UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2021-040
The Faculty of Sciences, Technology and Medicine

DISSERTATION

Defence held on 09/07/2021 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Susana MARTÍNEZ ARBAS

Born on 30 April 1986 in Madrid, (Spain)

INTEGRATED MULTI-OMIC ANALYSES OF MOBILE GENETIC ELEMENTS WITHIN A MIXED MICROBIAL COMMUNITY

Dissertation defence committee

Dr Paul Wilmes, dissertation supervisor
Professor, Université du Luxembourg

Dr Alexander Skupin, Chairman
Associate professor, Université du Luxembourg

Dr Patrick May, Vice Chairman
Université du Luxembourg

Dr Alexander Probst
Professor, University of Duisburg-Essen, Germany

Dr Laura Hug
Assistant professor, University of Waterloo, Canada

Integrated multi-omic analyses of the dynamics of mobile genetic elements within a mixed microbial community

A dissertation

by

Susana Martínez Arbas

Completed in the

Systems Ecology Group, Luxembourg Centre for Systems Biomedicine

To obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

Dissertation Defence Committee:

Supervisor:	Prof. Dr. Paul Wilmes
Comittee members:	Ass. Prof. Dr. Alexander Skupin
	Dr. Patrick May
	Prof. Dr. Alexander Probst
	Ass. Prof. Dr. Laura Hug

2021

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.

Susana Martínez Arbas,
Esch-sur-Alzette, Luxembourg
July 21, 2021

Acknowledgements

I would like to thank my supervisor Paul Wilmes for giving me the opportunity to dive into the microbiome world and for his guidance during the PhD. To Patrick May for his unconditional support and will to brainstorm at any time. To Alex Skupin for his support and for being part of my CET. To Karoline Faust for hosting me in her team for a little while. To Laura Hug and Alex Probst for being part of my thesis committee.
I would like to thank the Critical Transitions in Complex Systems (CriTiCS) Doctoral Training Unit (DTU), which has supported this PhD.
My admiration and gratitude goes to former LAOers; Emilie, Shaman, and Malte, for their will to always help and discuss the data and the plethora of interpretations, and a lot of technical and eventually emotional support.
I would like to thank former and current members of the group that have made my journey unforgettable for a bunch of different reasons, Laura, Susheel, Francesco, Javi, Loulou, Antoine, Rashi, Pedro, Ben, Valentina, Camille, Janine, Cedric, Oskar, Velma, Polina, Joanna, Kacy, Linda, Deepti.
Last but not least, I would like to thank my family and friends for their constant encouragement and for believing in me.

Abstract

Microbial communities are ubiquitous, complex and dynamic systems that constantly adapt to changing environmental conditions, while playing important roles in natural environments, human health and biotechnological processes. Invasive mobile genetic elements (iMGE) are considered as important biotic components of microbial communities, in particular (bacterio)-phages and plasmids are some of the most abundant and diverse biological entities, which may influence community structure and dynamics. Microbial populations within naturally occurring communities are constantly interacting with each other. Ecological interactions between those populations can be generally classified as competitive and cooperative relationships. To date, extensive studies on biotic interactions, i.e. relationships between microbial hosts with iMGEs and between microbial populations, have been somewhat limited, thus restricting our understanding of microbial community dynamics. Fortunately, high-throughput multi-omics derived from microbiomes, i.e. metagenomics and metatranscriptomics, enables access to both functional -potential and -expression information of those biotic components. Combining longitudinal multi-omics data with mathematical frameworks allows us to model microbial community interactions and dynamics, unlike ever before. Here, I present a longitudinal integrated multi-omics analysis of biotic components within foaming activated sludge, spanning ~1.5 years to unravel i) iMGE-host dynamics and ii) ecological interactome. In the first part of this work, empirical host-iMGE CRISPR-based links in combination with mathematical modelling highlighted the importance of plasmids, relative to phages, in shaping community structure, while also showing that plasmids vastly outnumbered, and were more targeted via CRISPR-Cas systems, compared to their phage counterparts. In the second part of this work, mathematical modelling is used to provide ecological contexts for the relationships between microbial community members. In general, we observed a dynamic interactome, with higher cooperative interactions, despite these populations encoding highly similar functional potential. In summary, this work demonstrates the potential of longitudinal multi-omics in expanding our understanding of microbial community dynamics, which could be expanded to other microbial ecosystems and potentially lead to applications in human health and biotechnological processes.

Scientific output

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

Publications in peer-review journals

- **Susana Martínez Arbas**, Susheel Bhanu Busi, Pedro Queirós, Laura de Nies, Malte Herold, Patrick May, Paul Wilmes, Emilie E. L. Muller, Shaman Narayanasamy (2021). Challenges, strategies and perspectives for reference-independent longitudinal multi-omic microbiome studies.
  Frontiers in Genetics **12**:666244. DOI: https://doi.org/10.3389/fgene.2021.666244. [**Appendix A.1**]

- **Susana Martínez Arbas**[†], Shaman Narayanasamy[†], Malte Herold, Laura A. Lebrun, Michael R. Hoopmann, Sujun Li, Tony J. Lam, Benoît J. Kunath, Nathan D. Hicks, Cindy M. Liu, Lance B. Price, Cedric C. Laczny , John D. Gillece, James M. Schupp, Paul S. Keim, Robert L. Moritz, Karoline Faust, Haixu Tang, Yuzhen Ye, Alexander Skupin, Patrick May, Emilie E. L. Muller and Paul Wilmes (2021) Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics.
  Nature Microbiology, **6**:123–135. DOI: https://doi.org/10.1038/s41564-020-00794-8. [**Appendix A.2**]

- Malte Herold, **Susana Martínez Arba**s, Shaman Narayanasamy, Abdul R. Sheik, Luise A. K. Kleine-Borgmann, Laura A. Lebrun, Benoît J. Kunath, Hugo Roume, Irina Bessarab,

[†]Co-first author

Rohan B. H. Williams, John D. Gillece, James M. Schupp, Paul S. Keim , Christian Jäger, Michael R. Hoopmann, Robert L. Moritz, Yuzhen Ye, Sujun Li, Haixu Tang, Anna Heintz-Buschart, Patrick May, Emilie E. L. Muller, Cedric C. Laczny and Paul Wilmes 2020. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance.
Nature Communications **11**:5281. DOI: https://doi.org/10.1038/s41467-020-19006-2. [**Appendix A.3**]

- Emilie E.L. Muller, Karoline Faust, Stefanie Widder, Malte Herold, **Susana Martínez Arbas**, Paul Wilmes (2018). Using metabolic networks to resolve ecological properties of microbiomes.
Current Opinion in Systems Biology **8**: 73-80. DOI: https://doi.org/10.1016/j.coisb.2017.12.004. [**Appendix A.4**]

Manuscripts in preparation

- **Susana Martínez Arbas** *et al.* Time-series integrated multi-omic analyses for ecological interactome inference of microbiomes [**Chapter 3**]

Oral presentations in scientific conferences, symposia and workshops

- A multi omic view of phage-host dynamics within a mixed microbial community (2017). *Life sciences PhD Days*, Belval, Luxembourg.

- Understanding mobile genetic elements as biotic drivers of critical transitions in microbial communities (2018). *Critical Transitions workshop*, Belval, Luxembourg.

- A multi-omic view of invasive genetic elements and their linked prokaryotic population dynamics within a mixed microbial community (2018). *Microbiology Day congress*, Belval, Luxembourg.

- A multi-omic view of invasive genetic elements and their linked prokaryotic population dynamics within a mixed microbial community (2018). *International Society for Microbial Ecology conference (ISME17)*, Leipzig, Germany.

Poster presentations in scientific conferences, symposia and workshops

- Comparative integrated-omic analyses of phage-host interactions within natural and engineered microbial communities (2017). *International Society for Computational Biology student symposium*, Prague, Czech Republic.

- Understanding mobile genetic elements as biotic drivers of critical transitions in microbial communities (2018). *Critical Transitions workshop*, Belval, Luxembourg.

- A multi-omic view of invasive genetic elements and their linked prokaryotic population dynamics within a mixed microbial community (2018). *European Conference on Computational Biology (ECCB)*, Athens, Greece.

- Critical transitions in microbial communities: mobile genetic elements as drivers of the microbial community dynamics within activated sludge of wastewater treatment (2018). *Life sciences PhD Days*, Belval, Luxembourg.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Part of this chapter was adapted and modified from the following first-author peer-review publication:

Susana Martínez Arbas, Susheel Bhanu Busi, Pedro Queiros, Laura De Nies, Malte Herold, Patrick May, Paul Wilmes, Emilie EL Muller, Shaman Narayanasamy

2021

Challenges, strategies and perspectives for reference-independent longitudinal multi-omic microbiome studies

Frontiers in Genetics, **12:666244**. **Appendix A.1**

1.1 Microbiomes

Microbiomes are ubiquitous, complex, heterogeneous and dynamic. A widely accepted definition of microbiome is that it is a collection of microorganisms living at the same place at the same time [Whipps et al., 1988; Berg et al., 2020]. This definition is context-specific and can be assumed to include the vast diversity of microbiomes per se, influenced by their environmental conditions. Microbiomes are made up of bacteria, archaea, fungi and (micro)eukaryotes, mobile genetic elements (MGEs) and relic DNA, i.e. extracellular DNA released into the environment when microorganisms die [Lennon et al., 2018]. In the context of the present work, microbiomes and microbial communities are referred interchangeably.

Microbiomes in natural environments can be used as quality/health indicators. For example, the microbiome composition can serve as indicator of e.g. seagrass health [Martin et al., 2020], eutrophication of coral reef ecosystems [Glasl et al., 2019], and soil quality [Schloter et al., 2018]. Microbiomes in biotechnological processes are key to efficient food production, e.g. cheese [Walsh et al., 2020] and wine [Liu et al., 2019], or wastewater treatment [Wu et al., 2019]. Microbiomes in biomedicine mainly involve the understanding of dysbiosis, i.e. microbiome composition that deviates from a "healthy" microbiome, i.e. enriched presence of pathogenic microorganisms. Dysbiosis is commonly observed in various diseases, such as diabetes[Musso et al., 2011], irritable bowel syndrome [Rodiño-Janeiro et al., 2018] and Parkinson's disease [Elfil et al., 2020], while differences in microbiome composition have been observed in naturally-born neonates versus those born via C-section [Wampach et al., 2017, 2018; Busi et al., 2021]. On the other hand, probiotics and prebiotics are commonly prescribed or recommended, through diet and/or supplementation to improve gut health in certain dysbiotic cases [Macfarlane and Cummings, 1999; Turnbaugh et al., 2007].

Microbiomes are highly dynamic systems that are constantly adapting to their environment [Gerber, 2014; Gonze et al., 2018]. For example, microbiome composition within the human gastrointestinal tract is known to vary between individuals and further compounded by experiencing daily fluctuations [Voigt et al., 2016]. Such microbiome dynamics are driven by both abiotic factors, i.e environmental conditions that include temperature, pH, or nutrient availability, and biotic factors including community composition and interactions between community members [Machineni, 2020].

Considering the important role of microbiomes in various aspects of our lives, microbiome research has been rapidly expanding over the last decade. Additionally, these highly dy-

namic and complex systems require extensive interdisciplinary research initiatives, spanning the fields of microbiology, chemistry, mathematics, engineering and medicine, among others. In this chapter, I cover various topics involving the study of microbiome dynamics.

1.1.1 Mobile genetic elements

Mobile genetic elements (MGEs) are genetic material that move within and between genomes, i.e. intracellular and extracellular mobility, respectively. MGEs are key components that drive evolution. The combined presence of MGEs in a genome is known as the mobilome, and it is made up of bacteriophages, plasmids and transposons. MGEs that move between cells can also be considered as invasive MGEs (iMGEs), e.g. bacteriophages and plasmids. iMGEs usually carry genes encoding functions that allow their transfer to- and replication within- other cells [Leplae et al., 2004]. iMGEs immediate effects on the bacteria are typically detrimental, resulting in either cell destruction or suicide. However, iMGE effects can also enhance the survival of a bacteria via acquisition of antimicrobial resistance genes [Koonin, 2016; Rios Miguel et al., 2020]. My work hereafter focuses on the influence of iMGEs, namely plasmids and phages, on community dynamics.

Bacteriophages (or phages), are viruses that infect bacteria. They are considered to be the most abundant and diverse biological entities on earth [Bergh et al., 1989; Weinbauer, 2004]. Genomes of phages are made up of RNA or DNA, with either; linear or circular configuration; single- or double- stranded, and vary widely in size [Hatfull and Hendrix, 2011; Hay and Lithgow, 2019]. Lytic phages adsorb into host cells and utilize host cellular machinery for self-replication, which eventually leads to the destruction of the host cell. Certain phages may also integrate their genetic material into the host genome, i.e. lysogenic phages. These phages either remain dormant or influence bacterial gene expression [Zrelovs et al., 2020]. However, under stressful situations, lysogenic phages can trigger their replication and become lytic [Nanda et al., 2015]. The diverse genetic material of phages are a key factor that provides them with a wide range of regulation and replication mechanisms, often adverse for the host. However, those replication mechanisms are capable of transferring genes from one bacteria to another via transduction (**Figure 1.1**).

Plasmids are usually highly abundant double stranded, mostly circular DNA, which greatly vary in their size. They are found within cells, but are independent from the bacterial chromosome, due to their ability to replicate as autonomous genetic elements [del Solar et al., 1998]. Plasmids may carry genes that either complement the bacterial metabolism

[Lee et al., 1992; Rios Miguel et al., 2020] or enhance bacterial survival, thus justifying the cost for a given bacteria to carry a plasmid [San Millan and MacLean, 2017]. Plasmids are also capable of transferring genetic material from one bacteria to another via conjugation (**Figure 1.1**). In general, genetic material transfer mechanisms involving MGEs are known as "horizontal gene transfer", which covers transduction, transformation and conjugation, and are key factors for evolution [Boto, 2010; Hall et al., 2020].

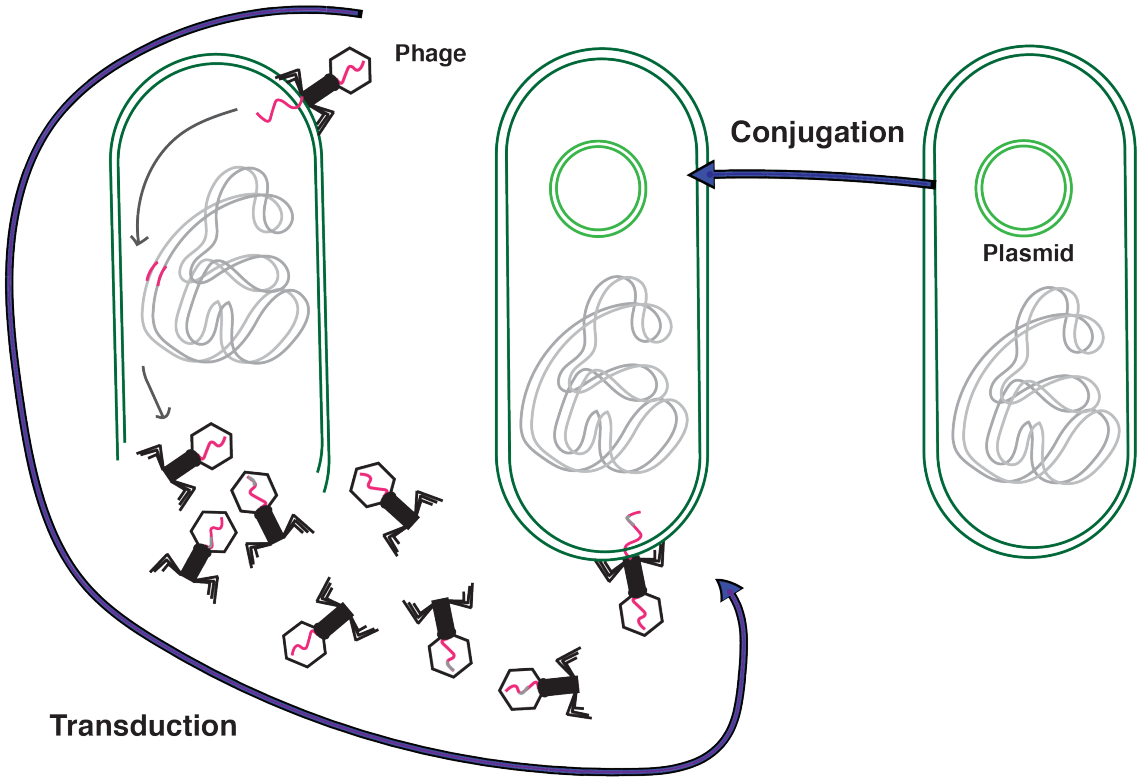


Figure 1.1: Conjugation and transduction as examples of horizontal gene transfer mechanisms.. On the left, a lytic phage infects a cell, during its life cycle can incorporate its genomic material to the bacterial chromosome, and when new bacteriophage material is being replicated and packed within the viral capsids, bacterial genomic material can be also included. Therefore, when new phages infect new bacteria, they can carry bacterial genes from previous infections. On the right, plasmids can be transferred from one bacteria to another through conjugation.

1.1.2 Prokaryotic defense mechanisms

Due to the negative effects of iMGEs on prokaryotic cells, they have evolved to develop several defence mechanisms against the activity of iMGEs (**Figure 1.2**). These include, but are not limited to, surface modification, superinfection exclusion, restriction-modification

systems, prokaryotic argonaute enzymes (pAgos), and CRISPR-Cas systems [van Houte et al., 2016].

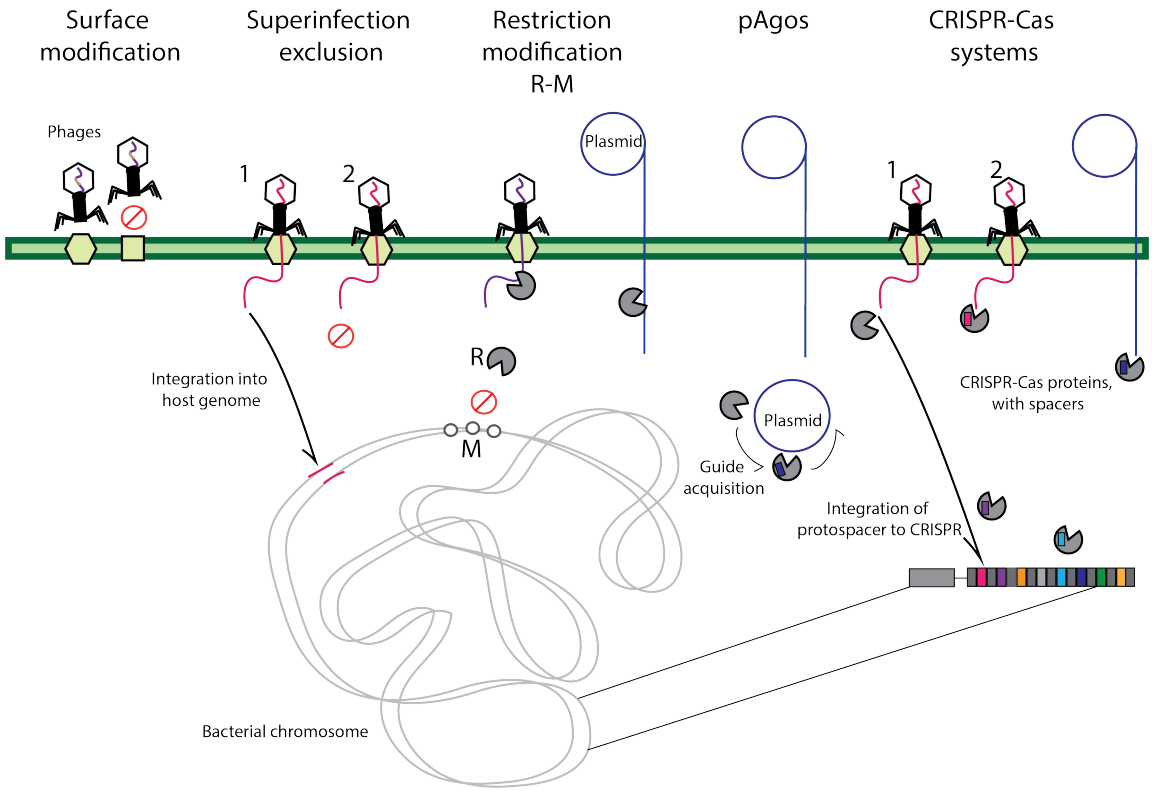


Figure 1.2: Bacterial immune mechanisms against iMGEs. Figure adjusted from [van Houte et al., 2016]. Overview of several prokaryotic defense mechanisms against phages, plasmids, or both phages and plasmids.

Surface modification consists of the complete loss, mutation, blocking or masking of the receptors that bacteriophages utilize to enter the cells [Bertozzi Silva et al., 2016]. Specifically, the surface modification of host receptors aims to inhibit the initial step of phage infection, which is the binding to specific surface proteins or cell wall components of the host cell. For example, in *Pseudomonas aeruginosa*, the receptor of pilus-specific phages, type IV pili, can be modified to prevent phage binding [Harvey et al., 2018].
Superinfection exclusion consists of the blockage of phage's genetic material injection to the host, or the phage replication. This mechanism is, indeed, encoded by already infecting phages, commonly in the form of prophages [Bondy-Denomy et al., 2016]. For example, the ϕ80 prophage of *Escherichia coli* encodes the Cor protein, that inactivates the cell surface receptor FhuA and inhibits the infection of related phages [Uc-Mass et al., 2004].
Restriction modification (R-M) systems protect the host against iMGEs by the cleavage of

foreign unmethylated DNA, while protecting native DNA by methylating specific sites or motifs. R-M systems are widely spread within bacteria and archaea, and there are several known types [Sneppen et al., 2015].

Prokaryotic argonaute enzyme (pAgo) proteins bind small nucleic acids and use them for sequence-specific cleavage, suppressing invasion and activity of iMGEs [Makarova et al., 2009; Ryazansky et al., 2018]. For example, the expression of these proteins in *E. coli* leads to plasmid degradation [Olovnikov et al., 2013].

CRISPR-Cas systems are a sequence-based recognition immune system within prokaryotes, especially prevalent within archaea (~90%) [Sorek et al., 2008]. CRISPR stands for clustered regularly inter-spaced short palindromic repeats [Jansen et al., 2002]. Briefly, a short sequence of the iMGEs, i.e. protospacer, is extracted and then integrated within the CRISPR locus of the prokaryotic chromosome. Upon integration, this short sequence is known as a spacer, and is flanked by repetitive sequences [Mojica et al., 2005]. There is a set of CRISPR associated genes, known as *cas* genes, usually near the CRISPR locus [Jansen et al., 2002]. Through the expression of the CRISPR locus and the activity of CRISPR associated proteins (Cas), the cell will recognize future infections and inactivate the iMGE (**Figure 1.3**). A large part of this work revolves around using CRISPR information, namely the unique characteristic of CRISPR loci and the complementarity between spacers and protospacers.

Last but not least, in order to overcome those prokaryotic defense mechanisms, iMGEs evolved ways to ensure their own survival [Samson et al., 2013]. For example, phages can escape the CRISPR-Cas recognition by mutating their protospacers, or by carrying anti-CRISPR cellular machinery [Faure et al., 2019], which are proteins that inhibit CRISPR-Cas systems.

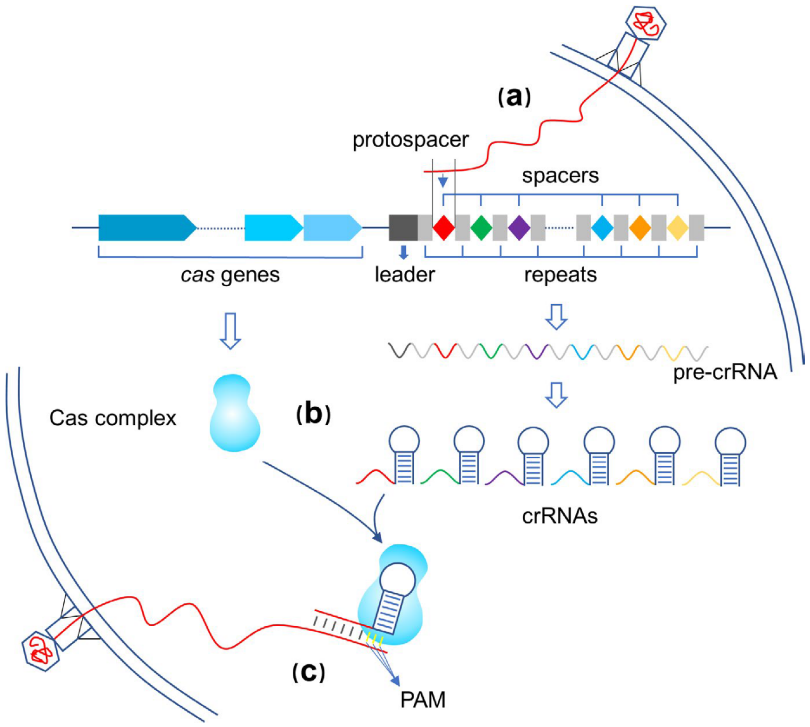


Figure 1.3: CRISPR-Cas system. Figure from [Gong et al., 2020]. The CRISPR locus is structured as spacer sequences (coloured diamonds) originated from protospacers, flanked by small repetitive sequences (grey rectangles) specific to the prokaryotic chromosome. Near the CRISPR locus, the *cas* genes encode CRISPR associated proteins that are involved in the **a**) integration, **b-c**) recognition and cleavage of the MGE sequences.

1.2 Model microbial community

Model organisms are typically used within the field of biology, including microbiome-related studies. Notable examples of model organisms are mouse (*Mus musculus*), yeast (*Saccharomyces cerevisiae*), fruit fly (*Drosophila melanogaster*), nematode worm (*Caenorhabditis elegans*), western clawed frog (*Xenopus tropicalis*), or zebrafish (*Danio rerio*). These organisms provide stable platforms to test hypotheses on fundamental insights into biolog-ical mechanisms which may result in beneficial biomedical, biotechnological or environmental outcomes. Similarly, the study of microbiomes and microbial communities could benefit from well-defined model systems. However, the high complexity, plasticity and challenging manipulability of microbiomes remain barriers to recapitulating microbiomes in the laboratory. In recent years, there have been important advances in the development of devices/platforms that simulate the conditions of natural microbiome systems, e.g. human gut [Shah et al., 2016] and bioreactors [Connelly et al., 2017].

The present work, hereby focuses on a model microbial community within the activated

sludge of a biological wastewater treatment plant (BWWTP).

1.2.1 Wastewater treatment

Communal wastewater is collected from residences, institutions, commercial and industrial establishments, ground water, storm-water and surface water. Wastewater treatment is of great importance for the protection of natural environments and human health, representing one of the most widely used biotechnological processes on our planet [Sheik et al., 2014].

Conventional wastewater treatment consists of a combination of physical, chemical, and biological processes, consisting of a preliminary treatment, a primary treatment, and a secondary treatment (**Figure 1.4**). The preliminary treatment removes large solids found in wastewater. The primary treatment removes organic and inorganic solids through physical sedimentation and flotation, with the use of clarifiers or settling tanks. The secondary treatment, or biological treatment, involves the so-called activated sludge process which removes or reduces the remaining organic material and suspended solids, by utilizing naturally occurring microorganisms in a controlled environment [Sonune and Ghate, 2004]. Specifically, during the activated sludge process, microbial populations remove dissolved organic matter through assimilation and oxidation. Then, the suspended biomass is separated from the treated wastewater by an aeration phase where the majority of the activated sludge is recycled [Sheik et al., 2014].

1.2.2 Activated sludge foaming

The process of activated sludge foaming takes place in anoxic tanks, in the form of foaming sludge islets. Microbial communities from activated sludge foam represent good models of microbial ecology due to the relatively homogeneous environment of BWWTPs, with well-defined physico-chemical boundaries, such as temperature, pH, oxygen and nutrient concentration [Narayanasamy et al., 2015]. They exhibit medial species richness, while being highly dynamic [Zhang et al., 2012; Yang et al., 2020]. The high biomass samples from this system allow for the extraction of relevant biomolecules. Moreover, foaming sludge represents a convenient and virtually unlimited source of spatially and temporally resolved samples. The continuous tracking of physico-chemical parameters, as standard procedure of the BWWTPs, allow for detailed complementary physicochemical information. The combination of the aforementioned hallmark characteristics enable extensive sample collections for multi-omic studies, alongside important measurements of environmental factors and metadata [Daims et al., 2006].

The exploration of the microorganisms involved in the activated sludge process could lead to potential for biofuel and bioplastic production [Sheik et al., 2014]. Foaming islets of activated sludge are partially composed of filamentous and lipid-accumulating microorganisms, and they are particularly suitable for energy recovery via transesterification to produce biofuel [Muller et al., 2014b].

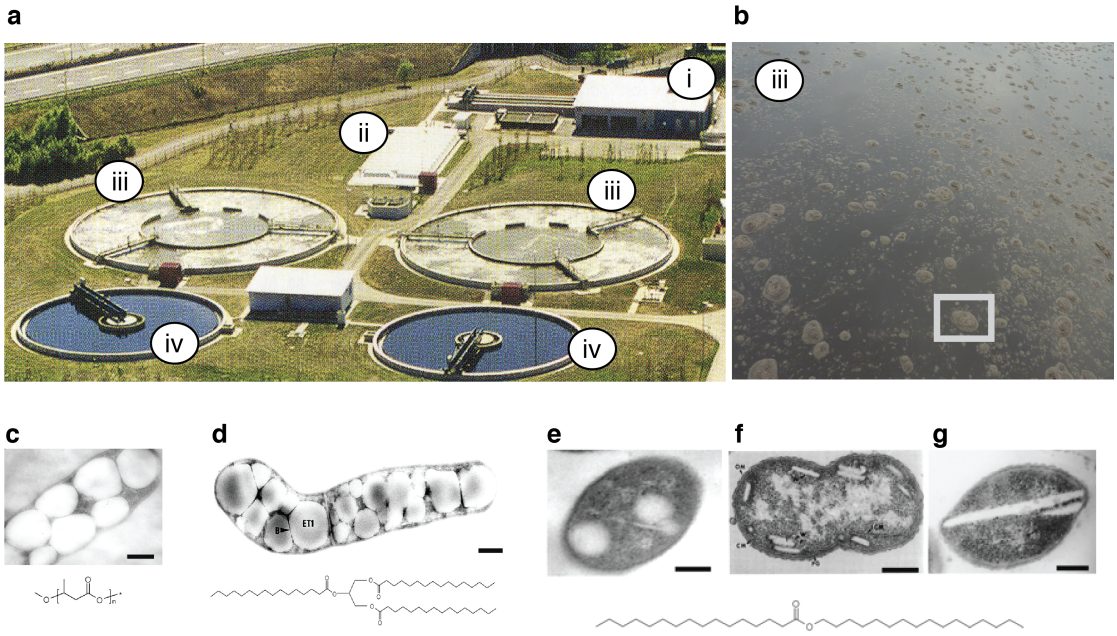


Figure 1.4: Biological wastewater treatment plant of Schifflange, Luxembourg. **a**, Overview of the infrastructure where the **i)** preliminary treatment, **ii)** primary treatment, **iii)** secondary treatment, and **iv)** the removal of remaining foam take place. **b**, Surface of the anoxic tank where the activated sludge takes place. **c-g** photos represent examples of lipid accumulating bacteria from [Wältermann and Steinbüchel, 2005]. **c**, Cell of *Ralstonia eutropha* H16 accumulating poly(3-hydroxybutyrate) (PHB) inclusions [Pötter et al., 2002]. **d**, Cell of *Rhodococcus opacus* PD630 from late stationary growth phase accumulating large amounts of triacylglycerols (TAG) inclusions [Alvarez et al., 1996]. **e**, Cell of *Acinetobacter calcoaceticus* ADP1 with three spherical wax esters (WE) inclusions. **f**, *Acinetobacter* sp. strain HO1-N accumulating small rectangular WE inclusions [Singer et al., 1985]. **g**, *Acinetobacter* sp. strain M1 accumulating large, disc-like WE inclusions [Ishige et al., 2002]. Abbreviations of the photos: B, boundary layer; CM, plasma membrane; ET, electron-transparent TAG inclusion; ICM, intracellular plasma membrane; OM, outer membrane; PHA, PHA inclusion; PG, peptidoglycan; W, wax ester inclusion. Bars in **c-g**, $0.2\,\mu m$.

1.3 Systems ecology of microbial community dynamics

The study of microbial community dynamics requires interdisciplinary and complementary approaches. Specifically, microbiomes are often analysed in a System biology approach, being communities and their emergent properties studied as a whole, rather than

the components alone, especially considering that a large majority of naturally occurring microbes remain either uncultured or unculturable within a laboratory setting [Staley and Konopka, 1985; Amann et al., 1995; Stewart, 2012]. As part of Systems Biology, Systems Ecology includes the study of dynamic processes of microbiomes and their interactions within ecosystems. These studies might involve interdisciplinary and integrative frameworks that include systematic sampling and multi-omic measurements, bioinformatics data processing and analysis, modelling and predictions, and ideally experimental validations to understand and control microbial ecosystems [Muller et al., 2013].
To resolve complex dynamics in microbiomes, longitudinal or time-series analyses are crucial. These dynamics stem from constant adaptation of a given community towards fluctuations of abiotic and biotic factors. However, the fates of these microbial consortia under perturbations are often not understood nor predictable [Muller, 2019]. Thus, longitudinal studies are capable of resolving the dynamics in microbiomes at the levels of composition, structure, and function, offering valuable insights into temporal trends, especially when used in tandem with environmental data [Law et al., 2016], which ultimately reveals how microbial consortia adapt to biotic and/or abiotic perturbations.

1.3.1 Sampling and biomolecular extraction

Sampling for longitudinal, systematic multi-omic studies requires the consideration of sampling frequency, availability of biomass, and metadata including complementary measurements [Gerber, 2014; Kumar et al., 2014; Eisenhofer et al., 2019], all of which affect downstream analyses. Since most microbial communities are heterogeneous, biomolecular extraction from a single sample is ideally required, over multiple extractions from subsamples [Peña-Llopis and Brugarolas, 2013; Roume et al., 2013].
Additionally, batch effects are often overlooked in omic studies [Duvallet et al., 2017], but can be minimized at any stage; i) by including randomisation, sample tracking, and extensive documentation during sample processing, handling, biomolecular extractions and high-throughput measurements [Leek et al., 2010], and ii) by adding specific downstream analytical and computational methods [Gibbons et al., 2018]. In longitudinal studies, sample randomisation can be implemented at several ensuing steps and could help discriminate between batch effects and temporal variation.

1.3.2 Systematic high-throughput measurements

Sequencing approaches are key to study microbial diversity of microbiomes. Initially, the 16S rRNA gene amplicon sequencing has made possible the characterization and study of

microbiomes, overcoming the need to cultivate every microorganism, individually. The 16S rRNA marker gene has a conserved region with a hypervariable sequence, which allows the differentiation between organisms. 16S rRNA sequencing, or amplicon sequencing, has been widely used to identify new taxa and characterize microbiomes, providing great insights for the development of many modelling methods. However, the 16S rRNA sequencing i) is unable to differentiate between strains, ii) leaves mobile genetic elements undetected, and iii) lacks a genomic context that provides insights about the function. To fill those gaps, untargeted sequencing of all the genomes from microbiome samples was developed, i.e. metagenomics (MG), which provides a broader picture of community composition and structure, while enabling the study of functional potential. More recently, the generation of high-throughput "function-omes", such as metatranscriptomics (MT), metaproteomics (MP), and (meta)metabolomics (MM) allow for the study of expressed community functions and exchange of substrates [Muller et al., 2018].

Leveraging the power of the entire high-throughput meta-omic spectrum, i.e. MG, MT, MP and MM, enables fine-grained observations of microbial community ecology. For example, the most abundant populations at the MG level may not necessarily be contributing the most to community function, assessed with additional functional omics, MT, MP, and MM [Muller et al., 2014a; Kaysen et al., 2017]. In addition, the integration of multiple omic readouts, e.g. the co-assembly of MG and MT sequencing data [Narayanasamy et al., 2016], could potentially decrease the biases and limitations stemming from a single omic readout [Narayanasamy et al., 2015].

1.3.3 Bioinformatic analyses

Bioinformatic analyses of microbiomes can be broadly separated into two approaches, reference-dependent and reference-independent. Reference-dependent methods rely on information or databases available *a priori*, especially if the components/features of a given microbial ecosystem are well characterized, e.g. those within the human gastrointestinal tract. Here, bioinformatic analysis is addressed from the context of a reference-independent approach, centered around *de novo* assemblies of MG and MT data. This is due to the asymmetric advantages and opportunities compared to reference-dependent approaches, namely the discovery of novel microbial taxa and/or functionalities.

Prior to downstream analyses, sequencing reads require preprocessing to remove artificial sequences, e.g. sequencing adapter, spike-ins [Bolger et al., 2014], or contaminants, e.g. human sequences in a human microbiome sample [Narayanasamy et al.,

2016], rRNA in metatranscriptomes [Kopylova et al., 2012], and laboratory contaminants [Heintz-Buschart et al., 2016]. *De novo* assembly can then be carried out to generate longer contiguous sequences (or contigs), which provide the basis for downstream analyses. Then, an estimation of the microorganisms that are present in a given sample can be assessed by reconstructing their genomes, into so-called metagenome-assembled genomes (MAGs), through binning. There are a multitude of binning tools that typically use compositional and coverage features of the contigs [Alneberg et al., 2014; Wu et al., 2014; Heintz-Buschart et al., 2016; Graham et al., 2017; Herath et al., 2017; Kang et al., 2015; Qian and Comin, 2019]. In addition, there are now tools to automate the process of refining the bins to select the highest-quality MAGs [Sieber et al., 2018; Uritskiy et al., 2018]. Importantly, these methods also enable ensemble binning approaches, balancing the strengths and weaknesses of different binning methods [Chen et al., 2020; Yue et al., 2020]. Taxonomic classification can be assessed at the contig or at the MAG level [Wu and Eisen, 2008; Wood and Salzberg, 2014; Kim et al., 2016], which can be used to further refine the MAGs and improve their quality [Sieber et al., 2018]. Given the integral roles of iMGEs within microbiomes (**Section 1.1.1**), identification/prediction of these features should be emphasised as part of microbiome analyses. For that purpose, there is a plethora of tools available to predict bacteriophages and plasmids from MG data [Carr et al., 2021]. Most of these tools rely on sequence similarity to known iMGE sequences, which results in a limited amount of predictions. However, the field is con-tinuously improving with novel state-of-the-art methods, e.g. a deep-learning method for alignment-free identification of phage sequences [Auslander et al., 2020]. In parallel, the assignment of functions to predicted open reading frames (ORFs) is per-formed via gene annotation [Seemann, 2014]. For that, there are specialized gene annota-tion tools for features of interest, such as antibiotic resistance and virulence genes [de Nies et al., 2020], or CRISPR type classification [Couvin et al., 2018]. Furthermore, the prediction of CRISPR loci [Zhang et al., 2020] from MG and MT data provides insights about microbial interactions with iMGEs, due to the spacer-protospacer complementarity [Edwards et al., 2016; Shmakov et al., 2020]. Finally, microbiomes may be studied from a gene-centric perspective, which does not re-quire MAG level delineation [Roume et al., 2015]. This work focuses on MAGs-centric analysis.

1.3.4 Bioinformatic analyses applied to longitudinal data

Different features of a defined system (bacterial taxa, iMGEs and CRISPR information), appear in varying quantities in different timepoints, and it is challenging to link and track them from one timepoint to another without any predefined reference. Therefore, the construction of a "representative longitudinal catalogue" (hereafter referred to as catalogue) of MAGs/genes, iMGEs and CRISPR information, provides a non-redundant representative base to link features from the different longitudinal samples [Herold et al., 2020]. The outcome of any downstream analysis is highly reliant on the quality of the MAGs and genes within a catalogue, which in turn depends on the quality large-scale bioinformatic processing (e.g. *de novo* assembly and binning).

Such a longitudinal catalogue can be obtained by an aggregated processing approach or by a sample-wise processing approach (**Figure 1.5**). The main advantage of an aggregated processing approach is its potential simplicity, due to the possibility of a single run for all the large-scale bioinformatic processing steps. While pooled sample sequencing assemblies have been shown to be effective [Magasin and Gerloff, 2015], especially in the advent of highly efficient *de novo* assemblers [Li et al., 2016] and digital normalisation [Brown et al., 2012]. Pooling sequencing reads from a large number of samples increases the complexity of the *de novo* assembly process, especially for complex communities. It also requires substantial computational resources, potentially resulting in lower quality contigs, MAGs and genes [Chen et al., 2020].

In a sample-wise approach, sequencing assemblies, binning and prediction of iMGEs and CRISPRs are performed per sample and pooled together for de-replication and clustering processes to construct the catalogue (**Figure 1.5**). Briefly, the de-replication method is applied after independent sample-wise large-scale bioinformatic processing [Evans and Denef, 2020]. On the one hand, a gene catalogue could be performed by predicting ORFs followed by de-replication through clustering [Li and Godzik, 2006; Edgar, 2010; Mirdita et al., 2019]. On the other hand, a MAG catalogue could be produced by the de-replication of MAGs, though it is more complex and requires several steps: i) definition of MAGs from sample-wise *de novo* assemblies, ii) curation of high-quality MAGs (high completeness, low contamination) [Parks et al., 2015], iii) de-replication of MAGs [Brito and Alm, 2016; Olm et al., 2017; Wampach et al., 2018; Evans and Denef, 2020] to select those which are most representative of the longitudinal data [Chen et al., 2020]. De-replication methods are particularly advantageous for longitudinal microbiome studies with many deeply sequenced MG and MT samples. This work relies on de-replication

based methods to generate the longitudinal catalogue.

Longitudinal catalogues provide compositional information of community taxa and po-

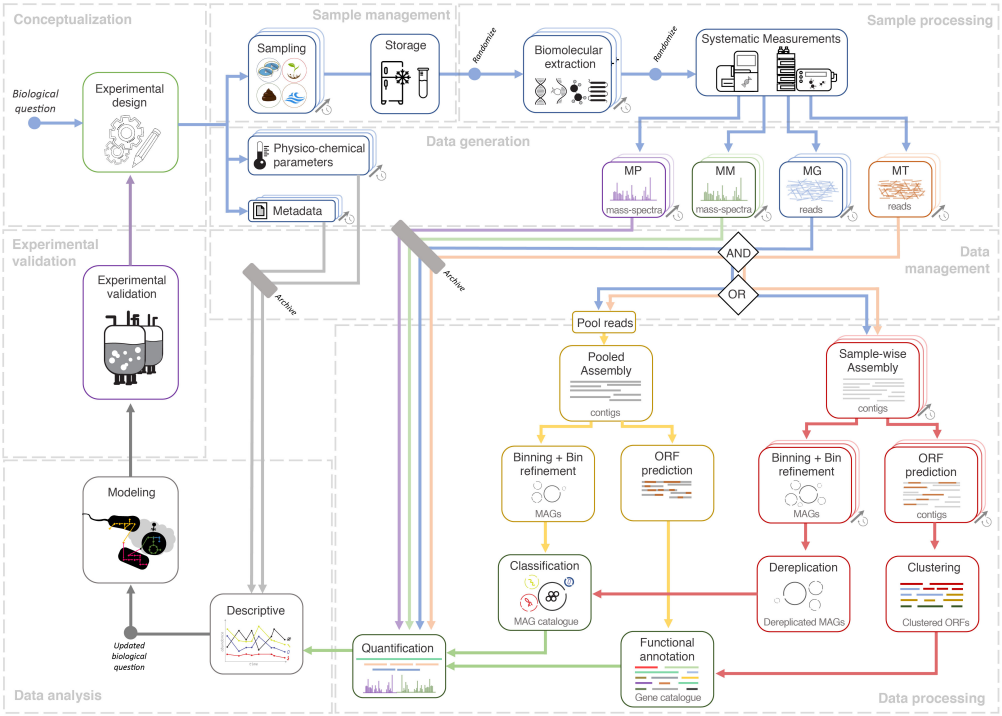


Figure 1.5: Generic workflow for a longitudinal and multio-mics microbiome data. From experimental design to validation.

tential functions. However, the relative quantification of community members and functionalities is key in harnessing the power of longitudinal microbiome data, as it allows the observation of community taxa/functional dynamics and could be used in downstream modelling. In that regard, quantifying MG and MT sequencing data is a standard process of aligning reads [Li and Durbin, 2009] to relevant catalogues, and then quantifying features of interest (e.g. population/gene relative genomic abundance, gene expression) based on those alignments, providing information on community structure, functional potential and gene expression. Complementally, MP data provides functional insights, whereby several methods are available for the quantification of such data [Delogu et al., 2020; Pible et al., 2020], while identification and quantification of metabolites through MM data [Kapoore and Vaidyanathan, 2016; Mallick et al., 2019; Røst et al., 2020] provide insights on the community phenotype(s). However, *in situ* measurements of substrate uptakes through labelling-based approaches [Starr et al., 2018] are challenging. There-

fore, specific metabolites of interest could be indirectly linked to members of a microbial community by proportionally assigning the relative contribution of a MAG to a given (re)constructed metabolic pathway based on genomic abundance or gene/protein expression [Noecker et al., 2016; Blasche et al., 2021].

1.3.5 Analysis of community characteristics and dynamics

Modelling community dynamics aims to make plausible predictions from previous observations and discernible patterns. However, a one-size-fits-all approach does not yet exist and has to be tailored for specific hypotheses and studies. Thus, data exploration is essential for a better understanding of what modelling approaches fit the data type, quality, and quantity.

Modelling microbial community omic data is challenging, especially when it involves short non-equidistant interval time-series with few samples, given the practicality and feasibility issues associated with *in situ* studies. In addition, microbiome omic data is i) compositional [Gloor et al., 2017], e.g. provided as relative abundances, which requires specific considerations when selecting statistical analyses, ii) highly sparse, such that the interpretation of zero-values generated from sampling, biological, or technical processes heavily affects data-derived conclusions [Silverman et al., 2020], and iii) high dimensional, which increases modelling difficulty due to the influence of feature selection that can heavily affect potential predictions [Bolón-Canedo et al., 2016]. Moreover, complementary use of different omics further increases complexity [Dahal et al., 2020; Thiele et al., 2020], but could add predictive power to these models [Fondi and Liò, 2015; Vasilakou et al., 2016]. Hence, the preprocessing, curation and transformation of the omics data, including metadata, is essential for downstream modelling efforts [Bordbar et al., 2014]. There are increasing efforts in the development of methods and frameworks towards improved longitudinal microbiome data analyses for a more accurate model selection [Bodein et al., 2019], e.g. a community model framework to test temporal structure and neutrality, and fitting to an interaction model [Faust et al., 2018].

To understand how environmental impact and variation of certain microbial populations may influence others, often resulting in shifts in microbial populations, quantitative characterizations of such dynamics are possible through mathematical models of observable patterns and their underlying mechanisms. Therefore, by leveraging the longitudinal microbial abundances one would be able to identify communal responses to perturbations, which, depending on the resilience of the community, may, for example, lead to a disbal-

ance of the microbial community [Gonze et al., 2017]. Initial exploration of the dynamics can be assessed by ordination analyses, where high dimensional population structure data is visualized in a two-dimensional space to observe the trajectory of the samples and the behaviour of the system, e.g. metastability, cycles, alternative states [Gonze et al., 2018]. Then, with correlation-based methods, mutualistic and competition relationships may be inferred by positive or negative correlations, whereby an ensemble approach of correlation methods has been shown to be useful [Weiss et al., 2016]. However, interactions within a community are complex and correlations are insufficient to infer them. Specifically, the application of other non-linear methods would be necessary to resolve those relationships, e.g. generalized Lotka-Volterra models [Fisher and Mehta, 2014]. Finally, most of the aforementioned approaches could also be extended to model microbiomes from a functional perspective, by complementing the community structure data with functional structure data, which could be further complemented or even replaced by functional models based on metabolic reconstructions [Biggs et al., 2015].

In the presence of multi-omics microbiome data, data integration have been routinely performed, with notable examples from studies in i) type-1 diabetes-derived microbiota [Heintz-Buschart et al., 2016], ii) healthy human gut [Tanca et al., 2017], iii) Crohn's disease [Erickson et al., 2012] and iv) activated sludge [Muller et al., 2014a; Roume et al., 2015]. However, equivalent studies within longitudinal microbiome analyses remain rather limited. Hence, this work includes an extensive multi-omic longitudinal data set derived from a model microbiome community (**Section 1.2.2**), which is used to study microbiome dynamics.

1.4 Objectives

The main objective of this work is to elucidate the microbial community dynamics of the model system under study, i.e. the foaming activated sludge, through the integration of longitudinal and multi-omics data. It is achieved through the generation of an extensive multi-omic longitudinal data set, followed by large-scale bioinformatics processing using state-of-the-art tools. Finally, modelling frameworks are applied, to characterize and understand community dynamics. This work specifically focuses on the interactions and dynamics between i) bacterial hosts and their associated iMGEs and ii) between bacterial populations.

In **Chapter 2** the main objective is to understand the roles of iMGEs on community dynamics. To achieve this, all the necessary components of the system were identified,

namely the microbial hosts and the iMGEs. To that end, the co-assembly MG and MT reads was performed, and then a pipeline combining several state-of-the-art tools was developed to identify MAGs and to construct a longitudinal catalogue of representative MAGs (rMAGs) that covers most of the abundant/dominant microbial community members. In parallel, the prediction of iMGEs, specifically phages and plasmids were carried out using state-of-the-art tools. In addition, CRISPR-based targeting via spacer sequences was used to further identify potential iMGEs. Therefore, CRISPR information was extensively extracted from the data. To that end, further classification of iMGEs and CRISPR-Cas systems was performed, resulting in a longitudinal catalogue for each of them, respectively. Consequently, iMGE-host interactions were inferred through the links defined via the CRISPR information. Quantitative information (such as rMAG and iMGE abundances) were then used to develop models of the dominant microbial family, and to construct iMGE and host interactions networks to explain the relative importance of iMGEs in the model system.

Chapter 3 aims to unravel the interactions between microbial populations of the model system over time, through an ecological context. Those ecological interactions were then explained using functional characteristics of interacting populations. To that end, modelling based on rMAG abundances over time, including physico-chemical recordings, were utilized to build an ecological interactome. This ecological interactome was inferred for the entire time-series and for shorter-time windows to reveal the nature of long-term and short-term interactions. Additionally, functional information, i.e. functional potential and expression, was analysed to explain the nature of the ecological interactions, with focus on the dominant population of *Candidatus*. Microthrix parvicella.

Chapter 2

Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics

This chapter describes longitudinal multi-omics microbiome data analyses to elucidate the dynamic interaction between (bacterial) hosts and iMGEs. This chapter includes material from the following published first author peer-reviewed publication.

Susana Martínez Arbas, Shaman Narayanasamy, Malte Herold, Laura A. Lebrun, Michael R. Hoopmann, Sujun Li, Tony J. Lam, Benoît J. Kunath, Nathan D. Hicks, Cindy M. Liu, Lance B. Price, Cedric C. Laczny, John D. Gillece, James M. Schupp, Paul S. Keim, Robert L. Moritz, Karoline Faust, Haixu Tang, Yuzhen Ye, Alexander Skupin, Patrick May, Emilie E. L. Muller and Paul Wilmes

2021

Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics

Nature Microbiology, 6:123-135. **Appendix A.2**

2.1 Abstract

Viruses and plasmids (invasive mobile genetic elements, iMGEs) have important roles in shaping microbial communities, but their dynamic interactions with CRISPR-based immunity remain unresolved. We analysed generation-resolved iMGE-host dynamics spanning one and a half years in a microbial consortium from a biological wastewater treatment plant using integrated meta-omics. We identified 31 bacterial metagenome-assembled genomes (MAGs) encoding complete CRISPR-Cas systems and their corresponding iMGEs. CRISPR-targeted plasmids outnumbered their bacteriophage counterparts by at least 5-fold, highlighting the importance of CRISPR-mediated defence against plasmids. Linear modelling of our time-series data revealed that the variation in plasmid abundance over time explained more of the observed community dynamics than phages. Community-scale CRISPR-based plasmid-host and phage-host interaction networks revealed an increase in CRISPR-mediated interactions coinciding with a decrease in the dominant *Candidatus* Microthrix parvicella population. Protospacers were enriched in sequences targeting genes involved in the transmission of iMGEs. Understanding the factors shaping the fitness of specific populations is necessary to devise control strategies for undesirable species and to predict or explain community-wide phenotypes.

2.2 Introduction

Microbial community dynamics are driven by both abiotic (environmental) and biotic (biological) factors (**Section 1.1**). The latter includes iMGEs which move between genomes, such as bacteriophages and plasmids [Zhang et al., 2013; Koonin et al., 2017] and are believed to play an important role in microbial community dynamics [Rizzo et al., 2013; Jassim et al., 2016] (**Section 1.1.1**), by transferring detrimental or beneficial genetic material to or between hosts [Zhang et al., 2013; Koonin et al., 2017]. They also represent key components in horizontal gene transfer, contributing to the spread of antimicrobial resistance [Zhang et al., 2011]. Here,we present a time-resolved, integrated meta-omic analysis aimed at elucidating CRISPR-mediated interactions and dynamics between iMGEs and their hosts, within a model mi-crobial community in activated sludge of a BWWTP, described in **Section 1.2**. We utilize correlation and linear regression methods to describe community dynamics under the lens of the dominant population *Candidatus* Microthrix parvicella and information from the CRISPR-Cas system to define host and iMGE interactions. Exploiting the sequence-based CRISPR links between spacers and protospacers, specific host populations can be linked n

to specific iMGEs, as well as to their corresponding invasion events[Zhang et al., 2013; Koonin et al., 2017].

2.3 Methods and Material

2.3.1 Sampling

Individual floating sludge islets within the anoxic tank of the Schifflange BWWT plant (Esch-sur-Alzette, Luxembourg) were sampled according to previously described protocols [Muller et al., 2014a]. Samples are indicated as dates (YYYY-MM-DD). Time-resolved sampling included two initial sampling dates (2010-10-04 and 2011-01-25) as previously reported [Muller et al., 2014a; Roume et al., 2015]. More frequent sampling was performed from 2011-03-21 to 2012-05-03, of which data from three samples (2011-10-05, 2011-10-25 and 2012-01-11) have been previously published [Muller et al., 2014a].

2.3.2 Concomitant biomolecular extraction and high-throughput metaomics

Concomitant biomolecular extraction of DNA, RNA and proteins as well as high-throughput measurements to obtain MG, MT, and MP data were carried out according to previously established protocols [Muller et al., 2014a; Roume et al., 2015].

2.3.3 Isolate culture, genome sequencing and assembly

85 isolate cultures of lipid accumulating bacterial strains were derived from the sludge islets sampled from the same anoxic tank described above. The isolation protocol, including the screening for lipid accumulation properties (via Nile Red staining), DNA extraction, and sequencing was performed as previously described [Roume et al., 2015; Muller et al., 2017]. The genomic data was assembled and analysed using an automated version of a previously described workflow [Muller et al., 2017] that spanned sequencing read preprocessing, *de novo* assembly and gene annotation (See **Section 2.3.16**). The genome of *Candidatus* Microthrix parvicella Bio17-1 was obtained from the publicly available NCBI BioProject PRJNA174686 [Muller et al., 2012].

2.3.4 Co-assembly of metagenomic and metatranscriptomic data

Sample-wise integrated MG and MT data analyses were performed using IMP version 1.3 [Narayanasamy et al., 2016] with customized parameters, i.e. i) Illumina Truseq2 adapters were trimmed, ii) the step involving the filtering of reads of human origin step was omitted

for the preprocessing, and iii) the MEGAHIT *de novo* assembler [Li et al., 2014] was used for the co-assembly of MG and MT data. Nonpareil2 [Rodriguez-R and Konstantinidis, 2014b] was applied to the preprocessed MG and MT data to assess the relative depth of coverage.

2.3.5 Metaproteomic analyses

Raw mass spectrometry files were converted to MGF format using "MSconvert" with default parameters. The resulting files were used to run the Graph2Pro pipeline [Tang et al., 2016] together with the corresponding assembly graphs from MEGAHIT, which allowed the integration of MG, MT and MP data. Assemblies often result in fragmented consensus contigs leading to a loss of information on strain variation, as well as, open-reading frames spanning multiple contigs. The Graph2Pro pipeline combines the Graph2Pep algorithm and FragGeneScan [Rho et al., 2010] to predict peptides from short and long edges of the graph even if the peptides span multiple edges. Graph2Pro further predicts protein sequences from the graphs of the IMP-based co-assemblies, using identified peptides as constraints. To produce the final protein identifications, MP data was searched against the sample-specific databases derived from Graph2Pro.

The combined set of tryptic peptides was used as the target database for peptide identification using the MS-GF+ search engine [Kim and Pevzner, 2014] using customized parameters. The instrument type was set to a high-resolution LTQ with a precursor mass tolerance of 15ppm and an isotope error range of -1 and 2. The minimum and the maximum precursor charges were set to 1 and 7, respectively. The false discovery rate (FDR) was estimated by using a target-decoy search approach where reverse sequences of the protein entries were generated while preserving the C-terminal residues (KR) and concatenated to the database. All identifications were filtered in order to achieve an FDR of 1%.

Identified peptides from the Graph2Pro pipeline were assigned using peptidematch [Chen et al., 2013] against Prokka-based [Seemann, 2014] predictions from IMP for protein-coding sequences of the rMAGs, and prodigal-based predictions [Hyatt et al., 2010] including fragmented genes (see **Section 2.3.11**) for protein-coding sequences of the iMGEs.

2.3.6 Binning, selection of representative genome bins, taxonomy and estimation of abundance

Co-assembled contigs from each timepoint were binned, as described previously [Heintz-Buschart et al., 2016]. Binning was based on nucleotide signatures, presence of single-copy essential genes and metagenomic depth-of-coverage. Bins from each timepoint with at least 28% completeness and contamination of less than 20% along with the 85 isolate genomes were subjected to a dereplication process, using dRep [Olm et al., 2017] v0.5.4, to select representative metagenome-assembled genomes (rMAGs). Accordingly, dRep parameters were set to: i) genome completeness of 0.6 (based on CheckM [Parks et al., 2015] v1.0.7), ii) strain heterogeneity of 101, iii) average nucleotide identity (ANI) threshold of 0.6 to form primary clusters, and iv) ANI threshold of 0.965 to form secondary clusters. Taxonomic classification was performed using a customized version of AMPHORA2 [Wu and Scott, 2012]. Additionally, taxonomic classification was performed with Sourmash [Brown and Irber, 2016] 2.0.0a1-lca-version with kmer-length of 21 and threshold of 4 using an existing database which included around 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 less than one third of the total probability scores, the assignment was discarded as a "low confidence assignment". Taxonomic assignments of AMPHORA2 and sourmash-lca were combined and then filtered to select a final taxonomic assignment for the rMAGs, giving priority to predictions from sourmash-lca due to higher expected specificity and an updated database. We then selected rMAGs with a "completeness - contamination" value of $\geq 50\%$ for further downstream analyses.
To represent population-level abundances and transcription levels, the preprocessed MG and MT paired- and single-end reads from all the time-series samples were mapped onto the collection of rMAGs using "bwa mem" [Li and Durbin, 2009] and contig-level average depth-of-coverage values were extracted for the MG and MT data. Gene-level MT read counts for all the predicted genes present within each rMAG were normalized to obtain the corresponding gene expression values, specifically, gene length and total read counts per sample were used as scaling factors to obtain normalized gene expression values.

2.3.7 Identification of CRISPR elements

CRISPR information, i.e. spacers, repeats and flanking sequences, were predicted using CRASS [Skennerton et al., 2013] version 0.3.8 based on the IMP-based preprocessed MG and MT paired- and single- end reads as input. MetaCRT [Rho et al., 2012] was used to predict spacers and repeats from IMP-based MT and co-assembled contigs. A custom script was used to extract flanking regions from the metaCRT results.

The redundancy of spacers, repeats and flanking sequences was reduced by clustering the sequences with CD-HIT-EST [Fu et al., 2012] version 4.6.7. Spacers were clustered using 90% sequence identity [Moller and Liang, 2017; Lam and Ye, 2019], covering the entire length of the compared sequences [Moller and Liang, 2017]. CRISPR-flanking regions were clustered using 99% sequence identity, with at least 97.5% coverage of both the compared sequences. On the other hand, the CD-HIT-EST clustering parameters for repeats were determined manually by clustering the known repeats belonging to a single CRISPR locus of *Ca*. M. parvicella Bio17-1 [Muller et al., 2012]. Specifically, the sequence identity parameter was first set to 99% and the sequence coverage was set to 100%. These parameters were reduced by 5% in the following iterations until all repeats were regrouped into a single cluster. Subsequently, all the known repeats of *Candidatus* Microthrix parvicella were clustered at i) 80% sequence identity, ii) covering the length of at least 75% of the shorter sequence. Thereby, these parameters were used for the clustering of all repeats. FASTA headers of all the sequences were left unchanged (i.e. -d parameter in CD-HIT-EST) because they contained information required for downstream analyses (e.g. sample name, contig name, CRASS-computed coverage, etc.). The clustering procedure for the different CRISPR elements yielded non-redundant sequences of repeats, spacers and flanking regions.

Spacer abundances were estimated by extracting their coverage values from CRASS. Equivalent information was obtained from metaCRT by using "bwa-mem" to map MG and MT reads from each of the time-resolved samples to the entire set of contigs predicted by metaCRT (i.e. contigs containing at least one CRISPR locus). The depth-of-coverage information was derived using bedtools [Quinlan and Hall, 2010]. Based on this, abundance values were extracted for each of the predicted spacers per timepoint. The depth-of-coverage information of the metaCRT contigs was then consolidated using CRASS coverage results by referring to the non-redundant spacer clusters (derived from CD-HIT-EST). The consolidated results are hereafter referred as "spacer abundance values". Specifically, the spacer abundance values from the specific timepoints were assigned to the non-

redundant spacers thereby allowing a temporal representation of spacer abundance values. Subsequently, the spacer abundance values were transformed to counts per million (CPM) [Robinson et al., 2009; Sha et al., 2015] per sample of non-redundant spacers that had at least one read count, in at least one sample were selected and the CPM values were calculated. Finally, to determine presence/absence of a given spacer, a minimum cutoff of CPM = 1 was applied. Applying standard cutoffs (i.e. above 3-5) caused loss of information from the short spacer sequences within the repetitive CRISPR regions, which usually do not recruit many reads during the mapping process.

2.3.8 Linking rMAGs to CRISPR elements

The non-redundant flanking regions and repeats were used to associate MAGs with specific CRISPR loci using BLASTN [Altschul et al., 1990]. Non-redundant CRISPR-flanking sequences and CRISPR-repeats were searched against the contigs of the MAGs. Flanking sequences and MAG-contig(s) exhibiting similarities of at least 95% of identity and coverage of either: i) 80% for flanking sequences >100bp or ii) 95% for flanking sequences <100bp were retained for the downstream filtering steps. Next, the aforementioned flanking sequences for which the associated repeats had at least 75% of identity and 80% coverage against the MAG-contig(s) were further retained for downstream processing. Upon defining the selected flanking repeats sequences linked to a MAG, spacers linked to the repeats flanking sequences were then associated to the MAG. Thereby, the composition of spacers per MAG was determined. Finally, all the CRISPR information belonging to a MAG was linked to its rMAG to preserve the maximum amount of CRISPR information.
cas genes and CRISPR types and subtypes were predicted from all the assembled contigs using CRISPRone [Zhang and Ye, 2017]. The *cas* genes and CRISPR types were then assigned to their respective MAGs.
We then selected rMAGs predicted as *Candidatus* Microthrix parvicella (see **Section 2.3.6**) to inspect the *cas* genes and CRISPR type predictions. Next, we used CRISPRCasFinder [Couvin et al., 2018] to further confirm the selected *cas* genes and CRISPR type predictions of *Candidatus* Microthrix parvicella. We performed manual curation on all the rMAGs predicted as *Candidatus* Microthrix parvicella. We identified a contig (D47_L1.43.1_contig_476300) of 10,224 bps that encoded a complete CRISPR operon which was highly similar to the CRISPR operon of the isolate genome of *Candidatus* Microthrix parvicella Bio17-1. This contig was incorporated with rMAG-165.

2.3.9 Identification of protospacers and protospacer-containing contigs

A BLASTN [Altschul et al., 1990] search was performed using all non-redundant spacers as queries against the contigs from all timepoints using the parameters defined in CRISPRtarget [Biswas et al., 2013]. Spacer matches with at least 95% coverage and 95% identity were selected for further analysis [Shmakov et al., 2017]. Any IMP-based MT or co-assembled contigs containing repeat sequences and/or identified by metaCRT to encode CRISPR sequences were excluded from downstream analyses. Accordingly, the remaining spacer matches (or complements) were defined as protospacers, and the respective contigs that contained at least one protospacer were defined as protospacer-containing contigs (PSCC) and were retained as iMGEs.

2.3.10 Classification of iMGEs

Bacteriophage sequences were predicted by analysing all co-assembled contigs using VirSorter [Roux et al., 2015] version 1.0.3 and VirFinder [Ren et al., 2017] version 1.0.0. Similarly, plasmid sequences were predicted using cBar [Zhou and Xu, 2010] version 1.2 and PlasFlow [Krawczyk et al., 2018] version 1.0.7. The predictions were consolidated by annotating candidate iMGE sequences as "plasmid" if the sequences were positively predicted by cBar and/or PlasFlow, as "phage" if the sequences were positively predicted by VirSorter and/or VirFinder, as "ambiguous" if the sequences were predicted as both plasmid and phages by any combination of the aforementioned tools, and finally as "unclassified" if they contained at least one protospacer and were not annotated as phage or as plasmid. Following this step, all iMGEs (i.e. phages, plasmids, ambiguous and unclassified) were clustered using CD-HIT-EST with clustering parameters of 80% identity and at least 50% coverage, generating the non-redundant set of iMGEs. The classification/annotation of representative clusters was retained for the downstream analyses. Finally, BLASTN [Altschul et al., 1990] was performed on the clustered contigs against NCBI plasmids and virus databases to retrieve their taxonomy.

Genomic and transcriptomic abundances of the iMGEs were obtained by mapping the IMP-preprocessed MG and MT paired- and single-end reads from all timepoints to the iMGE representative contigs using bwa-mem [Li and Durbin, 2009]. The contig-level average depth of coverage derived from the MG and MT data represented the iMGE abundance and iMGE gene expression, respectively.

2.3.11 Gene annotation of phage- and plasmid-derived contigs

Open reading frames within iMGEs were predicted using Prodigal [Hyatt et al., 2010] v2.6 with "meta" and "incomplete gene" settings. Predicted genes were annotated using hmmsearch [Johnson et al., 2010] against an in-house licensed version of the KEGG database [Kanehisa et al., 2016]. KEGG function identifiers were then converted to the higher-level COG functional categories [Tatusov et al., 2000]. Finally, ARGs were annotated using "hmmsearch" against ResFam's full HMM database [Gibson et al., 2015].

2.3.12 Linear model of community dynamics

Correlations between family-level groups, whereby plasmids and phages were assigned to bacterial families based on their previous contig assignments to MAGs were calculated using the "rcorr" function within the "Hmisc" R package. Euclidean distances of the correlation vectors were calculated using the "dist" function ("stats" R package). Next, a hierarchical clustering was applied on the calculated Euclidean distances, using the "hclust" function ("stats" R package). The tree was then cut with a height parameter of four (i.e. H = 4), using the "cutree" function from R "stats" package [R Core Team, 2013].
The "lm" function from the R "stats" package was used to generate the models. To avoid overfitting, we restricted the linear models to a maximum of 15 family-level groups. Random sampling was performed for 100,000 model realisations and model quality was assessed by the adjusted R^2 value. In our first approach, we did not restrict the model composition and allowed all combinations with the same probability. Then, from the random sampling data, we ranked models based on the adjusted R^2 value and looked for enrichment in specific families within the best models (N=25, N=50, N=100). In a first iteration, we selected enriched families and iMGEs, i.e. plasmids and phages, to obtain a global model, and then, we selected the significant groups from the global model to obtain a reduced model. Once we had the models for the entire time-series and the shorter-time intervals, we identified the common significant groups in all the models. Next, we removed the group *Microthrixaceae*-plasmids from the reduced models for each time interval, to assess the influence of these plasmids within the performance of the model.

2.3.13 Network analyses and visualization

CRISPR-based plasmid-host and phage-host networks were defined by the co-occurrence of rMAGs, spacers and a targeted iMGEs in at least one timepoint. Thus, if a given non-redundant spacer was assigned to a specific rMAG and this specific rMAG did not co-occur

in at least one timepoint, this spacer was deemed inactive within this rMAG throughout the time-series. Consequently, a spacer was assigned to a rMAG if, and only if, the spacer co-occured with its assigned rMAG in at least one timepoint. Thus, the iMGEs targeted by the spacers assigned to rMAGs were used to build the CRISPR-based plasmid-host and phage-host networks. Finally, the timepoint-specific networks were built based on the presence/absence of the rMAGs and their linked plasmids or phages. Network properties such node degree, betweenness and closeness were estimated by the function "speciesLevel", within the "bipartite" R package [Dormann et al., 2008]. Modularity, defined by the value of Q [Newman, 2006], and nestedness, defined as the value of the "Nestedness matrix based on Overlap and Decreasing Fill" (NODF) [Almeida-Neto et al., 2008], were calculated using the functions "computeModules" and "nested", respectively. Visualization and manual inspection of the networks were performed with Cytoscape [Shannon et al., 2003] version 3.6.1. R version 3.4.1, together within the "tidyverse" framework, was used for processing data tables, statistical analysis and data visualization [Conway et al., 2017].

2.3.14 Estimation of spacer gain-loss and CRISPR locus dynamics

Based on the previously calculated CPMs per rMAG, their assigned spacers and iMGEs, dates of first and last occurrence within the timeseries were defined. We subsequently defined events of gain and loss of spacers, and possible secondary encounters of the iMGE with the rMAGs in order to resolve the variation within a given CRISPR array per population. These events were classified as follows: i) gain of a given spacer if its first detection within the timeseries occurred after the first occurrence of its targeted iMGE, ii) likely gain of a given spacer if both the spacer and its targeted iMGE occurred for first time at the same timepoint, iii) likely secondary encounter if the spacer occurred for first time before its linked iMGE, iv) loss of a given spacer if the spacer's last detection occurred after the last detection of its linked iMGE, v) likely loss of a given spacer if the last detection of both spacer and iMGE occurred at the same timepoint, vi) spacer loss before iMGE loss if the last occurrence of the spacer occurred before the last occurrence of the iMGE.

2.3.15 Workflows automation and computing platforms

The automation of workflows was made using several versions of snakemake [Köster and Rahmann, 2012], from 3.10.2 to 5.1.4. All computing was run on the University of Luxembourg High-Performance Computing (ULHPC) platform [Varrette et al., 2014].

2.3.16 Data and code availability

The genomic FASTQ files, rMAGs, and isolate genomes from this work are publicly available within NCBI BioProject PRJNA230567. Similarly, MP data from this work is publicly available in the PRIDE database under the accession number PXD013655.
Additional publicly available projects cited by this work include NCBI BioProject PRJNA174686.
The code is available on three separate repositories: i) the IMP, binning and population genomes can be found in `https://github.com/shaman-narayanasamy/LAO-time-series` (doi: 10.5281/zenodo.3988660), ii) the CRISPR and MGE analyses can be found in `https://github.com/susmarb/LAO_multiomics_CRISPR_iMGEs` (doi: 10.5281/zenodo.3988592), iii) the isolate assembly analyses can be found in `https://github.com/shaman-narayanasamy/Isolate_analysis` (doi: 10.5281/zen-odo.3988667).

2.4 Results

2.4.1 Time-resolved meta-omics of foaming sludge islets

53 samples of foaming sludge islets from the surface of an anoxic tank were collected from a BWWTP over a period of 578 days. The mean sampling frequency of eight days (SD=16 days) is equivalent to the doubling time of the dominant population, *Candidatus* Microthrix parvicella (*Ca*. M. parvicella) [Rossetti et al., 2005; Sheik et al., 2016], thereby facilitating the study of population dynamics on a generational timescale. Concomitant DNA, RNA and protein fractions were obtained from each sample [Roume et al., 2013], which is critical for coherent downstream systematic measurements and multi-omic data integration [Roume et al., 2013]. These biomolecular fractions were subjected to deep, high-throughput measurements resulting in time-resolved metagenomic (MG), metatranscriptomic (MT), and metaproteomic (MP) data. A total of 1.5×10^9 MG reads and 1.7×10^9 MT reads underwent sample-specific, large-scale bioinformatic processing, followed by MG and MT *de novo* co-assembly [Narayanasamy et al., 2016] yielding a total of 2.1×10^7 contigs (**Appendix B.1**). Additionally, we estimated ~50% average coverage of community members resolved for the individual timepoints (**Appendix E.1**). MP datasets yielded a total of 7.6×10^6 mass spectra, whereby a total of 9.6×10^7 redundant peptides were identified per sample using the 3.1×10^7 protein sequences predicted from the co-assembled contigs as the search database (**Appendix B.2**).

Contigs from the co-assembled MG and MT data from each sample were binned, producing a total of 26,524 metagenome-assembled genomes (MAGs) across all samples (**Appendix B.1**), of which 1,364 MAGs were selected for dereplication together with a collection of 85 isolate genomes (**Appendix E.2**). The dereplication yielded pools of MAGs for which we defined representative MAGs (rMAGs) [Olm et al., 2017]. These rMAGs underwent taxonomic classification, quality filtering and manual curation thereby yielding a total of 92 rMAGs which were retained for downstream analyses (**Appendix B.3**). In this work, rMAGs are assumed to represent pools of MAGs resulting from dereplication and are equivalent to populations. Therefore, our population-level analyses are, by default, on the rMAG-level unless otherwise specified.

2.4.2 CRISPR-Cas information over the entire meta-omics dataset

We resolved the CRISPR-Cas systems within rMAGs by extracting their respective *cas* genes and classifying the CRISPR-types [Zhang and Ye, 2017]. This resulted in a final set of 31 (37%) rMAGs that encoded classifiable and complete CRISPR-Cas systems, i.e. *cas* genes allowing CRISPR-Cas system classification, and CRISPR loci containing the required information for linking hosts to iMGEs [Edwards et al., 2016]. The most common CRISPR-Cas system within the community was type I, which was found in 21 rMAGs and across several taxonomic families, followed by type III assigned to nine rMAGs, while type II and V systems were identified in three and one rMAGs, respectively. Combinations of different CRISPR types within a single rMAG were also detected. Accordingly, we found type I and III to be present together in five rMAGs, thereby representing the most commonly detected combination [Crawley et al., 2018] (**Figure 2.1**, **Appendix B.4**). We used an ensemble of computational methods to extract CRISPR information on the read- and contig- level, resulting in an extensive set of detected CRISPR repeats and spacers (both collectively referred to as CRISPR elements) per sample. Overall, we retrieved 89,856 repeats and 525,579 spacers over the entire time-series. However, they are redundant because the same repeats or spacers may appear at multiple timepoints (**Figure 2.2**). Therefore, we removed redundancy by clustering CRISPR elements, resulting in 8,469 and 162,985 non-redundant repeats and spacers, respectively. Spacers were more highly represented on the MG-level whereas repeats were more highly represented on the MT-level (**Appendix E.3**). 778 (~9%) non-redundant repeats and 20,002 (~12%) non-redundant spacers could be directly assigned to at least one rMAG, in turn representing 196,159 (~37%) and 29,685 (~33%) redundant spacers and repeats, respectively. In order to re-

tain the maximum amount of information for downstream analyses, the entire collection of spacers and repeats from the entire pool of MAGs were linked to their corresponding rMAGs (**Appendix B.4**). Although this may result in high numbers of unfiltered spacers associated with certain rMAGs, e.g. rMAG-117 which represents 41 MAGs and is associated with 6,574 spacers, this approach allows comprehensive tracking of CRISPR and targeted iMGE dynamics.

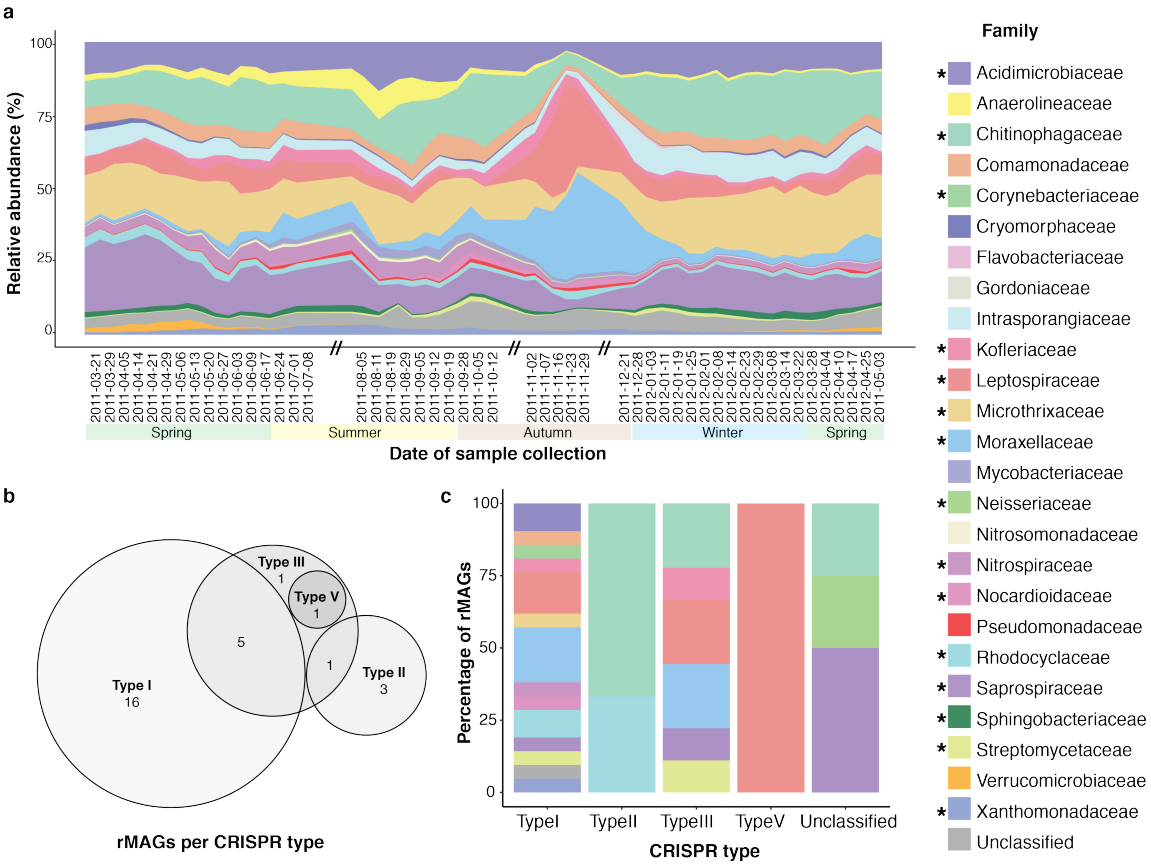


Figure 2.1: Community dynamics and CRISPR-Cas type distribution. a, Relative abundance of representative metagenome-assembled genomes (rMAGs) over time. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system. **b**, Venn diagram of CRISPR-Cas system types based on the numbers of rMAGs that encode them. Overlaps indicate single rMAGs carrying more than one CRISPR-Cas system. **c**, Distribution of taxonomic affiliations at family rank per CRISPR-Cas system type. (**a**) and (**c**) The legend colours marked with asterisks (*) represent families containing CRISPR-Cas systems.

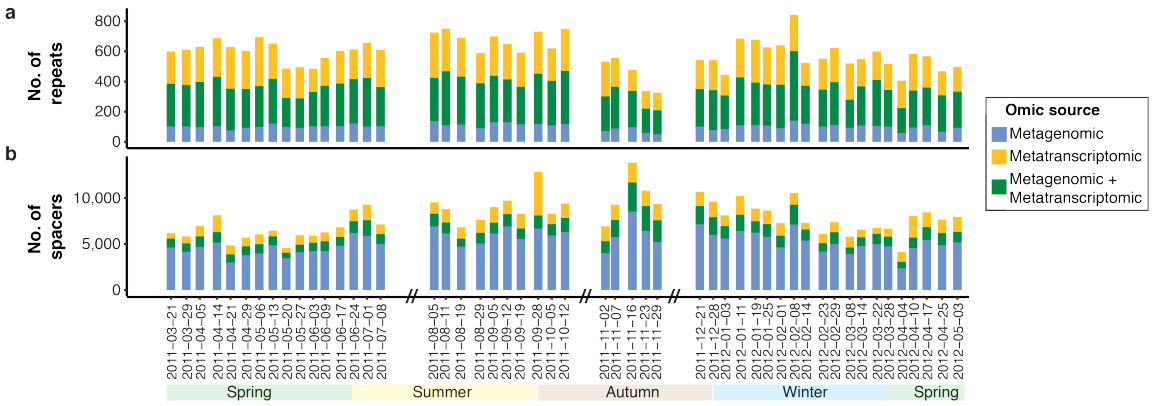


Figure 2.2: Non-unique CRISPR elements over time. Number of predicted **a**, repeats, and **b**, spacer per time point. The labels in the x-axis indicate the exact sampling dates, and the double slashes (//) represent absence of samples due to absence of foaming islets.

2.4.3 Protospacers in the entire meta-omics dataset

Protospacers may represent either the origin of the spacers or targets for iMGEs inhibition/splicing. Spacer information from the CRISPR loci can be used to detect iMGEs through complementary matching to their targeted protospacers [Marraffini and Sontheimer, 2010; Amitai and Sorek, 2016]. Single matches of spacers to targeted iMGEs have been shown to be sufficient for conferring immunity against such iMGEs [Bolotin et al., 2005; Mojica et al., 2005]. Thus, spacers were searched against all contigs. Those containing at least one protospacer match, i.e. protospacer-containing contigs (hereafter referred to as PSCCs), and lacking repeats to avoid self-matching, were considered as putative iMGEs. Accordingly, we detected 750,375 protospacers within 224,651 PSCCs (**Figure 2.3**), highlighting the large number of PSCCs that encode multiple protospacers (56%). It is noteworthy that the filtering of PSCCs with repeats (109,504 redundant PSCCs) resulted in the exclusion of potential iMGEs encoding CRISPR loci.

Upon the removal of redundancy with the iMGEs (next section and **Appendix E.4**), a total of 209,199 protospacers were retained within 49,306 non-redundant PSCCs (**Table 2.1**). Here, there were instances of single spacers targeting multiple protospacers, from either different or the same PSCCs. On average, one spacer targeted 21.85 protospacers (median=7, SD=51.27), while PSCCs tended to contain more than one protospacer (i.e. mean=3.29, median=2, SD=4.60).

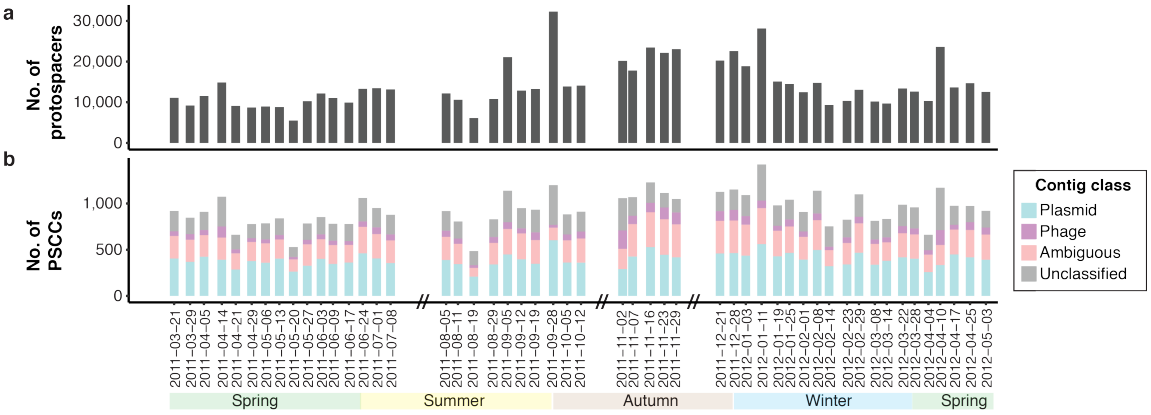


Figure 2.3: Non-unique protospacers, and protospacer-containing contigs (PSCC) over time. Number of predicted **a**, protospacers, and **b**, PSCCs per time point. The labels in the x-axis indicate the exact sampling dates, and the double slashes (//) represent absence of samples due to absence of foaming islets.

Table 2.1: Protospacer-containing contigs summary. Number of spacer matching protospacers, before and after filtering procedures.

Before search		**Unfiltered**	***Filter 1**	****Filter 2**	**Unique iMGE**
Unique Spacer	162,985	110,307	78,777	34,335	-
Non-unique (before redundancy removal)					
Protospacer	NA	18,599,466	2,400,670	750,375	209,199
PSCCs	NA	1,599,109	334,155	224,651	49,306

*** Filter 1: Filtering based on 95% identity and 95% query coverage**
**** Filter 2: Filtering sequences containing repeats**

2.4.4 Plasmids and phages in the entire meta-omics dataset

Based on the contigs from all timepoints, we predicted phage and plasmid sequences. The total number of annotated iMGEs represented 6.97% of all contigs, for which 2.22% contained at least one protospacer (i.e. PSCCs). Interestingly, we found that sequences annotated as plasmids outnumbered phages by ~16-fold (**Appendix E.4**). At this stage, there was a lack of predicted prophage sequences, likely due to the limitations of the available phage prediction methods. All the predicted iMGEs were clustered to yield non-redundant representative iMGEs traceable over time, which maintained similar proportions to the previously described redundant set, i.e. ~16-times more plasmid than phages (**Table 2.2**). Among these, we found 12,232 (1.7%) plasmids and 227 (0.5%) phages with similarities to sequences within the NCBI database, demonstrating a representation lack of those elements within public databases. A similar trend in proportions is reflected in the iMGEs

Table 2.2: Summary of redundancy removal of iMGEs. Number of candidate invasive mobile genetic elements (iMGEs) and protospacer-containing contigs (PSCC) before and after the removal of redundancy (using CD-HIT). The number of protospacers within the PSCCs is based on blastn matching of spacers and PSCC. Additionally, the classification of iMGEs as phage, plasmid, ambiguous and unclassified is based on predictions from VirSorter, VirFinder, PlasFlow and cBar.

iMGE	Redundant sequences		Non-redundant sequences		Proportion of redundancy (%)
	With protospacer	**Total**	**With protospacer**	**Total**	
Plasmids	30183	1389477	18778	707093	45,187
Phages	3923	83733	2518	42039	48,974
Ambiguous	3662	46863	2401	80618	45,793
Unclassified	11538	11538	6663	6663	42,252
Total	49306	1531611	30360	836413	43,627

targeted by spacers. Plasmids (12,412) are targeted 5-times more frequently than phages (2,351). Since we were interested in iMGEs that are interacting with hosts via CRISPR, we focused on the non-redundant iMGEs that were also PSCCs (henceforth we collectively refer to these as iMGEs) for downstream analyses. Additionally, the MG and MT co-assembled contigs allowed the detection of iMGEs exclusively present on the MT level, e.g. RNA phages [Callanan et al., 2018]. Accordingly, a total of 2,890 MT-only contigs assigned as iMGEs were retrieved, from which 2,102 and 387 were classified as plasmid and phage, respectively.

BWWTPs are thought to represent hotspots for the spread of antimicrobial resistance genes (ARGs) [Rizzo et al., 2013; Tong et al., 2019]. Therefore, we inspected plasmid and phage functions targeted by CRISPR systems [Davison et al., 2016; Shmakov et al., 2017] and screened those iMGEs for potential ARGs (**Appendix E.5**, **Appendix B.5**, **Appendix B.6**, **Figure 2.4**). We found 1,570 (0.22%) plasmids and 106 (0.25%) phages encoding 38 different ARGs, including tetracycline-resistance genes, which are known to be persistent in BWWTP [Che et al., 2019; Tong et al., 2019]. Additionally, we found ten plasmid-PSCCs. Among these, three were encoding ARGs that were being targeted by spacers, specifically aminoglycoside nucleotidyltransferase (ANT3), streptomycin phosphotransferase (APH3") and Class D beta-lactamases (ClassD) (**Table 2.3** and **Appendix B.7**). Apart from these specific cases, iMGEs encoding ARGs were not PSCCs and, thus, are likely not targeted by CRISPRs.

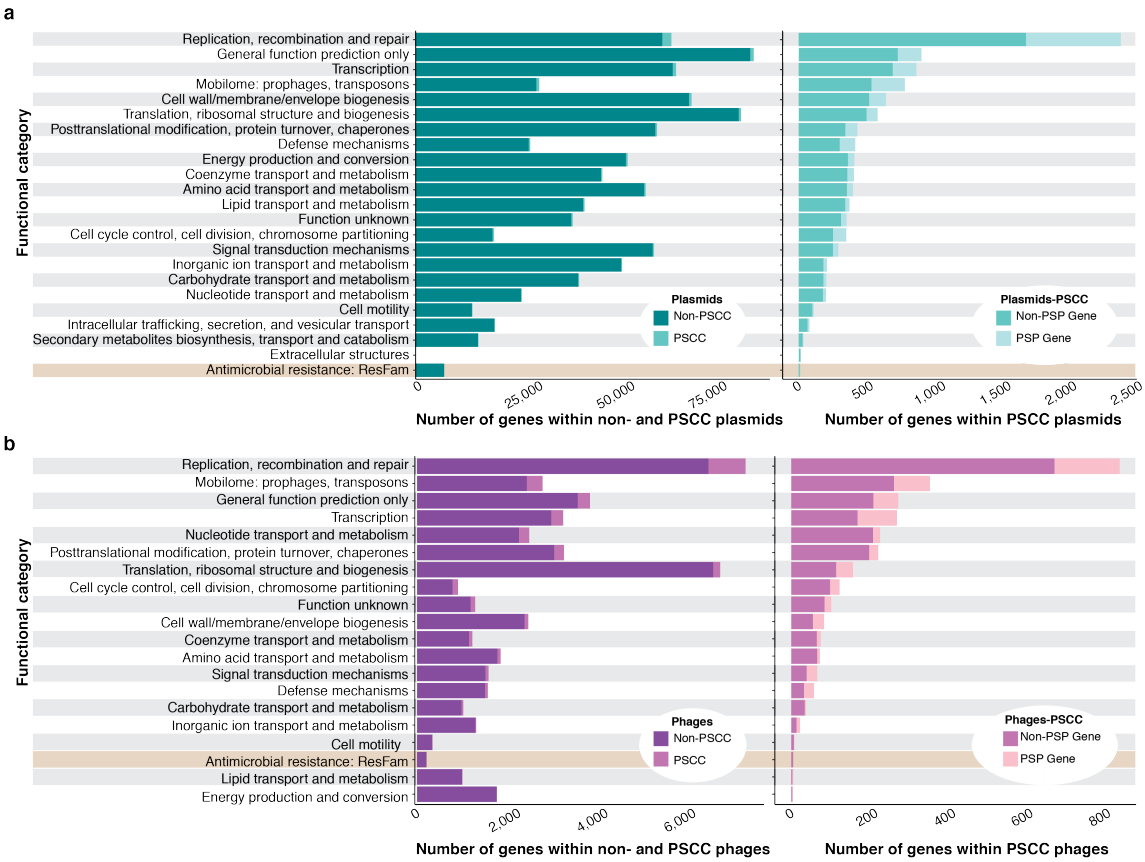


Figure 2.4: Functional gene categories encoded and targeted within plasmids and phages. **a**, Functional categories encoded by plasmids and **b**, by phages. **a**, **b** Each bar indicates the number of genes found per functional category. The left bar plots show the number of genes of specific functional categories within invasive mobile genetic elements (iMGEs) with and without protospacers (that is PSCCs). For those iMGEs that are PSCCs, the right bar plot highlights the number of genes of specific functional categories for which protospacers occurred within the intragenic regions.

Table 2.3: Antibiotic resistant genes (ARGs) within iMGEs.

	Plasmids		Phages	
	Number	**Percentage**	**Number**	**Percentage**
Total	707093	100	42039	100
Number of MGEs containing at least one annotated gene	516496	73,045	31283	74,414
Number of MGE-PSCCs containing at least one annotated gene	8977	1,270	1786	4,248
Number of MGEs carrying at least one ARG	1570	0,222	106	0,252
Number of MGE-PSCCs carrying at least one ARG	8	0,001	1	0,002
Number of ARGs carried by MGEs	1613	0,228	107	0,255
Number of ARGs carried by unique MGEs	38	0,005	9	0,021
Number of ARGs carried by MGE-PSCCs	8	0,001	1	0,002
Number of ARGs carried by unique MGE-PSCCs	5	0,001	1	0,002
Number of ARGs containing at least one protospacer	3	0	0	0

2.4.5 Community dynamics

The relative abundances of rMAGs and representative iMGEs were used to infer community dynamics over time (**Figure 2.1**, **Figure 2.5**, **Figure C.1**). We grouped rMAGs at the family level due to the large fraction of unclassified taxa. Families such as *Microthrixaceae*, *Moraxellaceae*, *Leptospiraceae* and *Acidimicrobiaceae*, which are known to be present within sludge communities [Muller et al., 2014a; Shchegolkova et al., 2016], were found to be prominent members. To further investigate the effects of iMGEs on the community dynamics, we linked iMGEs to their putative host families based on their assignments via binning. This resulted in a total of 79 family-level groups of bacteria, plasmids and phages. *Microthrixaceae* family showed a relative abundance average of 15.5% (median=15.9%, SD=5.2) with minor fluctuations throughout the time-series, except between 2011-11-16 and 2012-01-03, where there was a significant decrease. *Moraxellaceae* (mean=6.4%, median=3.6%, SD=7.5) and *Leptospiraceae* (mean=6.9%, median=5.9%, SD=6.4) showed relatively low abundances over time but increased with the decline in *Microthrixaceae* (**Figure 2.1**), thereby representing the shift in the community structure.
To further investigate the community dynamics, we defined three overlapping shorter-term intervals according to before, during and after the aforementioned community shift (**Appendix E.6**). Subsequently, correlation between the family-level groups, hierarchical clustering and linear modelling using *Microthrixaceae* family as the response variable were performed for the entire time-series and shorter-term intervals.
The correlation analysis showed 62 pairs of family-level groups consistently exhibiting significant correlations (**Figure 2.6**), whereby 10 families correlated ($r \leq -0.7$ or $r \geq 0.7$, p-value≤ 0.001) with their own plasmids and phages in the entire time-series as well as the shorter-term intervals, e.g. *Microthrixaceae*, *Moraxellaceae* and *Leptospiraceae* (**Appendix B.8**). Hierarchical clustering of correlation values from the entire time-series yielded a total of six clusters, whereby most bacteria, plasmids and phages assigned to the same families clustered together, demonstrating a predictable variation of these family-level groups. Further inspection of the dominant families showed *Microthrixaceae* clustering separately from *Leptospiraceae* and *Moraxellaceae* (**Figure 2.6**, **Figure 2.7** and **Figure C.2**). The latter two clustered together and exhibited significant negative correlation with *Microthrixaceae* (r=-0.63, p-value=8.3×10^{-7}, and r=-0.52, p-value=9.9×10^{-5}, respectively), further supporting their observed acyclical behaviour relative to *Microthrixaceae* (**Figure 2.6**, **Figure 2.7** and **Figure C.2**).
In addition, a selection of the best linear models showed an enrichment of *Microthrix-*

aceae-plasmids, *Acidimicrobiaceae*-phages and *Saprospiraceae*-plasmids groups and, in agreement with the enrichment analysis, the best model (adjusted R^2=0.9983) showed iMGEs from *Microthrixaceae*, *Saprospiraceae* and *Moraxellaceae* families exhibiting significant contributions (**Appendix E.7**). Thus, the longitudinal abundance data for *Microthrixaceae* exhibited good agreement with the models (**Figure 2.8**). Overall, the linear modelling analysis showed the appearance of *Microthrixaceae*-plasmids as the only common significant predictor in all the models (entire time-series and shorter-term intervals). This group was then removed from those models to assess its relative importance, resulting in a significant reduction of predictive power (**Figure C.3**, **Figure C.4**, **Appendix B.9**, **Appendix B.10** and **Appendix E.7**). Thereby, its plasmids had a stronger effect on the prediction of *Microthrixaceae* abundances compared to its phages, indicating a higher relative importance for plasmids in governing *Microthrixaceae* dynamics.

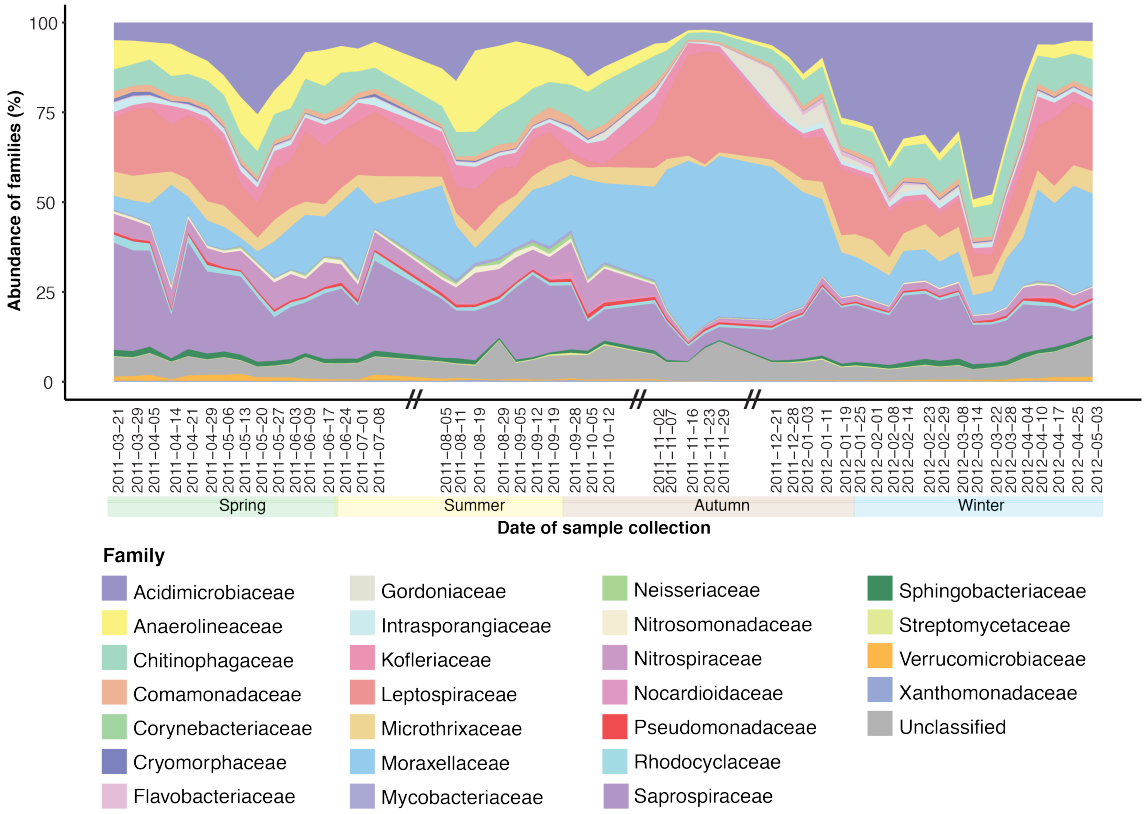


Figure 2.5: Community activity. Relative expression based on mapping MT data to representative metagenomic assembled genomes (rMAGs) over time. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system.

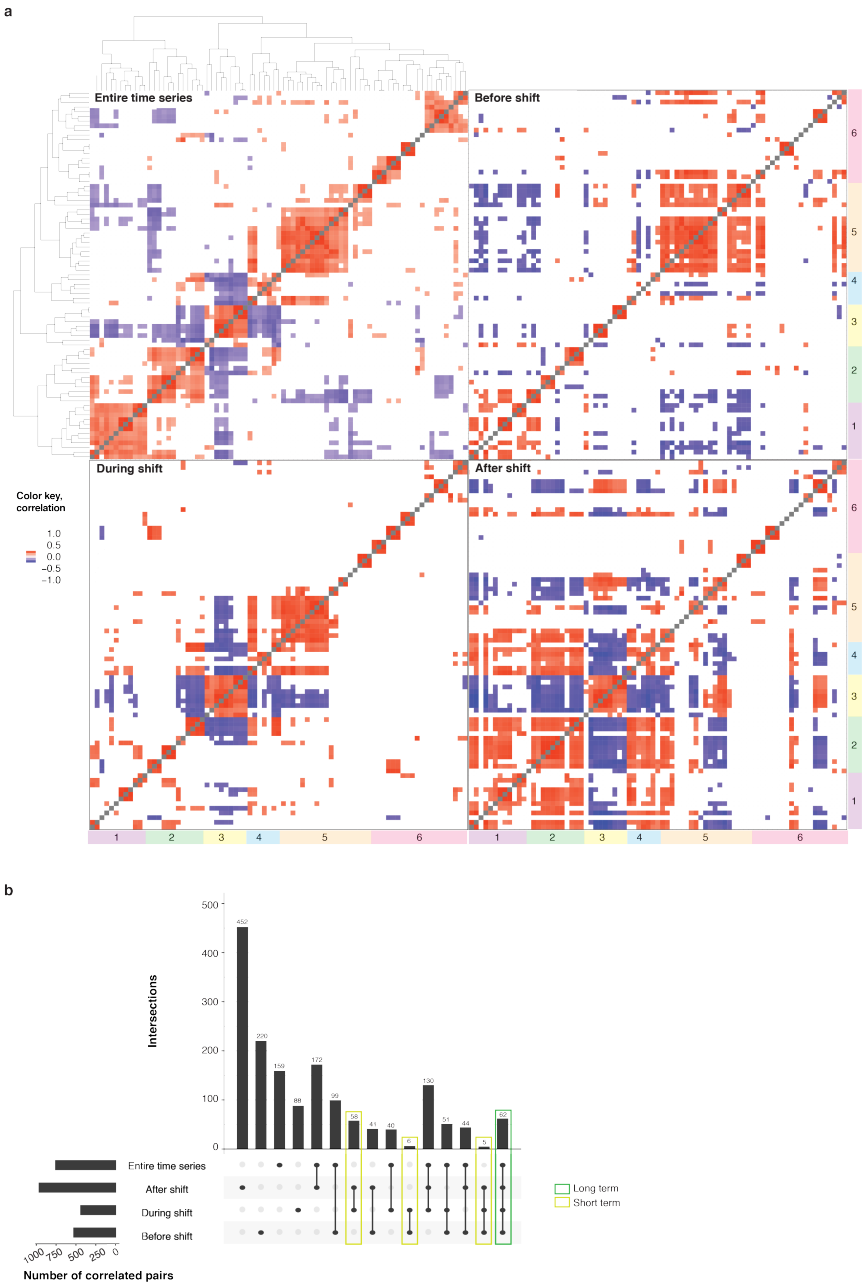


Figure 2.6: Correlation within longer- and shorter-term time intervals. **a**, Correlation heatmaps based on the entire time-series and shorter-term time intervals. The rows and columns preserve the order of the hierarchical clustering from the entire time-series. The coloured strip annotations on the right and bottom represent clusters 1-6. The values within the heatmaps represent significant (threshold: $p \leq 0.001$) correlations. Statistical tests were two-sided and adjusted for multiple comparisons. **b**, Significant correlated pairs within the entire time-series and shorter-term intervals (horizontal bars). Vertical bars represent the number of intersections between those pairs in different time intervals. Coloured boxes represent the intersections of longer-(entire time-series) and shorter-term dynamics, respectively. **Appendix B.8** provides detailed information on the common correlating pairs between the time intervals.

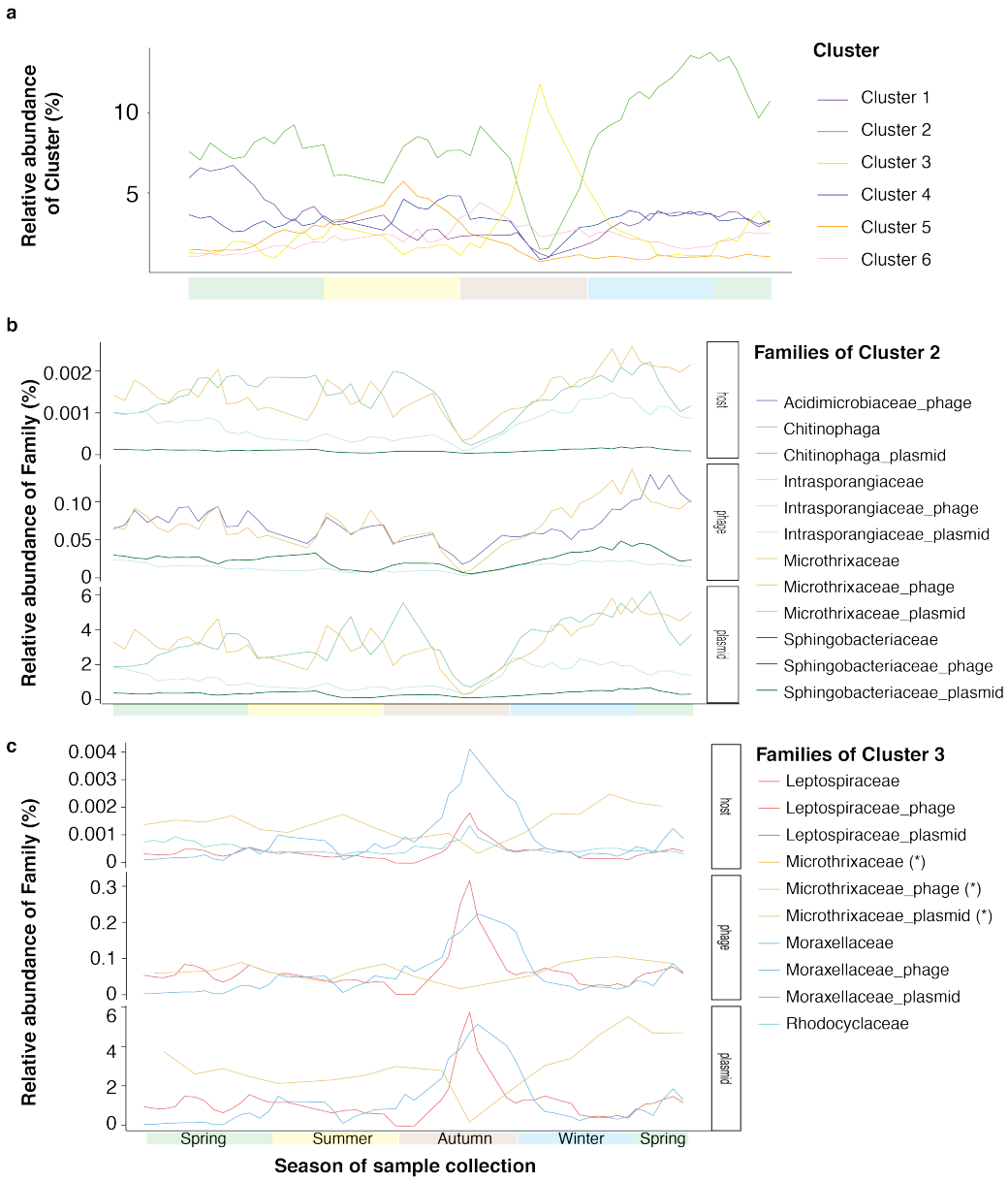


Figure 2.7: The rMAGs were grouped together at the family-level. Plasmids and phages were grouped based on their family-level association, i.e. binned together with an rMAG of a given family. The bacterial, plasmid and phage groups were clustered based on the correlation of their cumulative group-level abundance dynamics. **a**, Dynamics of all clusters based on cumulative abundance of each cluster members. **b**, Dynamics of the cluster 2 members, including *Microthrixaceae* and its associated plasmids and phages as cluster members. **c**, Dynamics of the cluster 3 members, including *Microthrixaceae* and its associated plasmids and phages as reference (these groups are marked with an asterix). Relative abundance values on the y-axis were derived from MG data. The x-axis represents time, colour coded by seasons as labelled in panel **c**. Please refer to **Figure 2.1** for the exact sampling dates within the seasons.

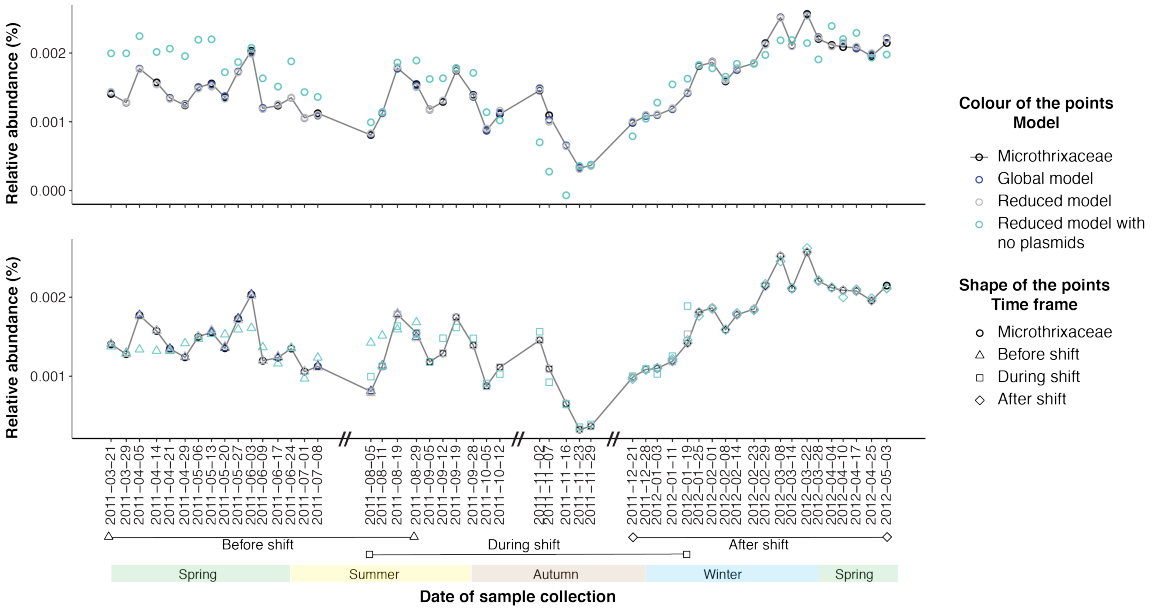


Figure 2.8: Microbial community dynamics. Top: models based on the longer-term dynamics. Bottom: models based on three shorter-term dynamics. The models are based upon the group-level relative abundance values. Longer-term dynamics are represented by all data points from the entire time-series. The shorter-term intervals were defined around the "shift" in community structure, where the abundance of *Microthrixaceae* family decreases drastically. Exact sampling dates of the shorter-term intervals are highlighted in the x-axis. Three models were applied to the longer- and shorter- term time intervals, respectively. The relative abundance of the *Microthrixaceae* family is included for reference.

2.4.6 CRISPR-Cas mediated iMGE-host interactions

To describe CRISPR-mediated interactions between iMGEs and their hosts, we retained 4,985 spacers that were encoded by at least one rMAG (host), co-occurred with its assigned rMAG in at least one timepoint, and targeted at least one iMGE at any given timepoint. We subsequently searched for iMGEs and corresponding spacers newly appearing during the time-series, i.e. spacer integration events, and observed that 2,377 spacers were detected either after or at the same timepoint as their corresponding targeted iMGEs. The mean spacer integration time, i.e. lag time between detection of an iMGE and its corresponding spacer, was 9.5 weeks (median=8, SD=8.5). Spacers which disappeared after the detection of their linked iMGEs were considered to be lost. We observed 1,616 spacers being lost with seven weeks as the average time for such deletions (median=5.5, SD=7.5). Interestingly, the average time for spacer integration and deletion was lower for phages compared to plasmids (**Appendix B.11**). Furthermore, there was a shift from spacer gain to loss on 2011-11-29, suggesting that a majority of integration events occurred during the summer to autumn transition and a majority of deletion events in late autumn, corresponding to the shift in community structure occurring in autumn-winter (**Figure C.5**).

We then separated the CRISPR-mediated interactions into a plasmid-host network comprising 18 hosts and 1,881 plasmids, with 2,274 interactions, and a phage-host network comprising 16 hosts and 472 phages, with 490 interactions (**Figure 2.9**). Additionally, we defined an occurring interaction within a given timepoint if a host and its interacting iMGE were detected in either MG or MT data, resulting in time-resolved network topology variations (**Appendix B.12**, **Appendix E.8**). We included orphan iMGEs and hosts, for which their associated counterparts were not detected within the same timepoint to visualize the dynamics (**Appendix D.1** and **Appendix D.2**).

The time-resolved plasmid-host interaction networks had an average modularity of Q=0.71 (median=0.73, SD=0.07), with two main modules of interactions: a group containing a core set of rMAGs classified as *Leptospira biflexa*, and a group containing rMAGs from different species, i.e. *Marinobacter hydrocarbonoclasticu*, *Acinetobacter* sp. ADP21, *Chitinophaga pinensis* and *Haliscomenobacter hydrossis*. *Ca.* M. parvicella was represented by rMAG-165. In contrast, the phage-host interaction networks had an average modularity of 0.69 (median=0.69, SD=0.07) and smaller interacting groups. However, the overall dynamics of both networks were similar, with the number of interactions increasing during November 2011, which co-occurred with the drop in *Ca.* M. parvicella (*Microthrixaceae*) and the increase in other populations, such as *Leptospira biflexa* or

Haliscomenobacter hydrossis. Based on these networks, we performed a one mode projection to resolve direct interactions between rMAGs with common iMGEs. For this, we observed a higher range of interactions between rMAGs from the plasmid-host network, suggesting a wide spread of plasmids across different families in contrast to the more restricted infection range of phages (**Figure 2.10**, **Appendix B.13**).

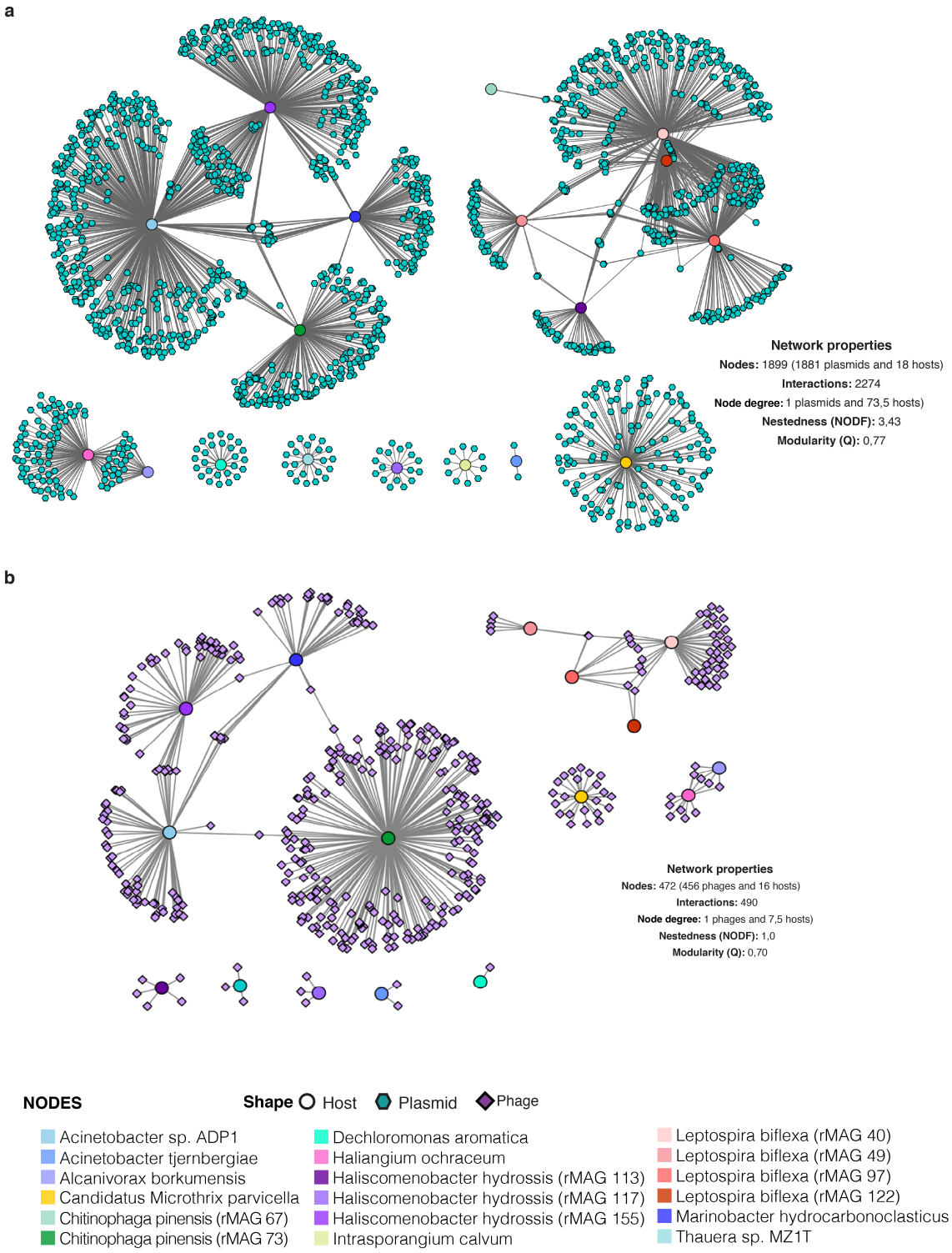


Figure 2.9: Networks of plasmid- and phage-host interactions. Bipartite networks representing global CRISPR-based interactions from the entire time-series involving bacterial hosts (multicolored circular nodes) and their associated **a**, plasmid (turquoise hexagonal nodes) and **b**, phage (purple diamonds) sequences. The edges represent at least one spacer at one timepoint from the host targeting the corresponding iMGE.

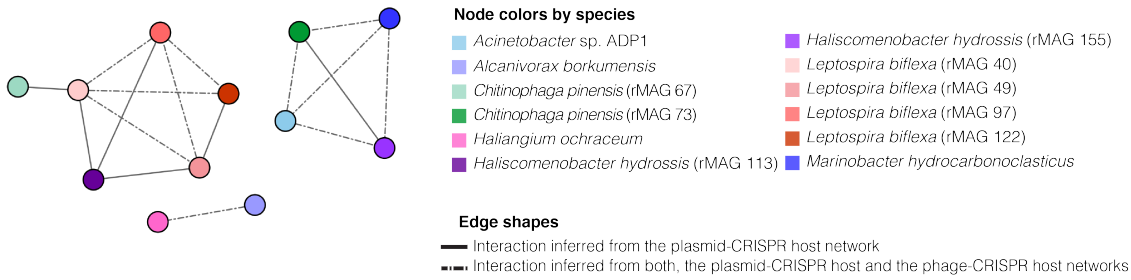


Figure 2.10: One-mode projection of the bipartite iMGE-CRISPR host networks. Interaction network between microbial populations, where the interactions are inferred from the iMGE-CRISPR host interactions.

2.4.7 Population-level iMGE-host dynamics

To further understand the iMGE-host dynamics in relation to the maintenance of microbial populations of interest, we focused on the dominant population within the community, *Ca*. M. parvicella [Blackall et al., 1996; Muller et al., 2012; McIlroy et al., 2013], which constitutes ~30% of the community at specific dates (**Figure 2.1**). More specifically, it showed distinct characteristics in the community and network dynamics, such that timepoints with decreased *Ca*. M. parvicella abundance exhibited a higher number of overall CRISPR-mediated interactions (**Figure 2.11**, **Appendix D.1** and **Appendix D.2**), further supported by the negative correlations with the total number of plasmid-host interactions over time (r=-0.33, p-value=0.017) and phage-host interactions over time (r= -0.40, p-value=0.004). However, upon focusing on the population-level CRISPR-based iMGE-host interactions of *Ca*. M. parvicella, we observed a positive correlation between the population abundance over time and its number of iMGEs-host interactions, i.e. plasmid-host (r=0.63, p-value$\approx$0) and phage-host (r=0.25, p-value=0.02). Finally, the iMGE-*Ca*. M. parvicella network exhibited highly modular structure, whereby a set of iMGEs interacted with its set of spacers (**Figure 2.11**).

We identified a single contig of 10,224 bp in length that encoded a complete CRISPR operon. This contig shared 97.62% sequence identity with *Candidatus* Microthrix parvicella Bio17-1 [Muller et al., 2012] (**Appendix E.9**, **Appendix B.14**). Briefly, the contig contained six *cas* genes and 11 CRISPR-repeats. Using the MT and MP data, we found the *cas* genes within the rMAG to be expressed over time, with Cas2 showing the highest level of gene expression while Cas7 was found more frequently at the protein level (**Figure 2.11**). We were able to link a total of 670 spacers across the entire time-series to this specific CRISPR locus. These spacers were present within an average of 25.5 timepoints (median=28.5, SD=14). Out of all the associated spacers, 433 lacked matches within the time-series and 246 could be linked to a protospacer in at least one timepoint. Among these, 64 targeted plasmids, 24 phages, and 12 both plasmids and phages (**Figure 2.11**). 10 of the 12 spacers targeting both had matches in protein coding genes including sigma 70 factor of RNA polymerase, GDSL-like lipase 2, and helix-turn-helix domain 23, which are genes known to be widely encoded by both plasmids and phages. Additionally, we inspected the spacers activity within the CRISPR loci and observed 45 spacers with gain or loss events (**Figure 2.12**). Similar to the community level, there was also a shift in gain to loss events, occurring after the community shift on 2011-12-28 (**Figure 2.13**). Overall, *cas* gene and Cas protein expression levels, coupled to spacer dynamics targeting more

plasmids (example in **Figure 2.14**) than phages, demonstrated a highly active CRISPR-Cas system within *Ca*. M. parvicella.

In contrast to *Ca*. M. parvicella, other populations exhibited more dynamic CRISPR loci, such as the rMAG-40 classified as *Leptospira biflexa*, and less dynamic loci, such as the rMAG-31 classified as *Intrasporangium calvum* (**Appendix E.10**). *L. biflexa* has eight putative CRISPR loci and a locus of *cas* genes classified as type V (**Appendix B.15**, **Figure 2.15**) and these contained a total of 680 spacers, of which 146 exhibited gain or loss within the time-series. The population with the highest amount of spacers was rMAG-73, classified as *Chitinophaga pinensis*, with CRISPR type III and a total of 1,119 spacers, of which 306 were active (i.e. with either gain or loss events). Overall, the size of the CRISPR locus did not directly relate to spacer gain/loss. Finally, we observed that different population-level CRISPR-Cas dynamics exist at the level of gene and protein expression as well as spacer integration activity. Based on our results, *Ca*. M. parvicella populations contain a functional CRISPR system but use it sparingly compared to other populations.

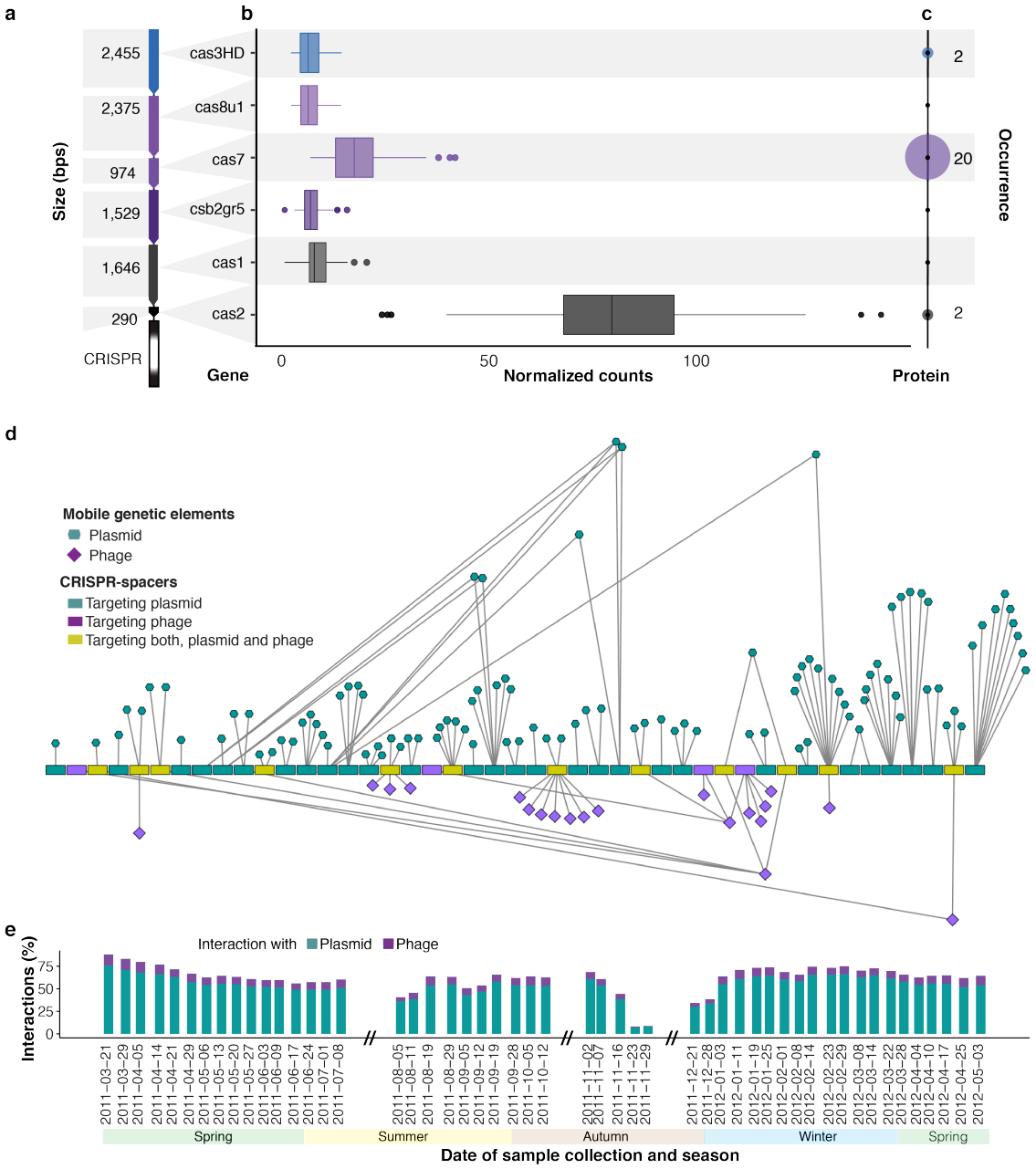


Figure 2.11: The CRISPR-Cas system of *Ca.* M. parvicella. a, CRISPR-cas locus predicted within a reconstructed population-level genome (rMAG-165) identified as *Ca.* M. parvicella. **b**, MT-based expression levels of the corresponding *cas* genes. Boxplot represents expression levels aggregated from 51 timepoints based on normalized read counts. Data are presented as median values, $Q1-1.5\times IQR$ and $Q3+1.5\times IQR$. **c**, MP-level representation of Cas proteins. The numbers represent the number of timepoints where at least one peptide of the corresponding Cas protein was detected. **d**, Representation of the active CRISPR spacers (gain or loss of spacer within the time-series) assigned to *Ca.* M. parvicella. The order of the spacers, is based on their first occurrence within the time-series. **e**, Spacer-iMGE-based interactions represented per timepoint as percentages of the global interactions of *Ca.* M. parvicella.

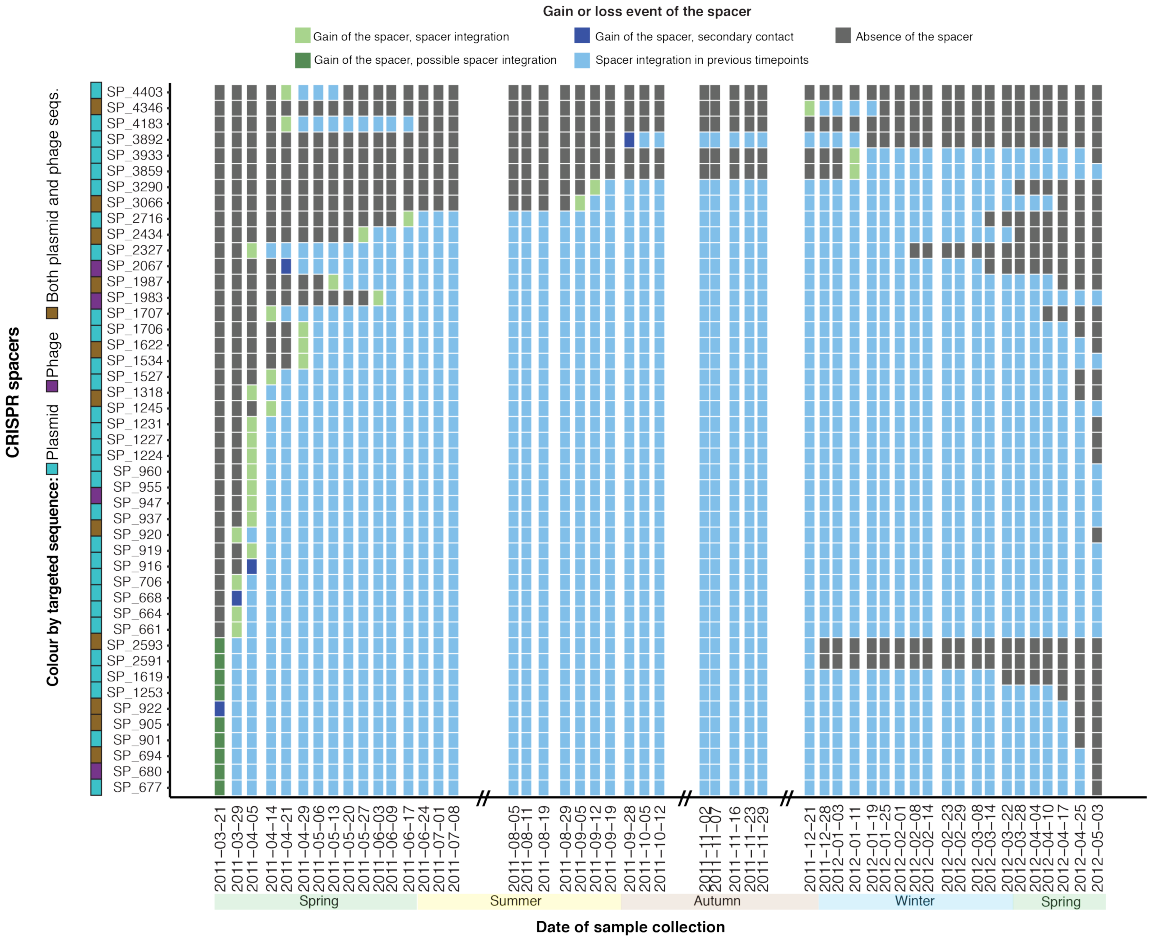


Figure 2.12: Spacer acquisition dynamics in *Candidatus* Microthrix parvicella population. Dynamics of spacers assigned to the *Ca.* M. parvicella population. The y-axis includes the spacer IDs. The coloured boxes next to the spacer IDs indicate the type of iMGE targeted by that spacer, i.e. turquoise for plasmids, purple for phages and brown for both. The boxes within the plot are coloured based on the presence (light blue) or absence (dark grey) of the spacer within the CRISPR array for each timepoint. Green boxes represent spacer gain events, specifically light green for spacer integration (iMGE is detected before the spacer) and dark green for a putative spacer integration (iMGE and spacer are detected at the same timepoint). Dark blue boxes represent potential secondary contact events (spacer detected before the iMGE). The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples from the sampled system.

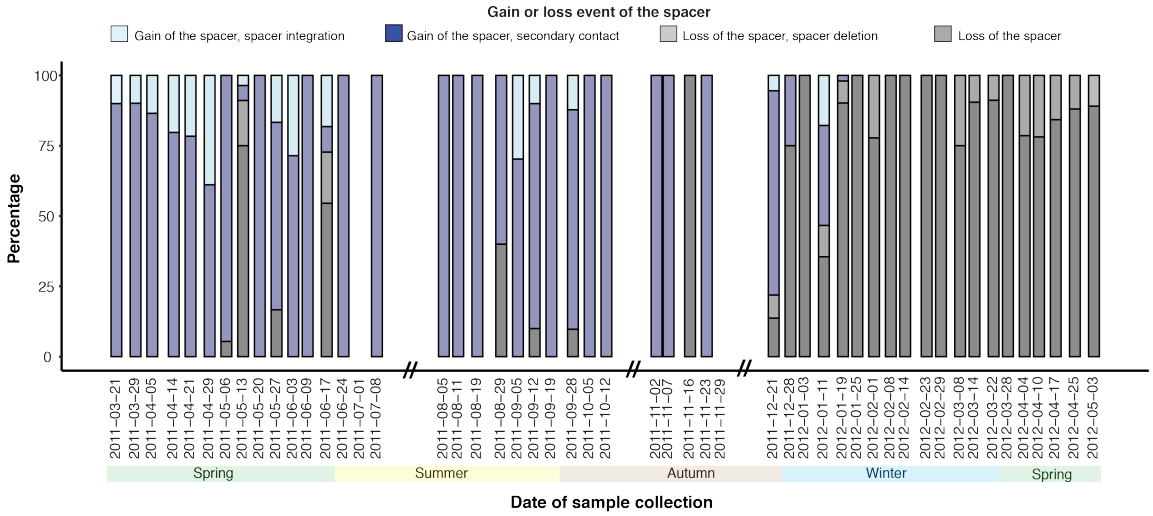


Figure 2.13: Spacer acquisition dynamics in *Candidatus* Microthrix parvicella population. Barplot representing the percentage of spacers per time-point reflecting a gain or loss events. Gain events are defined as: i) "Gain of the spacer, spacer integration" when the iMGE was detected before or at the same timepoint as its linked spacer, and ii) "Gain of the spacer, secondary contact" when the spacer was detected before the linked iMGE within the time-series. Loss events are defined as: i) "Loss of the spacer, spacer deletion" when both the spacer and the iMGE are not detected anymore within the remainder of the time-series, and ii) "Loss of the spacer" when the spacer is not detected within the time-series anymore but the iMGE is still detected after spacer loss. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples from the sampled system.

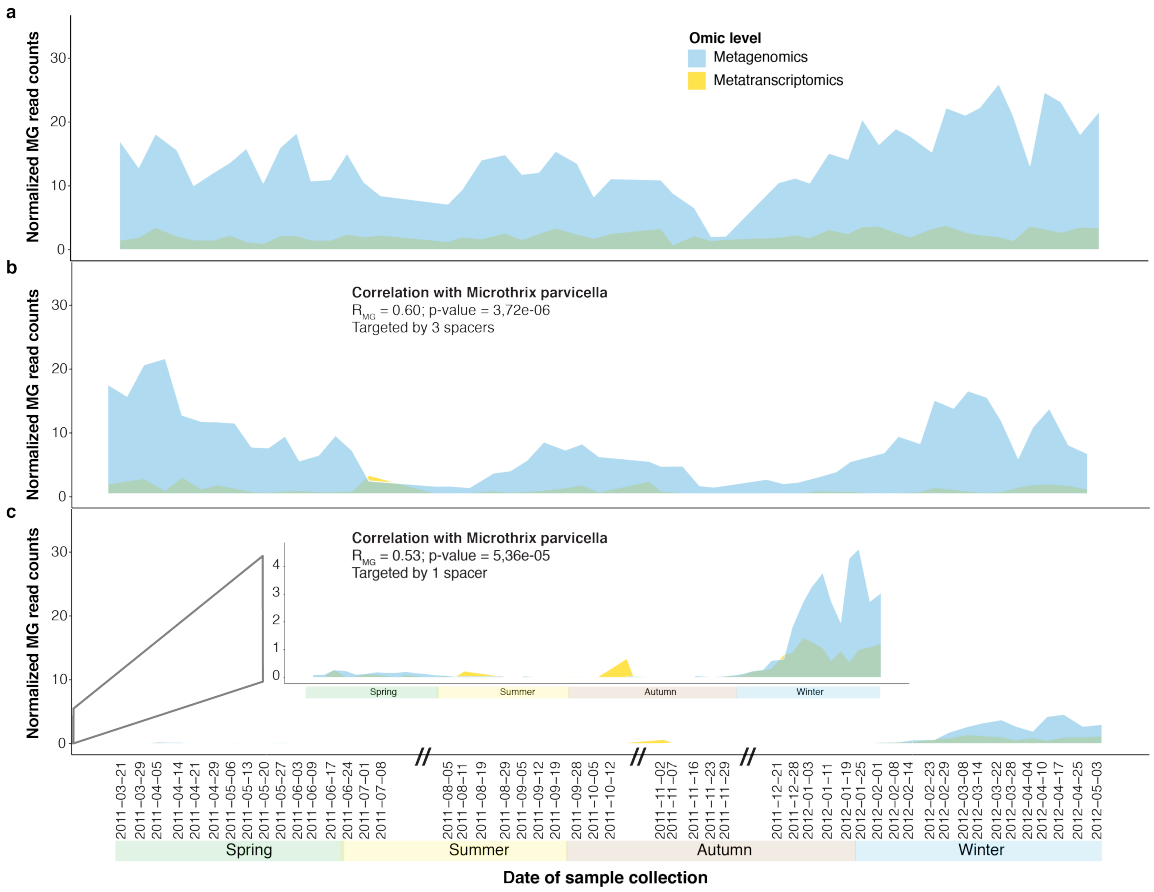


Figure 2.14: Abundance of *Ca*. M. parvicella and selected plasmid sequences targeted by the spacers of the same species. a, Metagenomics (MG)-based and metatranscriptomics-based (MT) abundance of *Ca*. M. parvicella over time. **b**, Abundance of plasmid contig "D28_L2.21_contig_56858", with a size of 2,503 bps which is targeted by three spacers within *Ca*. M. parvicella's CRISPR locus. **c**, Abundance of plasmid contig "D48_E1.25_contig_355826", with a size of16,151 bps which is targeted by one spacer within *Ca*. M. parvicella's CRISPR locus. Statistical tests were two-sided and adjusted for multiple comparisons.

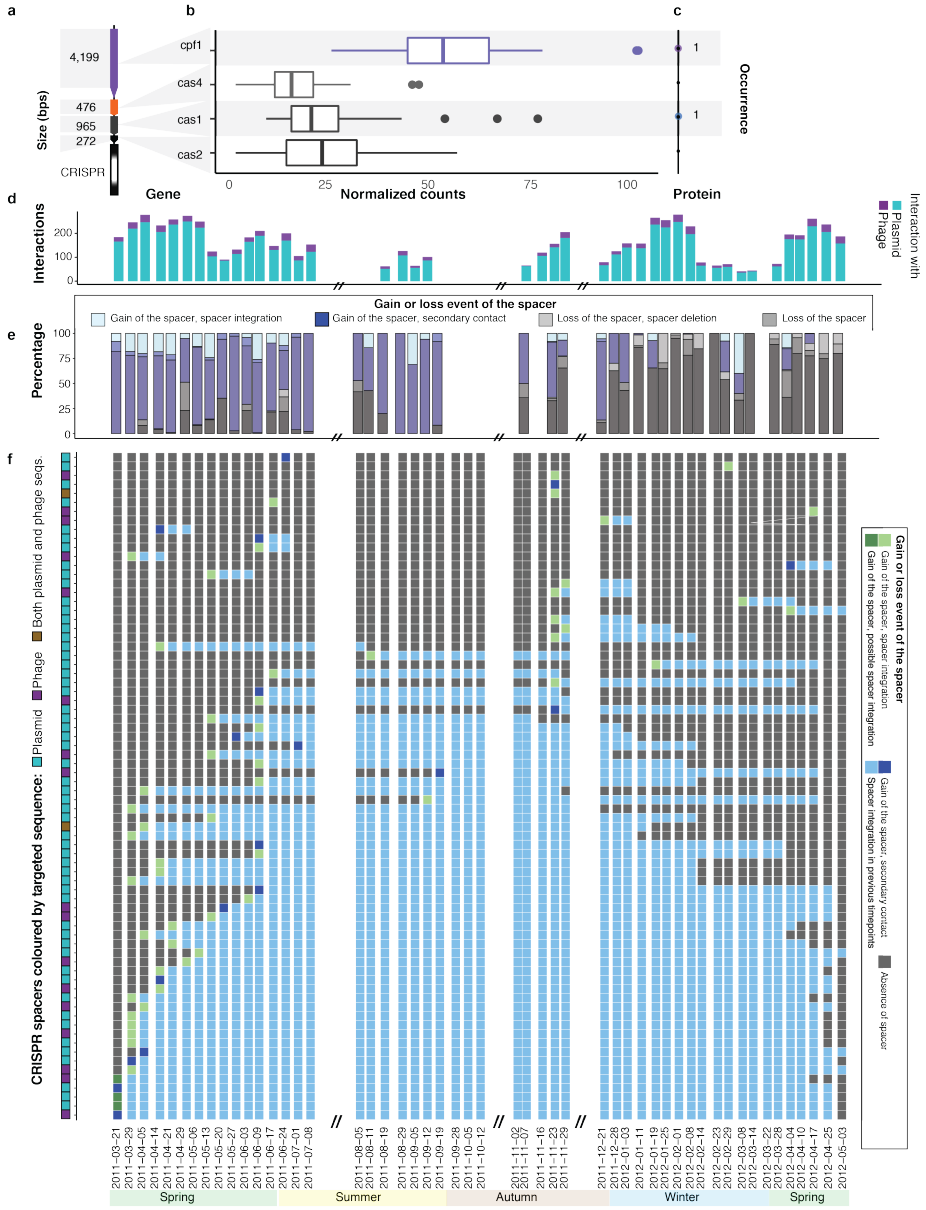


Figure 2.15: Spacers acquisition dynamics of the rMAG-40 population (*Leptospira biflexi*). **a**, CRISPR-Cas operon. **b**, *cas* genes expression at MT-level, aggregated from 51 timepoints. Presented data as median values, $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$. **c**, MP-level representation of Cas proteins; number of timepoints where at least one peptide was detected. **d**, Number of interactions between rMAG-40 and iMGEs. Purple section represents the number of interactions with phages; turquoise represent interactions with plasmids. **e**, Percentage of spacers per timepoint with a gain or loss event. Gain events are defined as: i) "Gain of the spacer, spacer integration", the iMGE was detected before, or at the same timepoint, as its linked spacer, and ii) "Gain of the spacer, secondary contact", the spacer was detected before the linked iMGE, within the time-series. Loss events are defined as: i) "Loss of the spacer, spacer deletion", both spacer and iMGE are not detected within the rest of the time-series, and ii) "Loss of the spacer", the spacer is not detected within the time-series anymore, but the iMGE is detected after spacer loss. **f**, Dynamics of spacers assigned to the rMAGs. Double slashes on x-axis (//) represent absence of samples.

2.5 Discussion

We present an extensive time-resolved integrated meta-omics analysis of CRISPR-mediated iMGE-host interactions. Given the vast extent of unresolved bacterial taxa as well as plasmid and phage sequences in this community, the reliance on existing sequence databases would have greatly limited the analysis of key community members. Our reference-independent approach, including *de novo* genomic assembly, binning and plasmid/phage prediction, were required to analyse this dataset. We were able to link microbial population genomes (rMAGS) to iMGEs using spacer-protospacer links [Edwards et al., 2016], unlike previous approaches that have relied on abundance levels [Brown et al., 2019]. Overall, our approach of resolving interaction dynamics between iMGEs and their hosts revealed an enrichment in CRISPR-based plasmid targeting relative to phages.

To extract coherent information across the time-series, we minimized redundancy concerning the: population-level genomes; CRISPR information; and iMGEs. The aforementioned procedures may potentially result in a dilution of information, especially regarding to the underlying species- and strain-level diversity. However, this trade-off was necessary, considering the inherent properties of the time-series dataset, namely in relation to the appearance, disappearance and/or reappearance of features over time. More importantly, our stringent methodology allowed us to balance the advantages of a *de novo* assembly-based approach, i.e. detecting novel microbial/iMGE populations, while enabling us to track the populations over time.

We systematically optimised the plasmid and phage prediction by applying an ensemble approach to reduce bias stemming from a single tool, establishing associations of iMGEs and specific rMAGs through binning, identifying strong correlations between iMGEs and their associated rMAGs, and using spacer-protospacer links to establish empirical evidence of interaction between rMAGs and iMGEs. Despite this, several limitations must be addressed, including the inherent inaccuracies of the plasmid and phage prediction tools, the inability to predict prophages within community, and the lack of reliable taxonomic classifications of iMGEs.

Our ensemble approach for iMGE identification demonstrated that plasmids are highly abundant within the community. The stepwise linear modelling approach demonstrated that plasmids have more pronounced impact on the dominant *Microthrixaceae* compared to phages. Furthermore, based on the extracted protospacer information, plasmids are also targeted more often than their phage counterparts by CRISPR systems. In contrast to previous studies focused on CRISPR-mediated immunity against phages, our results

thereby support the notion that plasmids also play key roles in adaptation and promotion of diversity [Davison, 1999]. In this context, BWWTPs are thought to be hotspots for the spread of ARGs through MGEs [Rizzo et al., 2013; Li et al., 2018]. Our data revealed a comparatively small fraction of plasmids encoding ARGs being targeted by CRISPR systems, suggesting that bacteria retain potentially beneficial plasmids [Jiang et al., 2013], e.g. those encoding ARGs [Murray et al., 2018], but further detailed investigation including data from longer-term time-series is required.

The period with decreased *Microthrixaceae* abundance (from 2011-11-02 to 2012-01-25) coincided with the increased in abundance of: other families (e.g. *Leptospiraceae* or *Moraxellaceae*); their corresponding plasmids; and overall CRISPR-mediated interactions. Based on this information, the increase in plasmids suggests a short-term fitness advantage for *Leptospiraceae* and *Moraxellaceae* populations on the one hand. On the other hand, CRISPR-mediated links indicate CRISPR-based suppression of those plasmids in a possible drive towards the normalisation of community structure and function, including the dominance of *Ca*. M. parvicella. However, any direct cause-effect relationships remain to be further explored under controlled laboratory conditions.

In relation to phages, we found that they tend to correlate with specific families, e.g. *Moraxellaceae* and *Leptospiraceae*, which exhibit acyclical dynamics in relation to the *Microthrixaceae* family, but showed a smaller effect in the linear models. Additionally, rMAG populations within the *Moraxellaceae* and *Leptospiraceae* families exhibited higher CRISPR activity in terms of phage-linked spacer gain/loss. In that regard, phages are known to affect specific populations, which, according to our data, does not include the dominant *Ca*. M. parvicella, as previously observed by Liu et al. [2017]. Therefore, future studies need to be directed at deciphering the roles of individual plasmids and phages on specific populations, as well as the community as a whole.

Based on our observations, a strong case can be made to include iMGEs and CRISPR-based interactions as additional features into models which incorporate abiotic parameters (e.g. temperature, pH, oxygen concentration, etc.) and biotic drivers (e.g. population dynamics, inter-microbial population interactions, etc.) [Roume et al., 2015; Coenen and Weitz, 2018; Brown et al., 2019], especially when such information can be extracted from metagenomic data. The inclusion of such additional features may provide a more comprehensive model of community dynamics and process performance.

Finally, the composition of CRISPR loci have been found to be highly environment-specific [Kunin et al., 2008], which should translate into environment-specific CRISPR-

mediated interactions. Therefore, the present study should be repeated on samples from other environments to provide a broader understanding of CRISPR-based interactions in relation to iMGEs [Bernheim et al., 2019].

2.6 Conclusions

This Chapter shows the power of longitudinal integrated-omics to provide perspectives of host and iMGE interactions. Empirical CRISPR-based information may be used to validate host-iMGE interactions derived from other analytical methodologies, such as correlation and linear modelling. In the next chapter, we demonstrate how the empirical CRISPR-based interactions devised in the present chapter can be expanded to support general models based on microbial community structure, function and ecological relationships.

Chapter 3

Time-series integrated multi-omic analyses for ecological interactome inference of microbiomes

This chapter is based on the following manuscript in preparation:
Martínez Arbas *et al.* (2021). Time-series integrated multi-omic analyses for ecological interactome inference of microbiomes.

3.1 Abstract

Microbial communities are highly dynamic systems that constantly react and adapt to changing environmental conditions which influence microbial community structure and function. Resilient microbial communities are able to overcome these environmental influences to continue playing their roles within the system. Here, we show which environmental, i.e. abiotic, factors influence the community dynamics in different seasons. We further infer ecological interactions using molecular signatures derived from longitudinal multi-omics data. These signatures are used to model a time-resolved community-wide interactome, which shows cooperation and competition in equivalent proportions, with a low presence of predation. We observe that cooperative interactions are stronger than competitive interactions. From the complete interactome, we further define a core subnetwork, i.e. ecological interactions occurring within the system in several time windows from the timeseries, that may demonstrate a shift in their ecological relationships, e.g. a cooperative relationship between two community members becoming a competitive relationship, likely due to their response to the environment. We further inspected the ecological relationships of the dominant taxa *Candidatus* Microthrix parvicella. In conclusion, we demonstrate the utility of longitudinal multi-omics to elucidate ecological relationships between community members.

3.2 Introduction

Structure and function of microbial communities are driven by the complex web of interactions between the members of the microbiome within a system, i.e. biotic relationships, and the influence of the environmental conditions, i.e. abiotic influence. This complex interactome is key for the resistance and resilience of microbiomes in response to perturbations [Coyte et al., 2015; Jiao et al., 2019]. Resilience is defined as the rate at which microbial populations recover from a perturbation, and resistance is defined by the effect that these perturbations have on the microbial abundances [Clark et al., 2021]. In general, resilience is conditioned upon functional redundancy among community members, when phenotypic traits are widely distributed, e.g. lipid-accumulating organisms within activated sludge are resilient through phenotypic plasticity and niche complementarity [Herold et al., 2020]. Ecological interactions can be i) asymmetric or symmetric, i.e. have effect in one or both of the interacting community members, and ii) have neutral, positive or negative effects [Faust and Raes, 2012; Coyte et al., 2021]. Symmetrical relationships, which affect both of the interacting members, are i) cooperation, i.e. both

interacting members benefit from the interaction, ii) competition, i.e. both members encounter detrimental effects [Guerrero et al., 2011], and iii) predation, i.e. one of the interacting members benefits from the relationship, while the other faces detrimental effects. Asymmetrical relationships, when there is neutral effect for one interacting member, while there is non-neutral effect for the other interacting member, are i) commensalism, i.e. the non-neutral effect is positive, and ii) amensalism, i.e. the non-neutral effect is negative. In response to environmental changes, these relationships may change, e.g. in human gut microbiomes, the presence of antibiotics may lead a shift between cooperative to competitive interactions [Seelbinder et al., 2020].

Mathematical frameworks are a useful methodology to infer ecological interactions and thus can be applied on microbiome data [Dohlman and Shen, 2019]. On one hand, correlation analyses are often applied to infer positive and/or negative pairwise associations [Weiss et al., 2016]. However, the directionality of these associations is not defined. On the other hand, the availability of timeseries (or longitudinal) data allows for the application of suitable algorithms to infer ecological relationships with directionality, to define more complex interactions [Weiss et al., 2016; Dohlman and Shen, 2019]. A notable example frequently used in longitudinal microbiome analyses is the generalized Lotka-Volterra algorithm [Fisher and Mehta, 2014]. Additionally, one can predict the abundance of one population based on a combination of abundances from different populations with regression methods [Trosvik et al., 2015; de Muinck et al., 2017; Silverman et al., 2018]. The direction and sign of the interactions are key to defining the above mentioned ecological relationships (i.e. cooperation, competition, predation, commensalism and amensalism). Although these interactions could be defined mathematically, the underlying biological mechanisms that explain those relationships need to be further interrogated, e.g. i) metabolic models have been widely used to explain cross-feeding [Deines and Bosch, 2016], ii) competition for the same resources [Guerrero et al., 2011], iii) competition between bacteria that are targeted by same invasive mobile genetic elements, such as phages [Refardt, 2011], or iv) cooperation through the transfer of plasmids that carry beneficial genes for their survival [Dimitriu et al., 2014].

Herein, we particularly describe the environmental factors influencing microbiome dynamics within the foaming activated sludge (i.e. model system defined in **Chapter 1** and **Chapter 2**). We then infer ecological networks based on the observed community structure and dynamics over, before, during, and after an observed shift of the community structure (defined in **Chapter 2**, **Section 2.4.2**), which further allowed us to identify core

interactions. Subsequently, we describe putative bio-molecular mechanisms underlying some of those interactions, based on their functional potential and functional expression.

3.3 Material and methods

3.3.1 Generation and analyses of the longitudinal and multi-omics dataset

The integrated longitudinal multi-omic data generation, processing and analyses of the model system are described in detail within **Section 2.3**.

3.3.2 Physico-chemical measurements

During the sample collection, physico-chemical parameters including pH, air and water temperature, conductivity, and oxygen levels were manually measured with a portable field kit (Hach). Additionally, automated measurements were recorded by the biological wastewater treatment plant (BWWTP), including pH, water temperature, oxygen, nitrates, phosphates, and ammonium levels, and dry matter.

3.3.3 Visualization of longitudinal sample trajectory

Non-parametric multidimensional scaling (NMDS) was applied to the normalized MG and MT rMAG abundance tables, with the constraints of the recorded physico-chemical parameters. Missing values of the manual measurements of pH were imputed using the R package "imputeTS" [Moritz and Bartz-Beielstein, 2017]. All analysis and visualizations were performed with R version 4.0.1, specifically the packages "lubridate" [Grolemund and Wickham, 2011], "vegan" [Oksanen et al., 2019] and "ggplot" [Wickham, 2016] were used.

The calculation of the Jensen-Shannon divergence from the rMAG abundances at the MG level was performed using a customised version of the publicly available code from Arumugam et al. [2011] (`https://enterotype.embl.de/enterotypes.html`).

3.3.4 Calculation of correlations

Pearson correlations were performed using the "psych" [Revelle, 2020] R package. To obtain correlation matrices at the MG and MT levels, the input tables were the normalized counts per gene per rMAG over time. The correlations were performed between the normalized gene counts of a given rMAG for timepoint t and timepoint $t + 1$ using the "corr.test" function, with the parameter "BH" (i.e. Benjamini Hochberg) as the adjustment

method for multiple testing. Only correlation values with an adjusted p-value ≤ 0.01 were kept. Finally, t-tests were performed on the correlation matrices to determine whether the mean of the correlation values in a given pair of timepoints t are different from the next pair $t + 1$; compared pairs with adjusted p-value ≤ 0.01 were considered as significantly different.

3.3.5 Modelling and ecological network inference

Centered log-ratio transformations [Aitchison et al., 2000], using the R package "compositions" [Boogaart and Tolosana-Delgado, 2008] were applied to the relative abundance values prior to the mathematical modelling step. To avoid overfitting and obtain a selection of the most significant rMAGs (predictor variables) influencing the abundance of a given rMAG (response variable), models based on linear regression with Ridge and Lasso penalizations, i.e. elastic nets [Friedman et al., 2010], were performed. These models were calculated for each rMAG in four different sets of overlapping time windows consisting of the following timepoints, i) the entire timeseries, i.e. from 2011-03-21 to 2012-05-03, ii) the first 20 timepoints, i.e. from 2011-03-21 to 2011-08-29, iii) 20 timepoints covering the previously observed shift of the community structure (**Section 2.4.2**), i.e from 2011-08-05 to 2011-01-19, and iv) the last 20 timepoints, i.e. from 2011-12-21 to 2012-05-03. The elastic net models were obtained using the "glmnet" [Friedman et al., 2010] and "caret" [Kuhn, 2008] packages in R, to find the penalization parameter values α and λ that calculate the best models. We then extracted the computed features for all the rMAGs from the linear models as depicted in eq. (3.1). Thus, these interactions were i) directed, i.e. the direction of the interaction is given by the predictors and reponse variables, ii) weighted, i.e. the interactions had assigned strength based on the coefficients of the models, and iii) signed, i.e. the effect of the interaction is positive (+) or negative (-), also assigned based on the coefficient of the models, see equation 1. These characteristics were then used to define ecological interactions between the rMAGs, as described by Faust and Raes [2012], i.e. interactions between hypothetical community members X and Y are defined as follows i) amensalism (0/-) - if X has no effect on Y, and Y has a negative effect on X, ii) commensalism (0/+) - if X has no effect on Y, and Y has a positive effect on X, iii) mutualism (+/+) - X and Y have reciprocal positive effects, iv) competition (-/-) - if X and Y have reciprocal negative effects, and v) parasitism or predation (-/+) - if X has positive effect on Y, and Y has a negative effect on X. Subsequently, the models resulted in four networks describing the interactions in the four overlapping time windows respectively.

$$Y_i = \beta_0 + \beta_i X_i + \epsilon \tag{3.1}$$

where X = predictor variable
Y = response variable
β_0 = intercept
β_i = slope
ϵ = random component

3.3.6 Network analysis and visualization

The node degree, i.e. the number of interactions per node, distributions of the inferred networks were then compared to the distributions of three different null network models [Connor et al., 2017]: i) random model based on Erdős–Rényi networks, ii) preferential attachment model based on Barabasi-Albert networks, and iii) stochastic-block models.
Node strength is calculated by first computing the average weight, of the absolute values, of interactions between two rMAGs, e.g. average of absolute values of the coefficients β. This results in the interactions of all rMAGs to be represented by a single edge value, without directionality. Then, these average interaction weights associated to a given node are summed up to obtain a cumulative node strength.
The R "igraph" package [Csardi and Nepusz, 2006] was used to perform network analyses and calculate centrality measures of the networks. To visualize the networks, the R "igraph" package [Csardi and Nepusz, 2006] and Cytoscape [Shannon et al., 2003] version 3.6.1 were used.

3.3.7 Functional annotation and profiling of the rMAGs

Open reading frames (ORFs) of complete genes were predicted using prodigal [Hyatt et al., 2010], and subsequent functional annotation was performed using "hmmsearch" from the HMMER suite [Johnson et al., 2010] against the KEGG HMM database [Kanehisa et al., 2016]. The normalized MG reads mapping to each timepoint (**Section 2.3.6**) were used to compute the average normalized read counts for all timepoints. A binary matrix was formulated based on the presence/absence per gene and per MAG, and then used to calculate the Jaccard distances between the rMAGs with the "vegdist" function of the R "vegan" package [Oksanen et al., 2019]. Then, multidimensional scaling (MDS) of the Jaccard distances was visualized, and k=4 was chosen to perform k-means clustering, using the

function “kmeans” from the R stats package [R Core Team, 2013]. Additionally, the normalized MT reads mappings were used to calculate the realized function of KO functional categories between different clusters and rMAGs, which is the proportion of genes of a given function that are expressed.

3.3.8 Code availability

The R code to perform the above described analyses and visualizations are available at `https://github.com/susmarb/Ecological-interactome-preliminary-`.

3.4 Results

3.4.1 Sample collection, multi-omic readouts and environmental measurements of the model system: foaming islets of activated sludge

Foaming sludge islets were sampled weekly over 1.5 years (**Chapter 2**, **Section 2.3.1**). **Figure 3.1** shows the amount of foaming sludge islets occurring on the surface of the anoxic tank of the Schifflange communal wastewater treatment plant, corresponding to the samples described in **Section 2.3.1**. The amount and appearance of the foam varied over time. Specifically, the foaming sludge appears almost nonexistent during the autumn season, while during winter a continuous layer of sludge is formed. In general, spring and summer exhibited amounts of sludge between the extremes in autumn and winter, with clear islets forming on the surface of the anoxic tank wastewater. Briefly, (and as described in **Chapter 2**), each sample was subjected to a concomitant biomolecular extraction of DNA, RNA, proteins and metabolites and posterior high-throughput measurements, resulting in a longitudinal and multi-omics dataset (**Section 2.3.1**). In this chapter, we further complemented the multi-omic longitudinal data with detailed physico-chemical parameters, i.e. abiotic factors. Specifically, we recorded two sets of physico-chemical parameters; manual and automated. On one hand, manually recorded physico-chemical measurements were carried out as part of the sampling procedure (**Figure 3.2**), which reflect the environmental conditions of the system at the time of sampling. On the other hand, automated physico-chemical measurements, e.g. organic compounds, were recorded as part of the routine operation of the wastewater treatment plant. These automated readouts, which are observed over a longer period of time (**Figure 3.2**), further depicting environmental seasonality. These environmental changes have been shown to af-fect the performance of the water treatment, e.g. the nitrification process [Johnston et al., 2019].

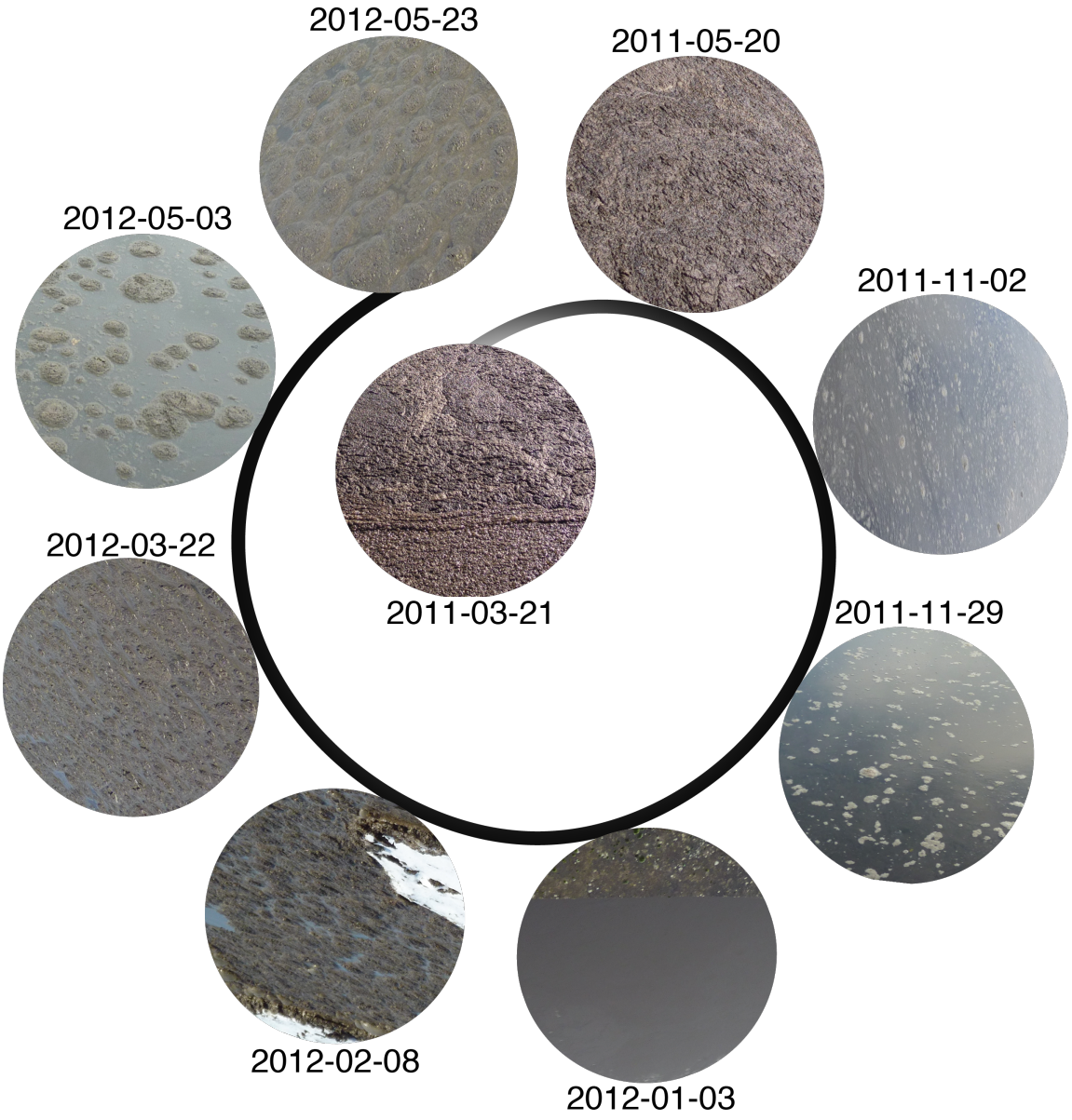


Figure 3.1: Foaming activated sludge islets. Pictures of the Schifflange communal wastewater treatment plant anoxic tank surface, located in Luxembourg, where the samples were collected.

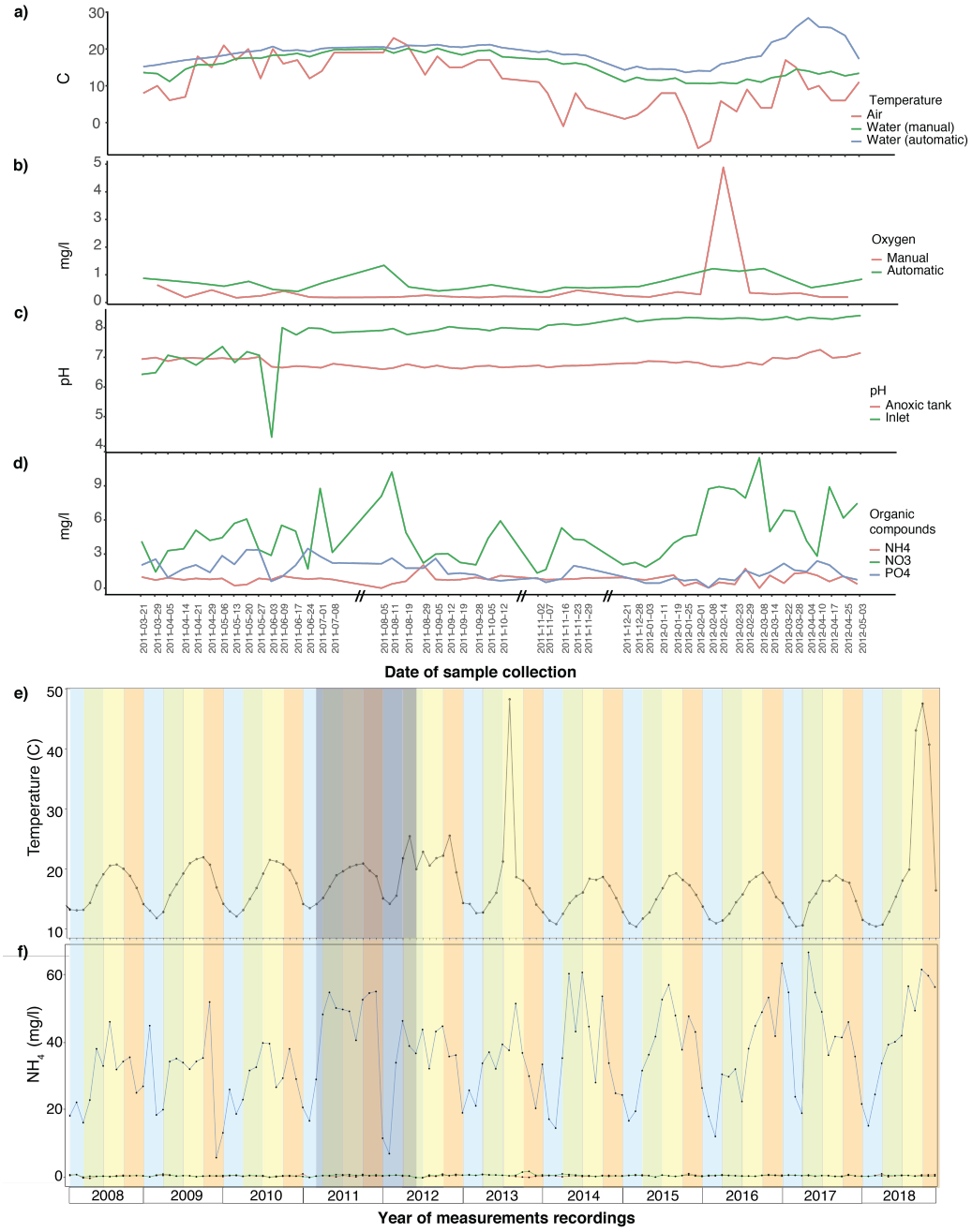


Figure 3.2: Physico-chemical parameters measured at the activated sludge wastewater treatment plant. a-d, Physico-chemical measurements of the anoxic tank during the weekly sample collection period spanning 1.5 years; **a**, air and water temperature, **b**, dissolved oxygen, **c**, water pH of inflow and anoxic tank, **d**, and concentration of organic compounds. Monthly average of daily monitored **e**, water temperature and **f**, NH4 from 2008 to 2018. **a-b**, "manual" and "automatic" refer to measurements by sample collector and automated monitoring sensors of the wastewater treatment plant, respectively. If manual or automatic are not specified, the measurement was then taken from the automated monitoring sensors. **e-f**, seasonal coloured backgrounds, i.e. blue represents winter, green spring, yellow summer and orange autumn. Grey box represents sampling period. Double slashes (//) on x-axis represent absence of samples/measurements.

3.4.2 Microbial community dynamics

Microbial community structure and functional dynamics were defined by reconstructing microbial genomes, i.e. representative metagenome-assembled genomes, or rMAGs, and estimating their genomic abundance (MG) and functional expression (MT) (**Section 2.3.6**). We observed fluctuations of the community structure (**Section 2.4.2**) with a striking shift over autumn (**Figure 2.1**) and an apparent posterior community structure recovery during winter. Similarly, the dynamics of the community expressed function (**Figure 2.5**) exhibits fluctuations alongside the community structure, including the drastic shift in microbial abundances during autumn. Additionally, constrained ordination analysis shows that temperature, dissolved oxygen and the pH of the incoming wastewater (i.e. inflow) are the abiotic factors that influence the trajectory of the community structure and expressed function the most (**Figure 3.3**). This is in line with previous knowledge that highlights how inflow composition and environmental conditions influence wastewater treatment operations [Liu et al., 2016]. The sample trajectory in **Figure 3.3** (at both MG and MT level) shows a few samples in autumn that are somewhat isolated from the rest of the samples, possibly indicating cyclic behavior. This apparent cyclic behavior and trajectory based on the omic profiles correspond to the varying amount of foaming sludge during the different seasons (**Figure 3.1**), which have been previously observed in activated sludge foaming [Wang et al., 2016; Johnston and Behrens, 2020].

Additionally, Jensen-Shannon distances between all the samples were calculated to observe the change in community diversity over time (**Figure 3.4**). An additional ordination analysis on the calculated Jensen-Shannon indices further supports the cyclical behavior observed based on the community structure (**Figure 3.3**), whereby the samples from autumn exhibited higher variation (JSD $\geq$ 2.5) relative to the other samples (JSD $<$ 2.5) (**Figure 3.4**), further supporting the notion of a system perturbation from which the community recovered afterwards.

Another indicator that is commonly used to describe system dynamics that may lead to alternative states is auto-correlation [Scheffer et al., 2009; Liu et al., 2015b; Dogra et al., 2020], i.e. correlations of consecutive timepoints. Here, we calculated the auto-correlation based on the gene level functional potential and expression of the rMAGs (**Figure 3.5**), and then computed the significance of those auto-correlation values. On one hand, the observed average trend of functional potential auto-correlations is R $\geq$ 0.7, with a large fluctuation corresponding to the community structure shift. On the other hand, the observed trend of the functional expression auto-correlations is also high, but it exhibits more fre-

quent large fluctuations compared to the functional potential auto-correlations. We found that the auto-correlation on the MG level is significantly different (p-value ≤ 0.01 of t-test) only in December-2011, immediately after the community shift. However, for the MT level significant differences were indicated in seven different instances covering April, July, August, November and December, with the latter coinciding with the significant difference in the aforementioned significant MG level auto-correlation of December-2011 (**Figure 3.5**). It is worth noting that the significant auto-correlation values in November to December, covers the community shift, indicating expression level dynamics during the community shift event. These observed short-term perturbations might be representing seasonal changes as described in more detail by Herold et al. [2020].

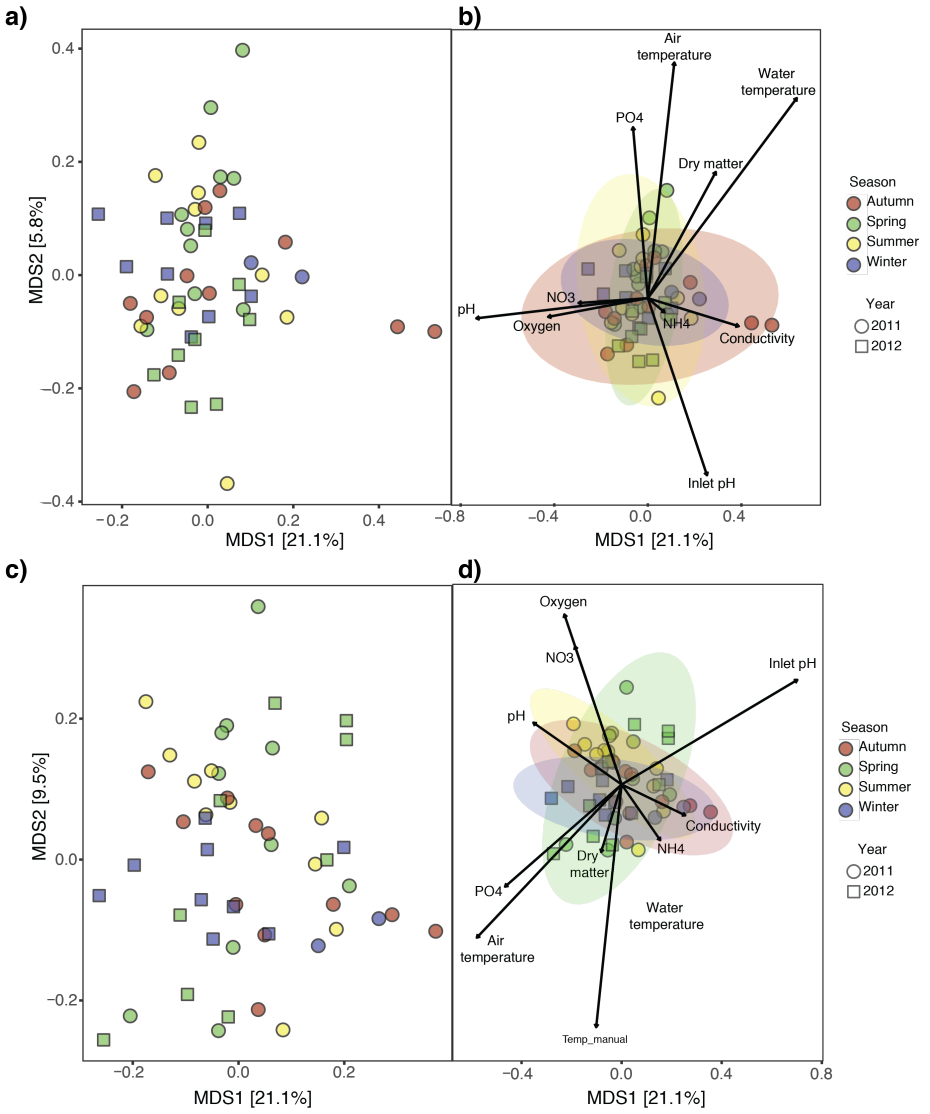


Figure 3.3: Ordination plots of community structure and functional profiles over time. Bray-Curtis dissimilarity ordination of relative abundances on the **a-b** metagenomics and **c-d** metatranscriptomics levels of individual rMAGs constrained by environmental conditions. **a** and **c** show the ordination of the samples, and **b** and **d** display direction and magnitude of influence of physico-chemical parameters relative to the trajectory of the samples.

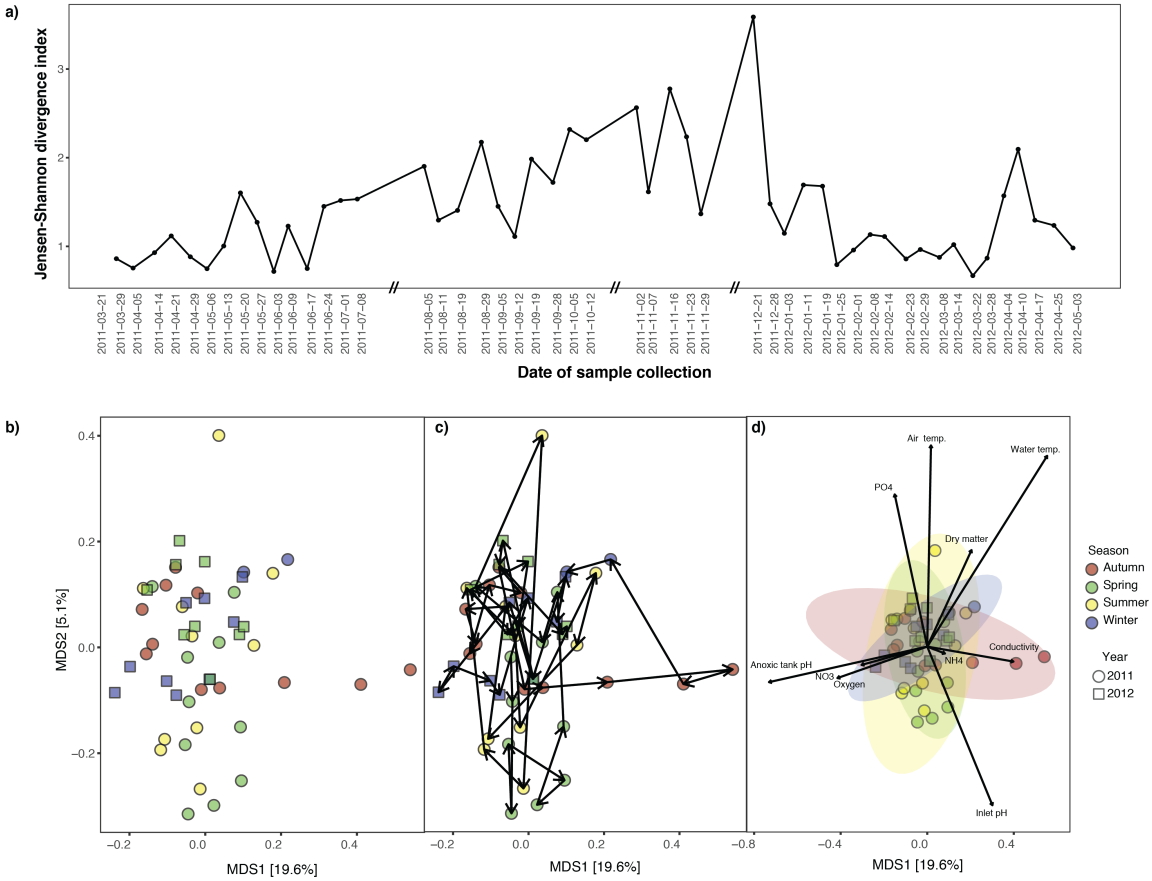


Figure 3.4: Community diversity dynamics. **a**, Jensen-Shannon divergence between consecutive time-points. **b-d**, Bray-Curtis dissimilarity ordination of the Jensen-Shannon distances between all samples including **b**, sample ordination, **c**, sample trajectory displayed as arrows, and **d**, direction and magnitude of physico-chemical parameter influence indicated as arrows. The labels on the x-axis indicate the sampling dates while the double slashes (//) represent absence of samples.

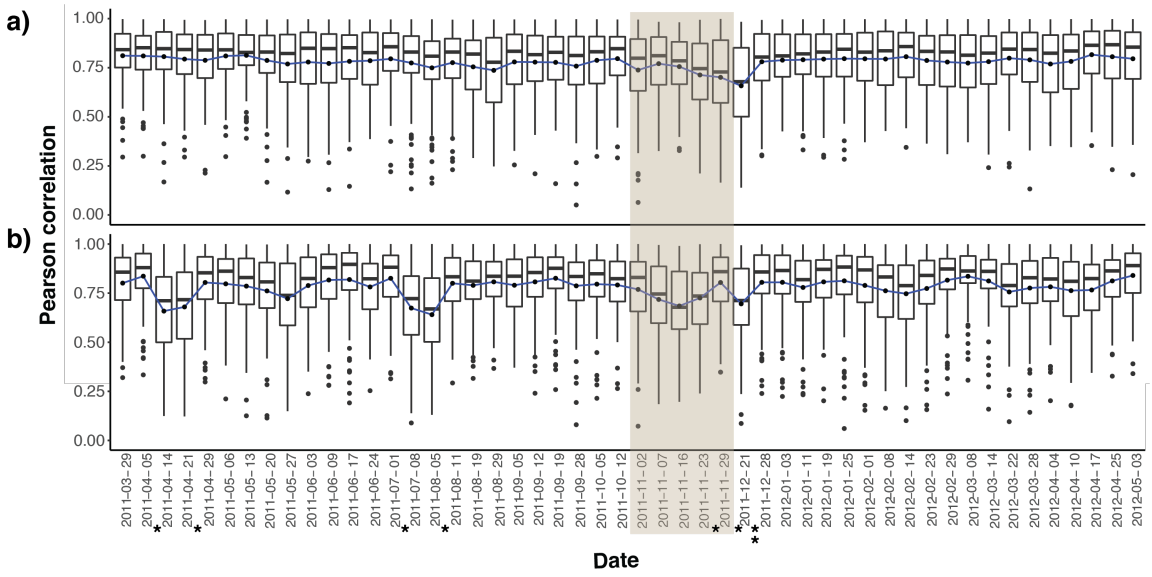


Figure 3.5: Autocorrelation of gene abundance and expression per representative MAG over time. a, Pearson correlations on the metagenomics level. **b**, Pearson correlations on the metatranscriptomics level. Correlation values are shown as absolute values, i.e. negative values are computed as positive values. The asterisks (*) on the x-axis, under the dates, refer to timepoint comparisons that were significantly different to the next timepoint comparison (t-test with p-value$\leq$0.05), where bottom asterisks corresponds to MG level significance, and top asterisks to MT significance.

3.4.3 Mathematical framework for defining ecological interactions and network

Inter-population ecological interactions were inferred to elucidate how the community members interact with each other. Briefly, those ecological interactions between community members such as cooperation, commensalism, competition, amensalism and predation were defined using a linear model based mathematical framework (**Section 3.3.5**). Given the nature of microbiome data such as sparsity and compositionality, a centered log-ratio transformation was performed to the rMAGs relative abundance data [Aitchison et al., 2000]. Furthermore, the high dimensionality of the data results in a higher number of variables (i.e. rMAGs and environmental parameters) compared to number of observations (i.e. timepoints). Therefore, elastic nets were utilised to avoid potential overfitting of the models. The elastic nets consist of linear models implemented with Ridge and Lasso regularization methods [Zou and Hastie, 2005], and aim to predict the response variable, i.e. transformed abundance of a selected rMAG, by selecting the most important group of predictors, i.e. all rMAGs other than the response variable and/or environmental parameters. The resulting models provide direction, weight and sign to the interactions between community members, i.e. between the rMAGs, (**Section 3.3.5**).

Elastic net modelling was applied separately to the entire timeseries and three additional overlapping time windows (defined in **Section 3.3.5** and **Figure 2.8**). The degree distributions of the inferred ecological networks were then compared to three degree distributions of null models, i.e. Erdős–Rényi random model, Barabasi-Albert preferential attachment model and stochastic-block model (**Figure 3.6**). We observed that the inferred ecological networks have a degree distribution trend between the preferential attachment and the stochastic-block models.

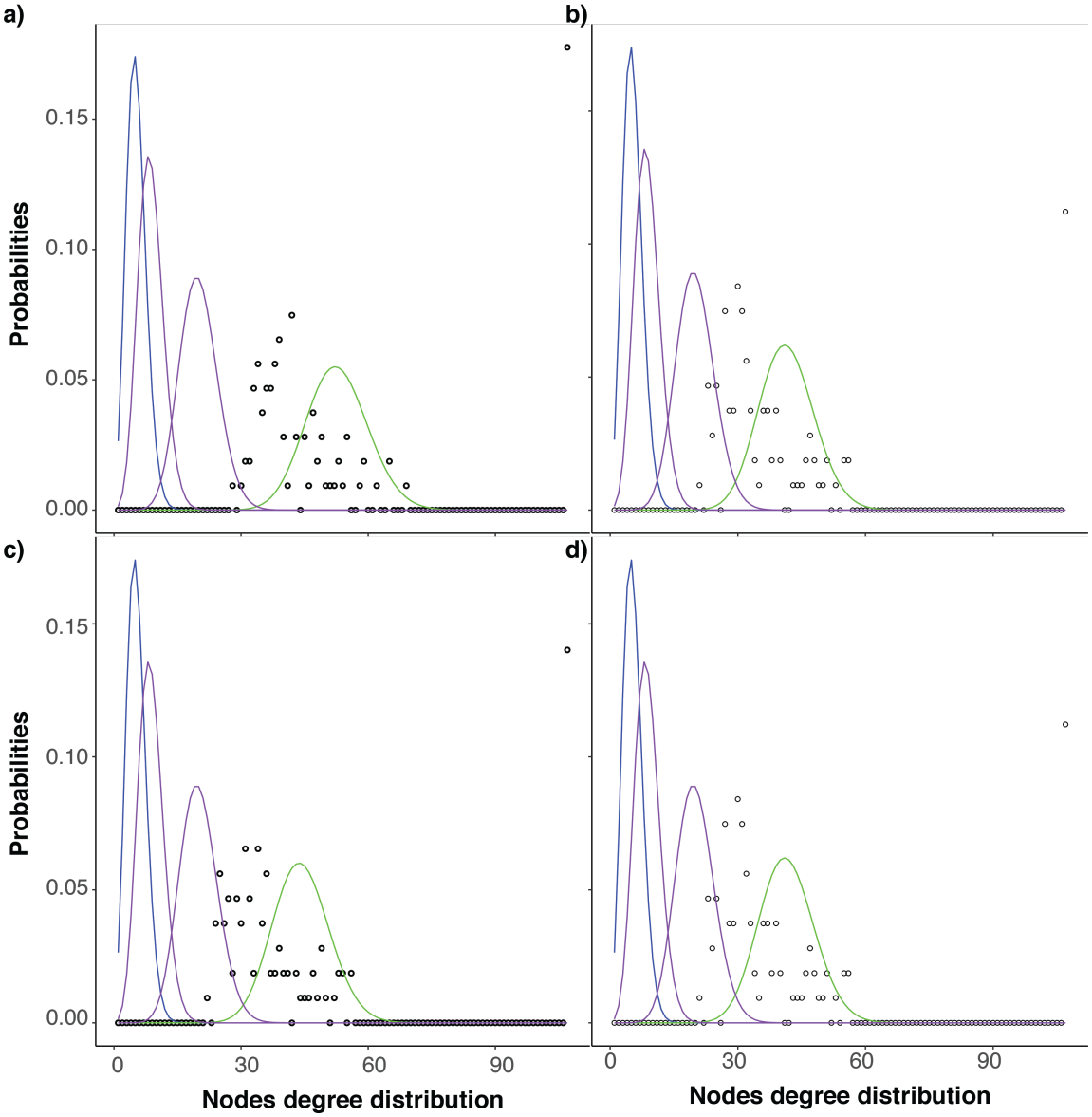


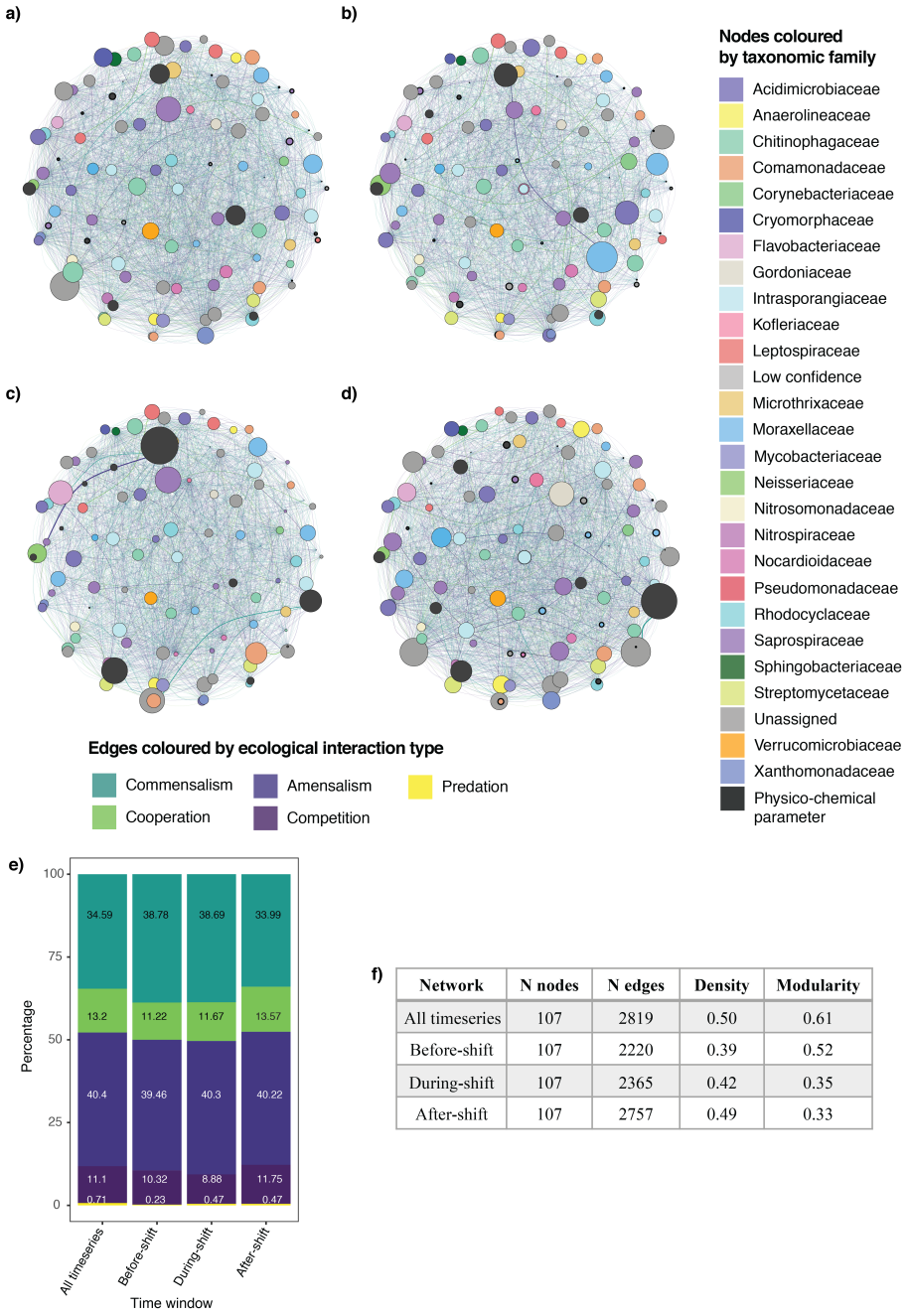
Figure 3.6: Comparison of ecological networks degrees with null models. Network degree distribution of the ecological network inferred from different overlapping time windows, including: **a**, entire timeseries data, **b**, before community structure shift, **c**, during-shift and, **d**, after-shift. The overlapping time windows are defined in **Section 3.3.5**. **a-d**, The points represent the observed degree distributions, and the coloured lines represent the trends of calculated degree distributions for networks representing three null models, i.e. Erdős–Rényi random model (blue line), Barabasi-Albert deterministic network of preferential attachment (green line), and stochastic-block model (purple lines).

3.4.4 Timeframe-specific ecological networks capture different dynamics

Each ecological network consisted of 107 nodes, which included both rMAGs and physico-chemical parameters (**Figure 3.7**), i.e. we treated both biotic and abiotic features as network nodes. There were 11,449 possible interactions per network, and we obtained an average of 2,540 interactions (MEDIAN=2,561, SD=293), whereby the average network density (i.e. the proportion of realized interactions among the potential interactions) was 0.465 (MEDIAN=0.495, SD=0.078). All the four networks collectively represented 4,795 unique interacting pairs, from which eight were predation interactions, 3,833 cooperative interactions, i.e. cooperation and commensalism, 641 competitive interactions, i.e. competition and amensalism, and 313 interactions changing their interaction type within the different time windows, e.g. two rMAGs were predicted as cooperative in one time window and competitive in any other.

The models representing the different overlapping time windows, which were built from different sets of timepoints (**Section 3.3.5** and **Figure 2.8**) captured interactions specific to a certain time window, e.g. interactions occurring exclusively during the community shift as a response to environmental changes and the growth spike of microorganisms that were lowly abundant just before the shift. We then computed the node strength, i.e. cumulative strength of weighted interactions per node, of the different time windows, which further reflected the differential node strength distributions (**Figure 3.8**) between the networks. Significant differences were observed between the entire timeseries and after-shift models (t-test p-value ≤ 0.05).

We then selected the top 20 rMAGs from each of the four networks with the highest node strengths(**Appendix B.16**) and defined them as the rMAGs with the most relative importance. Their comparison revealed that the importance of these rMAGs were highly specific to each network, represented by the little-to-no overlaps between the networks of different time windows (**Figure 3.8**). At the family level taxonomic classification, 66% of the total identified families were represented in the aforementioned important nodes (**Appendix B.17**). We then defined three core families that contained the most important rMAGs from all the four networks, which were *Saprospiraceae*, *Chitinophagaceae*, and *Moraxellaceae*, while *Microthrixaceae* was the taxa of highly important nodes in the entire timeseries network. Additionally, the physico-chemical parameters that appeared as important nodes within the networks included pH, water temperature and organic compounds, although their relative importance differ between networks.



f)

Network	N nodes	N edges	Density	Modularity
All timeseries	107	2819	0.50	0.61
Before-shift	107	2220	0.39	0.52
During-shift	107	2365	0.42	0.35
After-shift	107	2757	0.49	0.33

Figure 3.7: Inferred ecological networks. Each network represents the ecological relationships inferred from different time windows, i.e. **a**, the entire timeseries data, **b**, before, **c**, during, and **d**, after the community structure shift. **d**, Percentage of ecological interactions type per network. **e**, Summary of network properties. The node size is based on the node strength, i.e. the cumulative weight of interactions per node. The edge colour is based on the type of ecological interactions.

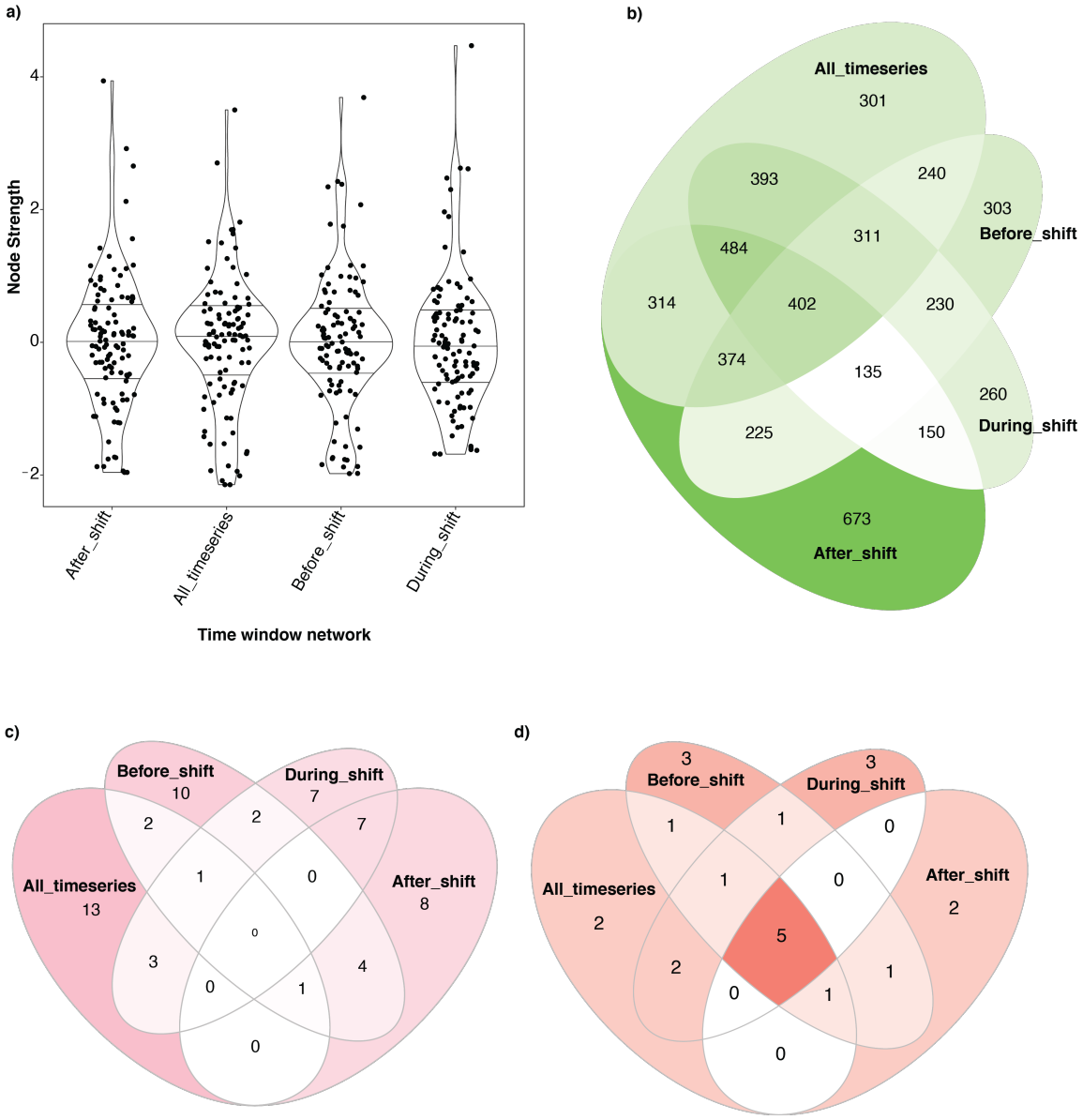


Figure 3.8: Summary of selected network features. **a**, Distribution of node strengths, i.e. strength of weighted interactions per node in the different time windows. **b**, Common and unique interactions between nodes (i.e. representative MAGs at species level) within the inferred ecological networks within different time windows. **c**, Top 20 common and unique nodes represented in **b**, of each overlapping time window network based on the highest node strength values. **d**, Taxonomic families of the top 20 nodes represented in **c**. Labels of "All_timeseries", "Before_shift", "During-shift", and "After_shift" refer to the four different time windows used to construct the networks and are defined in **Section 3.3.5**.

3.4.5 Cooperative relationships between populations are stronger than competitive relationships

In general, the distribution of ecological interactions within all the networks were categorized into two groups with similar distributions of ecological interactions: i) the entire timeseries and the after-shift time windows with ~37% of cooperative interactions (cooperation + commensalism), and ~52% of competitive interactions (competition + amensalism), and ii) the networks before- and during- shift with ~50% cooperative and competitive interactions (**Figure 3.7**). However, regardless of the distribution of cooperative/competitive interactions, the average strength of the cooperative interactions was significantly higher (Wilcoxon's test with p-value ≤ 0.05) compared to the strength of the competitive interactions in all the networks (**Appendix B.18**). Finally, predation interactions were higher within the entire timeseries network, and lower within the overlapping window networks, but in all the cases, they represented less than 1% of interactions, while their strength was lower than any other type of ecological interaction (**Appendix B.18**).

Based on family-level taxonomic classification (**Figure 3.9**), we found that the families *Corynebacteriaceae* and *Nitrosomonadaceae* were the ones with the highest percentage of cooperative interactions (above 50% of their interactions), followed by *Anaerolineaceae* and *Streptomycetaceae*. On the other hand, *Flavobacteriaceae*, *Moraxellaceae*, and *Nocardiaceae* were the families with the highest percentages of competitive interactions (above 50%). Notably, the highest percentage of predation interactions were also observed in the *Corynebacteriaceae* family (**Figure 3.9** and **Appendix B.19**). Additionally, the average strength of intra-family interactions exceeded the inter-taxa interactions (Wilcoxon's test with p-value ≤ 0.05) in all the networks (**Appendix B.20**). Finally, intra- and inter-taxa cooperative interactions were stronger than competitive interactions.

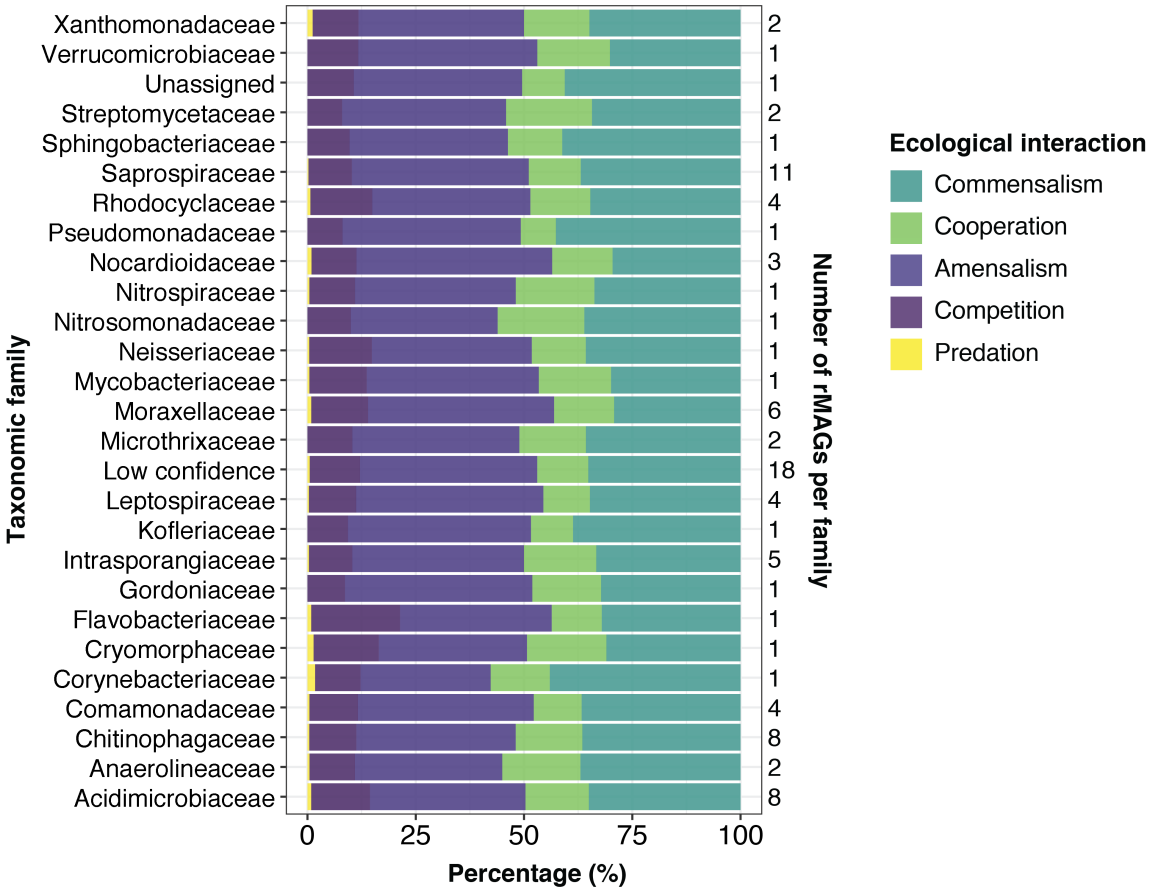
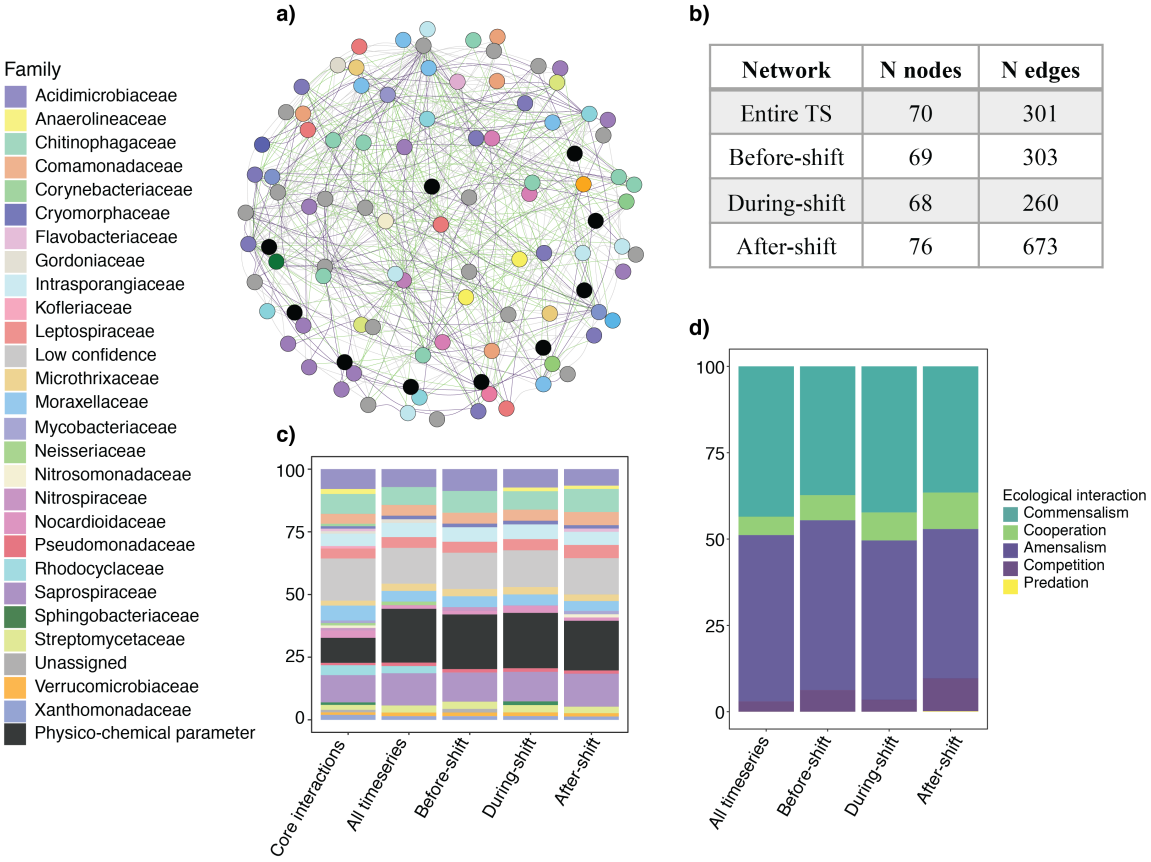


Figure 3.9: Type of ecological interactions per taxonomic family. The cumulative ecological relationships per family based on all inferred ecological networks from the different time windows, i.e. all timeseries, before-shift, during-shift, and after-shift ecological networks. The taxonomic family "Low confidence" refeers to unclassified rMAGs.

3.4.6 Core and unique subnetworks

In this work, we defined two distinct types of subnetworks, namely i) the core subnetwork and ii) the unique time window subnetworks, which covered different subsets of interactions. The core subnetwork was made up of ecological interactions found in all the different time window networks, and there were in total 402 core interactions (**Figure 3.10**). We further defined the following ecological interaction types: i) 134 stable cooperative interactions, i.e. cooperation or commensalism, ii) 95 stable competitive interactions, i.e. competition or amensalism, and iii) 173 dynamic interactions that switched from competitive to cooperative interactions or vice versa within the core subnetwork. Amongst those aforementioned stable and dynamic interactions, we found that predation interactions were always dynamic, i.e. not maintained in all the time windows.

The unique time window subnetworks, on the other hand, were defined by the ecological interactions found exclusively (or uniquely) within a given time window. Consequently, there are 301, 303, 260, and 673 interactions within the entire timeseries, before-, during- and after- shift unique time window subnetworks, respectively (**Figure 3.10**). The distribution of interactions within those unique subnetworks varied from their complete networks (**Section 3.4.4**), such that the i) the entire timeseries and the during-shift unique subnetworks were made up of ~50% percent of cooperative and ~50% competitive interactions, while ii) the before- and after-shift unique subnetworks showed higher percentage of competitive interactions (~55%) than cooperative interactions (~45%).
Finally, all the taxonomic families were involved in both core and unique time window subnetworks while the influence of physico-chemical parameters was lower in the core subnetwork, relative to the unique time window subnetworks (**Figure 3.10**).



Network	N nodes	N edges
Entire TS	70	301
Before-shift	69	303
During-shift	68	260
After-shift	76	673

Figure 3.10: Subnetworks information. **a**, Network representing the core interactions, i.e. those present in all the four time windows networks. **b**, Summary of the subnetworks representing unique interactions from each of the four time window networks. **c**, Percentage of taxonomic families and physico-chemical parameters involved in the subnetworks of core and unique interactions from **a**, and **b**. **d**, Percentage of ecological interactions involved in the **b** subnetworks.

3.4.7 Overview of community based on function

We interrogated the rMAGs functional capacities and expressions to explore the underlying functional mechanisms that could explain the predicted ecological interactions. To that end, we clustered the rMAGs based on their functional potential, i.e. presence/absence of genes covered at the MG level over the entire timeseries, as described by Herold et al. [2020]. Briefly, this clustering is based on the Jaccard similarity indexes of the rMAGs (**Section 3.3.7**), obtained from a binary table of presence/absence of genes in each rMAG, where presence/absence were defined based on the average of gene abundance from all the timepoints (**Section 3.3.7**). Accordingly, the clustering yielded four clusters of the rMAGs (**Figure 3.11**). All these clusters contained all the functional categories found within the community. Then, we defined the realized function by calculating the number of genes

that were expressed in a given functional category per cluster, in relation to the total number of genes of that category found within all the rMAGs (**Figure 3.11**).
The first cluster was dominated by rMAGs with unknown taxonomy, only predicted at the superkindom level as Bacteria. These rMAGs contained all the functions within the community, but exhibited low realized function for all the functional categories. The second cluster was dominated by rMAGs of *Saprospiraceae* and *Chitinophagaceae* families, and it contained the highest realized function of the functional categories, including the expression of 78.3% of the "Lipid metabolism" function. The third cluster was dominated by rMAGs of the families *Leptospiraceae* and *Moraxellaceae*, containing also *Anaerolinaceae*. While in general, the functional categories of the third cluster showed less realized functions than the second cluster, it had the highest realized function of the functional category "Drug resistance: antimicrobial", which contained antimicrobial resistance (AMR) genes. The fourth and last cluster was dominated by rMAGs of the families *Intrasporangiaceae* and *Acidimicrobiaceae*, and contained *Microthrixaceae* family rMAGs, which made up the dominant populations within the model community, in terms of abundance (**Section 2.4.2**). The realized function of the functional categories was similar to the third cluster, e.g. "Lipid metabolism" and "Metabolism of cofactors and vitamins". However, the expression varied in functions like "Glycan biosynthesis and metabolism", where the third cluster showed higher realized function, or "Xenobiotics biodegradation and metabolism", where the fourth cluster had more realized functions.

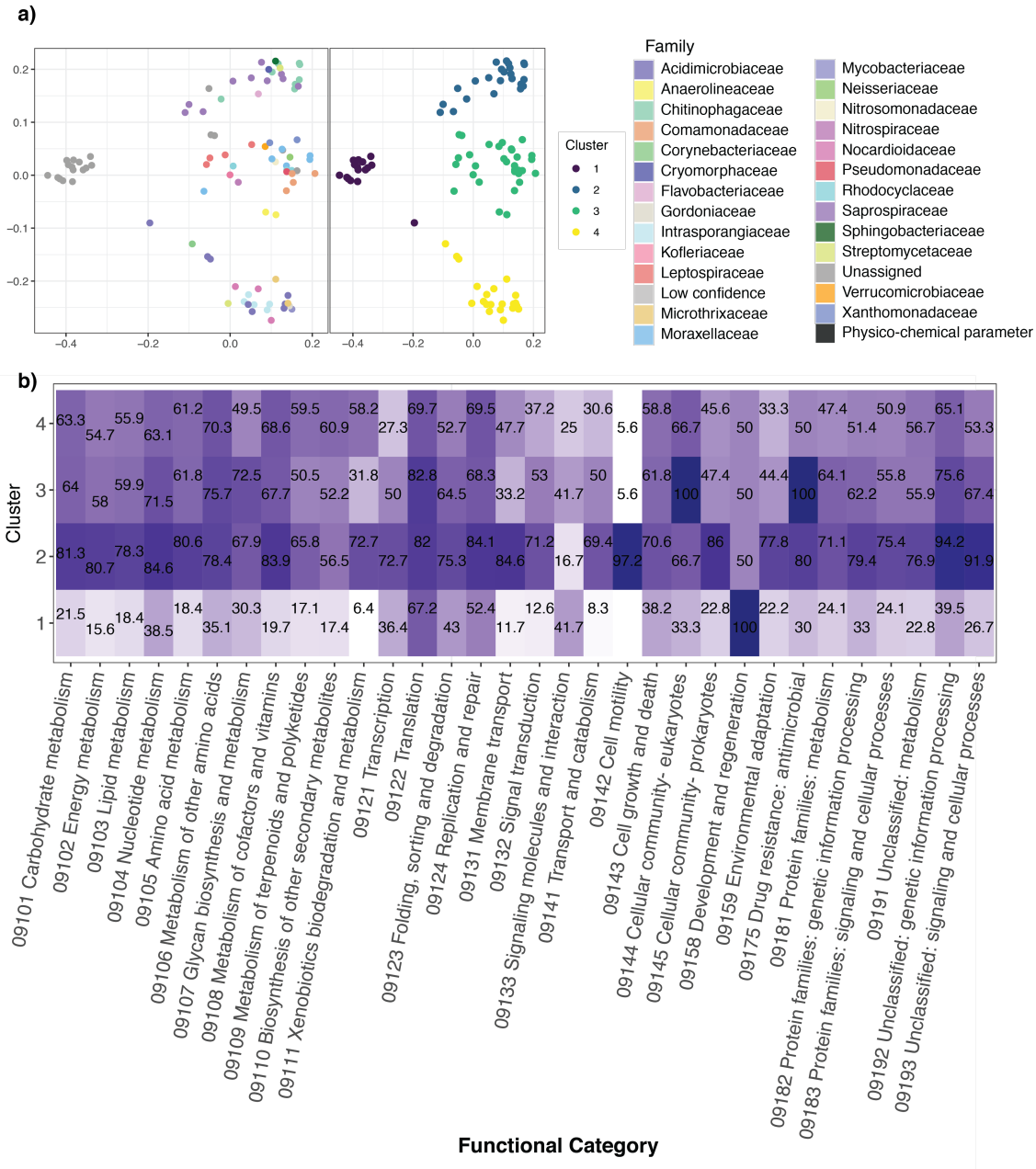


Figure 3.11: Functional potential and expression. **a**, NMSD of the rMAGs based on Jaccard similarity indexes, coloured by taxonomic family on the left, and by cluster membership on the right. **b** Percentage of genes expressed genes per functional category, within each cluster.

3.4.8 Core subnetwork involving dominant population; *Microthrix parvicella*

Microthrix parvicella is a filamentous lipid accumulating bacteria that plays an important role in the efficiency of the activated sludge process, due to its influence in bulking [Rossetti et al., 2005], while being dominant in terms of abundance within the model system (**Section 2.4.2**). We isolated interactions involving *Ca*. M. parvicella from the core subnetwork (**Figure 3.12**) because we were able to determine both stable and dynamic interactions with regards to *Ca*. M. parvicella. Additionally, those core interactions were less influenced by environmental fluctuations (**Figure 3.10**). Based on the relative abundance over time of the interacting rMAGs with *Ca*. M. parvicella, we observed similar abundance patterns for cooperative rMAGs, and nearly opposite trends of abundance for competitive rMAGs (**Figure 3.12**). These similar/opposite patterns were expected due to the linearity of the predicted relationships.

Based on the functional clustering of the rMAGS (**Section 3.4.7**), *Ca*. M. parvicella belonged to the fourth cluster and interacted with members of the first, second and fourth clusters, specifically with rMAGs of the taxonomic families *Acidimicrobiacea*, *Moraxellaceae*, *Xanthomonadaceae* and *Saccharimonadaceae* (**Figure 3.12**).

Its competitive interactions tended to occur with members of the second cluster, while it exhibited cooperative interactions with members of the first and fourth (its own) clusters. *Ca*. M. parvicella tended to maintain stable competitive/cooperative interactions with its counterparts, except for those rMAGs within the *Saccharimonadaceae* family. Interestingly, the overlaps of expressed genes between *Ca*. M. parvicella and its counterparts varied from 23% to 33%, with the smallest overlap with a competitor *Ca*. M. parvicella-centric interaction network. There was also no difference in percentages of functional overlap based on the type of ecological interaction, i.e. competitive/cooperative (**Figure 3.12**). We then focused on the "Lipid metabolism" functional category within *Ca*. M. parvicella, due to its well characterized phenotype of lipid accumulation [McIlroy et al., 2013]. Within this category most of the lipid metabolism pathways were expressed over time. Specifically, genes from the “Fatty acid degradation” subcategory were the highest expressed of the lipid metabolism function, while “Fatty acid elongation”, “Cutin, suberine and wax biosynthesis”, and “Alpha-linoleic acid metabolism” subcategories remained not-expressed over the entire timeseries (**Figure 3.13**). Interestingly, the cooperative rMAGs relative to *Ca*. M. parvicella (**Figure 3.12**) exhibited broad expression of the lipid metabolism genes, including the expression of “Cutin, suberine and wax biosynthesis

pathway" by the rMAGs of the *Acidimicrobiaceae* family (**Figure 3.13**).
The competitive rMAGs (**Figure 3.12**) also showed varying expression patterns of lipid metabolism pathways. For instance, rMAG D15_G1.18.2, from the *Moraxellaceae* family, exhibited an opposite relative abundance and expression trend compared to *Ca*. M. parvicella, (**Figure 3.13**). An additional example was rMAG D20_P23, from the family *Xanthomonadaceae*, which highly expressed the "Fatty acid biosynthesis" genes, but had very low expression levels of all other lipid metabolism subcategories (**Figure 3.13**).
We then performed a simple linear model using those isolated *Ca*. M. parvicella-associated rMAGs as predictor variables for the abundance of *Ca*. M. parvicella (response variable). We obtained a model with adjusted R^2=0.94, with D11_O1.7 (*Microthrixaceae*), D20_P23 (*Xanthomonadaceae*) and D15_G1.18.2 (*Moraxellaceae*) as significant rMAGs.

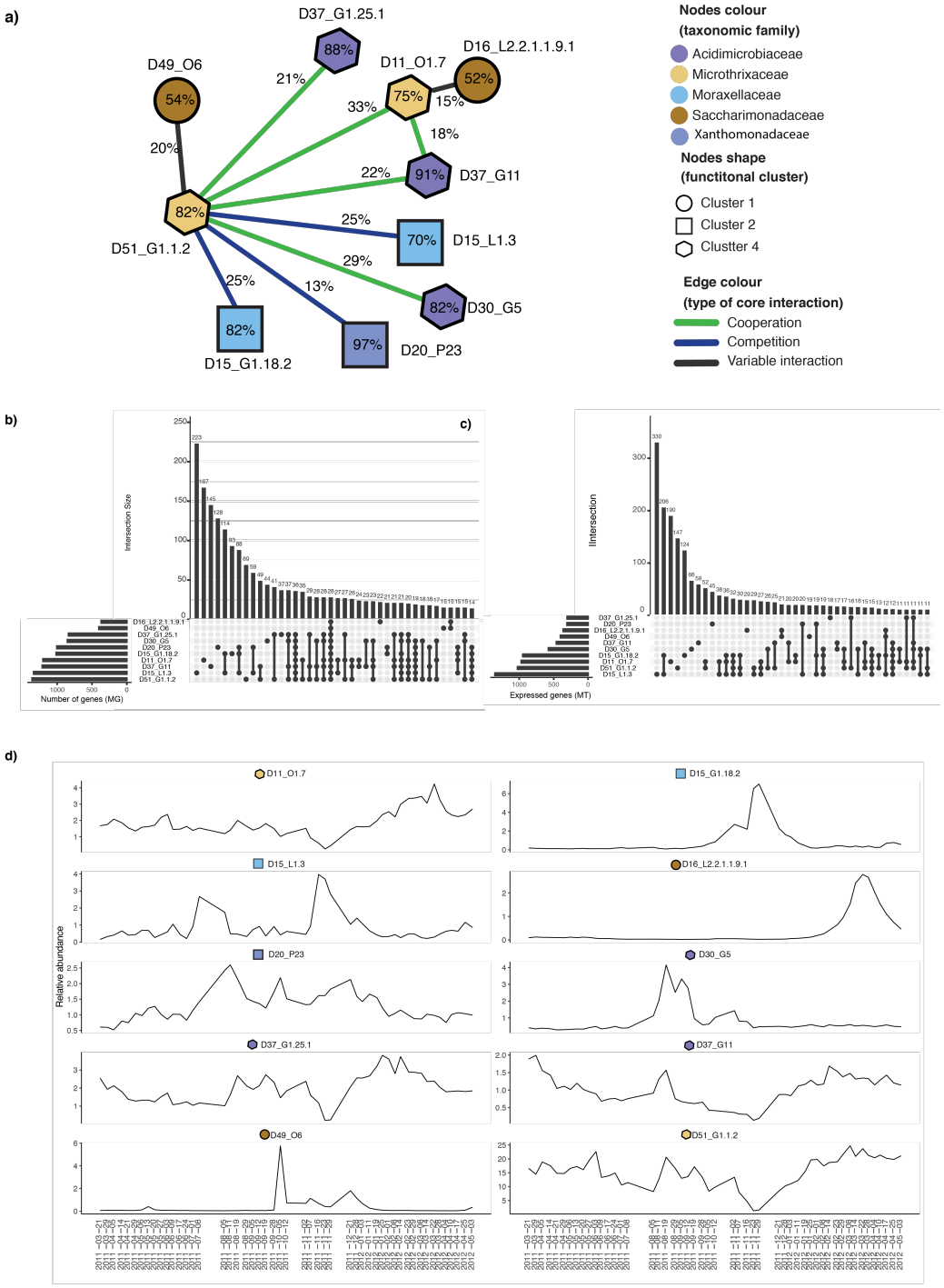


Figure 3.12: Core interactions of *Microthrix parvicella*. a, Subnetwork representing the core interactions in which *Ca*. M. parvicella (two rMAGs) is involved. **b**, Number of overlapping/non-overlapping genes of the rMAGs represented in **a**. **c**, Number of overlapping/non-overlapping expressed genes of the rMAGs represented in **a**. **d** Relative abundance over time of the rMAGs represented in **a**.

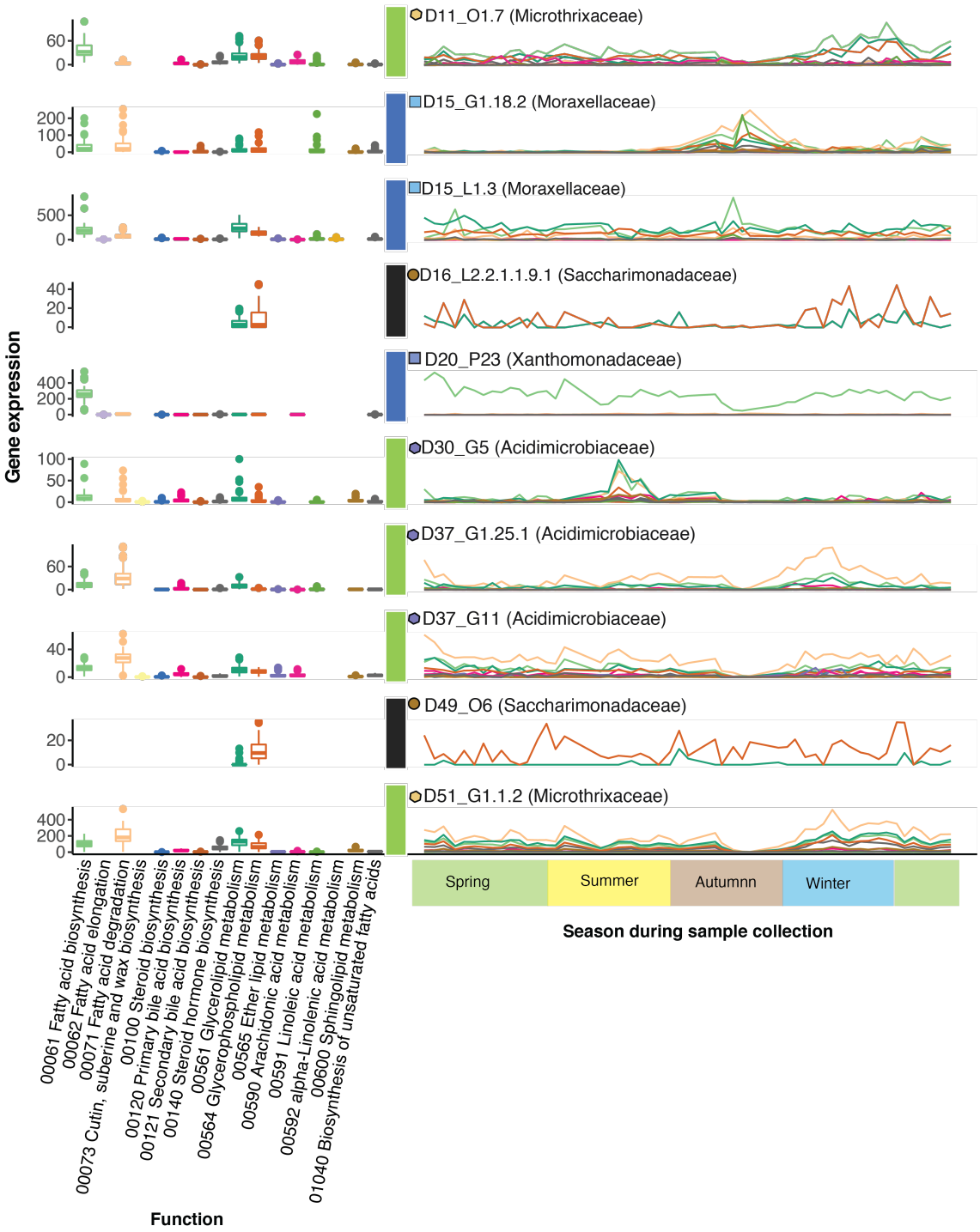


Figure 3.13: Lipid metabolism expression of selected rMAGs. **a**, Summary of functional pathways expression per rMAG. **b**, Functional pathway expression over time.

3.5 Discussion

In this chapter, we showed the seasonal and cyclical behavior exhibited by the model system in terms of i) amount of foam, ii) recorded long-term environmental parameters over 10 years and ii) omic-data trajectory, represented by community structure and diversity indices. Specifically, the sample trajectories showed outliers that might be consequence of short-term perturbations in the system, while the Jenssen-Shannon divergence indexes showed higher fluctuations revolving around the community shift event, i.e. lower index before and after the perturbation. Furthermore, the auto-correlation trends showed only one significantly different auto-correlation on the MG level while demonstrating multiple significantly different auto-correlations in the MT level, which might indicate rapid adaptation to short-term environmental perturbations through differential gene expression.

Linear models have been previously proposed for, and used to model ecological interactions within microbial communities [Trosvik et al., 2015; de Muinck et al., 2017]. To the best of our knowledge, most explorations in this direction relied on 16S data, while we used MG data. Additionally, we utilized elastic nets due to its general advantages and suitability for our data (**Section 3.4.3**). However, it must be noted that these models only capture linear relationships, while microbial interactions in nature can be more complex. Thus, the exploration of non-linear relationships is possible using e.g. generalized additive models [Trosvik et al., 2015]. In general, these regression methods are suitable for our data due to their capability of fitting to data trends such that missing data points need not be imputed in advance.

Considering the concordance of the community shift event occurring on 2011-11-02 to 2012-01-03 with the sample trajectory and MT-level auto-correlation, we computed models based on overlapping time windows centered around the community shift event. Accordingly, we identified unique time window subnetworks and highlighted the influence of physico-chemical parameters within those short-term overlapping time windows compared to the core subnetwork. To that end, we were able to show that the interactions captured by the different short-term overlapping time windows may indeed reflect dynamics responding to perturbations that are not captured over the entire timeseries. Finally, all the taxonomic families were involved in both core and unique time window subnetworks, covering all the functional potential of the community, regardless of the specific rMAGs within those subnetworks.

In our system, cooperative relationships appear to be more frequent and stronger than competitive relationships. Literature has suggested that closely related organisms tend

to compete against one another [Gómez et al., 2010], however we found the opposite in our system, whereby intra-familial relationships tend to be cooperative. Additionally, our functional cluster approach indicated that rMAGs with highly similar functional potential (i.e. same cluster) do not necessarily compete with each other. However, more detailed analysis is required in this direction, e.g. the use of higher resolution of functional exploration [Herold et al., 2020], to further establish general links between function and ecological relationships.

In this work, we used *Ca*. M. parvicella as an instance to explain ecological interactions in the context of their underlying functional mechanisms (**Section 3.4.8**). However, there is still a need to further explore the expression patterns of several functional categories to understand these ecological interactions. The performance of a linear model based on the core *Ca*. M. parvicella interactions was slightly lower compared to the reduced linear model defined in **Chapter 2**, which includes iMGEs. Abundance, functional, and relation to iMGEs information could be combined to form an integrated dynamic model to the entire microbial community.

Chapter 4

General conclusions and perspectives

Part of this chapter was adapted and modified from the following first-author peer-review publication:

Susana Martínez Arbas, Susheel Bhanu Busi, Pedro Queirós, Laura De Nies, Malte Herold, Patrick May, Paul Wilmes, Emilie EL Muller, Shaman Narayanasamy

2021

Challenges, strategies and perspectives for reference-independent longitudinal multi-omic microbiome studies

Frontiers in Genetics, **12**:858

(Accepted)

4.1 General overview

This work presents an extensive time-resolved integrated meta-omics analysis based on two different perspectives: i) iMGEs and host interactions and ii) ecological interactions between microbial community members. This and previous work [Herold et al., 2020] clearly demonstrate the advantages of the unprecedented wealth of information retrievable from a longitudinal and multi-omic microbiome data set. In general, both studies demonstrated the utility of mathematical frameworks supported by empirical information to elucidate a microbial community interactome.

On the one hand, **Chapter 2** revolves around the interactions between bacterial hosts and iMGEs, given that the latter is believed to influence the dynamics of microbial communities. Briefly, we detected an overwhelming abundance of plasmid sequences within this system. In addition, these plasmids were shown to be highly relevant compared to bacteriophages through a mathematical model that predicts the abundance of the dominant microbial family (*Microthrixaceae*). We were then able to support our model by incorporating CRISPR-based links, which showed that plasmids were indeed highly targeted, compared to phages.

On the other hand, **Chapter 3** involved the construction of an ecological interactome of all the identified bacterial community members and abiotic factors. We then defined a set of core interactions within the system that covered all the community predicted functions. We then used gene expression information to observe overlaps or complementarity of functionalities between community members, to explain competitive or cooperative interactions.

Overall, this work could serve as a baseline for further explorations of multi-omic longitudinal microbiome studies, extending to different systems, e.g. human gut microbiome [Lloyd-Price et al., 2019]. In this Chapter, I discuss the potential challenges, implications and longer-term impact of the present work. **Figure 4.1** shows an overview of this Chapter.

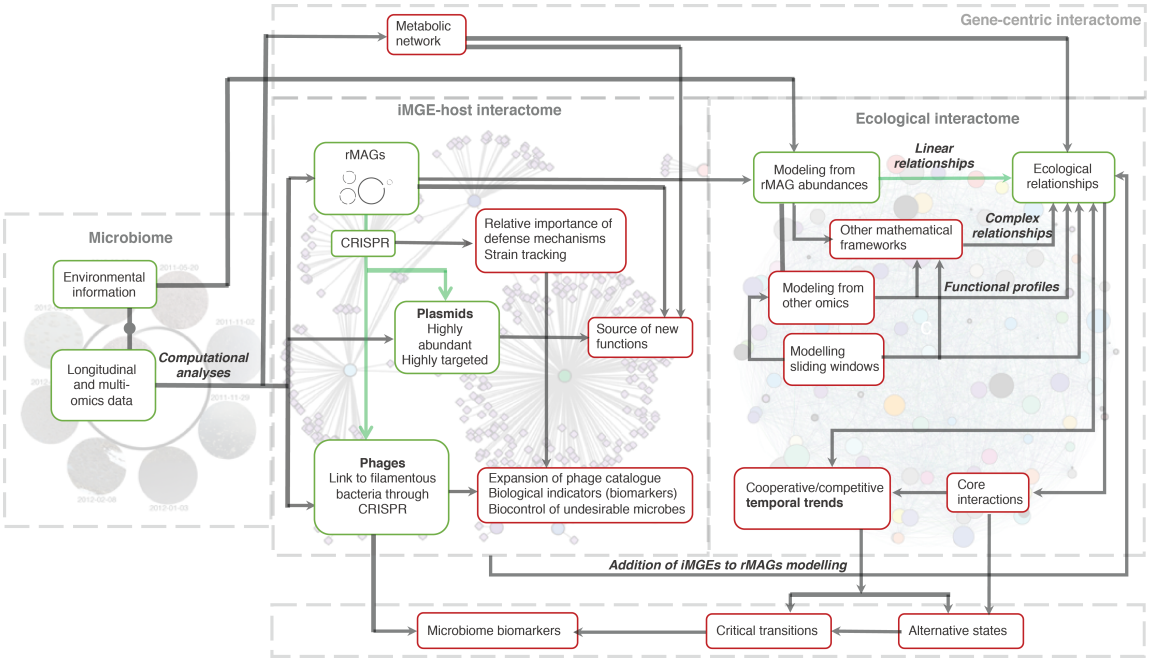


Figure 4.1: General overview of perspectives. Schematic summary of the present work and perspectives. Green boxes represent material shown in previous Chapters. Red boxes represent content discussed within the current Chapter.

4.2 Bacteriophages

As for other mixed microbial community-driven processes, biological wastewater treatment is influenced by abiotic factors, such as temperature or operational conditions, leading to for example seasonal changes in treatment efficiency [Jones and Schuler, 2010; Johnston and Behrens, 2020]. Moreover, biotic factors such as bacteriophages have been shown to influence activated sludge community structure [Brown et al., 2019], which can negatively affect the wastewater treatment quality. Specifically, the increase of bacteriophages (phages) may negatively affect the nitrification [Choi et al., 2011] and phosphorus removal processes [Barr et al., 2010], by provoking a decrease in populations responsible for those processes. Phages have been proposed to serve as bacterial pathogen indicators and biocontrol elements against undesirable bacteria, i.e. not only bacteria that affect the performance of the wastewater treatment, but also bacteria that could be released to rivers and lakes and pose a threat to public health, such as *Salmonella* [Jassim et al., 2016].

An important challenge of wastewater treatment is the bulking of the activated sludge, which involves filamentous bacteria, e.g. *M. parvicella* [Soddell and Seviour, 1990]. Bacteriophages have been proposed as elements to engineer filamentous bacteria involved in

bulking [Withey et al., 2005]. For instance NOC1, NOC2 and NOC3 phages could be used against Nocardioforms, another known agent of bulking [Khairnar et al., 2016]. Similarly, lytic phages could be used against *Haliscomenobacter hydrossis* [Kotay et al., 2011]. Interestingly, the use of parasitic bacteria has recently been proposed as a promising strategy for the biocontrol of bacteria responsible for stabilizing wastewater foams [Batinovic et al., 2021].

In this work, we demonstrated that filamentous bacteria involved in bulking, and in lipid accumulation, such as members of the *Anaerolineaceae* and *Microthrixaceae* families, have an active CRISPR-Cas system, which implied the existence and potential activity of bacteriophages that affect these populations. These predicted phages could be used to create, expand and ideally validate a bacteriophage catalogue with the purpose of engineering the overgrowth of filamentous bacteria within this system [Choi et al., 2011; Liu et al., 2015a]. However, significant refinement would first be required for these newly predicted phages, namely i) more complete and/or accurately reconstructed phage genomes, via *de novo* reassembly, ii) linking phage activity with environmental conditions, bacterial activity or functional processes i.e. under which circumstances do these phages peak, and iii) based on i) and ii) be defined as potential biological indicators (or biomarkers) for important processes within the activated sludge. Finally, multi-omics data, specifically the MT data, offers the unique opportunity to pre-dict, characterize and study the dynamics of RNA phages [Narayanasamy, 2017], while it simultaneously offers the opportunity to identify and explore novel CRISPR-Cas systems for targeting RNA sequences, i.e. RNA interference [Zhu et al., 2018].

4.3 Plasmids

Plasmid content has been typically assessed by focussing on plasmids associated with isolated strains/taxa [Ike et al., 1994; Bauda et al., 1995]. More recently, the availability of metagenomics data has allowed for the assessment of metagenomics-derived "plasmidomes". The work on the model system plasmidome (**Chapter 2**) showed an (over) abundance of plasmids, which were also highly targeted by CRISPR systems, further supporting plas-mids as highly diverse and dynamic components within the system [Sentchilo et al., 2013]. Interestingly, studies [Zhang et al., 2011; Perez et al., 2020] and specialized computational tools [Rozov et al., 2017; Krawczyk et al., 2018; Pellow et al., 2020] focusing on plas-midomes are also emerging, enabling a deeper understanding of these components and their roles within microbial ecosystems.

In this work, plasmid sequences were found to be enriched in functions related to DNA replication, recombination, repair and transfer (**Appendix E.5**), which is in line with a recent study of plasmids within a lake microbial ecosystem, where most of the known functions of plasmids were dominated by functions associated with replication and transposition [Perez et al., 2020]. Activated sludge systems were shown to carry a wide variety of antibiotic and heavy metal resistance genes [Zhang et al., 2011; Sentchilo et al., 2013]. In that regard, this work highlighted the depletion in functions related to antimicrobial resistance genes (ARGs) within plasmids targeted by CRISPR-Cas systems (**Chapter 2**). Previous work suggests that conjugative plasmids targeted by CRISPR-Cas systems often lack ARGs [Shmakov et al., 2017]. This might reflect the relevance of plasmids containing ARGs to shape the microbial community of the foaming activated sludge. Furthermore, this lack of ARGs within targeted plasmids has been proposed as a useful feature to decrease the spread of ARGs [Gholizadeh et al., 2020], and to engineer the microbial community by sequence-specific killing of pathogenic bacteria or removal of accessory genes [Westra et al., 2019]. Finally, plasmids are also highly interesting because they represent a source of novel functions within microbial communities, carrying between 40 to 60% genes of unknown functions [Zhang et al., 2011; Sentchilo et al., 2013; Dib et al., 2015].

4.4 CRISPR information as a multi-faceted analytical tool for microbiomes

In **Chapter 2**, information from CRISPR systems was mainly used as empirical links between host populations and corresponding iMGEs. Previous work has also relied on CRISPR information as a means to identify novel bacteriophages, and the identification of new phage marker genes [Davison et al., 2016]. However, CRISPR information within a microbiome is not limited to the prediction/classification of novel bacteriophages, specifically in the context of our multi-omic longitudinal data. We were able to use the information on spacer gain-loss, which could also serve as an indicator of selection pressure based on MGEs activity to better understand underlying mechanisms of bacterial (or archaeal) adaptation, i.e. during aggressive/active infection of bacteriophages, CRISPR-Cas systems might be less active relative to other defense mechanisms, given that complementary defense mechanisms are believed to act faster compared to CRISPR-Cas systems, for

instance surface modification mechanisms to avoid phage adsorption [Chevallereau et al., 2019]. However, this remains to be further characterized at a large scale in naturally occurring microbial communities, such as the model system of this work. Moreover, the combination of multi-omics and longitudinal data would allow us to further interrogate the dynamics of these defense mechanisms.

CRISPR-Cas information also serves as a way to differentiate strains that can be linked to different epidemiological traits, such as bacterial virulence, antimicrobial resistance and/or geographic origin [Karimi et al., 2018]. This strain information can be obtained through the order of spacers within the CRISPR locus [Shariat and Dudley, 2014], and could be used in complement with classic methods of strain tracking, such as single nucleotide variants, copy number variants, auxiliary gene content [Evans and Denef, 2020], and time-resolved strain tracking [Brito and Alm, 2016; Zlitni et al., 2020].

4.5 Expanding modelling approaches for the longitudinal multi-omics dataset

Modelling time series and multi-omics microbiome data is highly challenging due to its unique characteristics (**Chapter 1**). Classical mathematical algorithms used to model longitudinal data from other fields, such as finance and economics, may not be directly applicable to microbiome datasets because they are typically not as extensive or continuous, due to the cost of generating the data.

Chapter 1 briefly reviewed suitable methods to model longitudinal data to infer linear and non-linear relationships between community members (**Section 1.3.5**), based on their abundances at the MG level. However, extensive literature of statistical and mathematical frameworks for multi-omic and/or longitudinal microbiome data is currently available. For instance, Noor et al. [2019] review the integration of multi-omics data from data-driven and knowledge-based perspectives. Coenen et al. [2020] discuss approaches to characterize temporal dynamics and to identify periodicity of populations and putative interactions between them, while Faust et al. [2018] propose a classification scheme for better model selection. Bodein et al. [2019] provide a multivariate framework to integrate longitudinal and multi-omics data, while Park et al. [2020] discuss the development of models and software tools for time series metagenome and metabolome data. Ruiz-Perez et al. [2021] proposes bayesian dynamic networks to integrate time series multi-omics microbiome data. Overall, the application of these methodologies should be tailored towards

specific hypotheses and studies, for which data exploration is essential to select modelling approaches that fit the type, quality, and quantity of the data.

The emergence of studies which track microbiome dynamics of cohorts over time, i.e. multiple individuals/sites [Carmody et al., 2019; Lloyd-Price et al., 2019; Mars et al., 2020], necessitates the ability to discriminate variation stemming from the same individual/environment compared to those from different individuals/environments. In such cases, multi-level statistical modeling (also known as mixed-effects/hierarchical models) are able to account for repeated sampling or nested variation across a sample population [Sokal, 1995; Kuznetsova et al., 2017; Mallick et al., 2021]. Most notably, Lloyd-Price et al. [2019] extensively applied such methods to associate multi-omic microbiome signatures with host-derived molecular profiles in a cohort of 132 individuals. Other instances include multi-omic longitudinal studies that combine murine and human datasets to unveil the adaptation of gut microbiomes to raw and cooked food [Carmody et al., 2019] and the identification of therapeutic targets for irritable bowel syndrome [Mars et al., 2020]. Finally, there are newer methodologies that apply similar/related statistical frameworks to modelling multi-omic data [Mallick et al., 2021].

In the context of **Chapter 3**, the inference of ecological interactions based on microbial MG level abundances, i.e. cooperative and competitive interactions inferred through linear relationships, showed that cooperative interactions were stronger than competitive interactions intra- and inter- taxa. These models can be further expanded in several ways to answer different questions. For instance, the application of additional mathematical algorithms, such as generalized additive models (GAMS) [Trosvik et al., 2015], could be used to capture complex interactions, especially those that extend beyond linear relationships. Then, the results from different modelling approaches can be then integrated to select a group of core interactions [Röttjers et al., 2020].

We could also envision integrating the modelling-based approach in **Chapter 3** with the iMGEs information from **Chapter 2**. Specifically, the iMGEs could serve as predictor variables for rMAG abundances. Additionally, the link of such iMGEs through e.g. correlations, to environmental conditions, could add the assumption to the models that prophages are triggered to lytic lifestyle under environmental stress conditions.

4.6 Addition of biological information: functional clusters, metabolic networks and rMAG profiling

The advantage of additional functional information that the multi-omics data offers in this dataset is clear. On one hand, metabolic network reconstruction based on MG and MT data have been used to define ecological interactions of cooperation or competition based on the potential use of nutrient resources, and gene expression of specific functions, which also defines fundamental and realized niches [Muller et al., 2018]. The activated sludge system has demonstrated functional redundancy between the community members, and flexibility of the community members to respond to perturbations within the system [Herold et al., 2020]. The addition of such information for the ecological interactome model building could be used as another layer of information to define rMAG functional profiles. On the other hand, through the usage of the MP and MM data, one could prune the ecological network and further define cooperation or competition scores based on the relative amount of overlapping compounds that could be assigned to specific rMAGs [Wilmes et al., 2010]. Subsequently, one could also define functional profiles per rMAG/ iMGEs, per time point, i.e. functional potential based on the MG information and expressed function based on the MT information. To be able to further analyse temporal patterns, one could construct networks based on sliding windows, which is a standard methodology utilised in finance (e.g. stock market), rather than using predefined time windows, such as the performed models in **Chapter 2** and **Chapter 3** which were inferred from the entire time series and from overlapping time windows using as reference point the community shift observed in the autumn season (**Section 2.4.5**, **Appendix E.6**). The sliding window specific networks could reveal fine-grained dynamics, patterns or trends. However, one of the challenges with such an approach, specifically with the model data set, would be the selection of the right number of time points per sliding window and mitigating potential effects of uneven sampling. Overall, here a further integration of the multi-omics data is proposed, rather than analysing each omic independently and then comparing the results. This has the risk to add noise, but also the advantage to capture emergent properties of the system that might not be cap-tured when analysing the omics independently.

4.7 Expanding functional annotation

De novo assembly of MG and/or MT datasets yield a plethora of open reading frames, which are then functionally annotated *in silico*. These *in silico* functional annotation strategies are typically based on sequence homology [Ovchinnikov et al., 2017]. However, there are many ORFs that remain unannotated, e.g. "hypothetical proteins" or as "domains/proteins of unknown functions" [Robinson et al., 2021]. Based on the system of study, more than half of the predicted ORFs could remain unannotated.

Given the availability of quantitative gene-level MG and MT abundance information, one could envision utilizing such abundance information to construct gene-level interactomes. The general approach would be similar to that described in **Chapter 3**, but with the interacting components being genes, instead of populations/taxa. One could then infer functions of unknown genes based on "guilt by association", i.e. genes which are associated or interacting are more likely to share function [Gillis and Pavlidis, 2012]. While such approaches have been applied in model organisms [Lee et al., 2015; Kim et al., 2015, 2013], humans [Hwang et al., 2019], and mixed microbial communities [Jaffe et al., 2016; Corel et al., 2016]. Such strategies in combination with downstream *in silico* refinement, e.g. metabolic modelling [Magnúsdóttir et al., 2017], and laboratory validation, e.g. overexpression experiments [Narayanasamy et al., 2015], could lead to *de novo* discovery of enzymes, i.e. proteins from previously uncharacterized families/folds [Robinson et al., 2021], which could lead to promising enzymes for biotechnological processes, derived from unculturable mixed microbes [Berini et al., 2017].

4.8 Understanding microbial community dynamics for system prediction

Microbial communities are highly dynamic systems (**Chapter 1**) and their general properties, such as stability and alternative stable states, can be explored to predict responses to perturbations. Stability is a property widely used to define stable states of a system and their transition between them (if any). However, the definition of stability could vary based on various criteria, e.g. time scale, biological context, etc. [Gonze et al., 2018]. Moreover, microbial communities may exhibit a transition between multiple stable states, which are referred to as alternative stable states [Gonze et al., 2017]. The transition between alternative stable states may be smooth or drastic, i.e. critical transitions. Identification of trends may serve as early warning signals of imminent transitions between those alternative states

[Scheffer et al., 2009]. In general, these trends may be identified via i) statistical, or ii) mathematical model-based indicators [Scheffer et al., 2012].
Statistical indicators include, e.g. i) increased autocorrelations of dominant taxa within the system [Scheffer et al., 2009; Liu et al., 2015b; Fuhrman et al., 2006; Gilbert et al., 2012] and ii) bistability analysis (e.g. taxa distributions) [Lahti et al., 2014]. Additionally, other creative statistical measures could be applied based on the system of study and biological question. For instance, the computation of resilience indexes to quantify microbial populations response to environmental perturbations, i.e. the rate (how fast) and the extent (how close to the pre-pertubation state) a microbial population recovers from a given perturbation [Dogra et al., 2020], for which alpha-diversity has been used [Raymond et al., 2016]. Levy et al. [2020] have defined a permissivity parameter to calculate the relative tendency of the microbiome to permit (flexibility) or resist (rigidity) variation based on microbial population trajectories. Similarly, de Celis et al. [2020] defined a stability measure based on a core microbiome that makes up a large fraction of the total microbial abundance.
Community-level models which infer biological information, such as ecological interactions, are also suitable to identify early warning signals in predicting critical transitions. For example, Coyte et al. [2015]; Coyte and Rakoff-Nahoum [2019] simulated stability based on the amount of cooperative and competitive interactions within a system and observed that an increase in cooperative interactions leads to instability of the system. Although, efforts to explain and validate such transitions are being carried out [Hoek et al., 2016; Castellanos et al., 2020; Xu, 2021; Seelbinder et al., 2020].
In the context of the data analysed here, we have used longitudinal multi-omics with various methods to show that microbial populations within our model system are resilient against environmental perturbations [Herold et al., 2020]. Moreover, in **Chapter 3**, we computed the MG- and MT- level microbial population autocorrelations and observed fluctuating patterns that could be explained by seasonal changes. Additionally, our model-based ecological interactome revealed dynamical trends of cooperation and competition over time to further support the notion of stability and resilience within the system.

4.9 Perspective on microbial community control

Longitudinal multi-omics studies offer a plethora of new opportunities for further understanding microbial community dynamics. The methodology applied to the present work can be adapted and applied to different microbiomes to characterize their ecological interactome dynamics. Leveraging information of combined multi-omics modeling through

network topology analysis could lead to the identification of robust and cost-efficient microbiome-derived biomarkers. These biomarkers could serve as important indicators in microbial ecosystems, especially those relevant to health, biotechnology and nature.
For instance, a set of biomarkers could be defined to monitor wastewater treatment and perhaps predict bulking. Wastewater treatment management entities could then respond to these predictions by adjusting physico-chemical parameters (e.g. adjusting aeration periods) or through biological means (e.g introduction of specific bacteria or phages). On one hand, the latter could enable the process control of bulking to ensure smooth operations of the treatment plant, while on the other hand bulking could be maximized for optimized biofuel production. Consequently, such strategies could be helpful for engineering and control of sustainable wastewater treatment processes.

Bibliography

J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. Logratio Analysis and Compositional Distance. *Mathematical Geology*, 32(3):271–275, April 2000. ISSN 1573-8868. doi: 10.1023/A:1007529726302. URL https://doi.org/10.1023/A:1007529726302.

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, June 2013. ISSN 1546-1696. doi: 10.1038/nbt.2579. URL https://doi.org/10.1038/nbt.2579.

Mário Almeida-Neto, Paulo Guimarães, Paulo R Guimarães, Rafael D Loyola, and Werner Ulrich. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008. doi: 10.1111/j.0030-1299.2008.16644.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0030-1299.2008.16644.x.

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, November 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3103. URL http://www.nature.com/articles/nmeth.3103.

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990. ISSN 0022-2836 (Print). doi: 10.1016/S0022-2836(05)80360-2.

H. M. Alvarez, F. Mayer, D. Fabritius, and A. Steinbüchel. Formation of intracytoplasmic lipid inclusions by Rhodococcus opacus strain PD630. *Archives of microbiology*, 165(6):377–386, June 1996. ISSN 0302-8933. doi: 10.1007/s002030050341. Place: Germany.

R I Amann, W Ludwig, and K H Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169, March 1995. ISSN 0146-0749. URL https://pubmed.ncbi.nlm.nih.gov/7535888.

Gil Amitai and Rotem Sorek. CRISPR-Cas adaptation: insights into the mechanism of action. *Nature reviews. Microbiology*, 14(2):67–76, February 2016. ISSN 1740-1534 (Electronic). doi: 10.1038/nrmicro.2015.14.

Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R. Mende, Gabriel R. Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H. Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G. Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M. de Vos, Søren Brunak, Joel Doré, Jean Weissenbach, S. Dusko Ehrlich, and Peer Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, May 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09944. URL http://www.nature.com/articles/nature09944.

Noam Auslander, Ayal B Gussow, Sean Benler, Yuri I Wolf, and Eugene V Koonin. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Research*, 48(21):e121–e121, December 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa856. URL https://academic.oup.com/nar/article/48/21/e121/5921300.

Jeremy J. Barr, Frances R. Slater, Toshikazu Fukushima, and Philip L. Bond. Evidence for bacteriophage activity causing community and performance changes in a phosphorus-removal activated sludge. *FEMS microbiology ecology*, 74(3):631–642, December

2010. ISSN 1574-6941 0168-6496. doi: 10.1111/j.1574-6941.2010.00967.x. Place: England.

Steven Batinovic, Jayson J. A. Rose, Julian Ratcliffe, Robert J. Seviour, and Steve Petrovski. Cocultivation of an ultrasmall environmental parasitic bacterium with lytic ability against bacteria associated with wastewater foams. *Nature Microbiology*, 6(6): 703–711, June 2021. ISSN 2058-5276. doi: 10.1038/s41564-021-00892-1. URL https://doi.org/10.1038/s41564-021-00892-1.

P. Bauda, C. Lallement, and J. Manem. Plasmid content evaluation of activated sludge. *Water Research*, 29(1):371–374, January 1995. ISSN 0043-1354. doi: 10.1016/0043-1354(94)E0088-N. URL https://www.sciencedirect.com/science/article/pii/0043135494E0088N.

Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa de Souza, Leo van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):103, December 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00875-0. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00875-0.

Øivind Bergh, Knut Yngve BØrsheim, Gunnar Bratbak, and Mikal Heldal. High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468, August 1989. ISSN 1476-4687. doi: 10.1038/340467a0. URL https://doi.org/10.1038/340467a0.

Francesca Berini, Carmine Casciello, Giorgia Letizia Marcone, and Flavia Marinelli. Metagenomics: novel enzymes from non-culturable microbes. *FEMS Microbiology Letters*, 364(21), 2017. ISSN 0378-1097. doi: 10.1093/femsle/fnx211. URL https://doi.org/10.1093/femsle/fnx211. _eprint: https://academic.oup.com/femsle/article-pdf/364/21/fnx211/23930554/fnx211.pdf.

Aude Bernheim, David Bikard, Marie Touchon, and Eduardo P C Rocha. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic Acids Research*, 48(2):748–760, 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz1091. URL https://doi.org/10.1093/nar/gkz1091. _eprint: https://academic.oup.com/nar/article-pdf/48/2/748/31784873/gkz1091.pdf.

Juliano Bertozzi Silva, Zachary Storms, and Dominic Sauvageau. Host receptors for bacteriophage adsorption. *FEMS Microbiology Letters*, 363(4):fnw002, February 2016. ISSN 1574-6968. doi: 10.1093/femsle/fnw002. URL https://academic.oup.com/femsle/article-lookup/doi/10.1093/femsle/fnw002.

Matthew B. Biggs, Gregory L. Medlock, Glynis L. Kolling, and Jason A. Papin. Metabolic network modeling of microbial communities: Metabolic network modeling. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334, September 2015. ISSN 19395094. doi: 10.1002/wsbm.1308. URL http://doi.wiley.com/10.1002/wsbm.1308.

Ambarish Biswas, Joshua N Gagnon, Stan J J Brouns, Peter C Fineran, and Chris M Brown. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA biology*, 10(5):817–827, May 2013. ISSN 1555-8584 (Electronic). doi: 10.4161/rna.24046.

L L Blackall, H Stratton, D Bradford, T D Dot, C Sjorup, E M Seviour, and R J Seviour. "Candidatus Microthrix parvicella", a filamentous bacterium from activated sludge sewage treatment plants. *International journal of systematic bacteriology*, 46(1):344–346, January 1996. ISSN 0020-7713 (Print). doi: 10.1099/00207713-46-1-344.

Charles Bland, Teresa L Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C Kyrpides, and Philip Hugenholtz. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics*, 8:209, 2007. ISSN 14712105. doi: 10.1186/1471-2105-8-209. ISBN: 1471-2105 (Electronic)$\backslash$n1471-2105 (Linking).

Sonja Blasche, Yongkyu Kim, Ruben A. T. Mars, Daniel Machado, Maria Maansson, Eleni Kafkia, Alessio Milanese, Georg Zeller, Bas Teusink, Jens Nielsen, Vladimir Benes, Rute Neves, Uwe Sauer, and Kiran Raosaheb Patil. Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nature Microbiology*,

6(2):196–208, February 2021. ISSN 2058-5276. doi: 10.1038/s41564-020-00816-5. URL http://www.nature.com/articles/s41564-020-00816-5.

Antoine Bodein, Olivier Chapleur, Arnaud Droit, and Kim-Anh Lê Cao. A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. *Frontiers in Genetics*, 10:963, November 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00963. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00963/full.

Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu170. URL https://doi.org/10.1093/bioinformatics/btu170.

Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561, August 2005. ISSN 13500872. doi: 10.1099/mic.0.28048-0. URL http://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.28048-0. Publisher: Microbiology Society.

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Feature selection for high-dimensional data. *Prog Artif Intell*, 5:65–75, 2016. doi: 10.1007/s13748-015-0080-y. URL https://doi.org/10.1007/s13748-015-0080-y.

Joseph Bondy-Denomy, Jason Qian, Edze R Westra, Angus Buckling, David S Guttman, Alan R Davidson, and Karen L Maxwell. Prophages mediate defense against phage infection through diverse mechanisms. *The ISME Journal*, 10(12):2854–2866, December 2016. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2016.79. URL http://www.nature.com/articles/ismej201679.

K. Gerald van den Boogaart and R. Tolosana-Delgado. "compositions": A unified R package to analyze compositional data. *Computers & Geosciences*, 34(4):320–338, 2008. ISSN 0098-3004. doi: https://doi.org/10.1016/j.cageo.2006.11.017. URL https://www.sciencedirect.com/science/article/pii/S0098300407001010X.

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):107–120, February 2014. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3643. URL http://www.nature.com/articles/nrg3643.

Luis Boto. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683):819–827, March 2010. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2009.1679. URL https://royalsocietypublishing.org/doi/10.1098/rspb.2009.1679.

Ilana L. Brito and Eric J. Alm. Tracking Strains in the Microbiome: Insights from Metagenomics and Models. *Frontiers in Microbiology*, 7:712, May 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00712. URL http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00712/abstract.

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

C. Titus Brown, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, and Timothy H. Brom. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. *arXiv:1203.4802 [q-bio]*, May 2012. URL http://arxiv.org/abs/1203.4802. arXiv: 1203.4802.

M.R. Brown, J.C. Baptista, M. Lunn, D.L. Swan, S.J. Smith, R.J. Davenport, B.D. Allen, W.T. Sloan, and T.P. Curtis. Coupled Virus - Bacteria Interactions and Ecosystem Function in an Engineered Microbial System. *Water Research*, 152:264–273, January 2019. ISSN 0043-1354. Publisher: Pergamon.

Susheel Bhanu Busi, Laura de Nies, Janine Habier, Linda Wampach, Joëlle V. Fritz, Anna Heintz-Buschart, Patrick May, Rashi Halder, Carine de Beaufort, and Paul Wilmes. Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications*, 1(1):8, March 2021. ISSN 2730-6151. doi: 10.1038/s43705-021-00003-5. URL https://doi.org/10.1038/s43705-021-00003-5.

Julie Callanan, Stephen Stockdale, Andrey Shkoporov, Lorraine Draper, R. Ross, and Colin Hill. RNA Phage Biology in a Metagenomic Era. *Viruses*, 10(7):386, July 2018.

ISSN 1999-4915. doi: 10.3390/v10070386. URL http://www.mdpi.com/1999-4915/10/7/386.

Rachel N. Carmody, Jordan E. Bisanz, Benjamin P. Bowen, Corinne F. Maurice, Svetlana Lyalina, Katherine B. Louie, Daniel Treen, Katia S. Chadaideh, Vayu Maini Rekdal, Elizabeth N. Bess, Peter Spanogiannopoulos, Qi Yan Ang, Kylynda C. Bauer, Thomas W. Balon, Katherine S. Pollard, Trent R. Northen, and Peter J. Turnbaugh. Cooking shapes the structure and function of the gut microbiome. *Nature Microbiology*, 4(12):2052–2063, December 2019. ISSN 2058-5276. doi: 10.1038/s41564-019-0569-4. URL https://doi.org/10.1038/s41564-019-0569-4.

Victoria R. Carr, Andrey Shkoporov, Colin Hill, Peter Mullany, and David L. Moyes. Probing the Mobilome: Discoveries in the Dynamic Microbiome. *Trends in Microbiology*, 29(2):158–170, February 2021. ISSN 0966-842X. doi: 10.1016/j.tim.2020.05.003. URL http://www.sciencedirect.com/science/article/pii/S0966842X20301281.

Nazareth Castellanos, Gustavo G. Diez, Carmen Antúnez-Almagro, María Bailén, Carlo Bressa, Rocío González Soltero, Margarita Pérez, and Mar Larrosa. A Critical Mutualism – Competition Interplay Underlies the Loss of Microbial Diversity in Sedentary Lifestyle. *Frontiers in Microbiology*, 10:3142, January 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2019.03142. URL https://www.frontiersin.org/article/10.3389/fmicb.2019.03142/full.

You Che, Yu Xia, Lei Liu, An-Dong Li, Yu Yang, and Tong Zhang. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome*, 7(1):44, December 2019. ISSN 2049-2618. doi: 10.1186/s40168-019-0663-0. Publisher: BioMed Central.

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. ISBN: 1367-4811 (Electronic)$\backslash$r1367-4803 (Linking).

Lin-Xing Chen, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. Accurate and complete genomes from metagenomes. *Genome Research*, 30(3):315–333, March 2020. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.258640.

119. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.258640.119.

Anne Chevallereau, Sean Meaden, Stineke van Houte, Edze R. Westra, and Clare Rollie. The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374 (1772):20180094, May 2019. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2018.0094. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2018.0094.

Jeongdong Choi, Shireen Meher Kotay, and Ramesh Goel. Bacteriophage-based biocontrol of biological sludge bulking in wastewater. *Bioengineered bugs*, 2(4):214–217, August 2011. ISSN 1949-1026 1949-1018. doi: 10.1016/j.watres.2010.08.038. Place: United States.

Adam Thomas Clark, Jean-Francois Arnoldi, Yuval R. Zelnik, György Barabas, Dorothee Hodapp, Canan Karakoç, Sara König, Viktoriia Radchuk, Ian Donohue, Andreas Huth, Claire Jacquet, Claire Mazancourt, Andrea Mentges, Dorian Nothaaß, Lauren G. Shoemaker, Franziska Taubert, Thorsten Wiegand, Shaopeng Wang, Jonathan M. Chase, Michel Loreau, and Stanley Harpole. General statistical scaling laws for stability in ecological systems. *Ecology Letters*, page ele.13760, May 2021. ISSN 1461-023X, 1461-0248. doi: 10.1111/ele.13760. URL https://onlinelibrary.wiley.com/doi/10.1111/ele.13760.

Ashley R. Coenen and Joshua S. Weitz. Limitations of Correlation-Based Inference in Complex Virus-Microbe Communities. *mSystems*, 3(4):e00084–18, August 2018. ISSN 2379-5077. doi: 10.1128/msystems.00084-18. URL https://msystems.asm.org/content/3/4/e00084-18.abstract. Publisher: American Society for Microbiology Journals.

Ashley R. Coenen, Sarah K. Hu, Elaine Luo, Daniel Muratore, and Joshua S. Weitz. A Primer for Microbiome Time-Series Analysis. *Frontiers in Genetics*, 11:310, April 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00310. URL https://www.frontiersin.org/article/10.3389/fgene.2020.00310/full.

Stephanie Connelly, Seung G. Shin, Robert J. Dillon, Umer Z. Ijaz, Christopher Quince, William T. Sloan, and Gavin Collins. Bioreactor Scalability: Laboratory-Scale Bioreac-

tor Design Influences Performance, Ecology, and Community Physiology in Expanded Granular Sludge Bed Bioreactors. *Frontiers in Microbiology*, 8:664, May 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.00664. URL http://journal.frontiersin.org/article/10.3389/fmicb.2017.00664/full.

Nora Connor, Albert Barberán, and Aaron Clauset. Using null models to infer microbial co-occurrence networks. *PLOS ONE*, 12(5):e0176751, May 2017. doi: 10.1371/journal.pone.0176751. URL https://doi.org/10.1371/journal.pone.0176751. Publisher: Public Library of Science.

Jake R Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics (Oxford, England)*, 33(18):2938–2940, September 2017. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btx364.

Eduardo Corel, Philippe Lopez, Raphaël Méheust, and Eric Bapteste. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends in Microbiology*, 24(3):224–237, March 2016. ISSN 0966842X. doi: 10.1016/j.tim.2015.12.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0966842X15002796.

David Couvin, Aude Bernheim, Claire Toffano-Nioche, Marie Touchon, Juraj Michalik, Bertrand Néron, Eduardo P C Rocha, Gilles Vergnaud, Daniel Gautheret, and Christine Pourcel. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, 46(W1):W246–W251, July 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky425. URL https://academic.oup.com/nar/article/46/W1/W246/5001162.

K. Z. Coyte, J. Schluter, and K. R. Foster. The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–666, November 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aad2602. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aad2602.

Katharine Z. Coyte and Seth Rakoff-Nahoum. Understanding Competition and Cooperation within the Mammalian Gut Microbiome. *Current Biology*, 29(11):R538–

R544, June 2019. ISSN 09609822. doi: 10.1016/j.cub.2019.04.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982219304154.

Katharine Z. Coyte, Chitong Rao, Seth Rakoff-Nahoum, and Kevin R. Foster. Ecological rules for the assembly of microbiome communities. *PLOS Biology*, 19(2):e3001116, February 2021. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001116. URL https://dx.plos.org/10.1371/journal.pbio.3001116.

Alexandra B Crawley, James R Henriksen, and Rodolphe Barrangou. CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems. *The CRISPR Journal*, 1(2):171–181, 2018. doi: 10.1089/crispr.2017.0022. URL https://doi.org/10.1089/crispr.2017.0022.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL https://igraph.org.

Sanjeev Dahal, James T. Yurkovich, Hao Xu, Bernhard O. Palsson, and Laurence Yang. Synthesizing Systems Biology Knowledge from Omics Using Genome-Scale Models. *PROTEOMICS*, 20(17-18):1900282, September 2020. ISSN 1615-9853, 1615-9861. doi: 10.1002/pmic.201900282. URL https://onlinelibrary.wiley.com/doi/10.1002/pmic.201900282.

Holger Daims, Michael W. Taylor, and Michael Wagner. Wastewater treatment: a model system for microbial ecology. *Trends in Biotechnology*, 24(11):483–489, November 2006. ISSN 01677799. doi: 10.1016/j.tibtech.2006.09.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0167779906002162.

John Davison. Genetic Exchange between Bacteria in the Environment. *Plasmid*, 42(2):73–91, September 1999. ISSN 0147-619X. doi: 10.1006/PLAS.1999.1421. URL https://www.sciencedirect.com/science/article/pii/S0147619X9991421X?via%3Dihub. Publisher: Academic Press.

Michelle Davison, Todd J Treangen, Sergey Koren, Mihai Pop, and Devaki Bhaya. Diversity in a Polymicrobial Community Revealed by Analysis of Viromes, Endolysins and CRISPR Spacers. *PloS one*, 11(9):e0160574, 2016. ISSN 1932-6203 (Electronic). doi: 10.1371/journal.pone.0160574.

Miguel de Celis, Ignacio Belda, Rüdiger Ortiz-Álvarez, Lucía Arregui, Domingo Marquina, Susana Serrano, and Antonio Santos. Tuning up microbiome analysis to monitor WWTPs' biological reactors functioning. *Scientific Reports*, 10(1):4079, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61092-1. URL http://www.nature.com/articles/s41598-020-61092-1.

Eric J. de Muinck, Knut E. A. Lundin, and Pål Trosvik. Linking Spatial Structure and Community-Level Biotic Interactions through Cooccurrence and Time Series Modeling of the Human Intestinal Microbiota. *mSystems*, 2(5):mSystems.00086–17, e00086–17, October 2017. ISSN 2379-5077. doi: 10.1128/mSystems.00086-17. URL https://msystems.asm.org/content/2/5/e00086-17.

Laura de Nies, Sara Lopes, Anna Heintz-Buschart, Cedric Christian Laczny, Patrick May, and Paul Wilmes. PathoFact: A pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. preprint, Microbiology, March 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.03.24.006148.

Peter Deines and Thomas C. G. Bosch. Transitioning from Microbiome Composition to Microbial Community Interactions: The Potential of the Metaorganism Hydra as an Experimental Model. *Frontiers in Microbiology*, 7, October 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.01610. URL http://journal.frontiersin.org/article/10.3389/fmicb.2016.01610/full.

Gloria del Solar, Rafael Giraldo, María Jesús Ruiz-Echevarría, Manuel Espinosa, and Ramón Díaz-Orejas. Replication and Control of Circular Bacterial Plasmids. *Microbiology and Molecular Biology Reviews*, 62(2):434–464, June 1998. ISSN 1098-5557, 1092-2172. doi: 10.1128/MMBR.62.2.434-464.1998. URL https://MMBR.asm.org/content/62/2/434.

F. Delogu, B. J. Kunath, P. N. Evans, M. Ø. Arntzen, T. R. Hvidsten, and P. B. Pope. Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nature Communications*, 11(1):4708, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18543-0. URL http://www.nature.com/articles/s41467-020-18543-0.

Julián R. Dib, Martin Wagenknecht, María E. Farías, and Friedhelm Meinhardt. Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements? *Frontiers in Microbiology*, 6:463, 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00463. URL https://www.frontiersin.org/article/10.3389/fmicb.2015.00463.

T. Dimitriu, C. Lotton, J. Benard-Capelle, D. Misevic, S. P. Brown, A. B. Lindner, and F. Taddei. Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences*, 111(30):11103–11108, July 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1406840111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1406840111.

Shaillay Kumar Dogra, Joel Doré, and Sami Damak. Gut Microbiota Resilience: Definition, Link to Health and Strategies for Intervention. *Frontiers in Microbiology*, 11:572921, September 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.572921. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.572921/full.

Anders B Dohlman and Xiling Shen. Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. *Experimental Biology and Medicine*, 244(6):445–458, April 2019. ISSN 1535-3702, 1535-3699. doi: 10.1177/1535370219836771. URL http://journals.sagepub.com/doi/10.1177/1535370219836771.

C. F. Dormann, B. Gruber, and J. Fruend. Introducing the bipartite Package: Analysing Ecological Networks. *R News*, 8(2):8–11, 2008.

C L Dupont, D B Rusch, S Yooseph, M J Lombardo, R A Richter, R Valas, M Novotny, J Yee-Greenbaum, J D Selengut, D H Haft, A L Halpern, R S Lasken, K Nealson, R Friedman, and J C Venter. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J*, 6, 2012. doi: 10.1038/ismej.2011.189. URL https://doi.org/10.1038/ismej.2011.189.

Claire Duvallet, Sean M. Gibbons, Thomas Gurry, Rafael A. Irizarry, and Eric J. Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1):1784, December 2017. ISSN 2041-1723. doi:

10.1038/s41467-017-01973-8. URL http://www.nature.com/articles/s41467-017-01973-8.

Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, October 2010. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btq461. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq461.

Robert A. Edwards, Katelyn McNair, Karoline Faust, Jeroen Raes, and Bas E. Dutilh. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, 40(2):258–272, March 2016. ISSN 0168-6445. doi: 10.1093/femsre/fuv048. URL https://doi.org/10.1093/femsre/fuv048.

Raphael Eisenhofer, Jeremiah J. Minich, Clarisse Marotz, Alan Cooper, Rob Knight, and Laura S. Weyrich. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.*, 2:105–117, 2019. doi: 10.1016/j.tim.2018.11.003. URL https://doi.org/10.1016/j.tim.2018.11.003.

Mohamed Elfil, Serageldin Kamel, Mohamed Kandil, Brian B. Koo, and Sara M. Schaefer. Implications of the Gut Microbiome in Parkinson's Disease. *Movement Disorders*, 35 (6):921–933, June 2020. ISSN 0885-3185, 1531-8257. doi: 10.1002/mds.28004. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.28004.

Alison R. Erickson, Brandi L. Cantarel, Regina Lamendella, Youssef Darzi, Emmanuel F. Mongodin, Chongle Pan, Manesh Shah, Jonas Halfvarson, Curt Tysk, Bernard Henrissat, Jeroen Raes, Nathan C. Verberkmoes, Claire M. Fraser, Robert L. Hettich, and Janet K. Jansson. Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS ONE*, 7(11):e49138, November 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0049138. URL https://dx.plos.org/10.1371/journal.pone.0049138.

Jacob T Evans and Vincent J Denef. To Dereplicate or Not To Dereplicate? *mSphere*, 5(3):e00971–19, 2020. doi: 10.1128/mSphere.00971-19.

Guilhem Faure, Sergey A. Shmakov, Winston X. Yan, David R. Cheng, David A. Scott, Joseph E. Peters, Kira S. Makarova, and Eugene V. Koonin. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nature Reviews Microbiology*, 17(8):

513–525, August 2019. ISSN 1740-1534. doi: 10.1038/s41579-019-0204-7. URL https://doi.org/10.1038/s41579-019-0204-7.

Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, August 2012. ISSN 1740-1526, 1740-1534. doi: 10.1038/nrmicro2832. URL http://www.nature.com/articles/nrmicro2832.

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):120, December 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0496-2. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0496-2.

Charles K. Fisher and Pankaj Mehta. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS ONE*, 9(7):e102451, July 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0102451. URL https://dx.plos.org/10.1371/journal.pone.0102451.

Marco Fondi and Pietro Liò. Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. *Microbiological Research*, 171:52–64, February 2015. ISSN 09445013. doi: 10.1016/j.micres.2015.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S094450131500004X.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL http://www.jstatsoft.org/v33/i01/.

Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23):3150–3152, December 2012. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/bts565.

Jed A. Fuhrman, Ian Hewson, Michael S. Schwalbach, Joshua A. Steele, Mark V. Brown, and Shahid Naeem. Annually reoccurring bacterial communities are predictable from

ocean conditions. *Proceedings of the National Academy of Sciences*, 103(35):13104–13109, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0602399103. URL https://www.pnas.org/content/103/35/13104. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/103/35/13104.full.pdf.

Georg K. Gerber. The dynamic microbiome. *FEBS Letters*, 588(22):4131–4139, November 2014. ISSN 00145793. doi: 10.1016/j.febslet.2014.02.037. URL http://doi.wiley.com/10.1016/j.febslet.2014.02.037.

Pourya Gholizadeh, Şükran Köse, Sounkalo Dao, Khudaverdi Ganbarov, Asghar Tanomand, Tuba Dal, Mohammad Aghazadeh, Reza Ghotaslou, Mohammad Ahangarzadeh Rezaee, Bahman Yousefi, and Hossein Samadi Kafil. How CRISPR-Cas System Could Be Used to Combat Antimicrobial Resistance. *Infection and drug resistance*, 13:1111–1121, April 2020. ISSN 1178-6973. doi: 10.2147/IDR.S247271. URL https://pubmed.ncbi.nlm.nih.gov/32368102. Publisher: Dove.

Sean M. Gibbons, Claire Duvallet, and Eric J. Alm. Correcting for batch effects in case-control microbiome studies. *PLOS Computational Biology*, 14(4):e1006102, April 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006102. URL https://dx.plos.org/10.1371/journal.pcbi.1006102.

Molly K Gibson, Kevin J Forsberg, and Gautam Dantas. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, 9(1):207–216, January 2015. ISSN 1751-7370. doi: 10.1038/ismej.2014.106. URL https://pubmed.ncbi.nlm.nih.gov/25003965. Edition: 2014/07/08 Publisher: Nature Publishing Group.

Jack A Gilbert, Joshua A Steele, J Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, Alice C McHardy, Rob Knight, Ian Joint, Paul Somerfield, Jed A Fuhrman, and Dawn Field. Defining seasonal marine microbial community dynamics. *The ISME Journal*, 6(2):298–308, February 2012. ISSN 1751-7370. doi: 10.1038/ismej.2011.107. URL https://doi.org/10.1038/ismej.2011.107.

Jesse Gillis and Paul Pavlidis. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. *PLOS Computational Biology*, 8(3):1–13, 2012. doi: 10.1371/journal.pcbi.1002444. URL https://doi.org/10.1371/journal.pcbi.1002444. Publisher: Public Library of Science.

Bettina Glasl, David G. Bourne, Pedro R. Frade, Torsten Thomas, Britta Schaffelke, and Nicole S. Webster. Microbial indicators of environmental perturbations in coral reef ecosystems. *Microbiome*, page 13, 2019. URL https://doi.org/10.1186/s40168-019-0705-7.

Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8:2224, November 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02224. URL http://journal.frontiersin.org/article/10.3389/fmicb.2017.02224/full.

Tao Gong, Jumei Zeng, Boyu Tang, Xuedong Zhou, and Yuqing Li. CRISPR-Cas systems in oral microbiome: From immune defense to physiological regulation. *Molecular Oral Microbiology*, 35(2):41–48, 2020. doi: https://doi.org/10.1111/omi.12279. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/omi.12279. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/omi.12279.

Didier Gonze, Leo Lahti, Jeroen Raes, and Karoline Faust. Multi-stability and the origin of microbial community types. *The ISME Journal 2017 11:10*, 11(10):2159, May 2017. ISSN 1751-7370. doi: 10.1038/ismej.2017.60. URL http://www.nature.com/doifinder/10.1038/ismej.2017.60. Publisher: Nature Publishing Group.

Didier Gonze, Katharine Z Coyte, Leo Lahti, and Karoline Faust. Microbial communities as dynamical systems. *Current Opinion in Microbiology*, 44:41–49, August 2018. ISSN 13695274. doi: 10.1016/j.mib.2018.07.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1369527418300092.

Elaina D. Graham, John F. Heidelberg, and Benjamin J. Tully. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ*, 5:e3035, March 2017. ISSN 2167-8359. doi: 10.7717/peerj.3035. URL https://peerj.com/articles/3035.

Garrett Grolemund and Hadley Wickham. Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 14(3):25, 2011.

Javier Guerrero, Albert Guisasola, and Juan A. Baeza. The nature of the carbon source rules the competition between PAO and denitrifiers in systems for simultaneous bio-

logical nitrogen and phosphorus removal. *Water Research*, 45(16):4793–4802, October 2011. ISSN 00431354. doi: 10.1016/j.watres.2011.06.019. URL https://linkinghub.elsevier.com/retrieve/pii/S004313541100354X.

José M. Gómez, Miguel Verdú, and Francisco Perfectti. Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature*, 465(7300):918–921, June 2010. ISSN 1476-4687. doi: 10.1038/nature09113. URL https://doi.org/10.1038/nature09113.

Rebecca J. Hall, Fiona J. Whelan, James O. McInerney, Yaqing Ou, and Maria Rosa Domingo-Sananes. Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes. *Frontiers in Microbiology*, 11:1569, July 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.01569. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.01569/full.

Hanjeong Harvey, Joseph Bondy-Denomy, Hélène Marquis, Kristina M. Sztanko, Alan R. Davidson, and Lori L. Burrows. Pseudomonas aeruginosa defends against phages through type IV pilus glycosylation. *Nature Microbiology*, 3(1):47–52, January 2018. ISSN 2058-5276. doi: 10.1038/s41564-017-0061-y. URL https://doi.org/10.1038/s41564-017-0061-y.

Graham F Hatfull and Roger W Hendrix. Bacteriophages and their genomes. *Current Opinion in Virology*, 1(4):298–303, October 2011. ISSN 18796257. doi: 10.1016/j.coviro.2011.06.009. URL https://linkinghub.elsevier.com/retrieve/pii/S1879625711000514.

Iain D Hay and Trevor Lithgow. Filamentous phages: masters of a microbial sharing economy. *EMBO reports*, 20(6), June 2019. ISSN 1469-221X, 1469-3178. doi: 10.15252/embr.201847427. URL https://onlinelibrary.wiley.com/doi/abs/10.15252/embr.201847427.

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(October):16180, October 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2016.180. URL http://www.

nature.com/articles/nmicrobiol2016180. Publisher: Nature Publishing Group.

Christian Hennig. *fpc: Flexible Procedures for Clustering*. 2020. URL https://CRAN.R-project.org/package=fpc.

Damayanthi Herath, Sen-Lin Tang, Kshitij Tandon, David Ackland, and Saman Kumara Halgamuge. CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics*, 18 (S16):571, December 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1967-3. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1967-3.

Malte Herold, Susana Martínez Arbas, Shaman Narayanasamy, Abdul R. Sheik, Luise A. K. Kleine-Borgmann, Laura A. Lebrun, Benoît J. Kunath, Hugo Roume, Irina Bessarab, Rohan B. H. Williams, John D. Gillece, James M. Schupp, Paul S. Keim, Christian Jäger, Michael R. Hoopmann, Robert L. Moritz, Yuzhen Ye, Sujun Li, Haixu Tang, Anna Heintz-Buschart, Patrick May, Emilie E. L. Muller, Cedric C. Laczny, and Paul Wilmes. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nature Communications*, 11(1):5281, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19006-2. URL http://www.nature.com/articles/s41467-020-19006-2.

Tim A. Hoek, Kevin Axelrod, Tommaso Biancalani, Eugene A. Yurtsev, Jinghui Liu, and Jeff Gore. Resource Availability Modulates the Cooperative and Competitive Nature of a Microbial Cross-Feeding Mutualism. *PLOS Biology*, 14(8):e1002540, August 2016. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002540. URL https://dx.plos.org/10.1371/journal.pbio.1002540.

Sohyun Hwang, Chan Yeong Kim, Sunmo Yang, Eiru Kim, Traver Hart, Edward M Marcotte, and Insuk Lee. HumanNet v2: human gene networks for disease research. *Nucleic acids research*, 47(D1):D573–D580, January 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1126. URL https://pubmed.ncbi.nlm.nih.gov/30418591. Publisher: Oxford University Press.

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(1):119, December 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-119.

Michihiko Ike, Hitomi Suzuki, and Masanori Fujita. Distribution of Bacterial Plasmids in an Activated Sludge Plant. *Japanese Journal of Water Treatment Biology*, 30(2):65–71, 1994. ISSN 0910-6758, 1881-0438. doi: 10.2521/jswtb.30.65. URL http://www.jstage.jst.go.jp/article/jswtb1964/30/2/30_2_65/_article.

Takeru Ishige, Akio Tani, Keiji Takabe, Kazunori Kawasaki, Yasuyoshi Sakai, and Nobuo Kato. Wax Ester Production from n-Alkanes by Acinetobacter sp. Strain M-1: Ultrastructure of Cellular Inclusions and Role of Acyl Coenzyme A Reductase. *Applied and Environmental Microbiology*, 68(3):1192–1195, 2002. ISSN 0099-2240. doi: 10.1128/AEM.68.3.1192-1195.2002. URL https://aem.asm.org/content/68/3/1192. Publisher: American Society for Microbiology Journals _eprint: https://aem.asm.org/content/68/3/1192.full.pdf.

Alexander L. Jaffe, Eduardo Corel, Jananan Sylvestre Pathmanathan, Philippe Lopez, and Eric Bapteste. Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins: LGT among ultrasmall prokaryotes. *Environmental Microbiology*, 18(12):5072–5081, December 2016. ISSN 14622912. doi: 10.1111/1462-2920.13477. URL https://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13477.

Ruud Jansen, Jan D A van Embden, Wim Gaastra, and Leo M Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology*, 43 (6):1565–1575, March 2002. ISSN 0950-382X (Print).

Sabah A A Jassim, Richard G Limoges, and Hassan El-Cheikh. Bacteriophage biocontrol in wastewater treatment. *World journal of microbiology & biotechnology*, 32(4):70, April 2016. ISSN 1573-0972 (Electronic). doi: 10.1007/s11274-016-2028-1.

Wenyan Jiang, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R. Levin, and Luciano A. Marraffini. Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. *PLoS Genetics*, 9(9):e1003844, September 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003844.

Shuo Jiao, Weimin Chen, and Gehong Wei. Resilience and Assemblage of Soil Microbiome in Response to Chemical Contamination Combined with Plant Growth. *Applied and Environmental Microbiology*, 85(6):e02523–18, /aem/85/6/AEM.02523–18.atom, January 2019. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.02523-18. URL http://aem.asm.org/lookup/doi/10.1128/AEM.02523-18.

L Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11:431, August 2010. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-11-431.

Juliet Johnston and Sebastian Behrens. Seasonal Dynamics of the Activated Sludge Microbiome in Sequencing Batch Reactors, Assessed Using 16S rRNA Transcript Amplicon Sequencing. *Applied and Environmental Microbiology*, 86(19):e00597–20, /aem/86/19/AEM.00597–20.atom, July 2020. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.00597-20. URL https://aem.asm.org/content/86/19/e00597-20.

Juliet Johnston, Timothy LaPara, and Sebastian Behrens. Composition and Dynamics of the Activated Sludge Microbiome during Seasonal Nitrification Failure. *Scientific Reports*, 9(1):4565, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40872-4. URL https://doi.org/10.1038/s41598-019-40872-4.

Patricia A. Jones and Andrew J. Schuler. Seasonal variability of biomass density and activated sludge settleability in full-scale wastewater treatment systems. *Chemical Engineering Journal*, 164(1):16–22, October 2010. ISSN 1385-8947. doi: 10.1016/j.cej.2010.07.061. URL https://www.sciencedirect.com/science/article/pii/S1385894710006807.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1070. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1070.

Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*,

3:e1165, August 2015. ISSN 2167-8359. doi: 10.7717/peerj.1165. URL https://doi.org/10.7717/peerj.1165.

Rahul Vijay Kapoore and Seetharaman Vaidyanathan. Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2079):20150363, October 2016. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2015.0363. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0363.

Zahra Karimi, Ali Ahmadi, Ali Najafi, and Reza Ranjbar. Bacterial CRISPR Regions: General Features and their Potential for Epidemiological Molecular Typing Studies. *The Open Microbiology Journal*, 12(1):59–70, April 2018. ISSN 1874-2858. doi: 10.2174/1874285801812010059. URL https://openmicrobiologyjournal.com/VOLUME/12/PAGE/59/.

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, August 2017. ISSN 1878-1810. doi: 10.1016/j.trsl.2017.06.008. Publisher: Elsevier.

Krishna Khairnar, Rajshree Chandekar, Aparna Nair, Preeti Pal, and Waman N. Paunikar. Novel application of bacteriophage for controlling foaming in wastewater treatment plant- an eco-friendly approach. *Bioengineered*, 7(1):46–49, 2016. doi: 10.1080/21655979.2015.1134066. URL https://doi.org/10.1080/21655979.2015.1134066. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21655979.2015.1134066.

Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 26(12):1721–1729, December 2016. ISSN 1549-5469 1088-9051. doi: 10.1101/gr.210641.116.

Hanhae Kim, Junha Shin, Eiru Kim, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. YeastNet v3: a public database of data-specific and integrated func-

tional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, 42(D1): D731–D736, 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt981. URL https://doi.org/10.1093/nar/gkt981. _eprint: https://academic.oup.com/nar/article-pdf/42/D1/D731/16803898/gkt981.pdf.

Hanhae Kim, Jung Eun Shim, Junha Shin, and Insuk Lee. EcoliNet: a database of cofunctional gene network for Escherichia coli. *Database : the journal of biological databases and curation*, 2015, 2015. ISSN 1758-0463. doi: 10.1093/database/bav001.

Sangtae Kim and Pavel A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5(1):5277, December 2014. ISSN 2041-1723. doi: 10.1038/ncomms6277. ISBN: 2041-1723 (Electronic)$\backslash$r2041-1723 (Linking) Publisher: Nature Publishing Group.

Eugene V. Koonin. Viruses and mobile elements as drivers of evolutionary transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1701): 20150442, August 2016. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2015.0442. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2015.0442.

Eugene V Koonin, Kira S Makarova, and Feng Zhang. Diversity, classification and evolution of CRISPR-Cas systems. *Current opinion in microbiology*, 37:67–78, June 2017. ISSN 1879-0364 (Electronic). doi: 10.1016/j.mib.2017.05.008.

Evguenia Kopylova, Laurent Noé, and Hélène Touzet. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, December 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts611. URL https://doi.org/10.1093/bioinformatics/bts611.

Britt Koskella and Sean Meaden. Understanding bacteriophage specificity in natural microbial communities. *Viruses*, 5(3):806–823, March 2013. ISSN 1999-4915 (Electronic). doi: 10.3390/v5030806.

Shireen M. Kotay, Tania Datta, Jeongdong Choi, and Ramesh Goel. Biocontrol of biomass bulking caused by Haliscomenobacter hydrossis using a newly isolated lytic bacteriophage. *Water Research*, 45(2):694–704, January 2011. ISSN 0043-1354. doi: 10.1016/j.watres.2010.08.038. URL https://www.sciencedirect.com/science/article/pii/S0043135410006020.

Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic acids research*, 46(6):e35, April 2018. ISSN 1362-4962 (Electronic). doi: 10.1093/nar/gkx1321.

Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008. ISSN 1548-7660. doi: 10.18637/jss.v028.i05. URL https://www.jstatsoft.org/v028/i05.

Ranjit Kumar, Peter Eipers, Rebecca B. Little, Michael Crowley, David K. Crossman, Elliot J. Lefkowitz, and Casey D. Morrow. Getting Started with Microbiome Analysis: Sample Acquisition to Bioinformatics. *Current Protocols in Human Genetics*, 82 (1), July 2014. ISSN 1934-8266, 1934-8258. doi: 10.1002/0471142905.hg1808s82. URL https://onlinelibrary.wiley.com/doi/10.1002/0471142905.hg1808s82.

Victor Kunin, Shaomei He, Falk Warnecke, S. Brook Peterson, Hector Garcia Martin, Matthew Haynes, Natalia Ivanova, Linda L. Blackall, Mya Breitbart, Forest Rohwer, Katherine D. McMahon, and Philip Hugenholtz. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome research*, 18(2):293–297, February 2008. ISSN 1088-9051. doi: 10.1101/gr.6835308. ISBN: 1088-9051 (Print)$\backslash$n1088-9051 (Linking) Publisher: Cold Spring Harbor Laboratory Press.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13.

Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, October 2012. ISSN 1367-4803. URL http://dx.doi.org/10.1093/bioinformatics/bts480.

Cedric C Laczny, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Hera Margossian, Sergio Coronado, Laurens van 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. URL http://www.microbiomejournal.com/content/3/1/1. ISBN: 4016801400.

Leo Lahti, Jarkko Salojärvi, Anne Salonen, Marten Scheffer, and Willem M. de Vos. Tipping elements in the human intestinal ecosystem. *Nature Communications*, 5 (1):4344, September 2014. ISSN 2041-1723. doi: 10.1038/ncomms5344. URL http://www.nature.com/articles/ncomms5344.

Tony J. Lam and Yuzhen Ye. Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. *BMC Genomics*, 20(1):567, December 2019. ISSN 1471-2164. doi: 10.1186/s12864-019-5922-8. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5922-8.

Yingyu Law, Rasmus Hansen Kirkegaard, Angel Anisa Cokro, Xianghui Liu, Krithika Arumugam, Chao Xie, Mikkel Stokholm-Bjerregaard, Daniela I. Drautz-Moses, Per Halkjær Nielsen, Stefan Wuertz, and Rohan B. H. Williams. Integrative microbial community analysis reveals full-scale enhanced biological phosphorus removal under tropical conditions. *Scientific Reports*, 6(1):25719, May 2016. ISSN 2045-2322. doi: 10.1038/srep25719. URL http://www.nature.com/articles/srep25719.

Sébastien Leclercq, Clément Gilbert, and Richard Cordaux. Cargo capacity of phages and plasmids and other factors influencing horizontal transfers of prokaryote transposable elements. *Mobile genetic elements*, 2(2):115–118, March 2012. ISSN 2159-2543. doi: 10.4161/mge.20352. Publisher: Landes Bioscience.

K S Lee, W W Metcalf, and B L Wanner. Evidence for two phosphonate degradative pathways in Enterobacter aerogenes. *Journal of Bacteriology*, 174(8):2501–2510, 1992. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.174.8.2501-2510.1992. URL https://JB.asm.org/content/174/8/2501.

Tak Lee, Sunmo Yang, Eiru Kim, Younhee Ko, Sohyun Hwang, Junha Shin, Jung Eun Shim, Hongseok Shim, Hyojin Kim, Chanyoung Kim, and Insuk Lee. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic acids research*, 43(Database issue):D996–1002, January 2015. ISSN 1362-4962 0305-1048. doi: 10.1093/nar/gku1053.

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tack-

ling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, October 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2825. URL http://www.nature.com/articles/nrg2825.

J. T. Lennon, M. E. Muscarella, S. A. Placella, and B. K. Lehmkuhl. How, When, and Where Relic DNA Affects Microbial Diversity. *mBio*, 9(3):e00637–18, /mbio/9/3/mBio.00637–18.atom, June 2018. ISSN 2150-7511. doi: 10.1128/mBio.00637-18. URL https://mbio.asm.org/content/9/3/e00637-18.

Raphaël Leplae, Aline Hebrant, Shoshana J. Wodak, and Ariane Toussaint. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Research*, 32(suppl_1): D45–D49, January 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh084. URL https://doi.org/10.1093/nar/gkh084.

Roie Levy, Andrew T. Magis, John C. Earls, Ohad Manor, Tomasz Wilmanski, Jennifer Lovejoy, Sean M. Gibbons, Gilbert S. Omenn, Leroy Hood, and Nathan D. Price. Longitudinal analysis reveals transition barriers between dominant ecological states in the gut microbiome. *Proceedings of the National Academy of Sciences*, 117(24):13839–13845, June 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1922498117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1922498117.

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, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btv033. ISBN: 1367-4803 _eprint: 1401.7457.

Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, June 2016. ISSN 10462023. doi: 10.1016/j.ymeth.2016.02.020. URL https://linkinghub.elsevier.com/retrieve/pii/S1046202315301183.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://doi.org/10.1093/bioinformatics/btp324.

Liguan Li, Arnaud Dechesne, Zhiming He, Jonas Stenløkke Madsen, Joseph Nesme, Søren J. Sørensen, and Barth F. Smets. Estimating the Transfer Range of Plasmids Encoding Antimicrobial Resistance in a Wastewater Treatment Plant Microbial Community. *Environmental Science and Technology Letters*, 5(5):260–265, 2018. ISSN 23288930. doi: 10.1021/acs.estlett.8b00105.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl158. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158.

Di Liu, Pangzhen Zhang, Deli Chen, and Kate Howell. From the Vineyard to the Winery: How Microbial Ecology Drives Regional Distinctiveness of Wine. *Frontiers in Microbiology*, 10:2679, November 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.02679. URL https://www.frontiersin.org/article/10.3389/fmicb.2019.02679/full.

Mei Liu, Jason J. Gill, Ry Young, and Elizabeth J. Summer. Bacteriophages of wastewater foaming-associated filamentous Gordonia reduce host levels in raw activated sludge. *Scientific Reports*, 5(1):13754, September 2015a. ISSN 2045-2322. doi: 10.1038/srep13754. URL https://doi.org/10.1038/srep13754.

Rui Liu, Pei Chen, Kazuyuki Aihara, and Luonan Chen. Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. *Scientific Reports*, 5(1):17501, December 2015b. ISSN 2045-2322. doi: 10.1038/srep17501. URL http://www.nature.com/articles/srep17501.

Ruyin Liu, Rong Qi, Juan Wang, Yu Zhang, Xinchun Liu, Simona Rossetti, Valter Tandoi, and Min Yang. Phage-host associations in a full-scale activated sludge plant during sludge bulking. *Applied microbiology and biotechnology*, 101(16):6495–6504, August 2017. ISSN 1432-0614 (Electronic). doi: 10.1007/s00253-017-8429-8.

Tang Liu, Shufeng Liu, Maosheng Zheng, Qian Chen, and Jinren Ni. Performance Assessment of Full-Scale Wastewater Treatment Plants Based on Seasonal Variability of Microbial Communities via High-Throughput Sequencing. *PLOS ONE*, page 15, 2016.

Jason Lloyd-Price, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, Nadim J. Ajami, Kevin S. Bonham, Colin J. Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G. Graeber, A. Brantley Hall, Kathleen Lake, Carol J. Landers, Himel Mallick, Damian R. Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A. White, Jonathan Braun, Lee A. Denson, Janet K. Jansson, Rob Knight, Subra Kugathasan, Dermot P. B. McGovern, Joseph F. Petrosino, Thaddeus S. Stappenbeck, Harland S. Winter, Clary B. Clish, Eric A. Franzosa, Hera Vlamakis, Ramnik J. Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, May 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1237-9. URL http://www.nature.com/articles/s41586-019-1237-9.

G T Macfarlane and J H Cummings. Probiotics and prebiotics: can regulating the activities of intestinal bacteria benefit health? *BMJ (Clinical research ed.)*, 318(7189):999–1003, April 1999. ISSN 0959-8138. doi: 10.1136/bmj.318.7189.999. URL https://pubmed.ncbi.nlm.nih.gov/10195977. Publisher: British Medical Journal.

Lakshmi Machineni. Effects of biotic and abiotic factors on biofilm growth dynamics and their heterogeneous response to antibiotic challenge. *Journal of Biosciences*, 45(1):25, January 2020. ISSN 0973-7138. doi: 10.1007/s12038-020-9990-3. URL https://doi.org/10.1007/s12038-020-9990-3.

Jonathan D. Magasin and Dietlind L. Gerloff. Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics*, 31(3):311–317, February 2015. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btu546. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu546.

Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35 (1):81–89, January 2017. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3703. URL http://www.nature.com/articles/nbt.3703.

Kira S Makarova and Eugene V Koonin. Annotation and Classification of CRISPR-Cas Systems. *Methods in molecular biology (Clifton, N.J.)*, 1311:47–75, 2015. ISSN 1940-6029 (Electronic). doi: 10.1007/978-1-4939-2687-9_4.

Kira S Makarova, Yuri I Wolf, John van der Oost, and Eugene V Koonin. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology Direct*, 4(1):29, 2009. ISSN 1745-6150. doi: 10.1186/1745-6150-4-29. URL http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-4-29.

Himel Mallick, Eric A. Franzosa, Lauren J. Mclver, Soumya Banerjee, Alexandra Sirota-Madi, Aleksandar D. Kostic, Clary B. Clish, Hera Vlamakis, and Curtis Huttenhower. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun*, 10:3136, 2019. doi: 10.1038/s41467-019-10927-1. URL https://doi.org/10.1038/s41467-019-10927-1.

Himel Mallick, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, George Weingart, Boyu Ren, Emma H. Schwager, Suvo Chatterjee, Kelsey N. Thompson, Jeremy E. Wilkinson, Ayshwarya Subramanian, Yiren Lu, Levi Waldron, Joseph N. Paulson, Eric A. Franzosa, Hector Corrada Bravo, and Curtis Huttenhower. Multivariable Association Discovery in Population-scale Meta-omics Studies. preprint, Microbiology, January 2021. URL http://biorxiv.org/lookup/doi/10.1101/2021.01.20.427420.

Luciano A Marraffini and Erik J Sontheimer. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics*, 11(3):181–190, March 2010. ISSN 1471-0064 (Electronic). doi: 10.1038/nrg2749.

Ruben A.T. Mars, Yi Yang, Tonya Ward, Mo Houtti, Sambhawa Priya, Heather R. Lekatz, Xiaojia Tang, Zhifu Sun, Krishna R. Kalari, Tal Korem, Yogesh Bhattarai, Tenghao Zheng, Noam Bar, Gary Frost, Abigail J. Johnson, Will van Treuren, Shuo Han, Tamas Ordog, Madhusudan Grover, Justin Sonnenburg, Mauro D'Amato, Michael Camilleri, Eran Elinav, Eran Segal, Ran Blekhman, Gianrico Farrugia, Jonathan R. Swann, Dan Knights, and Purna C. Kashyap. Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. *Cell*, 182(6): 1460–1473.e17, September 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.08.007.

URL https://www.sciencedirect.com/science/article/pii/S0092867420309983.

Belinda C Martin, Marta Sanchez Alarcon, Deirdre Gleeson, Jen A Middleton, Matthew W Fraser, Megan H Ryan, Marianne Holmer, Gary A Kendrick, and Kieryn Kilminster. Root microbiomes as indicators of seagrass health. *FEMS Microbiology Ecology*, 96 (2):fiz201, February 2020. ISSN 0168-6496, 1574-6941. doi: 10.1093/femsec/fiz201. URL https://academic.oup.com/femsec/article/doi/10.1093/femsec/fiz201/5679015.

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, June 2013. ISSN 1751-7370 1751-7362. doi: 10.1038/ismej.2013.6.

Milot Mirdita, Martin Steinegger, and Johannes Söding. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, August 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty1057. URL https://academic.oup.com/bioinformatics/article/35/16/2856/5280135.

Francisco J.M. Mojica, César Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 60(2):174–182, February 2005. ISSN 0022-2844. doi: 10.1007/s00239-004-0046-3. URL http://link.springer.com/10.1007/s00239-004-0046-3. Publisher: Springer-Verlag.

Abraham G Moller and Chun Liang. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ*, 5:e3788, 2017. ISSN 2167-8359 (Print). doi: 10.7717/peerj.3788.

Steffen Moritz and Thomas Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207, 2017. ISSN 2073-4859. doi: 10.32614/RJ-2017-009. URL https://journal.r-project.org/archive/2017/RJ-2017-009/index.html.

Emilie E L Muller. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *mSystems*, 4:e00080–19, 2019. doi: 10.1128/mSystems.00080-19. URL https://doi.org/10.1128/mSystems.00080-19.

Emilie E L Muller, Nicolás Pinel, John D. Gillece, James M. Schupp, Lance B. Price, David M. Engelthaler, Caterina Levantesi, Valter Tandoi, Khai Luong, Nitin S. Baliga, Jonas Korlach, Paul S. Keim, and Paul Wilmes. Genome sequence of "Candidatus Microthrix parvicella" Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *Journal of bacteriology*, 194 (23):6670–6671, December 2012. ISSN 1098-5530. doi: 10.1128/JB.01765-12.

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):325–333, 2013. ISSN 0966842X. doi: 10.1016/j.tim.2013.04.009. URL http://dx.doi.org/10.1016/j.tim.2013.04.009. ISBN: 1878-4380 (Electronic)$\backslash$n0966-842X (Linking) Publisher: Elsevier Ltd.

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, December 2014a. ISSN 2041-1723. doi: 10.1038/ncomms6603. URL http://www.nature.com/articles/ncomms6603.

Emilie E. L. Muller, Shaman Narayanasamy, Myriam Zeimes, C.C. Cédric C. Laczny, L.A. Laura A. Lebrun, Malte Herold, N.D. Nathan D. Hicks, John D. J.D. Gillece, James M. Schupp, Paul Keim, Paul Wilmes, O Gascuel, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, G Sherlock, C Reich, R Stevens, O Vassieva, V Vonstein, A Wilke, O Zagnitko, P Goldstein, R Guralnick, D Haft, and D Hancock. First draft genome sequence of a strain belonging to the Zoogloea genus and its gene expression in situ. *Standards in Genomic Sciences*, 12(64), 2017. ISSN 1944-3277. doi: 10.1186/s40793-017-0274-y. Publisher: BioMed Central.

Emilie EL Muller, Abdul R Sheik, and Paul Wilmes. Lipid-based biofuel production from wastewater. *Current Opinion in Biotechnology*, 30:9–16, December 2014b. ISSN

09581669. doi: 10.1016/j.copbio.2014.03.007. URL https://linkinghub.elsevier.com/retrieve/pii/S095816691400072X.

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, April 2018. ISSN 24523100. doi: 10.1016/j.coisb.2017.12.004. URL https://linkinghub.elsevier.com/retrieve/pii/S245231001730197X.

Aimee K. Murray, Lihong Zhang, Xiaole Yin, Tong Zhang, Angus Buckling, Jason Snape, and William H. Gaze. Novel insights into selection for antibiotic resistance in complex microbial communities. *mBio*, 9(4), July 2018. ISSN 21507511. doi: 10.1128/mBio.00969-18. Publisher: American Society for Microbiology.

Giovanni Musso, Roberto Gambino, and Maurizio Cassader. Interactions Between Gut Microbiota and Host Metabolism Predisposing to Obesity and Diabetes. *Annual Review of Medicine*, 62(1):361–380, January 2011. ISSN 0066-4219. doi: 10.1146/annurev-med-012510-175505. URL https://doi.org/10.1146/annurev-med-012510-175505. Publisher: Annual Reviews.

Arun M. Nanda, Kai Thormann, and Julia Frunzke. Impact of Spontaneous Prophage Induction on the Fitness of Bacterial Populations and Host-Microbe Interactions. *Journal of Bacteriology*, 197(3):410–419, February 2015. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.02230-14. URL https://jb.asm.org/content/197/3/410.

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, Luxembourg, 2017. URL https://orbilu.uni.lu/handle/10993/29800.

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):363–8, May 2015. ISSN 1751-7915. doi: 10.1111/1751-7915.12255. URL http://www.ncbi.nlm.nih.gov/pubmed/25678254http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4408170. Publisher: Wiley-Blackwell.

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, December 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1116-8. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1116-8.

M. E. J. Newman. Modularity and community structure in networks. *Communications Law*, 19(2):56–62, 2006. ISSN 17467616. doi: 10.1073/pnas.122653799. ISBN: 0027-8424 (Print)$\backslash$r0027-8424 (Linking) _eprint: 0602124.

Cecilia Noecker, Alexander Eng, Sujatha Srinivasan, Casey M. Theriot, Vincent B. Young, Janet K. Jansson, David N. Fredricks, and Elhanan Borenstein. Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation. *mSystems*, 1(1): mSystems.00013–15, e00013–15, February 2016. ISSN 2379-5077. doi: 10.1128/mSystems.00013-15. URL https://msystems.asm.org/content/1/1/e00013-15.

Elad Noor, Sarah Cherkaoui, and Uwe Sauer. Biological insights through omics data integration. *Gene regulation*, 15:39–47, June 2019. ISSN 2452-3100. doi: 10.1016/j.coisb.2019.03.007. URL https://www.sciencedirect.com/science/article/pii/S2452310019300125.

Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, and Helene Wagner. *vegan: Community Ecology Package*. 2019. URL https://CRAN.R-project.org/package=vegan.

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, December 2017. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2017.126. URL http://www.nature.com/articles/ismej2017126.

Ivan Olovnikov, Ken Chan, Ravi Sachidanandam, Dianne K. Newman, and Alexei A. Aravin. Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA. *Molec-*

ular Cell, 51(5):594–605, September 2013. ISSN 1097-2765. doi: 10.1016/j.molcel.2013.08.014. URL https://doi.org/10.1016/j.molcel.2013.08.014. Publisher: Elsevier.

Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science (New York, N.Y.)*, 355(6322):294–298, January 2017. ISSN 1095-9203 0036-8075. doi: 10.1126/science.aah4043.

Seo-Young Park, Arinzechukwu Ufondu, Kyongbum Lee, and Arul Jayaraman. Emerging computational tools and models for studying gut microbiota composition and function. *Tissue, Cell and Pathway Engineering*, 66:301–311, December 2020. ISSN 0958-1669. doi: 10.1016/j.copbio.2020.10.005. URL https://www.sciencedirect.com/science/article/pii/S0958166920301543.

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, July 2015. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.186072.114. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.186072.114.

David Pellow, Itzik Mizrahi, and Ron Shamir. PlasClass improves plasmid sequence classification. *PLOS Computational Biology*, 16(4):1–9, 2020. doi: 10.1371/journal.pcbi.1007781. URL https://doi.org/10.1371/journal.pcbi.1007781. Publisher: Public Library of Science.

María Florencia Perez, Daniel Kurth, María Eugenia Farías, Mariana Noelia Soria, Genis Andrés Castillo Villamizar, Anja Poehlein, Rolf Daniel, and Julián Rafael Dib. First Report on the Plasmidome From a High-Altitude Lake of the Andean Puna. *Frontiers in Microbiology*, 11:1343, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020.01343. URL https://www.frontiersin.org/article/10.3389/fmicb.2020.01343.

Samuel Peña-Llopis and James Brugarolas. Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications. *Nat Protoc.*, 8(11):2240–2255, 2013. doi: 10.1038/nprot.2013.141.

Olivier Pible, François Allain, Virginie Jouffret, Karen Culotta, Guylaine Miotello, and Jean Armengaud. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome*, 8(1):30, December 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00797-x. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00797-x.

Markus Pötter, Mohamed H. Madkour, Frank Mayer, and Alexander Steinbüchel. Regulation of phasin expression and polyhydroxyalkanoate (PHA) granule formation in Ralstonia eutropha H16. *Microbiology (Reading, England)*, 148(Pt 8):2413–2426, August 2002. ISSN 1350-0872. doi: 10.1099/00221287-148-8-2413. Place: England.

Jia Qian and Matteo Comin. MetaCon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage. *BMC Bioinformatics*, 20 (S9):367, November 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2904-4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2904-4.

A R Quinlan and I M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 2010. doi: 10.1093/bioinformatics/btq033. URL https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

Frédéric Raymond, Amin A Ouameur, Maxime Déraspe, Naeem Iqbal, Hélène Gingras, Bédis Dridi, Philippe Leprohon, Pier-Luc Plante, Richard Giroux, Ève Bérubé, Johanne Frenette, Dominique K Boudreau, Jean-Luc Simard, Isabelle Chabot, Marc-Christian Domingo, Sylvie Trottier, Maurice Boissinot, Ann Huletsky, Paul H Roy, Marc Ouellette, Michel G Bergeron, and Jacques Corbeil. The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME Journal*, 10(3):707–720, March 2016. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2015.148. URL http://www.nature.com/articles/ismej2015148.

Dominik Refardt. Within-host competition determines reproductive success of temperate bacteriophages. *The ISME Journal*, 5(9):1451–1460, September 2011. ISSN 1751-

7362, 1751-7370. doi: 10.1038/ismej.2011.30. URL http://www.nature.com/articles/ismej201130.

Jie Ren, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1):69, 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0283-5. URL http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0283-5. ISBN: 4016801702 Publisher: Microbiome.

William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2020. URL https://CRAN.R-project.org/package=psych.

Mina Rho, Haixu Tang, and Yuzhen Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20):e191–e191, November 2010. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkq747. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq747.

Mina Rho, Yu Wei Wu, Haixu Tang, Thomas G. Doak, and Yuzhen Ye. Diverse CRISPRs evolving in human microbiomes. *PLoS Genetics*, 8(6), 2012. ISSN 15537390. doi: 10.1371/journal.pgen.1002441. ISBN: 1553-7404 (Electronic)$\backslash$n1553-7390 (Linking).

Ana B. Rios Miguel, Mike S.M. Jetten, and Cornelia U. Welte. The role of mobile genetic elements in organic micropollutant degradation during biological wastewater treatment. *Water Research X*, 9:100065, December 2020. ISSN 25899147. doi: 10.1016/j.wroa.2020.100065. URL https://linkinghub.elsevier.com/retrieve/pii/S2589914720300256.

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. *The Science of the total environment*, 447:345–360, March 2013. ISSN 1879-1026 (Electronic). doi: 10.1016/j.scitotenv.2013.01.032.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinfor-*

matics, 26(1):139–140, November 2009. ISSN 14602059. doi: 10.1093/bioinformatics/btp616. Publisher: Oxford University Press.

Serina L. Robinson, Jörn Piel, and Shinichi Sunagawa. A roadmap for metagenomic enzyme discovery. *Natural Product Reports*, page 10.1039.D1NP00006C, 2021. ISSN 0265-0568, 1460-4752. doi: 10.1039/D1NP00006C. URL http://xlink.rsc.org/?DOI=D1NP00006C.

Bruno K. Rodiño-Janeiro, María Vicario, Carmen Alonso-Cotoner, Roberto Pascua-García, and Javier Santos. A Review of Microbiota and Irritable Bowel Syndrome: Future in Therapies. *Advances in Therapy*, 35(3):289–310, March 2018. ISSN 1865-8652. doi: 10.1007/s12325-018-0673-5. URL https://doi.org/10.1007/s12325-018-0673-5.

Luis M Rodriguez-R and Konstantinos T Konstantinidis. Estimating coverage in metagenomic data sets and why it matters. *The ISME journal*, 8(11):1–3, May 2014a. ISSN 1751-7370. doi: 10.1038/ismej.2014.76. Publisher: Nature Publishing Group.

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, March 2014b. ISSN 1460-2059. doi: 10.1093/bioinformatics/btt584. Publisher: Oxford University Press.

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, January 2005. ISSN 1574-6976. doi: 10.1016/j.femsre.2004.09.005. URL https://academic.oup.com/femsre/article-lookup/doi/10.1016/j.femsre.2004.09.005.

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, January 2013. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2012.72. URL http://www.nature.com/articles/ismej201272.

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, December 2015. ISSN 2055-5008. doi: 10.1038/npjbiofilms.2015.7. URL http://www.nature.com/articles/npjbiofilms20157.

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. URL https://peerj.com/articles/985.

Roye Rozov, Aya Brown Kav, David Bogumil, Naama Shterzer, Eran Halperin, Itzhak Mizrahi, and Ron Shamir. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics (Oxford, England)*, 33(4):475–482, February 2017. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btw651.

Daniel Ruiz-Perez, Jose Lugo-Martinez, Natalia Bourguignon, Kalai Mathee, Betiana Lerner, Ziv Bar-Joseph, and Giri Narasimhan. Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data. *mSystems*, 6(2):e01105–20, /msystems/6/2/mSys.01105–20.atom, March 2021. ISSN 2379-5077. doi: 10.1128/mSystems.01105-20. URL https://msystems.asm.org/content/6/2/e01105-20.

Sergei Ryazansky, Andrey Kulbachinskiy, and Alexei A Aravin. The Expanded Universe of Prokaryotic Argonaute Proteins. *mBio*, 9(6):e01935–18, December 2018. ISSN 2150-7511. doi: 10.1128/mBio.01935-18. URL https://pubmed.ncbi.nlm.nih.gov/30563906. Publisher: American Society for Microbiology.

Lisa Röttjers, Doris Vandeputte, Jeroen Raes, and Karoline Faust. A framework for comparing microbial networks reveals core associations. *bioRxiv*, page 2020.10.05.325860, January 2020. doi: 10.1101/2020.10.05.325860. URL http://biorxiv.org/content/early/2020/10/05/2020.10.05.325860.abstract.

Lisa M. Røst, Lilja Brekke Thorfinnsdottir, Kanhaiya Kumar, Katsuya Fuchino, Ida Eide Langørgen, Zdenka Bartosova, Kåre Andre Kristiansen, and Per Bruheim. Absolute Quantification of the Central Carbon Metabolome in Eight Commonly Applied Prokaryotic and Eukaryotic Model Systems. *Metabolites*, 10(2):74, February 2020.

ISSN 2218-1989. doi: 10.3390/metabo10020074. URL https://www.mdpi.com/2218-1989/10/2/74.

Julie E. Samson, Alfonso H. Magadán, Mourad Sabri, and Sylvain Moineau. Revenge of the phages: defeating bacterial defences. *Nature Reviews Microbiology*, 11(10): 675–687, October 2013. ISSN 1740-1534. doi: 10.1038/nrmicro3096. URL https://doi.org/10.1038/nrmicro3096.

Alvaro San Millan and R. Craig MacLean. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiology spectrum*, 5(5), September 2017. ISSN 2165-0497. doi: 10.1128/microbiolspec.MTBP-0016-2017. Place: United States.

M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van de Koppel, I. A. van de Leemput, S. A. Levin, E. H. van Nes, M. Pascual, and J. Vandermeer. Anticipating Critical Transitions. *Science*, 338(6105):344–348, October 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1225244. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1225244.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, September 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08227. URL http://www.nature.com/articles/nature08227.

Michael Schloter, Paolo Nannipieri, Søren J. Sørensen, and Jan Dirk van Elsas. Microbial indicators for soil quality. *Biology and Fertility of Soils*, 54(1):1–10, January 2018. ISSN 0178-2762, 1432-0789. doi: 10.1007/s00374-017-1248-3. URL http://link.springer.com/10.1007/s00374-017-1248-3.

Bastian Seelbinder, Jiarui Chen, Sascha Brunke, Ruben Vazquez-Uribe, Rakesh Santhaman, Anne-Christin Meyer, Felipe Senne de Oliveira Lino, Ka-Fai Chan, Daniel Loos, Lejla Imamovic, Chi-Ching Tsang, Rex Pui-kin Lam, Siddharth Sridhar, Kang Kang, Bernhard Hube, Patrick Chiu-yat Woo, Morten Otto Alexander Sommer, and Gianni Panagiotou. Antibiotics create a shift from mutualism to competition in human gut communities with a longer-lasting impact on fungi than bacteria. *Microbiome*, 8(1):133, December 2020. ISSN 2049-2618. doi: 10.1186/s40168-020-00899-6.

URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00899-6.

T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, July 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu153. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu153.

Vladimir Sentchilo, Antonia P Mayer, Lionel Guy, Ryo Miyazaki, Susannah Green Tringe, Kerrie Barry, Stephanie Malfatti, Alexander Goessmann, Marc Robinson-Rechavi, and Jan R van der Meer. Community-wide plasmid gene mobilization and selection. *The ISME Journal*, 7(6):1173–1186, June 2013. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2013.13. URL http://www.nature.com/articles/ismej201313.

Ying Sha, John H. Phan, and May D. Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2015:6461–6464, 2015. ISSN 2694-0604 1557-170X 2375-7477. doi: 10.1109/EMBC.2015.7319872.

Pranjul Shah, Joëlle V. Fritz, Enrico Glaab, Mahesh S. Desai, Kacy Greenhalgh, Audrey Frachet, Magdalena Niegowska, Matthew Estes, Christian Jäger, Carole Seguin-Devaux, Frederic Zenhausern, and Paul Wilmes. A microfluidics-based in vitro model of the gastrointestinal human–microbe interface. *Nature Communications*, 7(1):11535, September 2016. ISSN 2041-1723. doi: 10.1038/ncomms11535. URL http://www.nature.com/articles/ncomms11535.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, November 2003. ISSN 1088-9051 (Print). doi: 10.1101/gr.1239303.

Nikki Shariat and Edward G Dudley. CRISPRs: molecular signatures used for pathogen subtyping. *Applied and environmental microbiology*, 80(2):430–439, January 2014. ISSN 1098-5336. doi: 10.1128/AEM.02790-13. URL https://pubmed.ncbi.

nlm.nih.gov/24162568. Edition: 2013/10/25 Publisher: American Society for Microbiology.

Nataliya M. Shchegolkova, George S. Krasnov, Anastasia A. Belova, Alexey A. Dmitriev, Sergey L. Kharitonov, Kseniya M. Klimina, Nataliya V. Melnikova, and Anna V. Kudryavtseva. Microbial Community Structure of Activated Sludge in Treatment Plants with Different Wastewater Compositions. *Frontiers in Microbiology*, 7, February 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00090. URL http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00090/abstract.

Abdul R Sheik, Emilie E L Muller, and Paul Wilmes. A hundred years of activated sludge: time for a rethink. *Frontiers in microbiology*, 5:47–47, March 2014. ISSN 1664-302X. doi: 10.3389/fmicb.2014.00047. URL https://pubmed.ncbi.nlm.nih.gov/24624120. Publisher: Frontiers Media S.A.

Abdul R Sheik, Emilie El Muller, Jean-Nicolas Audinot, Laura A Lebrun, Patrick Grysan, Cedric Guignard, and Paul Wilmes. In situ phenotypic heterogeneity among single cells of the filamentous bacterium Candidatus Microthrix parvicella. *The ISME journal*, 10 (5):1274–1279, May 2016. ISSN 1751-7370 (Electronic). doi: 10.1038/ismej.2015.181.

Sergey A. Shmakov, Vassilii Sitnik, Kira S. Makarova, Yuri I. Wolf, Konstantin V. Severinov, and Eugene V. Koonin. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio*, 8(5):e01397–17, /mbio/8/5/e01397–17.atom, November 2017. ISSN 2150-7511. doi: 10.1128/mBio.01397-17. URL https://mbio.asm.org/content/8/5/e01397-17.

Sergey A. Shmakov, Yuri I. Wolf, Ekaterina Savitskaya, Konstantin V. Severinov, and Eugene V. Koonin. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Communications Biology*, 3(1):321, December 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-1014-1. URL http://www.nature.com/articles/s42003-020-1014-1.

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, July 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0171-1. URL http://www.nature.com/articles/s41564-018-0171-1.

Justin D. Silverman, Heather K. Durand, Rachael J. Bloom, Sayan Mukherjee, and Lawrence A. David. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome*, 6(1):202, November 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0584-3. URL https://doi.org/10.1186/s40168-018-0584-3.

Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020. ISSN 20010370. doi: 10.1016/j.csbj.2020.09.014. URL https://linkinghub.elsevier.com/retrieve/pii/S2001037020303986.

M E Singer, S M Tyler, and W R Finnerty. Growth of Acinetobacter sp. strain HO1-N on n-hexadecanol: physiological and ultrastructural characteristics. *Journal of Bacteriology*, 162(1):162–169, 1985. ISSN 0021-9193. URL https://jb.asm.org/content/162/1/162. Publisher: American Society for Microbiology Journals _eprint: https://jb.asm.org/content/162/1/162.full.pdf.

Connor T. Skennerton, Michael Imelfort, and Gene W. Tyson. Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*, 41(10), 2013. ISSN 03051048. doi: 10.1093/nar/gkt183. ISBN: 0305-1048$\backslash$r1362-4962.

Kim Sneppen, Szabolcs Semsey, Aswin S. N. Seshasayee, and Sandeep Krishna. Restriction modification systems as engines of diversity. *Frontiers in Microbiology*, 6, June 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00528. URL http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00528/abstract.

J. A. Soddell and R.J. Seviour. Microbiology of foaming in activated sludge plants. *Journal of Applied Bacteriology*, 69(2):145–176, August 1990. ISSN 00218847. doi: 10.1111/j.1365-2672.1990.tb01506.x. URL http://doi.wiley.com/10.1111/j.1365-2672.1990.tb01506.x.

Robert R. Sokal. *Biometry : the principles and practice of statistics in biological research*. W.H. Freeman, New York, 3rd ed. edition, 1995. ISBN 0-7167-2411-1. Publication Title: Biometry : the principles and practice of statistics in biological research.

Amit Sonune and Rupali Ghate. Developments in wastewater treatment methods. *Desalination Strategies in South Mediterranean Countries*, 167:55–63, August 2004. ISSN 0011-9164. doi: 10.1016/j.desal.2004.06.113. URL https://www.sciencedirect.com/science/article/pii/S0011916404003558.

Rotem Sorek, Victor Kunin, and Philip Hugenholtz. CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature Reviews Microbiology*, 6:181, March 2008. URL https://doi.org/10.1038/nrmicro1793http://10.0.4.14/nrmicro1793. Publisher: Nature Publishing Group.

James T. Staley and Allan Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestial habitats. *Annual Review of Microbiology*, 39(1):321–346, October 1985. ISSN 0066-4227. doi: 10.1146/annurev.mi.39.100185.001541. URL https://doi.org/10.1146/annurev.mi.39.100185.001541. Publisher: Annual Reviews.

Evan P. Starr, Shengjing Shi, Steven J. Blazewicz, Alexander J. Probst, Donald J. Herman, Mary K. Firestone, and Jillian F. Banfield. Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. *Microbiome*, 6(1):122, July 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0499-z. URL https://doi.org/10.1186/s40168-018-0499-z.

Eric J. Stewart. Growing Unculturable Bacteria. *Journal of Bacteriology*, 194(16):4151, August 2012. doi: 10.1128/JB.00345-12. URL http://jb.asm.org/content/194/16/4151.abstract.

Alessandro Tanca, Marcello Abbondio, Antonio Palomba, Cristina Fraumene, Valeria Manghina, Francesco Cucca, Edoardo Fiorillo, and Sergio Uzzau. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome*, 5(1):79, December 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0293-3. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0293-3.

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):e1005224, December 2016. ISSN 1553-7358 (Electronic). doi: 10.1371/journal.pcbi.1005224.

R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33–6, January 2000. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/10592175http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC102395. Publisher: Oxford University Press.

Ines Thiele, Swagatika Sahoo, Almut Heinken, Johannes Hertel, Laurent Heirendt, Maike K Aurich, and Ronan MT Fleming. Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Molecular Systems Biology*, 16(5), May 2020. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20198982. URL https://onlinelibrary.wiley.com/doi/10.15252/msb.20198982.

Juan Tong, Anping Tang, Hongyan Wang, Xingxin Liu, Zhaohua Huang, Ziyue Wang, Junya Zhang, Yuansong Wei, ·Yanyan Su, and Yifeng Zhang. Microbial community evolution and fate of antibiotic resistance genes along six different full-scale municipal wastewater treatment processes. *Bioresource Technology*, 272:489–500, January 2019. ISSN 09608524. doi: 10.1016/j.biortech.2018.10.079. URL https://linkinghub.elsevier.com/retrieve/pii/S0960852418315116.

Pål Trosvik, Eric Jacques de Muinck, and Nils Christian Stenseth. Biotic interactions and temporal dynamics of the human gastrointestinal microbiota. *The ISME Journal*, 9(3): 533–541, March 2015. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2014.147. URL http://www.nature.com/articles/ismej2014147.

Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, October 2007. ISSN 1476-4687. doi: 10.1038/nature06244. URL https://pubmed.ncbi.nlm.nih.gov/17943116.

Augusto Uc-Mass, Eva Jacinto Loeza, Mireya de la Garza, Gabriel Guarneros, Javier Hernández-Sánchez, and Luis Kameyama. An orthologue of the cor gene is involved in the exclusion of temperate lambdoid phages. Evidence that Cor inactivates FhuA receptor functions. *Virology*, 329(2):425–433, November 2004. ISSN 0042-6822. doi: 10.1016/j.virol.2004.09.005. Place: United States.

Gherman V. Uritskiy, Jocelyne DiRuggiero, and James Taylor. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, 6 (1):158, December 2018. ISSN 2049-2618. doi: 10.1186/s40168-018-0541-1. URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0541-1.

Stineke van Houte, Angus Buckling, and Edze R. Westra. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiology and Molecular Biology Reviews*, 80 (3):745–763, September 2016. ISSN 1092-2172, 1098-5557. doi: 10.1128/MMBR.00011-16. URL https://mmbr.asm.org/content/80/3/745.

Sebastien Varrette, Pascal Bouvry, Hyacinthe Cartiaux, and Fotis Georgatos. Management of an academic HPC cluster : the UL experience. In *International conference on high performance computing & simulation*, pages 959–967, 2014.

Eleni Vasilakou, Daniel Machado, Axel Theorell, Isabel Rocha, Katharina Nöh, Marco Oldiges, and S Aljoscha Wahl. Current state and challenges for dynamic metabolic modeling. *Current Opinion in Microbiology*, 33:97–104, October 2016. ISSN 13695274. doi: 10.1016/j.mib.2016.07.008. URL https://linkinghub.elsevier.com/retrieve/pii/S1369527416300960.

R. M. Voigt, C. B. Forsyth, S. J. Green, P. A. Engen, and A. Keshavarzian. Circadian Rhythm and the Gut Microbiome. *International review of neurobiology*, 131:193–205, 2016. ISSN 2162-5514 0074-7742. doi: 10.1016/bs.irn.2016.07.002. Place: United States.

Aaron M. Walsh, Guerrino Macori, Kieran N. Kilcawley, and Paul D. Cotter. Meta-analysis of cheese microbiomes highlights contributions to multiple aspects of quality. *Nature Food*, 1(8):500–510, August 2020. ISSN 2662-1355. doi: 10.1038/s43016-020-0129-3. URL https://doi.org/10.1038/s43016-020-0129-3.

Linda Wampach, Anna Heintz-Buschart, Angela Hogan, Emilie E.L. Muller, Shaman Narayanasamy, Cedric C. Laczny, Luisa W. Hugerth, Lutz Bindl, Jean Bottu, Anders F. Andersson, Carine de Beaufort, and Paul Wilmes. Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life. *Frontiers in Microbiology*, 8:738, 2017. doi: doi:10.3389/fmicb.2017.00738.

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, and Paul Wilmes. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nature Communications*, 9(1):5091, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07631-x. URL http://www.nature.com/articles/s41467-018-07631-x.

Ping Wang, Zhisheng Yu, Jihong Zhao, and Hongxun Zhang. Seasonal Changes in Bacterial Communities Cause Foaming in a Wastewater Treatment Plant. *Microbial Ecology*, 71(3):660–671, April 2016. ISSN 0095-3628, 1432-184X. doi: 10.1007/s00248-015-0700-x. URL http://link.springer.com/10.1007/s00248-015-0700-x.

Markus G. Weinbauer. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, 28(2):127–181, May 2004. ISSN 1574-6976. doi: 10.1016/j.femsre.2003.08.001. URL https://academic.oup.com/femsre/article-lookup/doi/10.1016/j.femsre.2003.08.001.

Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, Amanda Birmingham, Jacob A Cram, Jed A Fuhrman, Jeroen Raes, Fengzhu Sun, Jizhong Zhou, and Rob Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669–1681, July 2016. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2015.235. URL http://www.nature.com/articles/ismej2015235.

Edze R. Westra, Stineke van Houte, Sylvain Gandon, and Rachel Whitaker. The ecology and evolution of microbial CRISPR-Cas adaptive immune systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1772):20190101, May 2019. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2019.0101. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0101.

J. M. Whipps, K. Lewis, and R.C Cooke. Mycoparasitism and plant disease control. *Fungi in Biological Control Systems, ed N. M. Burge (Manchester University Press)*, pages 161–187, 1988.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Paul Wilmes, Benjamin P. Bowen, Brian C. Thomas, Ryan S. Mueller, Vincent J. Denef, Nathan C. VerBerkmoes, Robert L. Hettich, Trent R. Northen, and Jillian F. Banfield. Metabolome-Proteome Differentiation Coupled to Microbial Divergence. *mBio*, 1(5): e00246–10, October 2010. ISSN 2150-7511. doi: 10.1128/mBio.00246-10. URL https://mbio.asm.org/content/1/5/e00246-10.

S. Withey, E. Cartmell, L.M. Avery, and T. Stephenson. Bacteriophages—potential for application in wastewater treatment processes. *Science of The Total Environment*, 339(1-3):1–18, March 2005. ISSN 00489697. doi: 10.1016/j.scitotenv.2004.09.021. URL https://linkinghub.elsevier.com/retrieve/pii/S0048969704006497.

Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, March 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-3-r46. URL https://doi.org/10.1186/gb-2014-15-3-r46.

Linwei Wu, Daliang Ning, Bing Zhang, Yong Li, Ping Zhang, Xiaoyu Shan, Qiuting Zhang, Mathew Robert Brown, Zhenxin Li, Joy D. Van Nostrand, Fangqiong Ling, Naijia Xiao, Ya Zhang, Julia Vierheilig, George F. Wells, Yunfeng Yang, Ye Deng, Qichao Tu, Aijie Wang, Dany Acevedo, Miriam Agullo-Barcelo, Pedro J. J. Alvarez, Lisa Alvarez-Cohen, Gary L. Andersen, Juliana Calabria de Araujo, Kevin F. Boehnke, Philip Bond, Charles B. Bott, Patricia Bovio, Rebecca K. Brewster, Faizal Bux, Angela Cabezas, Léa Cabrol, Si Chen, Craig S. Criddle, Ye Deng, Claudia Etchebehere, Amanda Ford, Dominic Frigon, Janeth Sanabria, James S. Griffin, April Z. Gu, Moshe Habagil, Lauren Hale, Steven D. Hardeman, Marc Harmon, Harald Horn, Zhiqiang Hu, Shameem Jauffur, David R. Johnson, Jurg Keller, Alexander Keucken, Sheena Kumari, Cintia Dutra Leal, Laura A. Lebrun, Jangho Lee, Minjoo Lee, Zarraz M. P. Lee, Yong Li, Zhenxin Li, Mengyan Li, Xu Li, Fangqiong Ling, Yu Liu, Richard G. Luthy, Leda C. Mendonça-Hagler, Francisca Gleire Rodriguez de Menezes, Arthur J. Meyers, Amin Mohebbi, Per H. Nielsen, Daliang Ning, Adrian Oehmen, Andrew Palmer, Prathap Parameswaran, Joonhong Park, Deborah Patsch, Valeria Reginatto, Francis L. de los

Reyes, Bruce E. Rittmann, Adalberto Noyola, Simona Rossetti, Xiaoyu Shan, Jatinder Sidhu, William T. Sloan, Kylie Smith, Oscarina Viana de Sousa, David A. Stahl, Kyle Stephens, Renmao Tian, James M. Tiedje, Nicholas B. Tooker, Qichao Tu, Joy D. Van Nostrand, Daniel De los Cobos Vasconcelos, Julia Vierheilig, Michael Wagner, Steve Wakelin, Aijie Wang, Bei Wang, Joseph E. Weaver, George F. Wells, Stephanie West, Paul Wilmes, Sung-Geun Woo, Linwei Wu, Jer-Horng Wu, Liyou Wu, Chuanwu Xi, Naijia Xiao, Meiying Xu, Tao Yan, Yunfeng Yang, Min Yang, Michelle Young, Haowei Yue, Bing Zhang, Ping Zhang, Qiuting Zhang, Ya Zhang, Tong Zhang, Qian Zhang, Wen Zhang, Yu Zhang, Hongde Zhou, Jizhong Zhou, Xianghua Wen, Thomas P. Curtis, Qiang He, Zhili He, Mathew Robert Brown, Tong Zhang, Zhili He, Jurg Keller, Per H. Nielsen, Pedro J. J. Alvarez, Craig S. Criddle, Michael Wagner, James M. Tiedje, Qiang He, Thomas P. Curtis, David A. Stahl, Lisa Alvarez-Cohen, Bruce E. Rittmann, Xianghua Wen, Jizhong Zhou, and Global Water Microbiome Consortium. Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology*, 4(7):1183–1195, July 2019. ISSN 2058-5276. doi: 10.1038/s41564-019-0426-5. URL https://doi.org/10.1038/s41564-019-0426-5.

Martin Wu and Jonathan A. Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology*, 9(10):R151, October 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-10-r151. URL https://doi.org/10.1186/gb-2008-9-10-r151.

Martin Wu and Alexandra J. Scott. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, 28(7):1033–1034, April 2012. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts079. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts079. Publisher: Narnia.

Yu-Wei Wu, Yung-Hsu Tang, Susannah G Tringe, Blake A Simmons, and Steven W Singer. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(1):26, August 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-26. URL https://doi.org/10.1186/2049-2618-2-26.

Marc Wältermann and Alexander Steinbüchel. Neutral Lipid Bodies in Prokaryotes: Recent Insights into Structure, Formation, and Relationship to Eukaryotic Lipid Depots.

Journal of Bacteriology, 187(11):3607–3619, June 2005. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.187.11.3607-3619.2005. URL https://JB.asm.org/content/187/11/3607.

Peng Xu. Dynamics of microbial competition, commensalism, and cooperation and its implications for coculture and microbiome engineering. *Biotechnology and Bioengineering*, 118(1):199–209, January 2021. ISSN 0006-3592, 1097-0290. doi: 10.1002/bit.27562. URL https://onlinelibrary.wiley.com/doi/10.1002/bit.27562.

Yongkui Yang, Longfei Wang, Feng Xiang, Lin Zhao, and Zhi Qiao. Activated Sludge Microbial Community and Treatment Performance of Wastewater Treatment Plants in Industrial and Municipal Zones. *International Journal of Environmental Research and Public Health*, 17(2):436, January 2020. ISSN 1660-4601. doi: 10.3390/ijerph17020436. URL https://www.mdpi.com/1660-4601/17/2/436.

Yi Yue, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, and Jian Tu. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics*, 21(1):334, December 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03667-3. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03667-3.

Quan Zhang and Yuzhen Ye. Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics*, 18(1):92, 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1512-4. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1512-4. ISBN: 1367-4811 (Electronic)$\backslash$r1367-4803 (Linking) Publisher: BMC Bioinformatics _eprint: 9605103.

Quan Zhang, Mina Rho, Haixu Tang, Thomas G. Doak, and Yuzhen Ye. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biology*, 14(4):R40, April 2013. ISSN 1474760X. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r40. Publisher: BioMed Central Ltd.

Tong Zhang, Xu-Xiang Zhang, and Lin Ye. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PloS one*, 6 (10):e26041, 2011. ISSN 1932-6203 (Electronic). doi: 10.1371/journal.pone.0026041.

Tong Zhang, Ming-Fei Shao, and Lin Ye. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *The ISME Journal*, 6(6):1137–1147, June 2012. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2011.188. URL http://www.nature.com/articles/ismej2011188.

Yuwei Zhang, Guofang Zhao, Fatma Yislam Hadi Ahmed, Tianfei Yi, Shiyun Hu, Ting Cai, and Qi Liao. In silico Method in CRISPR/Cas System: An Expedite and Powerful Booster. *Frontiers in Oncology*, 10:584404, October 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.584404. URL https://www.frontiersin.org/article/10.3389/fonc.2020.584404/full.

Fengfeng Zhou and Ying Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics (Oxford, England)*, 26(16):2051–2052, August 2010. ISSN 1367-4811 (Electronic). doi: 10.1093/bioinformatics/btq299.

Yifan Zhu, Sanne E. Klompe, Marnix Vlot, John van der Oost, and Raymond H. J. Staals. Shooting the messenger: RNA-targetting CRISPR-Cas systems. *Bioscience reports*, 38 (3), June 2018. ISSN 1573-4935 0144-8463. doi: 10.1042/BSR20170788.

Soumaya Zlitni, Alex Bishara, Eli L. Moss, Ekaterina Tkachenko, Joyce B. Kang, Rebecca N. Culver, Tessa M. Andermann, Ziming Weng, Christina Wood, Christine Handy, Hanlee P. Ji, Serafim Batzoglou, and Ami S. Bhatt. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Medicine*, 12(1):50, December 2020. ISSN 1756-994X. doi: 10.1186/s13073-020-00747-0. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00747-0.

Hui Zou and Trevor Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Nikita Zrelovs, Andris Dislers, and Andris Kazaks. Motley Crew: Overview of the Currently Available Phage Diversity. *Frontiers in Microbiology*, 11:579452, October 2020.

ISSN 1664-302X. doi: 10.3389/fmicb.2020.579452. URL https://www.frontiersin.org/articles/10.3389/fmicb.2020.579452/full.

Appendices

Appendix A

Article manuscripts

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

A.1 Challenges, strategies and perspectives for reference-independent longitudinal multi-omic microbiome studies

Susana Martínez Arbas, Susheel Bhanu Busi, Pedro Queirós, Laura de Nies, Malte Herold, Patrick May, Paul Wilmes, Emilie E. L. Muller, Shaman Narayanasamy

2021

Frontiers in Genetics

DOI: 10.3389/fgene.2021.666244

Contributions of author include:

- Coordination
- Writing and revision of manuscript

frontiers in Genetics

PERSPECTIVE
published: 14 June 2021
doi: 10.3389/fgene.2021.666244

Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies

Susana Martínez Arbas[1], Susheel Bhanu Busi[1], Pedro Queirós[1], Laura de Nies[1], Malte Herold[2], Patrick May[1], Paul Wilmes[1,3], Emilie E. L. Muller[4] and Shaman Narayanasamy[1]*

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [2]Department of Environmental Research and Innovation, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, [3]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [4]Université de Strasbourg, UMR 7156 CNRS, Génétique Moléculaire, Génomique, Microbiologie, Strasbourg, France

OPEN ACCESS

Edited by:
Himel Mallick, Merck, United States

Reviewed by:
Cecilia Noecker, University of California, San Francisco, United States
Siyuan Ma, University of Pennsylvania, United States

***Correspondence:**
Susana Martínez Arbas
susana.martinez@uni.lu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 February 2021*
Accepted: *30 April 2021*
Published: *14 June 2021*

Citation:
Martínez Arbas S, Busi SB, Queirós P, de Nies L, Herold M, May P, Wilmes P, Muller EEL and Narayanasamy S (2021) Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies. Front. Genet. 12:666244. doi: 10.3389/fgene.2021.666244

In recent years, multi-omic studies have enabled resolving community structure and interrogating community function of microbial communities. Simultaneous generation of metagenomic, metatranscriptomic, metaproteomic, and (meta) metabolomic data is more feasible than ever before, thus enabling in-depth assessment of community structure, function, and phenotype, thus resulting in a multitude of multi-omic microbiome datasets and the development of innovative methods to integrate and interrogate those multi-omic datasets. Specifically, the application of reference-independent approaches provides opportunities in identifying novel organisms and functions. At present, most of these large-scale multi-omic datasets stem from spatial sampling (e.g., water/soil microbiomes at several depths, microbiomes in/on different parts of the human anatomy) or case-control studies (e.g., cohorts of human microbiomes). We believe that longitudinal multi-omic microbiome datasets are the logical next step in microbiome studies due to their characteristic advantages in providing a better understanding of community dynamics, including: observation of trends, inference of causality, and ultimately, prediction of community behavior. Furthermore, the acquisition of complementary host-derived omics, environmental measurements, and suitable metadata will further enhance the aforementioned advantages of longitudinal data, which will serve as the basis to resolve drivers of community structure and function to understand the biotic and abiotic factors governing communities and specific populations. Carefully setup future experiments hold great potential to further unveil ecological mechanisms to evolution, microbe-microbe interactions, or microbe-host interactions. In this article, we discuss the challenges, emerging strategies, and best-practices applicable to longitudinal microbiome studies ranging from sampling, biomolecular extraction, systematic multi-omic measurements, reference-independent data integration, modeling, and validation.

Keywords: microbiome, metatranscriptomics, metaproteomics, time-series, metagenomics, metabolomics, *de novo* assembly

INTRODUCTION

Advances in the study of microbial communities have highlighted their important role in natural processes, including those considered as ecosystem services for humankind (Bodelier, 2011). Complex dynamics in microbiomes at the level of composition and structure, as well as function (Heintz-Buschart and Wilmes, 2018) stem from constant adaptation of a given community toward fluctuations of abiotic and biotic factors. However, the fate of these microbial consortia in the face of perturbations is often not understood nor predictable (Muller, 2019). Longitudinal approaches are necessary to understand microbial community dynamics, as they may offer valuable insights into temporal trends and consequences of environmental forcings, when used in tandem with host-derived (Heintz-Buschart et al., 2016; Lloyd-Price et al., 2019; Mars et al., 2020) or environmental (Law et al., 2016; Herold et al., 2020) data. Longitudinal studies can be conducted using diachronic or synchronic approaches (Costa Junior et al., 2013). Herein, we discuss the capacity of longitudinal diachronic approaches as a critical tool toward studying microbial communities. We will further focus on multi-omics longitudinal studies, which leverage the power of the entire high-throughput meta-omic spectrum, namely meta-genomics (MG), -transcriptomics (MT), -proteomics (MP), and -metabolomics (MM), as they are now more feasible and affordable than ever before (Narayanasamy et al., 2015).

Overall, longitudinal multi-omics will enhance our understanding of microbial community dynamics, which could potentially bring about positive outcomes in biomedicine, biotechnology, and for the environment. However, various aspects must be considered when conducting longitudinal multi-omic microbiome studies, ranging from experimental design, bioinformatic processing, modeling, and validation. In this article, we explore challenges, considerations, and potential solutions for such studies, based on recent advances and reports (Law et al., 2016; Lloyd-Price et al., 2019; Herold et al., 2020; Martínez Arbas et al., 2021), which are applicable to both microbe-centric (e.g., soil, water) or host-centric (e.g., human gut) systems. Finally, although this article focuses on specifically longitudinal multi-omic microbiome studies, the content is generally applicable to any large-scale microbiome studies.

MULTI-OMIC CONSIDERATIONS AND EXPERIMENTAL DESIGN FOR LONGITUDINAL STUDIES

Integration of multi-omic microbiome datasets has been routinely performed, with notable instances, including studies on type-1 diabetes (Heintz-Buschart et al., 2016), cancer (Kaysen et al., 2017), healthy human gut (Tanca et al., 2017), Crohn's disease (Erickson et al., 2012), and activated sludge (Muller et al., 2014; Roume et al., 2015; Yu et al., 2019). These studies clearly demonstrate the maturity of the current microbiome multi-omics toolbox. Despite this, and to the best of our knowledge, equivalent multi-omic surveys based on extensive longitudinal microbiome sampling remain rather limited. **Table 1** lists several relevant studies of longitudinal (at least six timepoints) and multi-omic (at least two omic levels, excluding 16S amplicon sequencing) microbiome datasets.

The famous adage "*absence of evidence is not evidence of absence*" (Altman and Bland, 1995) could likely be a prelude to most microbiome studies. Hence, we discuss these studies in the context of reference-independent bioinformatics

TABLE 1 | Longitudinal multi-omic microbiome datasets and studies.

System	Sample type	Duration*	Frequency*	Total of samples	MG	MT	MP	MM	Complementary data	Studies
Human gut microbiome	Stool samples from 132 humans; healthy or with Crohn's disease or ulcerative colitis	1 year	Bi-weekly	2,965	x	x	x	x	Host genomics, transcriptomics bisulfite sequencing, serologic profiles, diet surveys, and fecal calprotectin	Lloyd-Price et al., 2019 Ruiz-Perez et al., 2021
	Stool samples of 77 individuals	6 months	Monthly	474	x			x	Host transcriptome, metabolome, cytokines, methylome, dietary survey, and physiology	Blasche et al., 2021
Activated sludge	Floating sludge islets from a single anoxic tank	1.5 year	Weekly	53	x	x	x	x	Temperature, pH, oxygen concentration, conductivity, inflow, nitrate concentration, and extracellular metabolites	Herold et al., 2020 Martínez Arbas et al., 2021
	Full- and lab-scale activated sludge	2.5 months	Weekly	10	x	x			Temperature, pH, redox potential and dissolved oxygen	Law et al., 2016

Longitudinal multi-omic data must be of least six timepoints and at least two meta-omic readouts excluding 16S amplicon sequencing. Omics data derived from host(s) are considered separate from the microbial meta-omic spectra.
**Approximate values.*

approaches, centered around *de novo* assemblies of sequencing data (MG and MT), subsequently complemented by additional omics (MP and MM, depending on their availability; **Figure 1**). Reference-independent approaches offer asymmetric advantages and opportunities in discovering novel microbial taxa and/or functionalities (Celaj et al., 2014; Narayanasamy et al., 2015; Lapidus and Korobeynikov, 2021), compared to reference-dependent methodologies (Sunagawa et al., 2013; Treangen et al., 2013). Moreover, the integration of multi-omics has been shown to yield superior output compared to single omic studies. For instance, the co-assembly of MG and MT sequencing reads was shown to improve the quality of assembled contigs (Narayanasamy et al., 2016), which in turn improves taxonomic annotation, gene calling/annotation, binning, metabolic pathway (re) construction (Muller et al., 2018; Zhou et al., 2020; Zimmermann et al., 2021), and quantification of features, e.g., taxa/genes (Narayanasamy et al., 2016). Similarly, MP spectra searches are more effective when performed against gene databases derived from MG assemblies of the same sample/environment, compared to generic databases, thus improving the recruitment of measured peptides (Tanca et al., 2016; Heyer et al., 2017; Timmins-Schiffman et al., 2017). Moreover, such a reference-independent approach may be necessary for microbial communities that are not well characterized and lack extensive unified genome or gene catalogues, such as those available for the human gut microbiome (Li et al., 2014; Almeida et al., 2021). However, most microbial communities are heterogeneous, which further complicates downstream multi-omic data processing, integration, curation, transformation, and modeling (Jiang et al., 2019). Therefore, the adherence toward standards

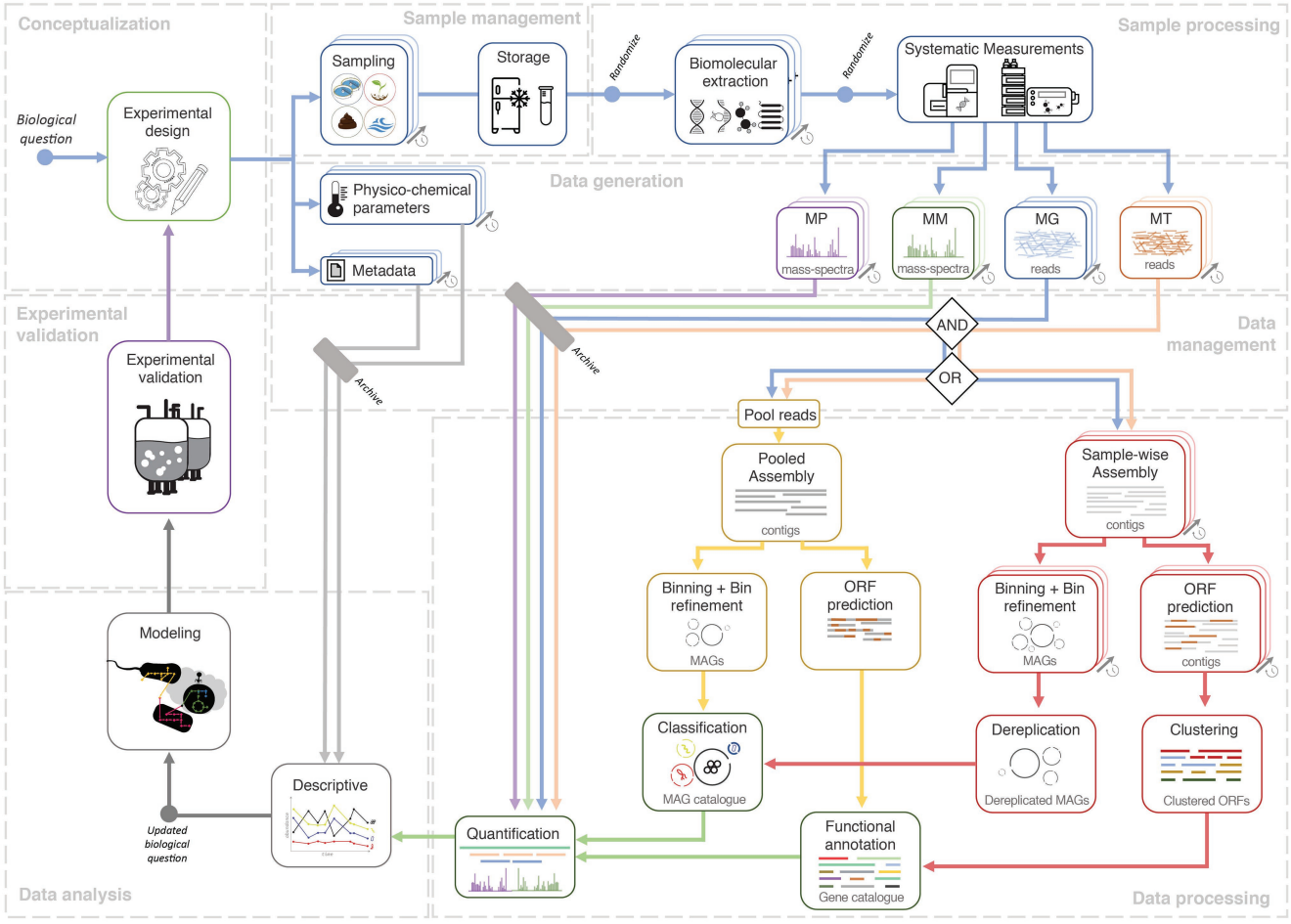


FIGURE 1 | Systems ecology workflow for longitudinal multi-omic microbiome studies. A study conceptualized *via* an experimental design phase and an initial biological question which is then followed by sample collection, sample management, and systematic high-throughput measurements. The next-generation sequencing (NGS) data could either undergo aggregated processing (yellow track) involving a pooled *de novo* assembly of NGS reads from all longitudinal samples, to eventually yield a metagenome assembled genome (MAG) and/or gene catalogue *via* binning and gene calling, respectively. In the dereplication approach (red track), data from each sample are first processed in a sample-wise manner, namely the steps of *de novo* assembly, binning, and gene calling. The resulting MAGs and predicted ORFs are then merged through a process called dereplication which generates the catalogue. The availability of a catalogue allows quantification whereby the output could be used for descriptive analyses which could potentially lead to updated or entirely novel biological questions. Quantified values, combined with descriptive analyses, could then be used within dynamic or metabolic models (gray track). Validation of models could lead to further *in situ* longitudinal experimental designs. Finally, all data (raw input, output, metadata) and code (not depicted) should be archived under a data and code management strategy. Free icons were used from https://www.flaticon.com (creators: Freepik, Gregor Cresnar, Freepik, and Smashicons).

and best-practices, spanning from sampling to data analyses is important to the outcome of a project. Accordingly, **Figure 1** illustrates the potential lifecycle of a longitudinal multi-omic microbiome study.

Longitudinal multi-omic studies require systematic and thorough study designs that consider sampling parameters (Gerber, 2014; Cao et al., 2017; Liang et al., 2020), metadata, and complementary measurements, such as physico-chemical parameters or questionnaires (Kumar et al., 2014), all of which affect downstream analyses. Sampling parameters, such as duration and frequency, are dictated by the inherent properties of a given microbial system. For instance, the sampling duration when studying gut microbiome development of neonates could span from birth until a "mature" gut microbiome composition is achieved (Stewart et al., 2018), which may vary from subject to subject. Naturally-occurring microbial systems that are exposed to the environment may exhibit annual cyclical behavior based on seasonality and, therefore, could be sampled for at least one complete season-to-season cycle (Johnston et al., 2019). Sampling frequency may be determined by the dynamics and/or generational-timescale of a given system. For instance, the human gut microbiome is known to exhibit daily fluctuations, and therefore could be sampled on a daily basis within a given temporal study (David et al., 2014), while activated sludge systems are known to exhibit (approximately) weekly doubling periods and thus could be sampled on a weekly basis (Herold et al., 2020; Martínez Arbas et al., 2021). Based on the recommendations of Sefer et al. (2016), if biological replicates are either not feasible (i.e., $n = 1$) or limited (i.e., low n) (Herold et al., 2020), one should ideally opt for higher frequency (dense) longitudinal sampling, and less dense sampling if biological replicates were available (i.e., high n), e.g., a cohort of patients (Lloyd-Price et al., 2019). Equidistant sampling is required by many downstream mathematical frameworks, such as cross-correlation or local similarity analysis (Faust et al., 2015), and thus should be strived for, as much as possible. However, the datasets listed in **Table 1**, albeit extensive and resource intensive, are not perfectly equidistant, further highlighting the practical challenges for longitudinal sampling *in situ*, including, but not limited to, accessibility, consistent biomass availability, and cost.

SAMPLE, DATA AND CODE MANAGEMENT

It is crucial to limit potential biases linked to longitudinal data, e.g., in extended time-series; samples are stored for long periods, while multiple personnel may be involved in sample collection, handling, storage, and documentation. Hence, clear guidelines and standardization must be established, as they are key factors that potentially affect downstream processes and overall outcome (Blekhman et al., 2016; Schoenenberger et al., 2016).

Biomolecular extraction from a single sample is ideal over multiple extractions from subsamples (Roume et al., 2013a). Advantageously, commercial kits for concomitant extraction of multiple biomolecules are available, including reports proposing adapted methods for extracting various biomolecules, such as DNA, total RNA, small RNA, protein, and metabolites (Peña-Llopis and Brugarolas, 2013; Roume et al., 2013b; Thorn et al., 2019). The availability of sufficient biomass (Eisenhofer et al., 2019) lysis-, homogenization-(Machiels et al., 2000; Santiago et al., 2014; Fiedorová et al., 2019) and preservation- (Borén, 2015; Hickl et al., 2019) methods are key factors that determine effectiveness to comprehensively recover all intracellular and/or extracellular biomolecules. Next, biomolecular extraction should be automated, whenever possible. While evaluations have shown that it may not necessarily provide better quality results compared to a human operator (Phillips et al., 2012), the output is more consistent (Fidler et al., 2020). In the same vein, omic readouts should also be generated on a single platform (s) as unique batches to ensure consistent output quality.

Batch effects are often overlooked in omic studies (de Goffau et al., 2021), but can be minimized during stages of sample processing by including randomization, sample tracking, and extensive documentation (Leek et al., 2010). Sample randomization implemented within batches of biomolecular extraction and high-throughput measurements could help discriminate batch effects and temporal variation, i.e., different sets of randomly selected samples from different timepoints could be treated together at each different step (Oh et al., 2019). Additionally, batch effects could be mitigated using downstream analytical (Wang and Cao, 2019) and computational methods (Gibbons et al., 2018; McLaren et al., 2019).

A potential effective experimental measure for minimizing and elucidating batch effects is the inclusion of mock/control samples during both the extraction and high-throughput measurements (Bokulich et al., 2016; Hornung et al., 2019; ATCC Mock Microbial Communities, 2020). Samples with low biomass, e.g., from neonates, glacier-streams, or acid-mine drainage, should include extraction blanks as negative controls, which are extremely valuable to discriminate contaminants arising from kits and reagents (Salter et al., 2014; Heintz-Buschart et al., 2018; Wampach et al., 2018; Weyrich et al., 2019). Furthermore, spike-ins could be helpful for downstream quantification (Zinter et al., 2019). Importantly, replicates can be used within downstream statistical frameworks (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021) to understand both within- and between-sample heterogeneity, thereby minimizing mischaracterisation of contaminants or findings driven by batch effects (de Goffau et al., 2021).

Longitudinal and multi-omic studies yield large datasets, where data processing and analyses are typically time and resource intensive. These rich datasets may be reused to study multiple aspects of a given microbial system (**Table 1**). Therefore, equal emphasis should be placed on designing bioinformatic workflows and code/data management strategies to improve reproducibility and transparency. For example, peer-review journals have begun mandating "data availability" sections and links to code repositories in adherence to project/coding best practices and standards (Sandve et al., 2013; Bokulich et al., 2020), further improving posterior data integration and analysis in the short-term, while improving scaling-up

and knowledge transfer in the long run (Shahin et al., 2017; Wilson et al., 2017). In addition, format-free archival repositories, such as Zenodo could be used for non-standard data types,[1] for instance simulated raw data, physico-chemical measurements, intermediate data, large tables, and archived Github repositories. Despite this, reports indicate that 26% of bioinformatics tools are no longer available (Mangul et al., 2019), while gaps in available raw data (Jurburg et al., 2020) and metadata (Schriml et al., 2020) still exist.

CONSTRUCTION OF LONGITUDINAL GENE AND GENOME REFERENCE CATALOGUES

Microbiomes may be studied from a gene-centric perspective (Roume et al., 2015), which requires read or contig-level taxonomic classification (Segata et al., 2012; Wood and Salzberg, 2014), ORF prediction (Hyatt et al., 2010; Rho et al., 2010), and gene annotation (Seemann, 2014; Buchfink et al., 2015; Franzosa et al., 2018; Queirós et al., 2020). Metagenome assembled genomes (MAGs) provide genomic context and can be obtained through binning (Chen et al., 2020; Yue et al., 2020) followed by taxonomic classification (Bremges et al., 2020; Chaumeil et al., 2020) and functional annotation. In that regard, several tools exist that improve the binning process by automating the selection of highest-quality MAGs (bins) and/or performing MAG refinement (Broeksema et al., 2017; Sieber et al., 2018; Uritskiy et al., 2018). These tools enable ensemble binning approaches, balancing out the strengths and weaknesses of different binning methods (Chen et al., 2020; Yue et al., 2020).

Features (i.e., taxa or genes) appear in varying quantities, in different timepoints of longitudinal meta-omic studies. It is challenging to link and track features from one timepoint to another without any given point of reference. Therefore, the construction of what we term as "representative longitudinal catalogues" (hereafter referred to as catalogues) of MAGs/genes, provides a non-redundant representative base to link features from the different longitudinal samples (Herold et al., 2020; Martínez Arbas et al., 2021). The outcome of any downstream analysis is highly reliant on the quality of the MAGs and genes within a catalogue, which further depends on the quality of large-scale bioinformatic processing (e.g., *de novo* assembly and binning). **Figure 1** illustrates two methods of constructing such catalogues, which are through aggregated processing of data from all samples or through de-replicating the output from individually processed sample data (i.e., sample-wise processing). A third alternative to these methods could be the representation of non-redundant genes in pangenomes from MAGs annotated at the species-level (Tettelin et al., 2005; Delmont and Eren, 2018), collected across all timepoints. This allows for identifying any varying patterns especially in the context of environmental factors and phylogenetic constraints influencing gene acquisition and/or genome-streamlining (Tettelin et al., 2005). Given that others have highlighted the catalogue building methodologies (Qin et al., 2010; Nayfach et al., 2020; Almeida et al., 2021); here, we elaborate methods discussed above in the context of both gene- and MAG-centric strategies.

The general advantage of the aggregated processing approach is simplicity, whereby a single run is required for all the large-scale bioinformatic processing steps (**Figure 1**). Moreover, pooled assemblies have been shown to be effective (Magasin and Gerloff, 2015), especially in the advent of highly efficient *de novo* assemblers (Li et al., 2016) and digital normalization (Brown et al., 2012). However, pooling reads from a large number of samples increases the complexity of the *de novo* assembly process, especially for complex communities. It also requires substantial computational resources, while potentially resulting in lower quality contigs, MAGs, and genes (Chen et al., 2020).

The dereplication method (**Figure 1**) is applied after independent sample-wise large-scale bioinformatic processing (Evans and Denef, 2020). Predicted ORFs could be de-replicated through clustering (Li and Godzik, 2006; Edgar, 2010; Mirdita et al., 2019), producing a gene catalogue (Li et al., 2014). On the contrary, the dereplication of MAGs is more complex, requiring several steps: binning from sample-wise *de novo* assemblies to generate MAGs, curation of high-quality MAGs (Parks et al., 2015), and dereplication of MAGs (Olm et al., 2017; Wampach et al., 2018) to select the most representative MAGs of the longitudinal data (Uritskiy et al., 2018; Chen et al., 2020). In general, dereplication methods are particularly advantageous for longitudinal microbiome studies with many deeply sequenced samples (Herold et al., 2020; Martínez Arbas et al., 2021).

Although not systematically evaluated, one caveat worth considering when constructing a catalogue based on *de novo* assemblies, binning, and dereplication is the potential loss of resolution in population-level diversity (Kashtan et al., 2014; Evans and Denef, 2020; Quince et al., 2020), which may include single nucleotide variants, copy number variants, strains, and auxiliary gene content (Evans and Denef, 2020) potentially impacting important downstream steps, such as integration of metaproteomic data (Tanca et al., 2016) or time-resolved strain tracking (Brito and Alm, 2016; Zlitni et al., 2020). To the best of our knowledge, the extent of the impact has yet to be systematically investigated. In our opinion, several strategies can be applied to overcome this issue, including the usage of a comparative genomics methodology, i.e., pangenomes (Delmont and Eren, 2018), even opt for (re) assemblies of read subsets associated to particular taxa or MAGs of interest (Albertsen et al., 2013), or the application of strain-level analysis tools (Anyansi et al., 2020).

Overall, choosing the specific methods for constructing a longitudinal catalogue depends on various factors, including the biological question, complexity of the community (van der Walt et al., 2017), number of samples, and sequencing depth. To the best of our knowledge, a comparison between an aggregated processing approach and a dereplication approach has yet to be conducted. Such a comparison would further help to inform researchers on selecting the best strategy for longitudinal analyses.

[1]https://zenodo.org

QUANTIFICATION AND NORMALIZATION

Longitudinal catalogues provide compositional information of community taxa and potential functions. However, the relative quantification of community members and functionalities is key in harnessing the power of longitudinal microbiome data, as it allows the observation of community taxa/functional dynamics and could be used in downstream modeling. In that regard, quantifying MG and MT sequencing data is a standard process of aligning reads (Li and Durbin, 2009) to relevant catalogues, and then quantifying features of interest (e.g., population/gene relative genomic abundance, gene expression) based on those alignments, providing information on community structure, functional potential, and gene expression. Complementally, MP data provide functional insights, whereby several methods are available for the quantification of such data (Delogu et al., 2020; Pible et al., 2020), while identification and quantification of metabolites through MM data (Kapoore and Vaidyanathan, 2016; Mallick et al., 2019; Røst et al., 2020) provide insights on the community phenotype (s). However, *in situ* measurements of substrate uptake through labeling-based approaches (Starr et al., 2018) are challenging. Therefore, specific metabolites of interest could be indirectly linked to members of a microbial community by proportionally assigning the relative contribution of a MAG to a given (re) constructed metabolic pathway based on genomic abundance or gene/protein expression (Noecker et al., 2016; Blasche et al., 2021).

Normalization of quantified values is required to enable community structure and function comparisons between timepoint samples. The selection of normalization methods is important as it affects downstream analytical steps. There are several methods to normalize longitudinal MG and MT data, from the generation of compositional data to log-ratios and differential rankings (Chen et al., 2018; Pereira et al., 2018; Morton et al., 2019). Additionally, one should also inspect the data for potential confounding batch effects and take it into consideration when performing normalization (Gibbons et al., 2018; McLaren et al., 2019; Coenen et al., 2020). In summary, effective relative quantification and normalization will serve as a strong basis for downstream modeling approaches, and the development of robust methods for absolute quantification will be decisive in the future.

ANALYSIS OF COMMUNITY CHARACTERISTICS AND DYNAMICS

Generally, microbiome omic data are complex, as it is (i) compositional, e.g., provided as relative abundances, which require specific considerations when selecting statistical analyses (Gloor et al., 2017), (ii) highly sparse, such that the interpretation of zero-values generated from sampling, biological, or technical processes heavily affects data-derived conclusions (Silverman et al., 2020), and (iii) high dimensional, which increases modeling difficulty due to the influence of feature selection that heavily affect potential predictions (Bolón-Canedo et al., 2016). Furthermore, multi-omic studies may contain gaps within the omic spectrum, such that certain samples may not be represented within a certain omic layer (Lloyd-Price et al., 2019). Despite introducing complexity, the complementary use of different omics could improve analysis outcomes and add predictive power to models (Muller et al., 2013; Fondi and Liò, 2015). Longitudinal data introduce another layer of complexity, i.e., time dependencies, such that one timepoint is dependent on the previous timepoints, rendering conventional statistical analyses unsuitable as they assume samples to be independent (Coenen et al., 2020). This is further compounded by the fact that samples from longitudinal *in situ* studies are often low in number and non-equidistant (Park et al., 2020). Imputation may be used to supplement missing values (i.e., omic measurements or timepoints; Jiang et al., 2020).

Initial exploration of the microbiome dynamics can be assessed through ordination analyses, where high dimensional population structure data are visualized in a two-dimensional space to observe the trajectory of the samples and the behavior of the system, i.e., metastability, cycles, and alternative states (Gonze et al., 2018). Then, community member relationships may be inferred using, e.g., correlation methods (Faust et al., 2012; Friedman and Alm, 2012; Weiss et al., 2016). Unfortunately, correlations may be insufficient to assess complex community interactions, whereby the application of modeling approaches would be necessary to resolve those relationships (Fisher and Mehta, 2014; Trosvik et al., 2015; Ridenhour et al., 2017). Modeling could serve as a means of integrating several layers of omic data (Lloyd-Price et al., 2019; Ruiz-Perez et al., 2021) further elucidating microbial interplay beyond species abundances and functional potential.

Extensive literature of statistical and mathematical frameworks for multi-omic and/or longitudinal microbiome data is currently available. For instance, Noor et al. (2019) review the integration of multi-omics data from data-driven and knowledge-based perspectives. Coenen et al. (2020) discuss approaches to characterize temporal dynamics and to identify periodicity of populations and putative interactions between them, while Faust et al. (2018) propose a classification scheme for better model selection. Bodein et al. (2019) provide a multivariate framework to integrate longitudinal and multi-omics data, while Park et al. (2020) discuss the development of models and software tools for time-series metagenome and metabolome data. Overall, the application of these methodologies should be tailored toward specific hypotheses and studies, for which data exploration is essential to select modeling approaches that fit the type, quality, and quantity of the data.

More recently, the emergence of studies which track microbiome dynamics of cohorts over time, i.e., multiple individuals/sites (Carmody et al., 2019; Lloyd-Price et al., 2019; Mars et al., 2020), necessitates the ability to discriminate variation stemming from the same individual/environment compared to those from different individuals/environments. In such cases, multi-level statistical modeling (also known as mixed-effects/hierarchical models) is able to account for repeated sampling or nested variation across a sample population (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021). Most notably Lloyd-Price et al. (2019) extensively applied such

methods to associate multi-omic microbiome signatures with host-derived molecular profiles in a cohort of 132 individuals. Other instances include multi-omic longitudinal studies that combine murine and human datasets to unveil the adaptation of gut microbiomes to raw and cooked food (Carmody et al., 2019) and the identification of therapeutic targets for irritable bowel syndrome (Mars et al., 2020). Finally, there are newer methodologies that apply similar/related statistical frameworks to modeling multi-omic data (Mallick et al., 2021).

The validation of the models remains one of the most challenging issues. Mathematical models combined with culture of synthetic microbial communities are commonly utilized to study mechanisms behind host-microbiome interactions (Moejes et al., 2017). It is also possible to validate interactions between microbes by, e.g., applying environmental perturbations in controlled conditions (Law et al., 2016; Herold et al., 2020). These explorations may result in a further understanding of the role of biotic and abiotic factors in shaping microbiomes, in relation to community phenotypes found in nature, biotechnological processes (Law et al., 2016; Herold et al., 2020), or host-associated microbiomes (Moejes et al., 2017; Garza et al., 2018).

CONCLUSION

Longitudinal microbiome studies combined with integrated multi-omic measurements provide unprecedented opportunities to study microbial community dynamics, both structurally and functionally. In tandem with evolving high-throughput technologies, e.g., long-read sequencing (Moss et al., 2020; Wickramarachchi et al., 2020), these studies will become important tools in the exploration and potential exploitation of microbial consortia. We described strategies to mitigate the various challenges associated with such studies, encompassing study design, best practices, practical considerations, and bioinformatics processing and modeling. While longitudinal multi-omics datasets are currently scarce (**Table 1**), we are confident that it will increasingly become more common, similar to how we are increasingly transitioning from single omics to multi-omic (Noor et al., 2019). Longitudinal microbiome multi-omics will serve as an important tool for further improving analytical methods, which will in turn lead to relevant biomedical, biotechnological, and environmental outcomes.

AUTHOR CONTRIBUTIONS

SMA and SN outlined the manuscript and coordinated the writing process. LdN, SN, and SMA prepared the figure. All authors contributed to the writing, reviewing, and editing of the manuscript. All authors approved the submitted version.

FUNDING

The Luxembourg National Research Fund (FNR) supported SMA, PQ, LdN, PM, and EELM through the PRIDE doctoral training unit grants (PRIDE15/10907093) and (PRIDE/18/11823097), the CORE Junior grant (C15/SR/10404839), and the CORE grant (CORE/17/SM/11689322). SBB was supported by the Sinergia grant (CRSII5_180241) through the Swiss National Science Foundation. PW was supported by the European Research Council (ERC-CoG 863664).

ACKNOWLEDGMENTS

We would like to thank Oskar Hickl for his input on metaproteomic analysis.

REFERENCES

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. 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

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., et al. (2021). A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114. doi: 10.1038/s41587-020-0603-3

Altman, D. G., and Bland, J. M. (1995). Statistics notes: absence of evidence is not evidence of absence. *BMJ* 311:485. doi: 10.1136/bmj.311.7003.485

Anderson, M. J. (2017). "Permutational multivariate analysis of variance (PERMANOVA)," in *Wiley Stats Ref: Statistics Reference Online*. eds. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels (Chichester, UK: John Wiley & Sons, Ltd.), 1–15.

Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., and Abeel, T. (2020). Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Front. Microbiol.* 11:1925. doi: 10.3389/fmicb.2020.01925

ATCC Mock Microbial Communities (2020). Available at: https://www.atcc.org/en/Products/Microbiome_Standards.aspx (Accessed November 30, 2020).

Blasche, S., Kim, Y., Mars, R. A. T., Machado, D., Maansson, M., Kafkia, E., et al. (2021). Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nat. Microbiol.* 6, 196–208. doi: 10.1038/s41564-020-00816-5

Blekhman, R., Tang, K., Archie, E. A., Barreiro, L. B., Johnson, Z. P., Wilson, M. E., et al. (2016). Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. *Sci. Rep.* 6:31519. doi: 10.1038/srep31519

Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* 10:963. doi: 10.3389/fgene.2019.00963

Bodelier, P. L. E. (2011). Toward understanding, managing, and protecting microbial ecosystems. *Front. Microbiol.* 2:80. doi: 10.3389/fmicb.2011.00080

Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., et al. (2016). Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 1:e00062-16. doi: 10.1128/mSystems.00062-16

Bokulich, N. A., Ziemski, M., Robeson, M. S., and Kaehler, B. D. (2020). Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Comput. Struct. Biotechnol. J.* 18, 4048–4062. doi: 10.1016/j.csbj.2020.11.049

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Prog. Artif. Intell.* 5, 65–75. doi: 10.1007/s13748-015-0080-y

Borén, M. (2015). "Sample preservation Through heat stabilization of proteins: principles and examples," in *Proteomic Profiling Methods in Molecular Biology*. ed. A. Posch (New York, NY: Springer), 21–32.

Bremges, A., Fritz, A., and McHardy, A. C. (2020). CAMITAX: taxon labels for microbial genomes. *Giga Science* 9:giz154. doi: 10.1093/gigascience/giz154

Brito, I. L., and Alm, E. J. (2016). Tracking strains in the microbiome: insights from metagenomics and models. *Front. Microbiol.* 7:712. doi: 10.3389/fmicb.2016.00712

Broeksema, B., Calusinska, M., McGee, F., Winter, K., Bongiovanni, F., Goux, X., et al. (2017). ICoVeR – an interactive visualization tool for verification and refinement of metagenomic bins. *BMC Bioinformatics* 18:233. doi: 10.1186/s12859-017-1653-5

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv [Preprint].

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Cao, H.-T., Gibson, T. E., Bashan, A., and Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: pitfalls and lessons. *BioEssays* 39:1600188. doi: 10.1002/bies.201600188

Carmody, R. N., Bisanz, J. E., Bowen, B. P., Maurice, C. F., Lyalina, S., Louie, K. B., et al. (2019). Cooking shapes the structure and function of the gut microbiome. *Nat. Microbiol.* 4, 2052–2063. doi: 10.1038/s41564-019-0569-4

Celaj, A., Markle, J., Danska, J., and Parkinson, J. (2014). Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome* 2:39. doi: 10.1186/2049-2618-2-39

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res.* 30, 315–333. doi: 10.1101/gr.258640.119

Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600

Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A primer for microbiome time-series analysis. *Front. Genet.* 11:310. doi: 10.3389/fgene.2020.00310

Costa Junior, C., Corbeels, M., Bernoux, M., Píccolo, M. C., Siqueira Neto, M., Feigl, B. J., et al. (2013). Assessing soil carbon storage rates under no-tillage: comparing the synchronic and diachronic approaches. *Soil Tillage Res.* 134, 207–212. doi: 10.1016/j.still.2013.08.010

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89

de Goffau, M. C., Charnock-Jones, D. S., Smith, G. C. S., and Parkhill, J. (2021). Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. *Microbiome* 9:6. doi: 10.1186/s40168-020-00949-z

Delmont, T. O., and Eren, A. M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* 6:e4320. doi: 10.7717/peerj.4320

Delogu, F., Kunath, B. J., Evans, P. N., Arntzen, M. Ø., Hvidsten, T. R., and Pope, P. B. (2020). Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat. Commun.* 11:4708. doi: 10.1038/s41467-020-18543-0

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* 2, 105–117. doi: 10.1016/j.tim.2018.11.003

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 7:e49138. doi: 10.1371/journal.pone.0049138

Evans, J. T., and Denef, V. J. (2020). To dereplicate or not to dereplicate? *mSphere* 5:e00971-19. doi: 10.1128/mSphere.00971-19

Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606

Fidler, G., Tolnai, E., Stagel, A., Remenyik, J., Stundl, L., Gal, F., et al. (2020). Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. *Sci. Rep.* 10:3419. doi: 10.1038/s41598-020-60304-y

Fiedorová, K., Radvanský, M., Němcová, E., Grombiříková, H., Bosák, J., Černochová, M., et al. (2019). The impact of DNA extraction methods on stool bacterial and fungal microbiota community recovery. *Front. Microbiol.* 10:821. doi: 10.3389/fmicb.2019.00821

Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 9:e102451. doi: 10.1371/journal.pone.0102451

Fondi, M., and Liò, P. (2015). Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiol. Res.* 171, 52–64. doi: 10.1016/j.micres.2015.01.003

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968. doi: 10.1038/s41592-018-0176-y

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Garza, D. R., van Verk, M. C., Huynen, M. A., and Dutilh, B. E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* 3, 456–460. doi: 10.1038/s41564-018-0124-8

Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037

Gibbons, S. M., Duvallet, C., and Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.* 14:e1006102. doi: 10.1371/journal.pcbi.1006102

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Gonze, D., Coyte, K. Z., Lahti, L., and Faust, K. (2018). Microbial communities as dynamical systems. *Curr. Opin. Microbiol.* 44, 41–49. doi: 10.1016/j.mib.2018.07.004

Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* 2:16180. doi: 10.1038/nmicrobiol.2016.227

Heintz-Buschart, A., and Wilmes, P. (2018). Human gut microbiome: function matters. *Trends Microbiol.* 26, 563–574. doi: 10.1016/j.tim.2017.11.002

Heintz-Buschart, A., Yusuf, D., Kaysen, A., Etheridge, A., Fritz, J. V., May, P., et al. (2018). Small RNA profiling of low biomass samples: identification and removal of contaminants. *BMC Biol.* 16:52. doi: 10.1186/s12915-018-0522-7

Herold, M., Arbas, S. M., Narayanasamy, S., Sheik, A. R., Kleine-Borgmann, L. A. K., Lebrun, L. A., et al. (2020). Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat. Commun.* 11:5281. doi: 10.1038/s41467-020-19006-2

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* 261, 24–36. doi: 10.1016/j.jbiotec.2017.06.1201

Hickl, O., Heintz-Buschart, A., Trautwein-Schult, A., Hercog, R., Bork, P., Wilmes, P., et al. (2019). Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut microbiome. *Microorganisms* 7:367. doi: 10.3390/microorganisms7090367

Hornung, B. V. H., Zwittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95:fiz045. doi: 10.1093/femsec/fiz045

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., et al. (2019). Microbiome multi-omics network analysis: statistical considerations,

limitations, and opportunities. *Front. Genet.* 10:995. doi: 10.3389/fgene.2019.00995

Jiang, R., Li, W. V., and Li, J. J. (2020). mbImpute: an accurate and robust imputation method for microbiome data. *Genomics* [Preprint]. doi: 10.1101/2020.03.07.982314

Johnston, J., LaPara, T., and Behrens, S. (2019). Composition and dynamics of the activated sludge microbiome during seasonal nitrification failure. *Sci. Rep.* 9:4565. doi: 10.1038/s41598-019-40872-4

Jurburg, S. D., Konzack, M., Eisenhauer, N., and Heintz-Buschart, A. (2020). The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Commun. Biol.* 3:474. doi: 10.1038/s42003-020-01204-9

Kapoore, R. V., and Vaidyanathan, S. (2016). Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374:20150363. doi: 10.1098/rsta.2015.0363

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344, 416–420. doi: 10.1126/science.1248575

Kaysen, A., Heintz-Buschart, A., Muller, E. E. L., Narayanasamy, S., Wampach, L., Laczny, C. C., et al. (2017). Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic hematopoietic stem cell transplantation. *Transl. Res.* 186, 79–94. doi: 10.1016/j.trsl.2017.06.008

Kumar, R., Eipers, P., Little, R. B., Crowley, M., Crossman, D. K., Lefkowitz, E. J., et al. (2014). Getting started with microbiome analysis: sample acquisition to bioinformatics. *Curr. Protoc. Hum. Genet.* 82, 18.8.1–18.8.29. doi: 10.1002/0471142905.hg1808s82

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lapidus, A. L., and Korobeynikov, A. I. (2021). Metagenomic data assembly – the way of decoding unknown microorganisms. *Front. Microbiol.* 12:613791. doi: 10.3389/fmicb.2021.613791

Law, Y., Kirkegaard, R. H., Cokro, A. A., Liu, X., Arumugam, K., Xie, C., et al. (2016). Integrative microbial community analysis reveals full-scale enhanced biological phosphorus removal under tropical conditions. *Sci. Rep.* 6:25719. doi: 10.1038/srep25719

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

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

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11. doi: 10.1016/j.ymeth.2016.02.020

Liang, Y., Dong, T., Chen, M., He, L., Wang, T., Liu, X., et al. (2020). Systematic analysis of impact of sampling regions and storage methods on Fecal gut microbiome and metabolome profiles. *mSphere* 5:e00763-19. doi: 10.1128/mSphere.00763-19

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi: 10.1038/s41586-019-1237-9

Machiels, B. M., Ruers, T., Lindhout, M., Hardy, K., Hlavaty, T., Bang, D. D., et al. (2000). New protocol for DNA extraction of stool. *Bio Techniques* 28, 286–290. doi: 10.2144/00282st05

Magasin, J. D., and Gerloff, D. L. (2015). Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics* 31, 311–317. doi: 10.1093/bioinformatics/btu546

Mallick, H., Franzosa, E. A., Mclver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10:3136. doi: 10.1038/s41467-019-10927-1

Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. *Microbiology* [Preprint]. doi: 10.1099/mic.0.001031

Mangul, S., Martin, L. S., Eskin, E., and Blekhman, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20:47. doi: 10.1186/s13059-019-1649-8

Mars, R. A. T., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., et al. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* 182, 1460–1473. doi: 10.1016/j.cell.2020.08.007

Martínez Arbas, S. M., Narayanasamy, S., Herold, M., Lebrun, L. A., Hoopmann, M. R., Li, S., et al. (2021). Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nat. Microbiol.* 6, 123–135. doi: 10.1038/s41564-020-00794-8

McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *elife* 8:e46923. doi: 10.7554/eLife.46923

Mirdita, M., Steinegger, M., and Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35, 2856–2858. doi: 10.1093/bioinformatics/bty1057

Moejes, F., Succurro, A., Popa, O., Maguire, J., and Ebenhöh, O. (2017). Dynamics of the bacterial community associated with *Phaeodactylum tricornutum* cultures. *Processes* 5:77. doi: 10.3390/pr5040077

Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* 10:2719. doi: 10.1038/s41467-019-10656-5

Moss, E. L., Maghini, D. G., and Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* 38, 701–707. doi: 10.1038/s41587-020-0422-6

Muller, E. E. L. (2019). Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* 4:e00080-19. doi: 10.1128/mSystems.00080-19

Muller, E. E. L., Faust, K., Widder, S., Herold, M., Arbas, S. M., and Wilmes, P. (2018). Using metabolic networks to resolve ecological properties of microbiomes. *Curr. Opin. Syst. Biol.* 8, 73–80. doi: 10.1016/j.coisb.2017.12.004

Muller, E. E. L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends Microbiol.* 21, 325–333. doi: 10.1016/j.tim.2013.04.009

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

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Heintz-Buschart, A., Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17:260. doi: 10.1186/s13059-016-1116-8

Narayanasamy, S., Muller, E. E. L., Sheik, A. R., and Wilmes, P. (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microb. Biotechnol.* 8, 363–368. doi: 10.1111/1751-7915.12255

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2020). A genomic catalog of earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi: 10.1038/s41587-020-0718-6

Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., et al. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 1:e00013-15. doi: 10.1128/mSystems.00013-15

Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Gene Regul.* 15, 39–47. doi: 10.1016/j.coisb.2019.03.007

Oh, S., Li, C., Baldwin, R. L., Song, S., Liu, F., and Li, R. W. (2019). Temporal dynamics in meta longitudinal RNA-Seq data. *Sci. Rep.* 9:763. doi: 10.1038/s41598-018-37397-7

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

Park, S.-Y., Ufondu, A., Lee, K., and Jayaraman, A. (2020). Emerging computational tools and models for studying gut microbiota composition and function. *Tissue Cell Pathw. Eng.* 66, 301–311. doi: 10.1016/j.copbio.2020.10.005

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Check M: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Peña-Llopis, S., and Brugarolas, J. (2013). Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications. *Nat. Protoc.* 8, 2240–2255. doi: 10.1038/nprot.2013.141

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274. doi: 10.1186/s12864-018-4637-6

Phillips, K., McCallum, N., and Welch, L. (2012). A comparison of methods for forensic DNA extraction: Chelex-100® and the QIAGEN DNA Investigator Kit (manual and automated). *Forensic Sci. Int. Genet.* 6, 282–285. doi: 10.1016/j.fsigen.2011.04.018

Pible, O., Allain, F., Jouffret, V., Culotta, K., Miotello, G., and Armengaud, J. (2020). Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome* 8:30. doi: 10.1186/s40168-020-00797-x

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

Queirós, P., Delogu, F., Hickl, O., May, P., and Wilmes, P. (2020). Mantis: flexible and consensus-driven genome annotation. *Bioinformatics* [Preprint]. doi: 10.1101/2020.11.02.360933

Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O. S., Summers, J. K., et al. (2020). Metagenomics strain resolution on assembly graphs. *Bioinformatics* [Preprint]. doi: 10.1101/2020.09.06.284828

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., et al. (2017). Modeling time-series data from microbial communities. *ISME J.* 11, 2526–2537. doi: 10.1038/ismej.2017.107

Røst, L. M., Brekke Thorfinnsdottir, L., Kumar, K., Fuchino, K., Eide Langørgen, I., Bartosova, Z., et al. (2020). Absolute quantification of the central carbon metabolome in eight commonly applied prokaryotic and eukaryotic model systems. *Metabolites* 10:74. doi: 10.3390/metabo10020074

Roume, H., Heintz-Buschart, A., Muller, E. E. L., May, P., Satagopam, V. P., Laczny, C. C., et al. (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *Npj Biofilms Microbiomes* 1:15007. doi: 10.1038/npjbiofilms.2015.7

Roume, H., Heintz-Buschart, A., Muller, E. E. L., and Wilmes, P. (2013b). "Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample," in *Methods in Enzymology*. ed. E. F. DeLong (Cambridge, Massachusetts, United States: Elsevier), 219–236.

Roume, H., Muller, E. E., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013a). A biomolecular isolation framework for eco-systems biology. *ISME J.* 7, 110–121. doi: 10.1038/ismej.2012.72

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., et al. (2021). Dynamic Bayesian networks for integrating multi-omics time series microbiome data. *mSystems* 6:e01105-20. doi: 10.1128/mSystems.01105-20

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* 9:e1003285. doi: 10.1371/journal.pcbi.1003285

Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., et al. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* 14:112. doi: 10.1186/1471-2180-14-112

Schoenenberger, A. W., Muggli, F., Parati, G., Gallino, A., Ehret, G., Suter, P. M., et al. (2016). Protocol of the Swiss Longitudinal Cohort Study (SWICOS) in rural Switzerland. *BMJ Open* 6:e013280. doi: 10.1136/bmjopen-2016-013280

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., et al. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* 7:188. doi: 10.1038/s41597-020-0524-5

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Sefer, E., Kleyman, M., and Bar-Joseph, Z. (2016). Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments. *Cell Syst.* 3, 35–42. doi: 10.1016/j.cels.2016.06.007

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Shahin, M., Ali Babar, M., and Zhu, L. (2017). Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* 5, 3909–3943. doi: 10.1109/ACCESS.2017.2685629

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843. doi: 10.1038/s41564-018-0171-1

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014

Sokal, R. R. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research. 3rd Edn.* New York: W.H. Freeman.

Starr, E. P., Shi, S., Blazewicz, S. J., Probst, A. J., Herman, D. J., Firestone, M. K., et al. (2018). Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. *Microbiome* 6:122. doi: 10.1186/s40168-018-0499-z

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588. doi: 10.1038/s41586-018-0617-x

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693

Tanca, A., Abbondio, M., Palomba, A., Fraumene, C., Manghina, V., Cucca, F., et al. (2017). Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* 5:79. doi: 10.1186/s40168-017-0293-3

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., et al. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* 4:51. doi: 10.1186/s40168-016-0196-8

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Thorn, C. E., Bergesch, C., Joyce, A., Sambrano, G., McDonnell, K., Brennan, F., et al. (2019). A robust, cost-effective method for DNA, RNA and protein co-extraction from soil, other complex microbiomes and pure cultures. *Mol. Ecol. Resour.* 19, 439–455. doi: 10.1111/1755-0998.12979

Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., et al. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* 11, 309–314. doi: 10.1038/ismej.2016.132

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

Trosvik, P., de Muinck, E. J., and Stenseth, N. C. (2015). Biotic interactions and temporal dynamics of the human gastrointestinal microbiota. *ISME J.* 9, 533–541. doi: 10.1038/ismej.2014.147

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1

van der Walt, A. J., van Goethem, M. W., Ramond, J.-B., Makhalanyane, T. P., Reva, O., and Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18:521. doi: 10.1186/s12864-017-3918-9

Wampach, L., Heintz-Buschart, A., Fritz, J. V., Ramiro-Garcia, J., Habier, J., Herold, M., et al. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* 9:5091. doi: 10.1038/s41467-018-07631-x

Wang, Y., and Cao, K.-A. L. (2019). Managing batch effects in microbiome data. *Brief. Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., et al. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi: 10.1111/1755-0998.13011

Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., and Lin, Y. (2020). MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics* 36, i3–i11. doi: 10.1093/bioinformatics/btaa441

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Comput. Biol.* 13:e1005510. doi: 10.1371/journal.pcbi.1005510

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Yu, K., Yi, S., Li, B., Guo, F., Peng, X., Wang, Z., et al. (2019). An integrated meta-omics approach reveals substrates involved in synergistic interactions in a bisphenol A (BPA)-degrading microbial community. *Microbiome* 7:16. doi: 10.1186/s40168-019-0634-5

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., et al. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 21:334. doi: 10.1186/s12859-020-03667-3

Zhou, Z., Tran, P. Q., Breiser, A. M., Liu, Y., Kieft, K., Cowley, E. S., et al. (2020). METABOLIC: high-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.27.357558

Zimmermann, J., Kaleta, C., and Waschina, S. (2021). gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* 22:81. doi: 10.1186/s13059-021-02295-1

Zinter, M. S., Mayday, M. Y., Ryckman, K. K., Jelliffe-Pawlowski, L. L., and DeRisi, J. L. (2019). Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* 7, 62. doi: 10.1186/s40168-019-0678-6

Zlitni, S., Bishara, A., Moss, E. L., Tkachenko, E., Kang, J. B., Culver, R. N., et al. (2020). Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Med.* 12:50. doi: 10.1186/s13073-020-00747-0

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.

A.2 Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics

Susana Martínez Arbas[†], Shaman Narayanasamy[†], Malte Herold, Laura A. Lebrun, Michael R. Hoopmann, Sujun Li, Tony J. Lam, Benoît J. Kunath, Nathan D. Hicks, Cindy M. Liu, Lance B. Price, Cedric C. Laczny , John D. Gillece, James M. Schupp, Paul S. Keim, Robert L. Moritz, Karoline Faust, Haixu Tang, Yuzhen Ye, Alexander Skupin, Patrick May, Emilie E. L. Muller and Paul Wilmes

2021

Nature Microbiology **6**:123–135

DOI: https://doi.org/10.1038/s41564-020-00794-8

Contributions of author include:

- Coordination
- Data analysis and visualization
- Writing and revision of manuscript

[†]Co-first author

nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-020-00794-8

Check for updates

OPEN

Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics

Susana Martínez Arbas[1,13], Shaman Narayanasamy[1,10,13], Malte Herold[1], Laura A. Lebrun[1], Michael R. Hoopmann[2], Sujun Li[3], Tony J. Lam[3], Benoît J. Kunath[1], Nathan D. Hicks[4,11], Cindy M. Liu[4,12], Lance B. Price[4,12], Cedric C. Laczny[1], John D. Gillece[4], James M. Schupp[4], Paul S. Keim[4,5], Robert L. Moritz[2], Karoline Faust[6], Haixu Tang[3], Yuzhen Ye[3], Alexander Skupin[1,7], Patrick May[1], Emilie E. L. Muller[1,8] and Paul Wilmes[1,9] ✉

Viruses and plasmids (invasive mobile genetic elements (iMGEs)) have important roles in shaping microbial communities, but their dynamic interactions with CRISPR-based immunity remain unresolved. We analysed generation-resolved iMGE–host dynamics spanning one and a half years in a microbial consortium from a biological wastewater treatment plant using integrated meta-omics. We identified 31 bacterial metagenome-assembled genomes encoding complete CRISPR–Cas systems and their corresponding iMGEs. CRISPR-targeted plasmids outnumbered their bacteriophage counterparts by at least fivefold, highlighting the importance of CRISPR-mediated defence against plasmids. Linear modelling of our time-series data revealed that the variation in plasmid abundance over time explained more of the observed community dynamics than phages. Community-scale CRISPR-based plasmid–host and phage–host interaction networks revealed an increase in CRISPR-mediated interactions coinciding with a decrease in the dominant '*Candidatus* Microthrix parvicella' population. Protospacers were enriched in sequences targeting genes involved in the transmission of iMGEs. Understanding the factors shaping the fitness of specific populations is necessary to devise control strategies for undesirable species and to predict or explain community-wide phenotypes.

Microbial community dynamics are driven by both abiotic (environmental) and biotic (biological) factors. The latter include mobile genetic elements that move within and/or between genomes[1,2] and are believed to play an important role in microbial community dynamics[3,4]. More specifically, invasive mobile genetic elements (iMGEs), such as bacteriophages and plasmids, may transfer detrimental or beneficial genetic material to or between hosts[1,2]. Bacteriophages (henceforth referred to as phages) are viruses that specifically infect and replicate within bacteria. Phages are considered to be the most abundant and diverse biological entities with single- or double-stranded DNA or RNA genetic material[5], and potentially play a role in shaping microbial community structure[6,7]. In contrast, plasmids are generally circular, double-stranded DNA molecules independent of the bacterial chromosome that encode their own origin of replication and are usually found in higher copy numbers[8]. Plasmids represent key components in horizontal gene transfer and are major contributors to the spread of antimicrobial resistance[9].

Prokaryotic hosts have several defence mechanisms[10] against iMGE invasion. One notable example is the CRISPR–Cas system, which is an adaptive immune process with mechanisms for acquired immunological memory[1,2]. It consists of genomic regions known as clustered regularly inter-spaced short palindromic repeats (CRISPRs) and a class of proteins referred to as CRISPR-associated (Cas) proteins. CRISPR–Cas systems recognize iMGEs and cleave short subsequences from these iMGEs, called protospacers, which are integrated as spacers within the CRISPR loci of prokaryotic genomes[11–13]. The spacer sequences serve as a genetic memory bank of infection history used to recognize and interfere with future invasions. By exploiting the sequence-based links between spacers and protospacers, specific host populations can be linked to specific iMGEs and to their corresponding invasion events[1,2].

The present work focuses on a model microbial community in an activated sludge biological wastewater treatment plant (BWWTP), which arguably represents the most widely used biotechnological process on our planet and is an essential component of future integrated energy and matter management strategies[14]. Foaming sludge, which occurs as floating islets on the surface of anoxic treatment tanks and is partially composed of populations of lipid-accumulating microorganisms, is particularly suitable for energy recovery via biodiesel production[15]. These communities also represent good models of microbial ecology because they exhibit medial species richness

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [2]Institute for Systems Biology, Seattle, WA, USA. [3]School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA. [4]TGen North, Flagstaff, AZ, USA. [5]The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. [6]Laboratory of Molecular Bacteriology, KU Leuven, Leuven, Belgium. [7]Department of Neuroscience, University of California, La Jolla, CA, USA. [8]Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA-CNRS, Université de Strasbourg, Strasbourg, France. [9]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [10]Present address: Megeno S.A., Esch-sur-Alzette, Luxembourg. [11]Present address: Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [12]Present address: Department of Environmental and Occupational Health, Miken Institute School of Public Health, George Washington University, Washington, DC, USA. [13]These authors contributed equally: Susana Martínez Arbas and Shaman Narayanasamy. ✉e-mail: paul.wilmes@uni.lu

while at the same time being highly dynamic. Foaming sludge represents a convenient and virtually unlimited source of spatially and temporally resolved samples with complementary detailed physicochemical information[16]. Here, we present a time-resolved, integrated meta-omics analysis aimed at elucidating CRISPR-mediated interactions and dynamics between iMGEs and their hosts. The resolved community and population interactions and dynamics highlight that CRISPR-based immunity within the studied community predominantly targets plasmid sequences.

Results

Time-resolved meta-omics of foaming sludge islets. A total of 53 samples of foaming sludge islets from the surface of an anoxic tank were collected from a BWWTP over a period of 578 days. The mean sampling frequency of 8 days (s.d. = 16 days) is equivalent to the doubling time of the dominant population '*Candidatus* Microthrix parvicella' *(M. parvicella)*[17,18], thereby facilitating the study of population dynamics on a generational timescale. Concomitant DNA, RNA and protein fractions were obtained from each sample[19], which is critical for coherent downstream systematic measurements and multi-omic data integration[20]. These biomolecular fractions were subjected to deep, high-throughput measurements resulting in time-resolved metagenomics (MG), metatranscriptomics (MT) and metaproteomics (MP) data. A total of 1.5×10^9 MG reads and 1.7×10^9 MT reads underwent sample-specific, large-scale bioinformatics processing, followed by MG and MT de novo co-assembly[21], yielding a total of 2.1×10^7 contigs (Supplementary Table 1). Additionally, we estimated ~50% average coverage of community members resolved for the individual time points (Supplementary Note 1 and Supplementary Fig. 1). MP datasets yielded a total of 7.6×10^6 mass spectra, whereby a total of 9.6×10^7 redundant peptides were identified per sample using the 3.1×10^7 protein sequences predicted from the co-assembled contigs as the search database (Supplementary Table 2).

Contigs from the co-assembled MG and MT data from each sample were binned, producing a total of 26,524 metagenome-assembled genomes (MAGs) across all samples (Supplementary Table 1), of which 1,364 MAGs were selected for dereplication together with a collection of 85 isolate genomes (Supplementary Note 2). The dereplication process yielded pools of MAGs for which we defined representative MAGs (rMAGs)[22]. These rMAGs underwent taxonomic classification, quality filtering and manual curation to yield a total of 92 rMAGs, which were retained for downstream analyses (Supplementary Table 3). In this work, rMAGs are assumed to represent pools of MAGs resulting from dereplication and are equivalent to populations. Therefore, our population-level analyses are, by default, on the rMAG level unless otherwise specified.

CRISPR–Cas information over the entire meta-omics dataset. We resolved the CRISPR–Cas systems within rMAGs by extracting their respective *cas* genes and classifying the CRISPR types[23]. This resulted in a final set of 31 (37%) rMAGs that encoded classifiable and complete CRISPR–Cas systems (that is, *cas* genes allowing CRISPR–Cas system classification) and CRISPR loci containing the required information for linking hosts to iMGEs[24]. The most common CRISPR–Cas system within the community was type I, which was found in 21 rMAGs and across several taxonomic families, followed by type III, which was assigned to 9 rMAGs, while type II and V systems were identified in 3 rMAGs and 1 rMAG, respectively. Combinations of different CRISPR types within a single rMAG were also detected. Accordingly, we found that types I and III were present together in five rMAGs, thereby representing the most commonly detected combination[25] (Fig. 1 and Supplementary Table 4).

We used an ensemble of computational methods to extract CRISPR information on the read- and contig- level, which resulted in an extensive set of detected CRISPR repeats and spacers (both collectively referred to as CRISPR elements) per sample. Overall, we retrieved 89,856 repeats and 525,579 spacers over the entire time series. However, they are redundant because the same repeats or spacers may appear at multiple time points (Extended Data Fig. 1). Therefore, we removed redundancy by clustering CRISPR elements, which resulted in 8,469 and 162,985 non-redundant repeats and spacers, respectively. Spacers were more highly represented on the MG level, whereas repeats were more highly represented on the MT level (Supplementary Note 3 and Supplementary Fig. 2). A total of 778 (~9%) non-redundant repeats and 20,002 (~12%) non-redundant spacers could be directly assigned to at least one rMAG, in turn representing 196,159 (~37%) and 29,685 (~33%) redundant spacers and repeats, respectively. To retain the maximum amount of information for downstream analyses, the entire collection of spacers and repeats from the entire pool of MAGs were linked to their corresponding rMAGs (Supplementary Table 4). Although this may result in high numbers of unfiltered spacers associated with certain rMAGs, for example, rMAG-117, which represents 41 MAGs and is associated with 6,574 spacers, this approach allows comprehensive tracking of CRISPR and targeted iMGE dynamics.

Protospacers in the entire meta-omics dataset. Protospacers may represent either the origin of the spacers or targets for iMGE inhibition/splicing. Spacer information from the CRISPR loci can be used to detect iMGEs through complementary matching to their targeted protospacers[26,27]. Single matches of spacers to targeted iMGEs are considered sufficient for conferring immunity against such iMGEs[28,29]. Thus, spacers were searched against all contigs. Those containing at least one protospacer match, that is, protospacer-containing contigs (hereafter referred to as PSCCs), and lacking repeats to avoid self-matching were considered as putative iMGEs. Accordingly, we detected 750,375 protospacers within 224,651 PSCCs (Extended Data Fig. 1), which highlights the large number of PSCCs that encode multiple protospacers (56%). It is noteworthy that the filtering of PSCCs with repeats (109,504 redundant PSCCs) resulted in the exclusion of potential iMGEs encoding CRISPR loci.

After removing redundancy with the iMGEs (see next section and Supplementary Note 4), a total of 209,199 protospacers were retained within 49,306 non-redundant PSCCs (Supplementary Table 5). Here, there were instances of single spacers targeting multiple protospacers from either different or the same PSCCs. On average, one spacer targeted 21.85 protospacers (median = 7, s.d. = 51.27), while PSCCs tended to contain more than one protospacer (that is, mean = 3.29, median = 2, s.d. = 4.60).

Plasmids and phages in the entire meta-omics dataset. On the basis of the contigs from all time points, we predicted phage and plasmid sequences. The total number of annotated iMGEs represented 6.97% of all contigs, for which 2.22% contained at least one protospacer (that is, PSCCs). Interestingly, we found that sequences annotated as plasmids outnumbered phages by ~16-fold (Supplementary Note 4 and Supplementary Table 6). At this stage, there was a lack of predicted prophage sequences, which is likely due to limitations of the available phage prediction methods. All the predicted iMGEs were clustered to yield non-redundant representative iMGEs that were traceable over time, which maintained similar proportions to the previously described redundant set; that is, ~16 times more plasmid (707,093) than phages (42,039). Among these, we found 12,232 (1.7%) plasmids and 227 (0.5%) phages with similarities to sequences within the National Center for Biotechnology Information (NCBI) database, which demonstrates the lack of representation of these elements within public databases. A similar trend in proportions was reflected in the iMGEs targeted by spacers. Plasmids (12,412) were targeted five times more frequently than phages (2,351). Since we were interested in iMGEs that are

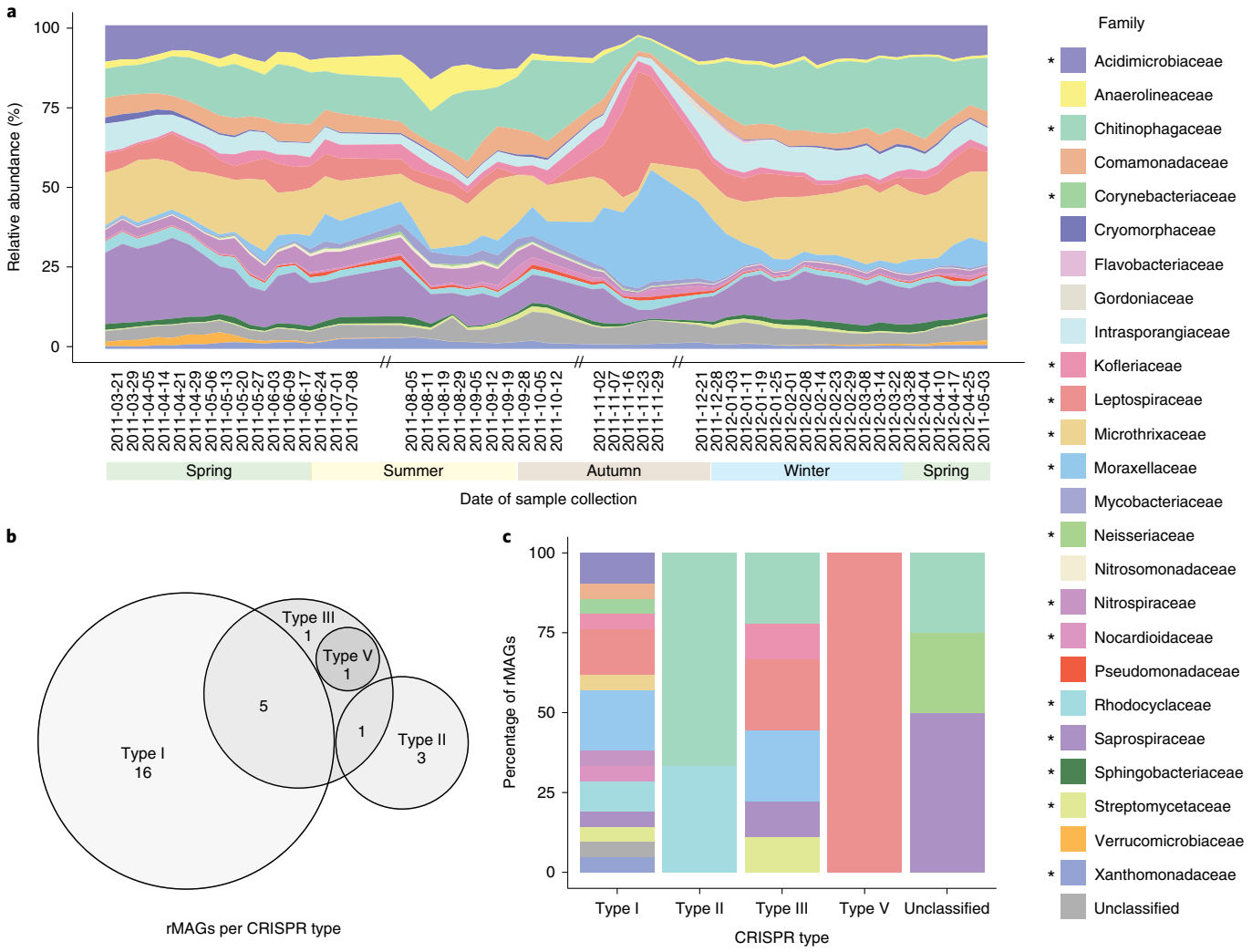


Fig. 1 | Community dynamics and CRISPR–Cas type distribution. a, The relative abundance of rMAGs over time. The labels on the *x* axis indicate the sampling dates and the double slashes (//) on the time axis represent the absence of samples in the sampled system (applicable to all the other figures). **b**, Venn diagram of CRISPR–Cas system types based on the numbers of rMAGs that encode them. Overlaps indicate single rMAGs carrying more than one CRISPR–Cas system. **c**, The distribution of taxonomic affiliations at the family rank per CRISPR–Cas system type. For **a** and **c**, the legend colours marked with asterisks represent families containing CRISPR–Cas systems.

interacting with hosts via CRISPR, we focused on the non-redundant iMGEs that were also PSCCs (henceforth, we collectively refer to these as iMGEs) for downstream analyses. Additionally, the MG and MT co-assembled contigs allowed the detection of iMGEs that were exclusively present on the MT level, for example, RNA phages[30]. Accordingly, a total of 2,890 MT-only contigs assigned as iMGEs were retrieved, from which 2,102 and 387 were classified as plasmid and phage, respectively.

BWWTPs are thought to represent hotspots for the spread of antimicrobial-resistance genes (ARGs)[3,31]. Therefore, we inspected plasmid and phage functions targeted by CRISPR systems[32,33] and screened those iMGEs for potential ARGs[34] (Supplementary Note 5, Supplementary Table 7 and Extended Data Fig. 2). We found 1,570 (0.22%) plasmids and 106 (0.25%) phages encoding 38 different ARGs, including tetracycline-resistance genes, which are known to be persistent in BWWTPs[31,35]. Additionally, we found ten plasmid PSCCs. Among these, three encoded ARGs that were being targeted by spacers, specifically aminoglycoside nucleotidyltransferase (ANT3), streptomycin phosphotransferase (APH3″) and class D beta-lactamases (ClassD) (Supplementary Tables 8 and 9). Apart from these specific cases, iMGEs encoding ARGs were not PSCCs; therefore, they are likely not targeted by CRISPRs.

Community dynamics. The relative abundance of rMAGs and representative iMGEs were used to infer community dynamics over time (Fig. 1, Extended Data Fig. 3 and Supplementary Fig. 3). We grouped rMAGs at the family level due to the large fraction of unclassified taxa. Families such as Microthrixaceae, Moraxellaceae, Leptospiraceae and Acidimicrobiaceae, which are present within sludge communities[15,36], were prominent members. To further investigate the effects of iMGEs on the community dynamics, we linked iMGEs to their putative host families based on their assignments via binning. This resulted in a total of 79 family-level groups of bacteria, plasmids and phages.

The Microthrixaceae family showed a relative abundance average of 15.5% (median = 15.9%, s.d. = 5.2) with minor fluctuations throughout the time series, except between 2011-11-16 and 2012-01-03, when there was a significant decrease. Moraxellaceae (mean = 6.4%, median = 3.6%, s.d. = 7.5) and Leptospiraceae (mean = 6.9%, median = 5.9%, s.d. = 6.4) showed

relatively low abundance over time, but increased with the decline in Microthrixaceae (Fig. 1), thereby representing the shift in the community structure.

To further investigate the community dynamics, we defined three overlapping shorter-term intervals according to before, during and after the aforementioned community shift (Fig. 2 and Supplementary Note 6). Subsequently, correlation between the family-level groups, hierarchical clustering and linear modelling using the Microthrixaceae family as the response variable were performed for the entire time series and for shorter-term intervals.

The correlation analysis showed 62 pairs of family-level groups that consistently exhibited significant correlations (Supplementary Fig. 4), whereby ten families correlated ($r \leq -0.7$ or $r \geq 0.7$, $P \leq 0.001$) with their own plasmids and phages in the entire time series as well as the shorter-term intervals, for example, Microthrixaceae, Moraxellaceae and Leptospiraceae (Supplementary Table 10). Hierarchical clustering of correlation values from the entire time series yielded a total of six clusters, whereby most bacteria, plasmids and phages assigned to the same families clustered together, which demonstrates that there is predictable variation of these family-level groups. Further inspection of the dominant families showed Microthrixaceae clustering separately from Leptospiraceae and Moraxellaceae. The latter two clustered together and exhibited significant negative correlation with Microthrixaceae ($r = -0.63$, $P = 8.3 \times 10^{-7}$, and $r = -0.52$, $P = 9.9 \times 10^{-5}$, respectively), which further supports their observed acyclical behaviour relative to Microthrixaceae (Extended Data Fig. 4 and Supplementary Fig. 5).

In addition, a selection of the best linear models showed an enrichment of Microthrixaceae plasmids, Acidimicrobiaceae phages and Saprospiraceae plasmids and, in agreement with the enrichment analysis, the best model (adjusted $R^2 = 0.9983$) showed iMGEs from Microthrixaceae, Saprospiraceae and Moraxellaceae families exhibiting significant contributions (Extended Data Fig. 5). Thus, the longitudinal abundance data for Microthrixaceae exhibited good agreement with the models (Fig. 2). Overall, the linear modelling analysis showed the appearance of Microthrixaceae plasmids as the only common significant predictor in all the models (entire time series and shorter-term intervals). This group was then removed from those models to assess its relative importance, and this resulted in a significant reduction of predictive power (Extended Data Fig. 6, Supplementary Tables 11 and 12, Supplementary Note 7 and Supplementary Fig. 6). Consequently, its plasmids had a stronger effect on the prediction of Microthrixaceae abundance compared to its phages, which indicates a higher relative importance of plasmids in governing Microthrixaceae dynamics.

CRISPR–Cas mediated iMGE–host interactions. To describe CRISPR-mediated interactions between iMGEs and their hosts, we retained 4,985 spacers that were encoded by at least one rMAG (host), co-occurred with its assigned rMAG in at least one time point and targeted at least one iMGE at any given time point. We subsequently searched for iMGEs and corresponding spacers newly appearing during the time series (that is, spacer integration events), and observed that 2,377 spacers were detected either after or at the same time point as their corresponding targeted iMGEs. The mean spacer integration time (that is, the lag time between the detection of an iMGE and its corresponding spacer) was 9.5 weeks (median = 8, s.d. = 8.5). Spacers that disappeared after the detection of their linked iMGEs were considered to be lost. We observed 1,616 spacers that were lost, with 7 weeks as the average time for such deletions (median = 5.5, s.d. = 7.5). Interestingly, the average time for spacer integration and deletion was lower for phages compared to plasmids (Supplementary Table 13). Furthermore, there was a shift from spacer gain to loss on 2011-11-29, suggesting that the majority of integration events occurred during the summer to autumn transition, while the majority of deletion events occurred in late autumn, which corresponds to the shift in community structure occurring in autumn to winter (Supplementary Fig. 7).

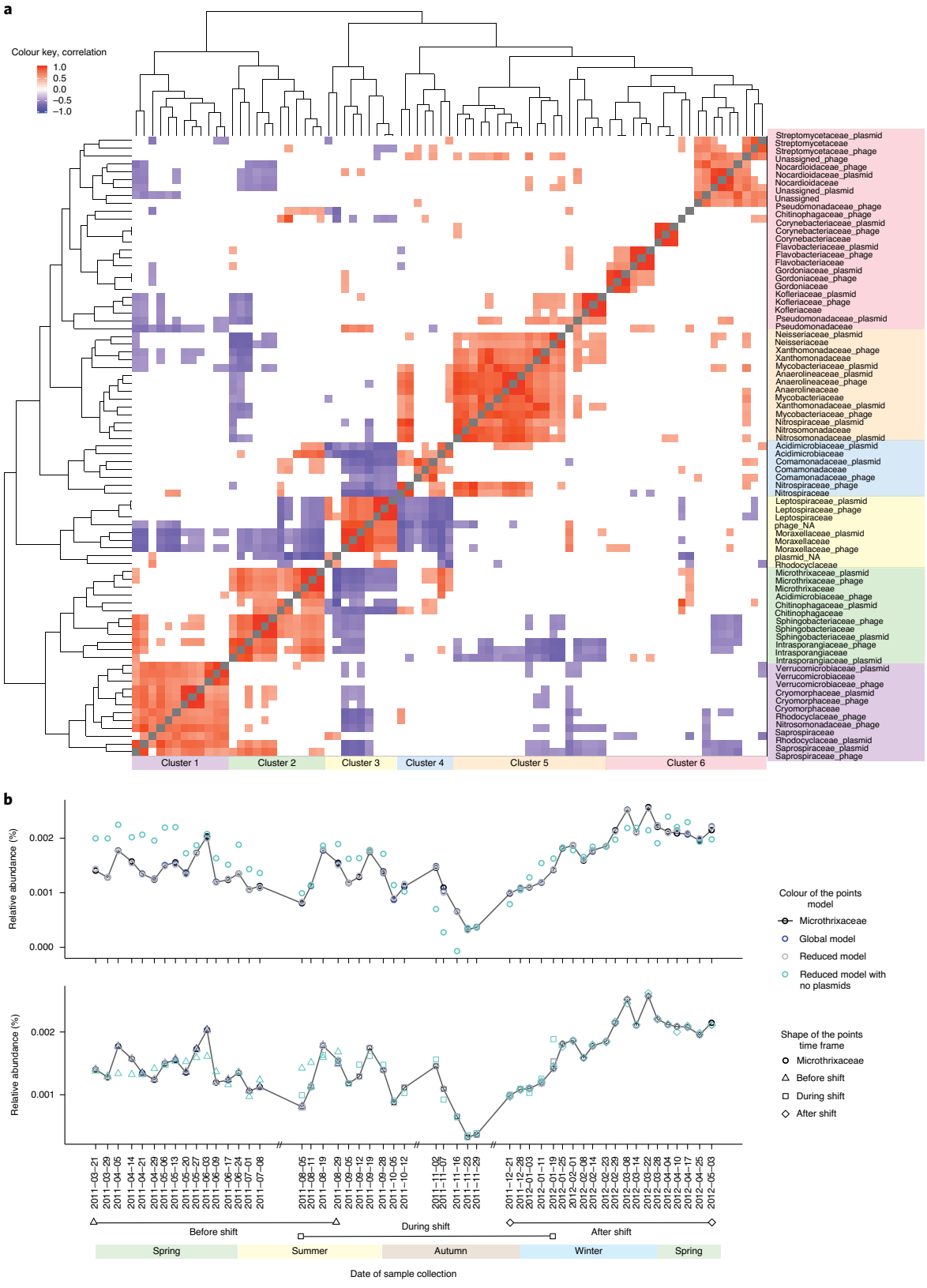
We then separated the CRISPR-mediated interactions into a plasmid–host network comprising 18 hosts and 1,881 plasmids, with 2,274 interactions (Fig. 3), and a phage–host network comprising 16 hosts and 472 phages, with 490 interactions (Extended Data Fig. 7). We also defined an occurring interaction within a given time point if a host and its interacting iMGE were detected in either MG or MT data, which resulted in time-resolved network topology variations (Supplementary Table 14 and Supplementary Note 8). We included orphan iMGEs and hosts for which their associated counterparts were not detected within the same time point to visualize the dynamics (Supplementary Videos 1 and 2).

The time-resolved plasmid–host interaction networks had an average modularity of $Q = 0.71$ (median = 0.73, s.d. = 0.07), with two main modules of interactions: a group containing a core set of rMAGs classified as *Leptospira biflexa* and a group containing rMAGs from different species, that is, *Marinobacter hydrocarbonoclasticus*, *Acinetobacter* sp. ADP21, *Chitinophaga pinensis* and *Haliscomenobacter hydrossis*. *M. parvicella* was represented by rMAG-165. In contrast, the phage–host interaction networks had an average modularity of $Q = 0.69$ (median = 0.69, s.d. = 0.07) and smaller interacting groups. However, the overall dynamics of both networks were similar, with the number of interactions increasing during November 2011, which co-occurred with the drop in *M. parvicella* (Microthrixaceae) and the increase in other populations, such as *L. biflexa* or *H. hydrossis*. Based on these networks, we performed a one mode projection to resolve direct interactions between rMAGs with common iMGEs. For this, we observed a higher range of interactions between rMAGs from the plasmid–host network, which suggests that there is a wide spread of plasmids across different families in contrast to the more restricted infection range of phages (Supplementary Fig. 8 and Supplementary Table 15).

Population-level iMGE–host dynamics. To further understand the iMGE–host dynamics in relation to the maintenance of microbial populations of interest, we focused on the dominant population within the community, *M. parvicella*[15,37–39], which constitutes ~30% of the community at specific dates (Fig. 1). More specifically, it showed distinct characteristics in the community and network dynamics, such that time points with decreased *M. parvicella*

Fig. 2 | Microbial community dynamics. a, The rMAGs were grouped together at the family level. Plasmids and phages were distinctly grouped on the basis of their family-level association (that is, binned together with a rMAG of a given family). The bacterial, plasmid and phage family-level groups were clustered on the basis of the correlation of their group-level abundance dynamics. The groups are displayed on the right of the heatmap. The coloured block on the right and bottom of the heatmap represents the six clusters emerging from the hierarchical clustering, represented by the trees at the top and left of the heatmap. The shown Pearson correlations have a significant level of $P < 0.001$ (that is, threshold). Statistical tests were two-sided and adjusted for multiple comparison. **b**, Upper: models based on the longer-term dynamics. Lower: models based on three shorter-term dynamics. The models are based on the group-level relative abundance values. Longer-term dynamics are represented by all data points from the entire time series. The shorter-term intervals were defined around the shift in community structure, at which the abundance of Microthrixaceae family drastically decreases. Exact sampling dates of the shorter-term intervals are highlighted in the *x* axis. Three models were applied to the longer- and shorter-term time intervals. The relative abundance of the Microthrixaceae family is included for reference.

abundance exhibited a higher number of overall CRISPR-mediated interactions (Fig. 4 and Supplementary Videos 1 and 2), which was further supported by the negative correlations with the total number of plasmid–host interactions over time ($r=-0.33$, $P=0.017$) and phage–host interactions over time ($r=-0.40$, $P=0.004$). However, after focusing on the population-level CRISPR-based iMGE–host interactions of *M. parvicella*, we observed a positive correlation between the population abundance over time and its number of iMGE–host interactions, that is, plasmid–host ($r=0.63$, $P\approx 0$) and phage–host ($r=0.25$, $P=0.02$). Finally, the



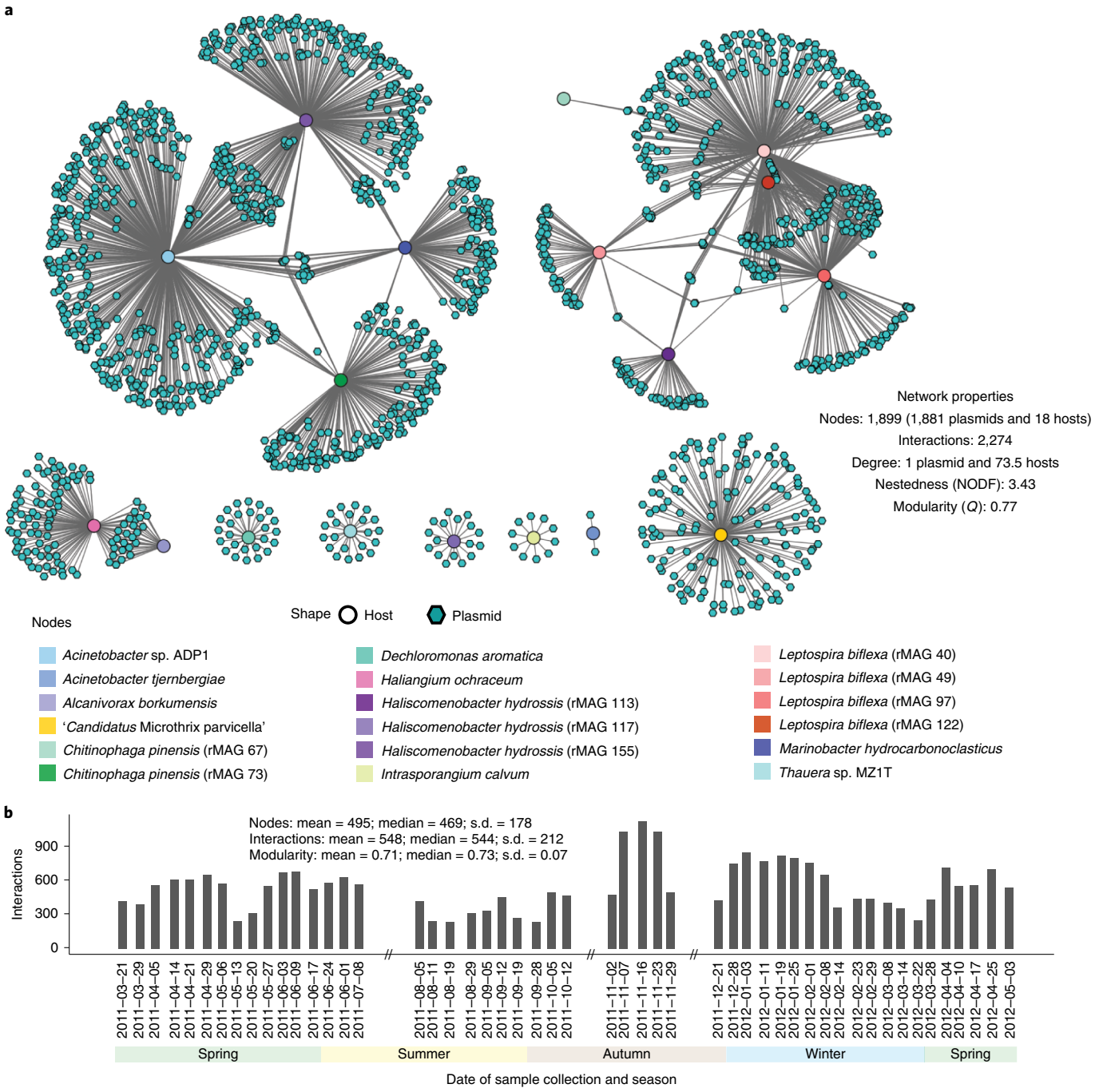


Fig. 3 | Networks of plasmid–host interactions. a, A bipartite network representing global CRISPR-based interactions from the entire time series involving bacterial hosts (multicoloured circular nodes) and their associated plasmids (turquoise hexagonal nodes). The edges represent at least one spacer at one time point from the host targeting the corresponding plasmid. **b**, Number of plasmid–host CRISPR-based interactions. Each bar represents the total number of interactions in a specific time point ($n=1$), for each of the 51 time points in the time series. The summary statistics within the panel represents the number of CRISPR-based interactions over the entire time series ($n=51$ in situ samples).

iMGE–*M. parvicella* network exhibited a highly modular structure, whereby a set of iMGEs interacted with its set of spacers (Fig. 4).

We identified a single contig of 10,224 base pairs in length that encoded a complete CRISPR operon[40]. This contig shared 97.62% sequence identity with '*Candidatus* Microthrix parvicella Bio17-1'[37] (Supplementary Note 9). Briefly, the contig contained 6 *cas* genes and 11 CRISPR repeats. Using the MT and MP data, we found that the *cas* genes within the rMAG were expressed over time, with *Cas2* showing the highest level of gene expression while *Cas7* was found more frequently at the protein level (Fig. 4). We were able to link a total of 670 spacers across the entire time series to this specific CRISPR locus. These spacers were present within an average of 25.5 time points (median = 28.5, s.d. = 14). Out of all the associated spacers, 433 lacked matches within the time series and 246 could be linked to a protospacer in at least one time point. Among these, 64 targeted plasmids, 24 targeted phages and 12 targeted both plasmids and phages (Fig. 4). Ten out of the 12 spacers targeting both had matches in protein-coding genes, including sigma 70 factor of

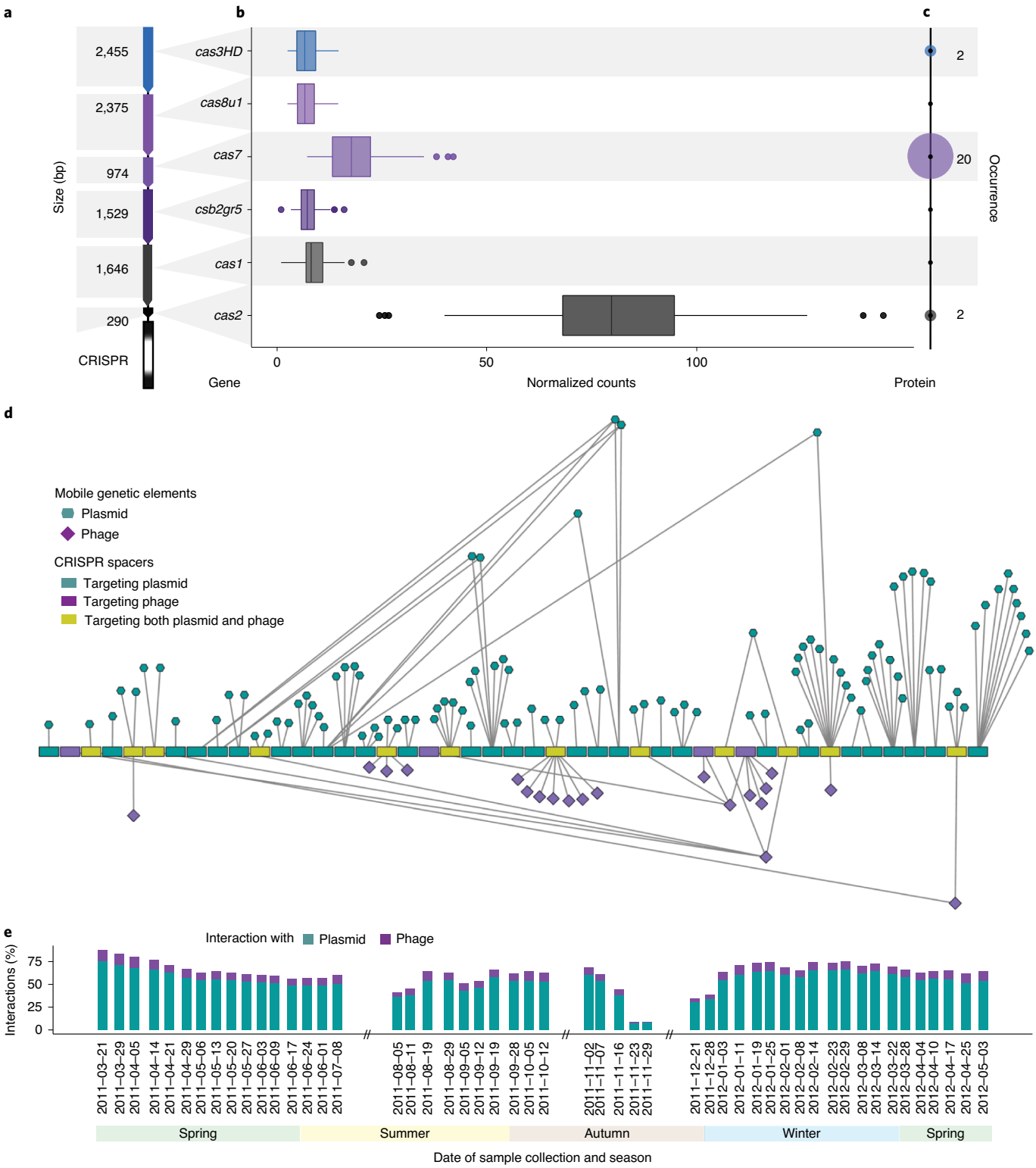


Fig. 4 | The CRISPR–Cas system of *M. parvicella*. **a**, The CRISPR-*cas* locus predicted within a reconstructed population-level genome (rMAG-165) identified as *M. parvicella*. **b**, MT-based expression levels of the corresponding *cas* genes. Boxplots represent expression levels aggregated from 51 time points based on normalized read counts. Data are presented as median values, Q1–1.5 × interquartile range (IQR) and Q3 + 1.5 × IQR. **c**, MP-level representation of Cas proteins. The numbers represent the number of time points at which at least one peptide of the corresponding Cas protein was detected. **d**, Representation of the active CRISPR spacers (gain or loss of spacer within the time series) assigned to *M. parvicella*. The order of the spacers is based on their first occurrence within the time series. **e**, Spacer-iMGE-based interactions represented per time point as percentages of the global interactions of *M. parvicella*.

RNA polymerase, GDSL-like lipase 2 and helix-turn-helix domain 23, which are genes known to be widely encoded by both plasmids and phages. Additionally, we inspected the activity of spacers within the CRISPR loci and observed 45 spacers with gain or loss events (Fig. 5). Similar to the community level, there was also a shift in gain to loss events occurring after the community shift on

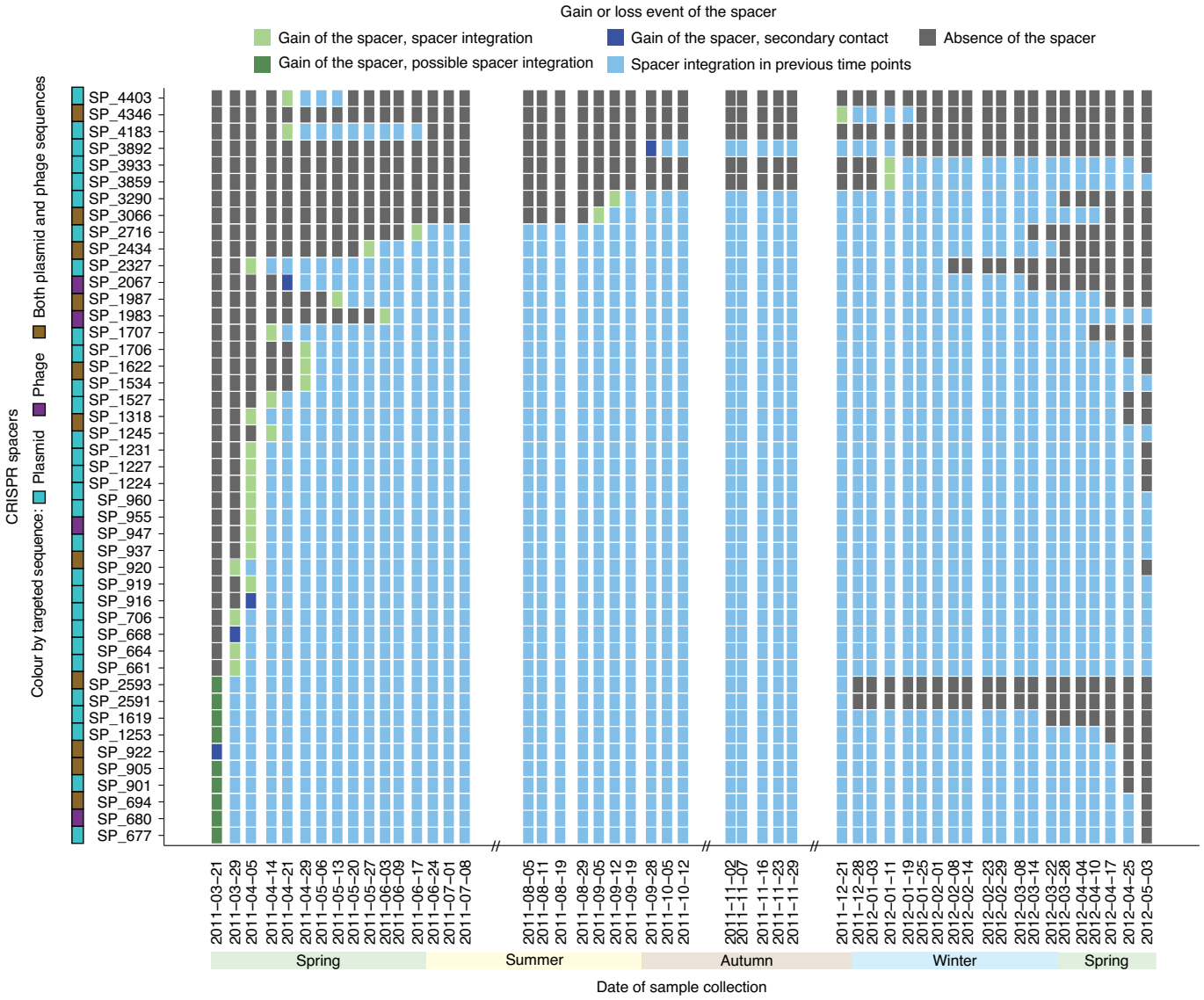


Fig. 5 | Spacer acquisition dynamics in the *M. parvicella* population. Dynamics of spacers assigned to the *M. parvicella* population. The *y* axis includes the spacer identities. The coloured boxes next to the spacer identities indicate the type of iMGE targeted by that spacer. The boxes within the plot are coloured based on the presence (light blue) or absence (dark grey) of the spacer within the CRISPR array for each time point. Green boxes represent spacer gain events, specifically light green for spacer integration (iMGE is detected before the spacer) and dark green for a putative spacer integration event (iMGE and spacer are detected at the same time point). Dark blue boxes represent potential secondary contact events (spacer detected before the iMGE).

2011-12-28 (Extended Data Fig. 8). Overall, the *cas* gene and Cas protein expression levels, coupled to spacer dynamics targeting more plasmids (example shown in Extended Data Fig. 9) than phages, demonstrate a highly active CRISPR–Cas system within *M. parvicella*.

In contrast to *M. parvicella*, other populations exhibited more dynamic CRISPR loci, such as the rMAG-40 classified as *L. biflexa*, and less dynamic loci, such as the rMAG-31 classified as *Intrasporangium calvum* (Supplementary Note 10). *L. biflexa* has eight putative CRISPR loci and a locus of *cas* genes classified as type V (Supplementary Table 16 and Extended Data Fig. 10), and these contained a total of 680 spacers, of which 146 exhibited gain or loss within the time series. The population with the highest amount of spacers was rMAG-73, which was classified as *C. pinensis*, with CRISPR type III and a total of 1,119 spacers, of which 306 were active (that is, with either gain or loss events). Overall, the size of the CRISPR locus did not directly relate to spacer gain or loss. Finally, we observed that different population-level CRISPR–Cas dynamics exist at the level of gene and protein expression as well as spacer integration activity. Based on our results, *M. parvicella* populations contain a functional CRISPR system, but use it sparingly compared with other populations.

Discussion

We presented an extensive time-resolved, integrated meta-omics analysis of CRISPR-mediated iMGE–host interactions. Given the vast extent of unresolved bacterial taxa as well as plasmid and phage sequences in this community, the reliance on existing sequence databases would have greatly limited the analysis of key community members. Our reference-independent approach, including de novo genomic assembly, binning and plasmid/phage prediction, were required to analyse this dataset. We were able to link microbial population genomes (rMAGS) to iMGEs using spacer–protospacer links[24], unlike previous approaches that have relied on

abundance levels[41]. Overall, our approach of resolving interaction dynamics between iMGEs and their hosts revealed an enrichment in CRISPR-based plasmid targeting relative to phages.

To extract coherent information across the time series, we minimized redundancy concerning population-level genomes, CRISPR information and iMGEs. The aforementioned procedures may potentially result in a dilution of information, especially regarding underlying species- and strain-level diversity. However, this trade-off was necessary considering the inherent properties of the time-series dataset, namely, in relation to the appearance, disappearance and/or reappearance of features over time. More importantly, our stringent methodology allowed us to balance the advantages of a de novo assembly-based approach, that is, detecting novel microbial and iMGE populations, while enabling us to track the populations over time.

We systematically optimized the plasmid and phage prediction process by applying an ensemble approach to reduce bias stemming from a single tool, establishing associations of iMGEs and specific rMAGs through binning, identifying strong correlations between iMGEs and their associated rMAGs and using spacer–protospacer links to establish empirical evidence of interactions between rMAGs and iMGEs. Despite this, several limitations must be addressed, including the inherent inaccuracies of the plasmid and phage prediction tools, the inability to predict prophages within community and the lack of reliable taxonomic classifications of iMGEs.

Our ensemble approach for iMGE identification demonstrated that plasmids are highly abundant within the community. The stepwise linear modelling approach demonstrated that plasmids have a more pronounced impact on the dominant Microthrixaceae compared to phages. Furthermore, based on the extracted protospacer information, plasmids are targeted more often than their phage counterparts by CRISPR systems. In contrast to previous studies focused on CRISPR-mediated immunity against phages, our results support the notion that plasmids also play key roles in the adaptation and promotion of diversity[42]. In this context, BWWTPs are thought to be hotspots for the spread of ARGs through iMGEs[3,43]. Our data revealed a comparatively small fraction of plasmids encoding ARGs that are targeted by CRISPR systems, which suggests that bacteria retain potentially beneficial plasmids[44], for example, those encoding ARGs[45], but further detailed investigation including data from longer-term time series is required.

The period with decreased Microthrixaceae abundance (from 2011-11-02 to 2012-01-25) coincided with the increased in abundance of other families (for example, Leptospiraceae or Moraxellaceae), their corresponding plasmids and overall CRISPR-mediated interactions. Based on this information, the increase in plasmids suggests a short-term fitness advantage for Leptospiraceae and Moraxellaceae populations, on the one hand. On the other hand, CRISPR-mediated links indicated CRISPR-based suppression of those plasmids in a possible drive towards the normalization of community structure and function, including the dominance of *M. parvicella*. However, any direct cause–effect relationships remain to be further explored under controlled laboratory conditions.

In relation to phages, we found that they tended to correlate with specific families, for example, Moraxellaceae and Leptospiraceae, which exhibited acyclical dynamics in relation to the Microthrixaceae family, but showed a smaller effect in the linear models. Additionally, rMAG populations within the Moraxellaceae and Leptospiraceae families exhibited higher CRISPR activity in terms of phage-linked spacer gain or loss. In that regard, phages are known to affect specific populations, which, according to our data, does not include the dominant *M. parvicella*, as previously observed[46]. Therefore, future studies need to be directed towards deciphering the roles of individual plasmids and phages on specific populations, as well as the community as a whole.

Based on our observations, a strong case can be made to include iMGEs and CRISPR-based interactions as additional features into models that incorporate abiotic parameters (for example, temperature, pH and oxygen concentration) and biotic drivers (for example, population dynamics and inter-microbial population interactions)[41,47,48], especially when such information can be extracted from MG data. The inclusion of such additional features may provide a more comprehensive model of community dynamics and process performance.

Finally, the composition of CRISPR loci is highly environment-specific[49], which should translate into environment-specific CRISPR-mediated interactions. Therefore, the present study should be repeated on samples from other environments to provide a broader understanding of CRISPR-based interactions in relation to iMGEs[50].

Methods

Sampling. Individual floating sludge islets within the anoxic tank of the Schifflange BWWT plant (Esch-sur-Alzette, Luxembourg; 49° 30′ 48.29″ N; 6° 1′ 4.53″ E) were sampled according to previously described protocols[15]. Samples are indicated as dates (YYYY-MM-DD). Time-resolved sampling included two initial sampling dates (2010-10-04 and 2011-01-25) as previously reported[15,48]. More frequent sampling was performed from 2011-03-21 to 2012-05-03, of which data from three samples (2011-10-05, 2011-10-05 and 2012-01-11) have been previously published[15].

Concomitant biomolecular extraction and high-throughput meta-omics. Concomitant biomolecular extraction of DNA, RNA and proteins as well as high-throughput measurements to obtain MG, MT and MP data were carried out according to previously established protocols[15,48,51].

Isolate culture, genome sequencing and assembly. A total of 85 isolate cultures of lipid-accumulating bacterial strains were derived from the sludge islets sampled from the same anoxic tank described above. The isolation protocol, including screening for lipid-accumulation properties (via Nile Red staining), DNA extraction and sequencing, was performed as previously described[48,51]. The genomic data were assembled and analysed using an automated version of a previously described workflow[51] that spanned sequencing read preprocessing, de novo assembly and gene annotation (see the section "Code availability"). The genome of '*Candidatus* M. parvicella Bio17-1' was obtained from the publicly available NCBI BioProject database PRJNA174686 (ref. [37]).

Co-assembly of MG and MT data. Sample-wise integrated MG and MT data analyses were performed using IMP[21] (v.1.3) with the following customized parameters: (1) Illumina Truseq2 adapters were trimmed; (2) the step involving the filtering of reads of human origin step was omitted for preprocessing; and (3) the MEGAHIT de novo assembler[52] was used for the co-assembly of MG and MT data. Nonpareil2 (ref. [53]) was applied to the preprocessed MG and MT data to assess the relative depth of coverage.

MP analyses. Raw mass spectrometry files were converted to MGF format using MSconvert with default parameters. The resulting files were used to run the Graph2Pro pipeline[54] together with the corresponding assembly graphs from MEGAHIT, which allowed the integration of MG, MT and MP data. Assemblies often result in fragmented consensus contigs, thus leading to a loss of information on strain variation and to open-reading frames spanning multiple contigs. The Graph2Pro pipeline combines the Graph2Pep algorithm and FragGeneScan[55] to predict peptides from short and long edges of the graph even if the peptides span multiple edges. Graph2Pro further predicts protein sequences from the graphs of the IMP-based co-assemblies using identified peptides as constraints. To produce the final protein identifications, MP data were searched against the sample-specific databases derived from Graph2Pro.

The combined set of tryptic peptides was used as the target database for peptide identification using the MS-GF+ search engine[56] and customized parameters. The instrument type was set to a high-resolution LTQ with a precursor mass tolerance of 15 ppm and an isotope error range of −1 and 2. The minimum and the maximum precursor charges were set to 1 and 7, respectively. The false discovery rate (FDR) was estimated by using a target-decoy search approach, whereby reverse sequences of the protein entries were generated while preserving the carboxy-terminal residues (KR) and concatenated to the database. All identifications were filtered to achieve an FDR of 1%.

Identified peptides from the Graph2Pro pipeline were assigned using peptidematch[57] against Prokka-based[58] predictions from IMP for protein-coding sequences of the rMAGs, and prodigal-based predictions[59], including fragmented genes (see section "Gene annotation of phage- and plasmid-derived contigs" below) for protein-coding sequences of the iMGEs.

Binning, selection of representative genome bins, taxonomy and estimation of abundance. Co-assembled contigs from each time point were binned as previously described[60]. Binning was based on nucleotide signatures, presence of single-copy essential genes and MG depth of coverage. Bins from each time point with at least 28% completeness and contamination of less than 20% along with the 85 isolate genomes were subjected to a dereplication process using dRep[22] (v.0.5.4) to select rMAGs. Accordingly, the following dRep parameters were set: (1) genome completeness of 0.6 (based on CheckM[61] (v1.0.7)); (2) strain heterogeneity of 101; (3) average nucleotide identity (ANI) threshold of 0.6 to form primary clusters; and (4) ANI threshold of 0.965 to form secondary clusters. Taxonomic classification was performed using a customized version[62] of AMPHORA2 (ref. [63]). Additionally, taxonomic classification was performed using sourmash[64] 2.0.0a1-lca-version with a kmer-length of 21 and a threshold of 4 using an existing database that included around 87,000 microbial genomes (downloaded on 09 November 2017 from https://osf.io/s3jx8/download).

AMPHORA2-based predictions for individual marker genes were combined via the summation of the associated assignment probabilities. If the summed probability scores for the highest-scoring taxonomic level constituted less than one-third of the total probability scores, the assignment was discarded as a 'low confidence assignment'. Taxonomic assignments of AMPHORA2 and sourmash-lca were combined and then filtered to select a final taxonomic assignment for the rMAGs, giving priority to predictions from sourmash-lca due to higher expected specificity and an updated database. We then selected rMAGs with a 'completeness − contamination' value of ≥50% for further downstream analyses.

To represent population-level abundance and transcription levels, the preprocessed MG and MT paired- and single-end reads from all the time-series samples were mapped onto the collection of rMAGs using bwa mem[65], and contig-level average depth-of-coverage values were extracted for the MG and MT data. Gene-level MT read counts for all the predicted genes present within each rMAG were normalized using R statistical software to obtain the corresponding gene expression values.

Identification of CRISPR elements. CRISPR information (that is, spacers, repeats and flanking sequences) were predicted using CRASS[66] (v.0.3.8) based on the IMP-based preprocessed MG and MT paired- and single-end reads as input. MetaCRT[67] was used to predict spacers and repeats from IMP-based MT and co-assembled contigs. A custom script was used to extract flanking regions from the metaCRT results.

The redundancy of spacers, repeats and flanking sequences was reduced by clustering the sequences with CD-HIT-EST[68] (v.4.6.7). Spacers were clustered using 90% sequence identity[69,70], covering the entire length of the compared sequences[69]. CRISPR-flanking regions were clustered using 99% sequence identity, with at least 97.5% coverage of both the compared sequences. Conversely, the CD-HIT-EST clustering parameters for repeats were manually determined by clustering the known repeats belonging to a single CRISPR locus of '*Candidatus* M. parvicella Bio17-1'[37]. Specifically, the sequence identity parameter was first set to 99% and the sequence coverage was set to 100%. These parameters were reduced by 5% in the subsequent iterations until all repeats were regrouped into a single cluster. Next, all the known repeats of *M. parvicella* were clustered at 80% sequence identity, covering the length of at least 75% of the shorter sequence. These parameters were used for the clustering of all repeats. FASTA headers of all the sequences were left unchanged (that is, *-d* parameter in CD-HIT-EST) because they contained information required for downstream analyses (for example, sample name, contig name and CRASS-computed coverage). The clustering procedure for the different CRISPR elements yielded non-redundant sequences of repeats, spacers and flanking regions.

Spacer abundance values were estimated by extracting their coverage values from CRASS. Equivalent information was obtained from metaCRT by using bwa-mem to map MG and MT reads from each of the time-resolved samples to the entire set of contigs predicted by metaCRT (that is, contigs containing at least one CRISPR locus). The depth-of-coverage information was derived using bedtools[71]. Based on this, abundance values were extracted for each of the predicted spacers per time point. The depth-of-coverage information of the metaCRT contigs was then consolidated using CRASS coverage results by referring to the non-redundant spacer clusters (derived from CD-HIT-EST). The consolidated results are hereafter referred as 'spacer abundance values'. Specifically, the spacer abundance values from the specific time points were assigned to the non-redundant spacers, thereby allowing a temporal representation of spacer abundance values. Subsequently, the spacer abundance values were transformed to counts per million (c.p.m.)[72,73] per sample, and non-redundant spacers that had at least one read count in at least one sample were selected and the c.p.m. values were calculated. Finally, to determine the presence/absence of a given spacer, a minimum cut-off value of c.p.m. = 1 was applied. Applying standard cut-offs (that is, above 3–5) caused loss of information from the short spacer sequences within the repetitive CRISPR regions, which usually do not recruit many reads during the mapping process.

Linking rMAGs to CRISPR elements. The non-redundant flanking regions and repeats were used to associate MAGs with specific CRISPR loci using BLASTN[74]. Non-redundant CRISPR-flanking sequences and CRISPR repeats were searched against the contigs of the MAGs. Flanking sequences and MAG contig(s) exhibiting similarities of at least 95% identity and coverage of either (1) 80% for flanking sequences >100 bp or (2) 95% for flanking sequences <100 bp were retained for the downstream filtering steps. Next, the aforementioned flanking sequences for which the associated repeats had at least 75% identity and 80% coverage against the MAG contig(s) were further retained for downstream processing. After defining the selected flanking repeat sequences linked to a MAG, spacers linked to the repeat flanking sequences were then associated to the MAG. In this way, the composition of spacers per MAG was determined. Finally, all the CRISPR information belonging to a MAG was linked to its rMAG to preserve the maximum amount of CRISPR information.

CRISPR types and subtypes and *cas* genes were predicted from all the assembled contigs using CRISPRone[23]. The *cas* genes and CRISPR types were then assigned to their respective MAGs.

We then selected rMAGs predicted as *M. parvicella* (see the section "Binning, selection of representative genome bins, taxonomy and estimation of abundance") to inspect the *cas* genes and CRISPR-type predictions. Next, we used CRISPRCasFinder[75] to further confirm the selected *cas* genes and CRISPR-type predictions of *M. parvicella*. We performed manual curation on all the rMAGs predicted as *M. parvicella*. We identified a contig (D47_L1.43.1_contig_476300) of 10,224 bp that encoded a complete CRISPR operon that was highly similar to the CRISPR operon of the isolate genome of '*Candidatus* M. parvicella Bio17-1'. This contig was incorporated with rMAG-165.

Identification of protospacers and protospacer-containing contigs. A BLASTN[74] search was performed using all non-redundant spacers as queries against the contigs from all time points using the parameters defined in CRISPRtarget[76]. Spacer matches with at least 95% coverage and 95% identity were selected for further analysis[32]. Any IMP-based MT results or co-assembled contigs containing repeat sequences and/or identified by metaCRT to encode CRISPR sequences were excluded from downstream analyses. Accordingly, the remaining spacer matches (or complements) were defined as protospacers, and the respective contigs that contained at least one protospacer were defined as PSCCs and were retained as iMGEs.

Classification of iMGEs. Bacteriophage sequences were predicted by analysing all co-assembled contigs using VirSorter[77] (v.1.0.3) and VirFinder[78] (v.1.0.0). Similarly, plasmid sequences were predicted using cBar[79] (v.1.2) and PlasFlow[80] (v.1.0.7). The predictions were consolidated by annotating candidate iMGE sequences as follows: 'plasmid' if the sequences were positively predicted by cBar and/or PlasFlow; 'phage' if the sequences were positively predicted by VirSorter and/or VirFinder; 'ambiguous' if the sequences were predicted as both plasmid and phage by any combination of the aforementioned tools; and, finally, 'unclassified' if they contained at least one protospacer and were not annotated as phage or as plasmid. Following this step, all iMGEs (that is, phages, plasmids, ambiguous and unclassified) were clustered using CD-HIT-EST with clustering parameters of 80% identity and at least 50% coverage, generating the non-redundant set of iMGEs. The classification/annotation of representative clusters was retained for the downstream analyses. Finally, BLASTN[74] was performed on the clustered contigs against NCBI plasmid and virus databases to retrieve their taxonomy.

Genomic and transcriptomic abundances of the iMGEs were obtained by mapping the IMP-preprocessed MG and MT paired- and single-end reads from all time points to the iMGE representative contigs using bwa-mem[65]. The contig-level average depth of coverage derived from the MG and MT data represented the iMGE abundance and iMGE gene expression, respectively.

Gene annotation of phage- and plasmid-derived contigs. Open reading frames within iMGEs were predicted using Prodigal[59] (v.2.6) with the "meta" and "incomplete gene" settings. Predicted genes were annotated using hmmsearch[81] against an in-house licensed version of the KEGG database[82]. KEGG function identifiers were then converted to the higher-level COG functional categories[83]. Finally, ARGs were annotated using hmmsearch against ResFam's full HMM database[84].

Linear model of community dynamics. Correlations of family-level groups, whereby plasmids and phages were assigned to bacterial families based on their previous contig assignments to MAGs, were calculated using the "rcorr" function within the Hmisc R package. Euclidean distances of the correlation vectors were calculated using the "dist" function (stats R package). Next, hierarchical clustering was applied on the calculated Euclidean distances, using the "hclust" function (stats R package). The tree was then cut with a height parameter of four (that is, $H = 4$), using the "cutree" function from R stats package[85].

The "lm" function from the R stats package was used to generate the models. To avoid overfitting, we restricted the linear models to a maximum of 15 family-level groups. Random sampling was performed for 100,000 model realizations, and model quality was assessed using the adjusted R^2 value. In our first approach, we did not restrict the model composition and allowed all combinations with the same probability. Then, from the random sampling data, we ranked models based on the adjusted R^2 value and looked for enrichments in specific families in the best models

($N = 25, 50, 100$). In the first iteration, we selected enriched families and iMGEs (that is, plasmids and phages) to obtain a global model, and then we selected the significant groups from the global model to obtain a reduced model. Once we had the models for the entire time series and the shorter-time intervals, we identified the common significant groups in all the models. Next, we removed the group Microthrixaceae plasmids from the reduced models for each time interval to assess the influence of these plasmids within the performance of the model.

Network analyses and visualization. CRISPR-based plasmid–host and phage–host networks were defined by the co-occurrence of rMAGs, spacers and a targeted iMGE in at least one time point. Thus, if a given non-redundant spacer was assigned to a specific rMAG and this specific rMAG did not co-occur in at least one time point, this spacer was deemed inactive within this rMAG throughout the time series. Consequently, a spacer was assigned to a rMAG if, and only if, the spacer co-occurred with its assigned rMAG in at least one time point. Thus, the iMGEs targeted by the spacers assigned to rMAGs were used to build the CRISPR-based plasmid–host and phage–host networks. Finally, the time-point-specific networks were built on the basis of the presence/absence of the rMAGs and their linked plasmids or phages.

Network properties such node degree, betweenness and closeness were estimated by the function "speciesLevel" within the bipartite R package[86]. Modularity, defined by the value of Q[87], and nestedness, defined as the value of the nestedness matrix based on overlap and decreasing fill (NODF)[88], were calculated using the functions "computeModules" and "nested", respectively.

Visualization and manual inspection of the networks were performed using Cytoscape[89] (v.3.6.1). R (v.3.4.1), together within the "tidyverse" framework, was used for processing data tables, statistical analyses and data visualization[90].

Estimation of spacer gain–loss and CRISPR locus dynamics. Based on the previously calculated c.p.m. per rMAG, their assigned spacers and iMGEs, the dates of the first and the last occurrence within the time series were defined. We subsequently defined events of gain and loss of spacers and possible secondary encounters of the iMGE with the rMAGs to resolve the variation within a given CRISPR array per population. These events were classified as follows: (1) gain of a given spacer if its first detection within the time series occurred after the first occurrence of its targeted iMGE; (2) probable gain of a given spacer if both the spacer and its targeted iMGE occurred for first time at the same time point; (3) probable secondary encounter if the spacer occurred for first time before its linked iMGE; (4) loss of a given spacer if last detection of the spacer occurred after the last detection of its linked iMGE; (5) probable loss of a given spacer if the last detection of both the spacer and the iMGE occurred at the same time point; (6) spacer loss before iMGE loss if the last occurrence of the spacer occurred before the last occurrence of the iMGE.

Workflow automation. Bioinformatics workflow automation was achieved using Snakemake[91] (v.3.10.2 to v.5.1.4).

Computing platforms. All computing was run on the University of Luxembourg High-Performance Computing (ULHPC) platform[92].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The genomic FASTQ files, rMAGs and isolate genomes from this work are publicly available within NCBI BioProject PRJNA230567. Similarly, MP data from this work are publicly available in the PRIDE database under the accession number PXD013655. Additional data are available via Zenodo (https://doi.org/10.5281/zenodo.3774024 and https://doi.org/10.5281/zenodo.3766442). Additional publicly available projects cited by this work include NCBI BioProject PRJNA174686. Source data are provided with this paper.

Code availability

The code is available on three separate repositories: (1) the IMP, binning and population genomes can be found in https://github.com/shaman-narayanasamy/LAO-time-series (https://doi.org/10.5281/zenodo.3988660); (2) the CRISPR and MGE analyses can be found in https://github.com/susmarb/LAO_multiomics_CRISPR_iMGEs (https://doi.org/10.5281/zenodo.3988592); and (3) the isolate assembly analyses can be found in https://github.com/shaman-narayanasamy/Isolate_analysis (https://doi.org/10.5281/zenodo.3988667).

Received: 8 December 2019; Accepted: 11 September 2020;
Published online: 02 November 2020

References

1. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR–Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14**, R40 (2013).
2. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR–Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).
3. Rizzo, L. et al. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review. *Sci. Total Environ.* **447**, 345–360 (2013).
4. Jassim, S. A. A., Limoges, R. G. & El-Cheikh, H. Bacteriophage biocontrol in wastewater treatment. *World J. Microbiol. Biotechnol.* **32**, 70 (2016).
5. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).
6. Samson, J. E., Magadan, A. H., Sabri, M. & Moineau, S. Revenge of the phages: defeating bacterial defences. *Nat. Rev. Microbiol.* **11**, 675–687 (2013).
7. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490 (2015).
8. del Solar, G., Giraldo, R., Ruiz-Echevarria, M. J., Espinosa, M. & Diaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).
9. Zhang, T., Zhang, X.-X. & Ye, L. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLoS ONE* **6**, e26041 (2011).
10. Houte, S. Van, Buckling, A. & Westra, E. R. Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.* **80**, 745–763 (2016).
11. Jansen, R., Embden, J. D. A., van, Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
12. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
13. Sun, C. L., Thomas, B. C., Barrangou, R. & Banfield, J. F. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J.* **10**, 858–870 (2016).
14. Muller, E. E. L., Sheik, A. R. & Wilmes, P. Lipid-based biofuel production from wastewater. *Curr. Opin. Biotechnol.* **30**, 9–16 (2014).
15. Muller, E. E. L. et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* **5**, 5603 (2014).
16. Narayanasamy, S., Muller, E. E. L., Sheik, A. R. & Wilmes, P. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microb. Biotechnol.* **8**, 363–368 (2015).
17. Rossetti, S., Tomei, M. C., Nielsen, P. H. & Tandoi, V. '*Microthrix parvicella*', a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiol. Rev.* **29**, 49–64 (2005).
18. Sheik, A. R. et al. In situ phenotypic heterogeneity among single cells of the filamentous bacterium Candidatus Microthrix parvicella. *ISME J.* **10**, 1274–1279 (2016).
19. Roume, H., Heintz-Buschart, A., Muller, E. E. L. & Wilmes, P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* **531**, 219–236 (2013).
20. Roume, H. et al. A biomolecular isolation framework for eco-systems biology. *ISME J.* **7**, 110–121 (2013).
21. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
22. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
23. Zhang, Q. & Ye, Y. Not all predicted CRISPR–Cas systems are equal: isolated *cas* genes and classes of CRISPR like elements. *BMC Bioinformatics* **18**, 92 (2017).
24. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
25. Crawley, A. B., Henriksen, J. R. & Barrangou, R. CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR–Cas systems. *CRISPR J.* **1**, 171–181 (2018).
26. Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
27. Amitai, G. & Sorek, R. CRISPR–Cas adaptation: insights into the mechanism of action. *Nat. Rev. Microbiol.* **14**, 67–76 (2016).
28. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
29. Bolotin, A., Quinquis, B., Sorokin, A. & Dusko Ehrlich, S. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
30. Callanan, J. et al. RNA phage biology in a metagenomic era. *Viruses* **10**, 386 (2018).

31. Tong, J. et al. Microbial community evolution and fate of antibiotic resistance genes along six different full-scale municipal wastewater treatment processes. *Bioresource Technol.* **272**, 489–500 (2019).
32. Shmakov, S. A. et al. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *mBio* **8**, e01397-17 (2017).
33. Davison, M., Treangen, T. J., Koren, S., Pop, M. & Bhaya, D. Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS ONE* **11**, e0160574 (2016).
34. Arbas, S. M. & Narayanasamy, S. *Number of genes per function within mobile genetic elements in Martinez Arbas, Narayanasamy et. al. (2020)* (Zenodo, 2020); https://doi.org/10.5281/zenodo.3774024
35. Che, Y. et al. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* **7**, 44 (2019).
36. Shchegolkova, N. M. et al. Microbial community structure of activated sludge in treatment plants with different wastewater compositions. *Front. Microbiol.* **7**, 90 (2016).
37. Muller, E. E. L. et al. Genome sequence of '*Candidatus* Microthrix parvicella' Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *J. Bacteriol.* **194**, 6670–6671 (2012).
38. Blackall, L. L. et al. 'Candidatus Microthrix parvicella', a filamentous bacterium from activated sludge sewage treatment plants. *Int. J. Syst. Bacteriol.* **46**, 344–346 (1996).
39. McIlroy, S. J. et al. Metabolic model for the filamentous '*Candidatus* Microthrix parvicella' based on genomic and metagenomic analyses. *ISME J.* **7**, 1161–1172 (2013).
40. Martinez Arbas, S. & Narayanasamy, S. *CRISPR locus information of M. parvicella in Martinez Arbas, Narayanasamy et. al. (2020)* (Zenodo, 2020); https://doi.org/10.5281/zenodo.3766442
41. Brown, M. R. et al. Coupled virus–bacteria interactions and ecosystem function in an engineered microbial system. *Water Res.* **152**, 264–273 (2019).
42. Davison, J. Genetic exchange between bacteria in the environment. *Plasmid* **42**, 73–91 (1999).
43. Li, L. et al. Estimating the transfer range of plasmids encoding antimicrobial resistance in a wastewater treatment plant microbial community. *Environ. Sci. Technol. Lett.* **5**, 260–265 (2018).
44. Jiang, W. et al. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet.* **9**, e1003844 (2013).
45. Murray, A. K. et al. Novel insights into selection for antibiotic resistance in complex microbial communities. *mBio* **9**, e00969-18 (2018).
46. Liu, R. et al. Phage–host associations in a full-scale activated sludge plant during sludge bulking. *Appl. Microbiol. Biotechnol.* **101**, 6495–6504 (2017).
47. Coenen, A. R. & Weitz, J. S. Limitations of correlation-based inference in complex virus–microbe communities. *mSystems* **3**, e00084-18 (2018).
48. Roume, H. et al. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms Microbiomes* **1**, 15007 (2015).
49. Kunin, V. et al. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.* **18**, 293–297 (2008).
50. Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. *Nucleic Acids Res.* **48**, 748–760 (2019).
51. Muller, E. E. L. et al. First draft genome sequence of a strain belonging to the Zoogloea genus and its gene expression in situ. *Stand. Genom. Sci.* **12**, 64 (2017).
52. 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 (2014).
53. Rodriguez-R, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**, 629–635 (2014).
54. Tang, H., Li, S. & Ye, Y. A graph-centric approach for metagenome-guided peptide and protein identification in metaproteomics. *PLoS Comput. Biol.* **12**, e1005224 (2016).
55. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191 (2010).
56. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
57. Chen, C., Li, Z., Huang, H., Suzek, B. E. & Wu, C. H. A fast peptide match service for UniProt knowledgebase. *Bioinformatics* **29**, 2808–2809 (2013).
58. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
59. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
60. 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).
61. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
62. Laczny, C. C. et al. Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract. *Front. Microbiol.* **7**, 884 (2016).
63. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
64. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
65. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
66. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* **41**, e105 (2013).
67. Rho, M., Wu, Y. W., Tang, H., Doak, T. G. & Ye, Y. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.* **8**, e1002441 (2012).
68. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
69. Moller, A. G. & Liang, C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ* **5**, e3788 (2017).
70. Lam, T. J. & Ye, Y. Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. *BMC Genomics* **20**, 567 (2019).
71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
72. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
73. Sha, Y., Phan, J. H. & Wang, M. D. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)* 6461–6464 (Institute of Electrical and Electronics Engineers, 2015).
74. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
75. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **2**, W246–W251 (2018).
76. Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget. *RNA Biol.* **10**, 817–827 (2013).
77. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
78. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
79. Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051–2052 (2010).
80. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35 (2018).
81. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
82. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
83. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
84. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
85. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2013).
86. Dormann, C. F., Gruber, B. & Fründ, J. Introducing the bipartite package: analysing ecological networks. *R News* **8**, 8–11 (2008).
87. Newman, M. E. J. Modularity and community structure in networks. *Commun. Law* **19**, 56–62 (2006).
88. Almeida-Neto, M., Guimarães, P., Guimarães, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
89. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
90. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).

91. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
92. Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. Management of an academic HPC cluster: the UL experience. In *Proc. 2014 International Conference on High Performance Computing & Simulation* (HPCS) 959–967 (Institute of Electrical and Electronics Engineers, 2014).

Acknowledgements

We thank the Luxembourg National Research Fund (FNR) for supporting this work through various funding instruments. Specifically, a PRIDE doctoral training unit grant (no. PRIDE15/10907093), CORE grants (nos. CORE/15/BM/10404093 and CORE/17/SM/11689322), a European Union ERASysAPP grant (no. INTER/SYSAPP/14/05), a proof-of-concept grant (no. PoC/13/02), a European Union Joint Programming in Neurodegenerative Diseases grant (no. INTER/JPND/12/01) and an ATTRACT grant (no. A09/03) all awarded to P.W., as well as an AFR Ph.D. (PHD-2014-1/7934898) grant to S.N. and a CORE Junior (C15/SR/10404839) grant to E.E.L.M. The project received financial support from the Integrated Biobank of Luxembourg with funds from the Luxembourg Ministry of Higher Education and Research. The work of P.M. was funded by the 'Plan Technologies de la Santé du Gouvernement du Grand-Duché de Luxembourg' through the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. The authors acknowledge the ULHPC for providing and maintaining the computing resources. We thank G. Bissen and G. Di Pentima from the Syndicat Intercommunal a Vocation Ecologique (SIVEC) for access to the Schifflange wastewater treatment plant.

Author contributions

S.M.A., S.N., E.E.L.M., P.M. and P.W. contributed to the planning and designing of the overall study and analyses. S.M.A., S.N., M.H., A.S., M.R.H., T.J.L., B.J.K., Y.Y. and S.L. contributed to the bioinformatics data analyses. E.E.L.M. and L.A.L. collected and performed the biomolecular extractions on the samples. N.D.H., C.M.L., L.B.P., J.D.G., J.M.S. and P.S.K. performed the DNA and RNA sequencing, while M.R.H. and R.L.M. performed the proteomic measurements. S.M.A., S.N., P.M., E.E.L.M., A.S., H.T., Y.Y., C.C.L., K.F. and P.W. participated in discussions related to this work. S.M.A., S.N., P.M., E.E.L.M. and P.W. wrote and reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

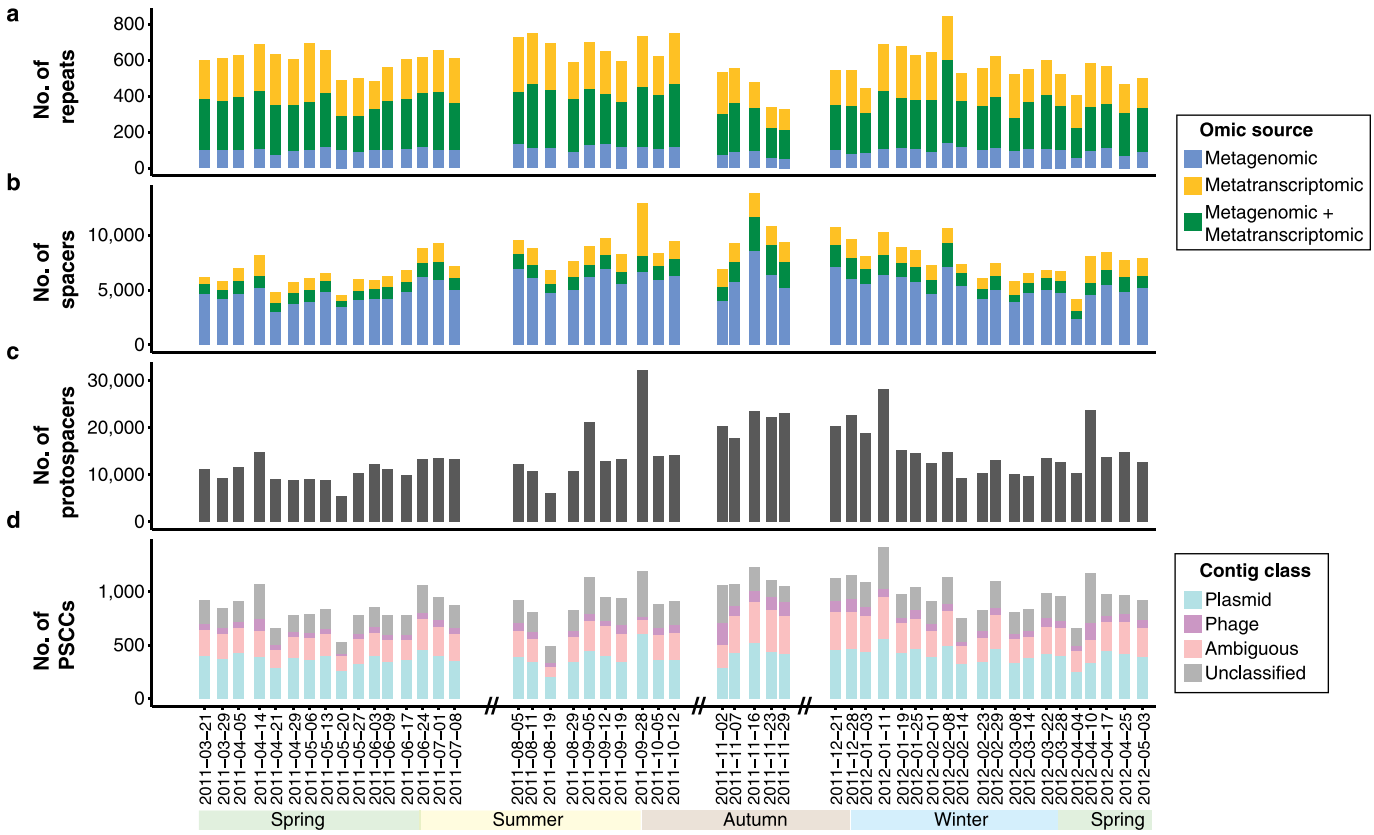
The authors declare no competing interests.

Additional information

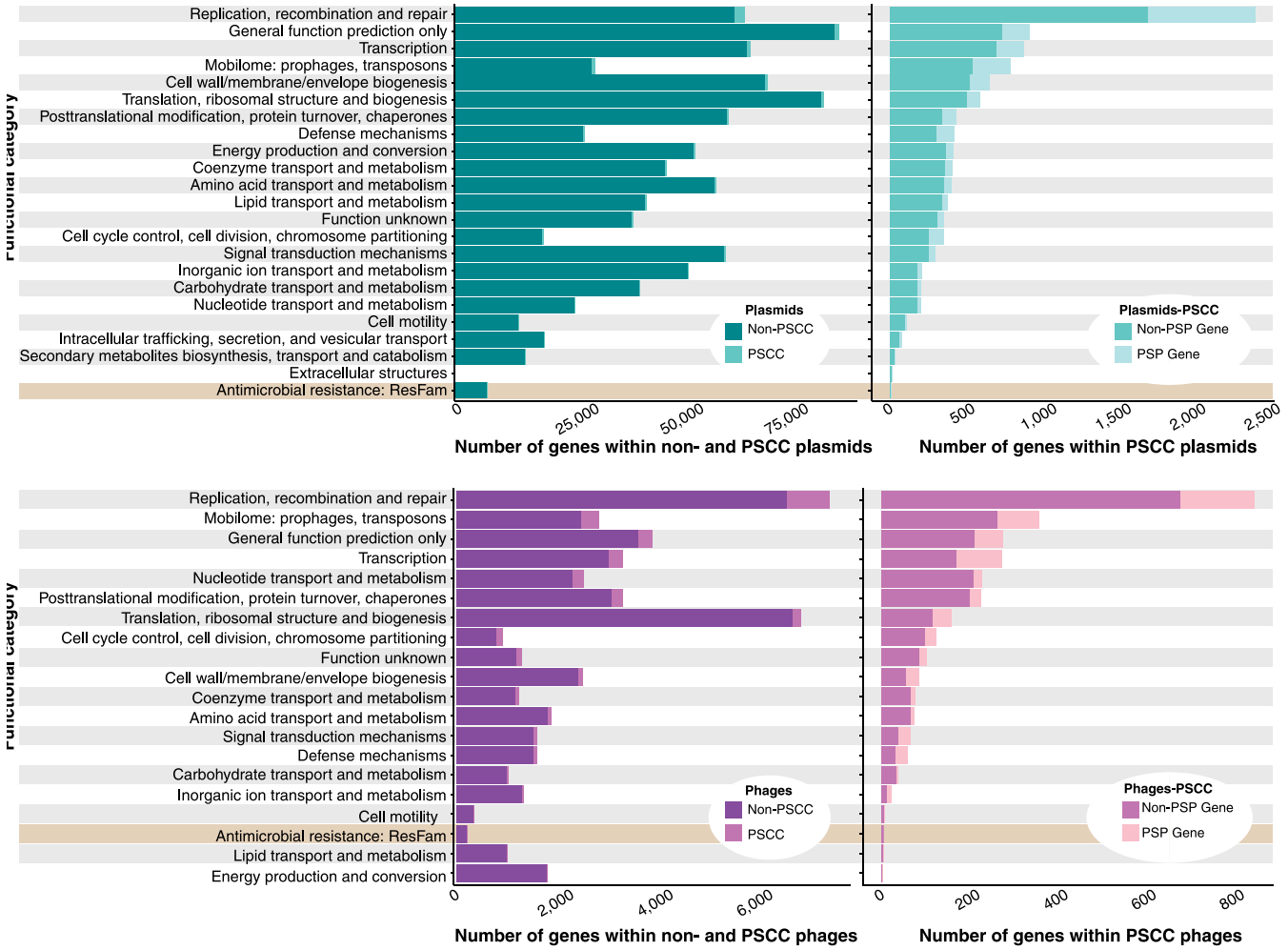
Extended data is available for this paper at https://doi.org/10.1038/s41564-020-00794-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-020-00794-8.

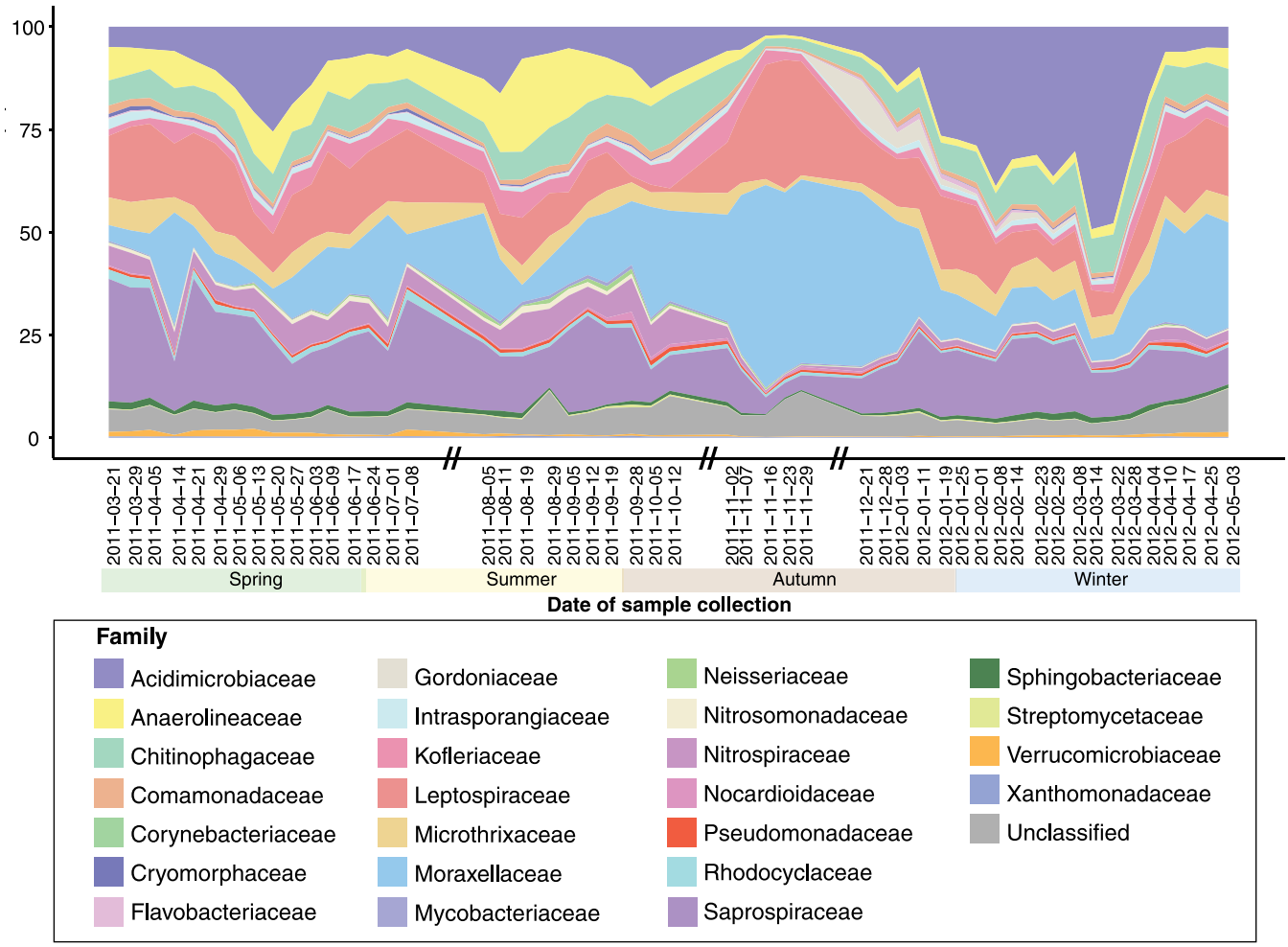
Correspondence and requests for materials should be addressed to P.W.



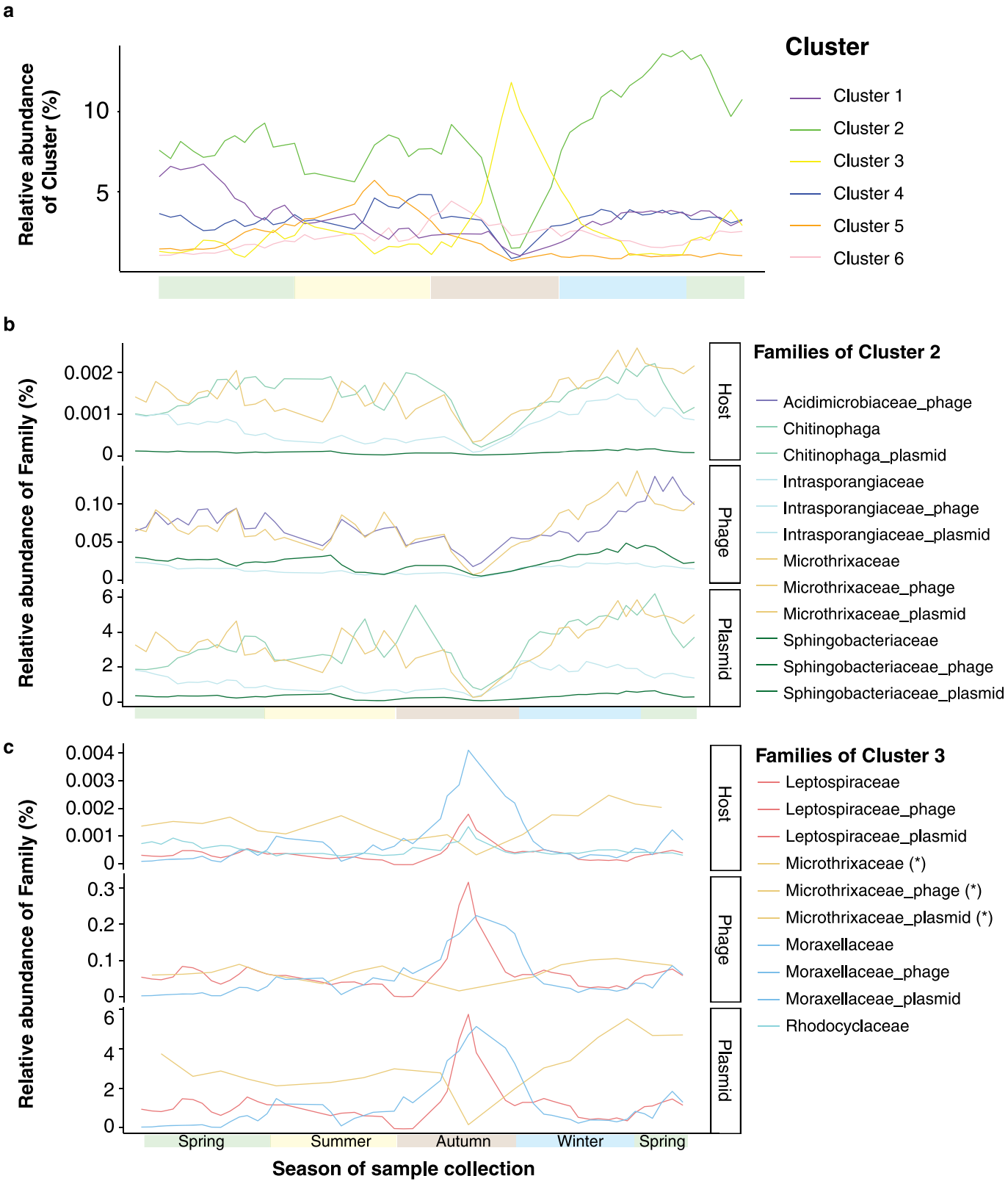
Extended Data Fig. 1 | Non-unique CRISPR elements, protospacers, and protospacer-containing contigs (PSCC) over time. Number of predicted **a**, repeats, **b**, spacers, **c**, protospacers, and d, PSCCs per time point. The labels in the x-axis indicate the exact sampling dates, and the double slashes (//) represent absence of samples due to absence of foaming islets.



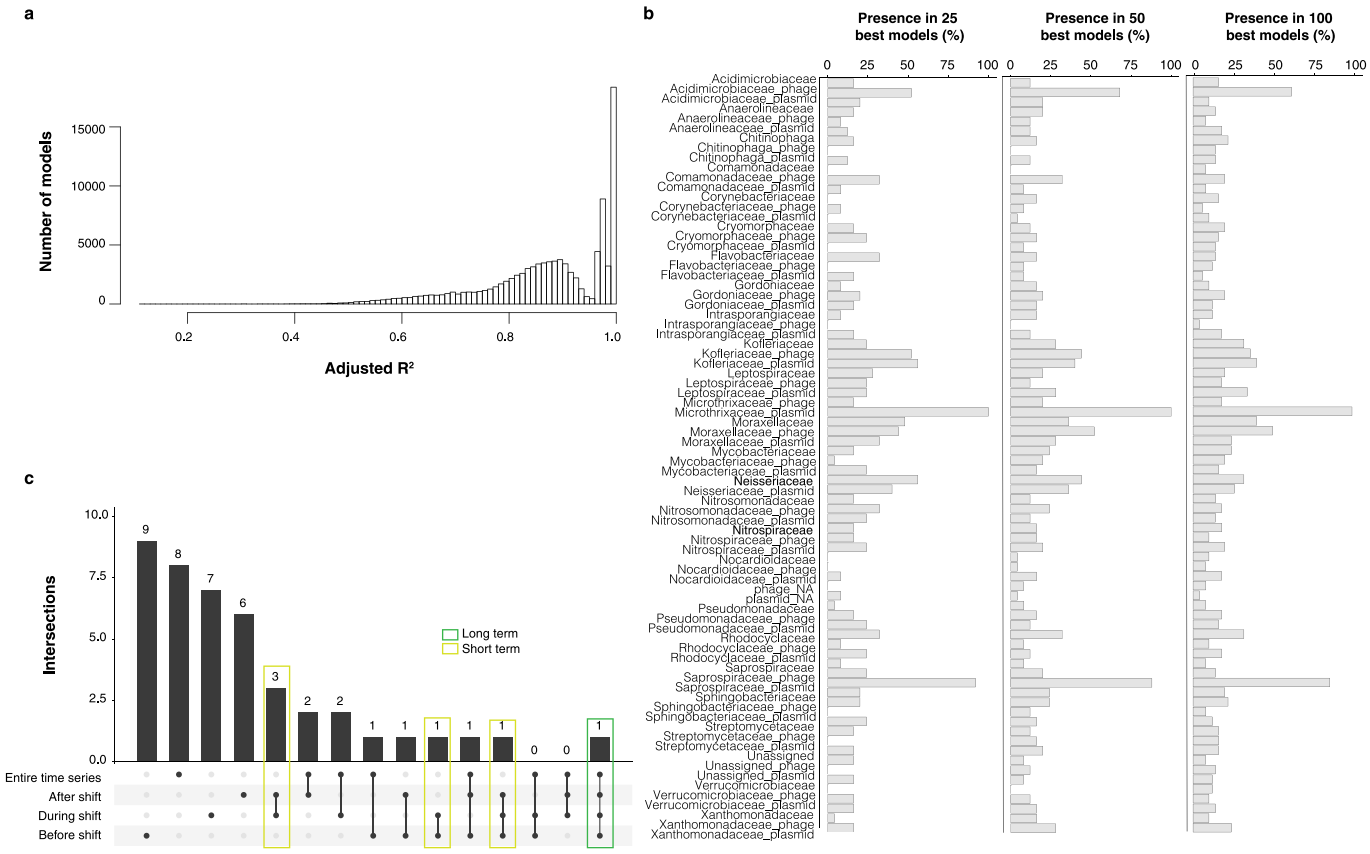
Extended Data Fig. 2 | Functional gene categories encoded and targeted within plasmids and phages. a, Functional categories encoded by plasmids and **b**, by phages. **a**, **b** Each bar indicates the number of genes found per functional category. The left bar plots show the number of genes of specific functional categories within invasive mobile genetic elements (iMGEs) with and without protospacers (that is PSCCs). For those iMGEs that are PSCCs, the right bar plot highlights the number of genes of specific functional categories for which protospacers occurred within the intragenic regions.



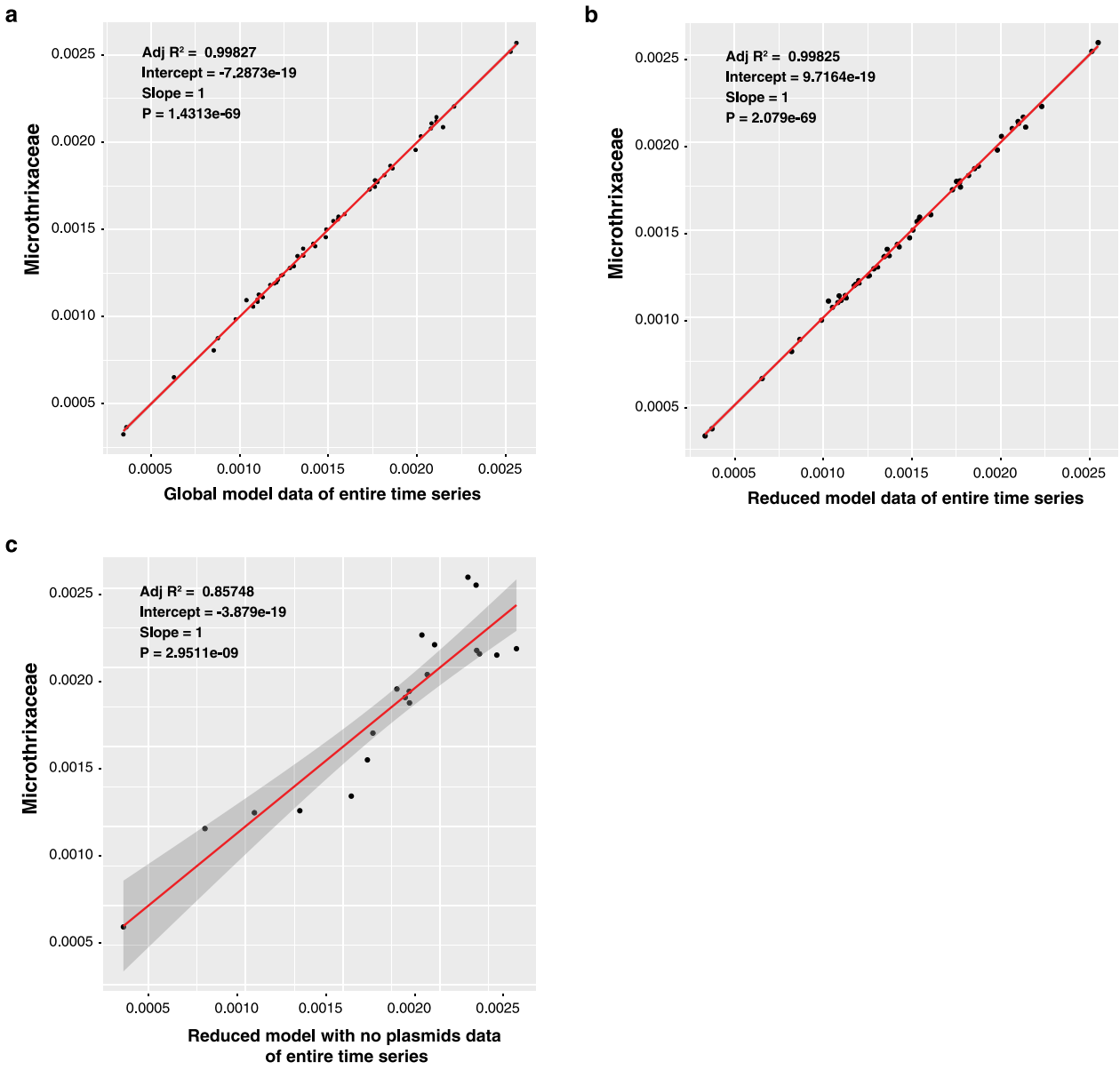
Extended Data Fig. 3 | Community activity. Relative expression based on mapping MT data to representative metagenomic assembled genomes (rMAGs) over time. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system.



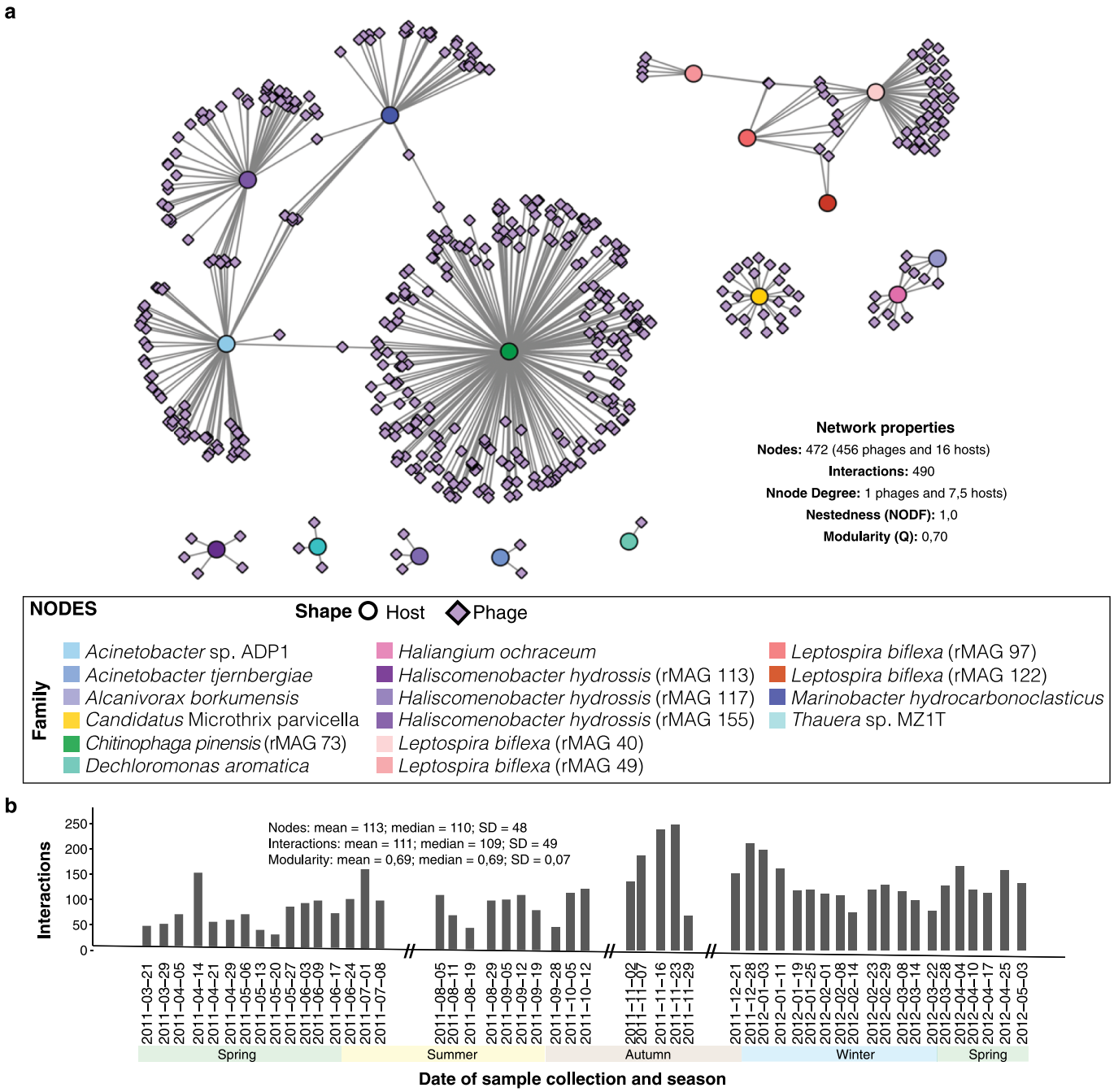
Extended Data Fig. 4 | Dynamics of clusters comprised of bacterial-, plasmid- and phage- groups. The rMAGs were grouped together at the family-level. Plasmids and phages were grouped based on their family-level association, that is, binned together with an rMAG of a given family. The bacterial, plasmid and phage groups were clustered based on the correlation of their cumulative group-level abundance dynamics. **a**, Dynamics of all clusters based on cumulative abundance of each cluster members. **b**, Dynamics of the cluster 2 members, including *Microthrixaceae* and its associated plasmids and phages as cluster members. **c**. Dynamics of the cluster 3 members, including *Microthrixaceae* and its associated plasmids and phages as reference (these groups are marked with an asterix). Relative abundance values on the y-axis were derived from MG data. The x-axis represents time, colour coded by seasons as labelled in panel **c**. Please refer to Fig. 1 for the exact sampling dates within the seasons.



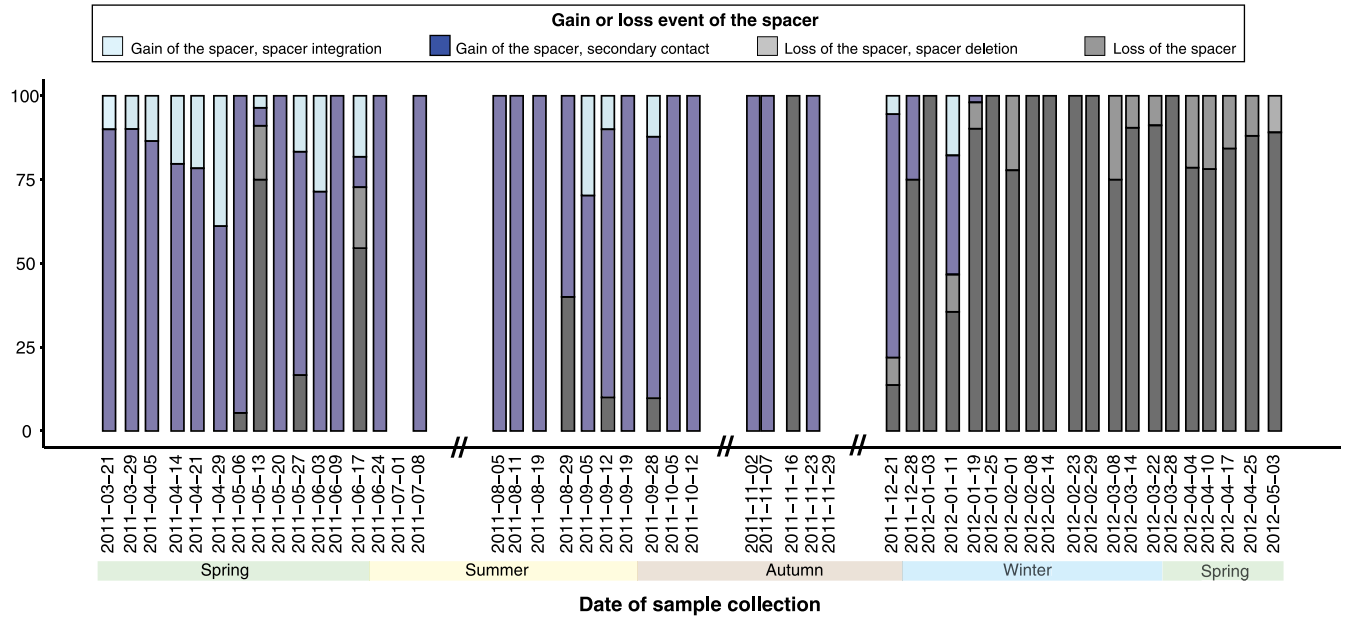
Extended Data Fig. 5 | Model fitness and family enrichment within the best models predicting *Microthrixaceae* family abundance. a, Distribution of the adjusted R^2 values of 100,000 model realizations. **b**, Enrichment of the family-level bacterial, plasmid and phage groups in the best 25, 50 and 100 models of the entire time-series. **c**, The upset plot represents the number of family-level bacterial, plasmid and phage groups (that is features) within the best model of different time intervals, that is the entire time-series and three time-windows (horizontal bars). The number of intersections between features in the best models in different long- and short-time intervals (vertical bars). The coloured boxes represent the intersections representing short- and long-term time dynamics, respectively.



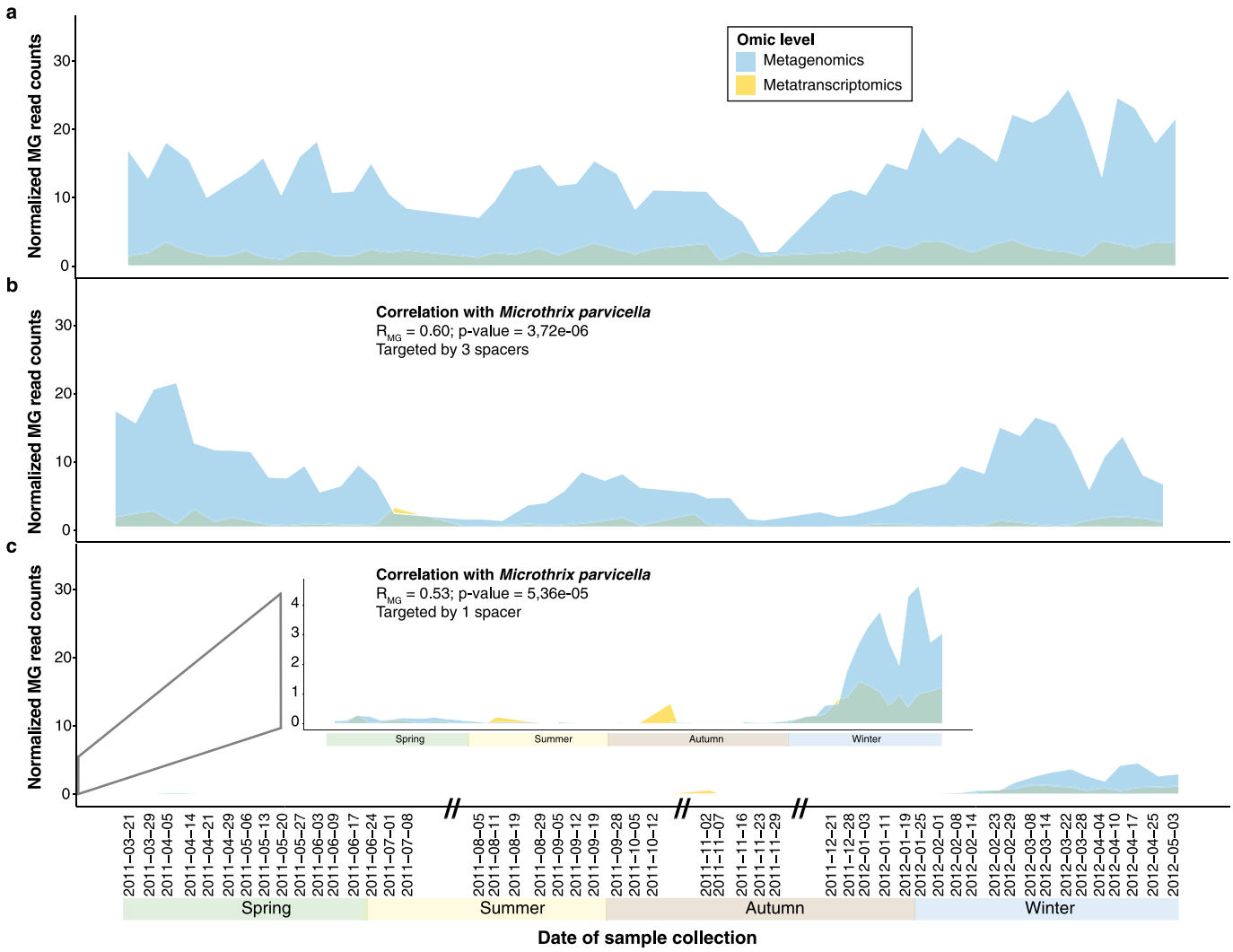
Extended Data Fig. 6 | Linear models predicting *Microthrixaceae* family abundance within the entire time-series. Model data fitted to the raw data of the entire time-series (n=51 *in situ* samples), specifically **a**, the best or global model, **b**, the reduced model, which lacks the non-significant families of the global model, and **c**, the reduced model without *Microthrixaceae*-plasmids. Gray bands represent the +/- standard error measurement of the regression line. Statistical tests were two-sided and adjusted for multiple comparisons.



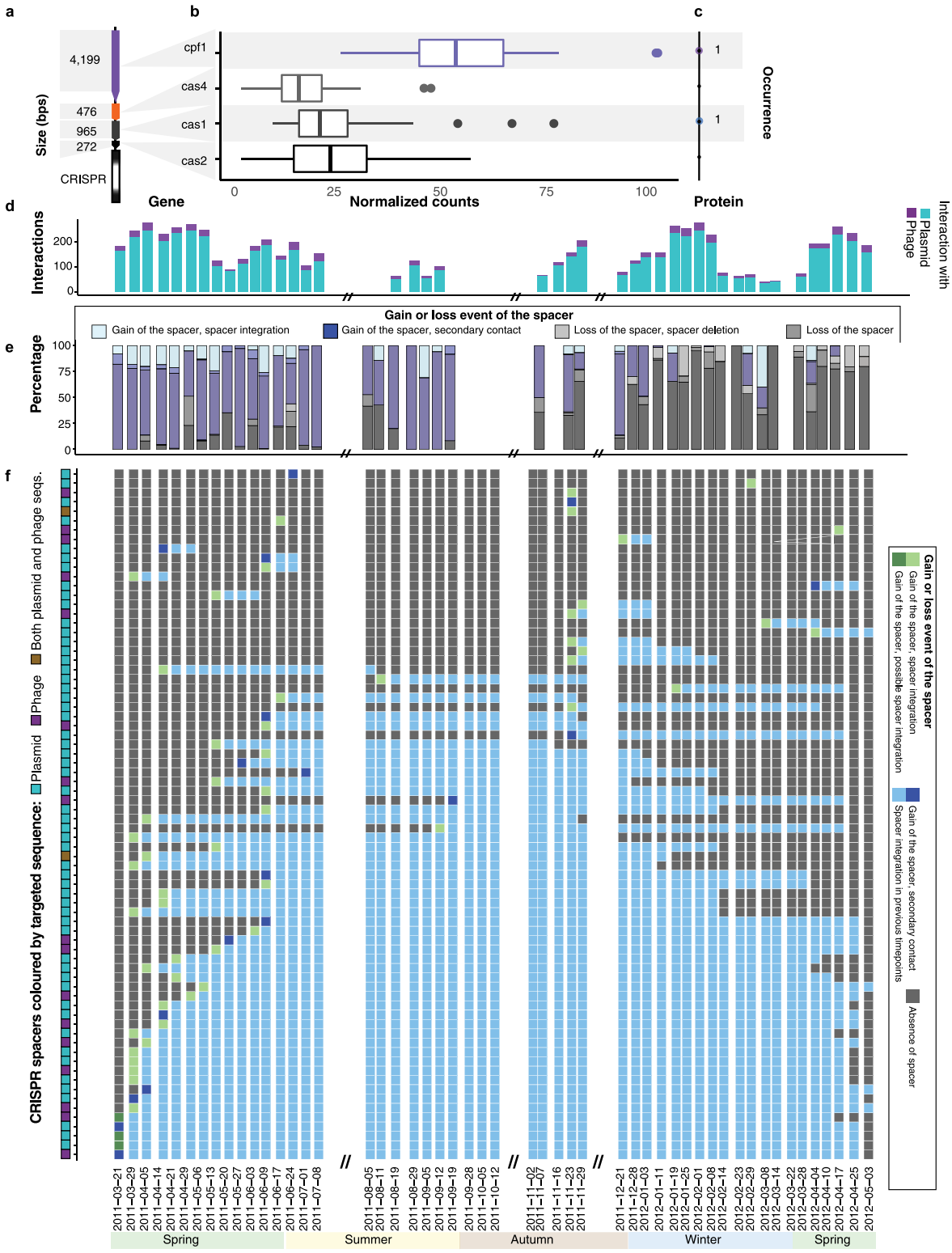
Extended Data Fig. 7 | Networks of phage-host interactions. a, Bipartite network representing global CRISPR-based interactions from the entire time-series between bacterial hosts (multicolored circular nodes) and their associated phages (purple diamond nodes). The edges represent at least one spacer from the host targeting the corresponding phage throughout the entire time-series. **b**, Number of phage-host CRISPR-based interactions. Each bar represents the total number of interactions in a specific timepoint (n=1), for each of the 51 timepoints in the time-series. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system. The summary statistics within the panel represents the number of CRISPR-based interactions in the entire time-series (n=51 *in situ* samples).



Extended Data Fig. 8 | Spacer acquisition dynamics in *Candidatus* Microthrix parvicella population. Barplot representing the percentage of spacers per time-point reflecting a gain or loss events. Gain events are defined as: i) "Gain of the spacer, spacer integration" when the iMGE was detected before or at the same timepoint as its linked spacer, and ii) "Gain of the spacer, secondary contact" when the spacer was detected before the linked iMGE within the time-series. Loss events are defined as: i) "Loss of the spacer, spacer deletion" when both the spacer and the iMGE are not detected anymore within the remainder of the time-series, and ii) "Loss of the spacer" when the spacer is not detected within the time-series anymore but the iMGE is still detected after spacer loss. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples from the sampled system.



Extended Data Fig. 9 | Abundance of *M. parvicella* and selected plasmid sequences targeted by the spacers of the same species. a, Metagenomics (MG)-based and metatranscriptomics-based (MT) abundance of *M .parvicella* over time. **b**, Abundance of plasmid contig "D28_L2.21_contig_56858", with a size of 2,503 bps which is targeted by three spacers within *M. parvicella*'s CRISPR locus. **c**, Abundance of plasmid contig "D48_E1.25_contig_355826", with a size of16,151 bps which is targeted by one spacer within *M. parvicella*'s CRISPR locus. Statistical tests were two-sided and adjusted for multiple comparisons.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Spacers acquisition dynamics of the rMAG-40 population classified as *Leptospira biflexi*. a, CRISPR-Cas operon. **b**, Metatranscriptomics-based expression levels of the cas genes. Boxplot represents expression levels aggregated from 51 timepoints based on normalized read counts. Data are presented as median values, Q1–1.5 x IQR and Q3 + 1.5 x IQR. **c**, Metaproteomic-level representation of Cas proteins. The numbers represent the number of time points where at least one peptide of the Cas protein was detected. **d**, Barplot representing the number of interactions between rMAG-40 and iMGEs. The purple section of the bars represent the number of interactions with phages, while in turquoise represent interactions with plasmids. **e**, Barplot representing the percentage of spacers per time-point with a gain or loss event. Gain events are defined as: i) "Gain of the spacer, spacer integration", when the iMGE was detected before, or at the same timepoint, as its linked spacer, and ii) "Gain of the spacer, secondary contact", when the spacer was detected before the linked iMGE, within the time-series. Loss events are defined as: i) "Loss of the spacer, spacer deletion", when both the spacer and the iMGE are not detected anymore within the rest of the time-series, and ii) "Loss of the spacer", when the spacer is not detected within the time-series anymore, but the iMGE is still detected after spacer loss. **f**, Dynamics of spacers assigned to the rMAG. The y-axis shows the IDs of spacers assigned to the rMAG. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system.

nature research

Corresponding author(s): Paul Wilmes

Last updated by author(s): 18/08/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Base calling of sequenced metagenomic (MG) and metatranscriptomic (MT) was processed using commercial software bundled within Illumina sequencing platforms to generate raw FASTQ data. Raw metaproteomic (MP) mass spectra were acquired using commercial software from Thermo Fischer Scientific. This work represents part of a larger ongoing multi-annual project. Please refer previous publications for detailed information on NGS and mass spectrometry platforms and the associated software for those platforms: https://doi.org/10.1038/ismej.2012.72 https://doi.org/10.1038/ncomms6603 https://doi.org/10.1038/npjbiofilms.2015.7 https://doi.org/10.1186/s40793-017-0274-y
Data analysis	All the code related to this work is available in three separate repositories: i) Integrated Meta-omics Pipeline (IMP), binning and population genomes: https://git-r3lab.uni.lu/shaman.narayanasamy/LAO-time-series, ii) CRISPR and mobile genetic element analyses: https://git-r3lab.uni.lu/susana.martinez/LAO_multiomics_CRISPR_iMGEs, iii) for the isolate assembly analyses: https://git-r3lab.uni.lu/shaman.narayanasamy/Isolate_analysis/activity. This information is included in the manuscript in the "Code availability" section. The software (and versions) used within this work include: IMP (ver. 1.3) Nonpareil (ver 2.0) Graph2Pro (no ver. number available) dRep (ver. 0.5.4) CheckM (ver. 1.0.7) R statistical package (ver. 3.4.1)

Cytoscape (ver 3.6.1)
bwa (ver. 0.7.17)
Crass (ver. 0.3.8)
metaCRT (no ver. number available)
CD-HIT (ver. 4.6.7)
VirSorter (ver. 1.0.3)
VirFinder (ver 1.0.0)
PlasFlow (ver 1.0.7)
cBar (1.2)
snakemake (ver from 3.10.2 to 5.1.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The metagenomic and metatranscriptomics FASTQ files, rMAGs, and isolate genomes are available as NCBI BioProject PRJNA230567 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA230567). MP data has been deposited in the PRIDE database under the accession number PXD013655 (https://www.ebi.ac.uk/pride/archive/projects/PXD013655). Supplementary Files 1 (https://doi.org/10.5281/zenodo.3774024) and 2 (https://doi.org/10.5281/zenodo.3766442) are available via Zenodo.
Additional publicly available projects cited by this work include NCBI BioProject PRJNA174686 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA174686).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A generation-resolved, integrated meta-omic analysis of invasive mobile genetic elements and microbial host dynamics within a microbial community from a biological wastewater treatment plant spanning one and a half years.
Research sample	Individual floating sludge islets from the surface of the anoxic tank of the Schifflange biological wastewater treatment plant were sampled due to their richness in lipid accumulating organisms. They were then subjected to a concomitant biomolecular extraction of DNA, RNA an proteins, and a high throughput measurements to obtain metagenomic, metatranscriptomic and metaproteomic datasets to be computationally analysed.
Sampling strategy	Samples were collected from Schifflange biological wastewater treatment plant (Esch-sur-Alzette, Luxembourg; 49°30′48.29″N; 6°1′4.53″E). Individual floating sludge islets were collected from the same spot of the anoxic tank, along with physico-chemical parameters of the water, i.e. pH, temperature, conductivity, oxygen. Two initial samples were collected on 2010-10-04 and 2011-01-25 in the context of previously published work (https://doi.org/10.1038/ncomms6603 and https://doi.org/10.1038/npjbiofilms.2015.7). More frequent sampling was performed from 2011-03-21 to 2012-05-03, of which data from three samples (2011-10-05, 2011-10-05 and 2012-01-11) have been previously published (https://doi.org/10.1038/ncomms6603). A total of 53 samples were collected over a period of 578 days. The mean sample frequency was 8 days (SD=16 days). The sampling procedure was designed to span at least one entire annual seasonal cycle (i.e. winter, spring, summer, autumn) while the sampling frequency corresponded to the doubling time of the dominant bacterial population of approximately 8 days, thus representing an approximate generational time scale. Sampling was performed by Laura A. Lebrun and Emilie E.L. Muller. This work represents part of a larger ongoing multi-annual project. Thus, all the samples were subjected to the same experimental protocols. Please refer to detailed methods on sampling procedures in previous publications: https://doi.org/10.1038/ncomms6603 https://doi.org/10.1038/npjbiofilms.2015.7
Data collection	Laura A. Lebrun and Emilie E.L. Muller performed the concomitant biomolecular extractions resulting in fractions of DNA, RNA, proteins and metabolites for each in situ sample. They also performed the bacterial strain isolation (re-plating), screening and genomic DNA extraction for lipid accumulating bacteria. Nathan D. Hicks, Cindy M. Liu, Lance B. Price, John D. Gillece, James M. Schupp and Paul S. Keim performed the DNA and RNA library preparation and next-generation sequencing (NGS) to obtain MG and MT data. They also performed the DNA library preparation and NGS of isolate genomic data.

Michael R. Hoopmann and Robert L. Moritz performed the mass-spectrometry measurements of the protein fractions.
This work represents part of a larger ongoing multi-annual project. For detailed information and descriptions about data collection, experimental protocols, experimental kit versions, DNA and RNA library preparation, proteomic sample preparation, high-throughput platforms, please refer to the following articles:
https://doi.org/10.1038/ismej.2012.72
https://doi.org/10.1038/ncomms6603
https://doi.org/10.1038/npjbiofilms.2015.7
https://doi.org/10.1186/s40793-017-0274-y

Timing and spatial scale

Individual floating sludge islets within anoxic tank number one of the Schifflange BWWT plant (Esch-sur-Alzette, Luxembourg; 49° 30'48.29"N; 6°1'4.53"E) were sampled always on the same spot. Sampling was carried out from 2010-10-04 to 2012-05-03. Two samples were collected on 2010-10-04 and 2011-01-25, to determine the sequencing conditions and the microbial diversity and was published in previous work. Subsequently, samples were collected on a weekly basis from 2011-03-21 to 2012-05-03, which approximately corresponds to the generational time scale of the sludge of eight days. The lack of samples in periods; from 2011-07-08 to 2011-08-05, from 2011-10-12 to 2011-11-02, and from 2011-11-20 to 2012-12-21 are due to absence of foaming islets as consequence of (i) heavy or continued rain and/or (ii) natural decrease of foam during summer and autumn seasons.

Data exclusions

The first two samples, collected on 2010-10-04 and 2011-01-25, were excluded from the all analyses after the "population abundance estimation" (in the "Binning, selection of representative genomic bins, taxonomy and estimation of abundance" section) because the sampling occurred before the period of weekly sample collection (i.e. 2011-03-21 to 2012-05-03) and therefore did not fit within the generational time-scale.

Reproducibility

Experimental procedures adhered to previously published protocols. Open source software was used in all the computational analyses. All custom scripts and commands are available within multiple Gitlab repositories. Wherever applicable, the software versions are reported in "Methods and Material" within the manuscript.

Randomization

Samples collected from 2011-03-21 to 2012-05-03 were randomized before biomolecular extractions. The biomolecular fractions were further randomized prior to the high-throughput measurements.
The two initial samples, collected on 2010-10-04 and 2011-01-25, were not included within the aforementioned randomization procedure(s) as they were collected in the context of previous work (https://doi.org/10.1038/ncomms6603 and https://doi.org/10.1038/npjbiofilms.2015.7) and were used to pilot the experimental protocols which was conducted prior to the higher frequency sampling (i.e. from 2011-03-21 to 2012-05-03).

Blinding

Blinding is not applicable in this study as it did not involve human subjects, but rather data from in situ samples from a naturally occurring environment.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions

Anoxic tank of an activated sludge (biological) wastewater treatment facility under seasonal climatic conditions (i.e. spring, summer, autumn and winter).

Location

Schifflange biological wastewater treatment plant (Esch-sur-Alzette, Luxembourg; 49°30′48.29″N; 6°1′4.53″E).

Access and import/export

Access was granted to the research personnel based on agreement between the principal investigator, Prof. Paul Wilmes (on behalf of the research institution), and the wastewater treatment facility management (Mr. Bissen and Mr. Di Pentima) from the Syndicat Intercommunal a Vocation Ecologique (SIVEC), Schifflange, Luxembourg. All research personnel are informally introduced to the management and personnel of the facility prior to conducting any work. Research personnel were not provided with keys or electronic access cards, and thus could only enter the premises upon the permission of personnel at the entrance of the facility.

Disturbance

Sampling had a minimum-to-no impact on the operations of the wastewater treatment facility. The work of the researchers did not require (complete or partial) shutdown or any operational disruption of the facility. Sampling was performed by the research personnel (Emilie E.L. Muller and Laura A. Lebrun) without any involvement of the staff of the facility. Research personnel either brought their own equipment or used equipment from the site, which was dedicated to them, thus not hindering any operations or personnel within facility. Researchers could access operational readings (e.g. temperature, inflow, outflow, etc.) of the facility directly via a dedicated web portal of the facility using login credentials provided by the facility management. Two formal meetings weres organized between researchers and management of the facility over the past five years.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

A.3 Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance

Malte Herold, **Susana Martínez Arba**s, Shaman Narayanasamy, Abdul R. Sheik, Luise A. K. Kleine-Borgmann, Laura A. Lebrun, Benoît J. Kunath, Hugo Roume, Irina Bessarab, Rohan B. H. Williams, John D. Gillece, James M. Schupp, Paul S. Keim , Christian Jäger, Michael R. Hoopmann, Robert L. Moritz, Yuzhen Ye, Sujun Li, Haixu Tang, Anna Heintz-Buschart, Patrick May, Emilie E. L. Muller, Cedric C. Laczny and Paul Wilmes

2020

Nature Communications **11**:5281

DOI: https://doi.org/10.1038/s41467-020-19006-2

Contributions of author include:

- Data analysis
- Revision of manuscript

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19006-2 OPEN

Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance

Malte Herold[1,2], Susana Martínez Arbas[1], Shaman Narayanasamy[1,3], Abdul R. Sheik[1], Luise A. K. Kleine-Borgmann[1], Laura A. Lebrun[1], Benoît J. Kunath[1], Hugo Roume[1,4], Irina Bessarab[5], Rohan B. H. Williams[5], John D. Gillece[6], James M. Schupp[6], Paul S. Keim[6], Christian Jäger[1], Michael R. Hoopmann[7], Robert L. Moritz[7], Yuzhen Ye[8], Sujun Li[8], Haixu Tang[8], Anna Heintz-Buschart[1,9,10], Patrick May[1], Emilie E. L. Muller[1,11], Cedric C. Laczny[1] & Paul Wilmes[1,12]✉

The development of reliable, mixed-culture biotechnological processes hinges on understanding how microbial ecosystems respond to disturbances. Here we reveal extensive phenotypic plasticity and niche complementarity in oleaginous microbial populations from a biological wastewater treatment plant. We perform meta-omics analyses (metagenomics, metatranscriptomics, metaproteomics and metabolomics) on in situ samples over 14 months at weekly intervals. Based on 1,364 de novo metagenome-assembled genomes, we uncover four distinct fundamental niche types. Throughout the time-series, we observe a major, transient shift in community structure, coinciding with substrate availability changes. Functional omics data reveals extensive variation in gene expression and substrate usage amongst community members. Ex situ bioreactor experiments confirm that responses occur within five hours of a pulse disturbance, demonstrating rapid adaptation by specific populations. Our results show that community resistance and resilience are a function of phenotypic plasticity and niche complementarity, and set the foundation for future ecological engineering efforts.

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, 4362 Esch-sur-Alzette, Luxembourg, Luxembourg. [2] Epidemiology and Microbial Genomics, Laboratoire National de Santé, 1 rue Louis Rech, 3555 Dudelange, Luxembourg. [3] Megeno S.A., 6A Avenue des Hauts-Fourneaux, 4362 Esch-sur-Alzette, Luxembourg. [4] MetaGenoPolis, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Université Paris-Saclay, Domaine de Vilvert, Bâtiment 325, 78350 Jouy-en-Josas, France. [5] Singapore Centre for Environmental Life Sciences Engineering, 60 Nanyang Dr, Singapore 637551, Singapore. [6] The Translational Genomics Research Institute, 3051 West Shamrell Boulevard, Flagstaff, AZ 86001, USA. [7] Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA. [8] School of Informatics, Computing and Engineering, Indiana University, 700 N. Woodlawn Avenue, Bloomington, IN 47405, USA. [9] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstr. 4, 04103 Leipzig, Germany. [10] Helmholtz Centre for Environmental Research GmbH – UFZ, Theodor-Lieser-Str. 4, 06120 Halle, Germany. [11] Equipe Adaptations et Interactions Microbiennes, UMR 7156 UNISTRA-CNRS, Université de Strasbourg, Strasbourg, France. [12] Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg. ✉email: paul.wilmes@uni.lu

Mixed-culture biotechnological processes are essential for humankind to achieve its sustainable development goals[1,2]. However, in order to engineer reliable processes, fundamental insights into microbial niche ecology are necessary. Biological wastewater treatment plants (BWWTPs) represent a ubiquitous biotechnological application and occupy a central position in sustainable resource management plans[3,4]. Oleaginous bacterial populations are commonly found as the main constituents of floating sludge in BWWTPs and include divergent taxa such as *Candidatus* Microthrix parvicella or *Acinetobacter* spp.[5]. Storage lipids, such as triacylglycerols (TAGs), wax esters (WEs), and polyhydroxyalkanoates (PHA), derived from the lipid-rich biomass can directly be transesterified to fatty acid alkyl esters (biodiesel)[5], whereby PHA also represents a suitable precursor for bioplastics[6]. In general terms, substrate storage provides microbial populations with a competitive advantage under rapidly fluctuating and oftentimes sparse substrate conditions[7,8]. Even though BWWTP operation is a controlled process, factors such as aeration cycles, seasonal changes in temperature, and composition of inflow wastewater fluctuate. These factors have a profound impact on population dynamics[9] as well as linked process efficiency[10]. For example, periods of inefficient operation have been linked to competition between polyphosphate and glycogen accumulating organisms[11]. However, for wastewater-borne lipid-accumulating populations, which have compelling potential to be used in circular economic models[3], community shifts have been observed[12–14] with yet unclear links to niche ecology in situ.

Integrated meta-omics approaches hold the potential to resolve the fundamental niches and realized niches of microbial populations in situ[15]. The former represents the exhaustive inventory of resource ranges and conditions sustaining viability in the absence of environmental stress, competition, or predation, while the latter represents the part of a fundamental niche that is actually utilized by a population in the presence of other species and in a particular environment. The reconstruction of the fundamental niches is possible by linking functional potential to metagenome-assembled genomes (MAGs)[16] obtained through metagenomic (MG) sequencing. Functional omics data, such as metatranscriptomics (MT) or metaproteomics (MP), allow the resolution of realized niches[16]. Meta-omics approaches have previously been used for comparative functional screening in different environments and to characterize microbial activity, e.g., by using MT/MG ratios[17,18]. In human gut-borne microbial communities, niche partitioning has been inferred based on transcriptional profiles[19]. Furthermore, the coupling of MT and MP to meta-metabolomic (MM) data allows the differentiation between niches of genetically closely related populations[20]. Resolving the functions of coexisting microbial populations is of particular interest in the context of the extensive functional redundancy within microbial ecosystems[21,22]. Based on their emergent properties[23], microbial communities are characterized by composite metabolic capabilities and increased robustness compared to individual strains[24,25]. Steering these complex systems towards a desired endpoint, e.g., increased lipid accumulation, requires in-depth understanding of niche space and stability.

Here, we study whether community resistance and resilience are a function of phenotypic plasticity and niche complementarity. We develop and apply a novel framework for the in situ characterization of fundamental and realized niches of individual populations providing an in-depth understanding of ecological processes within a microbial community. We delineate ecological niches by integrating longitudinal meta-omics data (MG, MT, MP, and MM) and study complementarity of the realized niches. The addition of functional omics data (MT, MP, and MM) enables the resolution of metabolic plasticity and we thereby reveal how microbial ecosystems respond to disturbance. Using ex situ experiments to simulate pulse disturbances, we assess the response of individual oleaginous populations to oleic acid addition under shifting dissolved oxygen concentrations. Our dataset and methods represent important resources for the emerging field of integrating meta-omics data to study mixed microbial communities. Our results contribute to applications beyond wastewater treatment such as informed ecological engineering or research on host-associated microbiota.

Results

A time-resolved meta-omics dataset. To characterize the niche space of lipid-accumulating populations as well as resistance and resilience of the microbial community, we sampled a municipal BWWTP weekly over a 14-months period (from 2011-03-21 to 2012-05-03). Additionally, two preliminary time-points outside of the time-series were included[13,26]. Samples were split into intracellular and extracellular fractions, followed by concomitant biomolecular extractions[27] and high-throughput measurements (Fig. 1). MG, MT, and MP data were obtained on the intracellular fractions and MM data was generated on both the intracellular and extracellular fractions.

After quality filtering, the per-sample averages of MG and MT reads were 5.3×10^7 ($\pm 7.7 \times 10^6$ s.d.) and 3.3×10^7 reads ($\pm 1.2 \times 10^7$ s.d.), respectively (Supplementary Data 1). We performed sample-specific genome assemblies (average of 4.1×10^5 contigs per sample) followed by binning[28] yielding a total of 1364 MAGs passing our quality filtering criteria (see "Methods" section). To track the abundance, gene expression, and activity of individual microbial populations over time, we dereplicated[29] the MAGs across samples to generate 220 representative MAGs (rMAGs). From these, we further selected those with the highest completeness resulting in 78 rMAGs (76.2% mean completeness, 2.2% mean contamination) (Supplementary Data 2). These genomes represent the major populations across the time-series, with an average mapping percentage of 26% ± 3% (s.d.) and 27% ± 3% (s.d.) of total MG reads and total MT reads per time-point, respectively, and are corroborated by a previous study based on 16S rRNA amplicon sequencing[13]. For the MP measurements, we obtained a per-sample average of $1.5 \times 10^5 \pm 8.2 \times 10^3$ (s.d.) MS2 spectra and a total of 7.6×10^6 MS2 spectra. Of 7.8×10^5 identified peptides, 3.3×10^5 (43%) could be matched to 2.1×10^5 predicted coding sequences of the 78 rMAGS. Per time-point, on average $1.5 \times 10^4 \pm 4.5 \times 10^3$ (s.d.) spectral matches, i.e., on average 94% of all rMAG-associated matches could be assigned to genes with predicted functions, i.e., assigned KEGG ortholog groups (KOs). To study the community-wide resource space and metabolite turnover, we measured metabolite levels by an untargeted approach using gas chromatography (GC) coupled with mass spectrometry (MS) (Supplementary Data 3). In total, 89% (58 of 65) of the identified metabolites could be linked to enzymes encoded by the rMAGs. We estimated resource uptake by calculating intracellular vs. extracellular metabolite ratios for 42 metabolites detected in both fractions (Supplementary Data 3). Additionally, six abiotic parameters were measured during sampling, as well as 34 parameters recorded continuously as part of the BWWTP online monitoring (Supplementary Data 4).

We also generated MG and MT data for the ex situ experiments. These simulated the fluctuating conditions within the BWWTP, namely the short-term response to pulse disturbances of oleic acid influx under shifting dissolved oxygen conditions. We sequenced DNA and RNA fractions obtained at 0, 5, and 8 h after addition of oleic acid, yielding on average 1.02×10^8 MG and 9.33×10^7 MT reads per sample. The increased sequencing depth compared to the in situ time-series was

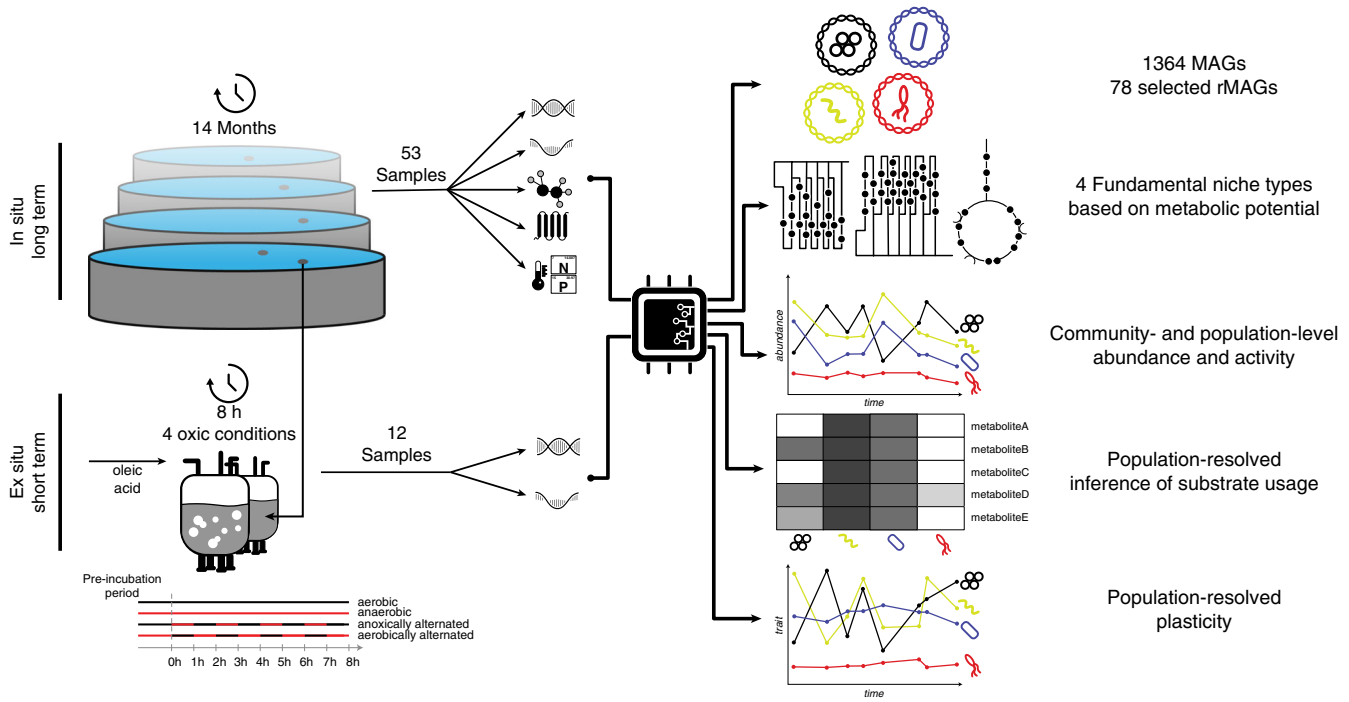


Fig. 1 Overview of the study design. Samples are derived from in situ sampling of an anoxic tank of a municipal biological wastewater treatment plant. Metagenomic (MG), metatranscriptomic (MT), metaproteomic (MP), and meta-metabolomic (MM) data is generated. Physicochemical data is also collected. Additionally, MG and MT data is generated for ex situ experiments using biomass from the same system and fed with oleic acid under different oxic conditions to evaluate short-term responses to pulse disturbance. The time-series meta-omics data is integrated to define metagenome-assembled genomes (MAGs) over all time points. Representative MAGs (rMAGs) across time are selected for further analysis. The rMAGs' functional potential is used to infer the fundamental niches. Abundance and activity data are derived from the functional omics and substrate usage is inferred per population. The variability of gene expression is used to assess the phenotypic plasticity of the individual populations.

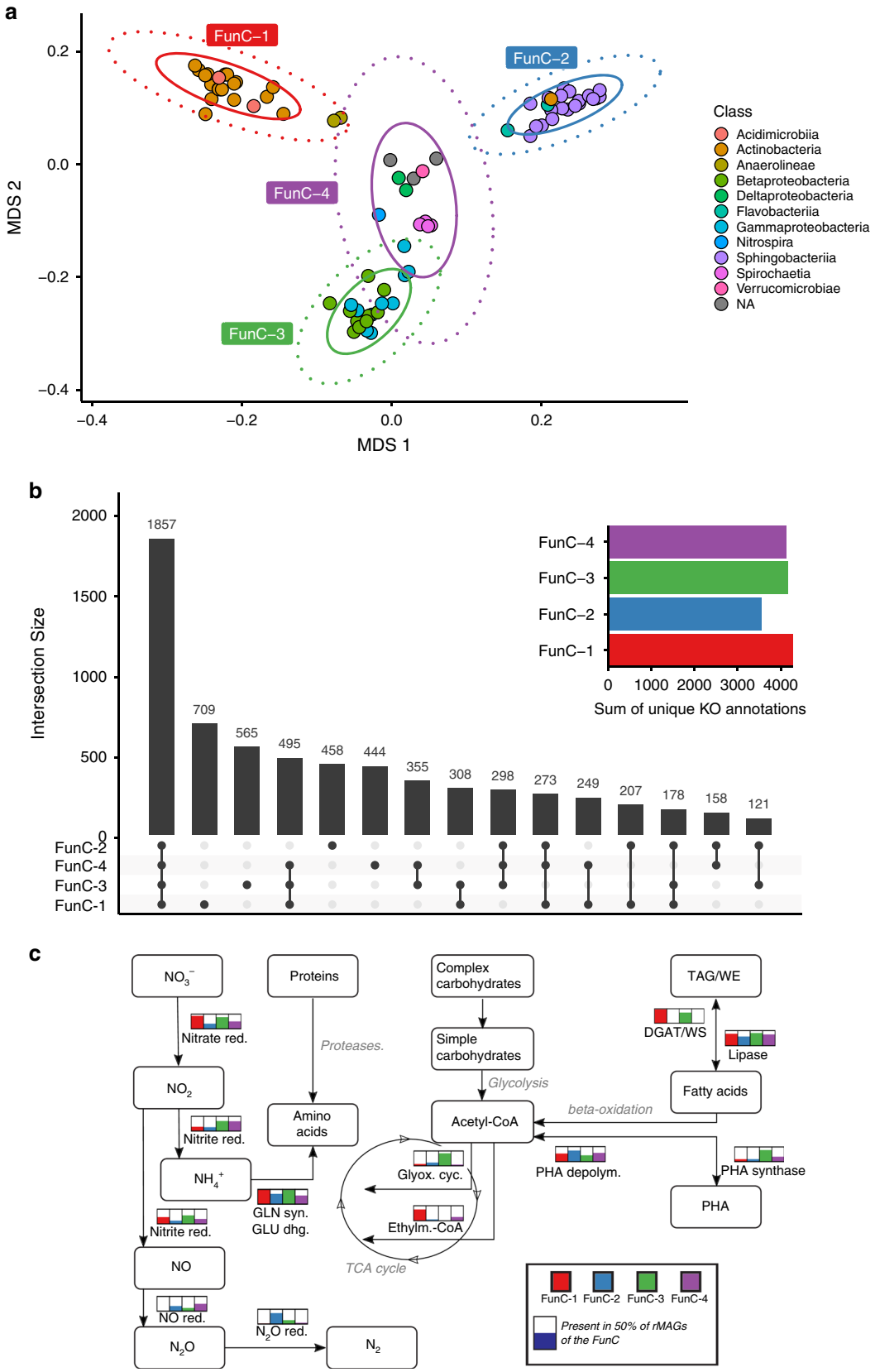
important to obtain a fine-grained view on short-term responses to oleic acid. Mapping of the sequencing reads to the selected set of rMAGs revealed mapping percentages comparable to the in situ time-series (mean: 21% ± s.d.: 1% for both MG reads and MT reads).

Overall, our meta-omics dataset comprehensively describes mixed microbial communities underlying lipid-accumulation processes in BWWTPs, and in particular their functional potential, composition, activity, as well as substrate availability and assimilation.

Distinct niche types. To resolve the fundamental niches of the pertinent bacterial populations through their functional genomic potential, we assigned KOs to the rMAGs' predicted coding sequences. We hypothesized that individual populations would form clusters based on the similarity/dissimilarity of their functional potential. We found four distinct clusters of rMAGs by projecting pairwise Jaccard distances of KO presence (Fig. 2a and Supplementary Fig. 1). These functional clusters (FunCs) represent differences of known, overall metabolic capabilities of the rMAGs and reflect their fundamental niches. FunC-1 consisted of Actinobacteria, and FunC-2 was primarily comprised of members of the Bacteroidetes phylum, mainly of the Sphingobacteriia class (Fig. 2a). FunC-3 contained Betaproteobacteria and Gammaproteobacteria whereas FunC-4 appeared more diverse, containing Spirochaetia as a subcluster, Deltaproteobacteria, and taxonomically unclassified rMAGs. We found mash-based genomic distance[30] to be strongly linked to FunC assignment (PROCRUSTES sum of squares: 0.399, correlation 0.775, PROTEST *p*-value 0.001, Supplementary Fig. 2a), highlighting that phylogeny is a strong determinant for FunC assignment. However, some distantly linked subgroups were defined by their shared functional complement, i.e., assigned to a different FunC than their neighbors in a corresponding phylogenetic tree (Supplementary Fig. 2b). This shows that KO profile similarity-based analyses provide important information in addition to phylogeny-based approaches[31].

A total of 1857 KOs was shared between all FunCs and we found that FunCs 1, 3, and 4 contained comparable numbers of nonredundant KOs with 4276, 4177, and 4129 KOs, respectively (Fig. 2b). FunC-2 exhibited a reduced number of KOs (3550), however it also represented the least taxonomically diverse FunC as it almost exclusively consisted of *Haliscomenobacter* spp. and *Chitinophaga* spp. (Supplementary Data 2). We tested for the molecular functions that were significantly enriched in individual FunCs and found, among others, functions related to lipid metabolism for FunC-1, amino sugar, and nucleotide sugar metabolism for FunC-2, and biofilm and secretion systems for FunC-3 to be enriched (Fig. 2c and Supplementary Data 5; one-sided Fisher's exact test, adjusted *p*-values < 0.05).

While lipid-accumulating organisms hold great potential for the recovery of high-value molecules[5], interactions between these organisms as well as the community at large are understudied in situ. We found that diacylglycerol O-acyltransferase (DGAT/WS), which is involved in lipid storage[32], was encoded in 23 out of 24 rMAGs of FunC-1, pointing to the importance of TAG accumulation in this cluster. Most FunC-3 members also encoded DGAT/WS (14 of 19). Moreover, PHA synthase was enriched in this cluster (15 of 19). All rMAGs encoded lipases, functions involved in fatty acid synthesis, or beta-oxidation. However, several acyl-CoA and acyl-ACP dehydrogenases were over-represented in FunC-1 and FunC-3. Additionally, acetyl-CoA acetyltransferases involved in the degradation and biosynthesis of



fatty acids were prevalent throughout all FunCs. The enrichment in FunC-1 and FunC-3 for genes involved in lipid accumulation are consistent with previous metabolic characterizations, with FunC-1 consisting mainly of Actinobacteria for which TAG accumulation has been described[33]. FunC-3 contains Betaproteobacteria and Gammaproteobacteria that have been characterized as TAG, WE, and/or PHA accumulators, e.g., *Thauera* spp., *Albidiferax* spp., or *Acinetobacter* spp.[33,34]. Importantly, we observed a difference between these FunCs in the utilization of acetyl-CoA. Specifically, FunC-1 members showed an enrichment in functions related to the ethylmalonyl-CoA pathway (crotonyl-CoA reductase and enoyl ACP

Fig. 2 Fundamental niche types. a Multidimensional scaling (MDS) of Jaccard distances for the functional repertoire (presence of KEGG ortholog groups [KOs]) for each rMAG. Ellipses containing 95% (inner) or 99% (outer) of cluster-assigned data points are shown resulting in four distinct functional clusters (FunCs). Colors indicate the class-level taxonomy of the rMAGs. **b** Numbers of shared and unique KO assignments between the FunCs. Colored bars show the total number of nonredundant KO assignments within the individual FunCs. Overlaps between different sets of FunCs and their unique KOs are represented by the central black bars with the points below defining the members of the respective sets. **c** Presence of key functions within the four FunCs. Bars next to metabolic conversions show the proportion of rMAGs encoding characteristic enzymes for the respective reaction or pathway adjusted for mean rMAG completeness. Pathways ubiquitously present across rMAGs are shown in gray color. Source data are provided as a Source Data file. red. reductase, GLN syn. glutamine synthetase, GLU dhg. glutmatate dehydrogenase, glyox. cyc. glyoxylate cycle, ethylm.-CoA ethylmalonyl-CoA pathway, PHA depolym. PHA depolymerase.

reductase), while FunC-3 members encoded key enzymes involved in the glyoxylate cycle (malate synthase and isocitrate lyase).

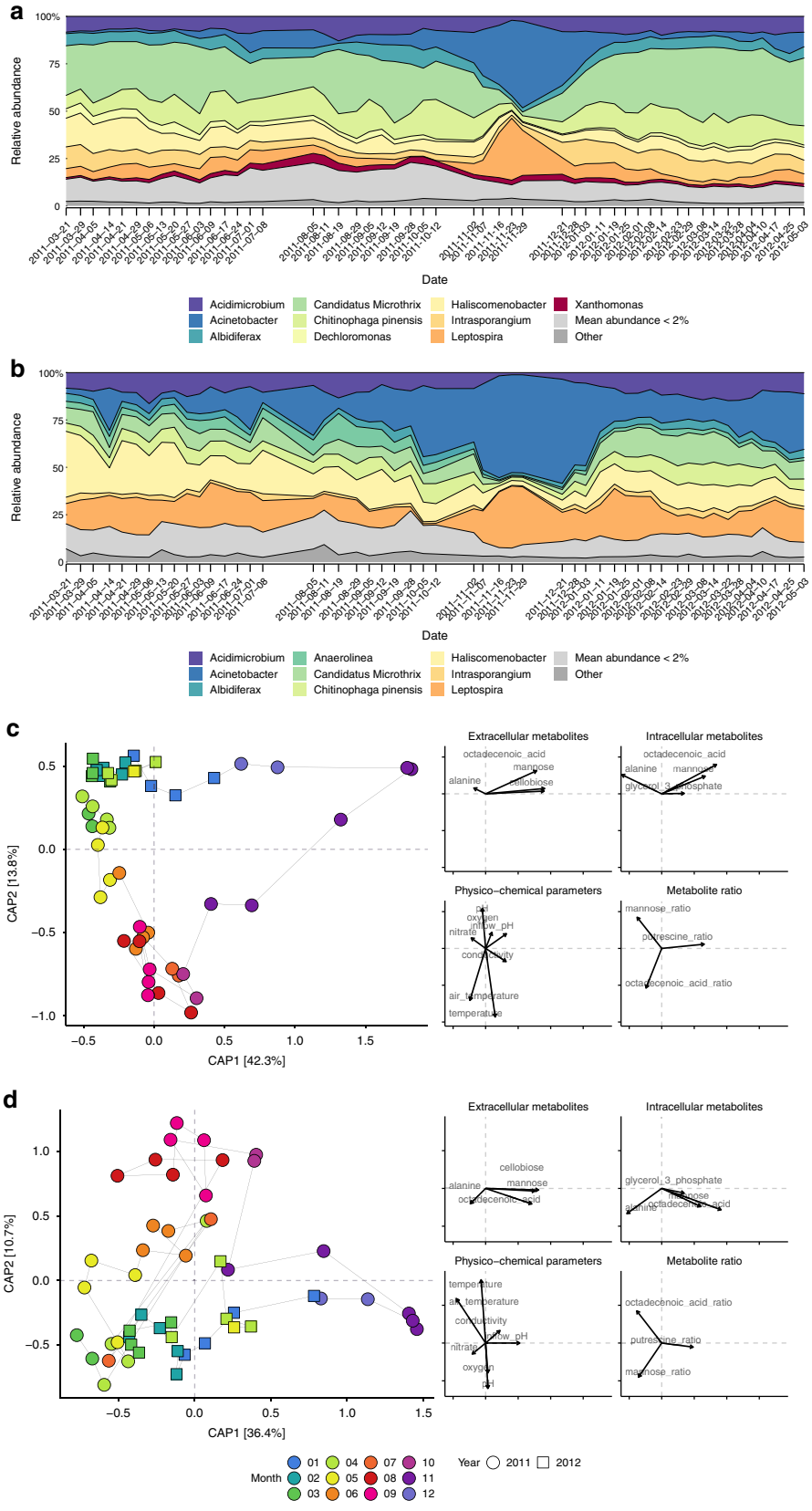
We further determined specific functional enrichment for the four FunCs in relation to the breakdown of other macromolecules (including CAZymes and proteases), nitrogen cycling, stress response, and motility (Supplementary Data 5). The discriminating functions point towards interdependencies between the different FunCs, e.g., in terms of denitrification (Fig. 2c). We found that the separation into taxonomically consistent groups is accompanied by specific conserved functions, e.g., strong enrichment in FunC-1 for WhiB transcriptional regulators characteristic of the Actinobacteria[35]. Overall, we observed a widely distributed set of core functions in foaming sludge microbiomes and identified groups of populations characterized by distinct functional potential in lipid metabolism, amino sugar, and nucleotide sugar metabolism as well as biofilm and secretion systems.

Community dynamics and stability. To understand whether population dynamics can be related to substrate availability and other abiotic factors[36], we used MG depth-of-coverage to infer rMAG population abundance across the time-series. We computed distances between the rMAGs' abundance profiles (based on their pairwise correlations) and found that the dynamics of rMAGs can be partially explained by the FunC assignment (PERMANOVA $R^2 = 0.12$, Pr > $F = 0.002$; no significant difference in dispersion; Supplementary Fig. 3), thereby linking FunC membership to temporal abundance shifts. The most abundant taxa (Supplementary Data 2) included *Candidatus* Microthrix (26.0% relative abundance across the time-series; referred to as *Microthrix* in the remaining text), *Acinetobacter* (8.1%), *Haliscomenobacter* (8.0%), *Intrasporangium* (7.2%), *Leptospira* (6.3%), *Albidiferax* (5.7%), and *Dechloromonas* (2.4%) (Fig. 3a). Several of the recovered rMAGs belonged to filamentous taxa according to the MiDAS field guide database for organisms in activated sludge[37], such as the highly abundant *Microthrix*, and *Haliscomenobacter*, as well as the less abundant *Anaerolinea* (1.1 %) and *Gordonia* (0.2 %).

Variations during the operation of BWWTPs occur largely due to changes in the influent wastewater composition and climatic conditions[38]. We observed gradual changes in the community structure with the seasons (Fig. 3a). In October 2011 (month seven of the timeseries), the community composition began to shift, with a markedly altered composition in late November 2011. This shift was characterized by spikes in the relative abundance of *Leptospira* (peak at 2011-11-23) and *Acinetobacter* (peak at 2011-11-29) (Fig. 3a), and co-occurred with a pronounced shift in substrates (Fig. 4 and Supplementary Fig. 4). The substrates included mainly nonpolar metabolites, including long-chain fatty acids (LCFAs) and glycerides, as well as polar metabolites mannose, glucose, disaccharides, ethanolamine, and putrescine. We found that the intersample distances of MG-based abundances could partially be explained by a subset of the abiotic factors (Fig. 3c). Summer samples were characterized by higher temperatures, phosphate levels and higher intracellular vs. extracellular oleic acid ratios. Higher extracellular mannose levels and a slight increase in conductivity marked the beginning of the autumn shift. During November, intracellular and extracellular levels for LCFAs increased, indicating a higher availability or turnover of LCFAs, but not necessarily an equivalent conversion to neutral storage lipids. In the subsequent winter time-points, substrate levels normalized and the community transitioned back to the predisturbance state.

The dominance of *Microthrix* was re-established within approximately ten generations, given estimates for in situ growth rates of 0.12–0.3 growth cycles per day[7,8]. The stability[39] of the individual rMAGs was heavily affected by the November shift (mean population stability: 1.43 ± 0.69 s.d.), compared to the stability when excluding the respective time-points (mean population stability: 2.39 ± 1.28 s.d.; 2011-11-02 to 2011-11-29; Supplementary Data 2). The observed population dynamics indicate that the community composition is resilient, i.e., recovers after pronounced changes in available substrates, and resistant to small-scale environment fluctuations over time.

While MG depth was used as a proxy for population abundance, MT depth enabled the analysis of transcriptional activity within the community and of individual populations (Fig. 3b). The comparison of intersample distances based on mean, relative MT depth showed similar patterns to MG-based results (Fig. 3c), albeit with a higher degree of variability indicated by increased inter-sample distances (Fig. 3d). A comparison of relative MP counts showed a more even distribution between populations with comparable overall trends (Supplementary Fig. 5). Samples collected in April 2011 and 2012 appeared to represent transition states between seasons. Additionally, a set of late winter and early spring samples in 2011 and 2012 showed higher similarities at the expression level than at the abundance level. Interestingly, the high abundance of individual genera, such as *Microthrix* or *Chitinophaga* was not necessarily reflected in their mean expression levels (Fig. 3b and Supplementary Fig. 5): populations assigned to *Leptospira, Haliscomenobacter, Anaerolinea*, and *Acinetobacter* showed higher mean expression overall. Spikes in relative MT depth as for *Acinetobacter* rMAGs (Fig. 3b; 2011-04-14, 2011-05-08, and 2012-04-25) point towards increased activity around these time-points, which however did not lead to major shifts in community structure. Notably, higher expression levels of *Acinetobacter* were succeeded by increased expression levels of *Haliscomenobacter* (2011-04-14 to 2011-05-20) or *Anaerolinea* (2011-05-08 to 2011-09-19). On average, MT-based stability values were less affected by the community shift than MG-based stability values (Supplementary Table 2). We also observed adaptation of metabolic pathway activity to environmental conditions (Fig. 5). Pentose to EMP pathway intermediates exhibited the highest correlation between MT and MP abundances, followed by Hydrogen metabolism and Fatty acid oxidation. Several pathways exhibited a characteristic drop during the November shift, e.g., hydrogen metabolism, hydrocarbon



degradation, and TCA cycle, while fatty acid oxidation showed a marked peak. This highlights the transition from dominance by generalist, lipid-accumulating populations towards a lipolytic community.

With each of the four FunCs comprising multiple organisms encoding similar KOs and, hence, metabolic capabilities, we studied how individual populations adapt to their environment. To this end, we linked changes in community structure and in the expression levels of individual populations to the influence of environmental parameters. While rMAG abundance patterns could be linked to FunC assignment (Supplementary Fig. 3), we could not identify an analogous categorization when correlating

Fig. 3 Community structure and function dynamics. a, **b** Relative abundance and expression levels of recovered populations represented by rMAGs over time based on MG depth (**a**) and MT depth (**b**) of coverage, respectively, representing mapping percentages of MG [26% ± 3% (s.d.)] and MT [27% ± 3% (s.d.)]. The relative abundance of individual rMAGs is grouped based on genus-level taxonomic assignment with rMAGs of unresolved taxonomy grouped in "Other". Recovered genera with mean abundance below 2% are summarized as a single group (light gray). **c**, **d** Ordination of Bray–Curtis dissimilarity of relative abundances, MG (**c**) and MT (**d**), of individual rMAGs constrained by selected abiotic factors (metabolite levels, metabolite-ratios, and physico-chemical parameters shown as black arrows with arrow lengths indicating environmental scores as predictors for each factor). Points are colored by month of sampling and point-shape reflects the year of sampling. Thin black lines connecting the points visualize the time course of sampling. Source data are provided as a Source Data file.

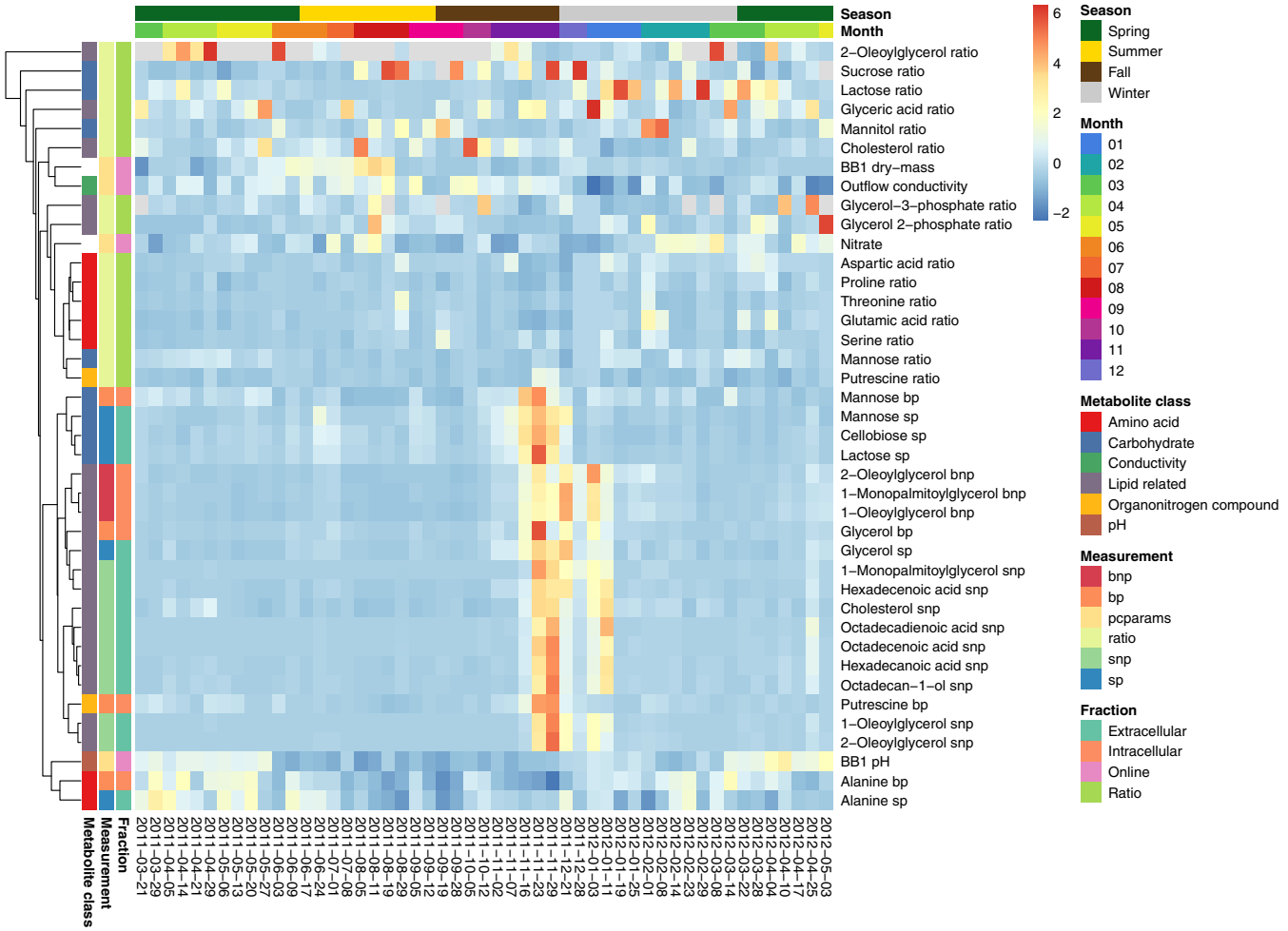


Fig. 4 Levels of metabolites and physico-chemical parameters. *Z*-score transformed metabolite intensities, metabolite ratios, and physico-chemical parameter levels over time are shown. Row annotations highlight classes of metabolites and parameters, measurement types (bnp: intracellular nonpolar metabolites, bp: intracellular polar m., pcparams: physico-chemical parameters, ratio: metabolite intrac./extrac. ratio, snp: extracellular nonpolar m., sp: extracellular polar m.), and the subtype or fraction of the measurement (manual: measured during sampling, online: measured during WWTP operation). Selected rows are shown (comprehensive heatmap shown in Supplementary Fig. 6). Source data are provided as a Source Data file.

rMAG abundances to abiotic factors. Instead, correlation patterns indicating similar preferences to environmental conditions emerged for subgroups of rMAGs across different FunCs (Supplementary Fig. 7). This shows that populations with a similar fundamental niche type responded differently to the environmental conditions pointing towards functional plasticity and, thus, adaptations of their realized niches

Niche characteristics of in situ and ex situ time-series. While we identified four fundamental niche types, it may be assumed that cohabiting species cannot occupy the same realized niches, leading to realized niche segregation within and between types. We hypothesized that different degrees of niche overlap, leading to variable levels of competition, must exist[40,41]. To better understand the complementarity of realized niches, we used the functional omics data to study how rMAGs overlapped in relation to their encoded genes and how rMAGs varied in their expression profiles. While the former represents competition between populations with overlapping profiles, the latter is an important factor for the adaptability and overall survival strategy of individual populations. We distinguished between expressed KOs and nonexpressed KOs based on MT/MG ratios as well as MP data and computed distances between the resulting time-point-specific expression profiles. While the separation based on the functional potential was preserved in a clustering of expression profiles (in particular for FunC-2), the expression profiles of FunCs-1, FunCs-3, and FunCs-4 overlapped to a greater extent than those of FunC-2 (Fig. 6a). Two *Anaerolinea* populations assigned to

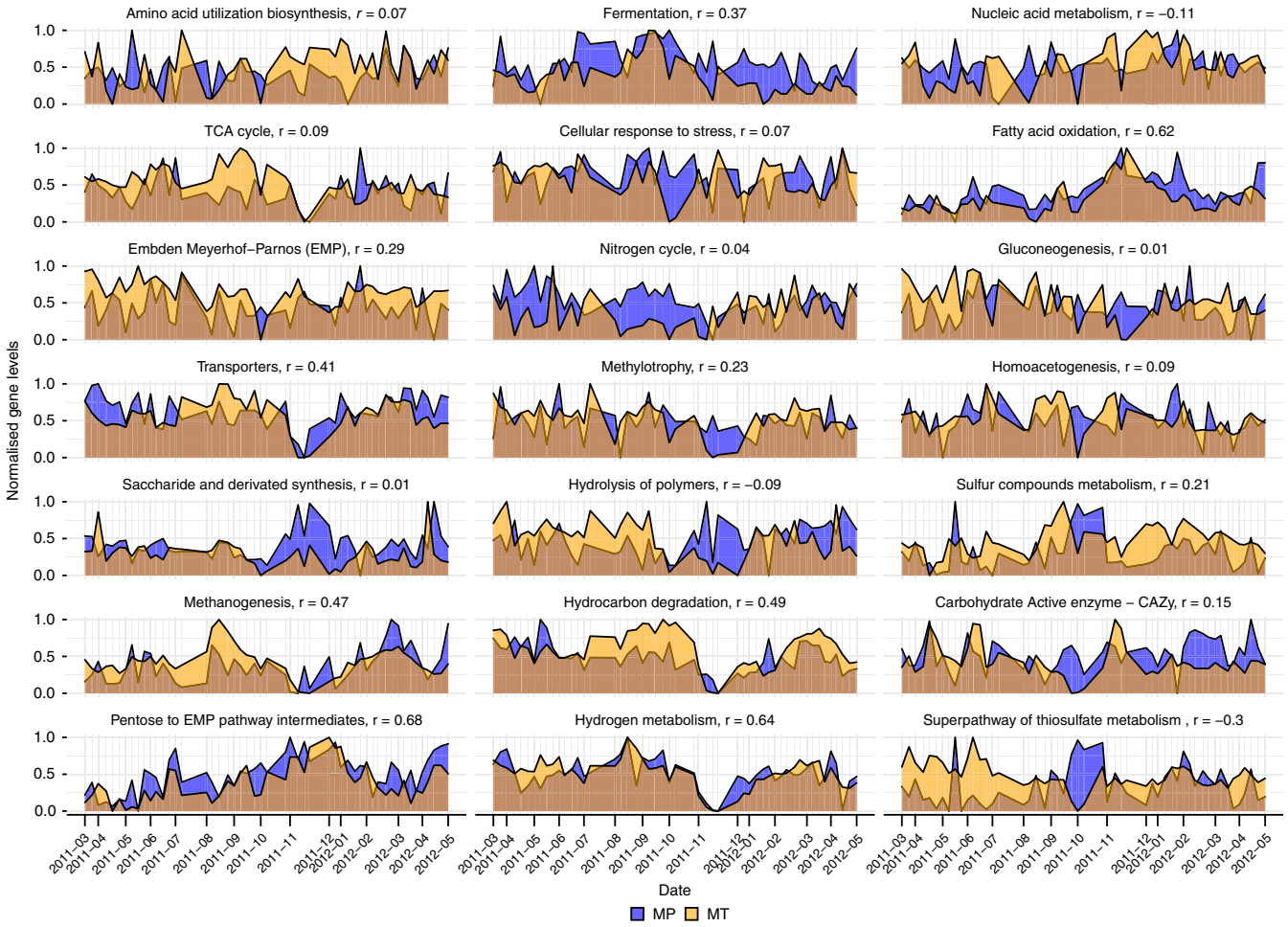


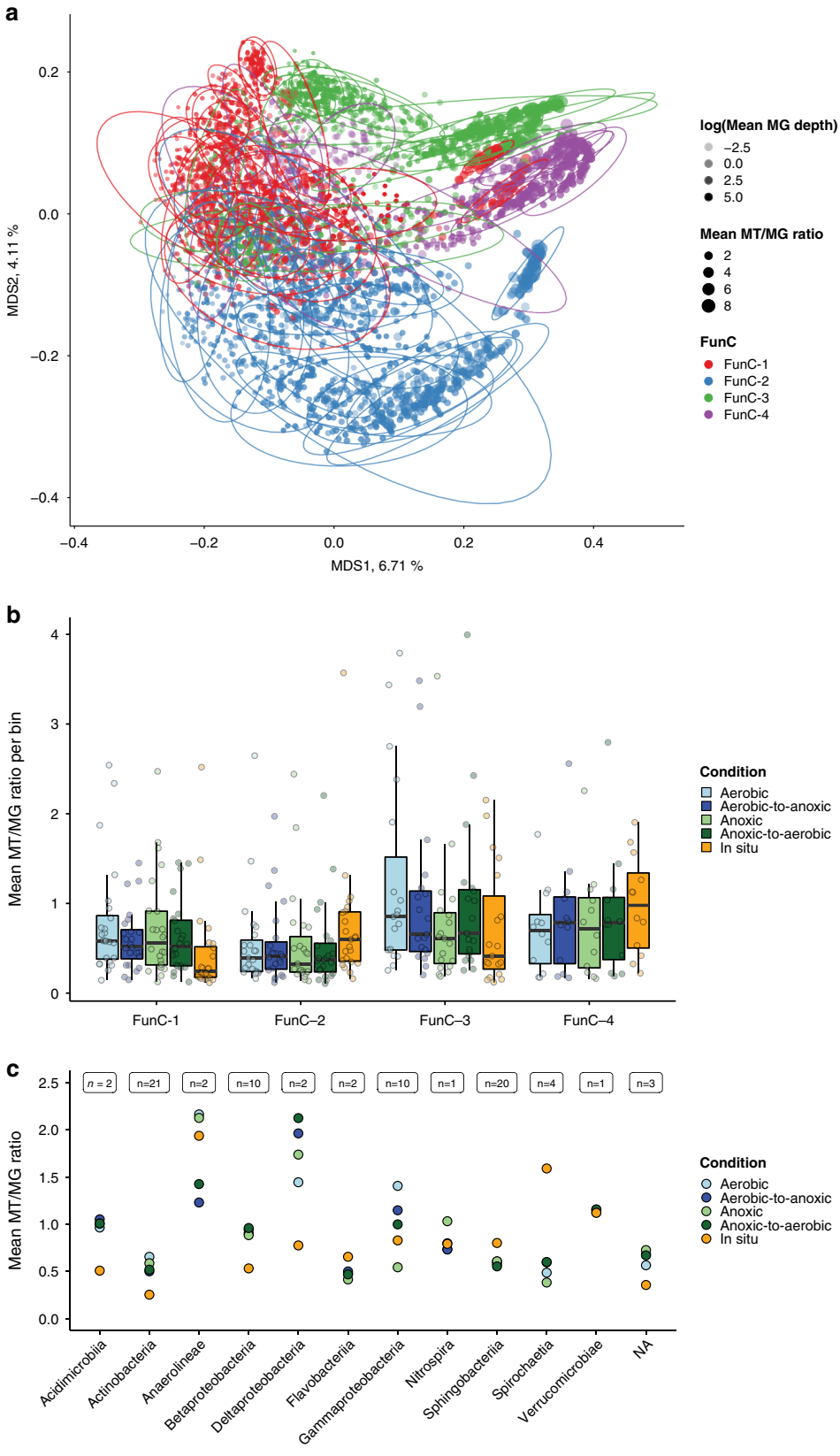
Fig. 5 Gene levels over time grouped by functional categories. Metatranscriptomic and metaproteomic levels (normalized relative expression for MT data and normalized relative spectral counts for MP data) of rMAG-derived genes assigned to FOAM ontology-based functional categories. Pearson correlation coefficients (*r*) of MT and MP values are shown in the title of each panel. Panels are ordered from highest to lowest mean MP relative count in row-major order. Source data are provided as a Source Data file.

FunC-1 appeared to express similar functions compared to the rMAGs of FunC-3 and FunC-4 and were found in a subgroup of rMAGs that showed a higher overall activity in terms of MT/MG ratios also when clustering expression profiles per time-point (Supplementary Fig. 8). Overall, the clusters based on KO expression status per time-point did not exhibit a separation according to the grouping into FunCs (Supplementary Fig. 8). This indicates a propensity of the respective rMAGs to more frequently express shared KOs than discriminatory KOs and, consequently, increased the competition for specific substrates.

To investigate the importance of individual, discriminatory functions, we selected rMAG clusters, based on gene expression and MP counts, to which the two most abundant rMAGs (D51_G1.1.2, A01_O1.2.4) had been assigned. We observed that clusters into which rMAG D51_G1.1.2 (*Microthrix*) was consistently categorized showed expression of few KOs with the majority being ribosomal proteins, TCA cycle-related enzymes such as pyruvate, malate, and glyceraldehyde 3-phosphate dehydrogenases, chaperones, and most frequently the WhiB family transcriptional regulator (19 time-points; Supplementary Data 6).

Clusters containing rMAG A01_O1.2.4 (*Acinetobacter*) frequently exhibited expression of genes related to motility and chemotaxis as well as stress response, but also functions related to phosphate accumulation, such as K08311 and K00937 (Supplementary Data 6). KOs related to lipid metabolism were also frequently expressed in these clusters e.g. acylglycerol lipase (in 35 time-points) or diacylglycerol O-acyltransferase (25 time-points). This indicates that high expression of key functionalities is an integral part of the strategies of the populations within these clusters even though they differed with respect to their encoded functions.

We next studied how the observed distinction between populations with high activity is linked to phenotypic plasticity. As alternating oxygen levels in BWWTPs play an important role in selecting for lipid accumulating populations[7,42], we added oleic acid, the preferred carbon source for *Microthrix*[43], in lab-scale experiments under different oxygen fluctuation conditions[8] (see "Methods" section; Fig. 1). These ex situ conditions involved aerobic, anoxic, aerobically preconditioned biomass followed by hourly anoxic alternations, and anoxically preconditioned followed by hourly aerobic alternations. The MT/MG ratios for FunC-1 and FunC-3 were higher ex situ when compared to the in situ samples, and vice versa for FunC-2 and FunC-4 rMAGs (Fig. 6b). Furthermore, especially for FunC-3, average MT/MG ratios were highest in the aerobic conditions and lowest in the anaerobic conditions. This is in line with FunC-3 being comprised mainly of Betaproteobacteria and Gammaproteobacteria, which include mostly aerobic genera[44]. A more fine-grained view on differences in specific activity was obtained, when grouping rMAGs based on taxonomic assignment (Fig. 6c). While rMAGs of the classes Acidimicrobia and Actinobacteria (FunC-1)



showed the lowest mean MT/MG ratios across the in situ time-series (0.5), the ratio was twice as high in the ex situ experiments across all conditions which can be attributed to the oleic acid pulse. Betaproteobacteria (FunC-3) behaved similarly, while Gammaproteobacteria (FunC-3) showed a tendency towards higher activity with increased oxygen levels. We observed high activity for rMAGs assigned to Anaerolineae and Spirochaetia in the in situ time-series. Interestingly, this was not the case for Spirochaetia in the ex situ experiments, which points towards the necessity for additional substrates. The Anaerolineae rMAGs, with taxonomically related species being mainly anaerobic[45], showed the lowest MT/MG ratio under the

Fig. 6 Realized niches. a MDS of time-point specific expression profiles based on MT/MG ratios or evidence at the MP level. Colors indicate FunC assignment of the individual rMAGs. Point shape represents cluster assignment based on automated clustering of the embedded points. Ellipses containing 95% of cluster-assigned data points are shown. Points size represents the average MT/MG depth ratios of the individual rMAGs. The amounts of variance explained by the first two dimensions are shown on the respective axes. **b** Mean MT/MG depth ratios over all time-points are shown per condition for 78 rMAGs (boxplots show: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; Each group of boxplots corresponds to a group of rMAGs (FunC-1 $n = 24$, FunC-2 $n = 23$, FunC-3 $n = 19$, FunC-4 $n = 12$), each boxplot represents an independent experiment.). **c** Mean MT/MG depth ratios grouped according to class-level taxonomic assignment of the rMAGs with the number of rMAGs for each group are shown in the top of the plot (*n*). Source data are provided as a Source Data file.

alternating conditions, while Deltaproteobacteria rMAGs showed high MT/MG ratios. Overall, the differentiated responses under alternating conditions point to distinct short-term and long-term adaptation strategies.

To study how fast the adaptations in response to the influx of oleic acid occur, we compared the baseline (0 h time-points, before oleic acid addition) against the 5 and 8 h time-points (after oleic acid addition). At 5 h, lipases, involved in TAG hydrolysis, for which high expression in the in situ samples was observed, were downregulated in the ex situ response to the addition of oleic acid (Supplementary Fig. 9a). An increased number of genes related to beta-oxidation were upregulated at 5 h, particularly in rMAGs assigned to FunC-3 (Supplementary Fig. 9b). Similar effects were observed when comparing the 0 h and 8 h time-points (Supplementary Fig. 10a, b). This suggests that responses in gene expression happen within the 5 h timeframe but on distinct time scales for different populations. In-depth analyses of the populations exhibiting the highest expression levels for TAG lipases, DGAT/WS, and PHB synthases (Supplementary Note 1 and Supplementary Figs. 11–14) underline the previously determined role of *Microthrix* as a key lipid accumulator in BWWTPs[13,46]. The results also indicate that populations such as *Anaerolinea*, *Leptospira* and *Acinetobacter* overlap with *Microthrix* in terms of their capacity to assimilate LCFAs and available neutral lipids. Niche complementarity and plasticity, i.e., overlapping fundamental and realized niches, as well as gene expression variability, impart population-independent processing of lipids in situ. From an ecosystem perspective, this community-wide trait confers functional resistance and resilience.

Discussion

The ability to reconstruct population-level genomes and infer their functional potential from metagenomes allows identification of the fundamental niches of distinct community members. Unprecedented views of realized niches are achieved by tracking functional gene expression via MT and MP analyses, as well as actual resource usage resolved via comparative metabolomics analyses of intracellular and extracellular metabolites. The joint resolution of fundamental and realized niche breadths of individual populations is key to understanding the ecological processes within microbial communities, including, but not limited to, how such consortia respond to disturbance.

Here, through the application of our novel framework for the integration of multi-meta-omics datasets, we were able to track community-wide and population-resolved traits longitudinally in situ as well as ex situ. We found four distinct fundamental niche types in this ecosystem. Populations assigned to a specific type shared common functional repertoires and largely shared a similar phylogenetic background, in line with previously observed metabolic repertoires[47,48]. Simultaneously, some functions, e.g., related to lipid accumulation, were found to be enriched in multiple niche types.

Despite our results showing a link between functional complement, realized niches, and phylogeny, we also observed distinct activities in response to the changing environmental conditions within individual niche types, e.g., some lowly abundant populations exhibited high activity. This suggests distinct adaptation strategies to variabilities in the resource space and is exemplified by the populations in the functional cluster that includes the dominant *Microthrix* population. *Microthrix* follows a strategy based on phenotypic heterogeneity for rapid adaptation to the prevailing environmental conditions[13]. Our ex situ validation experiments revealed the adaptations to changes in substrate availability and dissolved oxygen concentrations after as little as 5 h post-disturbance. This plasticity in gene expression allows the populations to be resistant to fluctuations in environmental conditions. Furthermore, this strategy was found to be unique to *Microthrix* as evidenced from the increased transcriptional response of other lipid-accumulating and/or lypolytic populations, e.g., *Acinetobacter*, *Leptospira*, or *Anaerolinea* spp., especially in the aerobic ex situ conditions.

Our work highlights the requirement to account for organism-specific adaptation strategies and time-frames within mixed communities. We observed that drastically altered community composition and gene expression patterns followed a severe disturbance in substrate levels within our time-series. We hypothesize that this community shift was a consequence of excess substrate availability, and it highlights a limit to the community's resistance. Individual populations recovered within ten sludge age cycles post-disturbance, which indicates that the resilience of the community is also linked to phenotypic plasticity. The overlap in realized niches reflects niche complementarity. This in turn is governed by interspecific competition over a set of substrates, such as oleic acid. Other independent work on the human gut microbiome has highlighted the importance of interspecific competition for the maintenance of stability under a constant feeding regimen[49]. How interspecific competition or lack thereof relates to resilience represents a key question for future work.

Overall, our framework demonstrates that multi-meta-omics data allows an in-depth characterization of ecological niches over time. Due to the observed plasticity in activity and the recovery after a major, transient perturbation, we confirm that the relationship between resistance and resilience is a function of fine-scale competition over resources in this environment. The resulting complementarity in both the fundamental and realized niches guarantees the provision of stable ecosystem services[50] and, thus, the long-term stable operation of mixed-culture biotechnological processes. These results are particularly relevant for the future engineering of niches within mixed-culture biotechnological processes[3], which are key to achieve humankind's sustainability goals[1,2]. In more general terms, it will be important to understand if phenotypic heterogeneity and niche complementarity play similarly important roles in the stability of other microbiomes.

Methods

Sampling and biomolecular extractions. Oleaginous biomass comprised of floating sludge islets was sampled from the surface of an anoxic tank at the Schifflange municipal biological wastewater treatment plant (BWWTP; Schifflange, Luxembourg; 49°30′48.29″N; 6°1′4.53″E)[27]. In situ sampling intervals of approximately one week were chosen to match the sludge age (the average time the

biomass remains in the entire system) as well as the average doubling time of the dominant *Microthrix* population[7,13].

Samples were collected with a levy cane, stored in 50 mL sterile Falcon tubes and flash-frozen on site. Biomolecules were extracted in randomized batches after the end of the sampling period. A total of 53 samples was extracted. The set included two preliminary samples (2010-10-04 and 2011-01-25) and 51 samples from a higher frequency sampling phase (2011-03-21 to 2012-05-03).

Polar and nonpolar metabolites, DNA, RNA, and proteins were extracted in a sequential co-isolation procedure[13,51]. Around 200 mg of frozen samples were weighed out. Extracellular metabolites were extracted from the supernatant with cold chloroform and methanol–water, and separated into polar and nonpolar fractions. Intracellular metabolites were isolated in the same way after a lysis step by cryomilling, followed by sequential spin column-based (Qiagen Allprep) purification of RNA, DNA, and proteins.

Abiotic factor measurements and data processing. At the time of sample collection, the following physico-chemical parameters were measured inside the tank with a portable field kit (Hach) on-site: pH, conductivity, oxygen-levels, and temperature.

Additionally, online monitoring measurements were recorded by the BWWTP operators including nitrate, phosphate, ammonium, dry-matter and dissolved oxygen levels at the outflow as well as conductivity and pH at the inlet, and pH and temperature inside the sampled 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 using the method *stine*[52].

Metaproteomic analyses. Protein samples were separated by 1D SDS-PAGE (Criterion precast 1D gel, Bio-Rad), stained and cut into 2 mm bands. Peptides were subjected to liquid chromatography (LC) after in-gel reduction, alkylation and tryptic digestion. An Easy-nLC column (Proxeon, Thermo Fisher Scientific) was used. The peptide mixture was separated with a binary solvent gradient for elution with 0.1% formic acid in water and 0.1% formic acid in acetonitrile. Mass spectrometry was performed with an LTQ-Orbitrap Elite (ThermoFisher Scientific) on an 11-scan cycle consisting of a single precursor scan at a mass range of 300−2000 m/z followed by ten data-dependent MS/MS scan events. MS/MS scans were carried out with an isolate width of 2 m/z and a normalized collision energy of 35. Additional details of the metaproteome preparations and measurements are described in a previous study[13].

We converted raw mass spectrometry files to MGF format using MSconvert[53] using default parameters. The Graph2Pro pipeline[54] was used to process the resulting files together with the corresponding MG and MT co-assembly graphs from MEGAHIT[55]. The Graph2Pro pipeline uses FragGeneScan[56] to predict proteins from the long edges in the assembly graphs (i.e., from the contigs). In addition, it predicts tryptic peptides that span multiple edges in the graphs. Search databases were constructed using the putative proteins and tryptic peptides, respectively. These were used for initial peptide identification with the MS/GF+ search engine[57]. Identified tryptic peptides were then combined and used as the constraints for Graph2Pro to predict protein sequences from the co-assembly graphs. The generated sample-specific databases produced by Graph2Pro were used for the final metaproteomic searches using MS/GF+ (second search) to produce the final identification results. MS–GF+ was used for the final peptide identification with custom parameters: the instrument type was set to high-resolution linear trap quadropole (LTQ) with a precursor mass tolerance of 15 ppm and an isotope error range of −1 and 2 and the minimum and the maximum precursor charges were set to 1 and 7, respectively. We estimated the false discovery rate (FDR) with a target-decoy search approach using reverse sequences of the protein entries while preserving the C-terminal residues (KR). An FDR threshold of 1% was used. Identified peptides from the Graph2Pro pipeline were assigned to coding sequences (CDS) of rMAGs from prokka-based[58] predictions (see below) by using the command line interface version of peptidematch[59]. Spectral counts for sample-specific peptide sequences were assigned to matching CDS.

Meta-metabolomic analyses. Four distinct measurements for the metabolite extracts were performed: i) nonpolar extracellular, ii) polar extracellular, iii) nonpolar intracellular, and iv) polar intracellular. Metabolite extracts were derivatized using a multipurpose 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 of N-methyl-N-trimethylsilyl-trifluoroacetamide (MSTFA; Macherey-Nagel), samples were incubated for additional 30 min at 40 °C under continuous shaking. Dried nonpolar samples were dissolved in 30 µL MSTFA and incubated under shaking for 60 min at 40 °C. For quality control, pool samples, i.e., a combination of all extracts of the same measurement[27], were introduced in the measurement sequence after every fifth measurement.

GC-MS analysis was performed 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 (intracellular and extracellular polar fraction) and split mode (10:1, intracellular non-polar fraction) 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 at a constant 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[60] (v. 2.5). 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 automatically annotated by retention time and mass spectrum using an in-house mass spectral library. Detected metabolite derivatives (_xMeOX_xTMS/_xTMS) were used for further statistical data analysis.

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 25% (90% for correlation analyses) of samples. Metabolite intensities were normalized with respect to pool samples to account for instrument drift as described previously[51] by dividing the intensity values by the mean of up to two preceding and subsequent pool samples according to the measurement sequence. Metabolite derivate names of identified metabolites were manually assigned to KEGG compound identifiers and CHEBI IDs.

Metagenomic and metatranscriptomic analyses. MG libraries were prepared as paired-end libraries with the AMPure XP/Size Select Buffer Protocol following a size selection step[13]. RNA libraries were prepared after washing stored extracts with ethanol and depletion of rRNAs with the Ribo-Zero Meta-Bacteria rRNA Removal Kit (Epicenter). The ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicenter) was used for cDNA library preparation. Libraries were sequenced on an Illumina Genome Analyser (GA) IIx instrument with a read-length of 100 bps paired-end. Downstream processing and assembly of MT and MG reads was carried out with IMP[28] version 1.3 with the following parameters: i) Illumina Truseq2 adapters were trimmed, ii) the filtering step for reads of human origin was omitted, and iii) the MEGAHIT v.1.0.6 de novo assembler[55] was selected for coassembly of the MG and MT data. Co-assembled contigs from each timepoint were binned based on nucleotide signatures, presence of single-copy essential genes and metagenomic depth of coverage[61]. MAGs from each timepoint with at least 28% completeness and with a contamination of less than 20% based on essential marker gene content[62] were retained for downstream selection of representative population-level genomes (rMAGs). To this end, MAGs were dereplicated with dRep[29] using the following parameters: i) completeness threshold of 0.6, ii) strain heterogeneity threshold of 101, iii) primary cluster identity of 0.6, and iv) secondary cluster nucleotide identity of 0.965, and other parameters at default settings. In a following step, a subset of rMAGs with the highest completeness rates was selected based on CheckM[63] completeness estimates, requiring at least 0.50 in the difference of completeness and contamination estimates. Furthermore, rMAGs without taxonomic assignment on kingdom level were removed as they could represent misassembled contigs, resulting in a set of 78 coherently taxonomically annotated rMAGs that were used for the time-series analysis. For downstream analyses, MG and MT reads from all time-points were mapped using bwa mem[64] per time-point using the rMAGs as references. MG and MT depth-of-coverage per time-point were computed on the gene and contig level by dividing the summed depth per base by the length of the respective sequence.

Assembled contigs from IMP were annotated with Prokka v1.11[58] including prediction of full-length coding sequences (CDS) with prodigal v2.60[65]. Predicted CDS were also searched with an in-house Hidden Markov Model (HMM) database[61] of KEGG ortholog groups (KO) using HMMer v.1.12b[66]. We inferred compounds linked to CDS through CDS-to-reactions links from predicted enzymes with their respective KO annotation and EC assignment. Links to FOAM ontology categories[67] were assigned to each CDS by matching KO annotations. To assign MIMAG classifications[68] for all MAGs, assembly statistics, e.g., N50, were computed with the R-package seqinr v3.6-1[69]. tRNAs were predicted with Aragorn v1.2.38[70] and MAGs were screened for rRNA genes with barrnap 0.9[71].

Taxonomic assignment of rMAGs was performed using AMPHORA2[72] in combination with sourmash-lca v. 2.0.0a1[73], kmer-length:21 and threshold:4 and an existing database including approximately 87,000 microbial genomes (downloaded on 2017-11-09 from https://osf.io/s3jx8/download). If no taxonomic assignment was possible by whole genome-comparison (sourmash-lca), predictions for unassigned levels were augmented with consensus predictions using AMPHORA2: Assignments based on individual marker genes were combined by summation of the associated assignment probabilities. The consensus assignment with the highest overall score was determined. If the consensus assignment scores constituted for less than one third of the total probability scores the assignment was discarded as “low confidence assignment”.

Ex situ experiments. The ex situ experiments were performed in bioreactors seeded with sludge samples and diluted 1:5 (v/v) with artificial wastewater with a

final volume of 2 l[8]. The mixed sludge was split into two aliquots subjected to aerobic or anoxic preconditioning for 2 h. For the anoxic preparations a gaseous dioxygen-free environment (<1000 ppm) was achieved and monitored within a glove box (Jacomex, Dagneux, France). Thirty milliliters of the sludge mixtures were transferred into 50 mL serum vials connected to a multifold valve system to generate alternating aerobic (compressed air) or anaerobic (nitrogen gas) conditions. Samples were subjected to aerobic, anoxic, or alternating conditions (in 1 h intervals) after 2 h of preconditioning. After the preconditioning (time-point 0 h), oleic acid was supplemented at 500 µM alongside nitrate (80 µM) and phosphate (16 µM). Additional samples for concomitant DNA and RNA extraction and sequencing were taken at 5 and 8 h. This resulted in 12 samples for the four conditions tested (aerobic, anoxic, aerobically preconditioned followed by alternating, and anoxically preconditioned followed by alternating).

Isolated DNA for the 12 samples was sequenced on an Illumina Hiseq 2500 with a read-length of 250 bps paired-end. Isolated RNA was reverse transcribed to cDNA and sequenced with a read-length of 100 bp paired-end. Resulting MT and MG reads were pre-processed with IMP[28] and mapped to the rMAGs reconstructed from the long-term time-series as described above. Raw read counts per CDS were determined with featureCounts[74] and compared with DESeq2[75] across all conditions.

Statistical analyses. Statistical analyses were performed using R 3.4.4 and R 3.6.1[76] with prevalent use of the tidyverse R-package[77].

Determination of niche types. Annotated KOs for the individual rMAGs were summarized in a binary presence/absence matrix, in which 0 s indicated absence and 1 s indicated presence of at least one gene annotated with the respective KO. Subsequently, pairwise binary Jaccard-distances between the rMAGs based on these KO profiles were calculated and projected into two-dimensional space by multidimensional scaling (MDS). To determine the clustering of rMAGs in the resulting embedding, the appropriate number of clusters was determined by utilizing the k-means function for a range of centroids (ranging from 1 to 9 centroids) and determining the total within-sum-of-squares error as a measure of variability of resulting clusters. Functional clusters (FunCs) were then determined by k-means clustering. Enrichment of individual KOs within the assigned FunCs was determined with Fisher's exact test based on the number of rMAGs with the assigned KO. Resulting *p*-values were adjusted by FDR correction (function p.adjust method = "fdr") and KOs with a *p*-value below 0.05 were considered as enriched within a FunC. To test the relationship of rMAGs abundance and FunC assignment, pairwise Pearson correlation (*cor.test* in R) was computed between the relative abundance values of the rMAGs across time. Resulting correlation coefficients (ρ) were transformed to distances with the following formula: $1 - \frac{\rho+1}{2}$. Dispersion of these distances was assessed with the *betadisper* function of the vegan package[78]. Association of the FunC assignment to the distances was tested using the *adonis* function.

Whole genome-based pairwise distances between all rMAGs were calculated with mash v.2.2.2[30] (-k 21 –s 10,000) and embedded in two dimensions using MDS. PROCUSTES from the vegan package[78] was used to map the whole genome-based embedding onto the KO profile-based embedding. PROTEST from the vegan package was used with 9999 permutations.

Linking abiotic factors to population abundances. Measurements of abiotic factors (metabolites and physico-chemical parameters), as well as ratios of intracellular and extracellular metabolite intensity ratios were transformed to *z*-scores. Relative abundances of rMAGs were associated to abiotic factor levels by testing for correlation (*cor.test* function, method = "spearman"). Additionally, abiotic factor levels were placed onto a 2D ordination of MG or MT-based abundance profiles (Bray–Curtis dissimilarity) applying the vegan function *scores*.

Correlation of gene levels. To assess the expression of pathways over time, MT depth and MP spectral counts for genes of rMAGs were summed for each L1 FOAM category[67]. Grouped values were divided by the total MT depth or MP counts of all rMAGs per time-point to obtain the relative contribution of genes assigned to a specific category. The relative values were scaled to values between 0 and 1. Correlations between the scaled MT and MP time-series for each functional category were calculated with the *cor* function in R.

Clustering of expression profiles. To characterize expression profiles of the distinct rMAGs over time, gene functions (KOs) were summarized as active or inactive depending on MT/MG ratios or evidence at the MP level. KOs were considered active if at least one gene with the KO matched the following criteria: either the MT/MG depth ratio of the gene was greater than 1 or at least 2 peptide spectral counts could be assigned to the gene. If the MG depth of a gene was below one, the MT depth was considered instead of the MT/MG ratio to avoid inflating active KOs for lowly abundant populations. Binary Jaccard distances of the resulting KO profiles were determined for each rMAG separately for each time-point. Clusters were determined in each of the resulting 51 ordinations with the *hdbscan* function of the dbscan package with a minimum of five members per cluster. The resulting clusters were used to assess which functions were expressed over time by different subsets of rMAGs with a focus on rMAGs assigned to the same clusters as *Microthrix* D51_G1.1.2 or *Acinetobacter* A01_O1.2.4.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Meta-omics data from five individual time-points has previously been published[13,26,51]. The MG and MT FASTQ files and the sample-wise MT-assembly and co-assembly contigs are available on NCBI BioProject PRJNA230567. MP data has been deposited in the PRIDE database under the accession number PXD013655. Raw metabolomics data is available at MetaboLights under the accession MTBLS2021, while processed intensities after identification are provided with this manuscript (Supplementary Data 3). Similarly, physico-chemical parameters are provided with this manuscript (Supplementary Data 4). Processed and intermediary data files from the combined multi-omic analyses, e.g., annotated and normalized MT, MG read counts, are available at Zenodo (https://doi.org/10.5281/zenodo.3961685). External databases were used in this study: KEGG (https://www.genome.jp/kegg/), CHEBI (https://www.ebi.ac.uk/chebi/).

Code availability

Code used for genome reconstruction and dereplication is available at the LCSB R3 GitLab (https://git-r3lab.uni.lu/shaman.narayanasamy/LAO-time-series). Code used for the processing and analyses of the meta-omics data, as well as for additional analyses and generation of plots for main and supplemental figures is also available at the LCSB R3 GitLab (https://git-r3lab.uni.lu/malte.herold/laots_niche_ecology_analysis).

Received: 27 February 2020; Accepted: 16 September 2020;
Published online: 19 October 2020

References

1. Timmis, K. et al. The contribution of microbial biotechnology to sustainable development goals. *Microb. Biotechnol.* **10**, 984–987 (2017).
2. Cavicchioli, R. et al. Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* **17**, 569–586 (2019).
3. Sheik, A. R., Muller, E. E. L. & Wilmes, P. A hundred years of activated sludge: time for a rethink. *Front. Microbiol.* **5**, 1–7 (2014).
4. van Loosdrecht, M. C. M. & Brdjanovic, D. Anticipating the next century of wastewater treatment. *Science* **344**, 1452–1453 (2014).
5. Muller, E. E., Sheik, A. R. & Wilmes, P. Lipid-based biofuel production from wastewater. *Curr. Opin. Biotechnol.* **30**, 9–16 (2014).
6. Castro, A. R. et al. Tuning culturing conditions towards the production of neutral lipids from lubricant-based wastewater in open mixed bacterial communities. *Water Res.* **144**, 532–542 (2018).
7. Rossetti, S., Tomei, M. C., Nielsen, P. H. & Tandoi, V. "Microthrix parvicella", a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiol. Rev.* **29**, 49–64 (2005).
8. Sheik, A. R. et al. In situ phenotypic heterogeneity among single cells of the filamentous bacterium Candidatus Microthrix parvicella. *ISME J.* **10**, 1274–1279 (2016).
9. Johnson, D. R., Lee, T. K., Park, J., Fenner, K. & Helbling, D. E. The functional and taxonomic richness of wastewater treatment plant microbial communities are associated with each other and with ambient nitrogen and carbon availability. *Environ. Microbiol.* **17**, 4851–4860 (2015).
10. Xu, S., Yao, J., Ainiwaer, M., Hong, Y. & Zhang, Y. Analysis of bacterial community structure of activated sludge from wastewater treatment plants in winter. *Biomed. Res. Int.* **2018**, 1–8 (2018).
11. Oehmen, A. et al. Advances in enhanced biological phosphorus removal: From micro to macro scale. *Water Res.* **41**, 2271–2300 (2007).
12. Xie, B., Dai, X.-C. & Xu, Y.-T. Cause and pre-alarm control of bulking and foaming by Microthrix parvicella—a case study in triple oxidation ditch at a wastewater treatment plant. *J. Hazard. Mater.* **143**, 184–191 (2007).
13. Muller, E. E. L. et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* **5**, 5603 (2014).
14. Hug, T., Gujer, W. & Siegrist, H. Modelling seasonal dynamics of Microthrix parvicella. *Water Sci. Technol.* **54**, 189–198 (2006).
15. Muller, E. E. L. et al. Using metabolic networks to resolve ecological properties of microbiomes. *Curr. Opin. Syst. Biol.* **8**, 73–80 (2018).
16. Muller, E. E. L. Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* **4**, 1–6 (2019).

17. Hultman, J. et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521**, 208–212 (2015).
18. DeAngelis, K. M., Silver, W. L., Thompson, A. W. & Firestone, M. K. Microbial communities acclimate to recurring changes in soil redox potential status. *Environ. Microbiol.* **12**, 3137–3149 (2010).
19. Plichta, D. R. et al. Transcriptional interactions suggest niche segregation among microorganisms in the human gut. *Nat. Microbiol.* **1**, 16152 (2016).
20. Wilmes, P. et al. Metabolome-proteome differentiation coupled to microbial divergence. *MBio* **1**, 3–7 (2010).
21. Prosser, J. I. Ecosystem processes and interactions in a morass of diversity. *FEMS Microbiol. Ecol.* **81**, 507–519 (2012).
22. Hester, E. R., Jetten, M. S. M., Welte, C. U. & Lücker, S. Metabolic overlap in environmentally diverse microbial communities. *Front. Genet.* **10**, 653881 (2019).
23. Comte, J., Fauteux, L. & Del Giorgio, P. A. Links between metabolic plasticity and functional redundancy in freshwater bacterioplankton communities. *Front. Microbiol.* **4**, 1–11 (2013).
24. Sabra, W., Dietz, D., Tjahjasari, D. & Zeng, A.-P. Biosystems analysis and engineering of microbial consortia for industrial biotechnology. *Eng. Life Sci.* **10**, 407–421 (2010).
25. Brenner, K., You, L. & Arnold, F. H. Engineering microbial consortia: a new frontier in synthetic biology. *Trends Biotechnol.* **26**, 483–489 (2008).
26. Roume, H. et al. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *npj Biofilms Microbiomes* **1**, 15007 (2015).
27. Roume, H. et al. A biomolecular isolation framework for eco-systems biology. *ISME J.* **7**, 110–121 (2013).
28. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
29. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
30. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
31. Heintz-Buschart, A. & Wilmes, P. Human gut microbiome: function matters. *Trends Microbiol.* **26**, 563–574 (2018).
32. Wältermann, M., Stöveken, T. & Steinbüchel, A. Key enzymes for biosynthesis of neutral lipid storage compounds in prokaryotes: properties, function and occurrence of wax ester synthases/acyl-CoA:diacylglycerol acyltransferases. *Biochimie* **89**, 230–242 (2007).
33. Alvarez, H. M. Triacylglycerol and wax ester-accumulating machinery in prokaryotes. *Biochimie* **120**, 28–39 (2016).
34. Alvarez, H. M. & Steinbüchel, A. Triacylglycerols in prokaryotic microorganisms. *Appl. Microbiol. Biotechnol.* **60**, 367–376 (2003).
35. Zheng, F., Long, Q. & Xie, J. The function and regulatory network of WhiB and WhiB-like protein from comparative genomics and systems biology perspectives. *Cell Biochem. Biophys.* **63**, 103–108 (2012).
36. Ju, F. & Zhang, T. Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant. *ISME J.* **9**, 683–695 (2015).
37. McIlroy, S. J. et al. MiDAS: the field guide to the microbes of activated sludge. *Database* **2015**, bav062 (2015).
38. Liu, T., Liu, S., Zheng, M., Chen, Q. & Ni, J. Performance assessment of full-scale wastewater treatment plants based on seasonal variability of microbial communities via high-throughput sequencing. *PLoS ONE* **11**, e0152998 (2016).
39. Tilman, D. The ecological consequences of changes in biodiversity: perspectives. *Ecology* **80**, 1455–1474 (1999).
40. May, R. M. & Arthur, R. H. M. Niche overlap as a function of environmental variability. *Proc. Natl Acad. Sci.* **69**, 1109–1113 (1972).
41. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
42. McIlroy, S. J. et al. Metabolic model for the filamentous 'Candidatus Microthrix parvicella' based on genomic and metagenomic analyses. *ISME J.* **7**, 1161–1172 (2013).
43. Kindaichi, T., Nierychlo, M., Kragelund, C., Nielsen, J. L. & Nielsen, P. H. High and stable substrate specificities of microorganisms in enhanced biological phosphorus removal plants. *Environ. Microbiol.* **15**, 1821–1831 (2013).
44. Teixeira, L. M. & Merquior, V. L. C. In *The Prokaryotes: Gammaproteobacteria* (eds. Rosenberg, E. et al.) 443–476 (Springer, Berlin, 2014).
45. McIlroy, S. J. et al. Culture-independent analyses reveal novel Anaerolineaceae as abundant primary fermenters in anaerobic digesters treating waste activated sludge. *Front. Microbiol.* **8**, 1134 (2017).
46. Nielsen, P. H., Roslev, P., Dueholm, T. E. & Nielsen, J. L. Microthrix parvicella, a specialized lipid consumer in anaerobic–aerobic activated sludge plants. *Water Sci. Technol.* **46**, 73–80 (2002).
47. Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P. & Thiele, I. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome* **3**, 55 (2015).
48. Plata, G., Henry, C. S. & Vitkup, D. Long-term phenotypic evolution of bacteria. *Nature* **517**, 369–372 (2015).
49. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: networks, competition, and stability. *Science* **350**, 663–666 (2015).
50. Shade, A. et al. Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* **3**, 1–19 (2012).
51. Roume, H., Heintz-Buschart, A., Muller, E. E. L. & Wilmes, P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* **531**, 219–236 (2013).
52. Moritz, S. & Bartz-Beielstein, T. imputeTS: time series missing value imputation in R. *R. J.* **9**, 207–218 (2017).
53. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
54. Tang, H., Li, S. & Ye, Y. A graph-centric approach for metagenome-guided peptide and protein identification in metaproteomics. *PLoS Comput. Biol.* **12**, 1–16 (2016).
55. Li, D., Liu, C., Luo, R., Sadakane, K. & Lam, T. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
56. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191–e191 (2010).
57. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
58. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
59. Chen, C., Li, Z., Huang, H., Suzek, B. E. & Wu, C. H. A fast peptide match service for UniProt knowledgebase. *Bioinformatics* **29**, 2808–2809 (2013).
60. Hiller, K. et al. Metabolite detector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal. Chem.* **81**, 3429–3439 (2009).
61. Heintz-Buschart, A. et al. Integrated multiomics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2016).
62. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
63. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
64. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
65. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
66. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
67. Prestat, E. et al. FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res.* **42**, e145–e145 (2014).
68. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
69. Charif, D. & Lobry, J. R. in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds. Bastolla, U. et al.) 207–232 (Springer, Berlin, 2007).
70. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
71. Seemann, T. barrnap 0.9: rapid ribosomal RNA prediction. https://github.com/tseemann/barrnap.
72. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
73. Brown, C. T. & Irber, L. Sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* **1**, 27 (2016).
74. 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).
75. 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).
76. Team, R. D. C. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2011).
77. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 168 (2019).
78. Oksanen, J. et al. The vegan package. *Community Ecol. Packag.* **190**, 719 (2008).

Acknowledgements

We thank Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal a Vocation Ecologique (SIVEC), for their permission to collect samples and access to the monitoring platform of the Schifflange wastewater treatment plant. We also thank Dr. Olivia Judson for her feedback on the title and abstract. Bioinformatic analyses presented in this paper were carried out using the high-performance computing facilities of the University of the Luxembourg. This work was supported by an ATTRACT programme grant (ATTRACT/A09/03) and INTER programme grant (INTER/SYSAPP/14/05), two CORE programme grants (C15/SR/10404839, C17/SR/11689322), and a PRIDE doctoral training unit grant (PRIDE15/10907093) to P.W., and a grant to A.R.S (PDR-2013-1/5748561) all funded by the Luxembourg National Research Fund (FNR). As one of the authors (A.R.S.) holds a position at the European Commission the following applies: "The views expressed are purely those of the writer and may not in any circumstances be regarded as stating an official position of the European Commission."

Author contributions

E.E.L.M., L.A.L., H.R., A.R.S. and P.W. performed the sampling and the biomolecular extractions. M.H., S.M.A., S.N., A.H.B., P.M., B.K., C.C.L., and P.W. analysed the data. A. R.S. performed short-term experiments and biomolecular extractions. I.B. and R.B.H.W. performed RNA and DNA sequencing for the ex situ experiment samples. J.D.G., J.M.S., and P.S.K. performed RNA and DNA sequencing for the in situ time-series samples. C.J. performed the metabolomics experiments and data analysis. M.R.H and R.L.M. measured the proteomic data. M.H., B.J.K., Y.Y., S.L., and H.T. analysed the proteomic data. M.H., E.E.L.M., and L.A.K.K.B. performed the metabolic characterization of the reconstructed genomes. M.H., E.E.L.M., A.H.B., P.M., C.C.L., and P.W. designed the study and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19006-2.

Correspondence and requests for materials should be addressed to P.W.

Peer review information *Nature Communications* thanks Janet Jansson, Grace Pold and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.4 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 **Uwe Sauer** and **Bas Teusink**

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

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].

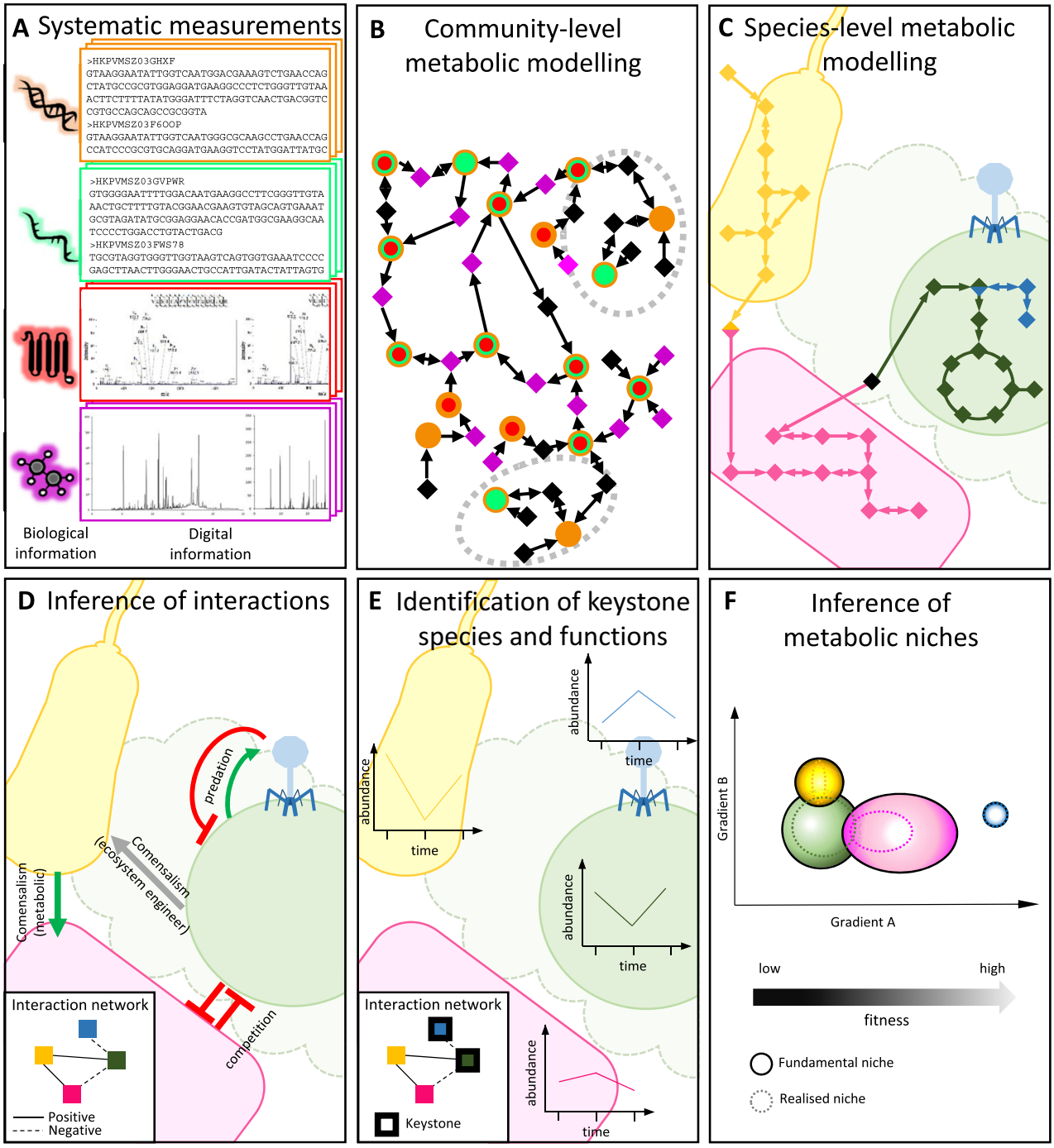
Different categories of keystone species have been proposed including ecosystem engineer (or modifier) keystone species (Figure 1E), trophic (prey or predator) keystone species or resource provider keystone species [48]. In any case, keystone species confer keystone functionalities to the ecosystem [49]. 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



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.

Appendix B

Additional tables

B.1 Sample-wise summary.

This table includes sampling dates, MG and MT sequencing information, large-scale bioinformatics processing from IMP, sample assessment results from Non-pareil and number of bins. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 1.

B.2 Metaproteomics summary.

Overview of the metaproteomic data per sample including the number of MS/MS spectra measured, number of spectra assigned to a peptide, the percentage of identified spectra and the total number of peptides. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 2.

B.3 Taxonomy of rMAGs.

Summary of representative metagenome-assembled genomes (rMAGs), including standard assembly statistics and taxonomic predictions. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 3.

B.4 CRISPR-Cas information of rMAGs.

Summary of CRISPR-Cas information per rMAG including number of CRISPR elements detected, sequence of CRISPR repeats and classification of the CRISPR-Cas system. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 4.

B.5 Function within iMGEs.

COG categories identified in plasmids and phages. Number of genes and number of protospacers per functional category. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community**

dynamics revealed using time-series integrated meta-omics (**Appendix A.2**) as Supplementary table 7.

B.6 Functions within iMGEs-PSCCs.

This information consists of a set of tables separated by COG functional categories and the type of iMGEs, i.e. phage or plasmid, in which they are found. Each table contains predicted gene functions for each COG category and information on protospacer-containing contigs (PSCCs) and non-PSCCs. These tables are publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Additional data, and can be found in the following link `https://zenodo.org/record/3774024`.

B.7 Specific ARGs within the iMGEs.

List of ARGs from ResFams database found within the iMGEs. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 9.

B.8 CRISPR-Cas information of rMAGs.

Pearson correlations between MG-derived relative abundances of family-level bacterial, plasmid and phage groups, filtered by p-value ≤ 0.001. Columns correspond to the analysed time intervals, i.e. entire time-series, before, during, and after the community shift. The table shows correlations only if they are significant in selected time intervals, i) all the analyzed time intervals, ii) before, during and after the community shift, and not in the entire time-series, iii) before and during the community shift, and not in the other time intervals, iv) during and after the community shift, and not in the other time intervals, and v) before and after the community shift, and not in the other time intervals. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 10.

B.9 Summary of the linear models predicting *Microthrixaceae* bacterial abundance over time.

This table contains the analyzed time intervals, i.e. longer-term (entire time-series) as well as shorter-term intervals corresponding to before, during, and after the community shift. Statistical tests were two-sided. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 11.

B.10 List of family-level bacteria, plasmid and phage groups within the optimal linear models.

The table contains the family-level groups that were significant within the linear models of *Microthrixaceae* bacterial family as response variable performed on the different time intervals, i.e. entire time-series as well as before, during, and after the community shift. Columns represent presence (1) or absence (0) per time interval, and the intersection of specific time intervals. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 12.

B.11 Summary of spacers activity.

CRISPR spacer gain and loss events, summarized by type of targeted iMGE, i.e. plasmid or phage, and per microbial population. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 13.

B.12 iMGE-host CRISPR based networks attributes.

This table contains the network properties of the plasmid-host and the phage-host CRISPR networks over time, i.e. network properties as number of nodes, number of interactions, modularity and nestedness over time, and node properties as betweenness, closeness and degree. This table is publicly available as part of the publication **Roles of bacteriophages,**

plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics (**Appendix A.2**) as Supplementary table 14.

B.13 One mode projection network from iMGE-CRISPR host networks.

The one-mode network represents the rMAGs, obtained from the plasmid- and phage-CRISPR host networks. The information within this table contains the number of common iMGEs between rMAGs and the network and node attributes. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 15.

B.14 CRISPR locus information of *Ca.* M. parvicella.

Comparative CRISPR locus analyses of *Candidatus* Microthrix parvicella Bio17-1 isolate genome and the contig containing *Ca*. M. parvicella -like CRISPR locus (D47_L1.43.1_contig_476300). This information is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Additional data, and can be found in the following link `https://zenodo.org/record/3766442`.

B.15 Summary information of spacers within rMAGs.

This table contains a summary of the final number of spacers within each rMAG containing CRISPRs, including spacers activity as gain and loss events. This table is publicly available as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**) as Supplementary table 16.

B.16 Top 20 nodes of the ecological networks.

This table contains the top 20 nodes of each ecological networks based on the node strength. The first column indicates the ID of the node, second, third, fourth and fifth columns indicate the timeframe of the network, and its values, 1 or NA, indicate presence or absence, respectively, of such node as one of the top 20 of each given network. This table is available in `https://doi.org/10.5281/zenodo.5113563`.

B.17 Taxonomic family of top 20 nodes of the ecological networks.

This table contains the taxonomic families assigned to the selected top 20 nodes of each ecological network. This table is available in https://doi.org/10.5281/zenodo.5113563.

B.18 Ecological interactions strength.

This table contains the results of the Wilcoxon test applied to define what type of ecological interactions were stronger within the ecological networks. This table is available in https://doi.org/10.5281/zenodo.5113563.

B.19 Ecological interactions per taxonomic family.

This table contains a summary of the amount of ecological interactions (percentage) in which taxonomic family is involved in all the ecological networks. This table is available in https://doi.org/10.5281/zenodo.5113563.

B.20 Taxonomic family interactions strength.

This table contains the results of the Wilcoxon test applied to define what type of ecological interactions were stronger within and between taxonomic families in the ecological networks. This table is available in https://doi.org/10.5281/zenodo.5113563.

Appendix C

Additional figures

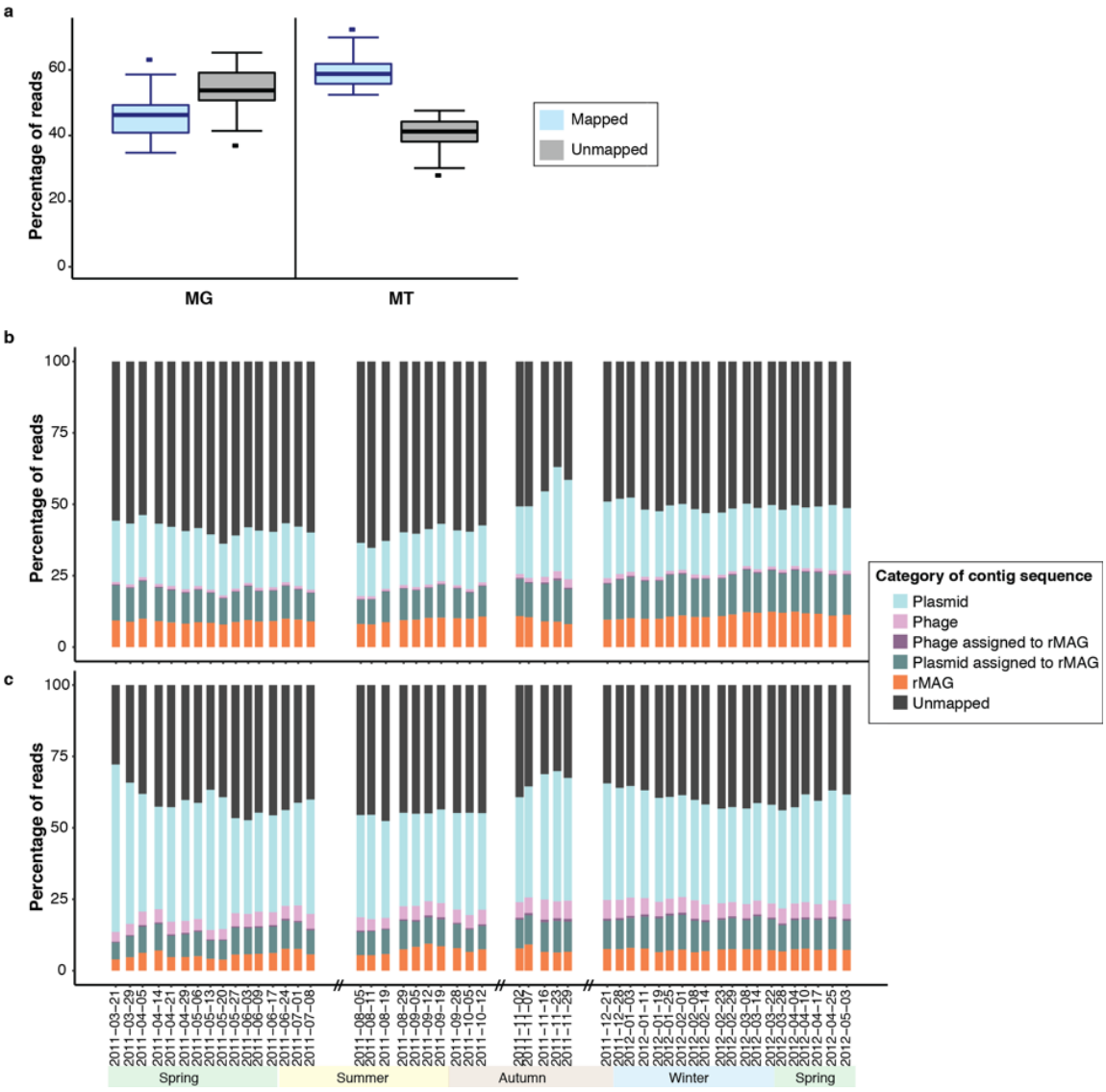


Figure C.1: Mapping of the MG and MT reads against representative metagenomics assembled genomes (rMAGs) and invasive mobile genetic elements (iMGEs). **a**, Boxplot representing the percentages of mapped and unmapped MG and MT reads within samples from the entire time-series (n=51 in situ samples). Data are presented as median values,$Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$. **b**, Barplot representing the percentages of mapped and unmapped MG reads per time point. c, Barplot representing the percentages of mapped and unmapped MT reads per time point. The labels in the x-axis indicate the exact sampling dates, and the double slashes (//) on the time axis represent absence of samples.

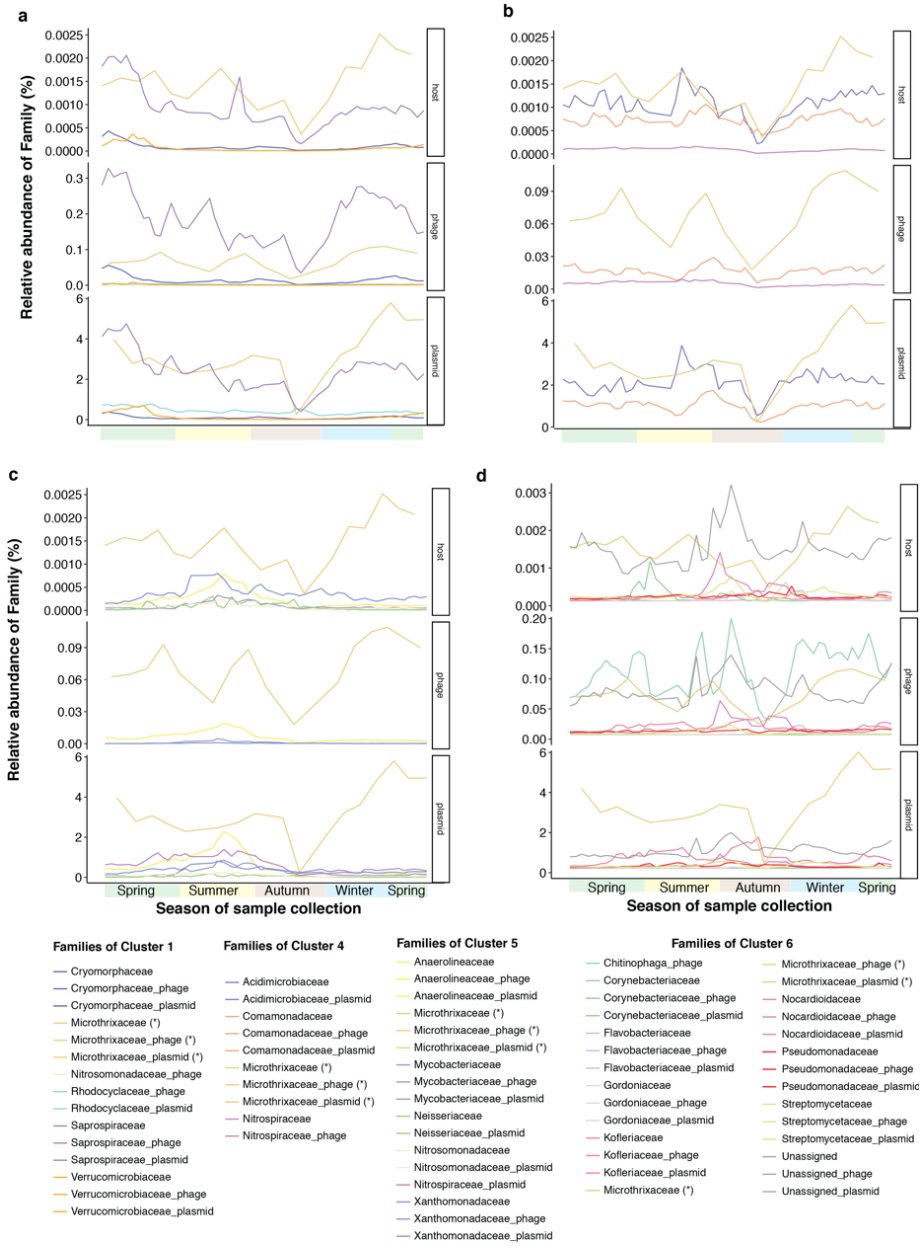


Figure C.2: Dynamics of clusters comprised of bacterial-, plasmid- and phage- groups. rMAGs were grouped at the family-level. Plasmids and phages were grouped based on their family-level association, i.e., binned together with an rMAG of a given family. The bacterial, plasmid and phage groups were clustered based on the correlation of their cumulative group-level abundance dynamics. Dynamics of the groups within **a**, Cluster 1, **b**, Cluster 4, **c**, Cluster 5, **d**, Cluster 6. *Microthrixaceae* family, and its associated plasmid and phage groups are found in Cluster 2 (shown in **Figure 2.7**). Plots display abundance dynamics of the *Microthrixaceae* family, plasmid and phage groups as proxy (i.e., not part of those clusters), being marked with an asterix (*) within those figures. The group "Unassigned" represents rMAGs that could not be classified on the family-level. Accordingly, the groups "Unassigned_plasmid" and "Unassigned_phage" represent plasmids and phages that were assigned to those "Unassigned" rMAGs. Unassigned (or unbinned) plasmids and phages were omitted from all figures. Relative abundance values on the y-axis were derived from MG data. The x-axis represents time, colour coded by seasons as labelled in panels **c** and **d**. Please refer to **Figure 2.1** for the exact sampling dates.

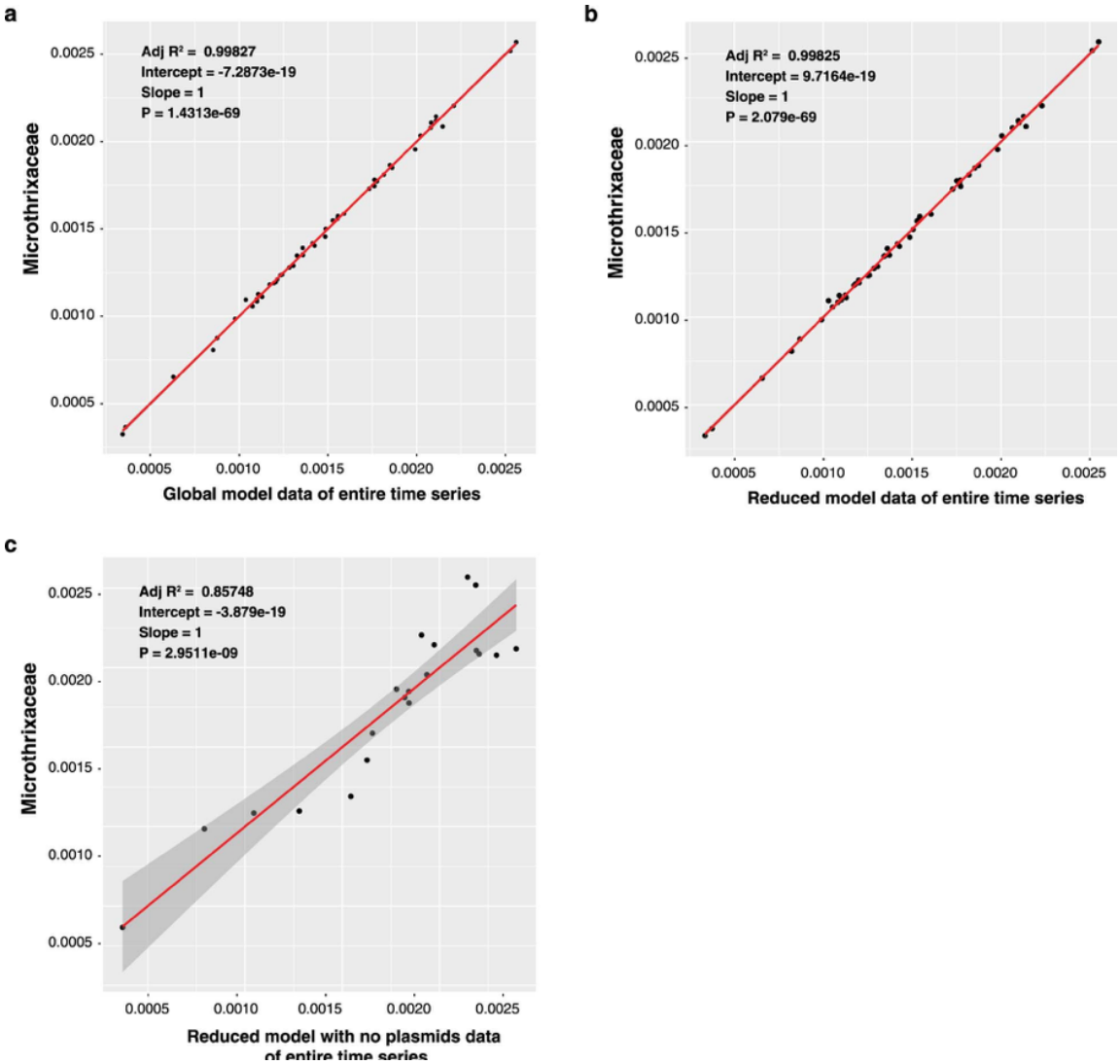


Figure C.3: Linear models predicting *Microthrixaceae* family abundance within the entire time-series. Model data fitted to the raw data of the entire time-series (n=51 in situ samples), specifically **a**, the best or global model, **b**, the reduced model, which lacks the non-significant families of the global model, and **c**, the reduced model without *Microthrixaceae*-plasmids. Gray bands represent the +/- standard error measurement of the regression line. Statistical tests were two-sided and adjusted for multiple comparisons.

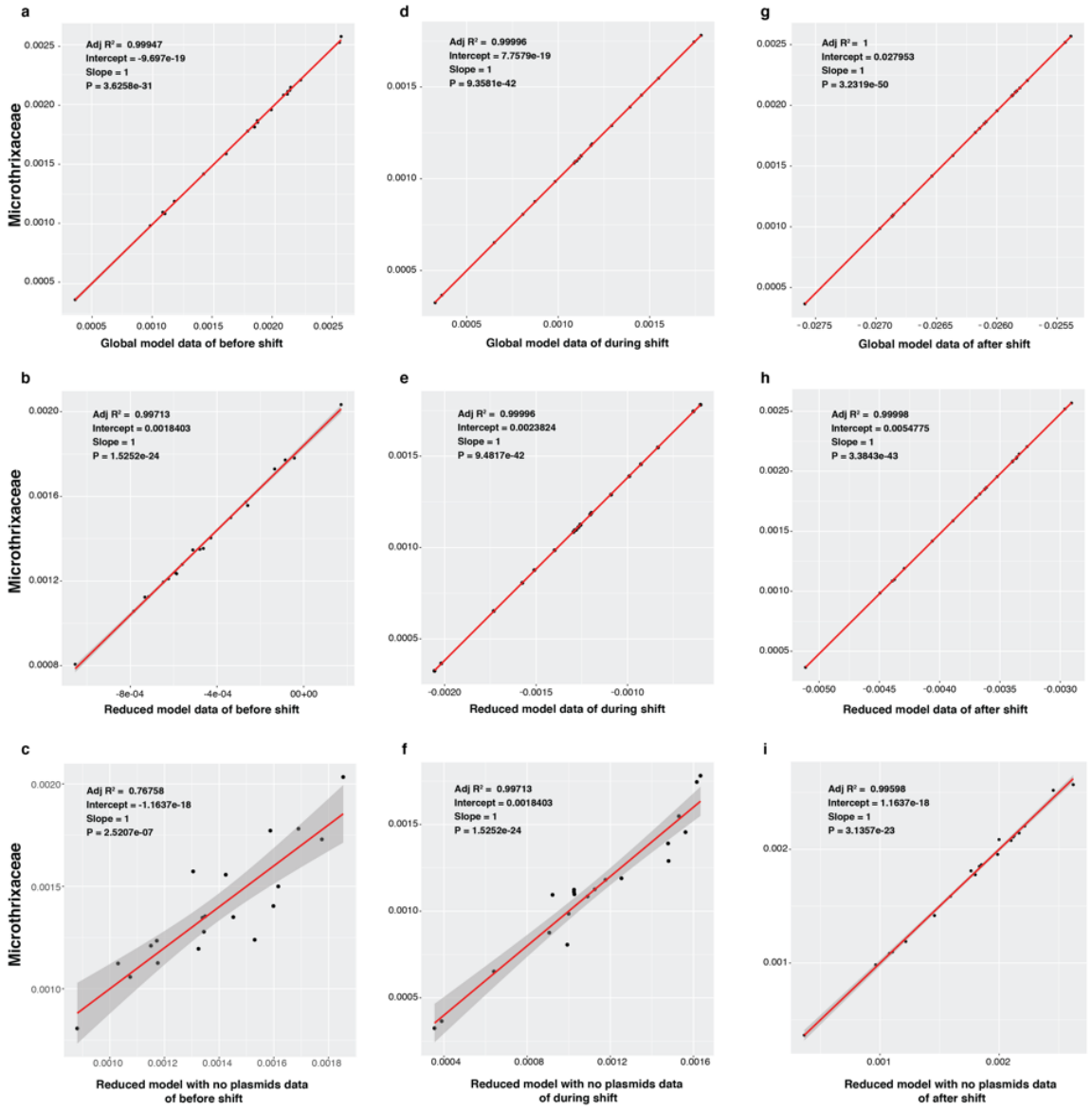


Figure C.4: Linear models predicting *Microthrixaceae* family abundance within different time intervals. Overlapping shorter-term intervals model data fitted to the raw data. All intervals consist of n=20 in situ samples, specifically **a**, **b**, **c**, Model data fitted to the raw data of the time interval before the community shift, between 2011-03-21 and 2011-08-29, specifically global, reduced and reduced without plasmids models, respectively. **d**, **e**, **f**, Model data fitted to the raw data of the time interval during the shift, between 2011-08-05 and 2011-01-19, specifically global, reduced and reduced without plasmids models, respectively, **g**, **h**, **i**, Model data fitted to the raw data of the time interval after the community shift, between 2011-12-21 and 2012-05-03, specifically global, reduced and reduced without plasmids models, respectively. Global linear models correspond to the optimal models in each time interval. Reduced linear models include only the significant features from the global models. Reduced models without plasmids include features from the reduced models but exclude plasmids of *Microthrixaceae*. Gray bands represent the +/- standard error measurement of the regression line. Statistical tests were two-sided and adjusted for multiple comparisons. Detailed results of the linear models, including model composition, residuals, and coefficients are shown in **Appendix B.9**

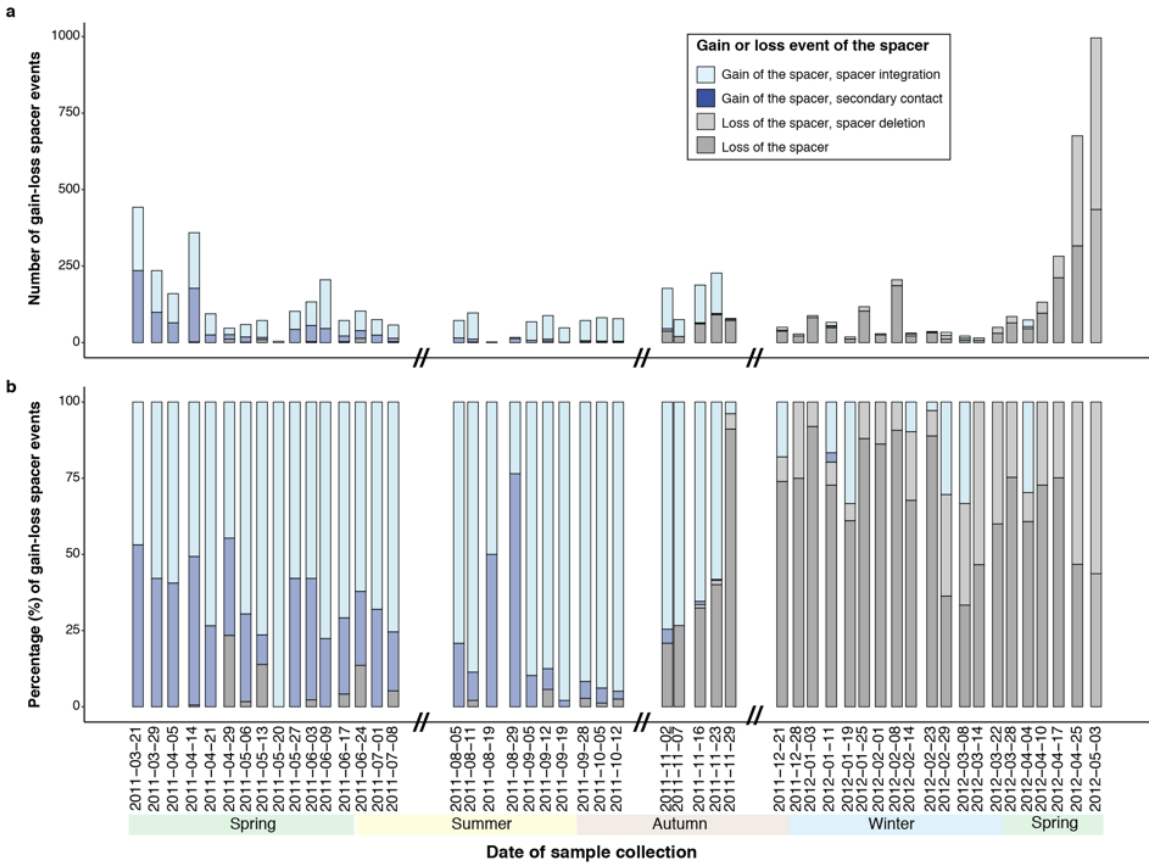


Figure C.5: Gain and loss of CRISPR spacers targeting iMGE. a, Barplot representing the number of spacers per time-point reflecting a gain or loss event. **b**, Representation of a in percentages. Gain events are defined as: i) "Gain of the spacer, spacer integration", when the iMGE was detected before, or at the same timepoint, as its linked spacer, and ii) "Gain of the spacer, secondary contact", when the spacer was detected before the linked iMGE within the time-series. Loss events are defined as: i) "Loss of the spacer, spacer deletion", when both the spacer and the iMGE are not detected anymore within the rest of the time-series, and ii) "Loss of the spacer", when the spacer is not detected within the time-series anymore, but the iMGE is still detected after spacer loss. The labels on the x-axis indicate the sampling dates and the double slashes (//) on the time axis represent absence of samples in the sampled system.

Appendix D

Videos

D.1 Video plasmid-host time-lapse networks

Host nodes (circles) are coloured based on their taxonomy. Material publicly available **here** as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**).

D.2 Video phage-host time-lapse networks

Host nodes (circles) are coloured based on their taxonomy. Material publicly available **here** as part of the publication **Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics** (**Appendix A.2**).

Appendix E

Additional notes

E.1 General assessment of the alignment of sequencing data

The MG and MT data provided the foundation for all the downstream analyses. We deemed it important to estimate the average community coverage based on the time-resolved MG and MT sequencing data. In this case, "community coverage" specifically refers to the estimated fraction of the genomes recovered in a given sequencing dataset, after accounting for factors such as community i) -richness, ii) -diversity and, most importantly, iii) sequencing depth [Rodriguez-R and Konstantinidis, 2014a]. The assessment was based on the output of Nonpareil [Rodriguez-R and Konstantinidis, 2014b] which estimates community coverage using unaligned (raw or pre-processed) sequencing reads. Specifically, we performed the assessment on IMP-based pre-processed MG and MT reads for each sample, rather than the raw sequencing reads, as the pre-processed reads are used for all downstream steps, including, but not limited to assembly, read mapping/alignment and inference of population sizes (rMAGs and iMGEs). Additionally, we performed the same assessment on the combined MG and MT data, given that IMP generates *de novo* assemblies based on these two data types. On average, the combined MG and MT sequencing depth achieved approximately 50% community coverage (**Figure E.1** and **Appendix B.1**) which allowed detailed, population-level study of the prominent community members. In general, we observed that when MG and MT reads were combined, the coverage estimations typically fell somewhere in between the MG and MT coverage values. The detailed coverage values are available in **Appendix B.1**.

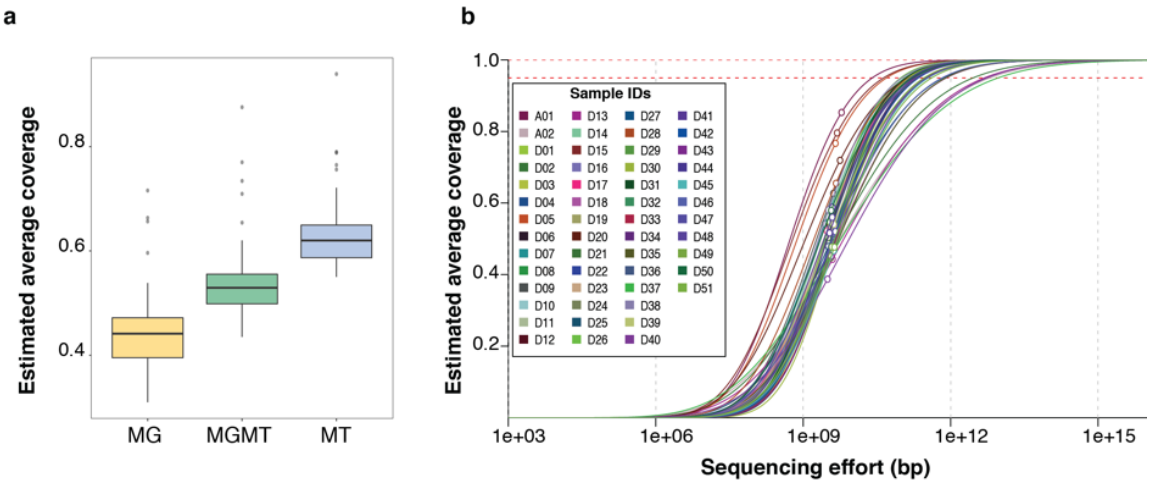


Figure E.1: Sequencing depth assessment of the metagenomic (MG) and metatranscriptomic (MT) datasets. a, Summary of Nonpareil results based on IMP-based preprocessed reads within all samples in the time series and the two initial samples (n=53 in situ samples). Data are presented as median values, $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$. Evaluation was performed for each of the following data sets: i) MG, ii) MT, and iii) combined MG and MT (i.e. MGMT) reads. **b**, Each curve represents a Nonpareil coverage model for each sample, whereby the x-axis value corresponds to the upper plateau of the curve (i.e. dotted red lines where y-axis value 1.0) and represents the theoretical sequencing effort required to cover all DNA and RNA elements within the community in terms of sequenced base pairs (bp). The circles on the curves represent the actual sequencing coverage of a given sample. The colors of the curves represent the individual samples.

E.2 Representative metagenome assembled genomes (rMAGs)

Given the time-resolved nature of our dataset, we aimed to link MAGs (i.e. bins) from different samples within the time-series that may be representative of the same microbial population. This was achieved through a systematic procedure to reduce the redundancy of the identified MAGs into representative MAGs (i.e. rMAGs), which are more suitable for downstream time-series analysis. The timepoint-specific binning was performed using the procedure described by Heintz-Buschart et al. [2016]. Briefly, two-dimensional penta-nucleotide-based maps for all contigs $\geq$ 1kbp were generated using Vizbin [Laczny et al., 2015] for assemblies of each timepoint. These maps were then clustered using the "dbscan" function in the R package "fpc" [Hennig, 2020]. Next, the number and multiplicity of 101 essential genes [Dupont et al., 2012; Albertsen et al., 2013] were used to assess the completeness and contamination of clusters generated from "dbscan". Clusters with multiple copies of the same essential genes were further divided by analysing the MG coverage depth of the essential genes. The two last steps are repeated three times on overcomplete bins [Heintz-Buschart et al., 2016]. Accordingly, the quality of bins, based

on their essential gene content are defined as follows: “P”: more than 100/109 essential genes, less than 115 essential genes in total (<14% duplicated genes, >92% complete), “G”: more than 71/109 essential genes (>65% complete), less than 20% in duplicate, “O”: more than 51/109 essential genes (>47% complete), less than 20% in duplicate, “L”: more than 31/109 essential genes (>28% complete), less than 20% in duplicate, “C”: at least 1/109 essential genes ($\leq$ 1% complete), less than 20% in duplicate, “E”: no essential genes, “B”: at least 1/109 essential genes (>= 1% complete), at least 20% in duplicate, “N”: noise (`https://git-r3lab.uni.lu/anna.buschart/MuStMultiomics/blob/master/autoCluster.R`). The binning procedure on all the timepoint assemblies yielded a total of 26,524 bins. Based on the quality metrics established by Heintz-Buschart et al. [2016], we selected 1,364 bins, now referred to as metagenomic species (MAGs) with quality criteria P, G, O and L and a collection of 85 isolate genomes for downstream dereplication [Olm et al., 2017]. The dereplication allowed us to link the selected MAGs and isolates from different time points. The dereplication process yielded 92 high-quality representative MAGs (i.e. rMAGs) which underwent taxonomic classification (detailed information of rMAGS available in **Appendix B.3**). Manual curation was carried out on rMAGs that were classified as *Candidatus* Microthrix parvicella. We then linked plasmids and phages to rMAGs based on the outcome of the binning cluster membership, i.e. if a plasmid or phage contig fell within a bin of a certain rMAG. Finally, we scanned the genomes for CRISPR operons, i.e. CRISPR loci with CRISPR-associated genes (cas genes) [Zhang and Ye, 2017].

E.3 Prediction of CRISPR elements

CRISPR information (i.e. repeats and spacers) were used to link host populations (i.e. rMAGs) to iMGEs. For this, we utilized two different tools to maximize the extraction of CRISPR information. First, CRASS [Skennerton et al., 2013] was used to extract CRISPR information (i.e. spacers, repeats and flanking sequences) directly from the preprocessed reads from IMP (both paired- and single- end reads). Next, we used metaCRT [Bland et al., 2007] to extract CRISPR information (i.e. spacers and repeats) on the contig level. We also used the contig-level information to extract flanking sequences from the metaCRT-derived CRISPR information to have equivalent information from both CRASS and metaCRT for further downstream processing. Overall, CRASS predicted more spacers (**Figure E.2**) while metaCRT predicted more repeats. This difference is likely due to the fact that spacers are more diverse elements, making direct extraction from sequencing reads particularly

effective compared to extraction from the *de novo* assembled, consensus-based contigs. Nevertheless, the use of both tools yielded complementary, contextual information. On the one hand, extraction of CRISPR information from sequencing reads using CRASS was necessary because i) *de novo* assemblers do not resolve repetitive regions, such as CRISPRs effectively, and ii) such approaches allow resolution of CRISPR information from lowly abundant and/or rare populations. On the other hand, the extraction of repeats using metaCRT from contigs allows the resolution of the CRISPR loci and linking these to the constituent rMAGs. Overall, the combination of these methods allowed for the extraction of comprehensive information regarding the different CRISPR loci which, in turn, allowed for detailed assessment of spacer complements and their linking to targeted iMGEs. We also inspected the representation of CRISPR elements on the MG and MT omic levels. In general, we found more repeats on the MT-level compared to the MG-level, while spacers were more abundant at the MG-level (**Figure 2.2** and **Figure E.2**). This may be due to the fact that the number of spacers is larger than the number of repeats. However, there may be different factors affecting the transcription of spacers including differential abundance of sub-populations and/or differential expression between and/or within CRISPR arrays. In particular, leading spacers are typically more highly transcribed compared to their lagging counterparts. A general assessment of the CRISPR elements highlighted that repeats had an average length of 30.9 bp (median=39 bp, SD=8.55 bp), while the shortest and longest sequences were 20 and 77 bp in length, respectively. In contrast, spacers had an average length of 33.22 bp (median=33, SD=6.26), whilst the shortest and longest spacers were 11 and 119 bps, respectively. We proceeded to reduce the redundancy of the spacers using CD-HIT-EST [Fu et al., 2012] and then BLASTN-searched (using parameters defined by previous work [Biswas et al., 2013]) the unique set of spacers against all IMP-based MT-assembled and co-assembled contigs. The parts of contigs that matched to spacers were defined as protospacers, while the contigs were defined as protospacer containing contigs (PSCCs). Despite the removal of redundancy among the spacers and stringent criteria used within the previous BLASTN search, we further clustered the search results by 95% identity and 95% query coverage, followed by filtering contigs that contained repeats to ensure removal of any possible self-matches.

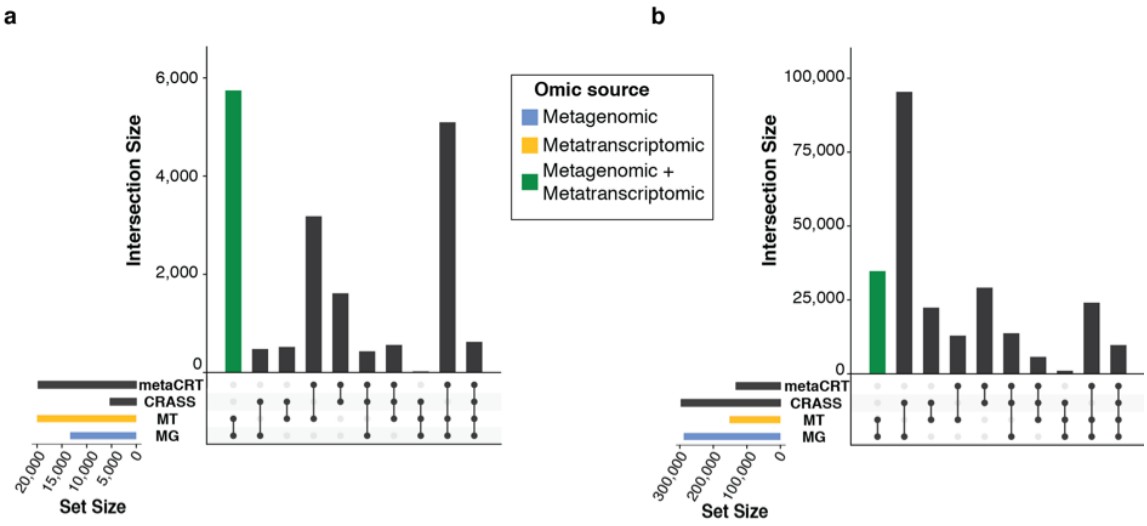


Figure E.2: Prediction of CRISPR elements. **a**, Upset plots representing the **a**, repeats and **b**, spacers predicted in the different omic levels and based on the different tools used. **a** and **b**, the vertical bars represent the intersection between the aforementioned omic levels and/or tools.

E.4 Prediction of invasive mobile genetic elements (iMGEs)

We utilized an ensemble approach to predict putative iMGEs with specifically focusing on plasmids and bacteriophages (phages). The first approach relied on the CRISPR information in the form of spacer-protospacer complements (**Appendix E.3**). Additionally, we used VirSorter [Roux et al., 2015] and VirFinder [Ren et al., 2017] to predict sequences derived from phages. Finally, cBar [Zhou and Xu, 2010] and PlasFlow [Krawczyk et al., 2018] were used to predict plasmid-derived contigs. We merged the results from all the aforementioned methods by assigning annotations to the sequences. A given sequence was annotated as a "plasmid" if it yielded a positive prediction by cBar or PlasFlow. Similarly, a sequence was annotated as "phage" if yielded a positive prediction by either VirSorter or VirFinder. A sequence was annotated as "ambiguous" if it was predicted as both plasmid and phage using any combination of the four aforementioned tools. However, all the aforementioned categories were not necessarily PSCCs because they had to contain at least one protospacer. Therefore, a contig was considered "unclassified" if it was a PSCC, but was not classified as a phage or a plasmid. Thereby, we extracted four classes of iMGEs annotated as either i) phage, ii) plasmid, iii) ambiguous or iv) unclassified. **Figure E.3** summarizes the outcome from the different methods and their classifications. Interestingly, we found that sequences annotated as plasmids (707,093) outnumbered phages (42,039) by around 16-fold. Additionally, 80,617 contigs were found to carry "ambiguous" pre-

dictions. 23,697 (2.86%) of those annotated contigs (i.e. phage, plasmid, ambiguous) contained at least one protospacer. Furthermore, a small number of contigs with protospacers (6,663) were annotated as "unclassified" due to their lack of plasmid and/or phage prediction. Overall, the total number of annotated sequences comprised 6.97% of the IMP-based co-assembled contigs. The redundancy of the identified iMGE sequences were reduced by clustering all the annotated sequences using CD-HIT-EST [Fu et al., 2012]. Upon clustering, the original annotations of the cluster representatives were retained. The non-redundant set of iMGEs retained similar proportions to those in the redundant set, i.e. approximately 17-fold more plasmids than phages. Importantly, a total of 30,360 unique PSCCs were retained for further analysis of iMGEs and the associated host dynamics. **Table 2.2** summarizes the redundant and unique (non-redundant) set of iMGEs that were predicted from the analyses. Finally, the absence of prophage predictions could be explained by the fact that i) VirSorter was the only tool that we applied which was capable of predicting prophage sequences, ii) short assembly contigs, and iii) limited prophage sequences within public databases for this specific environment.

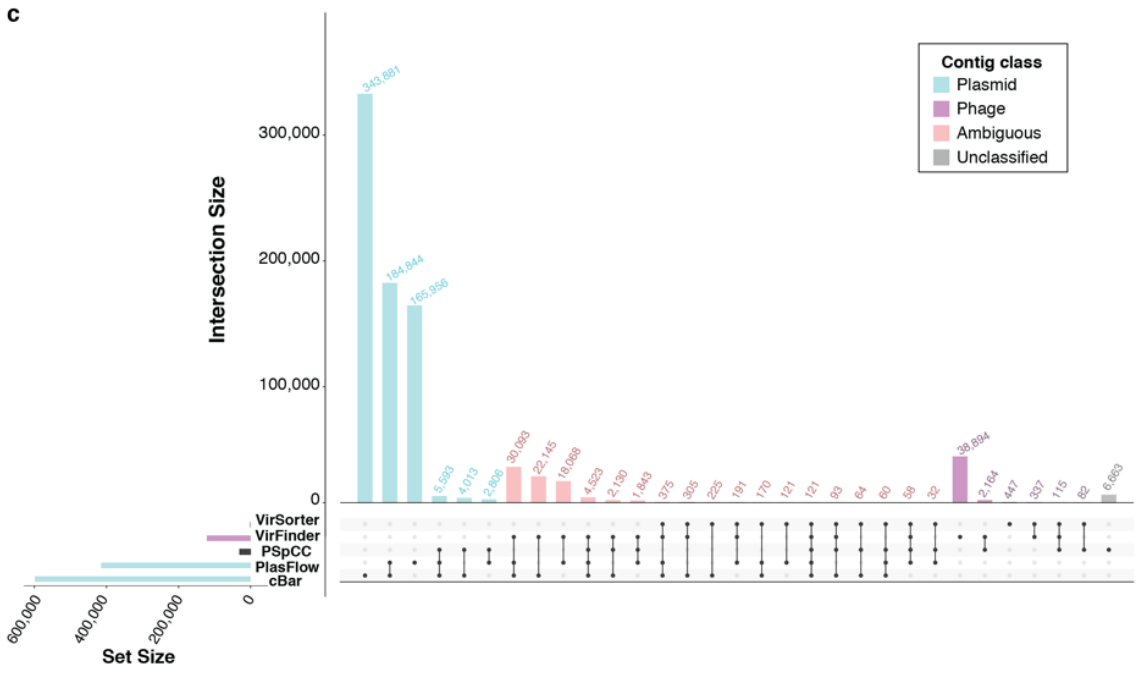


Figure E.3: Prediction of invasive mobile genetic elements (iMGEs). Upset plot representing the prediction of iMGEs using different methods (horizontal bars) and their relevant intersections (vertical bars). The colours of the vertical bars represent the designated annotation after consolidating the prediction of all the different tools.

E.5 Functional analysis of PSCCs

We inspected if CRISPR systems/immunity targeted genes with specific functions [Shmakov et al., 2017; Davison et al., 2016] within iMGEs. For this, we predicted annotated gene functions within both non-redundant plasmids and phages based on KEGG identifiers and linked those identifiers to the corresponding COG categories. The most frequently (top 4) targeted functional categories within plasmids containing protospacers (plasmid-PSCCs) were "Replication, recombination and repair", "General function prediction only", "Transcription" and "Mobilome: prophages, transposons" (**Figure 2.4**). The latter highlighting the potential limitations of phage prediction methods. However, the category of genes most enriched in protospacers was "Nucleotide transport and metabolism", which was not present within the top 4 functional categories of plasmid-PSCCs (**Appendix B.5**). Specific gene functions within this category include "Adenosine deaminase". The most frequently targeted phages (phage-linked PSCCs) also contained the same top 4 categories as plasmid-PSCCs, but differed in their order (**Figure 2.4**). Unlike plasmids-PSCCs, the category of genes containing the highest number of protospacers was "Replication, recombination and repair" and "Mobilome: prophages, transposons", which were the two most frequently targeted categories within phage-PSCCs. Gene functions within these categories included "DNA modification methylase" and "Phage terminase large subunit", respectively. Generally, protospacers were overrepresented within genes that play essential roles in replication, retention and transmission of the corresponding iMGEs. We also inspected the lists for potential depletions in functional categories. In general, we observed that the most depleted (i.e. $< 1\%$ in PSCCs compared to non-PSCCs) categories were "RNA processing and modification", "Extracellular structures", "Secondary metabolites biosynthesis, transport and catabolism" and "Antimicrobial resistance: ResFam". **Appendix B.6** contains the detailed information of the specific gene functions based on their COG categories. Using this broader functional analysis, we did not identify categories depleted specifically when comparing phages to plasmids. Overall, the proportion of targeted genes within the PSCC of phages is higher (30%) compared to their plasmid counterparts (25%), likely due to the lower cargo carrying capacity and dense coding regions of the phage genomes when compared to plasmids [Leclercq et al., 2012].

E.6 Correlation analysis

In general, temporal dynamics were analyzed based on the entire time-series, i.e. 2011-03-21 to 2012-05-03. However, we also inspected the shorter-term temporal dynamics by

manually defining three overlapping shorter-term intervals. These intervals are based on the shift in community structure when the abundance of *Microthrixaceae* family decreases drastically, as a reference point. The intervals also overlapped to ensure that sufficient data points were available for the downstream analyses of localised temporal dynamics. Accordingly, the intervals were defined as: i) before shift: 2011-03-21 to 2011-08-29, ii) during shift: 2011-08-05 to 2011-01-19, and after shift: 2011-12-21 to 2012-05-03. MAGs, plasmids and phages were merged on the family-level (i.e. family-level groups). Plasmids and phages that could not be assigned to any family-level group were collapsed into their own distinct groups (i.e., "plasmid_NA" and "phage_NA"). We calculated the Pearson correlation between the defined bacterial, plasmid and phage groups (all versus all). Next, a hierarchical clustering on the Euclidean distances was applied. This resulted in a total of six clusters. Based on the correlations, we first observed the cluster-level dynamics of the entire time-series. The dominant clusters 2, 3 and 4 were markedly affected by the community shift in the period between 2011-10-05 and 2012-01-11, during which the abundance of cluster 3 increased significantly, while clusters 2 and 4 reduced significantly, corresponding to the drastic reduction in the abundance of *Microthrixaceae*. Cluster 5 exhibited a peak on 2011-08-19, followed by a gradual decrease until the aforementioned community shift (i.e. drop in *Microthrixaceae*). Interestingly, cluster 4 peaked on 2011-09-28, while cluster 2 peaked just prior to the community shift on 2011-10-12 (**Figure 2.7** and **Figure C.2**).

E.7 Linear models

To further investigate the dynamics, we developed linear models based on the dominant *Microthrixaceae* as the representative family (i.e., the response variable), including a random sampling approach for linear model identification (**Section 2.3.12**). Linear models were applied based on the entire time-series and the predefined shorter-term intervals. We assessed the quality of the models with the distribution of the adjusted R^2 values. We observed a bimodal shape with a high number of optimal models and a long tail for non-predictive models (**Figure E.4**). Next, we selected the models with the highest adjusted R^2 values, and inspected these for potential enrichments in specific family-level groups, to select the best models. We observed that plasmids of *Microthrixaceae* were present in 100% of the best models (**Figure E.4**). We subsequently analyzed a globally optimal model with an adjusted R^2 value of 0.9983. In agreement with the enrichment analysis, the plasmids of *Microthrixaceae* and iMGEs assigned at family level, such as *Saprospiraceae*

and *Moraxellaceae*, exhibited significant contributions, while *Cryomorphaceae* plasmids and phages did not exhibit any significant contribution. We subsequently excluded non-significant families from the best global model, which led to a reduced model with an adjusted R^2 value of 0.997 and, compared to the global model, did not exhibit a significant decrease in the adjusted R^2 value (**Appendix B.9**). Overall, the longitudinal abundance data for *Microthrixaceae* exhibits good agreement in those models (**Figure 2.8**). To further validate the observed patterns, we repeated the linear modelling of the shorter-term intervals using the same procedure as for the models of the entire time-series. We obtained R^2 values of 1 in all the global short-term models, with no significant reductions in the R^2 values for the reduced models (**Appendix B.9**). We found that the plasmids of *Microthrixaceae* appeared as the only common significant predictor in all the models (entire time-series and short-term intervals). To further assess the relative importance of this group as main predictor of the *Microthrixaceae* family abundance dynamics, we specifically investigated the dependence of the model qualities with respect to *Microthrixaceae* plasmids (**Appendix B.10**). For this purpose, we excluded *Microthrixaceae* plasmids from the respective reduced models, which led to a reduction in predictive power in all models, especially in the short-term model before (from R^2 =0.99 to R^2 =0.21) the community shift. In contrast to the dynamics before and during the community shift, we found that the shorter-term models highlighted *Microthrixaceae* and *Moraxellaceae* phages as significant predictors after the community shift (**Figure C.3**) and **Figure C.4**. Overall, the longitudinal abundance data for *Microthrixaceae* is in good agreement with the global and reduced models, but less so when removing *Microthrixaceae* plasmids as predictors. Thereby, plasmids exhibited stronger effect on the prediction of *Microthrixaceae* abundances compared to phages. This in turn indicates a higher relative importance for plasmids in governing the *Microthrixaceae* dynamics.

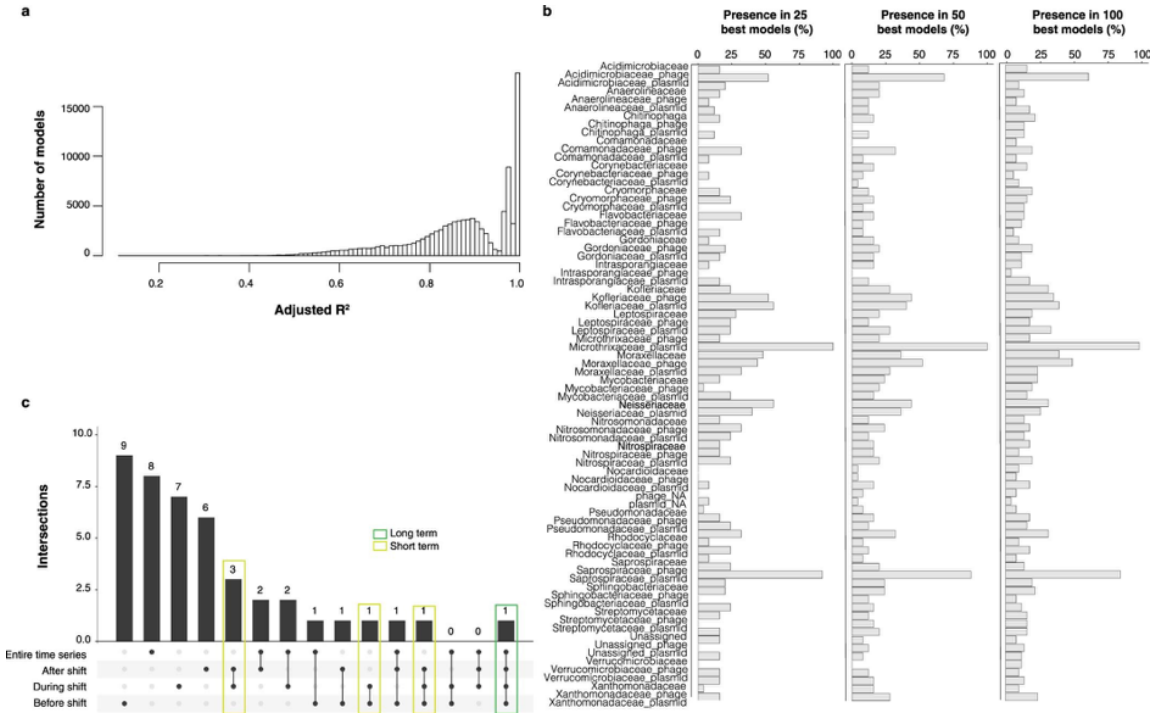


Figure E.4: Model fitness and family enrichment within the best models predicting Microthrixaceae family abundance. a, Distribution of the adjusted R^2 values of 100,000 model realizations. **b**, Enrichment of the family-level bacterial, plasmid and phage groups in the best 25, 50 and 100 models of the entire time-series. **c**, The upset plot represents the number of family-level bacterial, plasmid and phage groups (that is features) within the best model of different time intervals, that is the entire time-series and three time-windows (horizontal bars). The number of intersections between features in the best models in different long- and short-time intervals (vertical bars). The coloured boxes represent the intersections representing short- and long- term time dynamics, respectively.

E.8 iMGE-CRISPR-host based networks

The constructed iMGE-host networks are bipartite networks, i.e. there are two groups of nodes (rMAGs and iMGEs), where the considered interactions occur between elements of the different groups and not between elements of the same group (rMAG-iMGEs). Properties of bipartite networks include modularity and nestedness. Modularity (Q) relates to the connectivity between different groups [Newman, 2006], i.e. in this study it reflects connectivity between groups of iMGEs (phage or plasmid) and groups of hosts. On the other hand, nestedness is given as the value of the "Nestedness matrix based on Overlap and Decreasing Fill" (NODF) and represents the measure of structure in an ecological system, i.e. in this study it reflects the iMGE host range specificity [Koskella and Meaden, 2013]. The global phage-host network had a modularity Q of 0.7 (out of 1) and a nestedness NODF of 1.0 (out of 100), with similar values observed in the timepoint-specific

networks (**Appendix B.12**). Similarly, the global plasmid-host network had a modularity and nestedness of Q=0.77 and NODF=3.43, respectively. These network properties indicate high compartmentalization, i.e. specific groups of iMGEs interact with specific hosts, and a restricted host range. However and interestingly, the host range of plasmids is broader than for phages (**Figure 2.10**). Finally, it is important to note that these properties were exclusive for iMGE-host interactions via CRISPR, compared to networks including all interacting MGEs, not limited to those interacting via CRISPRs but also using other mechanisms.

E.9 The CRISPR-Cas genes of *Candidatus* Microthrix parvicella Bio-17

A complete *Candidatus* Microthrix parvicella CRISPR operon was also detected within a single contig of 10,224 bp (D47_L1.43.1_contig_476300). Specifically, the CRISPR operon contained i) CRISPR-associated endonuclease Cas1, ii) CRISPR-associated endonuclease Cas2, iii) CRISPR-associated endonuclease/helicase Cas3, iv) CRISPR-associated protein Cas7, v) CRISPR-associated proteins Cas8, vi) csb2gr5 and vii) a CRISPR locus with 11 repeats similar to those encoded by the *Ca*. M. parvicella Bio17-1 genome [Muller et al., 2012]. This combination of *cas* genes defined the CRISPR-Cas system as a type I and subtype I-U, whose signature genes are cas3HD and cas8u, respectively [Makarova and Koonin, 2015; Koonin et al., 2017]. Cas1 and Cas2 are universal proteins involved in CRISPR-spacer insertion. Furthermore, Cas7, Cas3HD and Cas8u1 are all involved in the interference step of CRISPR-based immunity, while the function of Csb2gr5 re-mains unclear [Makarova and Koonin, 2015; Koonin et al., 2017]. To ensure accuracy of the predicted CRISPR operon, we processed the genomes of *Candidatus* Microthrix parvicella Bio-17 and the contig containing the CRISPR operon with CRISPRCasFinder [Couvin et al., 2018], where we further confirmed a highly similar CRISPR operon (**Ap-pendix B.14**). The *cas* genes were found to be expressed at both the MT- and MP-levels while we observed spacer gain and loss events within the CRISPR locus during the time-series, which points towards an active CRISPR system (**Figure 2.11**, **Figure 2.12**, **Fig-ure 2.13**). Specifically, we detected 31 spacer gain events for which iMGE sequences were detected before their linked spacers, 9 spacer gain events for which iMGE sequences were detected at the same time as the spacers, and 5 spacer gain events for which the iMGE sequences were detected after their linked spacers.

Finally, the average lag time of an integration event (i.e. time between detection of iMGE

and detection of spacer) was 6 weeks (median=1, SD=12) for spacers targeting plasmid sequences and 4 weeks (median=1, SD=7) for spacers targeting phage sequences.

E.10 Contrasting *Candidatus* Microthrix parvicella's spacers and iMGEs with other populations

The *Ca*. M. parvicella-like rMAG-165 clearly demonstrated activity of its CRISPR system in terms of gene and protein expression, as well as spacer gain and loss activity (**Appendix E.9**). However, we were unable to assess the magnitude of the CRISPR system's activity in terms of iMGE targeting. Therefore, we performed an additional assessment to contrast CRISPR system activity of *Ca*. M. parvicella with other populations. Accordingly, we sourced additional rMAGs that encoded complete CRISPR systems by fulfilling the following criteria: i) CRISPR locus, ii) a set of *cas* genes, iii) CRISPR system type prediction and iv) occurrence within a single contig. Subsequently, we found rMAG-31 and rMAG-40, classified as *Intrasporangium calvum* and *Leptospira biflexi*, respectively, as suitable candidates for this assessment. rMAG-31 (*I. calvum*) encoded a type I CRISPR operon which in turn encoded seven *cas* genes and one CRISPR locus encoded on a single contig of 24,304 bp. However, the cas genes and Cas proteins were found to be lowly expressed, relative to those of *Ca*. M. parvicella. Its CRISPR locus contained 129 spacers, with only 7 spacers targeting iMGE sequences within the timeseries, exclusively plasmid sequences. Spacer integration events were lower when compared to *Ca*. M. parvicella with only one spacer gain event occurring in 18 weeks (**Appendix B.11**, **Appendix B.15**). MAG-40 (*L. biflexi*) encoded a type V CRISPR operon which contained four *cas* genes (**Figure 2.15**) and one CRISPR locus which was encoded on a single contig of 12,586 bp. The expression of *cas* genes was found to be comparable to *Ca*. M. parvicella, but the expression of Cas proteins was lower. Conversely, spacer integration events were significantly more frequent compared to *Ca*. M. parvicella, with rMAG-40 exhibiting acquisition of both plasmid- and phage-derived spacers alike. Plasmid-based spacer integration was the most prevalent (**Appendix B.15**). The average time for the integration of spacers targeting plasmids was 12 weeks (median=6, SD=15), while for spacers targeting phages was 11 weeks (median=5, SD=14) (**Appendix B.11**). In summary, we compared the CRISPR system activity of *Ca*. M. parvicella with other rMAGs, and observed that rMAG-31 (*I. calvum*) had lower activity compared *Ca*. M. parvicella. On the other hand, rMAG-40

demonstrated lower levels of CRISPR system activity when compared to *Ca*. M. parvicella, in terms of gene and protein expression, despite demonstrating enhanced spacer integration dynamics, both in terms of frequency (i.e. absolute number) and breadth (i.e. types of potential iMGEs targeted). In general, we show that different population-level CRISPR-Cas dynamics exist at the level of gene and protein expression as well as spacer integration activity. Based on our results, *Ca*. M. parvicella populations do contain a functional CRISPR system, but uses it rather sparingly compared to other population such as rMAG-40 (*L. biflexi*).