence against *Clostridium saccharobutylicum* DSM 13864; top hit), while MGS00113-CG02 and MGS00113-CG03 showed higher sequence similarities (77% identity over 93% of the query sequence against *Coprococcus* sp. ART55/1 and 75% identity over 92% of the query sequence against *Clostridium botulinum* A str. Hall, respectively; top hits). Given the PhyloPhlAn-based Cyanobacteria-like classification of MGS00113, CG01 was compared separately to genomes derived from representatives of the recently defined lineage of GIT-borne Cyanobacteria-like microorganisms, the Melainabacteria (Di Rienzi et al., 2013; Soo et al., 2014). This comparison revealed high sequence similarity with MEL.B1 (mean ANI of 97.11%, Supplementary Figure S3) and suggested that these two (MGS00113-CG01 and MEL.B1) represent closely related strains. MGS00113-CG01 and MGS00113-CG03 were omitted from further analysis due to high sequence similarity to the existing MEL.B1 genome or due to a high degree of contamination, respectively.

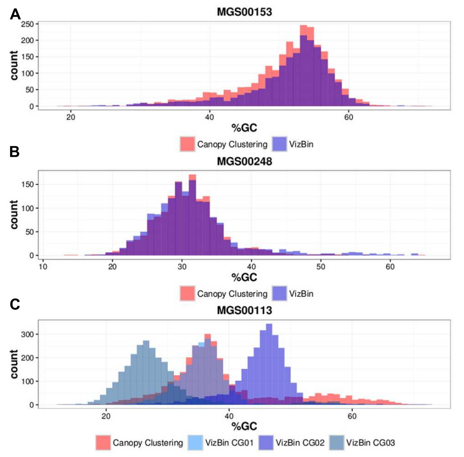


FIGURE 4 | Gene-wise %GC distributions of automatically generated vs. manually defined sequence clusters. (A) MGS00153. **(B)** MGS00248. **(C)** MGS00113.

While the automatically generated and the manually defined clusters were largely concordant with respect to MGS00153 and MGS00248, thus lending mutual support, the case of MGS00113 demonstrated the importance of *post hoc* inspection and refinement. In particular, the PhyloPhlAn-based taxonomic classification of MGS00113 suggested the sequences to be derived from a cyanobacterial organism. However, this classification was misleading as MGS00113 represented a mixture of genomic fragments of at least two classes and at least three distinct organisms.

Targeted Recovery of Genomic Complements Derived from Cyanobacteria-Like Organisms

Motivated by the high similarity between MGS00113-CG01 and MEL.B1, it was intriguing to assess whether further hitherto undescribed Cyanobacteria-like genomic complements could be recovered from the MuSt metagenomic data via *de novo* application of VizBin. To this end, Cyanobacteria-like sequences ("beacon contigs") were used to highlight candidate cyanobacterial clusters and three population-level genomes were subsequently recovered.

Identification of Cyanobacteria-Like Sequence Groups

Phylogenetic marker genes encoded by the *de novo* assembled contigs from the MuSt samples were first identified and taxonomically classified using AMPHORA2. The phylogenetic marker gene which was annotated to belong to the Cyanobacteria phylum most often was the gene encoding the DNA-directed RNA polymerase subunit beta (*rpoB*). A total of 139 cyanobacterial copies of this gene were found in 42 of the 55 MuSt samples. Subsequent sequence similarity-based clustering of complete genes resulted in three sequence clusters which are referred to herein as "CLSGs" and the respective representative genes are referred to as "CLSG marker genes" (**Figure 1**).

Fold-coverages of the CLSG marker genes were used as proxies for estimating population sizes to select for the sample with the highest fold-coverage for each CLSG. The CLSGs were numbered based on descending fold-coverage values with CLSG01 exhibiting the highest fold-coverage in any sample ($\approx$ 109 fold-coverage; M1-4V3, i.e., family 1, individual 4 of that family, sample 3 of that individual), CLSG02 was found to have a fold-coverage of $\approx$ 44 (M2-1V2), and CLSG03 was found to exhibit a quite low fold-coverage ($\approx$ 6.6-fold-coverage; M2-2V2; **Figure 5**). Pronounced intraindividual variations over time for each of the three CLSGs were observed, representing up to two orders of magnitude of differences for CLSG01 (M1-4V2 vs. M1-4V3, i.e., samples at timepoints 2 and 3 of the same individual, in **Figure 5A**).

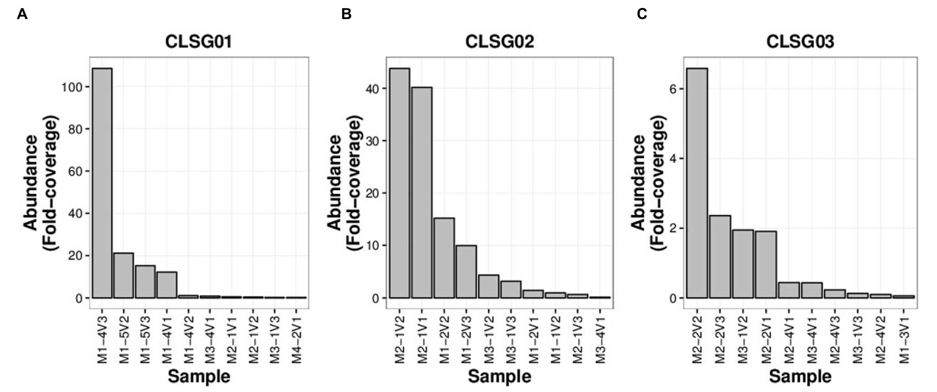


FIGURE 5 | CLSG marker gene rank-abundance plots for the ten samples with the highest apparent abundance of the respective populations. **(A)** CLSG01. **(B)** CLSG02. **(C)** CLSG03. The fold-coverage of CLSG genes was used as a proxy for the abundance of the respective CLSG. Samples are denoted as illustrated by the following example – M1-4V3: Family 1, individual 4 of that family, sample 3 of that individual.

In order to compare our CLSG populations to those previously reported in samples from other geographical locations (Di Rienzi et al., 2013) as well as other hosts (Soo et al., 2014), a phylogenetic tree based on the three herein identified CLSG *rpoB* genes and *rpoB* genes of 10 previously characterized Melainabacteria population-level genomes was constructed (**Figure 6**). Inspection of the tree revealed CLSG02 to be closely related to the MEL.B1 and MEL.C2 genomes. CLSG01 and CLSG03, however, exhibited far lower sequence similarity to previously characterized Melainabacteria and appear to be more distantly related. The environmental Melainabacteria population (ACD20) was found to be basal to the GIT-derived populations. None of the three CLSGs represented outgroups but rather shared phylogenetic relationships with the GIT-derived Melainabacteria.

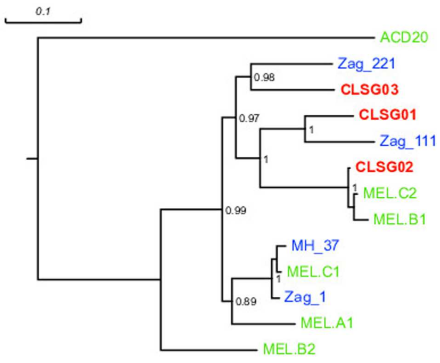


FIGURE 6 | *rpoB* gene-based maximum likelihood phylogenetic tree of previously published Melainabacteria genomes and herein recovered CLSG genomes. Green, blue, and bold red text represent sequences derived from genomes recovered in Di Rienzi et al. (2013) and Soo et al. (2014), and herein, respectively. The environment-derived Melainabacterium ACD20 was chosen as outgroup. Substitutions per site are indicated by the scale-bar on top. Branch support values ≥0.5 are shown for the respective splits.

Possible alternatives to the *rpoB* gene-based approach applied here were the use of *gyrA* or *gyrB* (Kasai et al., 2000; Aranaz et al., 2003; Baker et al., 2004; Menard et al., 2016), or the use of a gene set, e.g., as annotated by AMPHORA2. However, *gyrA* and *gyrB* seemed to be especially promising for the separation of particularly closely related organisms and the use of single marker genes provided important advantages in terms of simplicity over more complex marker gene sets.

Recovery of Population-Level Genomes Guided by CLSG Marker Genes

For each of the three CLSGs, the sample with the highest fold-coverage of the respective CLSG marker genes was selected and visualized using VizBin for cluster definition (**Figure 7**). The clusters containing the CLSG01 and CLSG02 beacon contigs were peripheral and well separated (**Figures 7A,B**). In addition, for CLSG03, the fold-coverage was added as an opaqueness value of the points to improve the delineation of cluster boundaries, thereby facilitating cluster selection (**Figure 7C**). The recovered population-level genomes are herein referred to as CLSG genomes and an overview of genomic and functional features is provided in **Table 1**.

The CLSG01 and CLSG02 genomes both exhibited a high degree of completeness (106/107 essential genes) and a low degree of contamination (3/107 essential genes in

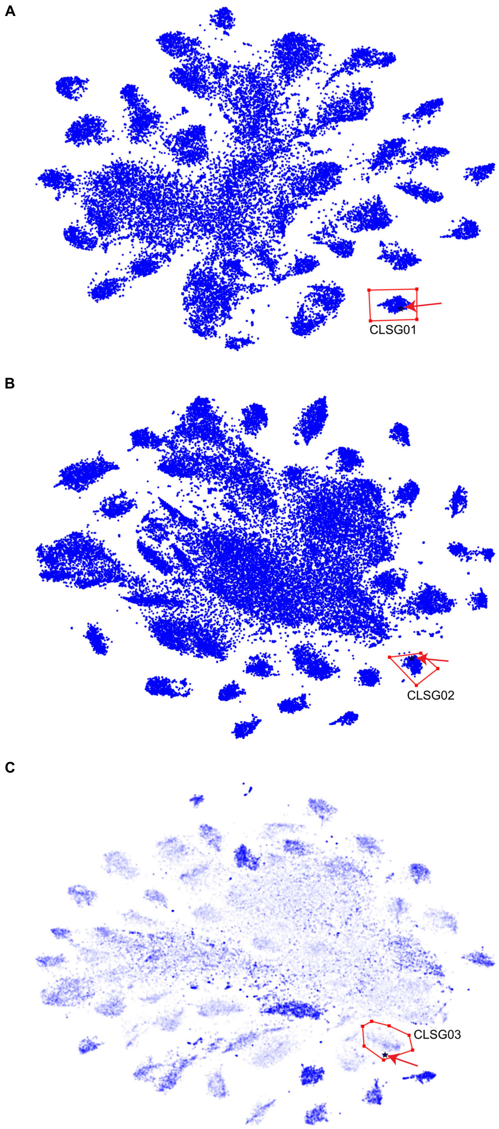


FIGURE 7 | Visualization and cluster selection of Cyanobacteria-like population-level genomes in VizBin. (A) CLSG01. **(B)** CLSG02. **(C)** CLSG03. **(A–C)** Original contig sequences $\geq$1,000 nt, cut into fragments of 5,000 nt. Blue points represent sequences from the samples' metagenomic assemblies. The black star-like shape represents the beacon contig encoding the CLSG marker gene and is highlighted by a red arrow for quicker detection. Red polygons delineate the selected sequence clusters. **(C)** The opaqueness is proportional to the natural logarithm of the fold-coverage and is used here to improve cluster boundary detection.

TABLE 1 | Genomic and functional features of refined and re-assembled population-level genomes.

Population-level genome	MGS00153	MGS00248	MGS00113-CG02	CLSG01	CLSG02	CLSG03
Originating sample	M2-3V2	M2-1V2	M2-1V1	M1-4V3	M2-1V2	M2-2V2
Size (bp)	1,954,779	1,555,611	2,970,300	1,871,540	2,180,307	1,916,257
# Contigs	157	113	408	47	83	551
%GC	50.81	30.48	44.49	32.25	35.21	35.98
# CDS	2,049	1,410	2,605	1,848	2,139	1,876
# Protein-coding CDS	2,006	1,371	2,560	1,809	2,095	1,852
# rRNAs (complete or partial)	2x 16S	0	4x 16S/23S	0	5S/23S	0
# tRNAs	41	39	40	39	42	24
tRNAs missing for [†]	I/F	F/N	I/F/Y	none	none	C/D/F/H/N/T/Y
# Essential genes (out of 107)	105	76	102	106	106	81
# Multi-copy essent. genes	1	1	17	3	3	3
EMP pathway complete [‡]	Yes	No	Yes	Yes	Yes	Yes
PP pathway complete [*]	No	No	No	No	No	No
TCA cycle complete	No	No	No	No	No	No
Entner–Doudoroff pathway complete	No	No	No	No	No	No
Predicted fermentation products [§]	ET/AC	ET/FO/AC	ET/LA/FO/AC	ET	ET/LA/FO	ET
Classical electron transport chain	No	No	No	No	Partial	No
Rnf electron transport complex	Yes	Partial	Partial	No	No	No
ATP synthase	Yes	Yes	Yes	Yes	Yes	Yes
# Flagellar genes	4	0	42	15	54	12
Vitamins: B1/B2/B3/B6/B9/B12/H [∇]	–/–/–/–/+/–/–	–/–/–/–/–/–/–	–/+/+/–(?)/+/–(?)/–	–/+/–/–/(?)/–/+	–/+/–/–/–(?)/–/+	–/+/–/–/–/–/+

[†]: C, cysteine; D, aspartic acid; F, phenylalanine; H, histidine; I, isoleucine; N, asparagine; T, threonine; Y, tyrosine. [‡]: EMP, Embden-Meyerhof-Parnas. [*]: PP, pentose phosphate. [¶]: TCA, tricarboxylic acid. [§] : ET, ethanol; AC, acetate; LA, lactate; FO, formate. [∇]: –, incomplete; +, complete; ?, partially complete.

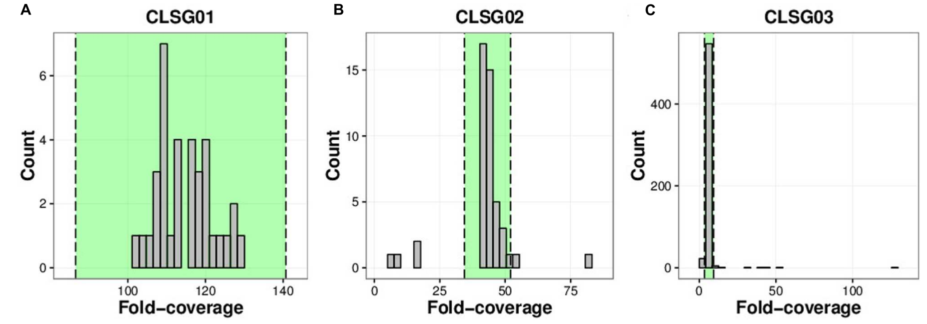


FIGURE 8 | Coverage distributions of individual CLSG genomes. (A) CLSG01 genome. **(B)** CLSG02 genome. **(C)** CLSG03 genome. The distributions of the average per-contig fold-coverages are depicted. Contigs with average fold-coverage values within the range highlighted in green between the two dashed lines are preserved. Contigs outside of this range exhibited exceptionally high or low fold-coverage values and were discarded, i.e., the absolute value of the modified *Z*-score was greater than 3.5.

multiple copies; **Table 1**). In contrast, the CLSG03 genome was less complete (81/107) and more contaminated (11/107 in multiple copies). The fold-coverage distributions of contigs of the individual CLSG genomes indicated that only few contigs exhibited divergent fold-coverage values, i.e., they were of exceptionally high or low fold-coverage (**Figure 8**). This supported the low degree of contamination as already indicated by the essential genes' abundance patterns. The CLSG03 genome exhibited a relatively low overall fold-coverage (<10×). Accordingly, the reduced degree of completeness for this CLSG genome could be due to insufficient sequencing depth for the corresponding population and thus suboptimal assembly of the population-level genome. In contrast, the almost completely recovered genome of the most abundant CLSG, CLSG01, accounted for around 5% of the reads in the originating sample (M1-4V3), thus constituting a sizeable fraction of the sample's metagenomic complement.

Taxonomic and Functional Characterizations of Recovered Genomes

We expected that all genomic complements from a single microbial population would exhibit similar fold-coverages in a given sample. Accordingly, the manually recovered genomic complements were refined by discarding contigs with extreme fold-coverages (**Figure 8**) and the reads constituting the preserved contigs were re-assembled (**Figure 1**). The refined and re-assembled population-level genomes were characterized taxonomically as well as functionally. The main results are summarized here with further details provided in the Supplementary Notes.

MGS00153 – Alphaproteobacteria-Like Population

The refined and re-assembled MGS00153 genome is likely derived from a member of the Alphaproteobacteria class (**Figure 2**; Supplementary Table S2). Unambiguous assignment to a lower taxonomic level, e.g., order or family, remained unresolved: two partially recovered 16S ribosomal RNA (rRNA) genes suggested a placement within the Kopriimonadales order while a MOLE-BLAST search of the recovered *rpoB* gene suggested the Rhizobiales to be the most closely related order (**Figure 9A**). This ambiguity could be due to the lack of reference sequences derived from more closely related microorganisms.

Functional characterization of the MGS00153 genome suggested that the identified organism has a fermentative lifestyle with ethanol and acetate as fermentation products, is auxotroph for most of the vitamins considered, and is unflagellated (**Table 1**; Supplementary Material).

MGS00248 – Mollicutes-Like Population

The Mollicutes class has been assigned to the Firmicutes phylum (Johansson and Pettersson, 2002) and has subsequently been reassigned to the Tenericutes phylum (Ludwig et al., 2009), an ambiguity which is reflected in the case of MGS00248 (**Figures 2** and **9B**, Supplementary Table 3, Supplementary Figure S4B). The genome of MGS00248 exhibited features which are typical for representatives of the Mollicutes class, e.g., genome size (1.5 Mbp), and lack of flagellar assembly genes (**Table 1**). Based on these inferred physiological traits, the organism is suggested herein to be derived from a member of the Mollicutes class.

MGS00113-CG02 – Clostridiales-Like Population

The organism represented by MGS00113-CG02 is likely a member of the Clostridiales order based on the taxonomic analysis results (Supplementary Table S4, Supplementary Figure S4C, **Figure 9C**). Furthermore, it is suggested to be a member of a butyrate-producing subgroup of organisms from the Lachnospiraceae family within the Clostridiales order as the MGS00113-CG02 genome was found to encode a butyrate-kinase (fig| 6666666.163361.peg.1219; Meehan and Beiko, 2014).

Cyanobacteria-Like Populations

CLSG02 and MEL.B1 shared a high sequence similarity (mean ANI of 97.03%, Supplementary Figure S6; **Figure 6**) and thus likely represent the same species. Moreover, MGS00113-CG01 and CLSG02 were found to be almost identical (mean ANI of 100%, Supplementary Figure S7). In contrast, the CLSG01 and CLSG03 genomes were more distantly related to previously recovered genomes from the Melainabacteria class (**Figure 6**; mean ANI of 77.63 and 77.77% to their respective closest relatives, Supplementary Figures S8 and S9) and thus constitute novel melainabacterial representatives.

Large overlaps in the functional potential with other GIT-derived Melainabacteria and limited overlap with an environment-derived Melainabacterium corroborated the taxonomic assignment of the CLSG genomes to the GIT-derived Melainabacteria class within the Cyanobacteria phylum (**Figure 10**; Supplementary Tables S5–S7). Most notably, no photosynthesis genes were found, the genomes were predicted to encode genes for vitamin B production (B_2, B_9, H), and represent obligate anaerobic fermenters (**Table 1**; Supplementary Material). While CLSG02 and MEL.B1 likely represent the same species, genome-specific functions were identified, e.g., a HigB/HigA toxin-antitoxin (TA) system (Christensen-Dalsgaard et al., 2010) was found to be encoded by the CLSG02 genome, yet, this is not encoded by the other melainabacterial genomes.

CONCLUSION

The concurrent characterization of community composition and functional potential through metagenomic sequencing is of great importance for the analysis of microbial communities. However, despite sequence assembly, metagenomic data typically remains fragmented which in turn hampers population-level analyses. Moreover, several microbial lineages are underrepresented in current reference genome catalogs. Therefore, reference-independent computational binning approaches are required for the deconvolution of metagenomic data into population-level genomes derived from hitherto undescribed microorganisms. Here, we applied two reference-independent binning approaches for the identification, recovery, and refinement of such genomes derived from the human GIT. The expansion of the repertoire of currently available reference genomes by the herein recovered representatives is expected to benefit human microbiome-based studies, eventually resulting in improved taxonomic profiling and functional characterization.

First, a multi sample-based, automated binning approach, CANOPY CLUSTERING, was used to perform an initial binning, which was followed by taxonomic classification of the automatically generated bins. The taxonomic classification enabled us to place a focus on sequence clusters likely derived from hitherto undescribed microbial populations for VIZBIN-based *post hoc* inspection and refinement. The importance of complementary approaches, such as VIZBIN, that enable human scrutiny of automatically generated sequence clusters is in particular highlighted by the required refinement of MGS00113

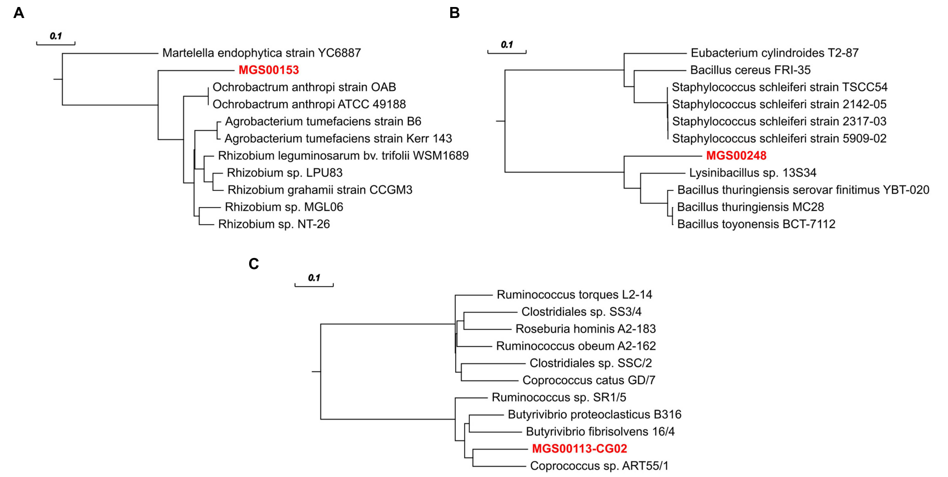

A
0.1
Martelella endophytica strain YC6887
MGS00153
Ochrobactrum anthropi strain OAB
Ochrobactrum anthropi ATCC 49188
Agrobacterium tumefaciens strain B6
Agrobacterium tumefaciens strain Kerr 143
Rhizobium leguminosarum bv. trifolii WSM1689
Rhizobium sp. LPU83
Rhizobium grahamii strain CCGM3
Rhizobium sp. MGL06
Rhizobium sp. NT-26
B
0.1
Eubacterium cylindroides T2-87
Bacillus cereus FRI-35
Staphylococcus schleiferi strain TSCC54
Staphylococcus schleiferi strain 2142-05
Staphylococcus schleiferi strain 2317-03
Staphylococcus schleiferi strain 5909-02
MGS00248
Lysinibacillus sp. 13S34
Bacillus thuringiensis serovar finitimus YBT-020
Bacillus thuringiensis MC28
Bacillus toyonensis BCT-7112
C
0.1
Ruminococcus torques L2-14
Clostridiales sp. SS3/4
Roseburia hominis A2-183
Ruminococcus obeum A2-162
Clostridiales sp. SSC/2
Coprococcus catus GD/7
Ruminococcus sp. SR1/5
Butyrivibrio proteoclasticus B316
Butyrivibrio fibrisolvens 16/4
MGS00113-CG02
Coprococcus sp. ART55/1

FIGURE 9 | *rpoB* gene-based phylogenetic trees for refined MGS genomes and their ten nearest neighbors according to MOLE-BLAST. **(A)** MGS00153 – Alphaproteobacteria-like. **(B)** MGS00248 – Mollicutes-like. **(C)** MGS00113-CG02 – Clostridiales-like. Bold, red text represents the sequences derived from population-level genomes recovered in this work. Scale bars on top represent substitutions per site.

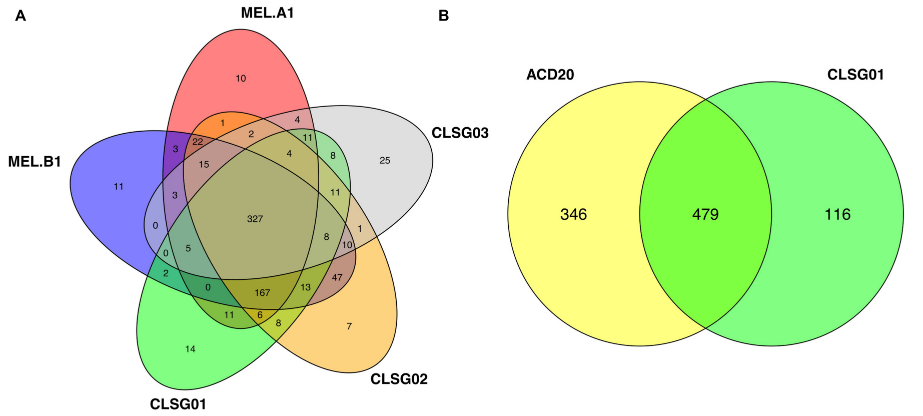


FIGURE 10 | Comparisons of SEED-based functional roles of previously recovered Melainabacteria genomes and CLSG genomes recovered herein. **(A)** The overlaps of SEED-based functional roles for two previously recovered, GIT-derived Melainabacteria population-level genomes (MEL.A1 and MEL.B1; Di Rienzi et al., 2013) and the three CLSG population-level genomes recovered in this work (CLSG01–03) are shown. **(B)** The overlap of SEED-based functional roles between a previously recovered groundwater-derived (ACD20) Melainabacteria population-level genome (Di Rienzi et al., 2013) and the CLSG01 genome is shown.

(multiple apparent clusters in the VizBin map, large spread in %GC of the gene content). Overall, the combination of the two binning approaches resulted in the recovery of one population-level genome from the Alphaproteobacteria class (MGS00153), one from the Mollicutes class (MGS00248), and one from the Clostridiales order (MGS00113-CG02).

Second, a targeted *de novo* recovery of population-level genomic complements from the Melainabacteria was performed, resulting in the recovery of two almost complete genomes and one partial genome (CLSG01–03). The assignment to the Melainabacteria was supported by phylogenetic, genomic, and functional analyses. While large fractions of the functional potential are shared between the herein recovered and the previously described melainabacterial genomes, individual genomes were found to encode genome-specific functions. Moreover, pronounced intraindividual population-abundance variations were observed over time which included differences in estimated population sizes spanning two orders of magnitude.

The observed intraindividual variations of population-abundances highlight the importance of longitudinal studies in the context of *in situ* genome recovery. Despite extensive efforts toward the recovery of microbial genomes from the human GIT, several hitherto undescribed GIT-derived population-level genomes were recovered in this work using the complementary combination of an automated and a user-driven binning approach. It is thus suggested that automated binning approaches should be supplemented with user-driven approaches to ensure the recovery of high-quality population-level genomes from longitudinally collected metagenomic data.

AUTHOR CONTRIBUTIONS

CdB and PW conceived the study, participated in its design and coordination, and drafted the manuscript. CL carried out the binning, performed the taxonomic and phylogenetic analyses, and wrote the paper. EM performed the functional analyses of the recovered and contributed to writing the paper. AH-B carried out the biomolecular extractions and contributed to writing the paper. MH participated in the generation of the functional annotations. LL participated in the biomolecular extraction. AH participated in the design of the study and in the sample and data collection. PM carried out the sequence assembly and gene prediction. CL, EM, AH-B, and PW interpreted the data. All authors read and approved the final manuscript.

FUNDING

The present work was supported by an ATTRACT programme grant (A09/03), a European Union Joint Programme in Neurodegenerative Disease research grant (INTER/JPND/12/01), and a European Union ERASysAPP grant (INTER/SYSAPP/14/05) to PW, an Aide á la Formation Recherche grant (AFR PHD/4964712) to CL, and a Core Junior grant (C15/SR/10404839) to EM, all funded by the Luxembourg National Research Fund (FNR). This work was carried out in the framework of the National Centre of Excellence in Research (NCER) on Parkinson's disease supported by the Luxembourg FNR.

ACKNOWLEDGMENT

In silico analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00884

REFERENCES

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. R. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi: 10.1038/nmeth.3103

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552. doi: 10.1080/10635150600755453

Aranaz, A., Cousins, D., Mateos, A., and Dominguez, L. (2003). Elevation of *Mycobacterium tuberculosis* subsp. caprae Aranaz et al. 1999 to species rank as *Mycobacterium caprae* comb. nov., sp. nov. *Int. J. Syst. Evol. Microbiol.* 53, 1785–1789. doi: 10.1099/ijs.0.02532-0

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Baker, L., Brown, T., Maiden, M. C., and Drobniewski, F. (2004). Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg. Infect. Dis.* 10, 1568–1577. doi: 10.3201/eid1009.040046

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bondoso, J., Harder, J., and Lage, O. M. (2013). *rpoB* gene as a novel molecular marker to infer phylogeny in Planctomycetales. *Antonie Van Leeuwenhoek* 104, 477–488. doi: 10.1007/s10482-013-9980-7

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., and Kjelleberg, S. (2007). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* 73, 278–288. doi: 10.1128/AEM.01177-06

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Christensen-Dalsgaard, M., Jørgensen, M. G., and Gerdes, K. (2010). Three new RelE-homologous mRNA interferases of *Escherichia coli* differentially induced by environmental stresses. *Mol. Microbiol.* 75, 333–348. doi: 10.1111/j.1365-2958.2009.06969.x

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., et al. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36, W465–W469. doi: 10.1093/nar/gkn180

Desper, R., and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.* 21, 587–598. doi: 10.1093/molbev/msh049

Di Rienzi, S. C., Sharon, I., Wrighton, K. C., Koren, O., Hug, L. A., Thomas, B. C., et al. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* 2:e01102. doi: 10.7554/eLife.01102

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10:R85. doi: 10.1186/gb-2009-10-8-r85

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J., Richter, R. A., Valas, R., et al. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* 6, 1186–1199. doi: 10.1038/ismej.2011.189

Eddy, S. (2007). *HMMER – Biosequence Analysis using Profile Hidden Markov Models*. Available at: http://hmmer.janelia.org

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi: 10.1099/ijs.0.64483-0

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Hyatt, D., LoCascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28, 2223–2230. doi: 10.1093/bioinformatics/bts429

Iglewicz, B., and Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. Milwaukee, WI: ASQC Quality Press.

Johansson, K.-E., and Pettersson, B. (2002). "Taxonomy of mollicutes," in *Molecular Biology and Pathogenicity of Mycoplasmas*, eds S. Razin and R. Herrmann (Boston, MA: Springer), 1–29. doi: 10.1007/b113360

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. doi: 10.7717/peerj.1165

Kasai, H., Ezaki, T., and Harayama, S. (2000). Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. *J. Clin. Microbiol.* 38, 301–308.

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., et al. (2012). MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* 7:e47656. doi: 10.1371/journal.pone.0047656

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578. doi: 10.1128/MMBR.00009-08

Laczny, C. C., Pinel, N., Vlassis, N., and Wilmes, P. (2014). Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.* 4, 4516. doi: 10.1038/srep04516

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., et al. (2015). VizBin – an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* 3, 1. doi: 10.1186/s40168-014-0066-1

Lagier, J. C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., et al. (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin. Microbiol. Infect.* 18, 1185–1193. doi: 10.1111/1469-0691.12023

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11070–11075. doi: 10.1073/pnas.0504978102

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi: 10.1093/bioinformatics/btp336

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Ludwig, W., Schleifer, K.-H., and Whitman, W. (2009). Revised road map to the phylum Firmicutes. *Bergeys Manual Syst. Bacteriol.* 3, 1–13.

Meehan, C. J., and Beiko, R. G. (2014). A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biol. Evol.* 6, 703–713. doi: 10.1093/gbe/evu050

Menard, A., Buissonniere, A., Prouzet-Mauleon, V., Sifre, E., and Megraud, F. (2016). The GyrA encoded gene: a pertinent marker for the phylogenetic revision of *Helicobacter* genus. *Syst. Appl. Microbiol.* 39, 77–87. doi: 10.1016/j.syapm.2015.09.008

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., et al. (2012). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Muller, E. E. L., Pinel, N., Laczny, C. C., Hoopmann, M. R., Narayanasamy, S., Lebrun, L. A., et al. (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* 5, 5603. doi: 10.1038/ncomms6603

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rajilić-Stojanović, M., and de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* 38, 996–1047. doi: 10.1111/1574-6976.12075

Roume, H., Heintz-Buschart, A., Muller, E. E. L., and Wilmes, P. (2013). Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* 531, 219–236. doi: 10.1016/B978-0-12-407863-5.00011-3

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* 4, 2304. doi: 10.1038/ncomms3304

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Soo, R. M., Skennerton, C. T., Sekiguchi, Y., Imelfort, M., Paech, S. J., Dennis, P. G., et al. (2014). An expanded genomic representation of the phylum Cyanobacteria. *Genome Biol. Evol.* 6, 1031–1045. doi: 10.1093/gbe/evu073

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). "Next generation sequence assembly with AMOS," in *Current Protocols in Bioinformatics*, 33rd Edn (Hoboken, NJ: John Wiley & Sons, Inc.). doi: 10.1002/0471250953.bi1108s33

Varrette, S., Bouvry, P., Cartiaux, H., and Georgatos, F. (2014). "Management of an academic HPC cluster: the UL experience," in *Proceedings of the International Conference on High Performance Computing Simulation* (Bologna: IEEE), 959–967. doi: 10.1109/HPCSim.2014.6903792

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35, D5–D12. doi: 10.1093/nar/gkl1031

Wu, M., and Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 28, 1033–1034. doi: 10.1093/bioinformatics/bts079

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2, 26. doi: 10.1186/2049-2618-2-26

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492107

Zhang, H., Gao, S., Lercher, M. J., Hu, S., and Chen, W. H. (2012). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res.* 40, 569–572. doi: 10.1093/nar/gks576

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNAsequences. *J. Comput. Biol.* 7, 203–214. doi: 10.1089/10665270050081478

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Laczny, Muller, Heintz-Buschart, Herold, Lebrun, Hogan, May, de Beaufort and Wilmes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

C.8 First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*.

Emilie E.L. Muller[†], Shaman Narayanasamy[†], Myriam Zeimes, Cédric C. Laczny, Laura A. Lebrun, **Malte Herold**, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul Keim, Paul Wilmes

2017
Standards in Genomic Sciences **12**: 64
DOI: 10.1186/s40793-017-0274-y

Contributions of author include:

- Data analysis and visualization
- Revision of the manuscript

[†] Co-first author

Muller *et al. Standards in Genomic Sciences* (2017) 12:64
DOI 10.1186/s40793-017-0274-y

Standards in
Genomic Sciences

EXTENDED GENOME REPORT **Open Access**

First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression in situ

Emilie E. L. Muller[1,3†], Shaman Narayanasamy[1†], Myriam Zeimes[1], Cédric C. Laczny[1,4], Laura A. Lebrun[1], Malte Herold[1], Nathan D. Hicks[2], John D. Gillece[2], James M. Schupp[2], Paul Keim[2] and Paul Wilmes[1*]

Abstract

The Gram-negative beta-proteobacterium *Zoogloea* sp. LCSB751 (LMG 29444) was newly isolated from foaming activated sludge of a municipal wastewater treatment plant. Here, we describe its draft genome sequence and annotation together with a general physiological and genomic analysis, as the first sequenced representative of the *Zoogloea* genus. Moreover, *Zoogloea* sp. gene expression in its environment is described using metatranscriptomic data obtained from the same treatment plant. The presented genomic and transcriptomic information demonstrate a pronounced capacity of this genus to synthesize poly-β-hydroxyalkanoate within wastewater.

Keywords: Genome assembly, Genomic features, Lipid metabolism, Metatranscriptomics, Poly-hydroxyalkanoate, Wastewater treatement plant

Introduction

Zoogloea spp. are chemoorganotrophic bacteria often found in organically enriched aquatic environments and are known to be able to accumulate intracellular granules of poly-β-hydroxyalkanoate [1]. The combination of these two characteristics renders this genus particulary interesting from the perspective of high-value resource production from wastewater [2, 3]. In particular, PHA may be used to synthesize biodegradable bioplastics or chemically transformed into the biofuel hydroxybutyrate methyl ester [2].

The genus name *Zoogloea* is derived from the Greek term; meaning 'animal glue', which refers to a phenotypic trait that was previously used to differentiate between *Zoogloea* species and other metabolically similar bacteria [1]. The polysaccharides making up this "zoogloeal matrix" have been proposed to act as a matrix for the adsorption of heavy metals [4].

To date, no genome sequence exists for any of the representative strains of the five presently recognised *Zoogloea* species and thus, limited information is available with regards to the genomic potential of the genus. Here we report the genome of a newly isolated *Zoogloea* sp. strain as a representative of the genus, with a focus on its biotechnological potential in particular for the production of biodiesel or bioplastics. Accordingly, we studied the *Zoogloea* core metabolism of the genus, particularly on the lipid accumulating properties of *Zoogloea* sp. LCSB751. Moreover, we integrate metatranscriptomic sequencing data to resolve gene expression of this genus in situ [5, 6]. Finally, we also analyze the clustered regularly interspaced palindromic repeats mediated defence mechanisms of *Zoogloea* sp. LCSB751 to infer putatively associated bacteriophages [7].

Organism information

Classification and features

Zoogloea sp. LCSB751 was isolated from an activated sludge sample collected from the surface of the first anoxic tank of the Schifflange communal wastewater treatment plant, Schifflange, Luxembourg (49°30′ 48.29′′N; 6°1′4.53′′E) on 12 October 2011. The activated sludge sample was processed by serial dilution with sterile physiological water to a factor of 10^4 and the biomass was then cultivated on solid MSV peptone

---

* Correspondence: paul.wilmes@uni.lu
†Equal contributors
[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
Full list of author information is available at the end of the article

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

medium [8] at 20 °C and under anoxic conditions (less than 100 ppm oxygen). Single colonies were iteratively re-plated until a pure culture was obtained. The newly isolated *Zoogloea* sp. LCSB751 was cryopreserved in 10% glycerol at −80 °C.

Zoogloea sp. LCSB751 is a facultative anaerobe as it was found to also grow aerobically at 20 °C - 25 °C with agitation in the following liquid media: R2A [9], MSV A + B [8] or Slijkhuis A [10]. Cell clumps were observed in all tested culture conditions. When grown on R2A agar or on MSV peptone agar at 25 °C under aerobic conditions, *Zoogloea* sp. LCSB751 colonies were initially punctiform and after three days, they were white, circular and raised with entire edges. The morphology of cells derived from these growth conditions indicates that these are short rod-shaped bacteria (Fig. 1a). The Gram-staining was negative which is in accordance with previously described isolates of *Zoogloea* spp. [11, 12] (Table 1).

Phylogenetic analysis based on 16S rRNA gene sequences confirmed that strain LCSB751 belongs to the *Zoogloea* genus of the beta-proteobacterial class (Table 1). However, this strain formed a distinct phyletic linage from the five recognized species of *Zoogloea*, that are represented by the type strains *Z. caeni* EMB43^{T} [13], *Z. oleivorans* Buc^{T} [11], *Z. oryzea* A-7^{T} [14], *Z. ramigera* Itzigsohn 1868 ATCC 19544^{T} [15] and *Z. resiniphila* DhA-35^{T} [16, 17] (Fig. 2).

Extended feature descriptions

The capacity of *Zoogloea* sp. LCSB751 to accumulate intracellular granules of lipids was tested using the dye Nile Red as described by Roume, Heintz-Buschart et al. [5]. Figure 1b shows the Nile Red positive phenotype of the described strain.

Additionally, the growth characteristics of the strain *Zoogloea* sp. LCSB751 were determined aerobically and at 25 °C with agitation in 3 different liquid media. Its generation time was the longest in Slijkhuis A medium with the highest biomass production. MSV A + B allowed a generation time of 4 h 30 min but lead to a poor biomass production as demonstrated by the low maximal optical density at 600 nm (OD$_{600}$) of 0.21. The tested liquid medium which allowed the fastest growth for *Zoogloea* sp. LCSB751 was R2A while the biomass production was close to those observed for Slijkhuis A (Table 2).

Genome sequencing information

Genome project history

Overall, 140 pure bacterial isolates were obtained from a single activated sludge sample, and screened for lipid inclusions using the Nile Red fluorescent dye. The genomes of 85 Nile Red-positive isolates were sequenced, of which isolate LCSB065 has already been published [5]. In particular, the genome of *Zoogloea* sp. LCSB751 was analyzed to obtain information about the functional potential of this genus, which has no publically available representative genome sequence, but also based on its particular phylogenetic position and to acquire knowledge on the genes related to lipid accumulation. The permanent draft genome sequence of this strain is available on NCBI with the GenBank accession number MWUM00000000 (BioSample: SAMN06480675). Table 3 summarizes the project information according to the MIGS compliance [18].

Growth conditions and genomic DNA preparation

Zoogloea sp. LCSB751 was grown on MSV peptone agar medium [8] at 20 °C under anoxic conditions. Half of the biomass was scrapped in order to cryopreserve the strain, while the second half was used for DNA extraction using the Power Soil DNA isolation kit (MO BIO, Carlsbad, CA, USA). This cryostock was used to distribute the strain to the Belgian Coordinated Collection of

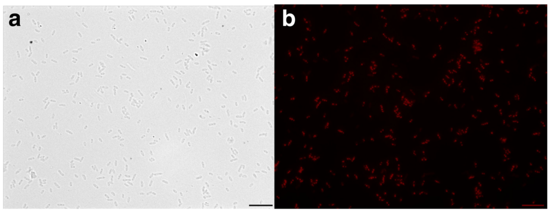


Fig. 1 Photomicrograph of *Zoogloea* sp. strain LCSB751. **a**: bright field of anaerobically grown colonies, Nile Red stained after heat fixation; **b**: same field observed with epifluorescence using an excitation light from a Xenon arc lamp. The beam was passed through an Optoscan monochromator (Cairn Research, Kent, UK) with 550/20 nm selected band pass. Emitted light was reflected through a 620/60 nm bandpass filter with a 565 dichroic connected to a cooled CCD camera (QImaging, Exi Blue). The images were taken using an inverted microscope (Nikon Ti) equipped with a 60× oil immersion Nikon Apo-Plan lambda objective (1.4 N.A) and an intermediate magnification of 1.5×. The scale represents 10 μm. All imaging data were collected and analysed using the OptoMorph (Cairn Research, Kent, UK) and ImageJ

Table 1 Classification and general features of *Zoogloea* sp. strain LCSB751 according to the MIGS recommendation [18]

MIGS ID	Property	Term	Evidence code[a]
	Classification	Domain *Bacteria*	TAS [34]
		Phylum *Proteobacterium*	TAS [35]
		Class *Betaproteobacterium*	TAS [36]
		Order *Rhodocyclales*	TAS [13]
		Family *Rhodocyclaceae*	TAS [13]
		Genus *Zoogloea*	IDA
		Species Unknown	IDA
		Strain: LCSB751	
	Gram stain	Negative	TAS [1]
	Cell shape	Rod	TAS [1]
	Motility	Motile	TAS [1]
	Sporulation	Not reported	NAS
	Temperature range	5–40 °C	TAS [11, 13, 14]
	Optimum temperature	25–30 °C	TAS [11, 13]
	pH range; Optimum	6.0–9.0; 6.5–7.5	TAS [11, 13]
MIGS-6	Habitat	Activated sludge	IDA
MIGS-6.3	Salinity	Inhibited at 0.5% NaCl (*w/v*)	TAS [14]
MIGS-22	Oxygen requirement	facultative anaerobe	IDA
MIGS-15	Biotic relationship	free-living	IDA
MIGS-14	Pathogenicity	non-pathogen	NAS
MIGS-4	Geographic location	Luxembourg	IDA
MIGS-5	Sample collection	2011	IDA
MIGS-4.1	Latitude	49°30′48.29″N;	IDA
MIGS-4.2	Longitude	6°1′4.53″E	IDA
MIGS-4.4	Altitude	275 m	IDA

[a] Evidence codes - *IDA* Inferred from Direct Assay, *TAS* Traceable Author Statement (i.e., a direct report exists in the literature), *NAS* Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [37]

Microorganisms collection center and deposited under number LMG 29444.

Genome sequencing and assembly

The purified DNA was sequenced on an Illumina Genome Analyzer IIx as previously described by Roume, Heintz-Buschart and colleagues [5]. Briefly, a paired-end sequencing library with a theoretical insert size of 300 bp was prepared with the AMPure XP/Size Select Buffer Protocol as previously described by Kozarewa & Turner [19], modified to allow for size-selection of fragments using the double solid phase reversible immobilization procedure [20] and sequenced on an Illumina HiSeq with a read length of 100 bp at TGen North (AZ, USA). The resulting 2,638,115 paired-end reads were trimmed of N bases (i.e. minimum phred quality score of 3 and filtered for Illumina TruSeq3 adapters), retaining 2,508,729 (~95%) of paired reads, 129,378 and eight forward- and reverse-singleton reads (i.e. mate pair discarded), respectively. All reads retained (paired-end and singleton reads) after the pre-processing were de novo assembled using SPAdes ver. 3.1.1, using the default *k*mer range and parameters [21].

The total number of contigs (776), the mean contig length (7497 bp) and the N50 value (180,423 bp) of the draft assembly of *Zoogloea* sp. LCSB751 (Table 3) indicate a fragmented assembly despite an estimated sequencing depth of ~150× fold coverage, ~100× based on 21-mer frequencies (using KMC2 [22]) and a ~ 120× average depth of coverage based on mapping reads back onto the de novo assembled contigs [23–25]. Assembled contigs above 1 kb are represented in Fig. 3.

Genome annotation

Gene (i.e. open reading frame) prediction and annotation was carried out on the assembled contigs using Prokka ver. 1.11 [26] and the RAST server [27], both executed using default parameters and databases. Briefly, Prokka predicted a total of 5200 features including 5118 CDS, 3 rRNA, 76 tRNA genes and one tmRNA genes as well as two repeat regions. Similarly, the RAST server predicted a total of 5202 features, of which 5125 represent coding sequences (CDS), 6 rRNA and 71 tRNA genes. The annotation derived from the RAST server was used for most of the genome descriptions and downstream analyses, unless explicitly mentioned. CDS on the forward and reverse strands within contigs above 1 kb are represented in Fig. 3. In addition, the proteins predicted by the RAST server were submitted to i) the WebMGA server [28], ii) the SignalP server v.4.1 [29] and iii) the TMHMM server v.2.0 [30], for COG functional annotation, signal peptides prediction and transmembrane helices prediction, respectively. 5202 of the predicted amino acid sequences were annotated with 13,030 Pfam IDs. Finally, metaCRT [31] was used to predict CRISPR loci and the resulting CRISPR-spacers were submitted to the CRISPRtarget server [32] for the identification of putatively associated bacteriophage sequences.

Genome properties

The draft genome assembly of *Zoogloea* sp. LCSB751 consists of 5,817,831 bp with a G + C content of 64.2%, distributed over 776 contigs (773 scaffolds) with an N50 value of 180,423 bp (Table 4), GC-skew and –deviation of contigs above 1 kb are represented in Fig. 3. The raw reads are available via the GenBank nucleotide database under the accession number MWUM00000000, while the assembly and the annotation (IDs 6666666.102999) can be accessed through the RAST server guest account.

The rRNA operon region is assumed to be occurring in multiple copies, because all reads from this region

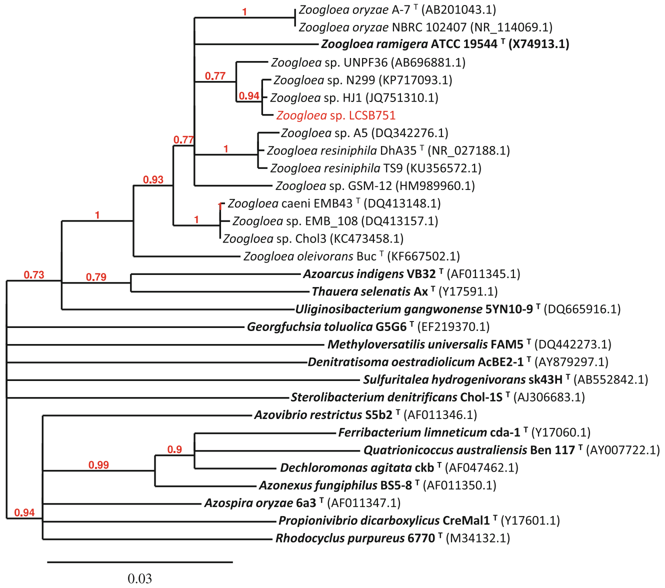

1
Zoogloea oryzae A-7 T (AB201043.1)
Zoogloea oryzae NBRC 102407 (NR_114069.1)
***Zoogloea ramigera* ATCC 19544 T (X74913.1)**
Zoogloea sp. UNPF36 (AB696881.1)
0.77
Zoogloea sp. N299 (KP717093.1)
0.94
Zoogloea sp. HJ1 (JQ751310.1)
Zoogloea sp. LCSB751
0.77
Zoogloea sp. A5 (DQ342276.1)
1
Zoogloea resiniphila DhA35 T (NR_027188.1)
Zoogloea resiniphila TS9 (KU356572.1)
0.93
Zoogloea sp. GSM-12 (HM989960.1)
Zoogloea caeni EMB43 T (DQ413148.1)
1
1
Zoogloea sp. EMB_108 (DQ413157.1)
Zoogloea sp. Chol3 (KC473458.1)
Zoogloea oleivorans Buc T (KF667502.1)
0.73
0.79
***Azoarcus indigens* VB32** T (AF011345.1)
***Thauera selenatis* Ax** T (Y17591.1)
***Uliginosibacterium gangwonense* 5YN10-9** T (DQ665916.1)
***Georgfuchsia toluolica* G5G6** T (EF219370.1)
***Methyloversatilis universalis* FAM5** T (DQ442273.1)
***Denitratisoma oestradiolicum* AcBE2-1** T (AY879297.1)
***Sulfuritalea hydrogenivorans* sk43H** T (AB552842.1)
***Sterolibacterium denitrificans* Chol-1S** T (AJ306683.1)
***Azovibrio restrictus* S5b2** T (AF011346.1)
***Ferribacterium limneticum* cda-1** T (Y17060.1)
0.9
***Quatrionicoccus australiensis* Ben 117** T (AY007722.1)
0.99
***Dechloromonas agitata* ckb** T (AF047462.1)
***Azonexus fungiphilus* BS5-8** T (AF011350.1)
***Azospira oryzae* 6a3** T (AF011347.1)
0.94
***Propionivibrio dicarboxylicus* CreMal1** T (Y17601.1)
***Rhodocyclus purpureus* 6770** T (M34132.1)
0.03

Fig. 2 Phylogenetic tree based on 16S rRNA gene sequences. The type species strains of every species of the *Rhodocyclaceae* family were used (in bold) as well as all the type strains of the genus *Zoogloea*, according to the List of Prokaryotic names with Standing in Nomenclature (LPSN; http://www.bacterio.net). Whole genome GenBank IDs are provided in brackets. The 16S rRNA sequences were aligned using ClustalW, the alignment was curated using Gblocks conserving 81% of the initial positions and the phylogeny was computed with BioNJ using 100 bootstraps and the default (K2P) substitution model, using the pipeline Phylogeny.fr [38]

were assembled into a single contig with a higher depth of coverage (~1200×, for RAST server features: fig|6666666.102999.rna.57, fig|6666666.102999.rna.60 and fig|6666666.102999.rna.61) compared to the rest of the genome. All 20 regular amino-acids were covered by tRNA-anticodons. The RAST server and Prokka annotated approximately 22% (1139) and 26% (1329) of the CDS as hypothetical proteins or proteins of unknown function, respectively. The distribution of COG functional categories are reported in Table 5, while subsystem-based functional classification are available via RAST server.

Table 2 Generation time, growth rate and maximum growth of *Zoogloea* sp. LCSB751 under different aerobic culture conditions

Medium	Generation time ± standard deviation[a]	Growth rate (min^{-1})	Maximum OD^{b}_{600}
R2A	1 h 54 min ± 3 min	0.0058	0.46
MSV A + B	4 h 30 min ± 53 min	0.0026	0.21
Slijkhuis A	10 h 42 min ± 1 h 51min	0.0011	0.73

[a] Values are an average of independent triplicate experiments
[b] OD_{600} stands for optical density measured at 600 nm with the spectrometer “Biochrom WPA CO 8000 Cell Density Meter” using BRAND disposable semi-micro UV cuvettes of 12.5 × 12.5 × 45 mm

Insights from the genome sequence

Genome-based inference of the central metabolism

The genome of *Zoogloea* sp. LCSB751 is predicted to encode for all the genes required for a complete TCA cycle, but is missing some or the complete set of genes for the EMP pathway, the pentose phosphate pathway and the Entner-Doudoroff pathway.

A periplasmic nitrate reductase as well as a nitrite reductase were identified, suggesting complete reduction of nitrate to ammonia by *Zoogloea* sp. LCSB751. Furthermore, a complete set of *nif* genes involved in nitrogen fixation were also encoded in the genome.

Genes for a complete electron transport chain were predicted as well as an alternative RNF complex [33].

The genome of *Zoogloea* sp. LCSB751 also encodes numerous genes for flagella synthesis and assembly, suggesting a motile lifestyle. Furthermore, the strain is predicted to be prototroph for all amino acids, nucleotides

Table 3 Project information

MIGS ID	Property	Term
MIGS 31	Finishing quality	Draft
MIGS-28	Libraries used	Illumina paired-end reads (insert size 30 bp)
MIGS 29	Sequencing platforms	Illumina HiSeq
MIGS 31.2	Fold coverage	150×
MIGS 30	Assemblers	SPAdes (version 3.1.1)
MIGS 32	Gene calling method	RAST server[a] and Prokka[b]
	Locus Tag	fig|6666666.102999
	Genbank ID	MWUM00000000
	GenBank Date of Release	15 March 2017
	GOLD ID	Gs0128811
	BIOPROJECT	PRJNA230567
MIGS 13	Source Material Identifier	LMG 29444
	Project relevance	Environmental, biodiversity, biotechnological

[a] Gene calling using GLIMMER [27, 39]
[b] Gene calling using Prodigal [26, 40]

and vitamins B_2, B_6, B_9, H, and is missing a single gene for the synthesis of B_{12}.

Additionally, the catechol 2,3-dioxygenase that has been studied in *Z. oleivorans*, was found to be encoded by the genome of *Zoogloea* sp. LCSB751 [11].

Lipid metabolism

The genome of *Zoogloea* sp. LCSB751 was further analysed with a focus on genes related to lipid metabolism, to better understand the lipid accumulation properties of *Zoogloea* spp. With 202 genes annotated with COG functional category I "Lipid transport and metabolism", more than 3.8% of the genome of *Zoogloea* sp. LCSB751 is potentially dedicated to lipid metabolism (Table 5 and Fig. 3). Using the SEED subsystem feature, similar results were obtained with 194 genes (3.8%) classified in the "Fatty acids, lipids and Isoprenoids" subsystem (Table 6).

Specifically, a complete set of predicted genes necessary for the synthesis, polymerisation and depolymerisation of PHA [2] was found as well as the genes of the MEP/DOXP pathway for terpenoid synthesis. However,

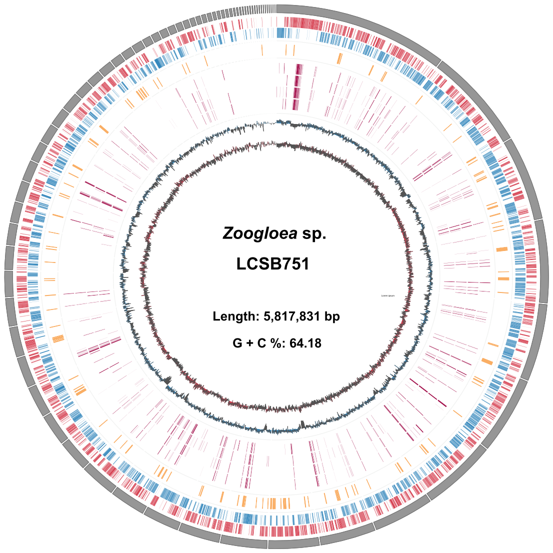


Fig. 3 Circular graphical map of the *Zoogloea* sp. LCSB751 draft genome assembly, annotation and in situ expression. Data shown on the map explained from the outer to inner circles (i-x): i) contigs above 1 kb. Accordingly, all subsequent information contained within inner circles are based on these contigs, including ii) forward strand coding sequences in red (CDS), iii) reverse strand CDS in blue, iv) CDS that are related to lipid accumulation in yellow (forward and reverse strands), v-viii) gene expression in situ based on metatranscriptomic data from four sampling dates (25 January 2011, 11 January 2012, 5 October 2011, and 12 October 2011 [6]) ix) GC-deviation (from overall G + C %) and x) GC-skew, respectively. Graphics were generated using Circos [41]. CDS were predicted and annotated using the RAST server [27]. Metatranscriptomic data from four sampling dates were aligned against the draft genome using BWA [42] and depth of coverage, computed using BEDtools [25] was used as a proxy for expression. Depth of coverage <0.3 were set to zero

Table 4 Genome statistics of *Zoogloea* sp. LCSB751

Attribute	Value	% of Total[a]
Genome size (bp)	5,817,831	100.00
DNA coding (bp)[b]	4,966,077	85.36
DNA G + C (bp)	3,733,728	64.18
DNA scaffolds	773	100.00
Total genes	5,202[c] / 5,200[d]	100.00[c] / 100.00[d]
Protein coding genes	5,125[c] / 5,118[d]	98.52[c] / 98.42[d]
RNA genes	77[c] / 80[d]	1.48[c] / 1.54[d]
Pseudo genes	unknown	unknown
Genes in internal clusters	unknown	unknown
Genes with function prediction [c]	3661	70.38
Genes assigned to COGs	4191	80.56
Genes with Pfam domains	4202	80.78
Genes with signal peptides	505	9.71
Genes with transmembrane helices	1157	22.24
CRISPR repeats	2[d] / 3[e]	2.85

[a]Total is based on either the size of the genome in base pairs, total number of scaffolds or the total number of genes in the annotated genome
[b]Cumulative length of genes, without considering overlaps
[c]As predicted by RAST server [27]
[d]As predicted by Pokka [26]
[e]As predicted by MetaCRT [31]

Table 5 Number of genes associated with general COG functional categories

Code	Value	%age	Description
J	182	3.50	Translation, ribosomal structure and biogenesis
A	3	0.06	RNA processing and modification
K	342	6.57	Transcription
L	204	3.92	Replication, recombination and repair
B	3	0.06	Chromatin structure and dynamics
D	52	1.00	Cell cycle control, Cell division, chromosome partitioning
V	69	1.33	Defense mechanisms
T	564	10.84	Signal transduction mechanisms
M	252	4.84	Cell wall/membrane biogenesis
N	177	3.40	Cell motility
U	142	2.73	Intracellular trafficking and secretion
O	189	3.63	Posttranslational modification, protein turnover, chaperones
C	362	6.96	Energy production and conversion
G	130	2.50	Carbohydrate transport and metabolism
E	305	5.86	Amino acid transport and metabolism
F	85	1.63	Nucleotide transport and metabolism
H	185	3.56	Coenzyme transport and metabolism
I	202	3.88	Lipid transport and metabolism
P	283	5.44	Inorganic ion transport and metabolism
Q	126	2.42	Secondary metabolites biosynthesis, transport and catabolism
R	520	10.00	General function prediction only
S	351	6.75	Function unknown
–	1011	19.43	Not in COGs

Percentage (%) is based on the total number of protein coding genes in the genome

the gene necessary to convert diacylglycerol in triacylglycerol or fatty alcohol in wax ester was not predicted, suggesting that PHA granules are the only lipid bodies accumulated in *Zoogloea* sp. LCSB751.

In situ gene expression

While genomic data provides information about the genetic potential of *Zoogloea* sp. LCSB751, it is possible to study expressed functions of the *Zoogloea* population in situ by using metatranscriptomic data derived from the biological wastewater treatment plant this strain originated from. Metatranscriptomic data derived from samples collected at four distinct time points (25 January 2011, 11 January 2012, 5 October 2011, and 12 October 2011), as studied by Muller and collaborators [6] was used herein. Genes with an average depth of coverage equal or higher than 0.3 were considered as expressed by mapping the rRNA-depleted transcripts on the genome of *Zoogloea* sp. LCSB751. 259, 312, 269 and 330 genes, respectively, were expressed, with 160 of them being expressed at all four time points (Fig. 3 and Additional file 1: Table S1). For the vast majority, (4732 genes), no transcripts were detected, which can be explained by the low population size of *Zoogloea* sp. in situ. This was estimated by phylogenetic marker gene (16S rRNA) amplicon sequencing on the sample collected on 25 January 2011 (data from [6]), for which the *Zoogloea* sp. population size was estimated at 0.1%. Similarly, metagenomic data from all the samples further support the low abundance of this strain in situ (Additional file 1: Table S2).

Nitrate reductase encoding genes (specifically the periplasmic nitrate reductase; NapA) were found to be expressed in all the four time points, while nitrite reductase or nitrogen fixation genes were sporadically expressed in those four time points. Interestingly, at least one copy of the acetoacetyl-CoA reductase and of the polyhydroxyalkanoic acid synthase were found to be expressed at each time point, possibly suggesting PHA accumulation by the population of *Zoogloea* sp. in this environment. Additionally, the third most expressed gene of *Zoogloea* sp. in this environment is a “granule associated protein (phasin)” typically known to be associated with PHA granules.

Table 6 Gene abundance and frequency related to the lipid metabolism of *Zoogloea* sp. LCSB751

Subsystem	Subsystem feature count	Subsystem feature (%)
Fatty acids, lipids and isoprenoids	**194**	**100**
Phospholipids	**30**	**15.46**
Cardiolipin synthesis	2	6.67
Glycerolipid and glycerophospholipid metabolism in bacteria	28	93.33
Triacylglycerols	**3**	**1.55**
Triacylglycerol metabolism	3	100
Fatty acids	**71**	**36.60**
Fatty acid biosynthesis FASII	30	42.25
Fatty acid metabolism cluster	41	57.75
Fatty acids, lipids and isoprenoids - no subcategory	**56**	**28.87**
Polyhydroxybutyrate metabolism	56	100
Isoprenoids	**34**	**17.53**
Isoprenoids for quinones	5	14.71
Isoprenoid biosynthesis	18	52.94
Polyprenyl diphosphate biosynthesis	4	11.76
Nonmevalonate branch of isoprenoid Biosynthesis	7	20.59

The different categories (in **bold**) and subcategories of the subsystem "Fatty acids, lipids and isoprenoid" are represented

CRISPR-*Cas* system and putative bacteriophages

A total of three CRISPR loci were detected with metaCRT, accompanied by six CRISPR-associated (*cas*) genes. Five of the predicted *cas* genes occur consecutively, within the same contig and all of the predicted *cas* genes occur adjacent to a CRISPR locus [7]. Two of CRISPR repeats types were 37 bp in length (sequence: GTTTCAATCCACGTCCGTTATTGCTAACGGACGAATC; GTGGCACTCGCTCCGAAGGGAGCGACTTCGTTGAAGC) while one of them is 32 bp (sequence: CACTCGCTCCGGAGGGAGCGACTTCGTTGAAG). These CRISPRs contain 175, 51 and 11 spacers, respectively, ranging from lengths of 33 to 46 bp. A total of 77 matches were found when searching the spacers against the ACLAME phage/viral/plasmid gene database, NCBI phage and NCBI virus databases using the CRISPRtarget tool [32]. 51 of the spacers match to bacteriophages, 6 to viruses, 11 to genes within plasmids and six to genes within prophages (Additional file 1: Table S3). Based on the available metatranscriptomic data, minute to no expression of the *cas* genes was observed, while the detected CRISPR regions were not covered by the metatranscriptomic data (Additional file 1: Table S1). This is likely due to the overall low abundance of this species in situ (Additional file 1: Table S2).

Conclusions

We describe the first draft genome of a strain potentially belonging to a novel species within the genus *Zoogloea*. The genetic inventory of *Zoogloea* sp. LCSB751 makes it of particular interest for future wastewater treatment strategies based around the comprehensive reclamation of nutrients and chemical energy-rich biomolecules around the concept of a "wastewater biorefinery column" [3] as well as for industrial biotechnological applications. Future comparative genomics studies would allow the scientific community to further confirm if the reported genomic repertoire is indeed typical of this genus. Using metatranscriptomic data, we further show that *Zoogloea* sp. populations are active in the studied wastewater treatment plant despite being low in abundance and likely accumulate PHA in situ.

Additional file

Additional file 1: Table S1. Metatranscriptomic coverage for the predicted features of *Zoogloea* sp. LCSB751. **Table S2.** Metagenomic coverage for the assembly contigs of *Zoogloea* sp. LCSB751. **Table S3.** *Zoogloea* sp. LCSB751 CRISPR spacer complements (protospacer) as per reported by CRISPRTarget [32]. (XLSX 182 kb)

Abbreviations

COG: Clusters of Orthologous Groups; CRISPR: Clustered regularly interspaced palindromic repeats; PHA: Poly-β-hydroxyalkanoate; *Cas*: CRISPR-associated

Acknowledgements

The authors thank Mr. Bissen and Mr. Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC) for their permission to collect samples and gain access to the monitoring platform of the Schifflange wastewater treatment plant; Dr. Nicolás Pinel for assisting with the large data transfers and initial analysis of the isolates. Dr. Aidos Baumuratov for his support for microscopic analyses of the isolates; and the high performance computing facilities of the University of Luxembourg. This work was supported by an ATTRACT program grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) both to PW, and Aide à la Formation Recherche (AFR) grants to both EELM (PRD-2011-1/SR) and SN (PHD-2014-1/7934898) as well as a CORE Junior grant to EELM (C15/SR/10404839), all funded by the Luxembourg National Research Fund (FNR).

Authors' contributions

EELM and LAL isolated the strain, LAL prepared the DNA, NDH prepared the library and sequenced it, SN, MZ, CCL and EELM performed the bioinformatics analyses. MZ performed growth experiments. MH and EELM visualized data. EELM and PW designed and coordinated the project. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg. [2]TGen North, 3051 West Shamrell Boulevard, Flagstaff, AZ 86001, USA. [3]Present address: Department of Microbiology, Genomics and the Environment, UMR

7156 UNISTRA – CNRS, Université de Strasbourg, Strasbourg, France. [4]Present address: Saarland University, Building E2 1, 66123 Saarbrücken, Germany.

Received: 11 April 2017 Accepted: 21 September 2017
Published online: 18 October 2017

References

1. Dugan PR, Stoner DL, Pickrum HM: The Genus *Zoogloea*. In The Prokaryotes: Vol. 7: Proteobacteria: Delta and Epsilon Subclasses. Deeply Rooting Bacteria. New York: Springer Science & Business Media; 2006:1105.
2. Muller EEL, Sheik AR, Wilmes P. Lipid-based biofuel production from wastewater. Curr Opin Biotechnol. 2014;30C:9–16.
3. Sheik AR, Muller EEL, Wilmes P. A hundred years of activated sludge: time for a rethink. Front Microbiol. 2014;5(March):47.
4. Sağ Y, Kutsal T. Biosorption of heavy metals by *Zoogloea ramigera*: use of adsorption isotherms and a comparison of biosorption characteristics. Chem Eng J Biochem Eng J. 1995;60:181–8.
5. Roume H, Heintz-Buschart A, Muller EEL, May P, Satagopam VP, Laczny CC, Narayanasamy S, Lebrun LA, Hoopmann MR, Schupp JM, Gillece JD, Hicks ND, Engelthaler DM, Sauter T, Keim PS, Moritz RL, Wilmes P. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. NPJ Biofilms Microbiomes. 2015;1:15007.
6. Muller EEL, Pinel N, Laczny CC, Hoopman MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, Heintz-Buschart A, Wampach L, Liu CM, Price LB, Gillece JD, Guignard C, Schupp JM, Vlassis N, Baliga NS, Moritz RL, Keim PS, Wilmes P. Community integrated omics links the dominance of a microbial generalist to fine-tuned resource usage. Nat Commun. 2014;5:5603.
7. Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol. 2016;14:67–76.
8. Levantesi C, Rossetti S, Thelen K, Kragelund C, Krooneman J, Eikelboom D, Nielsen PH, Tandoi V. Phylogeny, physiology and distribution of "*Candidatus* Microthrix calida", a new *Microthrix* species isolated from industrial activated sludge wastewater treatment plants. Environ Microbiol. 2006;8:1552–63.
9. Reasoner DJ, Geldreich EE. A new medium for the enumeration and subculture of bacteria from potable water. Appl Environ Microbiol. 1985;49:1–7.
10. Slijkhuis H. *Microthrix parvicella*, a filamentous bacterium isolated from activated sludge: cultivation in a chemically defined medium. Appl Environ Microbiol. 1983;46:832–9.
11. Farkas M, Táncsics A, Kriszt B, Benedek T, Tóth EM, Kéki Z, Veres PG, Szoboszlay S. *Zoogloea oleivorans* sp. nov., a floc-forming, petroleum hydrocarbon-degrading bacterium isolated from biofilm. Int J Syst Evol Microbiol. 2015;65:274–9.
12. Huang T-L, Zhou S-L, Zhang H-H, Bai S-Y, He X-X, Yang X. Nitrogen removal characteristics of a newly isolated indigenous aerobic denitrifier from oligotrophic drinking water reservoir, *Zoogloea* sp. N299. Int J Mol Sci. 2015;16:10038–60.
13. Shao Y, Chung BS, Lee SS, Park W, Lee S-S, Jeon CO. *Zoogloea caeni* sp. nov., a floc-forming bacterium isolated from activated sludge. Int J Syst Evol Microbiol. 2009;59(Pt 3):526–30.
14. Xie C-H, Yokota A. *Zoogloea oryzae* sp. nov., a nitrogen-fixing bacterium isolated from rice paddy soil, and reclassification of the strain ATCC 19623 as *Crabtreella saccharophila* gen. nov., sp. nov. Int J Syst Evol Microbiol. 2006;56(Pt 3):619–24.
15. Unz R. Neotype strain of *Zoogloea ramigera* Itzigsohn. Int J Syst Bacteriol. 1971;21:91–9.
16. Mohn WW, Wilson AE, Bicho P, Moore ER. Physiological and phylogenetic diversity of bacteria growing on resin acids. Syst Appl Microbiol. 1999;22:68–78.
17. International Journal of Systematic Bacteriology. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. List No. 70. Int J Syst Bacteriol 1999, 49:935–936.
18. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26:541–7.
19. Kozarewa I, Turner DJ. 96-plex molecular barcoding for the Illumina Genome Analyzer. Methods Mol Biol. 2011;733:279–98.
20. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW. Unlocking short read sequencing for metagenomics. PLoS One. 2010;5:e11840.
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol A J Comput Mol Cell Biol. 2012;19:455–77.
22. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal k-mer counting. Bioinformatics. 2015;31:1569–76.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
24. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.
25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
26. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
27. Aziz RKK, Bartels D, Best AAA, DeJongh M, Disz T, Edwards RAA, Formsma K, Gerdes S, Glass EMM, Kubal M, Meyer F, Olsen GJJ, Olson R, Osterman ALL, Overbeek RAA, McNeil LKK, Paarmann D, Paczian T, Parrello B, Pusch GDD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
28. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics. 2011;12:444.
29. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.
30. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.
31. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.
32. Biswas A, Gagnon JN, Brouns SJJ, Fineran PC, Brown CM. CRISPRTarget. RNA Biol. 2013;10:817–27.
33. Biegel E, Schmidt S, González JM, Müller V. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell Mol Life Sci C. 2011;68:613–34.
34. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87:4576–9.
35. Garrity GM, Bell JA, Lilburn T. Phylum XIV. phyl. nov. In: DJ Brenner, NR Krieg, JT Staley, GM Garrity (eds), Bergey's Manual of Systematic Bacteriology. Second Edition, Volume 2, Part B. New York: Springer; 2005, p. 1.
36. Garrity GM, Bell JA, Lilburn T. Class II. class. nov. In: DJ Brenner, NR Krieg, JT Staley, GM Garrity (eds), Bergey's Manual of Systematic Bacteriology. Second Edition, Volume 2, Part C. New York: Springer; 2005, p. 575.
37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
38. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36(Web Server issue):W465–9.
39. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27:4636–41.
40. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
41. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:589–95.

C.9 Birth mode determines earliest strain-conferred gut microbiome functions and immunostimulatory potential.

Linda Wampach [†], Anna Heintz-Buschart [†] , Joëlle V. Fritz, Javier Ramiro-Garcia, Janine Habier,
Malte Herold, Shaman Narayanasamy, Anne Kaysen, Angela H. Hogan, Lutz Bindl, Jean Bottu,
Rashi Halder, Conny Sjöqvist, Patrick May, Anders F. Andersson, Carine de Beaufort, Paul
Wilmes
2018
Nature communications **9**: 5091

Contributions of author include:

- Linking the obtained genome reconstructions in distinct samples by essential marker gene sequence homology
- Writing respective method parts and revision of the manuscript

---

[†] Co-first author

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-07631-x OPEN

Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential

Linda Wampach[1,9], Anna Heintz-Buschart [1,2,3], Joëlle V. Fritz[1,4], Javier Ramiro-Garcia[1], Janine Habier[1], Malte Herold[1], Shaman Narayanasamy[1,5], Anne Kaysen[1,4], Angela H. Hogan[6], Lutz Bindl [4], Jean Bottu[4], Rashi Halder [1], Conny Sjöqvist[7,8], Patrick May [1], Anders F. Andersson [7], Carine de Beaufort[4] & Paul Wilmes[1]

The rate of caesarean section delivery (CSD) is increasing worldwide. It remains unclear whether disruption of mother-to-neonate transmission of microbiota through CSD occurs and whether it affects human physiology. Here we perform metagenomic analysis of earliest gut microbial community structures and functions. We identify differences in encoded functions between microbiomes of vaginally delivered (VD) and CSD neonates. Several functional pathways are over-represented in VD neonates, including lipopolysaccharide (LPS) biosynthesis. We link these enriched functions to individual-specific strains, which are transmitted from mothers to neonates in case of VD. The stimulation of primary human immune cells with LPS isolated from early stool samples of VD neonates results in higher levels of tumour necrosis factor (TNF-α) and interleukin 18 (IL-18). Accordingly, the observed levels of TNF-α and IL-18 in neonatal blood plasma are higher after VD. Taken together, our results support that CSD disrupts mother-to-neonate transmission of specific microbial strains, linked functional repertoires and immune-stimulatory potential during a critical window for neonatal immune system priming.

---

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, avenue des Hauts-Fourneaux 7, 4362 Esch-sur-Alzette, Luxembourg. [2] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. [3] Helmholtz Centre for Environmental Research GmbH – UFZ, Theodor-Lieser-Str. 4, 06120 Halle (Saale), Germany. [4] Centre Hospitalier de Luxembourg, rue Nicolas Ernest Barblé 4, 1210 Luxembourg, Luxembourg. [5] Megeno S.A., avenue des Hauts-Fourneaux 9, 4362 Esch-sur-Alzette, Luxembourg. [6] Integrated BioBank of Luxembourg, rue Louis Rech 1, 3555 Dudelange, Luxembourg. [7] KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Tomtebodavägen 23a, 17165 Solna, Sweden. [8] Environmental and Marine Biology, Åbo Akademi University, Tykistökatu 6, 20520 Turku, Finland. [9] Present address: Laboratoire National de Santé, rue Louis Rech 1, 3555 Dudelange, Luxembourg. These authors contributed equally: Linda Wampach, Anna Heintz-Buschart, Joëlle V. Fritz. Correspondence and requests for materials should be addressed to P.W. (email: paul.wilmes@uni.lu)

The past decades have witnessed steadily increasing rates in caesarean section deliveries (CSD) performed largely in the absence of medical necessity and reaching proportions of 19.1% worldwide and 25% in Europe[1,2]. During vaginal birth, specific bacterial strains are transmitted from mothers to infants[3–6] and differences in microbial colonization in neonates born by CSD have been identified[7–10] as early as 3 days postpartum[7,10]. However, due to conflicting results, which principally imply a negligible impact of delivery mode on the colonizing neonatal microbiome in the gut[11], it remains unclear whether disruption of mother-to-infant transmission of microbiota through CSD occurs and whether it affects human physiology early on, with potentially persistent effects in later life. As the first few days after birth represent a 'critical window' in neonatal health and development[12–14], there is growing concern that disruption of microbial transmission from mother to neonate is linked to conditions more frequently observed in CSD-born individuals, including allergies[15], chronic immune disorders[16] and metabolic disorders[17]. To address these concerns, it is essential to determine if there are differences in the functional complement conferred by the earliest colonizing microbiota in relation to CSD, if any differences result from changes in the transmission of strains from mothers to neonates, and if these impact neonatal physiology.

While the majority of studies so far indicate that delivery mode is the strongest factor determining early neonatal gut microbiome colonization[3,7–10,18], these effects are either extenuated or largely absent in other studies[11,19]. In this context, it is important to consider that CSD may be performed as a result of underlying maternal or foetal medical conditions (e.g., multiple gestation, foetal malpresentation or suspected foetal macrosomia)[20] and can co-occur with other microbiome-influencing factors. More specifically, CSD is most often accompanied by the administration of antibiotics to mothers due to local health regulations or hospital practices (e.g., in case of a positive screening of the mother for group B *Streptococcus*)[21]. Being born small for gestational age (SGA) frequently coincides with CSD as well (i.e., more than 50% of all SGA neonates)[22]. SGA neonates have an elevated propensity for developing metabolic disorders during childhood or adulthood, which has been associated with alterations to the gut microbiome[23], and may be linked to the elevated rate of CSD in this population.

Apart from confounding factors, the methods and study designs employed over the past years may in part explain some of the conflicting results regarding the effect of delivery mode on the early gut microbiome. Notably, taxonomic profiling based on 16S rRNA gene amplicon sequencing does not offer sufficient resolution to assess the direct effect of the delivery mode at the level of strain transmission, which is expected to be a determinant of succession. Although recent studies have focused on mother-to-neonate strain transfer and have shown that maternal strains do colonize the neonatal gut, non-vaginal delivery was not assessed comprehensively[4–6]. In addition, although single nucleotide variants (SNVs) have been tracked over time, no such studies have so far covered the earliest time points after delivery (days 0–5) in well-matched mother–neonate pairs in relation to a direct comparison of delivery modes. Consequently, there is a strong need for adequate high-resolution metagenomic analyses capable of resolving the vertical transmission of individual-specific strains and encoded functions from mothers to neonates on an individual basis, while also supplementing observed in silico findings with further in vitro validation experiments.

Independent of whether prenatal colonization of the foetus takes place or not[24], delivery marks the moment of extensive exposure to microbial communities of faecal, vaginal, skin and environmental origins and this event thereby has a profound impact on the colonization of the neonatal gut[9,25]. The initial low microbial biomass in the earliest neonatal stool samples[7] makes the sequencing data prone to the over-representation of putative artefactual reads[24,26] and may contribute to the generation of inconsistent results. Therefore, the removal of any artefactual sequences is essential to ensure unambiguous, high-resolution overviews of the earliest microbial colonization of the neonatal gut.

Here, we performed a detailed analysis of the earliest microbial colonization of the neonatal gut using a combination of 16S rRNA gene amplicon sequencing and high-resolution metagenomics. Our results highlight differences in gut microbiome composition according to delivery mode as well as concurrent differences in the encoded functional potential, which in turn are linked to differences in the transfer of strains from mother to neonate. Based on the enrichment of the LPS biosynthesis pathway in VD neonates, we performed LPS extractions from neonatal stool samples, in vitro immune stimulation assays as well as extensive assessments of LPS purity. The stimulation of primary human immune cells with purified LPS from the faeces of VD neonates collected at day 3 postpartum resulted in the production of higher levels of TNF-α and IL-18. In accordance with these results, the levels of TNF-α and IL-18 in neonatal blood plasma were also higher in VD neonates when compared to CSD. Taken together, we observe a microbiome-driven relationship between delivery mode and endotoxin-induced immune system priming with the potential for lasting effects in later life.

Results

Study design and cohort characteristics. To characterize the temporal patterns of earliest microbial colonization in relation to delivery mode, we recruited and sampled a total of 33 neonates (Supplementary Data 1). The neonatal gut microbiome of some of these neonates had previously been characterized using a combination of 16S rRNA gene amplicon sequencing and quantitative real-time PCR[7]. For a subset of neonates, well-matched neonatal and maternal samples were subjected to high-resolution metagenomic analyses. To differentiate between potential effects of CSD and/or SGA, neonates born by CSD and neonates born by CSD and being SGA were included in the cohort and analysed separately. For each mother–neonate pair, we sampled microbiomes of maternal body sites, which are indicated to be important in relation to neonatal gut colonization (collection of stool and vaginal swabs; Methods) less than 24 h before delivery. Additionally, earliest neonatal stool samples were collected at ≤ 24 h, 3 days and 5 days postpartum (63 samples; Supplementary Data 1). Extracted genomic DNA from all samples was subjected to 16S rRNA gene amplicon sequencing and extracted DNA from the samples of the subset of mother–neonate pairs were subjected to random shotgun sequencing. The 16S rRNA gene amplicon sequencing data were processed using NG-Tax[27], while the resulting metagenomic data were processed using a reproducible, reference-independent bioinformatic pipeline[28].

Removal of artefactual sequences. Neonatal stool samples from days 1–5 postpartum contain limited amounts of microbial DNA[7], and low-biomass samples are prone to over-representation of artefactual DNA that is introduced during the extraction procedure or preparation of sequencing libraries[24,26]. For the 16S rRNA gene amplicon sequencing data, any possible effect from putative artefactual reads was restricted by applying the methodology previously described in Wampach et al[7]. To account for the presence of artefactual sequences in the metagenomic data, we devised an additional, combined in vitro and in silico strategy to identify and remove artefactual sequences from

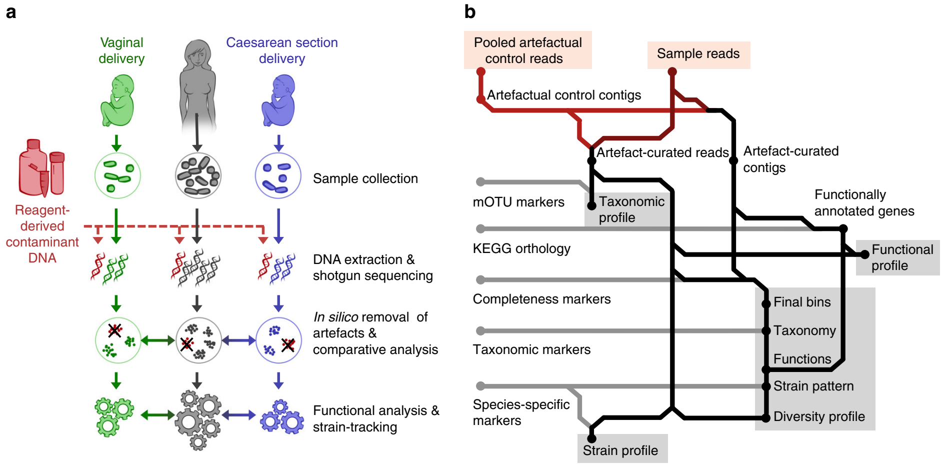


Fig. 1 Curation of metagenomic data. **a** Schematic representation of the workflow for removal of artefacts introduced during genomic extraction or preparation of sequence libraries in the low-biomass neonatal samples. **b** Sample-wise bioinformatic workflow for removal of artefactual sequences from metagenomic data, extraction of taxonomic and functional profiles, and reconstruction of genomes and strain-resolved analyses. The resulting data sets used for inter-sample comparisons are highlighted in grey. mOTU, metagenomic operational taxonomic unit

the metagenomic data (Fig. 1a). For the in vitro part, DNA was extracted from a human gut epithelial cell line using the same procedure as for the neonatal stool samples and diluted to the levels of DNA extractable from the collected low-biomass samples (Methods). The choice of human DNA as a negative control was based on the following criteria: (i) the inability to generate a sequencing library from blank water control samples due to the inherent very low amounts of DNA (these are typically below the threshold for library construction); (ii) the ability to clearly differentiate signal (in the titration series: human sequences) from artefacts (non-human sequences); microbial DNA was not chosen as the homology between contaminant and bona fide sequences may have confounded delineation; (iii) the removal of human sequences is common practice when performing metagenomic analyses on human samples and appropriate methods exist to distinguish between human and microbial sequences in silico; (iv) the blinding of the variability originating from the laboratory environment or sequencing facility due to the nature of the samples (i.e., human control samples were treated with the exact same reagents as the faecal study samples). Our in silico workflow for the identification and removal of artefacts from metagenomic data (Fig. 1b) first clusters[29] contigs from the artefact control samples and the study samples together (Supplementary Fig. 1a). It subsequently removes contigs from study samples that cluster with the artefactual contigs, i.e., that fall into the same bin (Supplementary Note 1). After subsequent filtering steps and the successful removal of artefactual contigs from all study samples, we observed differences in the number of removed reads according to sample type (Supplementary Fig. 1b; Supplementary Data 2). On the basis of this essential data curation step, sequences from *Achromobacter xylosoxidans* or *Burkholderia* spp. taxa were for example identified and subsequently eliminated from the bona fide metagenomic data. Using the curated metagenomic data, we obtained taxonomic profiles (Supplementary Data 3), functionally annotated gene sets (Supplementary Data 4), reconstructed genomes following binning[30] and strain-determining variant patterns[31] (Supplementary Data 5).

Earliest microbial taxonomic profiles. The 16S rRNA gene amplicon and the metagenomic sequencing data, which were generated for a subset of mother–neonate pairs, showed highly similar succession trends in terms of diversity, evenness and richness measures (Supplementary Fig. 2a & b; Supplementary Note 2). The taxonomic profiles derived from the 16S rRNA gene amplicon and metagenomic sequencing were highly correlated (Supplementary Fig. 3a). The differences in taxonomic profiles according to delivery mode reflected results from previous studies, notably the higher relative abundance in *Bacteroides* and *Parabacteroides* and lower levels in *Staphylococcus* in VD neonates at days 3 and 5 postpartum[7,10] (Supplementary Data 6 to 8; Supplementary Note 3). In order to resolve the effect of delivery mode in relation to other potentially contributing factors such as maternal antibiotic intake prior to delivery, gestational age, feeding regime and sampling time point, differentially abundant taxa for both 16S rRNA gene amplicon and metagenomic sequencing data were determined separately using a multivariate additive general model approach (MaAsLin[32]). Taking into account the effects of the above-mentioned factors, delivery mode was found to be the dominant driver of neonatal gut microbiome colonization, with other measured factors having considerably less of an effect (Supplementary Note 4; Supplementary Data 9).

Earliest functional differences according to delivery mode. To assess whether the apparent taxonomic differences between the gut microbiomes of VD and CSD neonates are reflected at the level of functional potential, we used the metagenomic sequencing data to calculate Jensen-Shannon divergences for all samples (Supplementary Fig. 4a). Overall, comparison of the functional profiles of all neonates to the gut microbial potential of their respective mothers highlighted that the neonatal gut microbiota were more divergent from the maternal vaginal microbiota than the corresponding gut microbiota (Supplementary Fig. 4a, b & c). We also compared the CSD (±SGA) gut microbiota at day 3 and day 5 postpartum to those of VD neonates (Fig. 2a). CSD (±SGA)

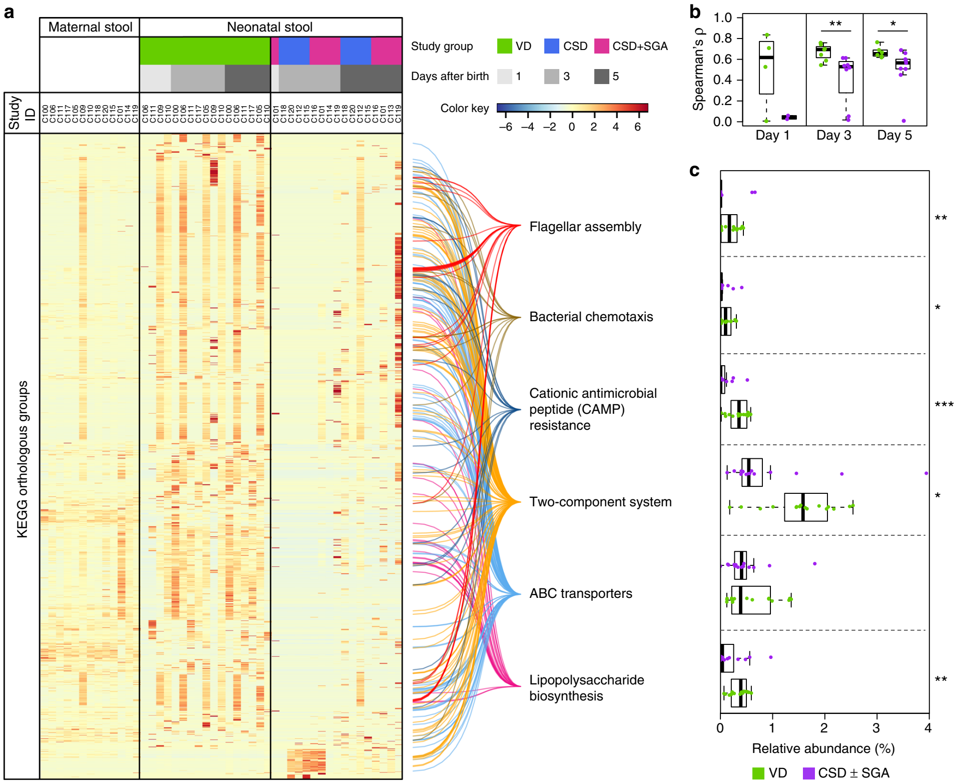


Fig. 2 Maternal and neonatal gut microbiome functional profiles. **a** Heatmap of relative abundance of gut microbial orthologous gene groups with significant differential abundances in neonates born by vaginal delivery (VD) compared to either caesarean section delivery (CSD) or CSD with small for gestational age (SGA) status (CSD + SGA) groups and having the same direction of $\log_2$ fold change (calculated with the R package DESeq2[33]; false-discovery-rate (FDR)-adjusted $P < 0.05$). Colour key indicates row-wise z-scores. The six significantly enriched pathways are indicated (FDR-adjusted $P < 0.05$). **b** Spearman correlation coefficients of functional profiles of neonatal and maternal gut microbiomes in VD and CSD (±SGA) neonates. **c** Cumulative relative abundance of enriched pathways indicated in **a** for VD and CSD (±SGA) groups. **b**, **c** Comparison by Wilcoxon rank-sum test for two-group comparisons with multiple testing adjustment; *FDR-adjusted $P < 0.05$, **FDR-adjusted $P < 0.01$, ***FDR-adjusted $P < 0.001$; Boxplots: centre line – median, bounds – first and third quartile, whiskers <= 1.5 × interquartile range

neonates lacked most functions at day 3 compared to VD neonates (Supplementary Fig. 5–10), while some appeared at day 5. Notably, neonatal-maternal correlations between community-wide functional potentials of the gut microbiomes at days 1, 3 and 5 postpartum were higher for VD than for CSD (±SGA) (Fig. 2b; Wilcoxon rank-sum test, FDR-adjusted $P = 6.0 \times 10^{-3}$ for day 3 and $P = 1.8 \times 10^{-2}$ for day 5).

We detected a total of 1,697 functional categories from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) database that were differentially abundant in the comparisons of the gut microbiome of CSD or CSD + SGA neonates to VD neonates. These presented the same directionality of change using the R package DESeq2[33] with a linear model considering the different collection time points containing at least 1,000 KOs (days 3 and 5) as covariates (Fig. 2a; Supplementary Data 10). Among the differentially abundant genes, there was an enrichment in genes involved in LPS biosynthesis (Fig. 2a; hypergeometric test, false discovery rate (FDR)-adjusted $P = 1.5 \times 10^{-9}$), and the proportion of reads mapping to genes involved in this pathway was larger in VD neonates compared to both CSD groups neonates (Fig. 2c; Wilcoxon rank-sum test, FDR-adjusted $P = 9.6 \times 10^{-3}$). Other important microbial metabolic pathways, which were enriched with differentially abundant genes between VD and CSD (±SGA), included flagellar assembly (Fig. 2a; hypergeometric test, FDR-adjusted $P = 4.9 \times 10^{-12}$), bacterial chemotaxis (Fig. 2a; hypergeometric test, FDR-adjusted $P = 1.5 \times 10^{-2}$), cationic antimicrobial peptide (CAMP) resistance (Fig. 2a; hypergeometric test, FDR-adjusted $P = 4.0 \times 10^{-3}$), two-component system (Fig. 2a; hypergeometric test, FDR-adjusted $P = 2.5 \times 10^{-5}$) and ABC transporters (Fig. 2a; hypergeometric test, FDR-adjusted $P = 1.3 \times 10^{-4}$). As comparisons between VD and CSD as well as VD and CSD + SGA were largely matching independent of SGA status, we combined both groups (CSD and CSD + SGA) to increase statistical power (CSD ± SGA). Notably, all pathways also showed higher relative gene abundances in VD compared to CSD (±SGA) neonates except for the ABC transporter pathway (Fig. 2c; Wilcoxon rank-sum test, FDR-adjusted $P = 4.1 \times 10^{-3}$, 3.8×10^{-2}, 2.2×10^{-4}, 2.1×10^{-2}, respectively).

To corroborate the apparent higher propensity of the VD microbiome for LPS biosynthesis, we annotated the OTUs resulting from the 16S rRNA gene amplicon sequencing data according to their attributed Gram staining information. Hereby, we observed that the gut microbiomes of VD neonates harboured significantly higher relative abundances of Gram-negative bacteria at days 3 and 5 compared to CSD (±SGA) neonates (Wilcoxon rank-sum test, FDR-adjusted $P = 1.7 \times 10^{-3}$ and $P = 4.0 \times 10^{-3}$ for day 3 and 5 respectively; Supplementary Fig. 3b). Additionally, the relative abundances of 7,000 KO functional categories were predicted using PanFP[34] based on the extensive 16S rRNA gene amplicon data (Supplementary Data 11). A multivariate analysis (MaAsLin[32]) was performed to compare the functional profiles of CSD (±SGA) to VD neonates and for both generated data sets (i.e., predicted KO functional categories based on 16S rRNA gene amplicon sequencing data and annotated KOs based on metagenomic sequencing data). Results from the multivariate analyses demonstrated that delivery mode was the strongest determining factor in both data sets (i.e., predicted and metagenomics-based KOs) for explaining the differentially abundant genes (Supplementary Data 9). Whilst not statistically significant, the trends for the predicted microbial pathways obtained with PanFP were largely concordant with the enriched pathways in VD neonates based on the differential analysis of the metagenomic data. Nevertheless, predictions of functional potentials based on 16S rRNA gene amplicon sequencing data are likely unreliable as a significant fraction of the gut microbiome (i.e., up to 40%) is represented by microorganisms without a sequenced isolate genome[35]. In contrast, the metagenomic data, through resolving the actual functional gene complement, allows a detailed comparison of the functional potential of the earliest gut microbiomes, as well as the tracking of individual-specific single-nucleotide variants (SNVs).

Vertical transfer of enteric strains from mothers to neonates. To determine if the observed differences in microbial functions were encoded by specific strains that were vertically transferred from the mother to the neonate, we mined the metagenomic sequencing data to identify microbial taxa and strains that both members of any of the 16 maternal–neonatal pairs, for which we had generated metagenomic data, had in common. We devised an ensemble approach to link reconstructed genomes on the basis of taxonomic annotations[30], similarity of phylogenetic marker genes and the presence of SNVs[31] (Methods). This enabled the tracking of specific strains from mothers to neonates (Supplementary Data 5). Given the high degree of specificity, the presence of transferred strains is highly relevant on a pair-by-pair, individual basis to assess mother-to-neonate transfer. This is all the more important given the extensive inter-individual variability of the neonatal gut microbiome (Supplementary Fig. 3c). Mother-to-neonate transfer differed between VD and CSD (±SGA), with significantly more maternal strains being shared by VD neonates than CSD (±SGA) neonates (Fig. 3a; Wilcoxon rank sum test: $P = 3 \times 10^{-3}$). While the reconstructed genomes of 25 taxa belonging to the phyla Proteobacteria and Firmicutes were identified in maternal–neonatal pairs for all birth modes (mostly the skin-derived and upper-gastrointestinal tract-inhabiting genera *Streptococcus and Staphylococcus* spp.), the reconstructed genomes of 23 enteric taxa belonging to the phyla Bacteroidetes and Actinobacteria (notably *Bacteroides* and *Bifidobacterium* spp.) were exclusively observed in VD pairs. Notably, in the case of vaginal delivery, multiple strains of Gram-positive bacteria (e.g., *Bifidobacterium*) were transferred from mother to neonate (Fig. 3a; transmission in 71% of all VD neonates, 0% in CSD ± SGA on days 3 and 5), as well as Gram-negative bacteria (e.g., Bacteroidetes; Fig. 3a; transmission in 79% of all VD neonates, 0% in CSD and 20% in CSD + SGA on days 3 and 5).

Linking differentially abundant functions to transferred strains. To compare the levels of genetic divergence between early colonizing microbial populations over time, we calculated fixation indices (F_{ST}) and intra-population diversity (π), on the basis of the resolved SNVs. Our results reflected a shift in population structures during the transfer from mothers to neonates. We observed higher fixation indices (Wilcoxon signed-rank test: $P < 4 \times 10^{-3}$) between maternal and neonates' strains (M:3 and M:5; Fig. 3b; Supplementary Fig. 11a) compared to the same transferred strains observed at different times within the neonates (i.e., 3:5). Moreover, intra-population diversity tended to increase during the first days of life for strains that were not transmitted from the mother and belonged to the phylum Firmicutes, as seen in CSD compared to VD neonates (Wilcoxon rank sum test for day 5 versus day 3: $P = 1 \times 10^{-2}$; Fig. 3c; Supplementary Fig. 11b), suggesting that new strains invaded the neonatal gut during this short time period. The differences in relative abundance of taxa corresponded to the inferred routes of transmission linked to birth mode. For example, the metagenomic operational taxonomic unit (mOTU) *Bacteroides dorei/vulgatus* was more abundant in VD neonates, whereas *Staphylococcus epidermidis* was more abundant in CSD (±SGA) neonates (Fig. 3d, e). While the same strain of *B. vulgatus* was present in paired maternal and neonatal samples in the VD group (Fig. 3f), *S. epidermidis* strains were observed only in CSD neonates (Fig. 3g), suggesting an origin outside the maternal gut or vaginal environment. Taken together, these results are consistent with the transmission of strains from the maternal gut microbiome during vaginal delivery, resulting in relatively stable colonization of the neonatal gut during the earliest days, in contrast to CSD neonates.

To assess whether the transferred strains conferred specific functional traits to the neonate or not, we assessed the genomic complements of the earliest microbiota. Analysis of reconstructed genomes that were linked to maternal metagenomes showed that vertically transmitted strains were more likely to be enriched in functions that were depleted in CSD neonates (odds ratio (OR) 5.0, Fisher's exact test $P = 2.4 \times 10^{-11}$). Among these strains, Bacteroidetes (*B. vulgatus* and other *Bacteroides* species) and *Clostridium* spp. were common. Most strikingly, a *B. vulgatus* genome, which shared the majority of SNVs with the corresponding maternal reconstructed genome, was enriched in functions that were significantly more abundant in VD neonates compared to CSD (±SGA) neonates (OR = 3.7, FDR-adjusted $P = 3.0 \times 10^{-186}$; Fig. 3h). By contrast, strains of *Staphylococcus aureus*, an uncharacterized Actinobacterium and *S. epidermidis* encoded functions that were more prevalent in association with CSD. The reconstructed genome of *S. epidermidis* (Fig. 3i) was enriched (OR = 4.1, FDR-adjusted $P = 5.9 \times 10^{-57}$) in functions with higher relative abundances in CSD (±SGA) neonates, and considerably fewer SNVs were shared with the respective mother. These results indicate that vaginal delivery not only favours the vertical transfer of enteric strains from mother to neonate, but also results in the transfer of specific functional traits to the neonate, which are involved in important microbial pathways such as LPS biosynthesis and may be relevant in stimulating the developing immune system during the first days of life.

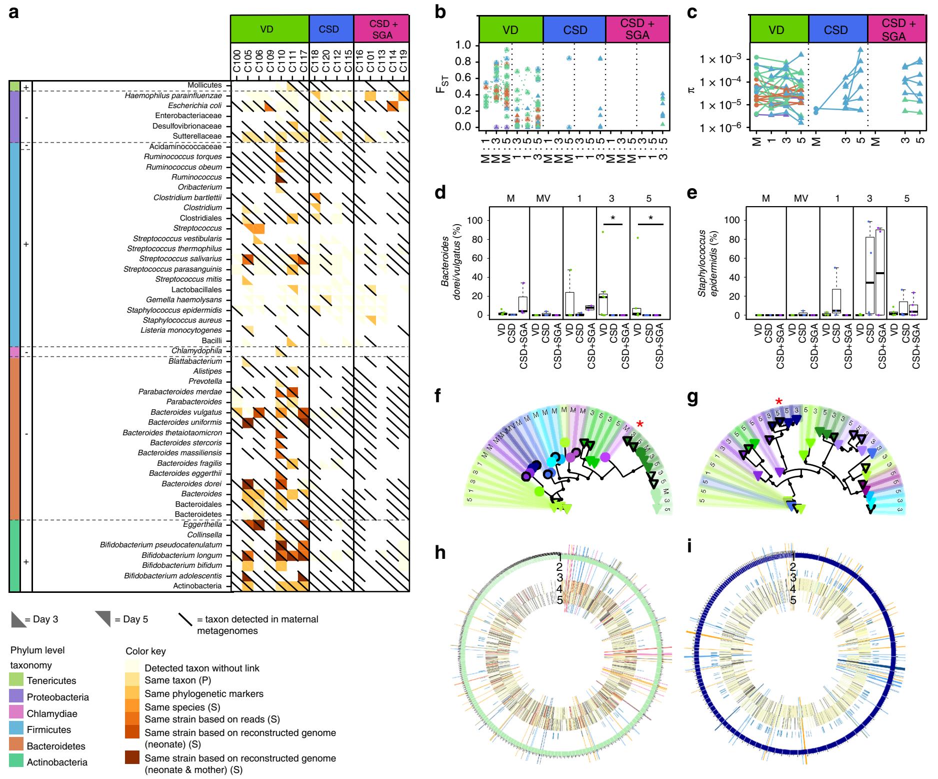


Fig. 3 Transmission of functions by distinct microbial strains. **a** Taxa which were detected in gut microbiomes of mothers (diagonal line) and neonates (on postnatal day 3 (below the line) and/or day 5 (above the line), indicated by shading) in vaginal delivery (VD), caesarean section delivery (CSD) and CSD with small for gestational age (SGA) status (CSD + SGA) groups. The level of evidence of transmission is indicated by the shading colour, with darker shading for stronger evidence. A taxon without link describes a taxon that was found in the maternal samples, but not shared between mother and neonate. P based on PhyloPhlAn; S based on StrainPhlAn. Neonates C115 and C116 are twins. **b** Inter-population fixation indices (F_{ST}) comparing maternal (M) and neonatal (days 1, 3, 5) faecal samples. Phylum-level colour key is given in **a**. Encircled symbols highlight strains that are shared with the respective mother. **c** Intra-population diversity index (π). Circles and triangles represent maternal and neonatal faecal samples, respectively. **d**, **e** Relative abundance of the metagenomic operational taxonomic units (mOTU) belonging to *Bacteroides dorei/vulgatus* (**d**) and *Staphylococcus epidermidis* (**e**) in maternal faecal (M), maternal vaginal (MV) and neonatal faecal (days 1, 3, 5) samples from VD, CSD and CSD + SGA groups; *false discovery rate (FDR)-adjusted $P < 0.05$ in Wilcoxon rank-sum test for two-group comparisons; boxplots: centre line - median, bounds - first and third quartile, whiskers <= 1.5 x interquartile range. **f**, **g** Strain-level phylogenetic trees of *B. vulgatus* (**f**) and *S. epidermidis* (**g**); black bordered and borderless symbols represent genome reconstructions and read-based strain-level identity, respectively; genome reconstructions marked with a red asterisk are represented in **h** and **i**. **h**, **i** Genome reconstructions of *B. vulgatus* from neonatal faeces (C117; VD; day 3; bin P2.2.1) (**h**) and *S. epidermidis* from neonatal faeces (C112; CSD; day 5; bin P2.2) (**i**). Circular tracks represent: assembled contigs (1), single-nucleotide variants (black), shared between mother and neonate (red) (2), positions of strain markers (3), abundance fold-changes between VD and CSD (±SGA) neonates for functionally annotated genes in forward (4) and reverse directions (5); long spokes highlight genes affiliated with enriched pathways as depicted and colour-coded in Fig. 2a

Immunostimulatory potential of the earliest gut microbiome. As LPS forms part of the outer membrane of Gram-negative bacteria, the attributed Gram staining information of microorganisms directly corresponds to their propensity to synthetize LPS. Importantly, LPS is a highly potent innate immune activator that is recognized by the Toll-like receptor (TLR) 4. The earliest VD gut microbiome exhibited an enrichment in the microbial LPS biosynthesis pathway (Figs. 2a and 3h) as well as in Gram-negative taxa, which were frequently transmitted from the mother (Fig. 3a). This observation is supported by the 16S rRNA gene amplicon sequencing data (Supplementary Fig. 3b). Consequently, an apparent higher microbial synthesis of LPS likely results in an increased immunostimulatory potential of the developing gut microbiome. To test whether the VD-associated colonizing gut microbiota, which encode a specific functional complement (including an enrichment in genes involved in LPS biosynthesis), drives early physiological differences in VD neonates, we focussed on

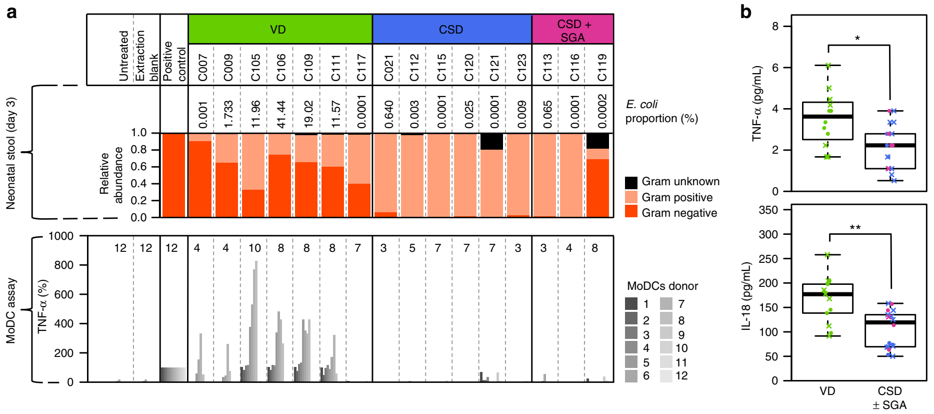


Fig. 4 Cytokines of monocyte-derived dendritic cells after stimulation with LPS from neonatal stool and in neonatal plasma. **a** Lipopolysaccharide (LPS) was isolated from faecal samples collected on day 3 postpartum from neonates in groups of vaginal delivery (VD), caesarean section delivery (CSD) and CSD with small for gestational age (SGA) status (CSD + SGA), and incubated for 24 h with human monocyte-derived dendritic cells (MoDCs) isolated from a total of 12 adult donors. MoDCs were stimulated with the exact same LPS volume that was extractable from the same initial amount of faecal material from each neonate sample (Methods). Exact numbers of donors used per sample are given in the plot. Positive control: LPS isolated from *E. coli* overnight culture. Neonates C115 and C116 are twins. **b** Plasma levels of TNF-α and IL-18 in samples collected at day 3 after birth from VD and CSD (±SGA) neonates. Comparison by Wilcoxon rank-sum test with multiple testing adjustment; *false discovery rate (FDR)-adjusted $P < 0.05$ and **FDR-adjusted $P < 0.01$. Circles correspond to neonates with metagenomic data, crosses represent neonates without metagenomic data. Boxplots: centre line – median, bounds – first and third quartile, whiskers <= 1.5 × interquartile range

the early immunostimulatory potential of LPS from the neonatal gut.

Based on our data, the microbial composition differed most strongly in VD and CSD neonates on day 3 postpartum and thereby may critically affect the developing immune system at this time[12]. We isolated LPS from faecal samples from day 3 with sufficient biomass from 16 neonates (7 VD, 7 CSD, 2 CSD + SGA; Supplementary Data 12; Methods). We subsequently used several approaches to assess the purity of the isolated LPS fractions (Supplementary Note 5; Methods). Using agarose and polyacrylamide gel electrophoresis, we successfully visualized the isolated LPS and did not find traces of protein contamination but observed minor traces of fragmented DNA. However, this DNA did not contribute considerably to the immunostimulatory effect of the LPS fractions (Supplementary Fig. 12). No contamination with peptidoglycan or other bacterial molecules containing the D-glutamyl-meso-diaminopimelic acid moiety were detected in the isolated LPS fractions using the highly sensitive HEK-Blue™ reporter cells overexpressing the receptors hTLR2, hNOD1 and hNOD2, respectively (Supplementary Fig. 13; Methods). Some microbial products detected by hTLR2 (e.g., lipoteichoic acid, lipoprotein from Gram-positive bacteria, lipoarabinomannan from Mycobacteria or zymosan from yeast cell walls) were likely present in LPS samples for which high amounts of LPS were obtained from faecal samples. Conclusively, the isolated LPS fractions were assessed to be of high purity based on the HEK-Blue™ cell assays, although some unknown microbial products may play a stimulatory role in the high-yield LPS fractions (1 ng of standard LPS and an average of 2.9 ng of LPS isolated from VD neonates; Supplementary Fig. 13b). Consequently, the composition of the different isolated LPS fractions played a role in the subsequently triggered immune response.

To assess the stimulatory effect at the interface between the neonatal innate and adaptive immune systems[36], we stimulated monocyte-derived dendritic cells (MoDCs) from 12 adult donors with neonatal LPS extracts (Methods). In order to reflect the in vivo situation as closely as possible, we stimulated the MoDCs with the exact same LPS volume that was obtainable from the same initial amount of faecal material from each neonate sample and subsequently measured levels of the LPS-inducible cytokine TNF-α in the supernatants using an ELISA assay (Fig. 4a; Supplementary Data 13). In parallel, a panel of additional cytokines was measured using an approach for quantifying and normalizing the employed LPS fractions (Methods). This was based on a maximum stimulation of MoDCs with 100 Endotoxin Units (EU) of LPS in order to mimic the amount of LPS an immune cell may encounter within a given neonatal sample (Supplementary Figure 14; Supplementary Data 14 and 15). The levels of all measured cytokines, and especially of TNF-α and IL-18, were higher in culture supernatants from MoDCs treated with LPS from VD neonates (Supplementary Fig. 14; Supplementary Data 16; Supplementary Note 6). Based on the outcome of all applied methods, we observed that the nature and composition of the different LPS subtypes contributed to the level of immune activation triggered by the LPS fractions from the different samples. Taken together, these results demonstrate higher immunostimulatory potentials of the earliest gut microbiome in neonates born by VD compared to CSD.

To test whether potential effects of differential immunostimulation are apparent early on in vivo, we assessed cytokine levels in plasma samples from a total of 31 neonates (Supplementary Data 1). Plasma samples were collected on the same day as the neonatal faecal samples from which LPS was isolated, i.e., day 3 postpartum. Levels of IL-18 were significantly higher in the VD group compared to the CSD group (Wilcoxon rank-sum test, FDR-adjusted $P = 2.4 \times 10^{-3}$; Fig. 4b), as were the levels of TNF-α (FDR-adjusted $P = 3.0 \times 10^{-2}$; Fig. 4b, Supplementary Data 17; Supplementary Note 7), which is consistent with previous results in neonates at a similar age[37] and in healthy infants aged 0–3 months[38].

Discussion

Here we employed high-resolution, artefact-curated metagenomic analyses of paired, high-quality samples from mothers and neonates to resolve the neonatal gut microbiome over the first few days of life. While previous studies have used analogous analytical approaches (16S rRNA gene amplicon sequencing and metagenomics) to resolve the early neonatal gut microbiome, these studies did not involve the systematic collection and appropriate preservation of paired mother–neonate samples[11], they did not specifically track vertical strain transfer[11,18], they did not include provisions for the removal of artefactual sequences[3,5,6,11,18], they did not focus on the earliest time points after delivery[3,18], nor did they resolve differences in functional potential according to delivery mode[3,5,6,11,18]. However, consideration of these factors is essential to assess the effect of delivery mode on the earliest transfer of community structure and function, subsequent microbiome colonization patterns and the resulting effects on neonatal physiology. Our cohort included paired sample sets from mothers (i.e., vaginal swabs and stool collected prior to delivery) and their respective neonates (i.e., stool collected at days 1, 3 and 5 and blood plasma at day 3) born by distinct delivery modes. All microbiome samples yielded high-resolution, artefact-curated sequencing data, which was analysed at the strain level.

As earliest neonatal gut microbiome samples are naturally of low biomass, the accurate identification and curation of potential artefactual sequences is essential. In the absence of appropriate controls, sequences derived from contaminant taxa in reagents may be relatively prominent, thereby masking actual signals and confounding results regarding in particular the transfer of taxa and functions from mothers to neonates. In our study, adequate controls were included and putative artefactual reads removed based on a combined in vitro and in silico workflow. In order to reach the required specificity to unambiguously address the question of vertical transmission of microbial community structure and function from mother to neonate, the use of curated, high-resolution metagenomic sequencing data (rather than solely performing 16S rRNA gene amplicon sequencing) is imperative. More specifically, the applied methodological approach allows the highly specific tracking of individual microbial functions and strains from mother to neonates on a case-by-case basis. Our results based on both 16S rRNA gene amplicon and metagenomic sequencing, and supported by multivariate analyses, demonstrate that early differences exist in the gut microbiomes of neonates and that these differences are predominantly driven by the mode of delivery. Our data agrees with previously reported differences in microbial composition related to birth mode, notably the increased relative abundance of *Bacteroides* and *Parabacteroides* in VD neonates as well as OTUs assigned to *Staphylococcus* being enriched in CSD neonates[3,4,7,39]. No fundamental differences in taxonomical compositions nor functional potentials were apparent when comparing CSD and CSD + SGA neonates, indicating that delivery mode is a stronger determinant for neonatal gut microbial colonization than SGA status.

In line with the broad taxonomic differences between VD and CSD, our findings demonstrate that CSD significantly affects the functional gene complement of the earliest neonatal gut microbiome by impeding the vertical transfer of specific bacterial strains from the maternal gut microbiome to the neonate. Consequently, the gut of CSD neonates is most likely colonized by strains derived from other sources, such as breast milk, skin or saliva, as suggested in previous studies[40–42]. Notably, a selection of enteric strains from the mother was found to be exclusively transferred to the VD neonate (e.g., Bacteroidetes). When measuring population differentiation (F_{ST}), we found evidence that strains acquired from the mother are capable of quickly adapting to the new environment, as observed previously[5]. In contrast, vaginal strains harboured a low potential to stably colonize the neonatal gut thereby further adding to recently published data demonstrating that vaginal taxa do not play a prominent role in the initial colonization of the neonatal gut[5]. This may in part be explained by the distinct niches of the anaerobic gastrointestinal environment and the microaerophilic vaginal environment. With respect to ongoing clinical interventions aimed at restoring the earliest neonatal gut microbiome in the case of caesarean section[43], our findings raise questions over the expected efficiency of microbiota engraftment from a purely vaginal source and suggest that gut-derived strains may be more efficacious.

Independent of the precise mechanism of strain transfer, we observed that several functional pathways were significantly under-represented in CSD neonates, while these were in turn enriched in VD neonates and linked to vertically transmitted strains, in particular the LPS biosynthesis pathway (Fig. 2a). LPS, an outer surface membrane component of Gram-negative bacteria, promotes the secretion of pro-inflammatory cytokines and thereby sits at the interface of the earliest gut microbiome colonization and neonatal immune system priming. Following the apparent enrichments in LPS biosynthesis in VD neonates due to higher amounts of Gram-negative bacteria, the subsequent extraction and quantification of LPS from neonatal stool and stimulation of primary human immune cells therewith demonstrated a reduced immunostimulatory potential of the earliest gut microbiome in CSD neonates. The differences in earliest immune system priming may result in persistent effects on human physiology in later life, which has also been recently suggested based on work in a murine model[44]. On the basis of the observed immunogenicity of the purified LPS fractions, it has not escaped our attention that other factors, including the actual LPS composition, may additionally contribute to the difference in the immunostimulatory potential of the colonizing gut microbiome (Fig. 4a; Supplementary Note 6). Furthermore, other bacterial products, triggering for example TLR2, may contribute towards the observed higher immunostimulatory potential of faecal LPS from VD neonates (Supplementary Fig. 13; Supplementary Note 5). Considering the potential repercussions on neonatal physiology, the detailed elucidation of these additional factors will be the subject of future work.

Our study highlights differences in immunostimulatory potential of the earliest gut microbiome according to delivery mode. This occurs during a critical window of immune system priming. Notably, alterations to early immune system stimulation may be linked to the higher propensity of CSD infants to develop chronic diseases in later life[2]. For example, previous studies focusing on environmental exposure in early life have suggested that the exposure to Gram-negative bacteria and/or environmental endotoxins (such as LPS) could confer protective effects towards allergy development[45,46]. In this context, LPS is likely closely involved in the priming of the neonatal immune system and the subsequent tolerance towards the colonizing gut microbiome during a most critical window in early neonatal life[12–14]. Using a mouse model, it has been shown that strongly immunostimulatory LPS can contribute to the protection from immune-mediated diseases such as diabetes[47] and that disruption of host-commensal interactions in early-life can lead to persistent defects in the development of specific immune subsets[12]. On the basis of additional cytokine measurements in neonatal plasma, VD neonates displayed higher levels of IL-18 and TNF-α, thereby indicating a link between the immunostimulatory potential of microbial LPS in the gut and the overall immune status of the neonatal host early on. Investigations of the longer-term consequences of these differences between CSD and VD neonates will be necessary to assess their possible impact on the development of chronic diseases in later life.

Apart from LPS biosynthesis, other pathways that were significantly enriched in the gut microbiome of VD neonates included genes involved in membrane transport, i.e., ATP-binding cassette (ABC) transporters. On the one hand this may reflect the adaptation of the colonizing microbiome of VD neonates to the gut environment through enhanced nutrient intake. On the other hand, associated ABC transporter proteins for both Gram-positive and Gram-negative bacteria have previously been shown to be immunogenic[48], which may suggest that they play a role in the activation of the neonatal immune system. Additionally, enrichments in pathways relating to bacterial motility were observed. These included the two-component system pathway, which is an important mediator of signal transduction, flagellar assembly and bacterial chemotaxis. These pathways are essential for bacterial motility in response to external stimuli and consequently also for competition with other members of the gut microbiome[49]. Additionally, flagellin, the main structural component of the flagellum, is an effective stimulator of innate immunity[50] and promotes mucosal immunity through the activation of TLR5[51]. Another functional pathway that is potentially interacting with the human immune system early on is the resistance to cationic antimicrobial peptides (CAMP). While this resistance has been observed in all major commensal phyla and across all members of the phylum Bacteroidetes, this pathway is essential to evade detection by the human immune system through the modification of the microbial LPS structure[52]. In the context of our study, an enrichment in CAMP resistance may therefore prevent the dominant colonizers (i.e., Bacteroidetes) from being recognized by the immune system and subsequently removed from the VD neonatal gut. Future studies are needed to assess whether the gut microbiome of VD neonates harbours more modified LPS moieties linked to CAMP resistance and which potential effects the altered LPS structures may have on the neonatal immune system. In accordance with the observation of an apparent enrichment in flagellar biosynthesis, bacterial chemotaxis and CAMP resistance, other microbiota-derived molecular factors, apart from LPS, may be involved in early immune system priming.

Our results imply that a more comprehensive understanding of the effect of the earliest microbial exposure on innate and adaptive immune responses and the different molecular factors involved in neonatal immune system priming is necessary. Future long-term follow-up studies based on larger cohorts, high-resolution multi-omic analyses, detailed immunological screening and tracking of health status will be essential to unravel the interdependencies between mode of delivery, other potential confounding factors, mother-to-neonate transmission, microbiome colonization, exposure to microbial factors, immune system priming and long-term health status. Furthermore, additional sources of maternal strains of importance in relation to microbiome-conferred molecular factors, besides the maternal vagina and gut, have to be considered to assess their relative importance in relation to their impact on neonate physiology. For this, additional samples may be obtained from maternal milk, skin, the oral cavity, and the hospital environment[5,11,40–42]. An additional focus should be placed on uncovering the source and mode of transfer of gut strains from mothers to neonates. Such mechanistic understanding will be important for devising future clinical interventions principally aimed at restoring a VD-like pioneering microbiota in the case of CSD. An alternative approach may consist of ensuring appropriate early priming of the neonatal immune system by the controlled provision of microbial antigens. Both avenues may provide the basis for the development of preventative strategies for adverse health effects in CSD neonates in the future.

Methods

Ethics. Written informed consent was obtained before specimen collection from all enrolled mothers after a detailed consultation. All aspects of recruitment as well as collection, handling, processing and storing of samples and data were approved by the Luxembourg ethics board, the Comité national d'éthique de recherche, under reference number 201110/06 and by the Luxembourg National Commission for Data Protection under reference number A005335/R000058.

Clinical metadata. All study participants were enrolled and gave birth at the Centre Hospitalier de Luxembourg (CHL). Exclusion criteria for mother–neonate pairs included the administration of antibiotics to neonates immediately postpartum, birth prior to 34 weeks of gestation, and maternal gestational diabetes. Clinical metadata for all the analysed time points (days 1, 3 and 5 postpartum) are listed in Supplementary Data 1. Recorded metadata include information on the delivery mode, classification of caesarean section as elective or emergent, birth weight, gestational age, identification of the neonate as small for gestational age (SGA status) where relevant, gender, body length, weight and feeding regime. If a neonate received formula milk at any collection time point, the neonate was considered having received combined feeding for the remainder of the study, as even short-term formula feeding has been shown to cause profound and long-lasting shifts in the gastrointestinal microbiome composition[53]. Enrolled pairs of mothers and neonates ($n = 16$ pairs) included one twin birth (C115 (CSD) and C116 (CSD + SGA)).

Sample collection. Neonatal faecal samples were collected during the first 24 h as well as at days 3 and 5 after birth. Samples and data were collected at the CHL until day 3 after birth; subsequent samples were collected at home by trained study nurses. From the 33 neonates that were recruited into the study, the gut microbiome of 15 (Supplementary Data 1) had previously been characterized using a combination of 16S rRNA gene amplicon sequencing and quantitative real-time PCR[7]. For a subset of neonates, the mother was sampled additionally. Maternal samples (vaginal swabs and faeces) were collected less than 24 h before delivery. Samples were collected into sterile plastic tubes, immediately flash-frozen in liquid nitrogen and stored at −80 °C until further processing. Neonatal blood was collected by capillary or venous sampling, and plasma was isolated and stored at −80 °C from 31 healthy neonates (13 VD, 13 CSD, five CSD + SGA) at day 3 (28 samples) or day 5 (three samples) after birth, including 15 of the 16 neonates for whom metagenomic data were analysed (six VD, four CSD, five CSD + SGA), two neonates for whom no metagenomic but 16S rRNA gene amplicon sequencing data were available (two CSD) and 14 neonates (seven VD, seven CSD) that were sampled under the same conditions[7] (Supplementary Data 1). Clinical data were stored on secure servers at the Luxembourg Centre for Systems Biomedicine (LCSB), and biological samples were stored until further processing at the Integrated BioBank of Luxembourg (IBBL), which is NF S96-900:2011 certified.

Sample processing and extraction of nucleic acids. Genomic DNA was isolated from vaginal swabs with the PowerSoil DNA isolation kit (MO BIO Laboratories; Antwerp, Belgium) with an additional step to increase extraction yield involving the incubation of the samples in PowerSoil tubes with solution C1 at 65 °C for 10 min prior to homogenization for 5 min at 20 Hz in an Oscillating Mill MM 400 (Retsch, Haan, Germany). DNA was subsequently extracted following the manufacturer's instructions.

Faecal samples and cell-culture pellets were processed with the Powerlyzer PowerSoil DNA isolation kit (MO BIO Laboratories), optimized for low-yield samples. Bead solution (500 μl), C1 (60 μl), UltraPure™ Phenol:Chloroform:Isoamyl Alcohol (25:24:1, v/v; Invitrogen, Aalst, Belgium; 200 μl) and 50 mg neonatal stool or 150 mg maternal stool were added to a dry glass bead tube, incubated at 65 °C for 10 min, and homogenized by milling for 45 s at 4 m s^{-1} in a FastPrep-24 5 G (MP Biomedicals, Illkirch-Graffenstaden, France). Samples were centrifuged for 1 min at 12,000 *g*. Solutions C2 (250 μl) and C3 (100 μl) were added to the supernatant and incubated at 4 °C for 5 min, centrifuged for 1 min at 12,000 *g*, then 700 μl of solution C4 and 600 μl of 100% ethanol were added to the supernatant and mixed. 650 μl were loaded onto a Spin Filter and centrifuged at 10,000 × *g* for 1 min. This step was repeated until all lysate had passed through the filter. For the higher input-mass maternal faecal samples, the same isolation procedure was followed except that the filters were washed with a mix of 300 μl solution C4 and 370 μl 100% ethanol, with centrifugation at 10,000 × *g* for 1 min. This latter step was omitted for the low input neonatal samples. All filters were washed with 650 μl 100% ethanol, then 500 μl solution C5. After drying, 60 μl solution C6 was added to the centre of the filter and incubated at room temperature for 5 min. DNA was eluted by centrifugation at 10,000 × *g* for 30 s. RNase A (100 μg ml^{-1}, 2 μl) was added and incubated at 37 °C for ≥ 30 min. Then, one-tenth volume 3 M sodium acetate (pH 6.8) and two volumes isopropanol were added to precipitate the DNA on ice prior to centrifugation. The pellet was washed with 150 μl 70% ethanol, before the dried DNA was dissolved in 50 µl (neonatal faecal samples) or 100 µl (maternal faecal samples) RNase-free water. To obtain an artefact control sample, DNA was extracted from 800,000 trypsinized Caco-2 cells/ml. Caco-2 cells were grown in Dulbecco's Modified Eagle's Medium (Thermo Fisher Scientific, Ghent, Belgium) containing 20% v/v foetal bovine serum and 1% penicillin–streptomycin (Invitrogen) to prevent microbial growth. DNA was extracted with the low-biomass protocol described above, subsequently titrated and samples with 480, 240, 120, 60 and 30 ng total mass were sequenced. DNA integrity and quantity were determined for extracted samples of all origins on 1% agarose gels and in a Qubit 2.0 fluorometer (Thermo Fisher Scientific). Extracted DNA was stored at −80 °C until further use.

DNA sequencing. All DNA samples (along with 8 controls) underwent standard amplicon sequencing of the V4 region of 16S rRNA genes using primers 515F- 5′-GTGBCAGCMGCCGCGGTAA-3′ and 805R- 5′-GACTACHVGGGTATCTAATCC-3′ at the Center for Analytical Research and Technology–Groupe Interdisciplinaire de Génoprotéomique Appliquée (CART-GIGA; Liège, Belgium). Selected DNA samples of maternal (vaginal and faecal extracts), neonatal (faecal extracts at days 1, 3 and 5) and cell-culture origins were subjected to random shotgun sequencing (Supplementary Data 1). Metagenomic libraries were constructed with an optimized low-quantity DNA library preparation kit and sequenced on a HiSeq 2500 platform (Illumina) at GATC Biotech (Konstanz, Germany). For neonatal samples collected from C105, C109, C110 and C119 metagenomic libraries were prepared using TruSeq DNA Nano kit (Illumina) and sequenced on a NextSeq 500 platform (Illumina) at LCSB Sequencing Platform. A total of 84% of the study samples (63 of 75) collected from the mother–neonate pairs yielded sufficient DNA for metagenomic sequencing and sufficient artefact-curated metagenomic data for subsequent analyses.

Metagenomic data processing. Metagenomic data sets were processed with the Integrated Meta-omic Pipeline (IMP; version 1.3), which performs pre-processing, assembly, functional annotation of predicted genes and downstream analyses of Illumina next-generation sequencing metagenomic data in a single, reproducible workflow[28]. Illumina TruSeq3-PE-2 adapter sequences were trimmed from the reads in the pre-processing step (including the removal of human reads), and the de novo assembly step used the MEGAHIT[54] metagenome assembler. The IMP parameters were customized for different sample types: default parameters were retained for maternal faecal samples; for low-biomass samples (maternal vaginal swabs, neonatal faecal samples from days 1, 3 and 5 and cell culture sample), the integrated VizBin[29] sequence cut-off length was set to 1.

Curation of metagenomic data from artefacts. To identify and exclude artefactual sequences in the low biomass samples, contigs were assembled from the sequencing reads obtained from the DNA extracts of the Caco-2 cells after the removal of human reads. Given that the Caco-2 cells were cultured in the presence of 1% penicillin–streptomycin, that the routine surveys for *Mycoplasma* were negative, and that the metagenomic sequencing data did not include any *Mycoplasma* sequences, any bacterial contamination of the mammalian cell culture could be confidently excluded. Then, metagenomic reads from each study sample were mapped against these contigs using Bowtie 2[55] (version 2.0.2). Matching sequences were excluded prior to taxonomic profiling of metagenomic reads by phylogenetic markers[35]. As the artefactual sequences identified in the control samples did not represent full genomes, we further used a binning-based approach to identify additional potential artefactual sequences of the same organism among the de novo assembled contigs of the study samples. After removing the rRNA sequences from the contigs[56], we performed joint binning of control cell-culture contigs with each of the samples' contigs individually using VizBin[29] without any length cut-off. Bins were identified based on VizBin embeddings[56] using density-based spatial clustering of applications with noise (DBSCAN), without correction for the depth of coverage and completeness. All distinct bins (total length < 10 Mbp) that contained > 0.01% of the total contig length of the cell-culture control sample were considered putative reconstructed genomes of artefactual DNA, and the corresponding contigs were removed from the study samples in silico.

Functional profiling. Genes were predicted from contigs assembled with IMP and, after removing artefactual contigs, these genes were functionally annotated with hidden Markov models (HMMs)[56] trained for all KO[57] groups. The functional KO HMMs were aligned using HMMER 3.1[58,59]. The best hit KO (if multiple KOs could be assigned to a gene, the KO with the highest bit score was chosen) for every gene was assigned if the bit score was higher than the binary logarithm of the number of target genes. The FeatureCounts[60] tool with arguments –p and –O was used to extract the number of reads per KO (Supplementary Data 4; representing mean ± standard deviation 77 ± 13 % of all mapping reads).

Linking genome reconstructions by marker gene sequence homology. The curated contigs were binned based on the VizBin embeddings using DBSCAN as well as correction by the depth of coverage and completeness[56]. The reconstructed genomes of all samples belonging to a mother–neonate pair were merged into a union set. For each sample set, predicted amino acid sequences were searched against and annotated using a defined set of essential marker genes[61] using HMMER 3.1[58]. Protein sequences assigned to 35 specific marker genes that form the cross-section of previously suggested sets of phylogenetic marker genes[61,62] were selected. These marker amino acid sequences were clustered with CD-HIT[63] at 97.5% identity. The frequencies of genes from different genome reconstructions co-occurring in the same clusters were determined. A simple graph network representation was constructed with the reconstructed genomes as nodes and counts of co-occurrences between two reconstructed genomes as weighted, undirected edges. Highly interlinked sub-networks, representing related reconstructed genomes, were detected with the cluster_fast_greedy algorithm[64] implemented in the R package igraph (v.1.0.1). The resulting reconstructed genomes from a given sub-network were manually inspected, and the taxonomy of reconstructed genomes was assigned using PhyloPhlAn[30] (Supplementary Data 5).

Strain-level analysis. Strains that occurred in multiple samples were determined with StrainPhlAn[31], using the pre-processed sequencing read data and reconstructed genomes. For each sample, taxonomic profiles were generated from pre-processed reads with MetaPhlAn2[65] using default settings. Strain reconstructions were extracted with the sample2markers.py script in StrainPhlAn with default arguments. StrainPhlAn was used to extract the clades detected in all samples and to construct reference databases for each clade. The sample-based strain reconstructions and reference databases of each clade and all reconstructed genomes were analysed with StrainPhlAn to build multiple sequence alignments and phylogenetic trees. The neonatal samples were considered to share strains with maternal samples if the cophenetic distance between the neonatal microbiome read-based or reconstructed genome-based markers and the maternal markers was less than the distance to the markers of any other individual. Trees were visualized with GraPhlAn (https://bitbucket.org/nsegata/graphlan/wiki/Home). To visualize the positions of markers in genome reconstructions, the reference markers of the species assigned to the reconstructed genomes in StrainPhlAn were aligned to the genome reconstructions post hoc, using blastn and an *E* value cut-off of 1×10^{-10}, as in StrainPhlAn.

Fixation index and intra-population diversity calculation. For all neonatal reconstructed genomes that were estimated to be > 65% complete and linked to at least one other sample of the same neonate or their mother, the fixation index (F_{ST}) and the intra-population diversity (π) were assessed by the presence of SNVs. Metagenomic sequencing reads were mapped against the reconstructed genomes using MOSAIK[66] (version 2.2), with default parameters. A minimum alignment identity of 95% was applied to restrict the mapping to reads of the same species[67]. Genome–sample combinations generating alignments with a median coverage < 20X and/or a breadth < 40% were not included in downstream analyses. To reduce bias stemming from variation in coverage, alignments were down-sampled to a median coverage of 20X using Picard tools (version 1.85; http://broadinstitute.github.io/picard/). SNV calling was performed with FreeBayes[68] (version 1.1.0) using the -pooled-continuous option on the merged alignment files containing all samples for the same genome. Potential SNVs were required to be supported by four or more reads and to have an allele frequency ≥ 1%.

The output from FreeBayes (VCF-file) was used as input for POGENOM (https://github.com/EnvGen/POGENOM), a Perl-based tool that enables population-genomic analysis of metagenome samples. POGENOM was used to calculate the intra-population nucleotide diversity (π), which is defined as the average number of nucleotide differences per site between any two sequence reads chosen randomly from the sample population ($0 \le \pi < 1$). When reads of two or more samples mapped with sufficient coverage to the same genome, the fixation index (F_{ST}) was calculated, reflecting the population differentiation between a pair of samples. F_{ST} is defined as one minus the average intra-population diversity of the samples divided by the nucleotide diversity between the samples (inter-population diversity). POGENOM was tuned to include only the loci recovered in all samples mapped to the same reference genome, assuring a valid comparison of the intra-species variation.

Processing of amplicon sequencing data. Analysis of the 16S rRNA gene amplicon sequences was performed with NG-Tax[27], with default parameters. Operational taxonomic units (OTUs) were assigned to the taxonomy in an open reference approach, using USEARCH[69] against the SILVA[70] 16S rRNA gene amplicon reference database (version 128; Supplementary Data 6). To exclude sequencing artefacts, only dominant phylotypes were examined by removing OTUs that were represented by fewer than 10 reads in the study samples.

Analyses of taxonomic profiles. To determine the Gram staining of the bacteria, we used the NCBI microbial attributes, which can be downloaded from http://www-ab2.informatik.uni-tuebingen.de/megan/taxonomy/microbialattributes.zip. Final Gram staining was assessed by main staining trends per genus and manually curated at the family and order levels. Functional community profiles were predicted based on OTU abundances using PanFP[34].

Statistical data analysis. The R statistical software package (version 3.3.3) was used for statistical analyses and visualization of the taxonomic profiles derived from metagenomic and amplicon sequencing. Sum normalization and calculations of taxon richness (number of metagenomic OTUs (mOTUs) for metagenomic data or OTUs for amplicon sequencing data), diversity (Shannon), evenness (Pielou) indices and Spearman correlation coefficients were performed using the vegan R package. To discover differences in the data sets between the birth modes at the different collection time points postpartum, Wilcoxon rank-sum tests were applied, with FDR multiple-testing adjustment if applicable. Differential taxonomic abundances (according to delivery mode) were also calculated using ANCOM[71] with Benjamini-Hochberg multiple testing correction at 0.05 false-discovery rate. To determine the effect of the variables within the metadata, differentially abundant taxa were also determined using MaAslin[32] with default parameters and a $q < 0.05$ threshold for multi-testing correction. The model used was genus ~ sampling day + maternal antibiotic intake + feeding regime + gestational age. Differential analysis of KO abundance, comparing VD to CSD and VD to CSD + SGA with a linear model, which considered the different collection time points containing at least 1000 KOs (days 3 and 5) as covariates, was performed with the R package DESeq2 version 1.10.1[33]. KOs were considered significantly differentially abundant in VD and CSD (±SGA) if the FDR-adjusted P value of the Wald test was < 0.05 for at least one comparison (CSD vs. VD or CSD + SGA versus VD) and the directionality of change in both comparisons was the same. Principal coordinate analysis (PCoA) graphs were generated using Jensen-Shannon distances as implemented in the R package phyloseq[72]. Differentially abundant pathways were detected through pathway enrichment analysis using a custom R script[56]. Tests for the enrichment of reconstructed genomes with differentially abundant KOs were performed using Fisher's exact test and FDR-adjustment for multiple testing in R.

LPS isolation from neonatal faecal samples. LPS was isolated from 16 selected neonatal faecal samples on the basis of availability of sufficient material. Samples (7 VD, 7 CSD, 2 CSD + SGA; Supplementary Data 12) were collected on day 3 after birth, and from overnight cultures of *Escherichia coli* strain K-12 (sub-strain MG1655). To maximise yields, LPS was purified from three aliquots of 50 mg of each neonatal faecal sample using the hot phenol–water method[73] and further purification was performed using a modified phenol re-extraction protocol[74]. For the *E. coli* control samples, three 5 ml overnight cultures were diluted to an optical density (600 nm) of 0.5 and centrifuged. LPS was isolated from cell pellets by the same protocol as above. LPS for each individual was pooled and quantified using an ELISA-based endotoxin detection assay (Endolisa; # 609033, Hyglos GmbH, Germany). From the 16 neonatal faecal samples, 11 produced measurable amounts of LPS, whereas 5 were under the detection limit (Supplementary Data 12). An extraction blank was generated using the same LPS isolation protocol.

Quantitative real-time PCR to determine bacterial loads. DNA from all neonatal faecal samples used for LPS isolation was diluted (when applicable) to a concentration of 1 ng l^{-1} and amplified in duplicates with universal prokaryotic 16S rRNA gene primers 926F and 1062R[75] and with specific *Escherichia coli* primers Ec461F and Ec780R[76] . Primer sequences, annealing temperatures and cycle details are specified in Supplementary Data 14. Genomic DNA isolated from *Salmonella* Typhimurium LT2 and *E. coli* strain K-12 (sub-strain MG1655) was used to prepare standard curves for universal prokaryotic and specific *E. coli* primers, respectively. Reaction mixture, measurements and calculations of bacterial load (nanograms bacterial DNA per milligram stool and nanograms *E. coli* DNA per milligram stool) were performed as previously described[7] (Supplementary Data 14). The proportion of *E. coli* DNA in comparison to total bacterial DNA was subsequently calculated.

In vitro immunostimulation using LPS from neonatal faecal samples. Primary human monocytes were isolated from blood samples obtained from the Luxembourg Red Cross originating from twelve healthy adult donors. Human neonatal dendritic cells (DCs) were previously shown to be competent in MHC class I antigen processing and presentation to the same extent than adult DCs[77]. Most importantly, the NF-κB-dependent pathway in TLR-4 signalling is intact in neonatal MoDCs as they produce pro-inflammatory cytokines upon LPS stimulation, while adult and neonatal DCs are both able to produce comparable levels of TNF-α, IL-6 and IL-8 in response to LPS[78]. Isolated monocytes were differentiated into dendritic cells (MoDCs) in 12-well plates for 5 days in RPMI 1640 medium (Thermo Fisher Scientific) supplemented with 10% foetal bovine serum (Thermo Fisher Scientific), 20 ng ml^{-1} each of granulocyte-macrophage colony-stimulating factor (Peprotech, London, UK), 20 ng ml^{-1} IL-4 (Peprotech) and 1% penicillin–streptomycin (Invitrogen). To assess the immune stimulatory potential of isolated LPS, we treated MoDCs for 24 h with LPS extracted from VD or CSD (±SGA) neonatal faecal samples using two different methods; one based on LPS volume and one based on the normalization of LPS concentration with the bacterial load (see below for more information).

As we started from the same amount of material for all the neonatal stool samples and used the exact same extraction protocol to isolate all LPS fractions for all samples, we assumed that if we treated MoDCs from the same donor with the exact same volume of yielded LPS (independent of the concentration of LPS present), we would realistically emulate the microbial LPS load which immune cells would be exposed to in vivo and thus be representative of the immunostimulatory potential of a given sample at 3 days postpartum. To stimulate MoDCs, 7.5 μl of LPS extract per 10^5 MoDCs was added per well. For the negative control, MoDCs were incubated with 7.5 μl of LPS extraction blank, and for the positive control, MoDCs were treated with 15 endotoxin units (EU) LPS isolated from *E. coli* cultures. MoDCs were treated for 24 h to assess the immunostimulatory potential of the isolated LPS. Treatments were performed in duplicates and tested on at least three different donors. Culture supernatants from stimulated MoDCs were diluted 1/10 (or 1/50, if above standard curve range) and analysed for the presence of TNF-α using a commercial ELISA reagent set (Human TNF alpha uncoated ELISA, Life Technologies, Belgium) and a microplate reader (Biotek instruments, Germany).

For the second method, we verified our results using the bacterial load for normalizing LPS concentration values. Naturally, all faecal samples have a different bacterial load within the 150 mg of starting material that is used to isolate LPS. In order to assess if the differences that we observed before with equal volumes of LPS (see above) were due to the fact that some samples have a much lower bacterial load or if also the bacterial composition (and proportion of Gram-negative bacteria) plays a role in the immunostimulation, we normalized the amount of LPS used to stimulate MoDCs with the bacterial load. For example, the bacterial load was highest for VD neonate C105 (Supplementary Data 15; Supplementary Fig. 14), and the corresponding bacterial load was 51.5 μg DNA per 150 mg stool. Therefore, this load was divided by the load in each other sample to yield a normalization factor. To stimulate MoDCs with 100 EU of LPS, 2.51 μl C105 LPS was added. For other samples, the LPS load was calculated by multiplying 2.51 μl by the previously determined bacterial normalization factor. For the negative control, 2×10^5 MoDCs were incubated with 15 μl of LPS extraction blank, and for the positive control 2×10^5 MoDCs were treated with 100 EU LPS isolated from *E. coli* cultures. Treatments were performed on cells from four distinct MoDCs donors (2×10^5 MoDCs/donor), except for LPS isolated from C120, which was only sufficient to stimulate donor 4's MoDCs in duplicate. MoDCs and isolated LPS samples were incubated for 24 h. Culture supernatants from stimulated MoDCs were diluted twofold and analysed for the presence of seven cytokines (CXCL8/IL-8, IL-1β, IL-6, IL-10, IL-12p70, IL-18 and TNF-α) using a Human Premixed Multi-Analyte Kit (R&D Systems Europe; UK) and a MagPix multiplex reader (Luminex, Netherlands), according to the manufacturers' instructions (Supplementary Data 16). Statistical significance between the different cohorts was determined using the Wilcoxon rank-sum test.

Coomassie blue and silver staining of LPS extracts. On the basis of availability of sufficient extracted LPS material, 0.5 μg of extracted LPS from the stool samples of two VD neonates (C007 and C111) collected on day 3 after birth, were prepared with Laemmli sample buffer (Bio-Rad, Belgium), heated for 5 min at 95 °C and separated on 12 % Bis-Tris precast gel (Bio-Rad, Belgium) at 200 V for 45 min. As positive controls, 0.5 μg, 1 μg and 10 μg of commercially available LPS (*Escherichia coli* O55:B5, gel-filtration chromatography; Sigma-Aldrich, Belgium) and 10 μg of *E. coli* protein extract were used. A precast gel was loaded with the LPS samples and stained with Coomassie (Imperial protein stain, ThermoFisher, Belgium) to check for protein contaminations. Silver staining of the gel was performed using a corresponding kit (SilverQuest, ThermoFisher, Belgium) according to the manufacturer's instructions.

Ethidium bromide staining of LPS extracts. To check if LPS extracts were contaminated with immunostimulatory nucleic acids, 0.5 μg of extracted LPS from the stool samples of two VD neonates (C007 and C111), which presented highly concentrated LPS fractions that could be visualised on agarose gel, were prepared with DNA loading dye (ThermoFisher, Belgium) and loaded onto a 1% agarose gel. In addition, 0.5, 1 and 10 μg of commercially available LPS (*Escherichia coli* O55: B5, gel-filtration chromatography; Sigma-Aldrich, Belgium) were used to compare the purity of the LPS samples. As a positive control, 100 ng of *E. coli* DNA extract was used. The gel was stained with ethidium bromide, separated at 100 V for 50 min and analysed using a BioDocAnalyse system (Biometra, Germany). To check if nucleic acid contaminations could be identified in isolated LPS samples and would result in a TNF-α response, agarose bands were cut out (Supplementary Fig. 12) and purified using NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, France). As controls, bands of *E. coli* DNA and commercially available LPS (10 μg) were cut out and purified. In addition, a purification blank was generated. The purified DNA fractions were used to stimulate MoDCs following the same protocol as for the stimulation with extracted LPS.

HEK-Blue™ cell assay. In order to verify the purity of the extracted LPS fractions, HEK-Blue™ reporter cell lines overexpressing one of the receptors hTLR2, hTLR4, NOD1 or NOD2 (InvivoGen, France), were stimulated with LPS extracted from five selected neonatal faecal samples (three VD and two CSD), which presented sufficient amounts of extractable LPS. HEK-Blue™ TLR and NOD cells are designed to detect stimulants of the human receptors by induction of secreted embryonic alkaline phosphatase (SEAP). For all the cell lines, the levels of SEAP were determined with HEK-Blue™ Detection (InvivoGen, France), a cell culture medium that allows for real-time detection of SEAP.

While the hTLR4 receptor only recognizes LPS, hTLR2 recognizes peptidoglycan, lipoteichoic acid and lipoprotein from gram-positive bacteria, lipoarabinomannan from mycobacteria, and zymosan from the yeast cell wall, the receptor NOD1 binds to bacterial molecules containing the D-glutamyl-meso-diaminopimelic acid (iE-DAP) moiety and NOD2 recognizes bacterial molecules (peptidoglycans) and stimulates an immune reaction. HEK-Blue™ cells were grown and maintained in DMEM (4.5 g l^{-1} glucose, L-glutamine, Sigma-Aldrich, Belgium), supplemented with 10% foetal bovine serum (Thermo Fisher Scientific), 1% penicillin–streptomycin (Sigma-Aldrich, Belgium), 100 μg ml^{-1} Normocin (InvivoGen, France) and respective selective antibiotics according to the user's manual.

To monitor the activation of NF-κB, HEK-Blue™ cells were seeded according to the user's manual in HEK-Blue™ Detection medium (InvivoGen, France), in flat-bottom 96-well plates and stimulated for 22 h with LPS samples. We used two conditions: first, using the same concentration of LPS, where 1 μl of extracted LPS (0.01 ng μl^{-1}) was added per well, and second, using the same volume of LPS, where 7.5 μl extracted LPS was added to 10^5 HEK-Blue™ cells. To convert endotoxin activity (EU) into mass (ng), we considered that around 10 EU are equivalent to 1 ng endotoxin[79]. For positive controls, HEK-Blue™ NOD1 cells were stimulated with 1 μl TriDAP (10 μg μl^{-1}, InvivoGen, France), HEK-Blue™ NOD2 cells with 1 μl Murabutide (10 μg μl^{-1}, InvivoGen, France), HEK-Blue™ hTLR2 cells with 1 μl of Pam3CSK4 (1 μg μl^{-1}; InvivoGen, France) and HEK-Blue™ hTLR4 cells with 1 μl ultrapure LPS (5 μg μl^{-1}, source strain: ATCC 12014; CDC 5624-50 [NCTC 9701], InvivoGen, France). In addition, all cell lines were treated with 1 μl ultrapure LPS (5 μg μl^{-1}, InvivoGen, France) and 1 μl ultrapure LPS (0.01 ng μl^{-1}, InvivoGen, France) as well as with commercially available LPS (standard LPS; *Escherichia coli* O55:B5, gel-filtration chromatography; Sigma-Aldrich, Belgium): 1 μl of 5 μg μl^{-1} and 1 μl of 0.01 ng μl^{-1}. For the negative control, HEK-Blue™ cells were incubated with 1 μl of endotoxin-free H_2O (InvivoGen, France). All conditions were performed in duplicates and SEAP expression was monitored using a microplate reader at 655 nm (Biotek instruments, Germany) except for LPS isolated from C117 where only 7.5 μl extracted LPS/10^5 HEK-Blue™ cells was added to the cells and tested in duplicates.

Cytokine profiling of neonatal plasma samples. Plasma samples ($n = 31$) collected 3 or 5 days postpartum (13 VD, 13 CSD and five CSD + SGA; 28 samples collected at day 3, 3 samples collected at day 5 postpartum; Supplementary Data 1) were diluted twofold and analysed for 18 cytokines using a Human Premixed Multi-Analyte Kit (R&D Systems Europe) and a Bio-Plex analyser multiplex reader (Bio-Rad, Belgium), according to the manufacturers' instructions. The kit is able to detect CXCL8/IL-8, IL-1β, IL-6, IL-10, IL-12/23 p40, IFN-β, IL-15, IL-21, IL-5, Galectin-1, IFN-γ, IL-18, IL-27, Granzyme B, IL-13, IL-2, IL-4 and TNF-α. Of these cytokines, 11 were above the detection limit (CXCL8/IL-8, IL-6, IL-10, IL-15, IL-21, Galectin-1, IL-18, IL-13, IL-2, IL-4 and TNF-α; Supplementary Data 17).

Code availability. All custom scripts written for this study are available online at https://git-r3lab.uni.lu/Cosmic/Earliest.

Data availability

The pre-processed, non-human metagenomic sequencing data and the amplicon sequencing data generated during the current study are available from NCBI under bioproject accession number PRJNA379120. A reporting summary for this Article is available as a Supplementary Information file.

Received: 3 August 2017 Accepted: 13 November 2018
Published online: 30 November 2018

References

1. Betrán, A. P. et al. The increasing trend in caesarean section rates: global, regional and national estimates: 1990-2014. *PLoS ONE* **11**, e0148343 (2016).
2. Keag, O. E., Norman, J. E. & Stock, S. J. Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis. *PLoS Med.* **15**, e1002494 (2018).
3. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
4. Asnicar, F. et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* **2**, e00164–16 (2017).
5. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host. Microbe* **24**, 133–145.e5 (2018).
6. Yassour, M. et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host. Microbe* **24**, 146–154.e4 (2018).
7. Wampach, L. et al. Colonization and succession within the human gut microbiome by archaea, bacteria, and microeukaryotes during the first year of life. *Front. Microbiol.* **8**, 738 (2017).
8. Jakobsson, H. E. et al. Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut* **63**, 559–566 (2014).
9. Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M. & Hidalgo, G. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* 11971–11975 (2010).
10. Rutayisire, E., Huang, K., Liu, Y. & Tao, F. The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life: a systematic review. *BMC Gastroenterol.* **16**, 86 (2016).
11. Chu, D. M. et al. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).
12. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–544 (2016).
13. Cahenzli, J., Köller, Y., Wyss, M., Geuking, M. B. & McCoy, K. D. Intestinal microbial diversity during early-life colonization shapes long-term IgE levels. *Cell Host. Microbe* **14**, 559–570 (2013).
14. Arrieta, M.-C. et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, 307ra152 (2015).
15. Eggesbø, M., Botten, G., Stigum, H., Nafstad, P. & Magnus, P. Is delivery by cesarean section a risk factor for food allergy? *J. Allergy Clin. Immunol.* **112**, 420–426 (2003).
16. Sevelsted, A., Stokholm, J., Bonnelykke, K. & Bisgaard, H. Cesarean section and chronic immune disorders. *Pediatrics* **135**, e92–e98 (2015).
17. Huh, S. Y. et al. Delivery by caesarean section and risk of obesity in preschool age children: a prospective cohort study. *Arch. Dis. Child.* **97**, 610–616 (2012).
18. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host. Microbe* **17**, 690–703 (2015).
19. Montoya-Williams, D. et al. The neonatal microbiome and its partial role in mediating the association between birth by cesarean section and adverse pediatric outcomes. *Neonatology* **114**, 103–111 (2018).
20. Caughey, A. B. et al. Safe prevention of the primary cesarean delivery. *Am. J. Obstet. Gynecol.* **210**, 179–193 (2014).
21. Aagaard, K., Stewart, C. J. & Chu, D. Una destinatio, viae diversae: Does exposure to the vaginal microbiota confer health benefits to the infant, and does lack of exposure confer disease risk? *EMBO Rep.* **17**, 1679–1684 (2016).
22. Werner, E. F. et al. Mode of delivery and neonatal outcomes in preterm, small-for-gestational-age newborns. *Obstet. Gynecol.* **120**, 560–564 (2012).
23. Groer, M. W. et al. Development of the preterm infant gut microbiome: a research priority. *Microbiome* **2**, 38 (2014).
24. Perez-Muñoz, M. E., Arrieta, M.-C., Ramer-Tait, A. E. & Walter, J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, 48 (2017).
25. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, Da & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
26. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
27. Ramiro-Garcia, J. et al. NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes. *F1000Res.* **5**, 1791 (2016).
28. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
29. Laczny, C. C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.* **4**, 4516 (2014).
30. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
31. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res. Gr.* **216242**, 116 (2017).
32. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
34. Jun, S.-R., Robeson, M. S., Hauser, L. J., Schadt, C. W. & Gorin, A. A. PanFP: pangenome-based functional profiles for microbial communities. *BMC Res. Notes* **8**, 479 (2015).

35. Sunagawa, S. et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
36. De Wit, D. et al. Blood plasmacytoid dendritic cell responses to CpG oligodeoxynucleotides are impaired in human newborns. *Blood* **103**, 1030–1032 (2004).
37. Malamitsi-Puchner, A. et al. The influence of the mode of delivery on circulating cytokine concentrations in the perinatal period. *Early Hum. Dev.* **81**, 387–392 (2005).
38. Berdat, P. A. et al. Age-specific analysis of normal cytokine levels in healthy infants. *Clin. Chem. Lab. Med.* **41**, 1335–1339 (2003).
39. Milani, C. et al. Exploring vertical transmission of bifidobacteria from mother to child. *Appl. Environ. Microbiol.* **81**, 7078–7087 (2015).
40. Brooks, B. et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1814 (2017).
41. Jost, T., Lacroix, C., Braegger, C. P., Rochat, F. & Chassard, C. Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. *Environ. Microbiol.* **16**, 2891–2904 (2014).
42. Pannaraj, P. S. et al. Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatr.* **171**, 647 (2017).
43. Dominguez-Bello, M. G. et al. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat. Med.* **22**, 250–253 (2016).
44. Fulde, M. et al. Neonatal selection by Toll-like receptor 5 influences long-term gut microbiota composition. *Nature* **560**, 489–493 (2018).
45. Riedler, J. et al. Exposure to farming in early life and development of asthma and allergy: a cross-sectional survey. *Lancet* **358**, 1129–1133 (2001).
46. Sordillo, J. E. et al. Multiple microbial exposures in the home may protect against asthma or allergy in childhood. *Clin. Exp. Allergy* **40**, 902–910 (2010).
47. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
48. Garmory, H. S. & Titball, R. W. ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. *Infect. Immun.* **72**, 6757–6763 (2004).
49. Zhao, K., Liu, M. & Burgess, R. R. Adaptation in bacterial flagellar and motility systems: from regulon members to'foraging'-like behavior in E. coli. *Nucleic Acids Res.* **35**, 4441–4452 (2007).
50. Lu, Y. & Swartz, J. R. Functional properties of flagellin as a stimulator of innate immunity. *Sci. Rep.* **6**, 18379 (2016).
51. Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157**, 121–141 (2014).
52. Cullen, T. W. et al. Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* **347**, 170–175 (2015).
53. Guaraldi, F. & Salvatori, G. Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front. Cell. Infect. Microbiol.* **2**, 94 (2012).
54. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
56. Heintz-Buschart, A. et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2016).
57. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
58. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
59. Finn, R. D. et al. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
60. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
61. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
62. Wu, D., Jospin, G. & Eisen, J. A. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS ONE* **8**, e77033 (2013).
63. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
64. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).
65. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
66. Lee, W.-P. et al. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE* **9**, e90581 (2014). https://doi.org/10.1371/journal.pone.0090581. eCollection 2014.
67. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10**, 504–509 (2007).
68. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at http://arxiv.org/abs/1207.3907 (2012).
69. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
70. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
71. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).
72. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).
73. Davis, Jr., M. R. & Goldberg, J. B. Purification and visualization of lipopolysaccharide from gram-negative bacteria by hot aqueous-phenol extraction. *J Vis Exp.* **28** pii: 3916. https://doi.org/10.3791/3916 (2012).
74. Hirschfeld, M., Ma, Y., Weis, J. H., Vogel, S. N. & Weis, J. J. Cutting edge: repurification of lipopolysaccharide eliminates signaling through both human and murine toll-like receptor 2. *J. Immunol.* **165**, 618–622 (2000).
75. Bacchetti De, Gregoris,T., Aldred, N., Clare, A. S. & Burgess, J. G. Improvement of phylum- and class-specific primers for real-time PCR quantification of bacterial taxa. *J. Microbiol. Methods* **86**, 351–356 (2011).
76. Hermann-Bank, M. L., Skovgaard, K., Stockmarr, A., Larsen, N. & Mølbak, L. The gut microbiotassay: a high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. *BMC Genom.* **14**, 788 (2013).
77. Gold, M. C. et al. Human neonatal dendritic cells are competent in MHC class i antigen processing and presentation. *PLoS ONE* **2**, e957 (2007).
78. Goriely, S. et al. Deficient IL-12(p35) gene expression by dendritic cells derived from neonatal monocytes. *J. Immunol.* **166**, 2141–2146 (2001).
79. Schaumberger, S., Ladinig, A., Reisinger, N., Ritzmann, M. & Schatzmayr, G. Evaluation of the endotoxin binding efficiency of clay minerals using the Limulus Amebocyte lysate test: an in vitro study. *AMB Express* **4**, 1 (2014).
80. Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. Management of an academic HPC cluster: The UL experience. In: *2014 International Conference on High Performance Computing & Simulation (HPCS)* 959–967 (IEEE, Bologna, Italy, 2014).

Acknowledgements

In silico analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg[80]. We are grateful to all the parents and neonates who participated in the study. We thank the dedicated clinical staff and neonatologists of the paediatric clinic and gynaecologists at the CHL for participant recruitment and sample collection, especially Alain Noirhomme and all involved study nurses of the Clinical and Epidemiological Investigation Center (CIEC) who performed sample and data collection at the CHL and at home as well as the scientific staff of the IBBL for sample storage. We are thankful for the assistance of Audrey Frachet and Janine Habier (LCSB) for laboratory support, Jochen Schneider (LCSB) for discussions, and Lola Kourouma for metadata entry and cross-checking of medical data. Lisa Morgenstern from GATC Biotech AG in Constance is thanked for her work and discussion on random shotgun sequencing and Wouter Coppieters and Latifa Karim from CART-GIGA for their work on 16S rRNA gene amplicon sequencing. We thank the MO BIO technical support team for recommendations on extraction kit optimizations. The present work was partially financed by the Fondation André et Henriette Losch. It was further supported by an ATTRACT programme grant (ATTRACT/A09/03) and CORE programme grants (CORE/15/BM/104040 and CORE/C15/SR/10404839) to P.W. and (CORE Junior/14/BM/8066232) to J.V.F., Aide à la Formation Recherche grants to L.W. (AFR PHD-2013-5824125) and S.N. (AFR PHD-2014-1/7934898), all funded by the Luxembourg National Research Fund (FNR). A.K. was funded by the University of Luxembourg (ImMicroDyn1). Sample collection, processing and storage were co-funded by the Integrated BioBank of Luxembourg under the Personalised Medicine Consortium Diabetes programme.

Author contributions

L.W. carried out the biomolecular extractions, sequence annotation, did the comparative analyses of metagenomic and 16S rRNA gene amplicon sequencing data. A.H-B. coordinated the metagenomic measurements, carried out the strain-level analysis and performed genome reconstructions. Both L.W. and A.H-B. curated the metagenomic data from artefactual sequences, performed binning, annotated called genes and interpreted the data. J.V.F. was involved in all the immunological assays, participated in the study design and data interpretation. J.R-G. carried out the processing of the 16S rRNA gene amplicon sequencing data and the subsequent multivariate analysis and functional predictions. J.H. carried out the LPS isolations, purity controls including gel staining and HEK-Blue™ cell assays and participated in all immunological assays. M.H. linked the obtained genome reconstructions in distinct samples by essential marker gene sequence homology, and S.N. carried out the sequence assembly and gene prediction. A.K. participated in biomolecular extractions and the bioinformatic assessment of metagenomic data. A.H.H. participated in the design of the study and storage of samples. C.d.B., L.B. and J.B. enrolled and consulted mothers and participated in sample and data collection. R.H. performed part of the metagenomic sequencing, P.M. carried out the functional profiling, and A.F.A. and C.S. performed the calculations of the fixation indices and intra-population diversities and participated in data interpretation. C.d.B. and P.W. conceived the study, participated in its design, performed data interpretation and coordinated the study. L.W., A.H-B., J.V.F. and P.W. wrote the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07631-x.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018