PhD-FSTM-2024-010

The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 21/02/2024 in Esch-sur-Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Oskar HICKL

Born on 1 April 1993 in Hamburg, (Germany)

Integrative Multi-omics Analysis Methods of Microbiomes

Dissertation defence committee

Dr. Patrick May, dissertation supervisor

Senior scientist, University of Luxembourg, Luxembourg

Dr. Alexander Skupin, Chairman

Professor, University of Luxembourg, Luxembourg

Dr. Paul Wilmes, Vice Chairman

Professor, University of Luxembourg, Luxembourg

Dr. Thilo Muth

Group Leader, Robert Koch Institute, Germany

Dr. Nicola Segata

Professor, University of Trento, Italy

Affidavit

I hereby confirm that the PhD thesis entitled 'Integrative Multi-omics Analysis Methods of Microbiomes' has been written independently and without any other sources than those cited.

Luxembourg, ____________________ ____________________

Name

Acknowledgements

I would like to extend my sincere thanks to my family and friends for their support during the years of study and the preparation of this thesis. I am very grateful to my family for their steady encouragement and support and ensuring that I always had someone to turn to. I would also like to thank my friends, for their companionship and support during this process. Their presence and assistance have been valuable in various ways. Additionally, I am thankful for the support from my colleagues, who have provided a productive and engaging environment during my studies. Finally, I want to thank my girlfriend for always being there with affirmation, aid and comfort in times of need. Completing this thesis has been a significant task, and the assistance I have received from everyone has played an important role in this achievement.

Contents

1. List of Publications

- Published (part of this thesis)

  - Hickl *et al.*, “binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets,” *Briefings in Bioinformatics*, vol. 23, no. 6, 2022, pp. 1–14. [Online]. Available: https://doi.org/10.1093/bib/bbac431

  - Queirós *et al.*, “Mantis: flexible and consensus-driven genome annotation,” *GigaScience*, vol. 10, 2021, pp. 1–14. [Online]. Available: https://doi.org/10.1093/gigascience/giab042

  - Hickl, Kunath *et al.*, “Alterations of oral microbiota and impact on the gut microbiome in type 1 diabetes mellitus revealed by integrated multi-omic analyses,” *Microbiome*, vol. 10, 2022, p. 243. [Online]. Available: https://doi.org/10.1186/s40168-022-01435-4

- Unpublished (available as preprint)

  - Queirós *et al.*, “UniFuncNet: a flexible network annotation framework,” *bioRxiv*, 2022.03.15.484380. [Online]. Available: https://doi.org/10.1101/2022.03.15.484380

  - Wilmes, Trezzi, Aho *et al.*, “Artefactual source of 2-hydroxypyridine,“ *Research Square* . [Online]. Available: https://doi.org/10.21203/rs.3.rs-1827631/v2

  - Trezzi, Aho *et al.*, “An archaeal compound as a driver of Parkinson’s disease pathogenesis,“ *Research Square* . [Online]. Available: https://doi.org/10.21203/rs.3.rs-1827631/v1

List of Abbreviations

ACSS	Acyl-coenzyme A synthetase short-chain family member
AD	Alzheimer's Disease
AI	Artificial Intelligence
AMP	Antimicrobial Peptide
BBB	Blood-Brain Barrier
DNA	Deoxyribonucleic Acid
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
EM	Expectation–Maximization
FFAR	Free fatty Acid Receptor
GC	Guanosin-Cytosin
GDM	Gestational Diabetes Mellitus
GPR	G Protein-coupled Receptor
HDAC	Histone Deacetylase
HMM	Hidden Markov Model
HQ	High-Quality
IBD	Inflammatory Bowel Disease
IL	Interleukin
IMP	Integrated Meta-omic Pipeline
KOfam	Kyoto Encyclopedia of Genes and Genomes (KEGG) protein families
LCA	Lowest Common Ancestor
LPS	Lipopolysaccharide
MAG	Metagenome-Assembled Genome
MIMAG	Minimum Information about a Metagenome-Assembled Genome
NC	Near-Complete
NGS	Next Generation Sequencing
NPFM	NCBI Protein Family Models
NRF2/NFE2L2	Nuclear factor erythroid 2-related factor 2
ORF	Open Reading Frame
PD	Parkinson's Disease
PFA	Protein Functional Annotation
PP	Peyer's Patch
PRR	Pattern Recognition Receptor
Pfam	Protein domain families
RNA	Ribonucleic Acid
SCFA	Short-Chain Fatty Acid
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
TCA	Tricarboxylic Acid Cycle
TIGRfam	The Institute for Genomic Research's database of protein families

2. Abstract

In recent years, our understanding of the pivotal role played by microbiota in shaping various environments has significantly expanded. Particularly, the microbial communities residing on and within humans hold great importance not just for human health but also for comprehending the intricacies of complex biological systems in general. As microbial ecology evolves, meta-omics techniques have solidified their position as indispensable tools for probing such biological niches. However, given the intricate community structures of these environments, there is a need for high-performance analytical methods to extract, process, and make sense of the vast amounts of information they contain.

This thesis advances the development and utilization of integrated multi meta-omics approaches aimed at enhancing our understanding of microbial ecology, with a primary focus on the human intestinal tract. Two novel tools were introduced: binny and Mantis. binny allows for the recovery of high-quality genomes directly from metagenomic datasets. This is instrumental in obtaining the actual members of microbial communities that shape their environment. Mantis offers a flexible platform for high-quality functional annotations of genomic data through a consensus approach.

Furthermore, through the integration of matched metagenomes, metatranscriptomes, and metaproteomes from stool and saliva samples, this work provides a linkage between oral and gut microbiota using multiple levels of evidence. This intersection provides unique insights, especially in the context of diseases. It contributes to bridging the crucial gap in our understanding of how human-associated microbial communities interact and influence the host's health.

In conclusion, the methods and findings presented in this thesis contribute to the field of microbial ecology and help shedding light on the intricate relationships between human-associated microbiota and health.

3. Aims and Objectives

This research endeavors to explore the microbial communities within the human body, with a focus on their roles and potential influences, particularly in the context of health and disease. The specific aims and objectives are outlined as follows:

1. **Understanding microbial communities:** The thesis tries to extend the understanding of microbial communities in humans, focusing on both intra-community and host-microbe interactions. Key for this would be investigating how functional capacities of microbes shape microbial communities through mechanisms of e.g. competition, cooperation, nutrient acquisition.

2. **Developing and utilizing methods to use with omics technologies:** Recognizing the complexities of microbial ecology, there is a need to use and develop high-performance methods to allow the study of complex environmental samples. These methods should contribute to the tools available for probing these communities and enhance our capabilities of understanding them. Especially, reproducibility and scalability in probing function, taxonomy, and the principal actors in microbial communities is the aim.

3. **Deciphering community dynamics in human disease:** The methodologies developed will be applied to a multi-meta omics dataset from a case-control study. Leveraging the multiple levels of information will allow deep insights into the microbiota and host-microbiota dynamics in the context of disease. Drawing from the principal biomolecules, DNA, RNA, and proteins, information about the functional potential, expression, and biological activity of the microbiota in the host environment will be gained and the complementary of this information leveraged to gain a more comprehensive understanding of actors and interactions relevant to a disease. This will be achieved by taxonomic identification, in combination with knowledge of regulation and activity, of microbes of interest. Finally, understanding of the possible interactions between communities in different, connected body sites, here the oral and gut microbiota in the context of disease, might provide insights into dynamics of multiple host-associated communities and how this impacts host health.

In essence, this thesis hopes to contribute to the broader understanding of microbial ecology and the relationships between human-associated microbiota and health.

4. Synopsis

4.1 Introduction

Microorganisms exhibit a substantial level of diversity. This allows them to inhabit almost every environment on Earth, ranging from extreme environments like arctic landscapes or hydro-thermal vents to oceans, lakes, vast amounts of different soils, and in and on animals [1, 2, 3]. Accordingly, they play key roles in Earth's ecosystems, driving almost all biogeochemical processes [2, 4].

To understand what role microorganisms play and how, harnessing high-resolution information is indispensable. It is essential to understand not only who is doing what but also the context in which these activities occur. To that end, the vast array of interactions between organisms and environments needs to be disentangled to understand how microorganisms with individually limited capabilities, shape every environment on Earth. This cannot be achieved without learning what genetic information these microorganisms harbor and what activities they perform using this information. Key is gathering and interpreting data from various 'omes' – including the metagenome, metatranscriptome, and metaproteome – from an environmental sample, as they provide the essential information on functional potential, expression and biological activity, all of which are essential in answering the aforementioned fundamental questions [2, 5, 6].

While understanding fundamental ecological questions elucidates the formation of the world humans inhabit, its significance extends profoundly to human lives directly and tangibly. It touches intimately on key matters of health and disease [7], development [8, 9], environmental stability [10, 11], as well as food security [10, 11, 12], and many processes in industry [13, 14, 15, 16, 17].

4.2 Microbial ecology and how microbial life shapes Earth

Understanding the dynamics of microbial communities forms the foundation of microbial ecology. Their constituents' interactions shape, and are shaped by, not only their environment but also form a balance of survival and reproduction within the microbial community itself. Without an understanding of these interactions, it is impossible to understand the emergent properties of ecosystems and also the biology of most life forms themselves, since they always evolved in the context of coexistence with other life [18].

Members of the various microbial communities inhabiting virtually all environments on Earth evolved capabilities designed to function within the framework of their respective communities and take part in intricate and continuous processes of resources and information exchange. This can be clearly seen in microbial communities playing key roles in shaping the planet's fundamental biochemical processes such as carbon, nitrogen, and sulfur cycling, which all are multi-stage, multi-organism processes [4]:

The microbial carbon cycle plays a pivotal role in supporting life on Earth, with its central process, primary production, being the starting point for most of Earth's food webs. Through photosynthesis, microorganisms such as cyanobacteria and algae, together with plants, convert carbon dioxide into organic compounds, initiating the flow of energy and carbon through ecosystems. These organic compounds form the basis of nutrition for a wide range of organisms, making primary production essential for the sustenance of both aquatic and terrestrial food networks . Beyond its biological importance, the carbon cycle regulates Earth's climate. Microbial processes of decomposition and respiration are responsible for the release of carbon dioxide and methane, two major greenhouse gases, back into the atmosphere. These gases play a significant role in Earth's heat retention and global temperature regulation [4, 19, 20, 21, 22, 23, 24].

The nitrogen cycle, mediated mainly by microbial activities, is fundamental to life due to nitrogen's essential role in biological molecules. Nitrogen is a key component of amino acids, the building blocks of proteins, which are needed for the function of all living organisms. Furthermore, nitrogen is a vital constituent of nucleic acids, such as DNA and RNA, central to genetic information storage and transmission. The transformation of nitrogen into biologically accessible forms, primarily facilitated by nitrifying bacteria, is therefore critical for the synthesis of these fundamental components of life. Nitrogen availability is often the limiting factor in ecosystem productivity, making the nitrogen cycle a key determinant in the growth and survival of organisms. The process of denitrification, wherein denitrifying bacteria convert nitrate back to molecular nitrogen, plays a dual role in maintaining ecosystem nitrogen balance and in regulating atmospheric composition by mitigating the accumulation of nitrous oxide, a potent greenhouse gas [4, 25, 24].

Finally, the ubiquitousness of microbial communities can be illustrated by different extremophiles: Thermophiles and psychrophiles cope with heat and cold, respectively, by ensuring protein stability and cell membrane integrity are maintained expressing genes for specialized proteins and varying cell membrane composition. Halophiles possess specialized mechanisms to accumulate organic solutes to stabilize cellular components [3].

In conclusion, the importance of microbial communities in shaping Earth's environments through their involvement in fundamental biochemical processes cannot be overstated. Their intricate interactions with their surroundings and with each other play a central role in maintaining the balance of life on Earth, influencing everything from nutrient cycles to global climate patterns.

To fully comprehend complex and dynamic systems like these, it is essential to investigate microorganisms in the context of their environments and extract relevant information in a comprehensive and holistic manner. This approach is crucial for understanding not just the individual organisms, but the entire ecosystem in which they function. Fundamental questions that need to be answered are: Who is present in the community? What potential functions do they possess? And most importantly, what are they actually doing in a given environmental context? Answering these questions requires tools capable of investigating the principal biomolecules involved. 'Omics technologies provide this capability as they allow their characterization and interpretation.

4.3 Meta-omics: Unraveling the who and how in microbial communities

The advancement of omics technologies, driven by high-throughput sequencing and mass spectrometry, has dramatically transformed our understanding of microbial communities. These methods provide comprehensive insights into the genetic and functional diversity within various environments based on nucleic acids, peptides and proteins, and metabolites, crucial for studying microbial ecosystems [26, 27] (Figure 4.1).

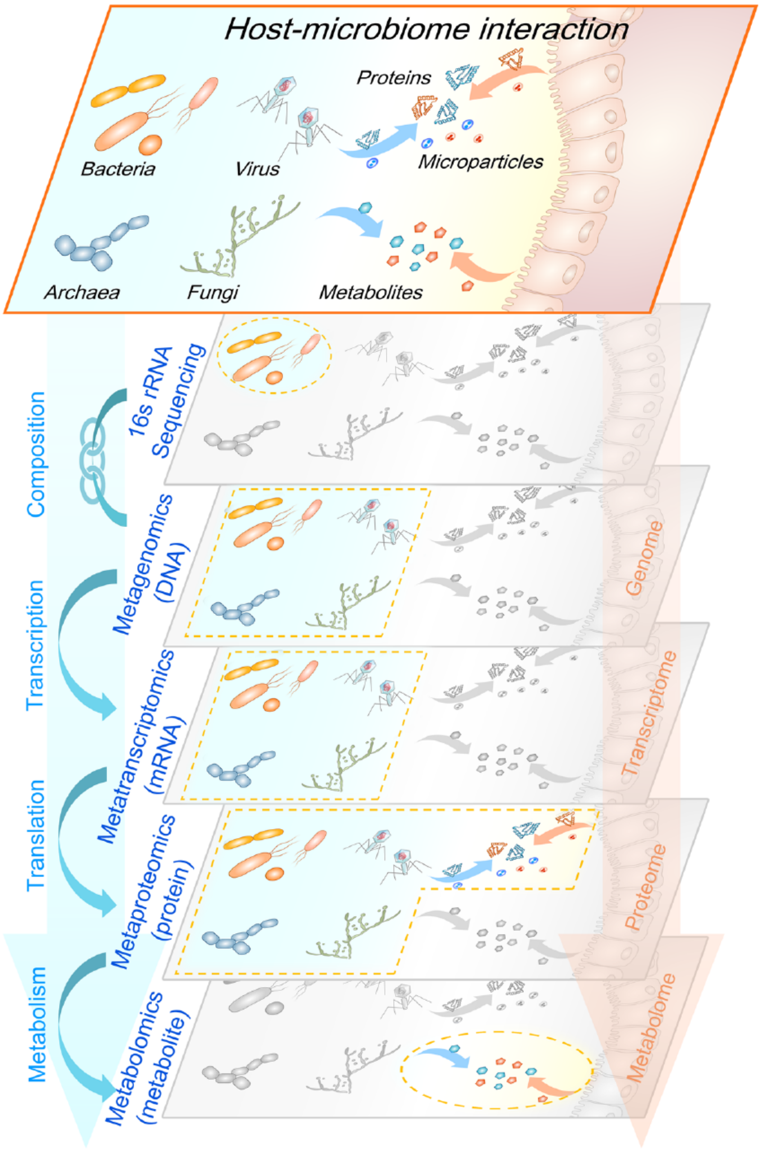


Figure 4.1: **Meta-omics methods to study microbial communities.** Different approaches provide different types of complementary information. Modified under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/) from Figure 1, Zhang *et al*, Advancing functional and translational microbiome research using meta-omics approaches. Microbiome 7, 154 (2019) [28].

4.3.1 Metagenomics

Metagenomics enables the recovery and analysis of genomic information directly from environmental samples, bypassing the need for traditional culturing methods. The transition from first-generation Sanger sequencing to second/next-generation sequencing (NGS) [29] and third-generation long-read sequencing has been critical in this field [30]. Methods that sequence only identity-conferring sequences, such as parts of 16s rRNA genes, provide an overview of the organisms in an environmental sample at usually intermediate resolution and also allow estimation of relative abundances. In contrast, whole-genome sequencing methods provide crucial additional information about function and regulation [31, 32]. The new generation of sequencing technologies have brought about more affordable, high-throughput sequencing capabilities with longer reads and fewer errors, essential for the accurate assembly of genetic material from single organisms [33] to diverse microbial populations [34]. The process involves assembling raw sequencing reads into more contiguous genome parts, a task complicated by genomic complexity from closely related organisms, uneven coverage of genomes, and incomplete information from sampling only a subset of the genetic information present in a sample [35, 36]. Recent advancements in assembly methods are trying to address these challenges, improving the ability to accurately reconstruct genomes from environmental DNA [35, 37, 38].

4.3.2 Metatranscriptomics

Metatranscriptomics allows analyzing RNA transcripts to determine active gene expression in microbial communities from environmental samples. It offers insights into a community's functional state under specific conditions. Transcriptomes encompass a wide array of RNA molecules, from those essential for protein synthesis (mRNAs, rRNAs, tRNAs) to an array of non-coding RNAs (sRNAs, miRNAs, lncRNAs, circRNA), each playing crucial roles in diverse regulatory mechanisms. The advent of NGS has also been transformative in metatranscriptomics, facilitating the high-throughput identification and quantification of these varied RNA types. It enabled the analysis of microbial community gene expression and regulation in response to environmental stimuli such as their interaction with host organisms [27, 39]. Assembly of RNA-Seq data suffers from similar problems as DNA-Seq, with the added hurdle of having to deal with inherently uneven read coverage depth from varying gene expression, which is exacerbated by the activities of potentially large numbers of organisms [40].

4.3.3 Metaproteomics

Proteins serve as direct indicators of microbial activity, elucidating the functions actually being performed within a community. The identification and quantification of proteins allows not only the measurement of activity in the most direct way but also to investigate other characteristics of great importance to understand community structure, such as estimating biomass, and, through labeling techniques, identification of key element's flows such as carbon or nitrogen [41, 42]. The advancements in mass spectrometry and chromatography have enabled detailed analysis of complex protein mixtures from environmental samples. Identifying these proteins remains a challenge due to the diversity and complexity inherent in microbial communities [43, 42]. However, improvements in database search engines and spectrum identification strategies are progressively enhancing the precision and depth of metaproteomic analyses [44, 45].

4.3.4 Metabolomics

The examination of the metabolic landscape of microbial communities is another way of probing direct activity by attempting to identify the mostly small molecules produced and consumed in the metabolism of organisms. Utilizing mass spectrometry techniques, metabolomics systematically aims to characterize a broad spectrum of these small molecules, each of which is defined by unique structural and functional attributes.. This task is complicated by the extreme chemical diversity and large number of unknowns of metabolites. Targeted strategies try to attenuate this by focusing only on a manageable subset of molecules with known properties. The strength of metabolomics lies in its ability to identify and quantify biochemical pathways and metabolic fluxes to analyze an an organism's or community's physiology. This aids in deciphering the functional dynamics of microbial communities, contributing to the discovery of their members ecological roles and interactions [27, 46].

4.3.5 Recovering Metagenome-Assembled Genomes

High-resolution meta-omics data provides a comprehensive view of microbial communities, but understanding the dynamics of these communities necessitates discerning the individual microbial entities and their functionalities. Recovering complete and uncontaminated genomes, known as Metagenome-Assembled Genomes (MAGs), from these communities is pivotal. Each microbial entity contributes distinct functional potentials and activities to the community. By isolating and understanding their genomes, it can be elucidated how the combination of single functions or activities lead to complex community phenotypes. This understanding is crucial not only for deciphering microbial interactions but also for tracing the impact of specific microbial entities on the community's collective functions [47, 48].

However, the recovery of MAGs from metagenomics data, also called 'binning', is a complex task, hindered by challenges such as limited sequencing depth, assembly fragmentation, repeat elements, and the ambiguity caused by closely related genomes [49, 47, 48].

To group sequence fragments (contigs) belonging to the same organism together, mainly two sources of distinguishing information are used:*k*-mer frequencies and abundance profiles. The former tries to exploit potentially unique signatures in the frequency of DNA sequence sub-strings, most commonly of length 4, the latter the (average) read depth of coverage of contigs, which has been shown to be correlated between contigs assembled from reads of the same organism, since, during sequencing, DNA fragments of the same clonal line should share an abundance pattern [47, 50, 51, 52].

Manual binning, while feasible for small datasets, is limited by human capability to interpret processed data patterns and its analysis speed [53]. Thus, it is not widely applied today to analyze large data sets. Instead automated binning algorithms have been developed that make use of the aforementioned features to group contigs into MAGs. A wide variety of clustering algorithms with different data pre- and post-processing strategies have been proposed. Notable and widely used binning tools that have made significant contributions to the field are CONCOCT [50], which uses Gaussian mixture models and ideally information from multiple (related) samples, MaxBin2 [51], designed to work with co-assemblies and employing an Expectation–Maximization (EM) algorithm, and MetaBAT2 [52], which uses graph-based clustering and extensive parameter fine-tuning.

Still, many challenges remain unsolved: The effective use of *k*-mer frequencies is challenging due to the resulting high-dimensional data, which is problematic for clustering algorithms since distance metrics become increasingly less useful [54]. Read depth of contig coverage only works well if there is sufficient coverage to create a sufficient signal and because of the stochastic nature of capturing genomic data during DNA pre-processing for sequencing and the process itself, often only the most abundant organisms and/or the ones with the largest genomes get covered well [47, 55].

Determining the completeness and purity of a MAG is not straightforward since no 'gold-standard' exists, reference catalogs are incomplete and even if an organism has a representative in a database, genome plasticity, especially in prokaryotes, can be high [56]. Potential solutions include the use of single-copy marker genes [57, 58], deep learning approaches [59], and manual inspection [60]. However, next-generation tools like binny [61], SemiBin [62], and MetaDecoder [63] are now emerging, offering more complete and pure genome recovery, even distinguishing closely related strains. binny, featured in this work, achieves high performance by applying a dynamic and robust clustering approach with quality control using single-copy marker gene sets wrapped in an iterative, adaptive binning procedure.

In the future, MAG recovery will likely achieve close to complete recovery of genomes in a sample with the discovery of more sophisticated features from sequence data [64, 65], the integration of long-read sequencing technologies [38] and advanced wet-lab pre-separation methods [66]. Large biomolecule sequence language models [67] akin to language models [68, 67] trained on vast amounts of human written text and used in e.g. chatbots[69], could exploit a feature space too complex for human comprehension, potentially revolutionizing the field of metagenomics and microbial ecology [70].

4.3.6 Annotating meta-omes

Regardless of whether obtained genomic sequences have been binned into MAGs or not, it is essential to annotate these sequences to derive relevant information about their function, organism of origin and relation to other sequences or genomes. The identity in form of a taxonomic label or phylogeny and functional potential in form of expressible genes, regulatory and structural elements serve as the key framework for interpreting the vast data from various 'omes' [71, 72]. It is paramount that these annotations are both trustworthy and comprehensive. Incomplete or erroneous annotations could lead to misinterpretations, consequently skewing biological insights.

Taxonomic annotation

Taxonomic annotation plays a crucial role in meta-omic studies as it provides, usually based on metagenomics data, the foundation for understanding the composition and diversity of microbial communities in various environments. This process involves identifying and classifying organisms present in a sample, which is essential for deciphering their potential functions and interactions, especially in the absence of direct functional information. Well established taxonomic systems with ever-growing databases of annotated genomes are available to use as reference source when trying to identify organisms in environmental samples [73, 74, 75, 76].

One of the primary challenges in the context of attributing taxonomy is the need to also classify organisms not well represented in current reference sources. This can be the case for currently unculturable organisms or those with high genome plasticity. While matching sequences to databases, based on e.g. sequence similarity [77, 78, 79], is a powerful way to achieve potentially very precise classification, the advent of meta-omics with large scale sequencing of environmental samples has shown that there is a significant amount of organisms to be studied which are not or only poorly described in current reference sources [80, 81, 76]. For these, techniques solely relying on having a similar representative available in a database will fail. Instead, it is necessary to employ strategies which allow classification with little information (i.e. only fragments of a genome available) that can generalize well.

Advancements in k-mer based strategies, which extract sub-strings of a given size from already annotated genomes and usually use them to build an index of distinguishing features for a taxon, have significantly improved in terms of specificity, recall, and computational efficiency. Since it is now only necessary to match a fraction of sequence, and often a lowest common ancestor (LCA) scheme is applied, to at least label a sequence at a higher taxonomic level, if no unique (e.g. species-level) k-mers are found, the classification of large amounts of metagenomics/-transcriptomics sequences with much higher speed, precision, and recall is possible [82, 83, 84, 85].

Another approach, enabled especially by the wealth of metagenomic data generated in recent years, is taxonomic profiling using marker genes. Tools such as mOTUs [86] and MetaPhlAn [87] allow the estimation of the abundance of organisms contained in metagenomes. To achieve this, genomes are grouped by sequence similarity or taxonomic label and sets of markers uniquely distinguishing these groups are derived. This approach usually allows for fast processing and the data needed is light-weight and the estimates from recent version are mostly quite accurate [49]. The main drawback is that no per-sequence labels are produced, so different features, such as functions on metagenomic sequences, cannot be linked.

As mentioned earlier for the recovery of MAGs, in the future, deep learning approaches that leverage a potentially much more complex feature space might offer substantially more powerful generalization capabilities to classify novel or poorly described organisms while maintaining high precision [88, 89].

In summary, taxonomic annotation can provide deep insights into the structure of microbial communities and how they relate to their environments, but is both conceptually and computationally demanding with vast amounts of sequence data gathered still lacking attribution to an organism. Still, there is massive potential for synergy to expand the knowledge of the biology of microorganisms with other information, chiefly functions encoded on sequences.

Functional genome annotation

Deciphering the functional capabilities of microbial communities is foundational to microbial ecology. It enables understanding the functional potential of organisms and communities, or measuring activity. To obtain a functional description for a region on a sequence, first expressible genes, transcribed non-coding regions and regulatory elements need to be identified [72]. This is a complex task, especially for data from environmental samples since many of the organisms contained might not adhere closely to gene and/or regulatory sequence organisation known from the few well studied model organisms available [90, 91, 92]. Various methods attempt to automate these tasks, as data volumes have made manual annotation impossible. Gene calling tools try to identify sequence regions representing genes and the translated regions within called Open Reading Frames (ORFs) using features such as promoters, start/stop codons, ribosomal binding sites, and GC content, sometimes in combination with conserved motifs in form of Hidden Markov models (HMMs) [93, 94, 95]. Examples of other methods are: the identification of ribosomal RNA genes using HMMs [96] , tRNA genes based on combinations of specific and unique motives [97] or structural RNAs using a combination of HMM guided methods and covariance models [98].

Information gained by these predictions can in turn be used to infer function by querying reference sources almost exclusively relying on similarity to experimentally characterized biomolecules, like proteins. These databases have various resolutions and scopes, from e.g. protein families and single proteins to protein domains and sites [99, 100, 101, 102].

Like taxonomic annotation, with the influx of vast amounts of sequencing data from environmental samples, there is a dire need for methods to elucidate function of non-model, uncultured organisms. There are advancements producing larger and larger gene catalogues, which contain significant amounts of genes which lack homologues in reference sources and thus their precise function remains unknown. Still, one can infer their role based on synteny or structural similarity to other proteins, even when there is no substantial sequence similarity. [103, 104].

While diverse array of large-scale reference sources is now available, a significant impediment in protein function annotation remains the capacity to make effective use of this wealth of information. Mostly researchers rely on a singular (or few) source(s), which might be incomplete, of varying quality or insufficient scope/resolution. Mantis, featured in this work, tries to address this challenge by utilizing database identifier intersections in conjunction with text mining. This approach facilitates the integration of information from different references into a unified, consensus-driven output. The ensuing annotations are more exhaustive and minimize bias or error from a single source while leveraging the potential synergies of different resources [105].

4.3.7 Summary

The study of microbial communities has seen rapid advancements through new experimental and computational methods. These methods aim to understand the identity, interplay, and capabilities of community members. Despite these advancements, researchers still face challenges in grappling with the scope and complexity of the interactions and the vast amounts of information to be processed and interpreted. It is also clear that improving the methods to investigate microbial communities as well as continuing efforts to study them in their various environments is paramount to understand life on Earth. This involves more than just understanding the numerous ways in which they influence biotic and abiotic systems on a large scale on Earth. Using omics methods, researchers aim to elucidate which organisms and interactions shape human life and which activities are responsible. Ultimately, this knowledge could be pivotal in curing or preventing diseases, and improving quality of life.

4.4 Microbiomes in human health and disease

Microorganisms exert considerable influence on human development and health, with a profound capacity to both support wellness and drive disease. They impact key functions, including metabolism and immune system function. [106, 107, 108, 109, 110].

The interplay of microbial and human metabolism primarily involves the breakdown and assimilation of nutrients [111]. In particular, the gut microbiota plays a crucial role in decomposing complex carbohydrates, which are indigestible by humans alone [107, 112]. It is also indispensable in priming the immune system, thereby maintaining homeostasis between microbial communities and their host. As a result of this priming, the fine-tuned immune system reacts to commensals or mutualists with tolerance, while actively eliminating pathogens [113, 114, 8].

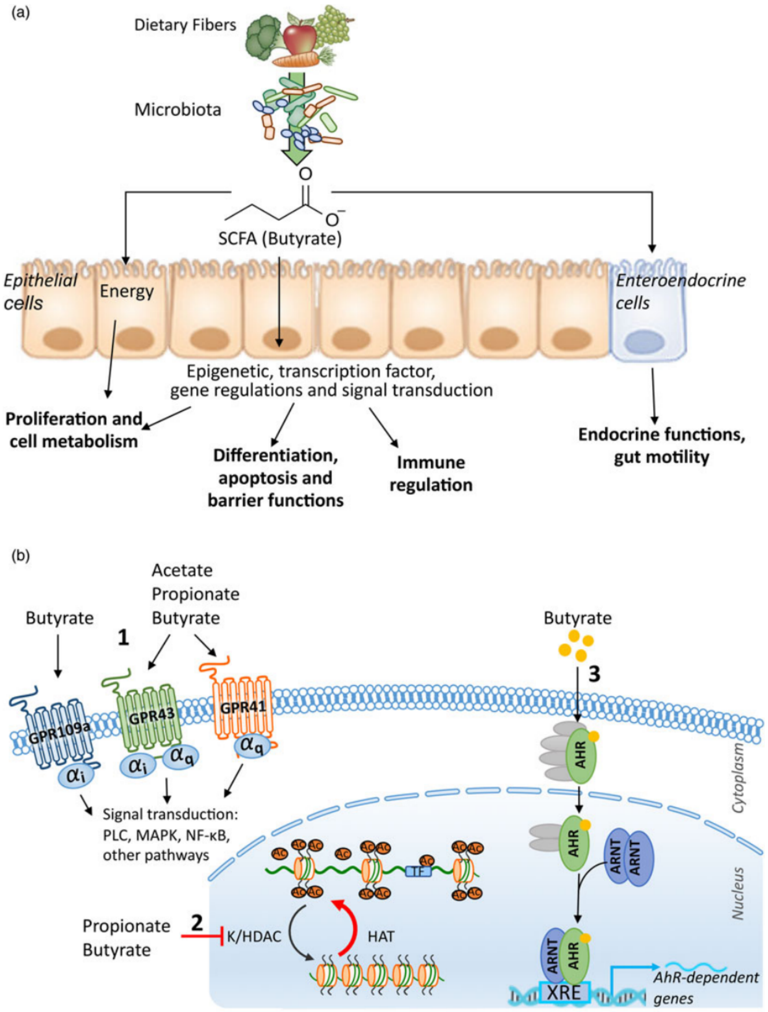


Figure 4.2: **Short-chain fatty acids (SCFAs) as key products of microbial metabolism in the intestine** a) Influence of microbiota derived SCFAs taken up through the intestinal epithelium on the human body, b) mechanisms through which they exert their effects: 1.) activation of G protein-coupled receptors (GPCRs), 2. modification of histones and transcription factors, 3. butyrate transcription factors binding. AhR, aryl hydrocarbon receptor; ARNT, aryl hydrocarbon receptor nuclear translocator; HAT, histone acetyltransferase; K/HDAC, lysine/histone deacetylase; MAPK, mitogen-activated protein kinase; PLC, phospholipase C; TF, transcription factor; XRE, xenobiotic response element. Modified with permission from Figure 1, Martin-Gallausiaux *et al*, SCFA: mechanisms and functional importance in the gut. Proceedings of the Nutrition Society. 2021;80(1):39) [115].

4.4.1 Influence of microbial fermentation products

The breakdown of complex carbohydrates by gut microbiota involves a variety of polysaccharides, including cellulose, hemicellulose, resistant starches, and non-starch polysaccharides like inulin [116, 117]. These polysaccharides are hydrolyzed by microbial enzymes into simpler sugars, which are then fermented to produce mainly short-chain fatty acids (SCFAs) [118]. SCFAs are absorbed by gut epithelial cells, contributing not only to their energy supply but also to various metabolic activities. They also seem to have wide ranging and substantial effects on the endocrine as well as immune system [111, 119] (Figure 4.2).

SCFAs are long recognized as being of critical importance to human health. Examples of the effect of the three most abundant ones, butyrate, propionate, and acetate, serve to illustrate the direct and substantial health effects of microbiota-produced compounds.

Butyrate

Butyrate serves as an energy source for colonocytes and is important in maintaining their oxidative metabolism [120, 121]. It seems critical for central metabolic pathways, such as the tricarboxylic acid (TCA) cycle of colon cells. In its absence, a lack of energy as observed in germ-free mice, impairs the colonocytes' ability to maintain the colonic mucosal barrier [120]. Additionally, butyrate acts as a histone deacetylase inhibitor, influencing gene expression and contributing to epigenetic regulation with potential effects on key regulators of cell proliferation, differentiation, and apoptosis [122, 120, 123]. These regulatory roles also potentially offer protective effects against diseases like colorectal cancer [124].

Furthermore, butyrate enhances the intestinal barrier by promoting the formation of tight junctions, essential for maintaining the integrity of the intestinal epithelium, thereby reducing the risk of gut permeability and systemic inflammation with multiple beneficial effects [125, 126, 124, 127, 128]: By aiding in the prevention of systemic inflammation, butyrate seems to contribute to cardiovascular health, potentially mitigating the progression of atherosclerotic disease [129, 130, 131]. Furthermore, maintaining a functional gut barrier prevents pathogenic bacteria from entering the bloodstream, which could otherwise promote cardiovascular diseases through the activation of the immune system and modulation of metabolic and inflammatory responses [131, 124]. Additionally, butyrate, among other SCFAs, has been implicated in blood pressure regulation, although the exact impact on hypertension is subject to ongoing research [131].

Finally, butyrate also modulates immune responses through several mechanisms. It interacts with GPR109A on intestinal dendritic cells as well as macrophages, which is implicated in various anti-inflammatory processes e.g. by increasing cytokine production, potentially promoting the development of regulatory T cells and restricting the proliferation of pro-inflammatory cells [132, 133, 134]. As a histone deacetylase (HDAC) inhibitor, butyrate increases histone acetylation, regulating the expression of genes relevant to immune responses [135]. One of the key human anti-inflammatory cytokines impacted through various mechanisms by butyrate is for example IL-10 [133, 136]. Butyrate also affects intestinal macrophages, increasing anti-microbial peptide production and encouraging differentiation of the anti-inflammatory M2 phenotype [137, 133, 138].

Propionate

Propionate is actively involved in metabolism, being processed in the liver where it contributes to gluconeogenesis, thus playing a part in the regulation of blood sugar levels [139].

Beyond its metabolic functions, propionate also impacts the central nervous system. It seems to exert a protective influence on the blood-brain barrier (BBB) and achieves this by blocking pathways related to microbial infection, as well as by being involved in the oxidative stress response through the activation of the NRF2/NFE2L2 signaling pathway [139]. Another BBB protective interaction involves the FFAR3 receptor on the brain endothelium [139]. Activation of the same receptor in gut enteroendocrine cells leads to hormone secretion, regulating appetite and satiety [139].

Furthermore, propionate may play a role in supporting cancer treatment. During intermittent fasting, its metabolism to indole-3-propionic acid can enhance the effectiveness of chemotherapy by modulating metabolic and immune pathways vital for cell regeneration and recovery [140], suggesting a potential role in supporting cancer treatment. Another aspect of propionate's function could be the mitigation of vascular calcification by modulating the intestinal microbiota composition, leading to enhanced SCFA production and improved intestinal barrier function, which might collectively reduce inflammation and calcification in the vascular system [141].

In terms of influencing immune responses, propionate is known to elicit an anti-inflammatory effect by being the main agonist of the already mentioned FFAR3 (GPR41) receptor as well as interacting with the G-protein coupled receptor 43 (GPR43/FFAR2), which seems to lead to a reduction of the production of pro-inflammatory cytokines as well as a suppression of the recruitment of inflammatory cells, such as neutrophils [142, 143, 144].

Acetate

Acetate, the most abundant SCFA in the colon, influences various metabolic processes. While it can be produced endogeneously and also taken up with the diet, microbiota production is a significant source as well [145, 146]. It can serve as a major energy source for peripheral tissues [147, 148, 149] and, as a substrate for lipid synthesis in the liver, it is relevant for de novo lipogenesis and cholesterogenesis [150, 149]. Its involvement in these processes has implications for overall lipid metabolism, influencing the balance between energy storage and expenditure [151, 149].

In addition to its metabolic roles, acetate is also investigated for its effects on the central nervous system as it can traverse the blood-brain barrier [152]. Acetate, among other things, modulates gene expression in the brain. Through its interaction with the key protein-acetylation enzyme ACSS2, it may influence cognitive functions, including appetite regulation, stress responses, and memory formation. [152, 149].

Furthermore, acetate impacts immune function and inflammation. It participates in the regulation of immune responses, particularly in the gut, where it, as other key SCFAs, seems to play a role in influencing regulatory T cells potentially via FFAR2/FFAR3 [153]. It also supports the function of dendritic cell cytokine production through ACCS2 [154, 149].

Summary SCFAs

In summary, microbiota-derived SCFAs significantly influence human physiology in key areas. They impact metabolism by regulating glucose and lipid processing, enhance immunity through modulation of inflammatory responses and support of gut barrier function, and affect behavior by influencing neurological pathways linked to stress and cognition. Additionally, they play a role in cardiovascular health, contributing to the regulation of blood pressure and vascular function. SCFAs are also implicated in the progression, prevention, or cure of various diseases such as colorectal cancer, where they inhibit tumor growth, and in type 2 diabetes, where they improve glucose metabolism and insulin sensitivity.

4.4.2 Essential nutrients

While diet is a primary source for most vitamins, the gut microbiota substantially contributes to the synthesis of specific vitamins, particularly vitamin C, K, and the B group vitamins (like B12, B6, folate, biotin, and riboflavin) [155, 156].

While there is no clear evidence of gut microbiota-produced ascorbate (vitamin C) adding in any substantial way to cover the human supply requirement, there is evidence of it modulating inflammatory immune responses by reducing the production of pro-inflammatory cytokines in T cells, inducing apoptosis in them, and inhibiting the aerobic glycolysis [157, 158].

The K vitamins in form of menaquinones, important for the maintenance of hemostasis [159] and potentially various other significant physiological processes [160], are produced in large amounts and taken up, but since the daily requirements are low, either microbiota-produced or nutrition based supply seems usually more than sufficient [161, 162, 158].

B group vitamins, a chemically diverse group of cofactors or cofactor precursors, are critical for human health, as they are necessary for various vital metabolic functions. Several are also sourced from microbial production.

Vitamin B6 (pyridoxine, pyridoxal, and pyridoxamine) is, among others, essential for the amino acid metabolism [163] and thus almost all aspects of human physiology. The gut microbiota seems to produce significant amounts of these vitamers and mechanisms of uptake by colonocytes have been found [164, 165].

Folates (vitamin B9) are critical for nucleotide synthesis and repair, amino acid metabolism, and methyl group transfers, among other processes [166]. While nutrition is the major source, humans do take up microbiota-produced folates [167, 168, 169]. Biotin (vitamin B7/H) is important for the metabolism of carbohydrates, fats, and amino acids as an essential cofactor of enzymes catalyzing carboxylation and decarboxylation reactions [170, 171, 172]. Microbiota-produced biotin is taken up in the large intestine, likely depending on its availability to the host [173, 169, 174].

Finally, riboflavin (vitamin B2), produced in considerable amounts by the gut microbiota [175] and taken up to some extent in the large intestine [176], is crucial for the electron transport chain of aerobic respiration and plays a significant part in the metabolism of amino acids [163].

It should be noted that many of the listed vitamins also likely impact gut microbiota structure and activity, resulting in a complex network of interactions between nutrition, host physiology and microbiota [177, 178]. For example, there is a wide variety of cobamides produced and exchanged by the gut microbiota and potential profound effects of nutritional cobalamin on them [179].

4.4.3 Community homeostasis: Microbial competition and host immunity

Microorganisms residing both within and on the body also play various roles regarding the functioning of the immune system. Co-evolution of the microbiota and the host led to the development of ways for microbial communities to modulate the immune responses of their host [180, 181] and in turn a multitude of mechanisms for the host to manage its inhabitants, differentiating between pathogens and commensals, producing defensive responses against the former and being tolerant to the latter [182]. As such, the 'healthy' microbiota, located in areas such as the skin, the oral cavity, and the gastrointestinal tract, for example not only form a physical barrier against pathogenic microorganisms but also contain microbial communities who compete with them, thereby limiting their growth [183, 184].

Bacteriocins

A central mechanism for intra-microbiota competition is anti-microbial molecules, primarily bacteriocins, which combine anti-bacterial effects with other activities that can impact other microbiome members or the host [185]. The variety of bacteriocins is vast and their functions diverse, here the focus shall be on their host-modulating functions and role in keeping an intra-microbiota structure homeostasis that benefits host and commensals. Some bacteriocins specifically target close relatives of the producer [186] or a narrow range of unrelated species [187], whereas many others, especially smaller peptides (produced by Gram-positive bacteria), have broader activity and affect unrelated bacteria [188].

In a homeostatic microbiota-host system, there should be a well established interplay between members of the microbiota, producing stable niches with e.g. sub-communities, all thriving within bounds set by nutrient availability, antimicrobial peptide (AMP) resistance, and host tolerance [7]. It is within this context, that the amount and variety of bacteriocins produce an anti-pathogenic effect, since overgrowth of existing opportunists is kept in check by limiting their expansion and growth. New organisms introduced, on the other hand, are likely susceptible to already present bacteriocins and thus eliminated directly or their growth strongly impaired, resulting in the host being resistant to colonization [185].

There are also various mechanisms through which members of the microbiota use bacteriocins to influence the host. In the nose, *Staphylococcus lugdunensis* produces the fibupeptide lugdunin, which acts against the competitor *Staphylococcus aureus* directly and also interacts with Toll-like receptor 2, activating human keratinocytes, promoting an increase in AMP production [189]. Some others, like Cytolysin [190], Colibactin [191], and Streptolysin S, impact the host by inducing cell damage [192]. Gassericin, produced by *Lactobacillus gasseri*, observed in human breast milk is, at least in piglets, also shown to influence intestinal epithelial cells, modulating secretion and absorption capabilities [192].

Another essential mechanism is the distribution, sharing, and competition for nutrients, in which bacteriocins are deeply involved in multiple ways. Firstly, they free the cellular contents of vulnerable cells they destroy and, secondly, they can also serve a dual role in nutrient acquisition, e.g. as siderophore-microcins, where the siderophore part is used as a 'trojan-horse' to bind to target cell outer membrane receptors [193].

Besides the numerous and complex interactions that keep a stable microbial community in and on the host, one key mechanism that is intricately related is the tuning of the host's immune system to both sides' benefit [7].

Immune system training and modulation

The immune system is essential for human survival [114]. Its adaptive and innate components protect against possible damaging elements and try to maintain homeostasis mutually beneficial for the various microbial communities associated with the host [180, 181].

For its correct functioning, tuning by exposure, especially in the early stages of life, is required. Immune responses are adapted to gradually shift to the adult phenotype, allowing for continuous commensal colonization without eliciting hyper-inflammatory responses [113, 194].

One of the most impactful influences of the gut microbiota on the immune system is the modulation of the maturation and operation of critical immune cells like T cells and B cells, essential for mounting effective immune responses [108, 195] (Figure 4.3). For example, in the intestine, regulatory T cells (T_{reg} cells) are key mediators of mucosal homeostasis [196]. They express immunoglobulins [197] and cytokines [198], modulate IgA activity [199] and also support the integrity of the epithelium, promoting repair [200] and barrier functions [201]. Because of these essential effects, it is a necessity for the microbiota to influence this type of T cell to achieve tolerance by the host, which they might accomplish in various ways: Through production of (secondary) bile acids [202, 203, 204], and, as described before, possibly through SCFAs [153, 205]. Deeply involved in (intestinal) immunity and directly affected by T_{reg} cells are T helper 17 cells (T_H17). They are central components of inflammatory responses and are again modulated by the microbiota through bile acids [202]. This dual effect of microbiota-produced bile acids, reducing T_H17 and increasing T_{reg} cell differentiation, seems to be a key mechanism shifting the host immune response from inflammatory and hostile to tolerant [202].

Also key to microbiota-host immunity homeostasis and linked to the former important actors is secretory immunoglobulin A (sIgA). It is the most abundant human antibody, with large amounts secreted by the intestinal mucosae [206, 207] and the fundamental mechanism to keep intestinal microbiota in check by neutralizing microbial cells and toxins, through binding, thus preventing their attachment [208]. In turn, stimulation by the intestinal microbiota is required for sufficient IgA production. An example of a specific and direct way to induce production of IgA in a favorable and homeostatic manner are members of the genus *Alcaligenes* expressing specific lipopolysaccharides (LPS) to maintain their niche in Peyer's patches (PP). They induce cytokine production by dendritic cells, which in turn leads to increased IgA production in general but also *Alcaligenes*-specific IgA, which potentially represents the main mechanism used for PP colonization [209, 210, 211].

On the side of innate immunity, recognition of peptidoglycans by pattern-recognition receptors (PRRs) is a fundamental mechanism of immunity by the human host [212]. There is strong evidence that the microbiota tunes innate immunity in various ways [213], by e.g. presenting peptidoglycans to the PRR Nod1, which results in higher functionality of neutrophils and a crucial base level of immune system activation [214].

All in all, it is evident that there is an intimate and extremely complex interplay both within the microbiota and between the microbiota and the host, shaping human-associated microbial ecosystems. It clearly illustrates, why shifts in community structure and activity can have strong

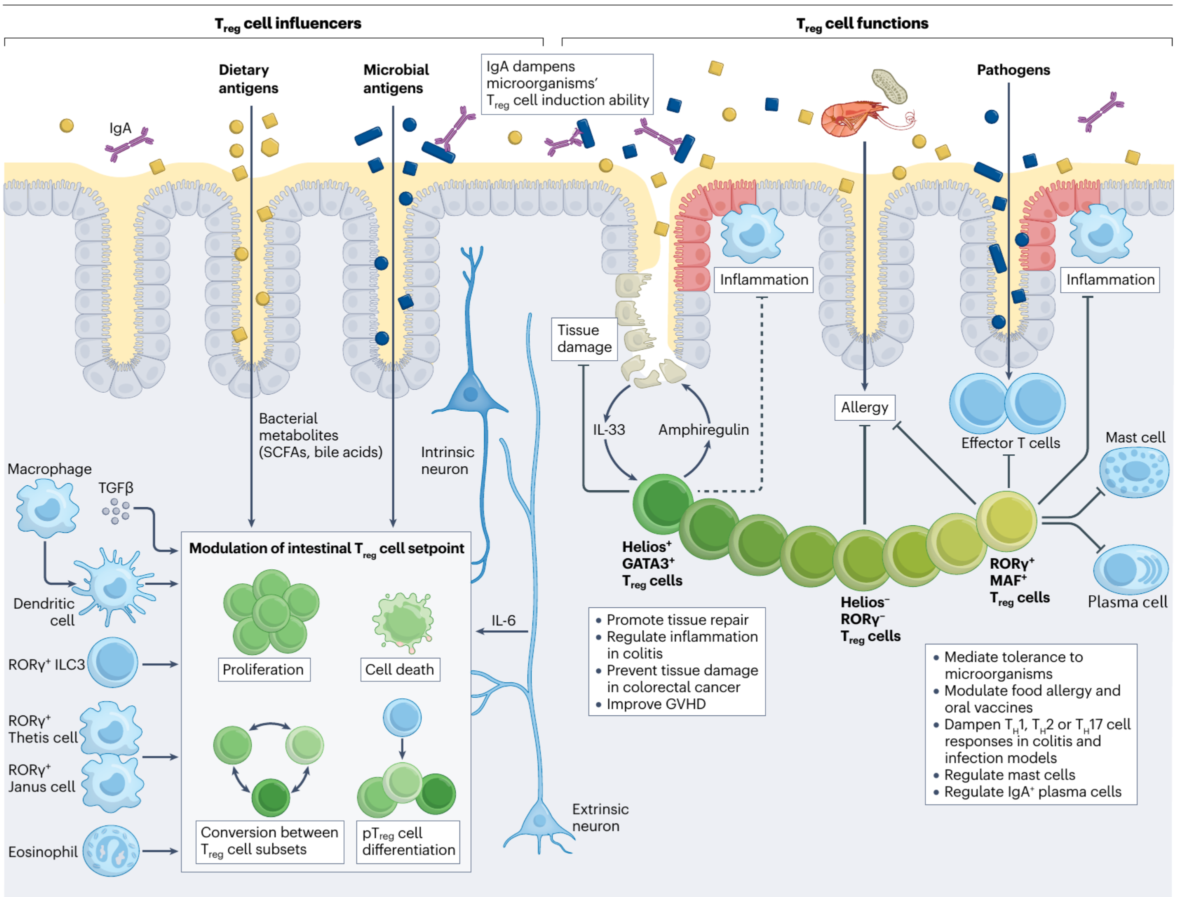

T_{reg} cell influencers
T_{reg} cell functions
IgA
Dietary antigens
Microbial antigens
IgA dampens microorganisms' T_{reg} cell induction ability
Pathogens
Inflammation
Inflammation
Tissue damage
IL-33
Amphiregulin
Allergy
Effector T cells
Mast cell
Bacterial metabolites (SCFAs, bile acids)
Intrinsic neuron
Macrophage
TGFβ
Dendritic cell
RORγ⁺ ILC3
RORγ⁺ Thetis cell
RORγ⁺ Janus cell
Eosinophil
Modulation of intestinal T_{reg} cell setpoint
Proliferation
Cell death
Conversion between T_{reg} cell subsets
pT_{reg} cell differentiation
IL-6
Extrinsic neuron
Helios⁺ GATA3⁺ T_{reg} cells
Helios⁻ RORγ⁻ T_{reg} cells
RORγ⁺ MAF⁺ T_{reg} cells
Plasma cell
• Promote tissue repair
• Regulate inflammation in colitis
• Prevent tissue damage in colorectal cancer
• Improve GVHD
• Mediate tolerance to microorganisms
• Modulate food allergy and oral vaccines
• Dampen T_H1, T_H2 or T_H17 cell responses in colitis and infection models
• Regulate mast cells
• Regulate IgA⁺ plasma cells

Figure 4.3: **Influencers and functions of intestinal regulatory T cell.** Intestinal regulatory T cells are core agents of intestinal immunity. Microbially derived and influenced factors (such as SCFAs, bile acids, sIgA) modulate immune responses by controlling T cell homeostasis. Modified with permission from Figure 1, Ramanan *et al*, Regulatory T cells in the face of the intestinal microbiota, Nat Rev Immunol, 23, 752, 2023, Springer Nature [195].

effects on the hosts health. Finally, due to the complexity of these interactions, it is often unclear what the cause and effect are (if a clear directional relationship exists at all). Therefore, caution is needed when attributing a specific phenotype (host and/or microbiota) to a particular agent or entity.

4.4.4 Chronic disease

Across different body sites, including mucosal surfaces in the gastrointestinal tract, the oral cavity and the skin, a diverse array of microbial taxa coexists with the human host. This symbiotic relationship, particularly with the immune system, is fundamental to maintaining health. However, disturbances in these microbial communities, known as dysbioses, can trigger inflammatory responses, potentially leading to chronic and autoimmune conditions [215, 216].

Inflammatory Bowel Disease (IBD), which encompasses conditions such as Crohn's disease and ulcerative colitis, represents a clinically significant group of chronic inflammatory conditions of the gastrointestinal tract. The interplay between the gut microbiota and the host in IBD is multifaceted, involving altered processing of gut microbiota-derived signals in addition to changes

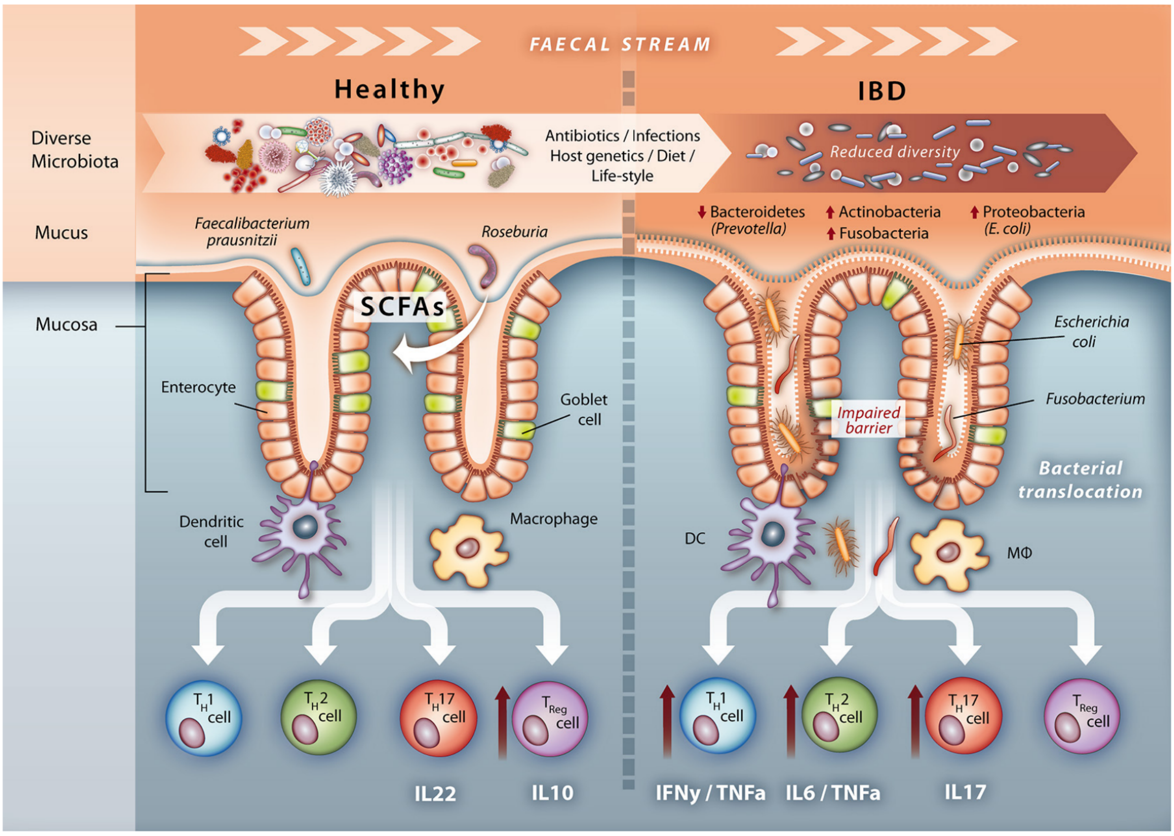


Figure 4.4: **Signatures of Inflammatory Bowel Disease.** In a healthy state, the gut microbiota is diverse and separated from the intestinal epithelium by a mucus layer which hosts mucus-resident bacteria. Immune cells promote a T$_reg$ profile ensuring homeostasis. In IBD patients, the diversity of the microbiota is decreased and the composition changed. Degradation of the epithelial barrier function lead to increased bacterial translocation, and inflammatory immune responses. Adapted by permission from BMJ Publishing Group Limited. Figure 1, Sommer *et al*, Microbiomarkers in inflammatory bowel diseases: caveats come with caviar, Gut 2017; 66:1735. [217].

in the composition and function of the gut microbiota itself [216, 218, 219] (Figure 4.4). As described earlier, gut microbiota-derived metabolites, such as SCFAs and bile acid metabolites, play pivotal roles in normal immune development, homeostasis, and have often been shown to be highly relevant to the pathophysiology of IBD [219]. For example, reduced fecal SCFA levels are often accompanied by the depletion of butyrate-producing bacterial genera like *Faecalibacterium* and *Roseburia* [219, 220]. Bile acids like cholate have been observed to be more abundant in fecal samples of IBD patients, hinting at a diminished ability of the microbiota to metabolize them with potential pro-inflammatory effect [221, 222].

Obesity, a global epidemic, is intricately linked to the composition of gut microbiota. In obese individuals, the gut microbiota composition often differs markedly from that of lean individuals. This difference is characterized by alterations in specific bacterial populations which, among other things, influence energy extraction from the diet, contributing to low-grade systemic inflammation and perpetuating a cycle of weight gain and further metabolic disturbances. Finally, this may lead to various metabolic and cardiovascular diseases [223, 224].

Diabetes, with Type 2 Diabetes Mellitus (T2DM) being the predominant form, has been closely linked to alterations in the gut microbiota. Epidemiological studies and mechanistic investigations in both humans and rodents suggest that T2DM and its precursor states, including prediabetes and gestational diabetes mellitus (GDM), may be influenced by changes in the gut microbiota. Meta-omics studies have illuminated some of the interactions between diet, (altered) gut microbiome, and host metabolism, revealing how dietary compounds are metabolized by the microbiome to produce a myriad of metabolites with systemic effects on the host. These microbiota alterations can affect key metabolic pathways, notably those involved in SCFA production and bile acid metabolism, which in turn influence insulin resistance, glucose intolerance, and overall metabolic health [225, 224].

Type 1 Diabetes Mellitus (T1DM) represents an autoimmune disorder marked by the destruction of insulin-producing β-cells in the pancreas, typically emerges early in life and leads to lifelong dependency on exogenous insulin [226]. While genetic predisposition plays a crucial role, the increasing incidence of T1DM points to a complex interplay between genetic and environmental factors, including gut microbiota dysbiosis. The gut microbiota's impact on T1DM seems to extend to modulating the host's immune system, particularly through molecular mimicry and the differentiation of immune cells like regulatory T cells (Tregs) [226]. However, the specific microbes and their metabolites involved in T1DM's onset and progression remain an active area of research, necessitating further investigation to develop microbiome-based therapeutic strategies. Findings from this work, using a multi-meta-omics approach, suggest a disease-specific dynamic between Streptococci species in the oral cavity that is linked to levels of those bacteria in the gut associated with T1DM. Alterations in the oral microbiome seemed to affect the microbial communities in the gut, particularly through reduced 'mouth-to-gut' transfer of S. *salivarius*, which may contribute to the inflammatory processes in T1DM [227].

Parkinson's Disease (PD) presents an example of how gut microbiota may significantly impact neurodegenerative disorders. Recent studies have indicated an alteration in the gut microbiome of PD patients, characterized by a decreased abundance of bacteria observed to be beneficial to health and mostly known for contributing anti-inflammatory effects. Changes in gut microbiota composition can influence gut motility, leading to conditions like constipation, dysphagia, and altered smell and taste, and are thought to affect the central nervous system via the gut-brain axis. These alterations are assumed to lead to gut inflammation, with intestinal barrier hyper-permeability, and the potential propagation of α-synuclein in the enteric nervous system, all of which may influence the risk and progression of PD, as well as the response to PD medication [228, 229, 230]. Similarly, in Alzheimer's Disease (AD), a form of dementia, emerging evidence suggests that gut microbiota plays a role in the disease's pathophysiology. Dysbiosis in dementia is again observed as a reduction in anti-inflammatory bacteria, possibly contributing to neuroinflammation and the formation of amyloid plaques, key features of Alzheimer's pathology. The gut-brain axis is considered a critical pathway for these microbial influences also here, with the potential to significantly affect neurodegenerative processes [228, 229].

4.4.5 Summary

In summary, the human microbiome, particularly the gut microbiota, is integral to various fundamental biological functions and overall health. Its role has been extensively examined, revealing impacts on almost all aspects of human health. Nonetheless, more comprehensive research applying integrated, high-resolution methods, is necessary to fully unravel the intricate interplays between the various microbiomes and the human body to devise strategies to not only avert diseases but also potentially improve the human condition, for example, by attenuating senescence or enhancing nutrient processing capabilities.

5

Results/Publications

5.1

binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets

5.1.1 Coversheet

Contributions of author Oskar Hickl

- Performed the conceptual design and planning of the study with P.M. and A.H.-B. This included identifying objectives and outlining methodologies to ensure a robust and coherent approach.
- Developed the application. This involved not only software design but also ensuring the application's functionality aligned seamlessly with the study's goals.
- Independently executed all experimental procedures. This encompassed setting up experiments, collecting data, and conducting preliminary analyses to validate the experimental outcomes.
- Led authoring the manuscript, detailing the study's objectives, methodologies, findings, and conclusions.
- Created all visual and tabular representations of the study's data. This task involved designing figures and tables that clearly and accurately conveyed complex information in an accessible format for readers.

5.1.2 Introduction

In the absence of our ability to perform single-cell sequencing on complex environmental samples or cultivate substantial amounts of community members individually, we have to rely on recovering genomes from metagenomes, the results of sequencing entire environmental samples. To understand the emergent properties of microbial communities, it is essential to have access to individual members' genomes to link individuals' functional potential back to the metagenome as a whole.

binny was designed with a focus on trying to address some of the prevalent challenges in binning metagenomes, where the recovery of substantial amounts of strain-resolved MAGs is still not possible. It adopts a unique, iterative and dynamic strategy to deal with the highly complex metagenomic data. It makes use of lineage-specific marker genes to allow immediate validation of potential MAGs.

Implemented as a Snakemake workflow, *binny* has at its core a binning algorithm developed in Python. The main routine operates iteratively; in each iteration, dimension reduction is performed based on the unbinned contigs' features. Subsequent clustering is then executed based on the low-dimensional coordinates extracted. Only clusters that pass purity and completeness thresholds are retained, while the rest undergo further iterations with modified parameters.

In benchmarks with both synthetic and real-world data against state-of-the-art binning methods, *binny* showed high performance. Over the semi-synthetic datasets, *binny* consistently surpassed or matched other methods in various metrics, especially in the recovery of pure and complete MAGs.

Two other aspects of the benchmarks highlighted *binny*'s performance: Accurately reconstructing highly fragmented genomes and distinguishing closely related organisms. Both are crucial abilities in recovering strain diversity from large, complex data sets.

When evaluated with real-world data from various environments, *binny* consistently performed well. It achieved the highest recovery of HQ MAGs and the second-highest of NC bins. Notably, *binny* recovered substantially more MAGs of MIMAG quality than any other methods.

In summary, the benchmarks suggest that *binny* could serve as a valuable addition to future metagenomic analyses, especially when dealing with large datasets.

5.1.3 Manuscript

Briefings in Bioinformatics, 2022, **23(6)**, 1–14
https://doi.org/10.1093/bib/bbac431
Advance access publication date 13 October 2022
Problem Solving Protocol

binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets

Oskar Hickl, Pedro Queirós, Paul Wilmes, Patrick May and Anna Heintz-Buschart

Corresponding authors: Patrick May, Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 1 Boulevard du Jazz, L-4370, Esch-sur-Alzette, Luxembourg. Tel: +352 46 6644 6263; E-mail: patrick.may@uni.lu; Anna Heintz-Buschart, Biosystems Data Analysis, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands. Tel: +31 020 525 6547; E-mail: a.u.s.heintzbuschart@uva.nl

Abstract

The reconstruction of genomes is a critical step in genome-resolved metagenomics and for multi-omic data integration from microbial communities. Here, we present *binny*, a binning tool that produces high-quality metagenome-assembled genomes (MAG) from both contiguous and highly fragmented genomes. Based on established metrics, *binny* outperforms or is highly competitive with commonly used and state-of-the-art binning methods and finds unique genomes that could not be detected by other methods. *binny* uses *k*-mer-composition and coverage by metagenomic reads for iterative, nonlinear dimension reduction of genomic signatures as well as subsequent automated contig clustering with cluster assessment using lineage-specific marker gene sets. When compared with seven widely used binning algorithms, *binny* provides substantial amounts of uniquely identified MAGs and almost always recovers the most near-complete (> 95% pure, > 90% complete) and high-quality (> 90% pure, > 70% complete) genomes from simulated datasets from the Critical Assessment of Metagenome Interpretation initiative, as well as substantially more high-quality draft genomes, as defined by the Minimum Information about a Metagenome-Assembled Genome standard, from a real-world benchmark comprised of metagenomes from various environments than any other tested method.

Keywords: metagenome-assembled genome, MAGs, embedding, dimensionality reduction, t-SNE, iterative clustering, marker gene sets

Introduction

High-throughput shotgun sequencing has become the standard to investigate metagenomes [1, 2]. Metagenome-assembled genomes (MAGs) allow the linking of the genetic information at species or strain level. In the absence of cultured isolates, MAGs form an important point of reference. Thereby, study-specific MAGs have led to the discovery of previously uncharacterized microbial taxa [3] and deepened insights into microbial physiology and ecology [4, 5]. In addition, large system-wide collections, which have been assembled recently, e.g. for the human microbiome [6] and several environmental systems [7], equip researchers with a common resource for short-read annotation. These collections also represent an overview of the pangenomic potential of microbial taxa of interest [8, 9]. In addition to facilitating the interpretation of metagenomic data, genome resolution also provides an anchor for the integration of functional omics [10, 11].

However, obtaining complete, un-contaminated MAGs is still challenging [12]. Most approaches start from assembled contigs, which are then binned by clustering, e.g. expectation-maximization clustering [13, 14] or graph-based clustering [15], of *k*-mer frequency or abundance profiles or both. Therefore, issues with metagenomic assemblies, such as fragmentation of the assembly because of insufficient sequencing depth, repeat elements within genomes and unresolved ambiguities between closely related genomes, are perpetuated to MAGs. In addition, the features based on which contigs are binned are not generally homogeneous over genomes: for example copy number, and thereby metagenomic coverage, may vary over the replicating genome; certain conserved genomic regions, and also newly acquired genetic material, can deviate in their *k*-mer frequency from the rest of the genome [12].

In the face of these challenges, the algorithms used to bin assembled metagenomic contigs into congruent groups, which form the basis for MAGs, can approximately be evaluated according to a set of criteria [16]. Most importantly, MAGs should be as complete as possible and contain as little contamination

Oskar Hickl is a PhD candidate in the Bioinformatics Core at the Luxembourg Centre for Systems Biomedicine. His research interests are metagenomics, -transcriptomics, and -proteomics method development and data analysis, as well as microbiome research.
Pedro Queiros was a PhD candidate in the Systems Ecology Group at the Luxembourg Centre for Systems Biomedicine. He now focusses on natural language processing and data integration as a machine learning engineer at Finquest, Foetz, Luxembourg.
Paul Wilmes is the Head of the Systems Ecology Group at the Luxembourg Centre for Systems Biomedicine and Professor in Systems Ecology at the Department of Life Sciences and Medicine of the University of Luxembourg. He uses advanced high-resolution molecular methods and experimental approaches to understand the functional ecology of microbiomes.
Patrick May is a Senior Researcher and the Head of the Genome Analysis group, Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. His research interests are human genetics and genomics and multi-omic microbiome research.
Anna Heintz-Buschart is Assistant Professor in Microbial Metagenomics at the University of Amsterdam. Her research interests connect data science, bioinformatics, and microbial ecology.

Received: June 9, 2022. **Revised:** September 3, 2022. **Accepted:** September 6, 2022

scale

as possible. In metagenomic datasets with defined compositions, such as those provided by the Critical Assessment of Metagenome Interpretation (CAMI) initiative [17–19], the evaluation can be achieved by comparison with the reference genomes. For yet unsequenced genomes, completeness and contamination can be assessed based on the presence and redundancy of genes that are expected to be present as single copies in many [20] or all [21] bacteria or archaea [22], or in specific lineages [23]. Contiguity and GC-skew provide further measures for highly complete genomes [12]. For reporting and storing MAGs in public repositories, the Minimum Information about a Metagenome-Assembled Genome (MIMAG) standard has been proposed [24]. In addition to completeness and contamination based on protein-coding genes, this standard also takes into account the presence of tRNA and rRNA genes. The latter present particular challenges for assembly and binning methods alike [12]. Nevertheless, the recruitment of rRNA genes to MAGs would improve the association with existing MAG collections [6, 25] and rRNA-gene-based databases [26], which are widely used for microbial ecology surveys. In addition to binning tools, refiners have been developed that complement results from multiple binning methods [27, 28]. These refiners generally improve the overall yield and quality of MAGs [29]. Finally, manual refinement of MAGs with the support of multiple tools is still recommended [12, 30–33].

Here, we present *binny*, an automated binning method that was developed based on a semi-supervised binning strategy [10, 34]. *binny* is implemented as a reproducible Python-based workflow using Snakemake [35]. *binny* is based on iterative clustering of dimension-reduced *k*-mer and abundance profiles of metagenomic contigs. It evaluates clusters based on the presence of lineage-specific single copy marker genes [23]. We benchmarked *binny* against six CAMI [17, 18] datasets and compared the results with the most popular binning methods MetaBAT2 [15], MaxBin2 [14], CONCOCT [13] and the recently developed VAMB [36], SemiBin [37] and MetaDecoder [38]. We evaluated the contribution of *binny* to automatic MAG refinement using MetaWRAP [27] and DAS Tool [28]. Finally, we evaluated the MAGs returned by all approaches from real-world metagenomic datasets from a wide range of ecosystems. We report that *binny* outperforms or is highly competitive with existing methods in terms of completeness and purity and improves combined refinement results. *binny* also returned most MIMAG-standard high-quality draft genomes from both highly fragmented and more contiguous metagenomes over a range of microbial ecosystems.

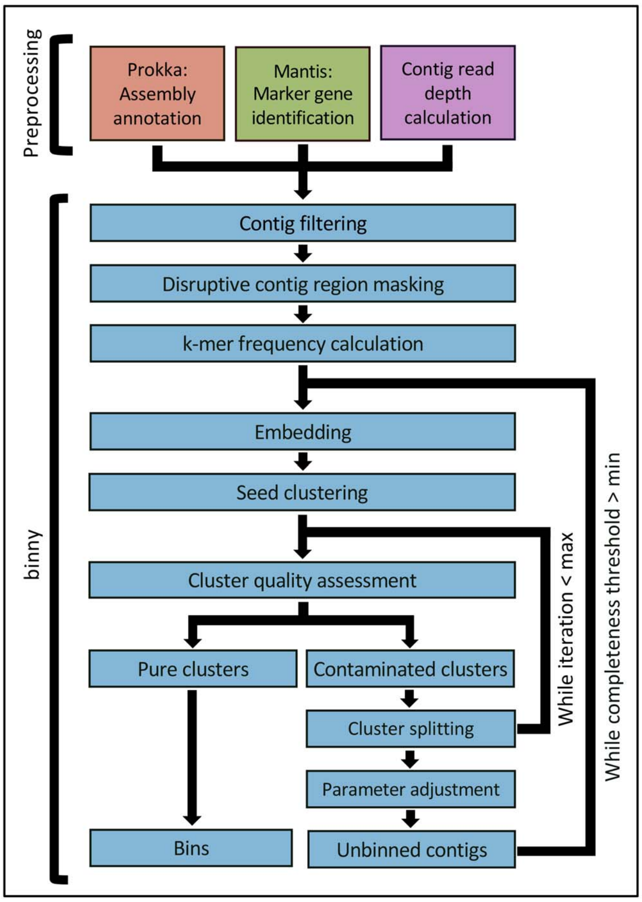

Preprocessing
Prokka: Assembly annotation
Mantis: Marker gene identification
Contig read depth calculation
binny
Contig filtering
Disruptive contig region masking
k-mer frequency calculation
Embedding
Seed clustering
Cluster quality assessment
Pure clusters
Contaminated clusters
Cluster splitting
Parameter adjustment
Bins
Unbinned contigs
While iteration < max
While completeness threshold > min

Figure 1. ***binny* workflow.** Overview of the Snakemake pipeline and of *binny*'s binning method. Preprocessing includes assembly annotation using Prokka, CheckM marker gene detection using Mantis, and (optional) average contig read coverage calculation. *binny* filters out contigs shorter than the specified value, masks potentially disruptive contig regions before calculating k-mer frequencies for the chosen k-mer size(s). In its main routine, *binny* iteratively embeds the contig data into two-dimensional space, forms clusters, assesses them with marker genes, and iteratively extracts clusters of sufficient quality as MAGs.

Material and Methods

binny workflow

binny is implemented as a Snakemake [35] workflow (Figure 1). At the centre of the workflow is the binning algorithm written in Python, which uses iterative, nonlinear dimension reduction of metagenomic read coverage depth and signatures of multiple *k*-mer sizes with subsequent automated contig clustering and cluster assessment by lineage-specific marker gene sets. Preparatory processing steps include the calculation of the average depth of coverage, gene calling using Prokka [39], masking of rRNA gene and CRISPR regions on input contigs and identifying CheckM [23] marker genes using Mantis [40].

Overview

binny operates in an iterative manner after processing of the annotated marker gene sets. Each iteration consists of nonlinear dimension reduction on the selected features (depths of coverage and *k*-mer frequencies) of the so far unbinned contigs and clustering based on the resulting two-dimensional coordinates. Clusters are selected if the contained marker gene sets indicate purity and completeness above defined thresholds. A new iteration is started on left-over un-binned contigs with dynamically adjusted parameters. Finally, clusters above the thresholds are output as MAGs.

Marker gene set processing

binny generates a directed graph database of the CheckM [23] taxon-specific marker sets annotated per contig in NetworkX [41]. This allows for fast access to the hierarchical (lineage-based) information. Some marker sets are omitted, as they are very small and/or led to imprecise assessments in testing (Supplementary Table 1).

Filtering of short sequences

By default, *binny* filters out all sequences shorter than 500 bp. For its main routine, further filtering is done based on an Nx value (default 90). For Nx filtering, the contigs are sorted by length in descending order and the first contigs that together make up x% of the assembly are retained. This size selection can be modified

scale

by setting minimum size values or ranges for contigs that do not contain marker genes (default 2250 bp) and those that contain them (default 2250 bp). This aims to maintain the maximum amount of information from an assembly because only contigs that have a low information content are omitted.

Masking of disruptive sequence regions

Certain regions on a sequence could skew the *k*-mer frequency and, thus adversely affect the binning process. For example, CRISPR regions contain foreign genetic elements, which have *k*-mer frequencies that can deviate substantially from the rest of the genome, whereas rRNA genes have highly conserved sequences whose *k*-mer profiles do not resemble the rest of a given genome. To avoid an impact on the *k*-mer frequency calculation and still keep sequences intact, *binny* by default masks sequence elements/regions such as rRNA genes and CRISPR regions, using Prokka-provided annotations from barrnap [39] and minced [42], respectively. The masked regions are ignored during the *k*-mer frequency calculation.

Single contig genome recovery

Genomes represented by single contigs might not be distinguishable from noise during clustering or be clustered together with highly similar contigs of other, fragmented genomes. Therefore, contigs with at least 40 different markers are extracted first and, if they are at least 90% pure and 92.5% complete, they are kept as single-contig MAGs and by default do not enter the iterative binning procedure.

Binning features

binny uses two contig features for dimensionality reduction and clustering: the *k*-mer frequencies of multiple sizes (default $k = 2$, 3 and 4) and the average read coverage (raw read counts of one or more samples), both centered log-ratio transformed. Coverage information can be included in form of bam files or a file with tab-separated average contig coverage values per sample.

Dimensionality reduction

To reduce the dimensionality of all features to two, the Fast Fourier Transform-accelerated Interpolation-based t-distributed Stochastic Neighbor Embedding implementation of openTSNE [43] is used. To decrease the computation time of the dimensionality reduction, Principal Component Analysis is used beforehand to lower the dimensionality of the initial feature matrix to either as many dimensions needed to explain 75% of the variation or to a maximum of 75 dimensions. To improve the embedding quality, especially with large datasets, multiple strategies are used: (i) a multi-scale kernel with perplexity ranges from 10 to 20 and 100 to 130 starting with 10 and 100, where each iteration the former is increased by 2 and the latter by 5, are used instead of a Gaussian model to balance out local and global structure, as described by Kobak and Berens [44]. (ii) An early exaggeration of *EX* for the number of unbinned contigs *NUC*:

$$\mathbf{EX} = min\{4, max\{100, \mathbf{NUC} \times 2.5 \times 10^{-4}\}\}, \tag{1}$$

with a learning rate *LR_EX* for the early exaggeration phase:

$$\mathbf{LR\_EX} = max\left\{2, \frac{\mathbf{NUC}}{\mathbf{EX}}\right\} \tag{2}$$

and a learning rate *LR*:

$$\mathbf{LR} = max\{200, min\{64 \times 10^{3}, \mathbf{NUC} \times 0.1\}\} \tag{3}$$

for the main phase are used. These values were chosen to achieve adequate embeddings of datasets of varying sizes [45, 46]. Additionally, the number of iterations to run early and main phase optimizations are based on the difference in Kullback–Leibler divergence (*KLD*) *KLD_DIFF*. The KLD is measured every 250 optimization iterations. The optimization ends [46], if:

$$\mathbf{KLD\_DIFF} < \mathbf{KLD} \times 0.01. \tag{4}$$

(iii) To avoid the impreciseness of Euclidean distance measures in high-dimensional space, Manhattan distance was chosen instead [47]. Default values were kept for all other openTSNE parameters.

Iterative clustering

binny uses hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [48] on the generated two-dimensional embedding, in iterations. *binny* will run clustering of the created embedding n times (default 3), each time extracting MAGs meeting the quality thresholds and continuing with the embedding containing only the leftover contigs. *n* is the number of values for HDBSCAN's `min_samples` parameter (default 1,5,10, hence n=3).

Other default clustering parameters are: the minimum cluster size is calculated with $\mathbf{ln}(n\ contigs)$, the cluster selection epsilon to merge micro-clusters is changed each *binny* iteration, cycling from 0.25 to 0.0 in 0.125 steps, and the distance metric used is Manhattan.

For each cluster, completeness and purity are assessed (see below). If a cluster passes the completeness threshold (by default starting with 92.5% and then decreasing to a minimum of 72.5%) and has a purity above 95%, if the completeness threshold is 90% or higher, otherwise it is set to 92.5%, it is kept as a MAG. Otherwise, *binny* will attempt to split that contig cluster iteratively using HDBSCAN a defined maximum amount of times (see above) but adding the raw depth(s) of coverage as additional dimension(s). Within each of these clustering rounds, the clusters below the quality threshold can be split again using HDBSCAN until no new clusters are identified and/or the maximum number of iterations is reached (default 1, no further splitting). To prevent the selection of low-purity clusters, the purity threshold is increased continuously to a maximum of 95% at completeness 70% or lower (99%, if the chosen marker set is Bacteria or Archaea).

Cluster assessment using marker gene sets

Clusters are assessed by calculating the purity and completeness based on the CheckM marker grouping approach, where marker genes known to be co-located in genomes of a lineage are collapsed into marker sets [23]. *binny* calculates MAG quality as in Parks *et al*. equation 1 and 2, respectively [23], except that instead of contamination purity is calculated. Let *P* be the purity for a set of collocated marker sets *MSS*, *MS* a marker set in *MSS*, *g* a single copy marker gene in *MS* and *C* the counts of *g* in

scale

a MAG:

$$P_{\mathrm{MSS}} = \frac{\sum_{\mathrm{M}\in\mathrm{MS}} \frac{\sum_{\mathrm{g}\in\mathrm{M}} \frac{1}{C_g}}{|M|}}{|\mathrm{MS}|}. \tag{5}$$

The taxonomic level and identity of the marker set are chosen dynamically. Assessment starts with completeness and purity of the domain-level marker sets and traverses the lineage down one taxonomic level at a time. At each level, completeness and purity for each taxon of the lineage are calculated. To combine the power of the domain level marker sets to give a general quality assessment with the specificity of lower level marker sets, the mean of purity and contamination for sub-domain level marker sets and their respective domain level set is used. If the marker set of the current taxon has an equal or higher completeness than the previously best-fitting marker set, it is set as the new reference. This choice is based on the assumption that the marker set with the highest completeness is least likely to be matching by chance and the larger the marker set size, the smaller the chance for miss-annotation. The lowest level to evaluate can be set by the user (default Class level).

Iterative binning

binny starts embedding and clustering the size-selected, un-binned contigs. The minimum contig size limit is decreased by 500 bp if less than half of the iterative clustering steps returned MAGs, until a minimum size of 500 bp is reached. In the next binning iteration, the completeness threshold will be decreased by 10% and the initial contig size threshold reset to the initial maximum value after which the cycle starts again. This will continue until the minimum completeness threshold is reached. At this point, the purity threshold is decreased to 87.5% for clusters with completeness ≥ 90% and the number of splitting attempts for contaminated clusters is increased to 2. This is done to recover as much information as possible in the final binning iteration. *binny* has a separate routine for co-assemblies, i.e. runs with depth of coverage information from more than one sample: here, *binny* creates embeddings and clusters of the un-binned contigs ≥ 500 bp of and runs subsequent binning iterations, for as long as it finds new MAGs that satisfy the purity and completeness thresholds. The completeness threshold is decreased by 10% in every binning iteration, down to the minimum completeness threshold (default 70% completeness). As with the single sample mode, the purity threshold is decreased to 87.5% for clusters with completeness ≥ 90% and the number of splitting attempts for contaminated clusters is increased to 2. Once no more MAGs are found at the minimum completeness threshold, *binny* runs final rounds with minimum contig sizes starting at 2000 bp, decreasing by 500 each round, until 500 bp or the minimum size set by the user is reached.

Contig depth of coverage calculation

If not provided explicitly, the average depth of coverage calculation can be performed directly from given BAM files within the Snakemake workflow using BEDTools [49] *genomeCoverageBed* and an in-house Perl script.

Coding sequence, RNA gene and CRISPR prediction by Prokka

A modified Prokka [39] executable is run with `–metagenome`, to retrieve open reading frame (ORF) predictions from Prodigal [50], rRNA and tRNA gene predictions from barrnap [39] and CRISPR region predictions from minced [42]. The modification improves speed by omitting the creation of a GenBank output and by the parallelization of the Prodigal ORF prediction step. Additionally, it allows the output of partial coding sequences without start and/or stop codons, which are frequently encountered in fragmented assemblies. No functional annotations of the called coding sequences are performed. The GFF output of Prokka is used in the subsequent steps.

Marker gene set annotation

Taxon-specific marker gene sets are acquired from CheckM (https://data.ace.uq.edu.au/public/CheckM_databases/) [23] upon installation of *binny*, hidden Markov profile models (HMM) of marker genes not found in `taxon_marker_sets.tsv` are removed, and `checkm.hmm` is split into PFAM [51] and TIGRFAM [52] parts. Mantis [40] is used to annotate coding sequences using the two HMM sets. Because both resources are of different scope and quality, consensus generation weights of 1.0 and 0.5 are used for PFAM and TIGRFAM models, respectively. Mantis' heuristic search algorithm is used for hit processing, the e-value threshold is set to 1 × 10^{-3}, and the `–no_taxonomy` flag is set.

Parameter customization

To optimize for their use case, a user can choose to change the sizes and number of *k*-mers used, the Nx value and/or minimum contig length to filter the assembly, as well as the minimum completeness and purity thresholds for MAGs. The user may choose not to mask potentially disruptive regions and can control the clustering process by adjusting several HDBSCAN parameters. Additionally, it is possible to choose between internal calculation of the average contig read depth or supplying a depth value file.

Requirements/dependencies

binny is implemented as a Snakemake pipeline and an installation script is provided that takes care of the installation of all necessary dependencies and required databases.

Benchmarking

Synthetic benchmark data

Binning performance was evaluated using datasets from the CAMI initiative [17, 19], each containing several hundreds of genomes at strain-level diversity. To benchmark against data of varying complexity, five short-read datasets with a total of 49 samples were chosen from the 2nd CAMI Toy Human Microbiome Project Dataset (https://data.cami-challenge.org/participate). Additionally, to test against a very large dataset, the five sample Toy Test Dataset High Complexity from the first CAMI challenge (https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_I_TOY_HIGH) was used.

To test the performance on co-assembled data, the pooled assemblies of each of the six CAMI datasets and the respective number of sample read files for each dataset, provided by the CAMI challenges, were used. Contig read depth per sample was calculated using *binny* and provided to all binning methods unless stated otherwise. Read files were de-interleaved (https://gist.github.com/nathanhaigh/3521724&#x2216;#file-deinterleave_fastq-sh) and mapped against the contigs using `bwa-mem` [53].

Real-world benchmark data

To assess the binning performance in different real-world scenarios with a variety of metagenome sizes, complexities and qualities, 105 metagenomes used in the MetaBAT2 publication [15] for benchmarking were chosen based on the availability of

scale

preprocessed read data at the Joint Genome Institute (JGI). The newest available assembly for the metagenomes and the respective preprocessed reads were retrieved from JGI (https://jgi.doe.gov/). The read data were processed in the same way as the CAMI data. For a full list with all sample information see Supplementary Table 2.

Binning and refinement methods

The performance of *binny* was compared to six other state-of-the-art binning methods, and to two binning refinement tools. *binny* and the other methods were all run using the default settings, unless specified otherwise:

MaxBin2 (2.2.7) [14] was run by providing the contig read depth files using the `-abund` option and with the `-verbose` option.

MetaBAT2 (2.2.15) [15] was provided the contig read depth files using the `-a` option and the options `-cvExt`, `–saveCls` as wel l as `-v`.

CONCOCT (1.1.0) [13] was run following the ’Basic Usage’ section in the documentation (https://concoct.readthedocs.io/en/latest/usage.html).

VAMB (3.0.2) [36] was run with the default parameters and using the Snakemake pipeline as described in the documentation (https://github.com/RasmussenLab/vamb/blob/master/README.md). Because VAMB is designed to achieve optimal performance through the combination of the data of multiple samples, the samples from each of the six CAMI datasets were concatenated and run together, as described by the authors (README sections `Recommended workflow` and `Snakemake workflow`). For the real-world metagenomes, samples sharing a JGI GOLD Study ID were run together. As VAMB could not be successfully run on some of the real-world samples using default values, or when trying with lower values of `-m` and `–minfasta`, the number of MAGs recovered was counted as zero for these samples. For a list of these samples see Supplementary Table 3.

SemiBin (1.0.2) [37] was run using the `single_easy_bin` mode with `–random-seed 0` and default parameters otherwise. For the single sample binning the global model was used, except for the CAMI 2 Gastrointestinal (GI) tract samples, for which `–environment human_gut` was used and the CAMI 2 Oral samples, for which `–environment human_oral` was used. For the real-world benchmark the respective models matching wastewater, ocean and soil samples were employed.

MetaDecoder (1.0.9) [38] was run using the default parameters, following the developers instructions, calling consecutively `coverage`, `seed` and `cluster`. To use `coverage`, the assemblies' respective bam files were converted to sam format using samtools.

DAS Tool (1.1.2) [28] was run using Diamond [54] as a search engine on the unfiltered binning method outputs.

MetaWRAP (1.2.2) [27] was set to output only contigs with less than 10% contamination and at least 70% completeness and was also provided the unfiltered binning method outputs. Both refinement tools, DAS Tool and MetaWRAP, were run: (i) per sample using the data of *binny*, MetaDecoder, and SemiBin and (ii) the two binning methods except *binny*, to asses how many MAGs *binny* contributes in an ensemble approach.

MAG quality standards

To match real-world workflows, all binning outputs were assessed using CheckM (1.0.12) [23] and filtered to contain only MAGs with a purity > 90% and a completeness > 70%. The latter threshold was set in accordance with the CheckM publication, which suggests that CheckM results are reliable at completeness equal or larger than 70%. MAGs above these thresholds are subsequently called ‘HQ’ MAGs. MAGs with a purity > 95% and a completeness > 90% are called ‘near-complete’ (NC) MAGs, as defined by Bowers *et al.* [24].

Additionally, the MIMAG definition of high-quality draft genomes was employed, requiring at least 18 unique tRNAs and three unique rRNAs to be present in the MAG in addition to a purity of >95% and a completeness of >90% [24].

Besides the recall in terms of bps of the assembly recovered, the read recruitment of MAGs was assessed. All reads mapping as primary mappings to contigs of a MAG were counted per sample and divided by the total read count (forward + reverse) using pysam (https://github.com/pysam-developers/pysam).

Assessment of benchmark results

Results for the CAMI benchmark were processed using AMBER (2.0.3) [55], a genome reconstruction evaluation tool, with the following parameters, `-x ‘50,70,90’` and `-k ‘circular element’`.

To evaluate a MAG, AMBER selects the gold standard genome with the highest share of bps in that MAG as the reference. In contrast to CheckM, where purity and completeness refer to the amount of marker genes present or duplicated, within AMBER and using an available gold standard, purity and completeness refer to the amount of bp of the reference genome recovered for completeness, and the share of bp of a given MAG with a given reference genome, respectively. Additionally, to assess one or multiple datasets taken together, AMBER defines overall completeness as ‘*Sum of base pairs coming from the most abundant genome in each predicted genome bin divided by the sum of base pairs in all predicted bins....*’ and overall purity as ‘*Sum of base pairs coming from the most abundant genome in each predicted genome MAG divided by the sum of base pairs in all predicted bins....*’.

Purity and completeness values are reported as the per dataset average, unless specified otherwise. For the real-world benchmarks, the average proportion of bp recovered or the number of MAGs recovered is reported together with the standard error of the mean (SEM). Another metric used is the adjusted Rand index (ARI), which is a commonly used metric to measure how similar two datasets are. Trying to make the comparisons between different binning methods as realistic, fair and transparent as possible, we report all metrics derived from the CheckM-filtered binning results, unless specified otherwise.

To assess the intersections of MAGs formed by the different binning methods on multi-sample datasets, genomes were counted separately for each sample. To this end, the gold standard genome name was concatenated with the sample id to yield unique identifiers for each genome in each sample. All other figures were created using the Python libraries *matplotlib* [56] and *Seaborn* [57], as well as *UpSetPlot* [58], setting the minimum intersection size to be shown to ten, for the UpSet plots. The remaining data analyses were performed and table outputs created using the Python *NumPy* and *pandas* libraries.

To evaluate if the binning methods could recover NC and HQ MAGs from organisms with closely related or highly similar genomes in the same sample, for each of the 54 samples of the six CAMI datasets all versus all Average Nucleotide Identity (ANI) calculations were performed using FastANI (1.33) [59]. Each genome was assigned the highest ANI to another genome in the same sample. The numbers of NC and HQ MAGs recovered per binning method with ANIs higher than 90.0–99.9% in 0.1 steps were counted.

Results

scale

Performance on synthetic datasets

To assess *binny*'s performance, six datasets from the CAMI initiative were chosen: the high complexity toy dataset of the first CAMI iteration to investigate how *binny* performs on very large datasets and the five toy human microbiome datasets of the second CAMI iteration to evaluate the performance on a wide range of microbiome sizes and complexities. Generally, a binning tool performs best, if it recovers the most complete MAGs with the highest purity, which corresponds to the highest ARI.

Over all six datasets (54 samples), *binny* with default settings recovered 35.5% (SEM 2.8%) of the reference genome lengths in the samples as NC MAGs (n = 1564) and 42.7% (SEM 3.0%) as HQ MAGs (n = 2021), with median recall values of 26.3% and 36.3%, respectively (Figure 2, Supplementary Table 4). In total, 45.1% of the reference genomes where recovered at a purity of 98.4% with an ARI of 0.977 (Supplementary Figure 1, Supplementary Table 5).

For the high complexity dataset, *binny* recovered 30.0% of the total reference genomes with a purity of 97.8% and an ARI of 0.970 (Supplementary Figure 2, Supplementary Table 6).

The lowest recall was observed for the CAMI 2 Airways dataset with 25.9%, a purity of 98.1%, and an ARI of 0.973 (Supplementary Figure 3), whereas the highest recall of 66.3%, with a purity of 98.6% and an ARI of 0.978 was reached with the CAMI 2 GI dataset (Supplementary Figure 4). For the other three datasets, *binny* achieved the following respective recall, purity and ARI numbers: 60.9%, 98.0% and 0.969 (CAMI 2 Urogenital); 48.0%, 98.9% and 0.983 (CAMI 2 Skin); and 33.2%, 98.6% and 0.982 (CAMI 2 Oral) (Supplementary Figures 5–7, Supplementary Table 6; for detailed metrics for MAGs and samples see Supplementary Tables 7 and 8, respectively).

The average read recruitment from the CAMI data of the *binny* output was 72.4%. The highest recruitment was achieved for the GI dataset sample 5 with 99.4%, whereas the lowest was observed for the skin dataset sample 19 (40.7%). Notably, a substantial proportion of the reads recruited were mapped to single contig MAGs for the CAMI 2 datasets (on average 60.7%), whereas for the CAMI 1 datasets, only about a fifth of the reads recruited by binned contigs, were mapped to single contig MAGs (Supplementary Tables 9 and 10).

Running *binny* with multiple depth files

When assessing the performance on co-assembled datasets with depth information from multiple samples, *binny* had a recall of 54.3% over the CAMI datasets with a purity of 98.4%. In total 1055 NC MAGs were produced, 413 of which contained more than five contigs (Supplementary Figures 8–10). The highest recall was achieved for the CAMI 2 GI co-assembly with 75.9% and a purity of 99.0%, whereas the worst performance was observed for the CAMI 2 Airways dataset with a recall of 32.6% and purity of 97.4% (Supplementary Tables 11–13).

To test to which degree *binny* makes use of the information from the multiple read depth files per co-assembly, *binny* was additionally run with only one depth file per co-assembly. *binny* using all available depth files had a 20.4% higher recall at a slightly higher purity, leading to a recovery of 25.0% more NC MAGs (211) in total and 102.5% more NC MAGs (209) of contig sizes larger than 5 (Supplementary Figures 8–10, Supplementary Tables 11–13).

Effect of masking potentially disruptive sequence regions

To test the effect of masking potentially disruptive sequences, we also ran *binny* on the 54 CAMI samples without the masking procedure. The unmasked run did not differ substantially from the one with the default settings regarding assembly recall and purity (Supplementary Table 14). In total, 29 fewer NC MAGs were recovered without the default masking (Supplementary Table 15). The amount of MAGs recovered matching the MIMAG standard was reduced by 5% from 1167 to 1112 (Supplementary Table 16).

Effect of lineage-specific marker gene sets

To evaluate the utility of using lower taxonomic level marker gene sets, we compared the difference in NC and HQ MAGs recovered between the default setting of a maximum depth at class-level to only using kingdom-level markers with the unfiltered output from the 54 CAMI samples. With the class-level marker sets and 8.5% more NC and 21.0% more HQ MAGs with a size of more than five contigs could be recovered, demonstrating the effectiveness of the lower level marker gene information with *binny*. Overall, with class-level markers the recall was 5.7% higher, whereas the purity was 1.2% and the ARI 1.8% lower (Supplementary Tables 17 and 18).

Run time

For all experiments, *binny* was run on compute nodes equipped with AMD Epyc ROME 7H12 CPUs, and for the run-time benchmark 32 cores and 56 GB of RAM were used. For the CAMI samples, the complete *binny* pipeline took on average 112 minutes to run, with a max of 413 minutes for sample five of the CAMI 1 high complexity dataset. The Prokka annotations took on average 28%, the Mantis annotations on average 15% and *binny* on average 57% of the total run time (Supplementary Table 19).

binny generally outperformed state-of-the-art binning methods on synthetic datasets

Over all six CAMI datasets *binny* recovered per sample the highest portion of the assembly (bps) as HQ (42.7%) or NC (35.5%) MAGs, followed by MetaDecoder (38.6%, 30.9%) and SemiBin (35.8%, 30.5%). Additionally, *binny* showed the highest median MAG counts with 23.8%, 36.8% more NC and 14.8%, 29.2% more HQ MAGs than MetaDecoder and SemiBin, respectively (Figure 2, Supplementary Table 4).

binny was the only binning method that resulted in high purity (97.3%) and high ARI (0.962) output over all datasets without additional CheckM filtering. Using CheckM filtering, *binny*'s purity and ARI were increased by 1.1% and 0.015, respectively, whereas the assembly recall was decreased by 3.0% (Supplementary Figure 1B, C, Supplementary Table 5). The binning method with the second highest NC MAG recall, MetaDecoder, had a purity of 84.6% natively and an ARI of 0.813. After CheckM filtering, the purity and ARI of VAMB was the highest among binning methods (99.5% purity and an ARI of 0.994, respectively), but at the same time the recall was reduced from 56.7% to 28.5% (Supplementary Figure 1B, C, Supplementary Table 5). For detailed metrics on the MAGs and samples see Supplementary Tables 7 and 8, respectively.

binny also outperformed the other binning methods on each of the individual datasets, except for the CAMI 1 High complexity dataset, where SemiBin produced 2.4% more NC MAGs (Figure 2, Supplementary Figures 2–7, Supplementary Table 7).

Many of the CAMI samples contain larger amounts of single-contig or almost contiguous genomes than are commonly

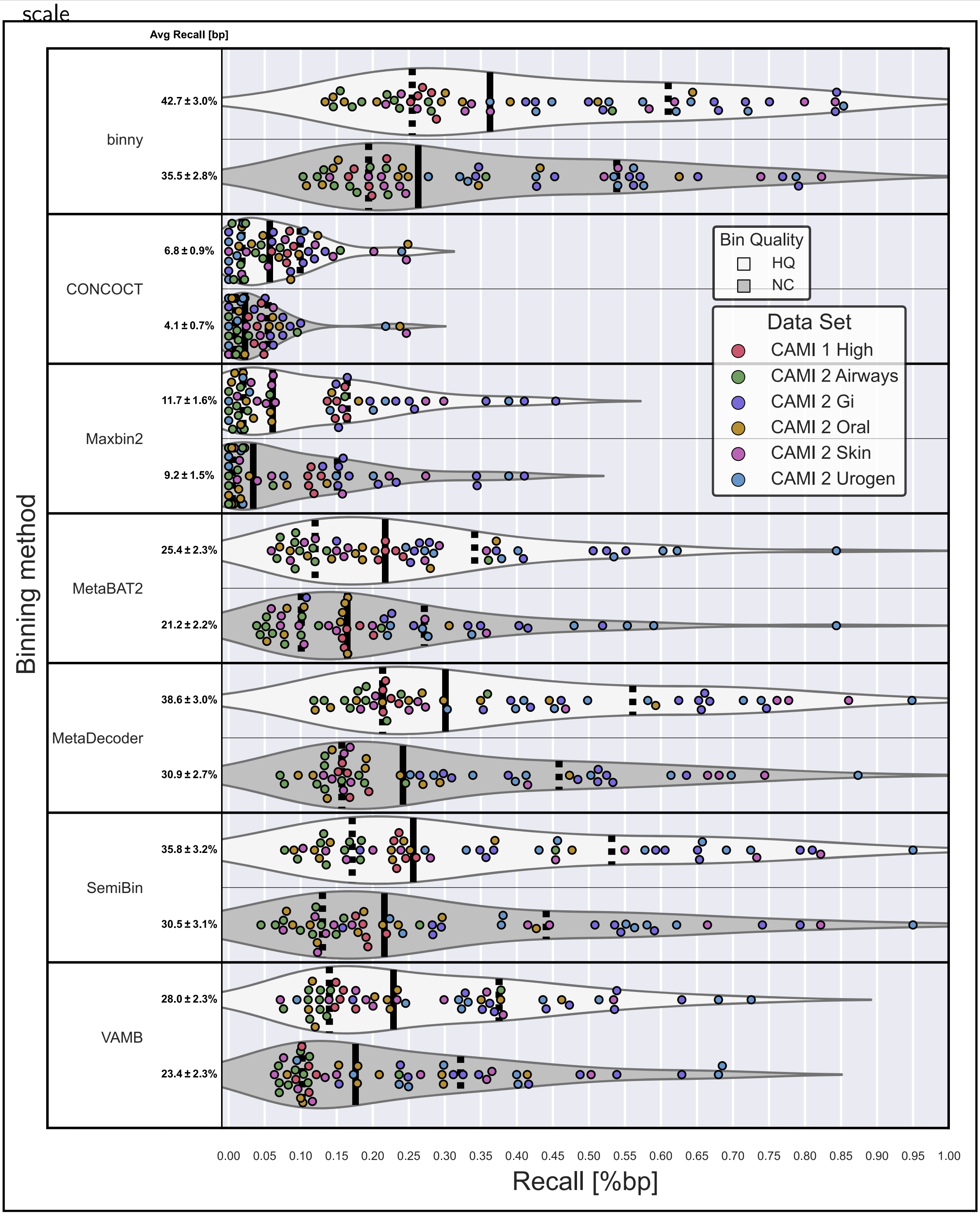


Figure 2. Performance of binning methods on CAMI datasets. Recall of bp assembled sequences as HQ and NC MAGs per binning method per sample, for the six CAMI datasets. The average recall is shown with the SEM.

observed in real-world samples. To evaluate *binny*'s performance without those, we considered the subset of genomes that consisted of more than five contigs. Here, *binny* also produced substantially more NC (13.1%) and HQ (25.3%) MAGs than the second best performing method, SemiBin (Supplementary Figure 11). *binny* recovered the largest amount of NC MAGs for the CAMI 2 GI, AW and Skin datasets, tied with SemiBin for the UG dataset and came second for the Oral dataset after VAMB (5.6% less) and the CAMI 1 High complexity dataset after SemiBin (0.4% less), respectively (Supplementary Figure 12, Supplementary Table 7). Looking at the assembly recall as NC and HQ MAGs, *binny* showed the best performance for all datasets (Supplementary Figure 13).

Additionally, *binny* recovered the most NC and HQ MAGs on co-assembly versions of the six datasets. It recovered 9.2% more NC and 13.9% more HQ MAGs than the second best method, MetaDecoder, and 7.6% more NC and 25.1% more HQ MAGs of genomes consisting of more than five contigs than the second best performer, SemiBin (Supplementary Figures 9, 10, 14, 15, Supplementary Tables 11–13).

scale

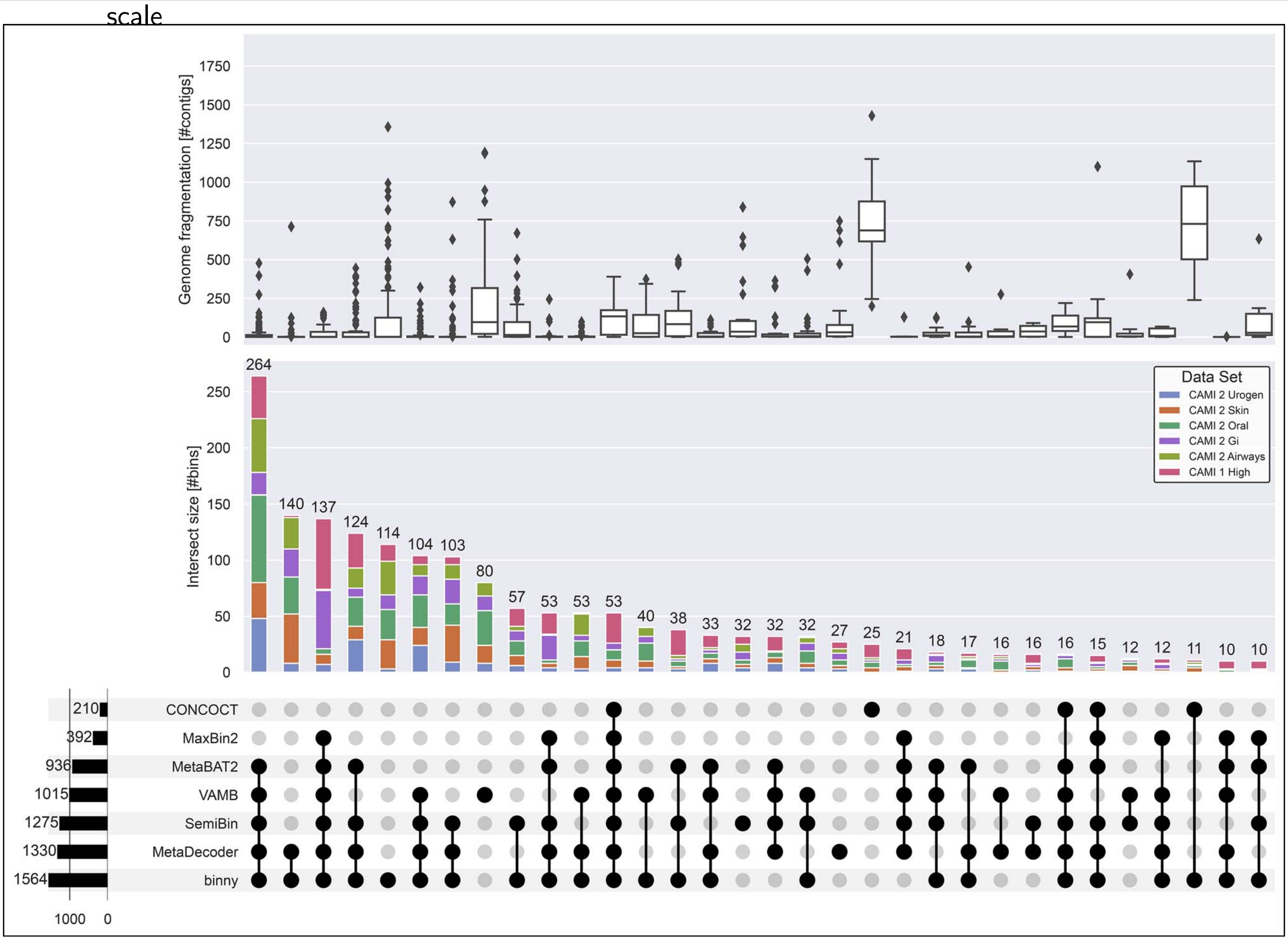


Figure 3. Intersections of recovered CAMI NC MAGs and reference genome fragmentation grade. Intersections of NC MAGs of seven CheckM-filtered binning methods for 54 samples from six CAMI datasets. Upper panel: Reference genome fragmentation in number of contigs. Middle panel: Intersection size in number of NC MAGs with proportions of MAGs stemming from the six CAMI datasets. Lower panel: Number of MAGs per binning method on the left, intersections > 9 in the centre.

Lastly, we assessed the amount of MAGs meeting the MIMAG draft standard. *binny* recovered the most MAGs of that quality for each CAMI dataset, recovering in total 20.3% more, with 1167, than the second best method, MetaDecoder, which produced 971 MIMAG drafts over all 54 samples from the six CAMI datasets (Table 1 and Supplementary Table 16).

binny recovered unique MAGs

To evaluate the performance of different binning tools, it is also of interest to see how much unique information is recovered by each individual binning method. *binny* yielded 42.5% more unique NC MAGs (114) than the next best, VAMB for the CAMI datasets. Additionally, the two largest sets of MAGs shared by two binning methods are both *binny* sharing MAGs with MetaDecoder (140) or SemiBin (57), respectively (Figure 3). For the HQ genomes, similar results were observed: *binny* recovered the second most unique MAGs after VAMB (5.8% less) and was present in all of the intersections with the largest numbers of genomes (Supplementary Figure 16, Supplementary Table 7). On the co-assemblies, *binny* recovered 31.3% more unique NC and 67.4% more unique HQ MAGs, than the method with the second most unique MAGs, MetaDecoder (Supplementary Figures 9 and 14).

binny produced complete and pure MAGs from contiguous as well as highly fragmented genomes

Next, we assessed the ability of different binning methods to recover genomes of different fragmentation grades. *binny* recovered substantially more highly fragmented genomes (defined here as genomes with more than 500 contigs) than almost all methods (50 NC MAGs). Only CONCOCT recovered more highly fragmented genomes than *binny* (54), whereas both methods shared the recovery of a large portion of these fragmented genomes. VAMB produced the third most with 27 highly fragmented NC MAGs (Supplementary Figure 17A, Supplementary Table 7). When looking at the number of fragmented HQ MAGs recovered, *binny* substantially outperformed all other methods, recovering 26.6% more than the second best method, CONCOCT, with 282 MAGs (Supplementary Figure 17B, Supplementary Table 7). For the co-assemblies, *binny* recovered 133.3% more NC and 101.2% more HQ MAGs than the second best method SemiBin (Supplementary Figure 18, Supplementary Table 13).

binny recovers MAGs from genomes with highly similar relatives

When assessing a binning methods' performance, it is also of interest to evaluate how well it is able to separate closely related organisms, as this would e.g. allow for the study of strain variation within a sample. Over all CAMI samples, *binny* recovered the largest amount of NC and HQ MAGs from genomes with highly similar relatives in the same sample over an ANI range from 90% to 99.9%. At an ANI of 90.0% *binny* recovered 730 NC and 840 HQ MAGs. The second and third highest performing methods were MetaDecoder (35.4% less NC, 20.2% less HQ MAGs) and SemiBin (52.1% less NC, 45.6% less HQ MAGs). When taking only into

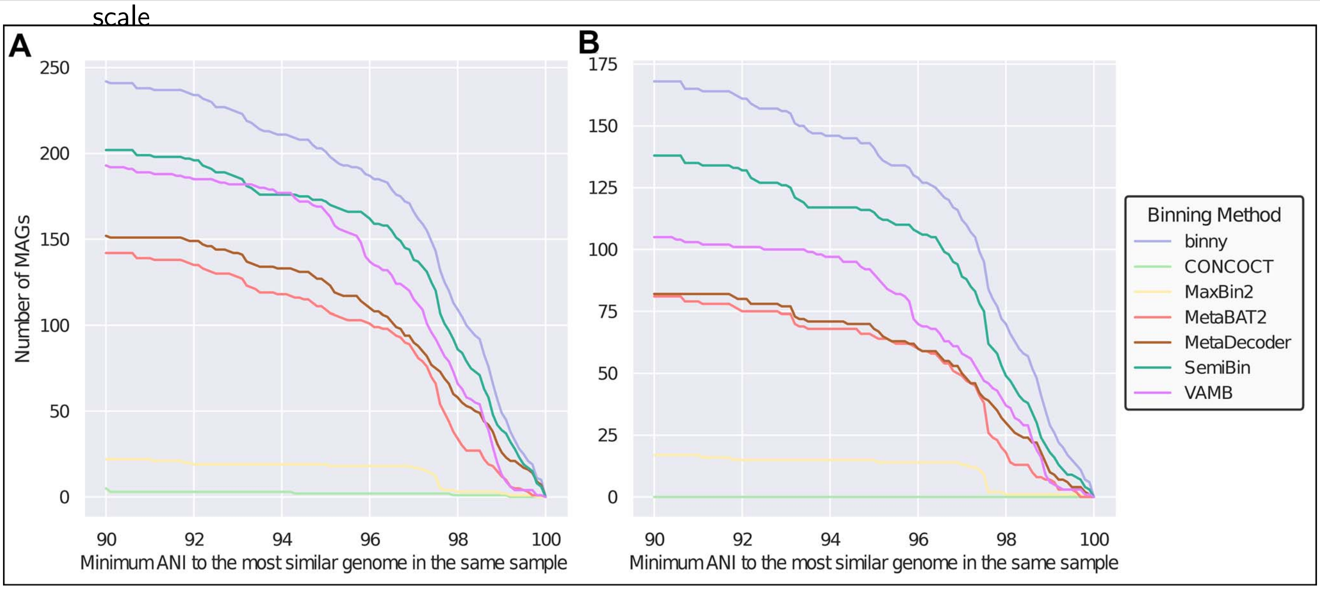


Figure 4. Performance of binning methods on recovering MAGs with close relatives. Number of (**A**) HQ and (**B**) NC MAGs with a minimum ANI to the most similar genome in the same CAMI sample of at least 90.0% up to 99.9% in 0.1% steps for seven CheckM-filtered binning methods. Includes genomes consisting of at least six contigs.

Table 1. MAGs matching the MIMAG standard. Rows represent values per binning method for the six CAMI datasets and the number for the real-world benchmark data. Bold values show the highest count per dataset, underlined values the second highest count.

	High	AW	GI	Oral	Skin	UG	IMG
binny	**164**	**192**	**202**	**243**	**215**	**152**	**629**
CONCOCT	17	10	19	37	18	6	142
MaxBin2	85	5	79	20	26	16	422
MetaBAT2	144	81	100	134	85	111	417
MetaDecoder	140	<u>147</u>	166	193	<u>181</u>	<u>144</u>	533
SemiBin	<u>148</u>	113	<u>168</u>	184	156	143	<u>553</u>
VAMB	107	121	138	<u>197</u>	123	113	406

AW: Airways, UG: Urogen, IMG: real-world data.

account genomes consisting of six or more contigs, *binny* still outperformed all other methods, followed by SemiBin (21.7% less NC, 19.8% less HQ MAGs) and VAMB (60.0% less NC, 25.5% less HQ MAGs). At an ANI cut-off of 95% *binny* recovered 36.8% and 22.6% more NC MAGs than the second highest performing method from genomes consisting of any number or at least six contigs, respectively. Finally, at an ANI of 99.0% *binny* was able to generate 41.8% and 61.1% more NC MAGs from genomes consisting of any number or at least six contigs, respectively, than the method placing second (Figure 4 and Supplementary Figure 19).

binny recovered the largest number of MIMAG drafts for real-world assemblies from different environments.

When benchmarking binning tools with real-world data from a wide variety of environments, *binny* recovered on average the second largest amount of the assembly (bp) as NC (19.8%) bins, after MetaDecoder (20.2%), and the largest amount of HQ (28.8%) MAGs. MetaDecoder in total recovered the most NC MAGs (1647), followed by *binny* (1523) and SemiBin (1513). Notably, there was a substantial gap in performance to the next best method, MetaBAT2, with 1223 NC MAGs recovered (23.7% less than SemiBin). *binny* recovered the largest amount of HQ MAGs (3013), followed by MetaDecoder (2969) and SemiBin (2747). As in the CAMI benchmarks, CONCOCT showed the lowest recall for both NC and HQ MAGs, whereas MaxBin2 performed comparatively better with these data than in the CAMI benchmark (Figure 5 and Supplementary Tables 20, 21). When counting the recovered MAGS matching the MIMAG standard, *binny* produced 629 MAGs, 13.7% more than the second best-performing method, SemiBin (553), with MetaDecoder ranking third with 533 (Table 1 and Supplementary Table 22).

binny improved ensemble binning/refinement approaches

To test if *binny* is able to improve refinements in combination with other binning methods, we ran the two most popular automatic refinement tools, DAS Tool and MetaWRAP, on the 54 samples of the six CAMI datasets, combining MetaDecoder and SemiBin either with or without *binny*.

When *binny* was excluded, a 1.9% and 2.9% lower recall was observed for DAS Tool (48.4%) and MetaWRAP (45.0%), respectively, whereas binny had marginally lower recall than DAS Tool with 48.1% (Supplementary Figure 20B, C, Supplementary Table 23). *binny* on its own, unfiltered, recovered 7.0% more NC MAGs than DAS Tool and 2.4% less than MetaWrap without the *binny* MAGs. When including *binny*, MetaWRAP was able to recover 8.8% more NC MAGs (1705) than *binny* on its own, whereas DAS Tool produced 2.4% more NC MAGs (1605) (Supplementary Figure 20A, Supplementary Figure 21, Supplementary Table 24). Only MetaWrap with *binny* input produced more HQ MAGs than *binny* alone with 2174 (6.0% more) (Supplementary Figure 22, Supplementary Table 24). Including only MAGs with more than five contigs, the DAS Tool and MetaWRAP without *binny* performed

scale

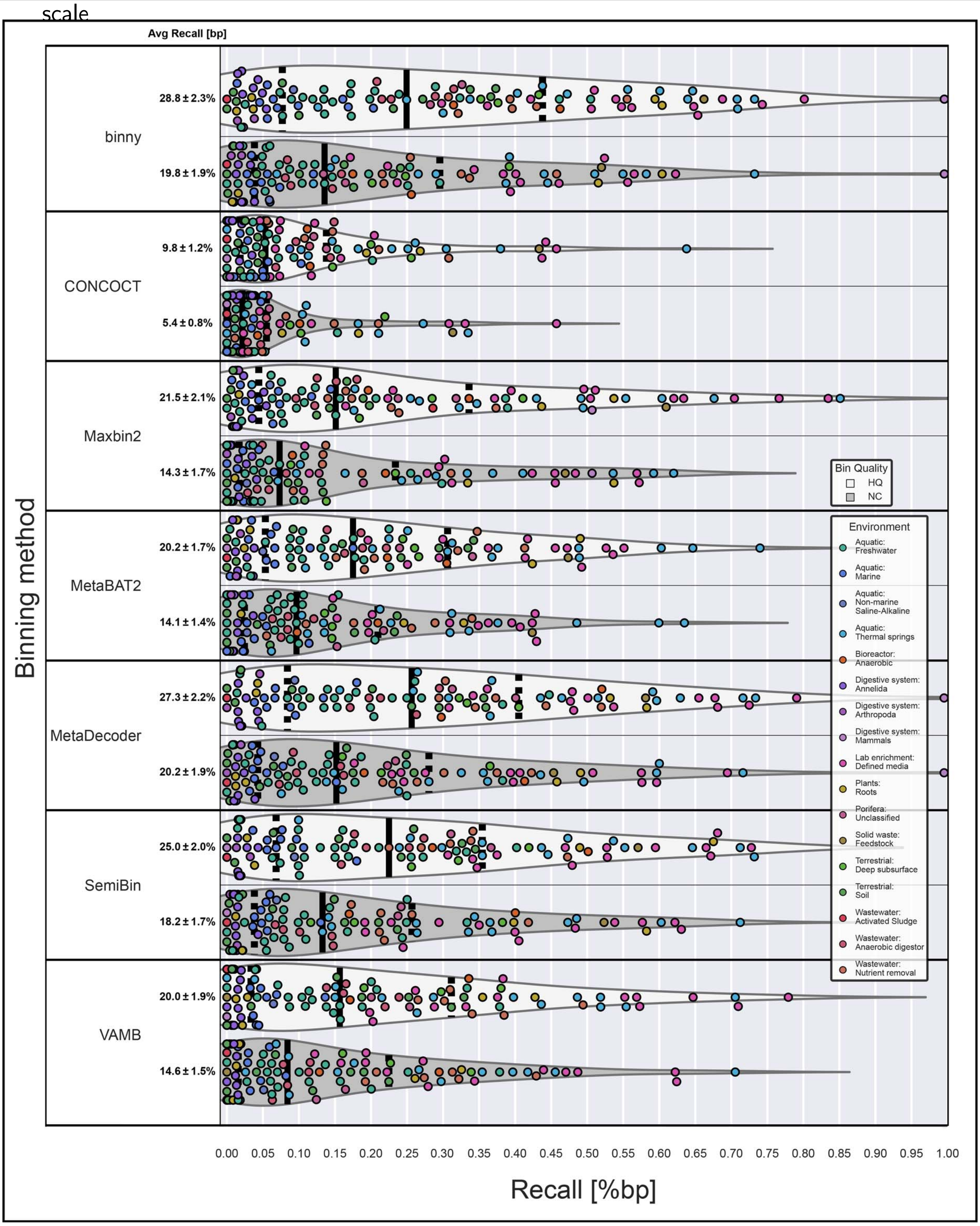


Figure 5. Performance of binning methods on real-world datasets from various environments. Assembly recovery as HQ and NC MAGs per binning method per sample from 105 real-world samples. The average recall (% bp) is shown with the SEM.

worse than *binny* alone (10.8% and 5.5% fewer NC MAGs, respectively). The runs including all three binning methods showed the highest performance overall, with MetaWRAP recovering the most MAGs (Figure 6). Evaluating the HQ MAG recovery the results were similar, but now only MetaWrap with all three binning methods outperformed *binny* (Supplementary Figure 23). While MetaWRAP produced almost no heavily contaminated MAGs, DAS Tool returned large numbers of MAGs with very low purity, despite showing over the entire CAMI benchmark data high purity (Supplementary Figure 20D, Supplementary Table 23).

Discussion

binny is a fully automated binning method, recovering unique information in form of complete, pure MAGs. It combines *k*-mer composition, read coverage and lineage-specific marker gene sets

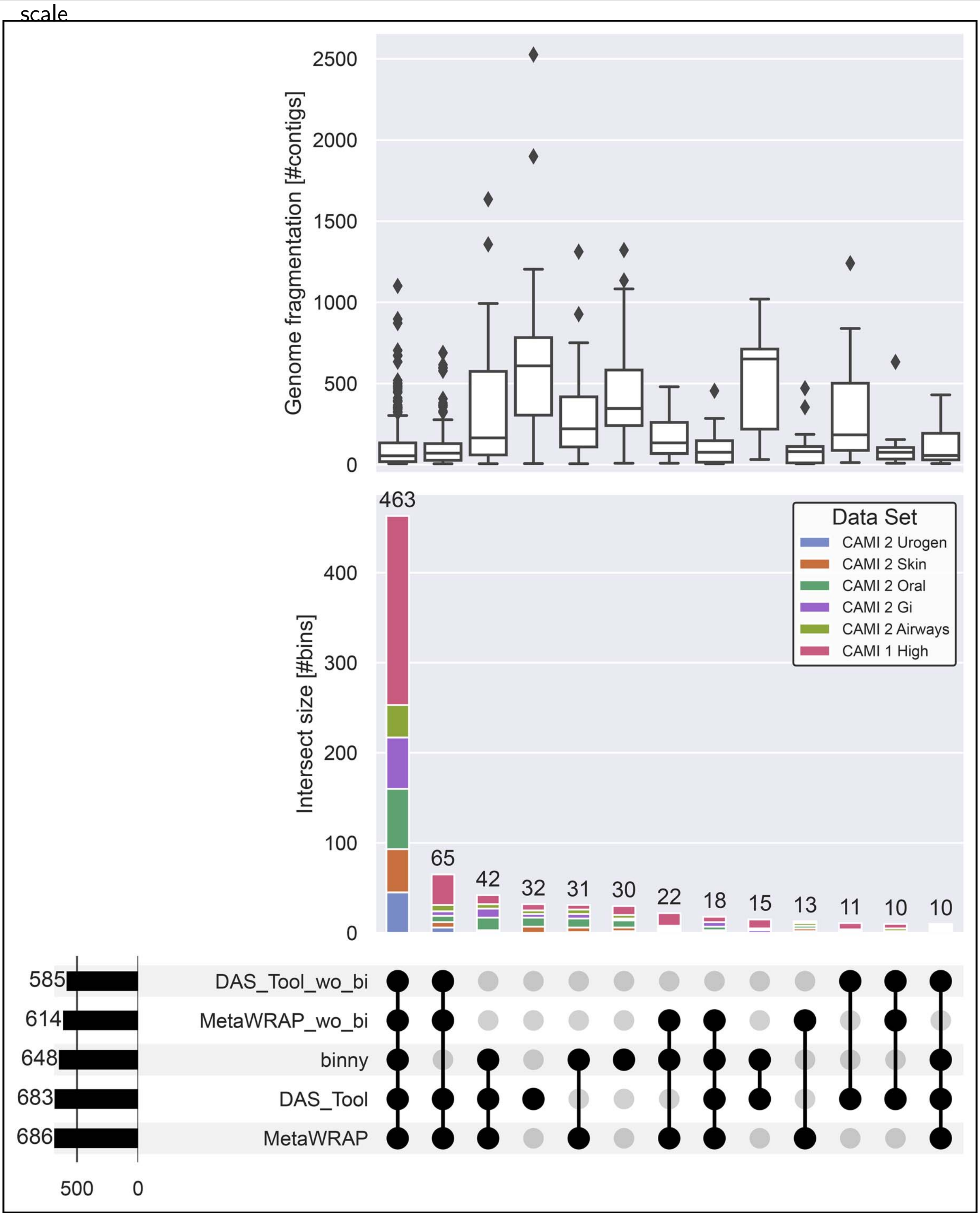


Figure 6. Intersections of recovered CAMI NC MAGs from bin refinement methods. Intersections of NC MAGs from genomes consisting of more than five contigs by *binny*, DAS Tool and MetaWrap for 54 samples from six CAMI datasets. Binning method output used by the refinement methods: *binny*, MetaDecoder and SemiBin or the latter two, but without *binny* (_wo_bi) Upper panel: Reference genome fragmentation in number of contigs. Middle panel: Intersection size in number of NC MAGs with proportions of MAGs stemming from the six CAMI datasets. Lower panel: Number of MAGs per binning method on the left, intersections > 9 in the centre.

for iterative, nonlinear dimension reduction of genomic signatures and subsequent automated contig clustering with cluster assessment. The low-dimensional embedding strategy to reduce large amounts of features has been used before for binning to aid the clustering of contigs [34, 60]. Clustering algorithms perform better in fewer dimensions, because distance information becomes increasingly imprecise at higher dimensions and the chance of random correlation between features rises [61].

While there are already binning methods available that make use of marker genes [14, 38, 62] and also lower dimensional embedding of contig features [62], *binny* uses a new and unique iterative embedding and clustering strategy. Importantly, it assesses clusters of contigs during its iterations, recognizing when further splitting of clusters is necessary. Of note, this lowers the complexity of each clustering task enabling *binny* to recover genomes that might not be separable with only a single scale embedding or clustering attempt. This seems to work particularly well for large, complex communities as shown with different CAMI datasets.

In combination with the ability (enabled by the marker gene approach) to incorporate also short informative contigs, which would be discarded by most other binning methods due to their applied contig length thresholds, *binny* is able to deal with highly fragmented genomes as shown for the CAMI samples. Of the tested binning methods, only CONCOCT was also able to deal with highly fragmented genomes. Although for the CAMI datasets, contigs below 1000 bp rarely made up >5% of the recovered MAGs size, *binny* assigned those usually with high precision (Supplementary Table 25). Additionally, *binny* performed also particularly well at recovering highly contiguous CAMI genomes. This can again be attributed to the ability to assess purity and completeness using the marker gene approach, here in particular for single-contig genomes.

binny also outperformed all other tested binning methods on the CAMI co-assemblies, where the added information provided by the coverage data from multiple samples substantially increased the overall performance. This is in line with previous studies observing additional discriminatory power of differential coverage depth compared with only sequence-based features [13, 15]. Here again, *binny*'s iterative, supervised strategy seems well suited to the complexity of assemblies that contain highly fragmented genomes.

We also evaluated the effect of masking potentially disruptive sequence regions for the calculation of *k*-mer profiles. While the difference in performance with and without masking was not substantial, we believe that it reduces noise in the *k*-mer distributions of contigs from the same genome. One key reason for the small impact in the current setting might be the strong effect of the read coverage depth on the embedding and clustering procedure, which could outweigh the impact of the masked *k*-mer profile. Masking reads mapping to the disruptive regions, also modifying the depth information, might increase its effectiveness and could be implemented in future versions.

It is generally advised [18, 63] to make use of refinement methods, such as DAS tool and MetaWRAP here, which employ ensemble approaches to recover more pure and complete MAGs than the single binning methods alone. *binny* was shown to be an excellent addition to such approaches, because of its ability to recover large amounts of unique pure and complete MAGs (Figures 3 and 5).

Finally, the results of the 105 metagenome benchmark show that *binny*'s performance translates to real-world scenarios, competing well with the latest methods on the recovery of NC and HQ MAGs, while massively outperforming all other methods on the number of MIMAG-standard MAGs recovered. Still, there are also many samples where all binning methods were unable to recover a sizeable proportion of the assemblies as MAGs of sufficient quality. This might hint at the still limited capabilities of binning methods or could be caused by low quality of these assemblies.

Conclusion

In conclusion, we demonstrate that *binny* outperforms or is highly competitive with currently available, state-of-the-art and/or popular binning methods based on established evaluation metrics, recovering unique, complete, and pure MAGs from simple and complex samples alike, while being able to handle contiguous, as well as fragmented genomes. Moreover, we could show that *binny* adds new MAGs when used in combination with other binning methods and binning refinement approaches, enabling researchers to further improve the recovery of genomes from their metagenomes.

Key Points

- *binny* outperforms or is highly competitive with commonly used and recently developed genome reconstruction tools.
- *binny* is benchmarked using community-standard simulations and a wide range of real-world metagenomes.
- *binny* efficiently and iteratively learns using lineage-specific markers and selected genomic features.

Data availability

The latest version of *binny* can be found at https://github.com/a-h-b/binny. Scripts used in this study and related data are available at https://github.com/ohickl/binny_manuscript and https://doi.org/10.5281/zenodo.6977322.

Author contributions statement

O.H., P.M. and A.H.-B. designed this study. O.H. and A.H.-B. created the application. O.H. performed all experiments. O.H., P.Q., P.M. and A.H.-B. were involved in creating the workflow. O.H., P.M. and A.H.-B. wrote the manuscript; P.Q. and P.W. contributed to the review of the manuscript before submission. All authors read and approved the manuscript.

Acknowledgements

The authors would like to thank Francesco Delogu and Benoit Kunath, for scientific discussions. Development and data analysis were performed using the research cluster of the Faculty of Science at the University of Amsterdam and the HPC facilities of the University of Luxembourg, both of whose administrators are thanked for their excellent support. We also thank Adrian Fritz of the CAMI team for support.

Funding

Luxembourg National Research Fund (FNR) (grant PRIDE/11823097); European Research Council (ERC-CoG 863664).

References

1. Quince C, Walker AW, Simpson JT, *et al.* Shotgun metagenomsics, from sampling to analysis. *Nat Biotechnol* 2017;**35**(9):833–44.
2. New FN, Brito IL. What Is Metagenomics Teaching Us, and What Is Missed? *Annu Rev Microbiol* 2020;**74**:117–35.
3. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 2017;**541**(7637):353–8.
4. Delmont TO, Quince C, Shaiber A, *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* 2018;**3**(7):804–13.
5. Shen L, Liu Y, Allen MA, *et al.* Linking genomic and physiological characteristics of psychrophilic arthrobacter to metagenomic data to explain global environmental distribution. *Microbiome* 2021;**9**(1):136.
6. Almeida A, Nayfach S, Boland M, *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 2021;**39**(1):105–14.

scale

7. Nayfach S, Roux S, Seshadri R, *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;**39**(4):499–509.
8. Tett A, Huang KD, Asnicar F, *et al.* The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* 2019;**26**(5):666–679.e7.
9. Karcher N, Nigro E, Punčochář M, *et al.* Genomic diversity and ecology of human-associated Akkermansia species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol* 2021;**22**(1):209.
10. Heintz-Buschart A, May P, Laczny CC, *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* 2016;**2**:16180.
11. Herold M, Arbas SM, Narayanasamy S, *et al.* Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun* 2020;**11**(1):5281.
12. Chen L-X, Anantharaman K, Alon Shaiber A, *et al.* Accurate and complete genomes from metagenomes. *Genome Res* 2020; **30**(3):315–33.
13. Alneberg J, Bjarnason BS, de Bruijn I, *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; **11**(11):1144–6.
14. Yu-Wei W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;**32**(4):605–7.
15. Kang DD, Li F, Kirton E, *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359.
16. Meziti A, Rodriguez-R LM, Hatt JK, *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* 2021;**87**(6):e02593–20.
17. Sczyrba A, Hofmann P, Belmann P, *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* 2017;**14**(11):1063–71.
18. Meyer F, Fritz A, Deng Z-L, *et al.* Critical assessment of metagenome interpretation: the second round of challenges. *Nat Methods* 2022;**19**(4):429–40.
19. Meyer F, Lesker T-R, Koslicki D, *et al.* Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat Protoc* 2021;**16**(4):1785–801.
20. Na S-I, Kim YO, Yoon S-H, *et al.* UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *Journal of Microbiology (Seoul, Korea)* 2018;**56**(4):280–5.
21. Brown CT, Hug LA, Thomas BC, *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015;**523**(7559):208–11.
22. Rinke C, Schwientek P, Sczyrba A, *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**(7459):431–7.
23. Parks DH, Imelfort M, Skennerton CT, *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**(7):1043–55.
24. Bowers RM, Nikos C, *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017; **35**(8):725–31.
25. Mitchell AL, Almeida A, Beracochea M, *et al.* (eds). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* 2019;gkz1035.
26. Almeida A, Mitchell AL, Tarkowska A, *et al.* Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience* 2018;**7**(5).
27. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018;**6**(1):158.
28. Sieber CMK, Probst AJ, Sharrar A, *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;**3**(7):836–43.
29. Yue Y, Huang H, Qi Z, *et al.* Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC bioinformatics* 2020;**21**(1):334.
30. Murat Eren A, Esen OC, Quince C, *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015; **3**:e1319.
31. Broeksema B, Calusinska M, McGee F, *et al.* ICoVeR - an interactive visualization tool for verification and refinement of metagenomic bins. *BMC bioinformatics* 2017;**18**(1):233.
32. Bornemann TLV, Esser SP, Stach TL, *et al.* uBin-a manual refining tool for metagenomic bins designed for educational purposes. preprint. *Genomics* 2020.
33. Murat Eren A, Kiefl E, Shaiber A, *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021; **6**(1):3–6.
34. Laczny CC, Sternal T, Plugaru V, *et al.* VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* 2015;**3**(1):1.
35. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)* 2018;**34**(20): 3600.
36. Nissen JN, Johansen J, Allesøe RL, *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;**39**(5):555–60.
37. Pan S, Zhu C, Zhao X-M, *et al.* A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat Commun* 2022;**13**(1): 2326.
38. Liu C-C, Dong S-S, Chen J-B, *et al.* Metadecoder: a novel method for clustering metagenomic contigs. *Microbiome* 2022;**10**(1):46.
39. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* 2014;**30**(14):2068–9.
40. Queirós P, Delogu F, Hickl O, *et al.* Mantis: flexible and consensus-driven genome annotation. *GigaScience* 2021;**10**(6): giab042.
41. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds). *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA, 2008, 11–5.
42. Bland C, Ramsey TL, Sabree F, *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics* 2007;**8**: 209.
43. Poličar PG, Stražar M, Zupan B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding-bioRxiv. 2019.
44. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* 2019;**10**(1):5416.
45. Linderman GC, Steinerberger S. Clustering with t-SNE, Provably. *SIAM Journal on Mathematics of Data Science* 2019;**1**(2):313–32.
46. Belkina AC, Ciccolella CO, Anno R, *et al.* Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun* 2019;**10**(1):5415.
47. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Goos G, Hartmanis J, van Leeuwen J, *et al.*, editors, *Database Theory-ICDT 2001*, volume **1973**, pages 420–34. Springer Berlin Heidelberg,

scale

Berlin, Heidelberg, 2001. Series Title: Lecture Notes in Computer Science.

48. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In Hutchison D, Kanade T, Kittler J, *et al.*, editors, *Advances in Knowledge Discovery and Data Mining*, volume **7819**, pages 160–72. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 2010;**26**(6):841–2.
50. Hyatt D, Chen G-L, Locascio PF, *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 2010;**11**:119.
51. Mistry J, Chuguransky S, Williams L, *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021;**49**(D1): D412–9.
52. Li W, O'Neill KR, Haft DH, *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;**49**(D1):D1020–8.
53. Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*arXiv:1303.3997 [q-bio], May 2013, arXiv: 1303.3997.
54. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**(1):59–60.
55. Meyer F, Hofmann P, Belmann P, *et al.* AMBER: Assessment of Metagenome BinnERs. *GigaScience* 2018;**7**(6).
56. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 2007;**9**(3):90–5.
57. Waskom M. seaborn: statistical data visualization. *Journal of Open Source Software* 2021;**6**(60):3021.
58. Lex A, Gehlenborg N, Strobelt H, *et al.* UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph* 2014;**20**(12): 1983–92.
59. Jain C, Rodriguez-R LM, Phillippy AM, *et al.* High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**(1):5114.
60. Ceballos J, Ariza-Jiménez L, Pinel N. Standardized approaches for assessing metagenomic contig binning performance from barnes-hut t-stochastic neighbor embeddings. In: González Díaz CA, González CC, Leber EL *et al.* (eds). *VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering*. Cham: Springer International Publishing, 2020, 761–8.
61. Kriegel H-P, Kröger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data* 2009;**3**(1):1–58.
62. Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* 2016;**6**:24175.
63. Chen L-X, Anantharaman K, Alon Shaiber A, *et al.* Accurate and complete genomes from metagenomes. *Genome Res* 2020; **30**(3):315–33.

5.2

Mantis: flexible and consensus-driven genome annotation

5.2.1 Coversheet

Contributions of Oskar Hickl

- Discussed and refined the implementation details and methodology behind the development of the tool. This involved collaborating with other authors to ensure the design was efficient and met its intended goals.
- Helped design the benchmarking process for evaluating the performance, efficiency, and scalability of the tool. This included selecting appropriate metrics and defining test scenarios that accurately reflect real-world usage conditions.
- Assisted writing the manuscript, detailing the development process, methodology, and results obtained from using the tool.
- Conducted testing of the tool with various parameters to identify any issues or limitations that might affect its performance and functionality. These tests focused on aspects such as installation, running, and overall stability, which are necessary for ensuring a positive user experience. Additionally, recommendations were made based on these findings to improve the usability and effectiveness of the tool.

5.2.2 Introduction

With the constant increase in high-resolution genomic data, the need for quality annotations, which facilitate interpretation of the data, is more pressing than ever.To address this, scalable, high-quality protein functional annotation (PFA) is necessary, as it is key to assessing the functional potential of organisms and thus understanding their biology as well as their environments. Numerous systems exist in this field, but challenges remain, including the need for ensuring high quality of annotations, computational efficiency, adaptability, and reproducibility.

One key challenge in PFA is extracting information from a variety of reference sources without over-relying on a single source, which may have limitations in quality, scope, or resolution. Mantis addresses this by using database identifier intersections in conjunction with text mining. This method integrates information from various references to produce a unified, consensus-driven output. The generated annotations are exhaustive and remain consistent across different research configurations.

Briefly, Mantis employs a six-step workflow for protein function annotation, involving: sample pre-processing, homology searches, hit processing at both intra and inter-HMM levels, metadata integration, and consensus annotation generation. The users can specify organism taxonomy for targeted annotations as well as reference sources of choice, which are automatically integrated, in addition to the standard reference sources such as Pfam, eggNOG, NPFM, KOfam, and TIGRfam.

Annotation performance was assessed through *in silico* tests with curated protein entries from the UniProt database. The evaluations considered various parameters, including e-value thresholds and hit-processing algorithms, and the influence of text mining.

Performance in real-world scenarios was evaluated by annotating known sequenced organisms, also taking into account the utility of taxon-specific HMMs for accurate annotations. Comparative evaluations positioned Mantis in relation to other PFA tools, underlining its precision in protein sequence annotation. In summary, Mantis integrates multiple reference databases to provide a systematic approach to PFA. Its evaluations in controlled and practical settings show its potential as a valuable addition to the functional annotation tool set.

5.2.3 Manuscript

GigaScience, 10, 2021, 1–14

https://doi.org/10.1093/gigascience/giab042
Technical Note

TECHNICAL NOTE

Mantis: flexible and consensus-driven genome annotation

Pedro Queirós [1,*], Francesco Delogu [1], Oskar Hickl [2], Patrick May [2,*] and Paul Wilmes [1,*]

[1]Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, 4367 Esch-sur-Alzette, Luxembourg and [2]Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, 4367 Esch-sur-Alzette, Luxembourg

*Correspondence address. Pedro Queirós, Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Université du Luxembourg 2, Avenue de l'Université L-4365 Esch-sur-Alzette. E-mail: pdqueiros@gmail.com http://orcid.org/0000-0002-0831-4261; Paul Wilmes, Systems Ecology, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Université du Luxembourg 7, Avenue des Hauts Fourneaux L-4362 Esch-sur-Alzette. E-mail: paul.wilmes@uni.lu http://orcid.org/0000-0001-8698-3770

Abstract

Background: The rapid development of the (meta-)omics fields has produced an unprecedented amount of high-resolution and high-fidelity data. Through the use of these datasets we can infer the role of previously functionally unannotated proteins from single organisms and consortia. In this context, protein function annotation can be described as the identification of regions of interest (i.e., domains) in protein sequences and the assignment of biological functions. Despite the existence of numerous tools, challenges remain in terms of speed, flexibility, and reproducibility. In the big data era, it is also increasingly important to cease limiting our findings to a single reference, coalescing knowledge from different data sources, and thus overcoming some limitations in overly relying on computationally generated data from single sources.
Results: We implemented a protein annotation tool, Mantis, which uses database identifiers intersection and text mining to integrate knowledge from multiple reference data sources into a single consensus-driven output. Mantis is flexible, allowing for the customization of reference data and execution parameters, and is reproducible across different research goals and user environments. We implemented a depth-first search algorithm for domain-specific annotation, which significantly improved annotation performance compared to sequence-wide annotation. The parallelized implementation of Mantis results in short runtimes while also outputting high coverage and high-quality protein function annotations.
Conclusions: Mantis is a protein function annotation tool that produces high-quality consensus-driven protein annotations. It is easy to set up, customize, and use, scaling from single genomes to large metagenomes. Mantis is available under the MIT license at https://github.com/PedroMTQ/mantis.

Keywords: bioinformatics; consensus; homology; HMM; protein function annotation

Background

On a cellular scale, life is, in essence, the activity and the interaction of a plethora of different molecules, among which proteins are responsible for the vast majority of processes. A primary task in understanding how biology works is to resolve its actors properly (e.g., the proteins) and place them into context. The past decades have seen the development of the (meta-)omics fields, unlocking an unprecedented amount of data and deepening our understanding in several fields of biology [1, 2]. Alongside the evolution of the technologies and the increase in data volume, the identification of proteins transitioned from purely experimental techniques (e.g., chemical assays and spectroscopy)

Received: 4 November 2020; **Revised:** 22 March 2021; **Accepted:** 14 May 2021

scale

toward computational-based sequence analysis thanks to the discovery of the relationship between conservation of proteins' functions and sequences [3]. Therefore, the current challenges are to make use of the vast number of protein sequences and annotations available and to link new protein sequences to the previously established knowledge. High-throughput methods, such as next-generation sequencing, are able to produce a large amount of data, which then need to be analysed and interpreted. One of the ways to make sense of these data is through protein function annotation (PFA), which is, in the context of this article, the identification of regions of interest (i.e., domains) in a sequence and assignment of biological function(s) to these regions. This strategy has proven effective in the study of single organisms, as well as consortia [4–9]. Function prediction is based on reference data, i.e., transferring the function from protein X to the unknown protein Y if they are highly similar [3]. Different approaches may be used, the most common being the comparison of an unknown protein sequence to reference data composed of well-studied and functionally annotated proteins (homology-based methods) [10–16]. Other methods may infer function through the use of machine learning [10, 17], protein networks [18, 19], protein structure [20], or genomics context-based techniques [21], but these are not covered in this article. For sequence alignment, BLAST [22] or Diamond [23] are commonly used, whereas, for hidden Markov models (HMM) profiles, HMMER [24] is most widely used. In PFA, these tools are often integrated into larger pipelines to provide enhanced output interpretability, workflow automation, and parallelization [14–16, 25]. Some PFA tools target specific taxa [26], while others are designed with large-scale omics analysis in mind [27–29]; indeed, each PFA tool is designed to cater to its niche research topic. While experimental validation remains the gold standard, PFA, despite its many shortcomings [30], is an increasingly valuable strategy that aims to tackle the progressively more difficult task of making sense of the large quantities of data being continuously generated.

The most common method of processing candidate annotations (i.e., sequences or HMM profiles that are highly similar to the query sequence) involves capturing only the most significant candidate ("best prediction only" [BPO] algorithm). This PFA approach works well for single-domain proteins, but multi-domain proteins may have multiple putative predictions [31–33], whose location in the sequence may or may not overlap. This selection criterion may potentially lead to missing annotations and is therefore not suitable in complex PFA scenarios. To tackle this problem, domain-specific PFA is necessary. A simple approach, previously discussed in Yeats et al. [31], would be to order the predictions by their significance and iteratively add the most significant one, as long as it does not overlap with the already added predictions (henceforth referred to as the "heuristic" algorithm). Owing to the biased selection of the first prediction, this algorithm does not guarantee an optimal solution (e.g., a protein sequence may have multiple similarly significant predictions). It has been previously shown that incorporating prediction significance and length may produce better results [34]. We implemented a depth-first search (DFS) algorithm that improves on the previous approaches.

The selection of reference HMMs is also critical because PFA will ultimately be based on the available reference data. Whilst using unspecific HMMs to annotate a taxonomically classified sample may result in a fair amount of true-positive (TP) results (correct annotations), depending on the confidence threshold used, it may also increase the rate of false-positive (FP) results (over-annotation, due to a less strict confidence threshold) or false-negative (FN) results (under-annotation, due to a more strict confidence threshold) [35]. Using taxon-specific HMMs (TSHMM) rather than unspecific HMMs should, in principle, provide better annotations on a taxonomically classified sample, a feature that is already integrated into some PFA tools such as eggNOG-mapper [15] and RAST [16]. In essence, TSHMM-based annotation limits the available search space, which may have positive and negative consequences. Because the search space is more specific, the annotations produced should be of higher quality; however, this higher specificity of the TSHMM could also lead to under-annotation (incomplete reference TSHMMs) or mis-annotations (low-quality reference TSHMM) [36]. This underlines the necessity to use specific (e.g., TSHMMs) and unspecific HMMs in a complementary manner. In this regard, the use of multiple sources of reference data remains a challenging aspect of PFA, and, with multiple high-quality reference data sources available, it is increasingly important to coalesce knowledge from different sources. While some PFA tools allow for the use of multiple reference data sources, either as a separate [25] or a unified [15, 37] database, it is still challenging to integrate multiple data sources dynamically.

When using reference data from multiple high-quality sources, the most common and straightforward approach is to consider the output from each reference data source independently (e.g., [25]). However, by doing so, we overlook that many sources can overlap and/or complement each other. Commonly this is compensated for via manual curation, which is feasible only for a limited number of annotations. An automated approach would be to assume only the most significant annotation source for any given sequence and disregard other sources; this may result in vast losses of potentially valid and complementary information (e.g., database identifiers). Because this is not desirable, the challenge is in both deciding which source(s) provide the best annotation as well as identifying complementary annotations. In the present context, complementary annotations can be defined as functional annotations that are functionally similar but originate from difference data sources; as such, while functionally similar, different data sources are likely to contain information that is absent in other data sources and vice versa. This unique functional information (i.e., database identifiers or functional descriptions) may prove essential in downstream data analysis. A straightforward approach to verify whether functional annotations are functionally similar is to check whether they share a database identifier (ID), e.g.,

(i) Function: "Responsible for glucose degradation"; IDs: K00844, EC:2.7.1.1, PF03727
(ii) Function: "Responsible for glucose degradation"; IDs: P52789, PF03727, IPR022673

We can observe that the annotations (i) and (ii) share the database ID PF03727, thus it can be concluded that these annotations are functionally similar. If we were only to select the first annotation, we would ignore potentially useful information (IDs P52789 and IPR022673). However, it may be the case that no IDs are shared between the different annotations, e.g.,

(i) Function: "Responsible for glucose degradation"; IDs: K00844, EC:2.7.1.1
(ii) Function: "Responsible for glucose degradation"; IDs: P52789, IPR022673

We can observe that even though the annotations (i) and (ii) no longer share an ID, they still have the same function "Responsible for glucose degradation." Humans can quickly sur-

scale

mise that these annotations are the same because they share the same function description. Should the descriptions be identical or very similar, a machine could achieve the same conclusion with relative ease. However, in our experience, these free-text functional descriptions are often moderately or heavily dissimilar [38, 39], with only a few keywords allowing us to ascertain that they are indeed the same. This then makes it more difficult to use multiple reference data sources. For example:

(i) Function: "Responsible for glucose degradation"; IDs: K00844, EC:2.7.1.1
(ii) Function: "Protein is an enzyme and it is responsible for the breakdown of glucose"; IDs: HXK2_HUMAN

In such a scenario, someone trained in a biology-related field can quickly identify the most important words ("degradation"/"breakdown" and "glucose") in both sentences and conclude that both annotations point to the same biological function. The challenge is now to enable a machine, deprived of any intellect and intuition, to eliminate confounders (ubiquitous words, e.g., "the"), identify keywords and their potential synonyms, and reach the same conclusion. A possible strategy is to use text mining, which is the process of exploring and analysing large amounts of unstructured text data aided by software, identifying potential concepts, patterns, topics, keywords, and other attributes in the data [40]. Text mining has been previously used with biological data [41–45], and even more specifically with regards to gene ontologies [46–51] and PFA [43]. However, to our knowledge, there is no tool for the dynamic generation of a consensus from multiple protein annotations. This article solves the problem of scaling the integration of different annotation sources, integrating a compact and flexible text-mining strategy. We implemented a 2-fold approach to build a consensus annotation, first by checking for any intersecting annotation IDs and second by evaluating how similar the free-text functional descriptions are. This approach attempts to address 3 relevant issues with PFA [35, 36, 52, 53]: over-annotation, under-annotation, and redundancy. Another challenge in PFA is the lack of flexibility of some tools, as these are often intrinsically connected to their in-house–generated reference data and therefore hard to customize. In contrast, we developed a tool that, while offering high-quality unspecific and specific HMMs, is independent of its reference data, thus being customizable and allowing dynamic integration of new data sources.

We hereby present Mantis, a Python-based PFA tool that overcomes the previously presented issues, producing high-quality annotations with the integration of multiple domains and multiple reference data sources. Mantis automatically downloads and compiles several high-quality reference data sources and efficiently uses the available hardware through parallelized execution. Mantis is independent of any of the default reference data, resulting in a versatile and reproducible tool that overcomes the challenge of high-throughput protein annotation coming from the many genome and metagenome sequencing projects.

Mantis

Mantis is available at https://github.com/PedroMTQ/mantis, and its workflow (see Fig. 1) consists of 6 main steps: (i) sample pre-processing, (ii) HMM profile-based homology search, (iii) intra-HMM hit processing, (iv) metadata integration, (v) inter-HMM hit processing, and (vi) consensus generation. For future reference, an instance when an HMM matches with a protein sequence is referred to as a "hit." The workflow starts with sample pre-processing, in which the sample(s) is/are split into chunks. This is followed by homology search, where query sequences are searched against the available reference data using HMMER. During intra-HMM hit processing the DFS algorithm is used to generate and select the best combination of hits per HMM source; Fig. 2 shows how different algorithms may lead to a different selection of hits. Metadata integration adds the metadata (functional description and IDs) to the respective hits. During inter-HMM hit processing, the DFS algorithm is used to generate all the combinations of hits from all HMM sources (in this step all hits are pooled together). Finally, consensus generation ensures that the best combination of hits among all hits from the multiple reference data sources is selected. This combination is expanded by adding additional hits with consistent metadata (intersecting identifiers or similar functional descriptions) (see Methods section for a detailed description of all these steps). We provide default execution parameters; however, the user is free to fully customize Mantis, not only the parameters but also the reference databases used. Mantis requires a FASTA-formatted protein sequence file as input, where the user can also provide the organism's taxon to allow for taxon-specific annotation (TSA). Reference databases are downloaded automatically. The MANTIS.config file allows for configuration of the reference data and its respective weights and enables the compilation of specific eggNOG TSHMMs. For more details, see the documentation [54]. Owing to issues with Python's multiprocessing in MacOS, and the fact that HMMER is not available on Windows, Mantis is only available on Linux-based systems.

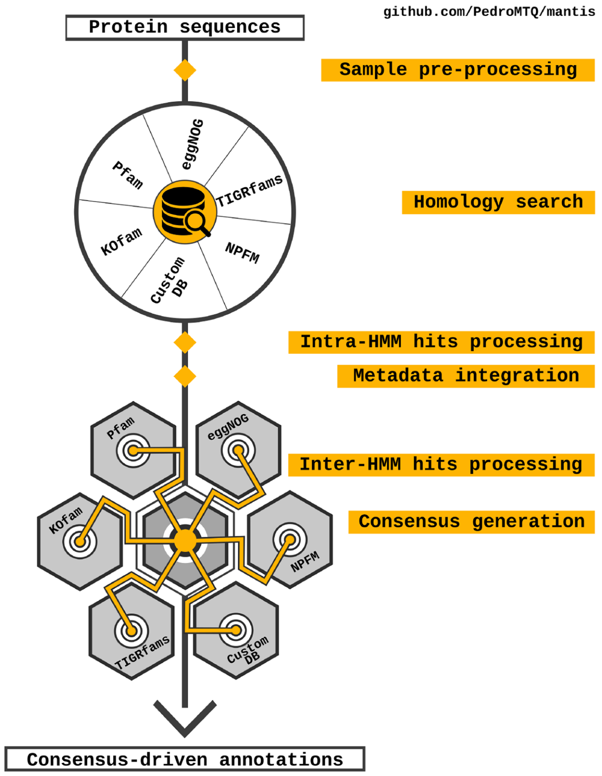


Figure 1: Overview of the Mantis workflow. KOfam [55], Pfam [56], eggNOG [57], NCBI protein family models (NPFM) [58], and TIGRfams [59] are the reference HMMs currently used in Mantis. CustomDB can be any HMM library provided by the user.

scale

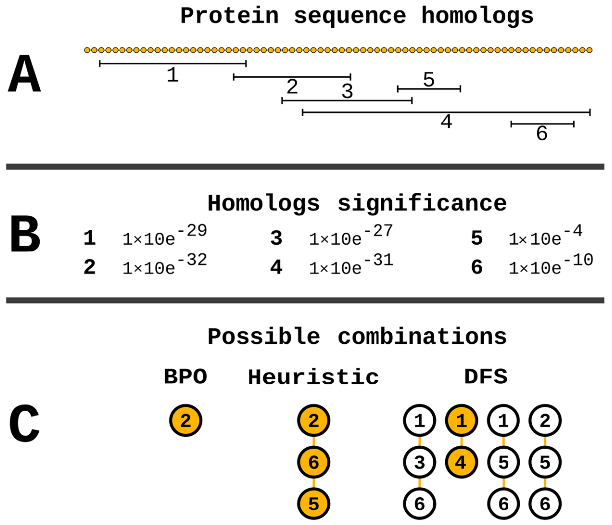


Figure 2: Homolog selection for the 3 hit-processing algorithms in Mantis. The selection of the hit(s) depends on the underlying algorithm. In the case of the portrayed protein with 6 hits (A) (which are overlapping to various degrees) that have varying significance values (B) the 3 algorithms would behave as follows: (i) BPO would select only the most significant hit (No. 2); (ii) the heuristic algorithm initially selects the most significant hit (No. 2), which then restricts (due to overlapping residues) the hits available for selection (hits 1, 3, and 4 can no longer be selected), leading to the selection of the next most significant hit (No. 6), and finally the selection of hit 5; (iii) the DFS algorithm generates all possible combinations of hits, which are then scored according to the e-value, hit coverage, and total combination coverage (for more details, see "Multiple hits per protein"). According to these parameters, the most likely combinations of hits would be hits 1 and 4.

Analysis

To analyse and validate the performance of Mantis, we performed several *in silico* experiments. We annotated a reference dataset containing curated protein entries from UniProt to set default parameters and evaluate the impaect of different Mantis features: (i) impact of the e-value threshold, (ii) impact of the hit-processing algorithm, (iii) how each reference data source contributed to the final output, and (iv) impact of the consensus generation on annotation quality. Furthermore, we annotated several sequenced organisms, with and without TSHMMs, thus evaluating the impact of using taxon-resolved reference data. Finally, we compared Mantis against eggNOG-mapper [15] and Prokka [14]. A description of the samples used for this benchmark is available in "Sample selection." Prokka was only used for the annotation of prokaryotic data (i.e., all except for *Saccharomyces cerevisiae* and *Cryptococcus neoformans*). To compare the performance between the different tests, we calculated a confusion matrix for each test. For future reference, a TP occurs when a functional annotation (predicted from a PFA tool) shares ≥ 1 database ID with the respective reference annotation (e.g., Pfam ID), an FP when no database IDs are shared, an FN when the PFA tool does not annotate a protein sequence but a reference annotation is available, and a true-negative (TN) when the PFA tool does not annotate a protein sequence and no reference annotation is available. Precision is defined as TP/(TP + FP), recall as TP/(TP + FN), and F1 score (harmonic mean of precision and recall) as 2 × [(precision × recall)/(precision + recall)]. The F1 score is used as a performance metric. Further details on the benchmark are available in "Establishing a test environment."

Initial quality control

Function assignment e-value threshold

It is known that the e-value threshold directly affects annotation quality; however, no gold standard threshold exists [34]. Depending on the reference data source's size, quality, and specificity, we may use more or less stringent thresholds. It is therefore essential to test annotation quality with different thresholds. As such, we tested different static e-value thresholds and a dynamic threshold, which is described in "Testing different e-value thresholds." As can be seen in Supplementary Table 1, precision was similar across the range of e-value thresholds tested, with recall/sensitivity decreasing with lower e-value thresholds. Unexpectedly, unlike recall, precision was not directly correlated with the e-value threshold; indeed a maximum precision of 0.747 was obtained for the e-value threshold $1e^{-6}$, with precision slightly decreasing with more stringent e-value thresholds. A maximum F1 score of 0.827 was observed for the e-value threshold $1e^{-3}$; as such, we chose this value as the default e-value threshold for Mantis.

Impact of hit-processing algorithms

To understand whether the different hit-processing algorithms resulted in statistically significant differences in F1 scores, we created synthetic samples and performed pairwise comparisons between the DFS and the other algorithms: (i) DFS and heuristic and (ii) DFS and BPO. We rejected the H_0: "no differences in F1 score between the tested algorithms" in both comparisons because $P < 0.01$. The DFS algorithm resulted in a greater F1 score (mean = 0.827) than the heuristic (mean = 0.826) and BPO (mean = 0.816) algorithms. Further details on results can be found in Supplementary Table 2, and further details on the testing method can be found in "Testing hit-processing algorithms."

Impact of sample selection

Testing exclusively against well-annotated organisms is a recurring issue with protein annotation benchmarking, resulting in the re-annotation of sequences already present in the reference data used, leading to a biased annotation quality evaluation. To avoid this bias, we downloaded all the curated UniProt (i.e., Swiss-Prot) protein entries (as of 14 April 2020) and selected entries by their creation date such that we have 4 samples that contain protein entries created in different date ranges (2010–2020, 2015–2020, 2018–2020, and 2020). Samples with more recent protein entries are increasingly more likely to lack any proteins used to generate Mantis's reference data, which increases the likelihood that potential annotations are due to true sequence homology (and not to circular re-annotations). We annotated these samples using 3 different hit-processing algorithms (DFS, heuristic, and BPO), determining the impact of each on the F1 score.

As seen in Fig. 3, the F1 score decreased as the sample was restricted to more recent data. As seen in Supplementary Table 3, when comparing the hit-processing algorithms, we found that the DFS algorithm consistently outperformed the other algorithms, with an average F1 score 0.021 and 0.003 higher than the BPO and heuristic algorithms, respectively. In addition, the F1 score difference between the multiple hits algorithms (DFS and heuristic) and the single hit algorithm (BPO) increased as the entries in a sample were restricted to more recent years.

Contribution of the different reference data sources

We analysed each reference data source's contribution to the output annotation for the UniProt 2010–2020 sample. By check-

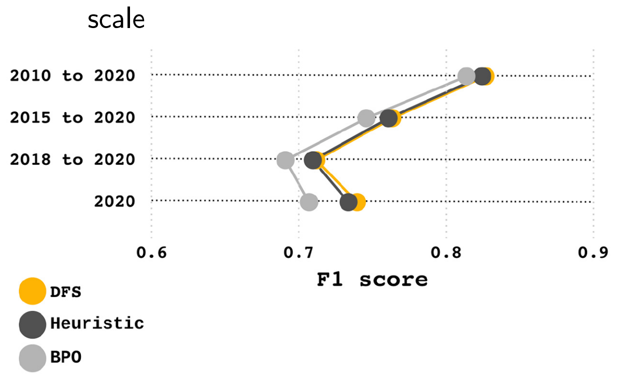


Figure 3: Annotation F1 score per hit-processing algorithm and sample. Overall, the DFS and heuristic algorithms achieve similar results, outperforming the BPO algorithm.

ing the column “HMM_files” in the consensus_annotation.tsv file, we found that Pfam was present in 24.4% of the sequence annotations, KOfam in 62.37%, eggNOG in 76.52%, NPFM in 13.91%, and TIGRfam in 12.96%. Note that, because multiple reference data sources may be present in 1 sequence (due to the consensus generation and hit-processing algorithms), the sum of the previous values is >100%.

Impact of consensus generation

During consensus generation, 2 methods are used for checking the consistency of the hit metadata: ID intersection and text mining. We analysed the contribution of both methods for the annotation of the UniProt 2010–2020 sample and found that 35.10% of the consistency checks were due to the text-mining approach, and the remaining were due to ID intersection.

We also tested the impact of text mining on annotation performance: to do so, we annotated the Uniprot 2010–2020 sample but restricted the consensus generation in different manners and with different algorithms. Six different test conditions were created: (i) DFS with default consensus generation, (ii) DFS with consensus generation restricted to IDs (i.e., ID intersection but no text mining), (iii) DFS without consensus generation (i.e., neither ID intersection nor text mining), (iv) BPO with default consensus generation, (v) BPO with consensus generation restricted to IDs, and (vi) BPO without consensus generation. We also annotated the same sample using eggNOG-mapper—condition (vii). Prokka was not used here because the present sample contains non-prokaryotic data. The F1 scores were as follows: (i) 0.827, (ii) 0.790, (iii) 0.774, (iv) 0.814, (v) 0.779, (vi) 0.763, and (vii) 0.703. Further details can be found in Supplementary Table 4.

Hit-processing approximation

During hit processing, 2 algorithms may be used, the DFS, and, as a backup (if the DFS algorithm’s runtime exceeds 60 seconds), the heuristic. We calculated how many times the heuristic algorithm was used as a backup during the hit processing of the 2010–2020 UniProt sample. We found that for the intra-HMM hit processing, the heuristic algorithm was used in 7.2% of the sequences, and for the inter-HMM hit processing in 0.5% of the sequences.

Quality control with sequenced organisms

As a secondary quality control, to assess the impact on F1 score when using taxon-resolved reference data, we annotated several sequenced organisms (for more details, see Supplementary Table 5) with and without TSHMMs. We also evaluated the impact of the different hit-processing algorithms on these samples. As seen in Fig. 4, well-studied organisms (e.g., *S. cerevisiae*) had better annotations, especially when applying TSHMMs, unlike poorly described organisms. The average F1 score gain with TSHMMs was 0.006. With TSHMMs, the DFS algorithm had, on average, 0.001 and 0.010 higher F1 scores than the heuristic and BPO algorithms, respectively. Without TSHMMs, the DFS algorithm had, on average, 0.008 and 0.013 higher F1 scores than the heuristic and BPO algorithms, respectively. Further details can be found in Supplementary Table 6.

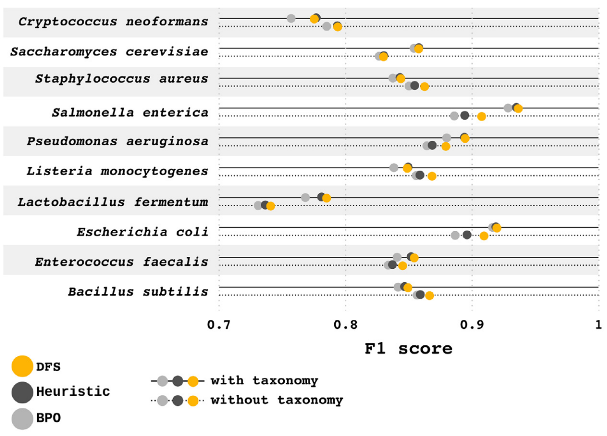


Figure 4: F1 score per hit-processing algorithm and organism, with and without using taxonomy information. F1 score was higher for well-studied organisms; TSHMMs also tend to perform better with these organisms.

Comparison between Mantis and other PFA tools

The sequenced organisms enumerated in Supplementary Table 5 were annotated with Mantis, eggNOG-mapper, and Prokka (for the latter non-prokaryote organisms were excluded). To evaluate the added value of using the very comprehensive eggNOG reference data source, we also assessed Mantis’s F1 score using different reference data. In total, 6 different tests were performed for each organism: (i) Mantis with default data sources and with taxonomy information, (ii) Mantis with default data sources except for eggNOG’s data and with taxonomy information, (iii) Mantis with default data sources but without taxonomy information, (iv) eggNOG-mapper without tax scope option, (v) eggNOG-mapper with tax scope option, and (vi) Prokka with default data sources and default execution.

On average, test (i) had F1 score and annotation coverage of 0.857% and 96.56%, respectively; (ii) 0.832% and 89.82%; (iii) 0.850% and 96.14%; (iv) 0.734% and 88.45%; (v) 0.725% and 88.02%; and (vi) 0.507% and 62.38%. As seen in Fig. 5, Mantis outperformed the other PFA tools in all tests (with 1 exception in the organism *S. cerevisiae*, where eggNOG-mapper without taxonomy had an F1 score of 0.841 and Mantis without taxonomy had an F1 score of 0.830). The mean Mantis F1 score with default execution and TSHMMs was 0.131 higher than eggNOG-mapper (with tax scope) and 0.360 higher than Prokka. Mantis’s setting without the eggNOG reference data had a mean F1 score 0.107 higher than eggNOG-mapper (both tools with taxonomy information) and a mean F1 score 0.025 lower than Mantis’s with the eggNOG reference data. Further details are available in Supplementary Table 7.

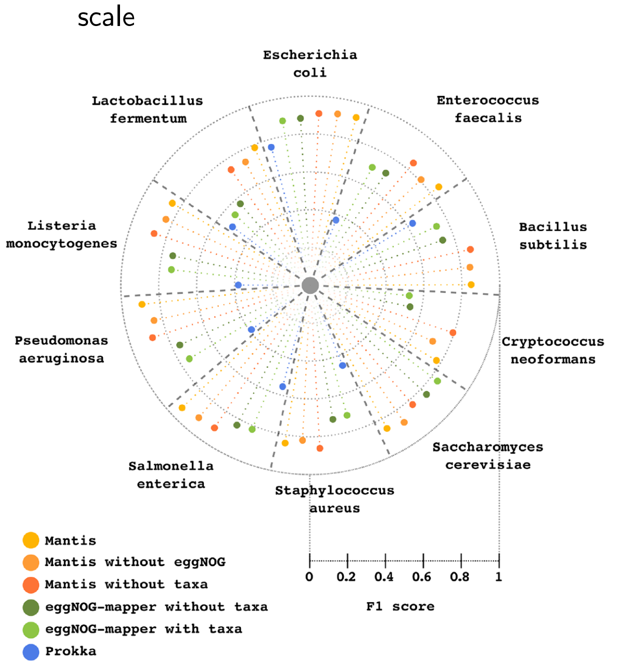


Figure 5: Annotation F1 score of Mantis, eggNOG-mapper, and Prokka using different reference data. Each slice represents an organism and contains the F1 score obtained between the different conditions.

Annotating metagenomes

To our knowledge, there are no manually curated metagenome annotations, therefore annotation validation was not performed; instead we only calculated the annotation coverage. We selected 4 samples from different environments and predicted the protein-coding genes with Prodigal v2.6.3 [60]. The annotated samples were:

- Biogas highly efficient cellulose-degrading consortium (SEM1b) [61, 62] with 39,411 sequences;
- Glacier-fed stream sediment (GFS) [63] with 270,341 sequences (phenol-chloroform extraction batch No. 37);
- Marine [64] with 605,043 sequences (ERR1726751);
- Human gut microbiome (MuSt [7]) with 692,061 sequences (M05-01-V1).

The performance of Mantis varied per metagenome sample; it annotated 213,539, 162,133, 33,016, and 559,792 sequences in the samples GFS, marine, SEM1b, and MuSt, respectively. The respective annotation coverage was as follows: 78.99%, 26.80%, 83.77%, and 80.89%. We repeated the same test for eggNOG-mapper and Prokka (in the case of Prokka by annotating the original nucleotide sequences); the coverage for the samples GFS, marine, SEM1b, and MuSt, was, respectively, 77.52% and 10.87%, 16.21% and 1.01%, 81.95% and 32.32%, and 78.72% and 20.37%.

Computational efficiency

We ran Mantis against samples with a different number of sequences and a different number of available CPUs. We performed this test for the DFS and heuristic algorithm only. As expected, we found that the heuristic algorithm was faster than the DFS algorithm. The heuristic algorithm was, on average, 1.42 times faster than the DFS algorithm. As expected, runtimes were inversely correlated to the number of CPUs and sequences. Further details can be found in Supplementary Table 8.

We also aimed at allowing Mantis to be run on personal computers, which requires removing the eggNOG dataset. However, as we have previously shown in “Comparison between Mantis and other PFA tools,” this does not have a large effect on F1 score. We annotated the previously enumerated sequenced organisms (Supplementary Table 5) on a Dell XPS 13-9370 with Ubuntu 20.04.1 LTS 64 bit, 16 GB RAM, 512 GB SSD, and an 8 core Intel Core it-8550U CPU. The mean runtime for prokaryotes and eukaryotes was 28 and 93 minutes, respectively. Further details are available in Supplementary Table 9.

Discussion

We herein presented Mantis, an open-access PFA tool that produces high-quality annotations and is easily installed and integrated into other bioinformatic workflows. Mantis uses a well-established homology-based method and produces high-quality consensus-driven annotations by relying on the synergy between multiple reference data sources and improved hit-processing algorithms.

Mantis addresses some major challenges in PFA, such as flexibility, speed, the integration of multiple reference data sources, and use of domain-specific annotations. It also addresses under-annotation through the use of multiple reference data sources, which implicitly leads to a wider search space. Additionally, redundancy, which is a drawback inherent to consensus-driven annotation, is ameliorated by removing duplicate database IDs and/or identical descriptions. We have attempted to avoid over-annotation through the generation of a consensus-driven annotation, which identifies and merges annotations that are consistent (i.e., similar function) with each other (e.g., if 3 of 5 independent sources point towards the same function and 2 others point towards other, unrelated functions, then these 3 annotations are more likely to be valid), and eliminating the remaining inconsistent annotations.

We have shown that a stricter/lower e-value threshold did not necessarily lead to a higher F1 score. As expected, a lower threshold restricted the amount of hits, lowering the recall. However, we also found that more stringent e-value thresholds may result in a lower precision; this behaviour is connected to Mantis’s consensus generation and hit combination scoring. A thorough explanation is available in the Supplementary PDF.

Well-curated and commonly used resources were chosen as the default reference data sources for Mantis, containing both unspecific and specific reference data (e.g., taxon-specific). As we have shown, no single reference data source accounted for most annotations, each offering both unique and overlapping insight into protein function, thus confirming their synergy and partial redundancy. These are integrated through a consensus-driven approach, which Mantis uses as an additional quality control step, and a means to automatically incorporate a broader variety of IDs. The intersection of IDs was, as expected, the main contributor towards this integration (because most databases provide cross-linking); however, we found that the text-mining approach still contributed considerably (35.12% for the UniProt 2010-2020 sample), which clearly highlights the need to use such a method.

We additionally evaluated the impact of not using text mining during consensus generation and removing the consensus generation altogether on the DFS and BPO algorithms. The benchmark using the BPO algorithm without consensus generation represented the baseline approach towards the integration of multiple reference data sources (merely selecting the most

scale

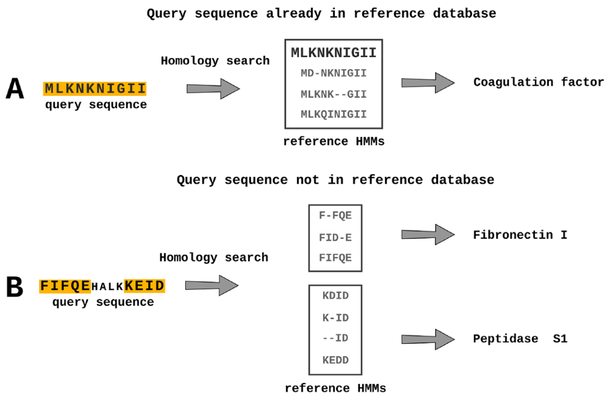


Figure 6: The impact of the reference data completeness on protein function annotation. **A.** The functional prediction is facilitated by the query sequence being previously identified and included in the reference HMMs. **B.** If the query sequence has not been previously annotated, multiple regions in the protein may match with different reference HMMs.

significant hit during inter- and intra-HMM hit processing). In contrast, the benchmark using the DFS algorithm with the consensus generation depicted the accumulation of all the features introduced by Mantis. Overall, we found a difference of 0.064 in F1 scores, which suggests the additive effect of Mantis's various data integration methods. Mantis, in respect to this specific benchmark, also obtained an F1 score higher than eggNOG-mapper in all conditions, which suggests the importance of using multiple reference data sources.

We have implemented 2 algorithms for domain-specific homolog search (DFS and heuristic as backup) and have not only shown that these algorithms perform better when annotating previously described protein sequences but that their impact on the F1 score increased when annotating previously uncharacterized protein sequences (e.g., average F1 score gain with DFS and BPO algorithms in the UniProt 2010–2020 and 2020 samples was 0.013 and 0.033, respectively). We hypothesize that for the latter, a homology search is not capable of finding whole-sequence homologs, finding, however, multiple domains that partially constitute the protein sequence. As such, we argue that by increasing the resolution (sequence homology to domain homology) of homology-based reference data, domain-specific algorithms may become increasingly valuable. We think that this would be especially important when annotating protein sequences without well-described homologs but that contain previously characterized conserved protein domains. In Fig. 6A, we can observe that the present query sequence is already used to generate the HMM profiles in the reference data, matching with the HMM profile containing it. Such a scenario is common when annotating well-described organisms (e.g., *Escherichia coli*). However, as is often the case when annotating non-model organisms and metagenomes, the query sequence is absent from the reference data (Fig. 6B), thus partially matching with several HMMs (which may correspond to multiple domains, depending on the resolution of the reference data). Unlike the BPO algorithm, the heuristic and DFS algorithms are able to incorporate multiple homologs. While these may not be enough to determine a protein's biological function, they still provide a better biological context than a single functional annotation.

Further improvements in annotation quality may also require the use of motif-based and/or genomic context–based (e.g., operon context information, co-expression, and subsystems) methods such as those described by Sigrist et al. [65], Mooney et al. [66], Mavromatis et al. [67], Overbeek et al. [21], and Hannigan et al. [68]. Nevertheless, the significantly higher F1 score seen when comparing the DFS and BPO algorithms highlights the need to adopt better hit-processing methods, especially for non-model organisms. With samples ranging from thousands to millions of protein sequences, sub-optimal hit-processing algorithms may cascade into unnoticeable pitfalls in downstream data analysis (e.g., accumulation of incomplete or low-quality genome annotation, which may lead to false biological interpretations). While we have shown that the DFS algorithm outperforms the heuristic algorithm, both achieve a very similar F1 score when applied to non-synthetic samples; because the heuristic algorithm is much more time efficient (as seen in Supplementary Table 8), a user may confidently set it as primary algorithm.

The use of TSHMMs resulted in a 0.006 higher F1 score; however, this improvement (as seen in Fig. 4) was not consistent across all the annotated organisms (as expected, a similar trend was also seen with eggNOG-mapper). We believe that this is due to a poorer quality of the TSHMMs for some organisms, which is a consequence of the issues with the current taxonomy classification system [69, 70] and lack of knowledge regarding highly resolved taxa [71]. Model organisms such as *E. coli* and *S. cerevisiae* clearly benefited from TSHMMs, both because the reference data already contain data specific to these organisms and because functions of proteins within model organisms are better experimentally described. Conversely, non-model organisms are often only computationally annotated by association, contributing to a weaker reference annotation (which can be observed by the higher rate of potentially new annotations in these organisms, as seen in Supplementary Table 6). Nonetheless, while experimental evidence remains the gold standard, it is unfeasible to ignore the need for computational methods to infer function. While steps in this direction have been taken [16, 57], taxon-resolved PFA remains a challenge.

We benchmarked Mantis against 2 other PFA tools—eggNOG-mapper and Prokka—and have shown that Mantis achieves a higher F1 score (0.131 higher than eggNOG-mapper and 0.350 higher than Prokka). Although Mantis's default execution heavily relies on the eggNOG reference data, we have also shown that even without it, it is possible to achieve an almost similar F1 score. This attests to the quality of the various reference data used, showcasing as well the possibility of running Mantis on a personal computer (something that would be impossible with eggNOG's prohibitive size).

We also evaluated the annotation coverage of Mantis and the other PFA tools when annotating metagenomes. Mantis had the highest annotation coverage among the tested PFA tools, but eggNOG-mapper was close behind. All PFA tools had a low annotation coverage for the marine sample. We believe that this may be due to a lack of reference HMMs for this specific environment. This metagenomic sample has data from varying ocean depths, with many novel sequences from viruses, prokaryotes, and picoeukaryotes [64].

Finally, as shown in "Accessibility and scaling," a conda environment and automated reference data download are provided. In addition, Mantis accepts several formats as input (i.e., protein FASTA file, TSV file with paths, directories, or compressed archives), outputting easy-to-parse TSV files. We believe that these features address some of the reproducibility challenges that the bioinformatics community still faces [72].

As discussed, there is still room for improvement in the hit-processing algorithm DFS (because it does not provide large F1 score gains over the heuristic algorithm). In the future, Mantis

scale

could also include genomic context–based annotation methods. Despite the aforementioned challenges, we have clearly shown that Mantis is a flexible tool that also produces annotations with high precision and recall.

Conclusion

By making use of the synergistic nature of differently sourced high-quality reference data, Mantis produces reliable homology-based annotations. By allowing for total customization of these reference data, Mantis is also flexible, easily integrated and adapted towards various research goals. In conclusion, we have shown that Mantis addresses a number of the current PFA challenges, resulting in a highly competitive PFA tool.

Methods

Accessibility and scaling

Mantis automatically sets up its reference data by downloading HMMs from different sources and, when necessary, reformatting the data to a standardized format and downloading any relevant metadata. Reference data can be customized via a config file. It also dynamically configures its execution depending on the resources available. A conda environment and extensive documentation [54] are available.

Mantis splits most of the workflow into sub-tasks and subsequently parallelizes them by continuously releasing tasks to workers from a global queue (via Python's multiprocessing module). During each main task of the annotation workflow, workers are recruited (the number of workers depends on the available hardware and work required); these will then execute all the queue tasks. When a worker has finished its job, it will execute another task from the queue until there are no more tasks to execute. If the queue is well balanced, minimal idle time (time spent waiting for workers to get a new task) can be achieved. Load balancing is achieved by splitting the sample and reference data into chunks. During set-up, large reference data sources (>5,000 HMM profiles) are split into smaller chunks; this enables parallelization and ensures that each annotation sub-task takes approximately the same time. Samples are equally split into chunks (sample chunk size is dynamically calculated). If the sample has ≤200,000 sequences, sequences are distributed by their length among the different chunks, so that each chunk has approximately the same number of residues. If the sample has >200,000 sequences, then sequences are distributed to each chunk independently of their length (this alternative method is an efficiency safeguard). This 2-fold splitting achieves quasi-optimal load balancing. With the sample and reference data in chunks, posterior workflow steps can be parallelized wherever applicable. It is noteworthy that Mantis uses HMMER's hmmsearch for homology search, which outputs an e-value scaled to the sample/chunk size. Because Mantis splits the samples into chunks, during hit processing, the e-value is scaled to the original sample size.

Input and output

MANTIS accepts protein sequence FASTA files as input. If the sample has been previously taxonomically classified, the user can add this information when running Mantis. For example, if annotating an *E. coli* sample, the user could add ”−od” followed by the NCBI ID or the organism name:

```
$ python mantis run_mantis -t sample.faa -od 562
```

Mantis outputs, for each sample, 3 TSV files, each corresponding to a different step in Mantis's workflow: (i) a raw output output_annotation.tsv (generated during Fig. 1 step "Intra-HMM hits processing"), with all the hits, their e-value, and coordinates; (ii) integrated_annotation.tsv (generated during Fig. 1 step "Metadata integration"), with the same information as output_annotation.tsv, but also with hits metadata (e.g., KEGG orthology IDs [KO], enzyme commission [EC] numbers, free-text functional description); and (iii) the main output file consensus_annotation.tsv (generated during Fig. 1 step "Consenus generation"), with each query protein ID and their respective consensus annotation from the different reference data sources (e.g., Pfam). These files provide contextualized output in a format that is both human and machine-readable. A Mantis.out file is also provided per sample, serving as a log file for each execution step.

Reference data and customization

Mantis, by default, uses multiple high-quality reference HMM sources: Pfam [56], eggNOG [57], NPFM [58], KOfam [55], and TIGRfam [59] (these default HMMs can be partially or entirely removed). To find more meaningful homologs through TSA, Mantis uses TSHMMs, originally compiled by eggNOG and NPFM. The eggNOG TSHMMs were compiled by downloading all the TSHMMs at http://eggnog5.embl.de/download/latest/per_tax_level/; their respective metadata originate from the metadata available in the aforementioned link, as well as the metadata within the eggNOG-mapper SQL database. NPFM TSHMMs were compiled by downloading all the NPFM HMMs at https://ftp.ncbi.nlm.nih.gov/hmm/current/ and assigning each HMM into their respective TSHMM. A general NPFM HMM was created by pooling all non-assigned HMM profiles and the TSHMMS from the following NCBI IDs: 2157 (*Archaea*), 2 (*Bacteria*), 2759 (*Eukaryota*), 10239 (*Viruses*), 28384 (Others), and 12908 (Unclassified). These IDs correspond to NCBI's top-level taxonomy rank IDs. A general eggNOG HMM was created by pooling together the TSHMMs from the same aforementioned NCBI taxon IDs. The user can customize which eggNOG TSHMMs are downloaded by Mantis by adding the line "nog_tax = NCBI_ID1, NCBI_ID2" to the config file. Custom HMM sources can also be added by the user; metadata integration of these is also possible (an example is available in Mantis's repository). Because some sources are more specific than others, the user can also customize the weight given to each source during consensus generation. HMM profiles often only possess an ID respective to the database from which they were downloaded, which may not directly provide any discernible information. Mantis, when necessary, ensures that the hits from these HMMs are linked to their respective metadata. For future reference, while an HMM is an individual profile, Mantis compiles all related HMM profiles into a single file, making it indexable by HMMER. Thus when a certain HMM source is mentioned, it refers to the collection of related HMM profiles.

Taxon-specific annotation

TSA uses the TSHMMs and unspecific HMM made available by eggNOG and NPFM. TSA, however, works differently from the annotation method of the other reference data. When given taxonomy information (either a taxon name or NCBI ID) the organism's taxonomic lineage is computed (e.g., for *E. coli* the lineage would be "2 - 1224 - 1236 - 91347 - 543 - 561 - 562"). TSA starts by searching for homologs in the most resolved TSHMM (in this case for taxon 562, if it exists). All valid homologs (respecting

scale

the e-value threshold) are extracted for each query sequence, and unannotated sequences are compiled into an intermediate FASTA file. A new homology search round starts with the sequences in the current intermediate FASTA, but now in the TSHMM 1 level above (in this case the TSHMM 561). This cycle repeats until all query sequences have valid homologs or until there are no more TSHMMs to search for. If there are still sequences to annotate, then these homologs are searched for in the general eggNOG and NPFM HMMs. If no taxonomy information is given, the homology search starts with the general NPFM and eggNOG HMMs. Non-taxon-specific HMMs (i.e., Pfam, KOfam, and TIGRfams) are always used, regardless of the sample's taxonomy.

Multiple hits per protein

HMMER outputs a "domtblout" file [24], where each line corresponds to a hit/match between the reference data and the query protein sequence. The e-value threshold within the HMMER command limits the amount of hits to be analysed in the posterior processing steps. Each hit, among other information, contains the coordinates where the query sequences matched with the reference HMM profiles and the respective confidence score (e-value) (Fig. 2A and B). Mantis uses HMMER's independent e-value when using the DFS and heuristic algorithms, whereas it uses the full sequence e-value when using the BPO algorithm (because only the best hit is extracted per protein sequence). For simplicity purposes, both are simply referred to as e-value throughout this article. The annotation of a protein sequence with multiple hits is a nontrivial problem, thus requiring the implementation of a method for the processing of hits. We designed a method that generates and evaluates all possible combinations of hits by applying the DFS algorithm [73]. This algorithm allows the traversal of a tree-structured search space (i.e., each node is a hit), whilst pruning solutions that do not respect predefined constraints (i.e., overlapping hit residue coordinates), backtracking from leaf to root until the possible solution space is exhausted. Our method generates all the possible combination hits with the following method: (i) get 1 hit from the collection of hits and define it as the combination root hit; (ii) check which other hits overlap up to 10% (default value) [31] with previous hits and select 1 to add to our present combination of hits; (iii) repeat step (ii) until no more hits can be added; (iv) repeat steps (i–iii) so that we loop over all the other hits and all possible combinations are generated. We used Cython [74] to speed up the DFS implementation. Cython is an optimizing static compiler for the Python programming language, allowing the compiler to generate C code from Cython code, in this case, functioning as a wrapper for the DFS algorithm. The total number of possible combinations is $2^N - X - 1$, where N is the number of hits the protein sequence has, X the number of impossible combinations (combinations with overlapping hits), and 1 the empty combination. Owing to exponential scaling, this method is not always computationally feasible (e.g., the query sequence is very large and has many small-sized hits). In such a scenario, the DFS algorithm may exceed the system's recursion limit or be unable to find a solution in optimal time (60 seconds by default, but customizable). Should this happen, Mantis uses the previously described heuristic algorithm, which scales linearly (a warning is written in the Mantis.out log).

After generating all the possible combinations, each combination is evaluated according to several parameters:

- $\text{query}_{\text{length}}$—number of residues in the query sequence.
- $\text{hit}_{\text{length}}$—number of residues in the hit.
- $\text{combo}_{\text{length}}$—number of hits in the respective combination.
- Total coverage (TC)—number of non-redundant residues in all the combination's hits divided by $\text{query}_{\text{length}}$. A high TC implies that the combination covers a large percentage of the protein sequence.
- Average hit coverage (HC)—sum of the coverage of each hit ($\text{hit}_{\text{length}}/\text{query}_{\text{length}}$). This sum is then averaged by dividing by $\text{combo}_{\text{length}}$. A high HC implies that the hits in the combination are large, thus benefiting combinations with few large hits rather than combinations with many small hits.
- Combination e-value (CE)—the e-value of each hit is scaled twice, once to reduce the range between different e-values ($\log_{10}$) and the second time to understand how each hit e-value compares to the best/lowest hit e-value found for a particular sequence (minmax scaling). The scaled e-values are then summed and divided by $\text{combo}_{\text{length}}$.

The "combination score" is defined by the following equation:

$$TC \times HC \times CE. \tag{1}$$

The combination with the highest combination score is then selected, where the available choices will ultimately depend on the algorithm used (Fig. 2c). Our intra-HMM hit-processing implementation thus applies a 2-fold quality control, initially by limiting the amount of hits in HMMER's domtblout (i.e., e-value threshold) and second by hierarchically ordering and selecting the most significant combination of hits.

Using multiple reference data sources

An unannotated protein sequence may match with 0, 1, or multiple reference HMM profiles, from 1 or more data sources. When a protein sequence has multiple hits from different data sources, it is important to identify functionally similar annotations so that no information is lost (i.e., functional descriptions or IDs that may be in 1 reference data source but not in another). By linking the metadata respective to the HMM profiles to the now annotated protein sequence, we can identify functionally similar annotations and integrate multiple reference data sources into 1 final consensus annotation. In this manner, functionally similar annotations are merged, and any complementary information they provide can then be used in downstream analysis (e.g., Annotation 1 has a Pfam and KO ID, Annotation 2 has an EC number and the same KO ID; merging these will result in a final annotation with more information).

For the integration of functional annotations from multiple data sources, a two -fold approach was used: (i) consensus between IDs and (ii) consensus between the free-text functional description. The latter is used as a backup because ID cross-linking is not universally available. Each reference data source includes metadata relevant to the HMM profiles herein; these metadata may include multiple intra- and/or inter-database IDs, as well as free-text functional descriptions. IDs are extracted either through source-specific metadata parsing or regular expressions. Free-text functional descriptions are extracted by source-specific metadata parsing. With this information it is then possible to identify annotations that are functionally similar/consistent and may thus be complementary to each other. The consensus between IDs is calculated by identifying intersections between the functional annotations of different reference data sources (e.g., both annotations have the same Pfam ID). IDs

scale

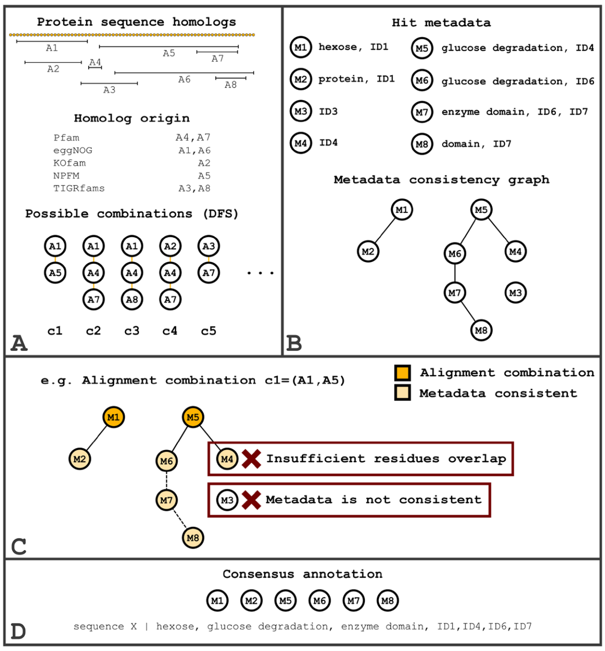

Protein sequence homologs
A1
A5
A7
A2
A4
A6
A8
A3
Homolog origin
Pfam A4,A7
eggNOG A1,A6
KOfam A2
NPFM A5
TIGRfams A3,A8
Possible combinations (DFS)
A1 A5
A1 A4 A7
A1 A4 A8
A2 A4 A7
A3 A7
...
c1 c2 c3 c4 c5
A
Hit metadata
M1 hexose, ID1
M2 protein, ID1
M3 ID3
M4 ID4
M5 glucose degradation, ID4
M6 glucose degradation, ID6
M7 enzyme domain, ID6, ID7
M8 domain, ID7
Metadata consistency graph
M1 M2 M5 M6 M4 M7 M3 M8
B
e.g. Alignment combination c1=(A1,A5)
Alignment combination
Metadata consistent
M1 M2 M5 M6 M7 M8
M4 Insufficient residues overlap
M3 Metadata is not consistent
C
Consensus annotation
M1 M2 M5 M6 M7 M8
sequence X | hexose, glucose degradation, enzyme domain, ID1,ID4,ID6,ID7
D

Figure 7: Inter-HMM hit-processing steps. Inter-HMM hit processing starts by pooling all hits [A1, AN] together (regardless of the reference data source) and generating all the possible (non-overlapping coordinates) combinations [c1, cN] (A). A metadata consistency graph (B) is also built by connecting all nodes [M1, MN] that have intersecting IDs or highly similar descriptions (e.g., A1's metadata M1 is consistent with A2's metadata M2 (shared ID1), and A5's metadata M5 is consistent with A6's metadata M6 (similar description ”glucose degradation”). With this metadata consistency graph, the hit consistency HCN score of each combination is calculated. For c1, for example, a sub-graph containing M1, M5, and all directly connected nodes (only M2 and M6 but not M4 because it has insufficient residue overlap—A4) would be created. The number of nodes in this sub-graph would then be divided by the total number of nodes in the original graph; therefore c1 would have an HCN of (2 + 2)/8 = 0.5. The remaining parameters would then be calculated and the best combination, according to equation 2, would be selected. Finally, if, for example, the best combination is c1, then this combination is expanded by merging all nodes directly or indirectly connected to M1 and M5 in the metadata consistency graph (C) and with sufficient residue overlap (i.e., M2, M6, M7, M8). The expanded combination is then merged into the final consensus annotation (D).

within the free-text functional descriptions are extracted (with regular expressions) and also used here. If no consensus between IDs is found, then we proceed with a consensus calculation between functional descriptions (further described in the Supplementary PDF).

Inter-HMM hit processing starts by pooling together all hits from the different reference data sources and generating all possible combinations of hits (Fig. 7A). The same method used in intra-HMM hit processing is applied, where the DFS algorithm is used by default (again using the heuristic algorithm as a backup), but the BPO and heuristic algorithms can also be used. We then check the metadata consistency (either through IDs or free-text functional descriptions) of each hit against the current sequence's other hits. With this information, a metadata consistency graph is generated (Fig. 7B). With the metadata consistency graph and all possible combinations of hits, we can then calculate the consensus combination score using equation 2. This requires calculation of the combination score, using equation 1. This score is then multiplied by an additional score, comprising the following parameters:

- Average hit consistency (HCN)—number of hits (among all hits) with metadata directly consistent (i.e., nodes directly connected in the metadata consistency graph) to the hits in the present combination. Consistency checks are restricted to other reference data sources besides the hit own's reference source (e.g., if a hit is from Pfam, we would only check hits that are not from Pfam). This number, plus the number of hits in the combination, is divided by the total number of hits for the respective query sequence [e.g., if a combination has 2 hits, with these having metadata consistent with 3 other hits, and if there are 10 hits in total, HCN would be equal to (2 + 3)/10 = 0.5]. This is an important parameter because it entails independent sources describing the same function.
- Reference HMM weight (HMMW)—mean weight of all the reference data sources within the combination. This is calculated by adding all hits' HMM weights and dividing this sum by the number of hits in the combination [e.g., if a hit comes from Pfam that has a weight of 1, and another from eggNOG that has a weight of 0.8, HMMW would be equal to (1 + 0.8)/2 = 0.9]). The default weight for each default reference data source has been set according to the authors' perception of the reference quality—creation method, curation level, and annotation completeness (eggNOG, 0.8; Pfam, 0.9; NPFM and KOfam, 0.7; and TIGRfam, 0.5). This weight is customizable; the default weight for custom reference data is 0.7 (which can also be customized).
- Metadata quality (MQ)—mean metadata quality of each hit in the combination. If a hit has no annotation data (IDs or description), it is given a score of 0.25; 0.5 if only the description; 0.75 if only the IDs; 1 if IDs and description. All hits' metadata quality score is summed and divided by the number of hits in the combination.

Note that hit metadata consistency (through IDs or descriptions) requires a minimum of 70% residue overlap (default but can be changed). Using the previously calculated combination score, we then calculate the consensus combination score using the following equation:

$$\text{Combination}_{\text{score}} \times \frac{\text{HCN} + \text{HMMW} + \text{MQ}}{3}. \qquad (2)$$

The combination with the highest consensus combination score is selected and expanded by concatenating additional metadata from other consistent hits (Fig. 7C). In this step, consistent hits can be either directly or indirectly connected in the metadata consistency graph (a minimum of 70% residue overlap is still required). This expanded combination is then merged into the final query sequence consensus annotation (Fig. 7D). Redundant (i.e., repeated identifiers or functional descriptions) or poor-quality information (e.g., “hypothetical protein”) is removed from the consensus annotation.

Sample selection

To select an initial testing dataset we started by downloading all the curated Uni-Prot [75] (i.e., Swiss-Prot) protein entries created after 2010 (until 14 April 2020), along with their respective sequences, annotations, and annotation scores. We then split these entries by date, 2010–2020, 2015–2020, 2018–2020, and 2020 only. For genomic sample benchmarking we selected organisms widely used in microbial community standards. The respective genomes, proteomes, and reference annotations were then downloaded from Uniprot on 26 May 2020 (Supplementary

scale

Table 5). These samples were also used for comparing Mantis to eggNOG-mapper and Prokka.

Establishing a test environment

For annotation quality benchmarking, we evaluate each annotation produced by Mantis and check whether it agrees (database IDs intersection) with the respective reference annotation, creating a confusion matrix. We created 2 main types of test samples, the first consisting exclusively of curated UniProt [75] protein entries (and the respective annotations), which were then split by date of creation (2010–2020, 2015–2020, 2018–2020, 2020); and the second type consisting of organism-specific UniProt protein entries, with a mix of curated and automatically generated annotations. Each sequence's reference annotation consists of the UniProt protein function annotations. Each sequence reference annotation and the respective PFA tool's annotation is composed of a set of identifiers (if available: enzyme ECs, Gene Ontology (GO) IDs, eggNOG IDs, KEGG orthology IDs, Pfam IDs, and TIGRfam IDs) and functional descriptions. During the benchmark process, each sequence's reference annotation (e.g., "glucose degradation ID1") is compared against the PFA tool (i.e., Mantis, eggNOG-mapper, and Prokka) annotation (e.g., "degrades glucose ID1"). This comparison entails checking whether any of the database IDs present in the reference annotation (i.e., ID1) are also present in the PFA tool annotation (i.e., ID1); if they are, we consider this annotation to be the same. This has some significant limitations: (i) the functional description is the same but the corresponding set of identifiers is not; and (ii) when annotating multiple regions of the protein (which is the case when using Mantis's DFS and heuristic algorithms), it is possible that only 1 of the annotated regions has IDs that intersect with the respective sequence reference annotation. Unfortunately, owing to the different resolutions of the reference HMMs, it is not always possible to understand whether an annotation refers to a specific domain or a partial whole-sequence hit. While a domain-centric benchmark would be feasible for Pfam, the same is not true for the remaining reference HMMs with broader resolutions (e.g., TIGRFams provides general functional annotations). However, as we have previously shown, even when using the BPO algorithm, Mantis has shown to output almost equally high F1 scores. Despite these limitations, because whole-sequence reference annotations contain comprehensive cross-linking with other databases, it provides clear benefits: (i) it fits better for the wide-ranging scopes of the reference data sources, and (ii) it allows for a more fair benchmark of the different PFA tools that may use different reference data sources (and thus output annotations with different database IDs). This method then allows for the construction of a confusion matrix, where each pairwise whole-sequence annotation comparison (PFA tool/reference annotation) corresponds to a single class. TPs occur when the PFA tool–generated annotation and the reference annotation share 1 or more database IDs (e.g., Pfam ID), and FPs when no database IDs are shared. FNs occur when the PFA tool does not annotate a protein sequence, although a reference annotation is available; and TNs when the PFA tool does not annotate a protein sequence, and no reference annotation is available. The functional text descriptions are not taken into account during the benchmark; therefore if an annotation has no IDs, we simply consider there to be no annotation. Protein sequences annotated with the descriptions "unknown function," "uncharacterized protein," "hypothetical protein," or with Pfam's "domain-unknown-function"/DUF IDs are not taken into account during benchmarking (for reference and PFA tool annotations). In addition, it is also possible for the reference or PFA tool not to have an annotation for a certain sequence. In any of the these 3 scenarios, if the PFA tool manages to annotate the sequence, this case is classified as a potentially new annotation (PNA). Because no ground-truth exists in these scenarios, PNAs are excluded from the confusion matrix classes (not used during any performance metrics) and are only used to calculate the annotation coverage. PNAs can potentially provide novel insight into protein sequences without any previous annotation. Because, by default, most sequences used during benchmarking will have an annotation, TNs, and ergo any metrics using TNs (e.g., specificity), are irrelevant.

"Annotation coverage" is defined here as the number of annotations produced by the PFA tool divided by the total number of protein sequences in a sample $\mathrm{Total_{seqs}}$. $\mathrm{Total_{seqs}}$ includes sequences with and without a reference annotation (because not all sequences have a reference annotation); the total number of the PFA tool annotations includes TPs, FPs, and PNAs. Annotation coverage is calculated by $(\mathrm{TP} + \mathrm{FP} + \mathrm{PNA})/\mathrm{Total_{seqs}}$. Numerous metrics can be calculated from the various confusion matrix categories; we considered precision and recall/sensitivity to be among the most important. Precision is defined as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and corresponds to the number of correctly annotated protein sequences out of all the protein sequences that the PFA tool managed to annotate. Recall is defined as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and corresponds to the number of correctly annotated protein sequences out of all the protein sequences that we know the function of (i.e., protein sequences that have a reference annotation). Both are equally important; a tool with low precision will incorrectly annotate protein sequences, whereas a tool with low recall will not produce sufficient annotations. A way to converge both scores into 1 is to use the F1 score, which is defined as $2 \times [(\mathrm{Precision} \times \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})]$. Unless otherwise stated, values shown in this article are shown as absolute values ranging from 0 to 1.

Finally, we benchmarked Mantis against 2 other PFA tools, eggNOG-mapper and Prokka. For homology search, Mantis uses HMMER [24], for eggNOG-mapper we used the Diamond-based [23] search (as suggested by the authors), and Prokka uses BLAST and HMMER.

All tests ran on high-performance computing resources with Dell C6320, 2 $*$ Intel Xeon E5-2680 v4 at 2.4 GHz [76]; each core had 4 GB of RAM. Unless specified, all tests ran with 25 cores and 100 GB RAM (actual Mantis minimum hardware requirements are much lower). In addition, the same methodology and nomenclature apply to any other benchmarked tools described in this article. Mantis used HMMER v3.2.1. The local version of eggNOG-mapper used was v2.0.6 with database v5.0.1 found at https://github.com/eggnogdb/eggnog-mapper/commit/41ec3566ab00fd437f905dfde592c553632a9eae. The local version of Prokka used was v1.14.6 found at https://github.com/tseemann/prokka/releases/tag/v1.14.6. For details on execution commands see the Supplementary PDF.

Testing different e-value thresholds

Different e-value thresholds were tested: $1e^{-3}$, $1e^{-6}$, $1e^{-9}$, $1e^{-12}$, $1e^{-15}$, $1e^{-18}$, $1e^{-21}$, $1e^{-24}$, $1e^{-27}$, $1e^{-30}$, and a dynamic threshold. The dynamic threshold was set according to the query sequence length, which was previously shown to provide better results with BLAST [34]. For the dynamic threshold, for sequences with $<$150 amino acids, the e-value threshold was set to $1e^{-10}$; if $>$150 and $<$250, $1e^{-\mathrm{sequence_{length}}/10}$; and if $>$250, $1e^{-25}$. The UniProt 2010–2020 sample was then annotated with all the different e-value

scale

thresholds, and each output was compared to the reference annotations.

Testing hit-processing algorithms

To understand whether the different hit-processing algorithms resulted in statistically significant differences in F1 scores, we created 5,000 randomized synthetic samples with 5,000 sequences each, which were randomly selected from the 2010–2020 UniProt sample. Per algorithm, we compared the Mantis annotations of each subset to the reference annotations (to allow for pairwise comparison of each algorithm, the same subsets were used in all algorithms). This resulted in a list of confusion matrices (5,000 per algorithm), from which we calculated the F1 score. We applied the Wilcoxon signed-rank test, with the H_0: no differences in F1 score between the tested algorithms. As a non-parametric test, this test makes no assumptions on the distribution of the data. A pairwise comparison was done between DFS and the other algorithms: (i) DFS and heuristic and (ii) DFS and BPO.

Availability of Source Code and Requirements

- Project name: Mantis
- Project home page: https://github.com/PedroMTQ/mantis
- Operating system: Linux
- Programming language: Python
- Other requirements: Python 3+, HMMER 3+, and several Python packages (see the provided environment for a full list)
- License: MIT
- RRID:SCR_021001
- Biotools ID: mantis_pfa

Data Availability

The data and code supporting the results of this article are available at https://git-r3lab.uni.lu/pedro.queiros/mantis_supplements. An archival copy of the code and supporting data is available via the *GigaScience* repository, GigaDB [77].

Additional Files

Supplementary pdf. (i) discussion on how the e-value threshold may change Mantis' output, (ii) execution commands, and (iii) information on how the similarity analysis was performed.
Supplementary Table 1. Function assignment e-value threshold
Supplementary Table 2. Impact of hit processing algorithms
Supplementary Table 3. Impact of sample selection
Supplementary Table 4. Impact of consensus generation
Supplementary Table 5. Quality control against sequenced organisms – list of samples
Supplementary Table 6. Quality control against sequenced organisms – results
Supplementary Table 7. Comparison between Mantis and other PFA tools
Supplementary Table 8. Annotation efficiency – random sequences
Supplementary Table 9. Annotation efficiency – personal PC
Supplementary Table 10. Metagenome coverage

Abbreviations

BLAST: Basic Local Alignment Search Tool; BPO: best prediction only; CE: combination e-value; CPU: central processing unit; DFS: depth first search; EC: enzyme commission; FP: false positive; FN: false negative; GFS: glacier-fed stream sediment; GO: gene ontology; HC: average hit coverage; HCN: hit consistency; HMM: hidden Markov models; HMMW: average reference HMM weight; ID: database identifier; KEGG: Kyoto Encyclopedia of Genes and Genomes; KO: KEGG orthology; MQ: metadata quality; NCBI: National Center for Biotechnology Information; NLP: natural language processing; NPFM: NCBI protein family models; PFA: protein function annotation; PNA: potentially new annotation; RAM: random access memory; TC: total coverage; TN: true negative; TP: true positive; TSA: taxon-specific annotation; TSHMM: taxon-specific HMM; TSV: tab-separated value.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Luxembourg National Research Fund PRIDE17/11823097. P.W. acknowledges the European Research Council (ERC-CoG 863664).

Authors' Contributions

Author contributions according to the contributor roles taxonomy (CRediT) were as follows: Conceptualization: P.Q. and P.M.; Data curation: P.Q.; Formal analysis: P.Q.; Funding acquisition: P.W. and P.M.; Investigation: P.Q.; Methodology: P.Q. and P.M.; Project administration: P.Q. and P.M.; Resources: P.Q.; Software: P.Q.; Supervision: P.M. and P.W.; Validation: P.Q. (lead), F.D., and O.H.; Visualization: P.Q.; Writing—original draft: P.Q. (lead), and P.M.; Writing—review and editing: P.Q., P.M., F.D., O.H., and P.W. All authors proofread and approved of the content in the manuscript.

Acknowledgments

The experiments presented in this article were carried out using the high-performance computing facilities of the University of Luxembourg [76]. P.W. acknowledges the European Research Council (ERC-CoG 863664). We thank Tomila Litvishko for proofreading the manuscript. We would like to acknowledge all the creators of the reference data and software used by Mantis—building upon the complementary knowledge of others truly moves the field forward.

References

1. Segata N, Boernigen D, Tickle TL, et al. Computational meta'omics for microbial community studies. Mol Syst Biol 2013;**9**:666.
2. Muller E, Glaab E, May P, et al. Condensing the omics fog of microbial communities. Trends Microbiol 2013;**21**(7): 325–33.
3. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys 2003;**36**(3):307–40.
4. Arias C, Weisburd B, Stern-Ginossar N, et al. KSHV 2.0: A comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. PLoS Pathog 2014;**10**(1):e1003847.

scale

5. Chapel A, Kieffer-Jaquinod S, Sagné C, et al. An extended proteome map of the lysosomal membrane reveals novel potential transporters. Mol Cell Proteomics 2013;**12**(6):1572–88.
6. Iorizzo M, Senalik DA, Grzebelus D, et al. De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics 2011;**12**(1):389.
7. Heintz-Buschart A, May P, Laczny CC, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol 2016;**2**:16180.
8. Mason OU, Scott NM, Gonzalez A, et al. Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. ISME J 2014;**8**(7):1464–75.
9. Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 2019;**176**(3):649–662.
10. Sureyya Rifaioglu A, Doğan T, Jesus Martin M, et al. DEEPred: Automated protein function prediction with multi-task feed-forward deep neural networks. Sci Rep 2019;**9**(1):7344.
11. Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;**21**(6):697–700.
12. Borgwardt KM, Ong CS, Schönauer S, et al. Protein function prediction via graph kernels. Bioinformatics 2005;**21**:i47–56.
13. Steinegger M, Meier M, Mirdita M, et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 2019;**20**(1):473.
14. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;**30**(14):2068–9.
15. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol 2017;**34**(8):2115–22.
16. Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics 2008;**9**(1):75.
17. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A 2019;**116**(28): 13996–4001.
18. Zhao B, Hu S, Li X, et al. An efficient method for protein function annotation based on multilayer protein networks. Hum Genomics 2016;**10**(1):33.
19. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;**47**:D607–13.
20. Deng L, Zhong G, Liu C, et al. MADOKA: An ultra-fast approach for large-scale protein structure similarity searching. BMC Bioinformatics 2019;**20**(19):662.
21. Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;**33**(17):5691–702.
22. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. J Mol Biol 1990;**215**(3):403–10.
23. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;**12**(1):59–60.
24. Eddy SR. HMMER; 2020, HMMER: biosequence analysis using profile hidden Markov models.Accessed 22th March 2021, http://hmmer.org.
25. Jones P, Binns D, Chang HY, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics 2014;**30**(9):1236–40.
26. Lohse M, Nagel A, Herter T, et al. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ 2014;**37**(5):1250–8.
27. Wu S, Zhu Z, Fu L, et al. WebMGA: A customizable web server for fast metagenomic sequence analysis. BMC Genomics 2011;**12**:444.
28. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: The microbiome analysis resource in 2020. Nucleic Acids Res 2020;**48**:D570–8.
29. Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol 2016;**1399**:207–33.
30. Pfeiffer F, Oesterhelt D. A manual curation strategy to improve genome annotation: Application to a set of haloarchael genomes. Life 2015;**5**(2):1427–44.
31. Yeats C, Redfern OC, Orengo C. A fast and automated solution for accurately resolving protein domain architectures. Bioinformatics 2010;**26**(6):745–51.
32. Ekman D, Bjorklund AK, Frey-Skott J, et al. Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. J Mol Biol 2005;**348**(1):231–43.
33. Lees JG, Lee D, Studer RA, et al. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. Nucleic Acids Res 2014;**42**:D240–5.
34. Treiber ML, Taft DH, Korf I, et al. Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. BMC Bioinformatics 2020;**21**(1):74.
35. Schnoes AM, Brown SD, Dodevski I, et al. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;**5**(12):e1000605.
36. Friedberg I. Automated protein function prediction—The genomic challenge. Brief Bioinform 2006;**7**(3):225–42.
37. Araujo FA, Barh D, Silva A, et al. GO FEAT: A rapid web-based functional annotation tool for genomic and transcriptomic data. Sci Rep 2018;**8**(1):1794.
38. Klimke W, O'Donovan C, White O, et al. Solving the problem: Genome annotation standards before the data deluge. Stand Genomic Sci 2011;**5**(1):168–93.
39. Standardizing data. Nat Cell Biol 2008;**10**(10):1123–4.
40. Gaikwad SV, Chaugule A, Patil P. Text mining methods and techniques. Intl J Comput Appl 2014;**85**(17):422–5.
41. Wang S, Ma J, Yu MK, et al. Annotating gene sets by mining large literature collections with protein networks. Pac Symp Biocomput 2018;**23**:602–13.
42. Pesquita C, Faria D, Falcão AO, et al. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009;**5**(7): e1000443.
43. Zeng Z, Shi H, Wu Y, et al. Survey of natural language processing techniques in bioinformatics. Comput Math Methods Med 2015;**2015**:674296.
44. Slater LT, Bradlow W, Ball S, et al. Improved characterisation of clinical text through ontology-based vocabulary expansion. J Biomed Semantics 2020;**12**, doi:10.1186/s13326-021-00241-5.
45. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: Success, failure and the future. Brief Bioinform 2016;**17**(1):132–44.
46. Benabderrahmane S, Smail-Tabbone M, Poch O, et al. IntelliGO: A new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics 2010;**11**:588.
47. Peng J, Uygun S, Kim T, et al. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics 2015;**16**:44.

scale

48. Liu M, Thomas PD. GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness. BMC Bioinformatics 2019;**20**(1):155.
49. Daraselia N, Yuryev A, Egorov S, et al. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinformatics 2007;**8**:243.
50. Ehsani R, Drabløs F. TopoICSim: A new semantic similarity measure based on gene ontology. BMC Bioinformatics 2016;**17**(1):296.
51. Kramer M, Dutkowski J, Yu M, et al. Inferring gene ontologies from pairwise similarity data. Bioinformatics 2014;**30**(12):i34–42.
52. Promponas VJ, Iliopoulos I, Ouzounis CA. Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure. Stand Genomic Sci 2015;**10**:108.
53. Ellens KW, Christian N, Singh C, et al. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res 2017;**45**(20):11495–514.
54. Queirós P. Mantis - Wiki. 2020. Acessed 22th March 2021, available at https://github.com/PedroMTQ/mantis/wiki.
55. Aramaki T, Blanc-Mathieu R, Endo H, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 2020;**36**(7):2251–2.
56. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. Nucleic Acids Res 2019;**47**:D427–32.
57. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;**47**:D309–14.
58. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res 2020;**48**(D1):D265–8.
59. Haft DH, Selengut JD, Richter RA, et al. TIGRFAMs and genome properties in 2013. Nucleic Acids Res 2013;**41**:D387–95.
60. Hyatt D, Chen GL, LoCascio PF, et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;**11**:119.
61. Delogu F, Kunath BJ, Evans PN, et al., Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. Nat Commun 2020;**11**:4708.
62. Kunath BJ, Delogu F, Naas AE, et al. From proteins to polysaccharides: Lifestyle and genetic evolution of *Coprothermobacter proteolyticus*. ISME J 2019;**13**(3):603–17.
63. Busi SB, Pramateftaki P, Brandani J, et al. Optimised biomolecular extraction for metagenomic analysis of microbial biofilms from high-mountain streams. PeerJ 2020;**8**:e9973.
64. Sunagawa S, Coelho LP, Chaffron S, et al. Structure and function of the global ocean microbiome. Science 2015;**348**(6237):1261359.
65. Sigrist CJA, de Castro E, Cerutti L, et al. New and continuing developments at PROSITE. Nucleic Acids Res 2013;**41**: D344–7.
66. Mooney MA, Nigg JT, McWeeney SK, et al. Functional and genomic context in pathway analysis of GWAS data. Trends Genet 2014;**30**(9):390–400.
67. Mavromatis K, Chu K, Ivanova N, et al. Gene context analysis in the integrated microbial genomes (IMG) Data Management System. PLoS One 2009;**4**(11):e7979.
68. Hannigan GD, Prihoda D, Palicka A, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res 2019;**47**(18):e110.
69. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018;**36**(10):996–1004.
70. Parks DH, Chuvochina M, Chaumeil PA, et al. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 2020;**38**(9):1079–86.
71. Buell R, Deutschbauer A, Adin D, et al. Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa; Workshop Report. United States. 2019. doi:10.2172/1616527
72. Mangul S, Mosqueiro T, Abdill RJ, et al. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol 2019;**17**(6), doi:10.1371/journal.pbio.3000333.
73. Kaur N, Garg D. Analysis of the depth first search algorithms. Data Mining Knowl Eng 2012;**4**:37–41.
74. Behnel S, Bradshaw R, Citro C, et al. Cython: The best of both worlds. Comput Sci Eng 2011;**13**(2):31–9.
75. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 2019;**47**:D506–15.
76. Varrette S, Bouvry P, Cartiaux H, et al. Management of an academic HPC cluster: The UL experience. IEEE HPCS; 2014; available at https://hpc.uni.lu.
77. Queirós P, Delogu F, Hickl O, et al. Supporting data for "Mantis: Flexible and consensus-driven genome annotation." GigaScience Database; 2021. http://dx.doi.org/10.5524/100903.

5.3

Alterations of oral microbiota and impact on the gut microbiome in type 1 diabetes mellitus revealed by integrated multi-omic analyses

5.3.1 Coversheet

Contributions of Oskar Hickl

- Shared primary authorship responsibilities with B.J.K., performing a lead role in the study by conducting data curation, analysis, interpretation, and visualization of results. Additionally, principally co-wrote and refined the manuscript.
- Responsible for the creation of visual elements of the manuscript, including figures 1 and 6, table 1, supplementary table 1, and supplementary figure 1.
- Collaborated with D.B. in the procurement and provision of the metaproteomic data. Also built and optimized the metaproteomics workflow.

5.3.2 Introduction

In this study, a multi-meta-omic approach was employed, integrating matched metagenomics, metatranscriptomics, and metaproteomics with clinical data. The aim was to identify differences in oral and gut microbiome composition, expression, and activity in the context of Type 1 Diabetes Mellitus (T1DM). Data from 35 subjects across eight families with multiple T1DM cases were analyzed to investigate the dynamics of the oral microbiota, its impact on the gut microbiome, and potential subsequent host inflammatory responses. Additionally, the study assessed mouth-to-gut microbial transmission in relation to T1DM, with the identification of involved strains and the comparison of transmission patterns between T1DM patients and healthy controls.

The integrative approach provided a comprehensive understanding of bacterial strain transfers and their functional activity between the oral cavity and the gut. Observations indicated transmission of genera like *Prevotella* and *Bacteroides* from the oral cavity to the gut, where evidence of their activity was found.

In the T1DM cohort, the oral microbiota exhibited characteristics hinting at an acidified environment, including a decrease in the abundance and activity of the usually acid-intolerant *S. salivarius* and an increase in the acid-tolerant opportunistic pathogen *S. mutans*. Corresponding observations in the gut revealed a decreased presence of *S. salivarius* and an increased abundance of *E. coli*. Additionally, genes associated with bacterial virulence and oxidative stress response, mainly linked to the *Enterobacteriaceae* family, showed increased activity. Additionally, elevated expression levels of specific human proteins, suggesting heightened immune responses and potential inflammation in the gut of T1DM patients, was observed.

The study employed a novel approach, enabled by multi-omics, to trace bacterial strain-variants across the multiple omic layers, elucidating bacterial colonization in the gut, particularly of those originating from the oral cavity. The findings highlight the complementary nature of metaproteomics to metatranscriptomics, in terms of validating identified bacterial strains and providing insights into bacterial activity.

While there is a significant correlation between bacterial abundance in the gut and transmission levels, such a trend is not as apparent in the oral cavity. This suggests that the efficiency of bacterial transmission may not solely rely on their initial abundance in the oral cavity. Physiological changes induced by T1DM in the oral environment may exert a significant influence on microbial transmission dynamics. As a follow-up to this study, a more detailed examination of these physiological parameters, in conjunction with the innovative strain-variant methodology, may provide deeper insights into microbiome transmission dynamics in T1DM patients.

This kind of integrated analysis underscores the importance of considering the oral-gut axis when trying to understand T1DM pathology. Further research is warranted to unravel the intricacies of this relationship and explore its potential implications e.g. for the development of therapeutic interventions.

5.3.3 Manuscript

Kunath *et al. Microbiome* (2022) 10:243
https://doi.org/10.1186/s40168-022-01435-4

Microbiome

RESEARCH

Open Access

Alterations of oral microbiota and impact on the gut microbiome in type 1 diabetes mellitus revealed by integrated multi-omic analyses

B. J. Kunath[1*†], O. Hickl[1†], P. Queirós[1], C. Martin-Gallausiaux[1], L. A. Lebrun[1], R. Halder[1], C. C. Laczny[1], T. S. B. Schmidt[2], M. R. Hayward[3], D. Becher[4], A. Heintz-Buschart[5], C. de Beaufort[1,6], P. Bork[2,7,8,9], P. May[1] and P. Wilmes[1,10*]

Abstract

Background: Alterations to the gut microbiome have been linked to multiple chronic diseases. However, the drivers of such changes remain largely unknown. The oral cavity acts as a major route of exposure to exogenous factors including pathogens, and processes therein may affect the communities in the subsequent compartments of the gastrointestinal tract. Here, we perform strain-resolved, integrated meta-genomic, transcriptomic, and proteomic analyses of paired saliva and stool samples collected from 35 individuals from eight families with multiple cases of type 1 diabetes mellitus (T1DM).

Results: We identified distinct oral microbiota mostly reflecting competition between streptococcal species. More specifically, we found a decreased abundance of the commensal *Streptococcus salivarius* in the oral cavity of T1DM individuals, which is linked to its apparent competition with the pathobiont *Streptococcus mutans*. The decrease in *S. salivarius* in the oral cavity was also associated with its decrease in the gut as well as higher abundances in facultative anaerobes including *Enterobacteria*. In addition, we found evidence of gut inflammation in T1DM as reflected in the expression profiles of the *Enterobacteria* as well as in the human gut proteome. Finally, we were able to follow transmitted strain-variants from the oral cavity to the gut at the individual omic levels, highlighting not only the transfer, but also the activity of the transmitted taxa along the gastrointestinal tract.

Conclusions: Alterations of the oral microbiome in the context of T1DM impact the microbial communities in the lower gut, in particular through the reduction of "mouth-to-gut" transfer of *Streptococcus salivarius*. Our results indicate that the observed oral-cavity-driven gut microbiome changes may contribute towards the inflammatory processes involved in T1DM. Through the integration of multi-omic analyses, we resolve strain-variant "mouth-to-gut" transfer in a disease context.

Introduction

Thousands of distinct microbial taxa colonise the different mucosal and skin habitats of the human body [1]. These communities and their functional gene complements directly interface with host physiology, most notably the immune system [2, 3]. Altered community compositions are thought to play crucial roles in

[†]B. J. Kunath and O. Hickl contributed equally to this work.

*Correspondence: benoit.kunath@uni.lu; paul.wilmes@uni.lu

[1] Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg
[10] Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Belvaux, Luxembourg
Full list of author information is available at the end of the article

triggering inflammatory processes which are most likely drivers of chronic diseases [1, 4, 5], including autoimmune diseases [6–8]. The human microbiome is influenced by biotic and abiotic factors specific to each body site, which leads to distinct microbial community compositions [9]. Although closely related taxa can be present at multiple sites, most species exhibit differentiation into locally adapted strains [10].

Bacterial species usually consist of an ensemble of strains which form coherent clades [11]. Thereby they are clearly distinguishable from the closest co-occurring related species based on their high genetic similarity [12, 13]. The classical metagenomic approach consists of assembling short DNA reads into contigs and to group them into different metagenome-assembled genomes (MAGs). However, the assembly produces a patchwork of consensus contigs corresponding to the most abundant genotypes in the sample and thus can lose strain variations. Multiple approaches exist to retrieve variant information which typically involves the mapping of the metagenomic reads against the assembled contigs or reference genomes. Variant calling is then performed to determine the alleles or haplotypes [14]. Despite the genetic similarity between strains of a single species, the individual strains can exhibit different phenotypes. Such cases are notably well documented in the context of pathogenicity where many species are known to have both pathogenic and commensal strains [11]. Therefore, strain-level resolution is highly relevant in the study of the human microbiome and its links to health and disease.

The gut microbiome has been extensively studied primarily in the context of chronic diseases including cardiovascular diseases [15], inflammatory bowel disease [16], obesity [17], cancers [18], neurodegenerative diseases [19] or autoimmune conditions such as rheumatoid arthritis [20] or type 1 [21], and type 2 diabetes [22]. Type 1 diabetes mellitus (T1DM) is a chronic disease characterised by insulin deficiency due to autoimmune destruction of insulin-producing β-cells within the pancreatic islets. T1DM often starts during the early years of life and is one of the most common chronic diseases in childhood [23]. Its incidence worldwide has reached 15 per 100,000 people and has been globally increasing in the last decades in most developed countries [24–26]. Despite a significant genetic influence, the rise in T1DM prevalence in individuals who are not genetically predisposed strongly suggests an interplay between genetic predisposition and environmental factors [27].

Among the possible different environmental factors, the gut microbiome modulates the function of the immune system via direct and indirect interactions with innate and adaptive immune cells [3, 28]. Several studies have shown alterations of the gut microbiome composition between individuals with T1DM compared to healthy controls [29–32]. However, contrasting findings between studies have not led to a generalisable microbiome signature for T1DM and it still remains unclear how microbiome changes affect the gastrointestinal tract and immune functions in T1DM.

The oral cavity and the colon sit at opposite sides of the gastrointestinal tract. The mouth is considered a gateway to different organs of the body, and therefore acts as a potential reservoir for different pathogens [33]. Poor dental health and dysfunctional periodontal immune-inflammatory reactions caused by bacterial pathogens may lead to periodontitis and are associated with increased risks of developing systemic inflammatory disorders [34]. The development of inflammation in the oral cavity has notably been found to be associated with systemic inflammation and cardiovascular disease [35], insulin resistance [36], and complications in type 1 and type 2 diabetes [37]. Despite the limited number of shared taxa between the oral cavity and the lower gut [38] due to the gastric bactericidal barrier, intestinal motility or bile and pancreatic secretions [39], a recent study has shown that the oral community type was predictive of the community recovered from stool [40]. Additionally, Schmidt, Hayward et al. recently found that a subset of 74 species were frequently transmitted from mouth to gut and formed coherent strain populations along the gastrointestinal tract [41]. Finally, it is known that the physiology of the oral cavity is altered in T1DM patients, notably with a decrease of salivary flow rate (dry-mouth symptom) and an increased concentration of glucose in the saliva and subsequent acidification of the oral cavity [42–44]. However, the effect of T1DM on the microbiome of the oral cavity, or the effect of the microbiome on T1DM in general is still poorly understood, with few, and regularly contradicting findings [45].

Here, we apply an integrated multi-omic approach, including matched meta- genomics, transcriptomics and proteomics together with available clinical data to characterise differences in the oral and gut microbiomes in the context of T1DM on 35 individuals from eight families with multiple case of T1DM per family. We identify distinct oral microbiota suggestive of competition between streptococcal species and an acidified oral cavity. We link these differences to alterations in the gut microbiome and the host's inflammatory response. Finally, we explore the level of mouth-to-gut transmissions in T1DM, highlight transferred and active strains, and identify differences in strain-level

scale

transmission profiles in T1DM patients compared to healthy controls.

Methods

Ethics

Written informed consent was obtained from all subjects enrolled in the study. This study was approved by the Comité d'Ethique de Recherche (CNER; reference no. 201110/05) and the National Commission for Data Protection in Luxembourg.

Sample acquisition

The study design was an observational study of eight selected families (M01-M06, M08, M11) containing at least two members with T1DM and healthy individuals in two generations or more, from existing patient cohorts from the Centre Hospitalier du Luxembourg. Individual patients are annotated as a combination of their family and a number for each individual per family (e.g. M05.1). Recruited families were seen three times (V1, V2, V3) at intervals of between 4 and 8 weeks for data and samples collection. On enrolment, study participant pedigrees were drawn, medical history was collected and a 'Food Frequency Questionnaire' was completed. During every visit, anthropometric data were recorded as previously described [46] (Supplementary Data 1). Donors collected 2–3 ml of saliva at home before dental hygiene and breakfast in the early morning. Faecal samples were also self-collected and both samples were immediately frozen on dry-ice, transported to the laboratory and stored at − 80 °C until further processing. Part of the cohort's raw data (families M01–04) [41, 46] as well as the oral and gut metagenomics (families M05–11) [41, 46] were previously studied and published. The following method sections describe the processing of the newly produced dataset.

Biomolecular extractions

For each individual and visit, faecal and saliva samples were subjected to comprehensive biomolecular isolations.

For the faecal samples, 150 mg of each snap-frozen sample was reduced to a fine powder and homogenised in a liquid nitrogen bath followed by the addition of 1.5 ml of cold RNAlater and brief vortexing prior to incubation overnight at − 20 °C. After incubation, the sample was re-homogenised by shaking for 2 min at 10 Hz in an oscillating Mill MM 400 (Retsch) and subsequently centrifuged at 700×*g* for 2 min at 4 °C. The supernatant was retrieved and the cells were pelleted by centrifugation at 14,000×*g* for 5 min. Cold stainless steel milling balls and 600 µl of RLT buffer (Qiagen) were added to the pellet and this was re-suspended via quick vortexing. Cells were disrupted by bead beating in an Oscillating Mill MM 400 (Retsch) for 30 s at 25 Hz and at 4 °C. Finally, the lysate was transferred onto a QIAshredder column and centrifuged at 14,000×*g* for 2 min and the eluate retrieved for multi-omics extraction. The subsequent biomacromolecular extractions were based on the Qiagen Allprep kit (Qiagen) using an automated robotic liquid handling system (Freedom Evo, Tecan) as described in Roume et al. and in accordance with the manufacturer's instructions [47].

For the saliva samples, the individual snap-frozen sample was thawed on ice, and 1 ml was subsampled and centrifuged at 18,000×*g* for 15 min at 4 °C. The supernatant was discarded and the pellet directly refrozen in liquid nitrogen. Cold stainless steel milling balls were added to the frozen pellet for homogenisation by cryomilling for 2 min at 25 Hz in an oscillating Mill MM 400 (Retsch). Subsequently, 300 µl of methanol and 300 µl of chloroform were added before a second passage through the Oscillating Mill at 20 Hz for 2 min. After centrifugation at 14,000×*g* for 5 min, two phases (polar and non-polar) and a solid interphase were visible. The two phases were discarded and the solid interphase kept for multi-omics extraction. Stainless steel milling balls and 600 µl of RLT buffer (Qiagen) were added to the pellet, re-suspended via quick vortexing and cells were disrupted by bead beating in an Oscillating Mill MM 400 (Retsch) for 30 s at 25 Hz at 4 °C. The lysate was transferred onto a QIAshredder column and centrifuged at 14,000×*g* for 2 min. The subsequent steps were performed as described for the faecal samples.

DNA sequencing

After extraction, the retrieved DNA was depleted of leftover RNA by RNAse A treatment at 65 °C for 45 min. After ethanol precipitation, the samples were re-suspended in 50 µl nuclease-free water. The quality and quantity of the retrieved DNA were assessed both before and after treatment via gel electrophoresis and Nanodrop analysis (ThermoFisher Scientific).

Sequencing libraries for salivary samples were prepared using the NEBNext Ultra DNA Library Prep kit (New England Biolabs, Ipswich) using a dual barcoding system, and sequenced at 150 bp paired-end on Illumina HiSeq 4000 and Illumina NextSeq 500 machines.

RNA sequencing

The extracted RNA was treated with DNase I at 37 °C for 30 min and purified using phenol-chloroform. From the aqueous phase, RNA was precipitated with isopropanol and re-suspended in 50 µl nuclease free water.

RNA integrity and quantity were assessed before and after treatment using the RNA LabChip GX II (Perkin

scale

Elmer). Subsequently, 1 µg of RNA sample was rRNA-depleted using the RiboZero kit (Illumina, MRZB12424). Further library preparation of rRNA-depleted samples was performed using TruSeq Stranded mRNA library preparation kit (Illumina, RS-122-2101) according to the manufacturer's instructions apart from omitting the initial steps for mRNA pull-down. Prepared libraries were checked again using the RNA LabChip GX II (Perkin Elmer) and quantified using Qubit (Invitrogen). A 10-nM pool of the libraries was sent to the EMBL genomics platform for sequencing on a Illumina NextSeq 500 machine.

Protein processing and mass spectrometry

The following section describes the procedures for samples from families M05, M06, M08, and M11. For a description of the protein processing of samples from families M01–M04 see Heintz-Buschart et al. [46].

Extracted proteins were processed and digested using the S-Trap^{TM} system (ProtiFi) following manufacturer's instructions. Briefly, protein suspensions were solubilised with SDS then reduced, alkylated and acidified for complete denaturation.

Approximately 200 µl of samples were transferred onto the S-Trap column and centrifuged until all of the sample volume was transferred. The columns were then washed twice with 180 µl S-Trap protein binding buffer. Protein digestion was performed by adding 20 µl of 0.04 µg/µl trypsin solution to each column, to achieve a trypsin to protein ratio of 1:50. Incubation was performed for three hours at 47 °C in a Thermomixer. Tryptic peptides were eluted with 40 µl 50 mM TEAB, 40 µl 0.1% acetic acid, and 35 µl 60% acetonitrile with 0.1% acetic acid at 4000×*g* for 1 min per elution. Samples were dried at 45 °C in a vacuum centrifuge and stored at − 20 °C.

Peptides were fractionated into eight fractions using the high pH reversed-phase peptide fractionation kit (Pierce^{TM} Thermo Fisher Scientific) according to the manufacturer's instructions and using self-made columns as previously described [48]. Digested, dried peptides were resuspended in 300 µl of 0.1% trifluoroacetic acid and suspensions transferred onto the columns. After centrifugation at 3000×*g* for 2 min the eluate was retained as "flow-through"-fraction. Columns were then washed with 300 µl water (ASTM Type I) at 3000×*g* for 4 min. Separation of samples into eight fractions was performed using 300 µl of elution solutions with increasing concentrations of acetonitrile in 0.1% trifluoroacetic acid at 3000×*g* for 4 min. Each elution fraction was collected in a separate microcentrifuge tube, dried at 45 °C in a vacuum centrifuge and stored at − 20 °C.

Peptide concentrations were measured for fraction two of each sample using the Quantitative Fluorometric Peptide Assay kit (Pierce^{TM} Thermo Fisher Scientific) according to the manufacturer's instructions.

Of each of the samples, for each fraction, the volume for 170 ng of peptides were loaded onto in-house built columns (100 µm × 20 cm), filled with 3 µm ReproSil-Pur material and separated using a non-linear 100 min gradient from 1 to 99% buffer B (99.9% acetonitrile, 0.1% acetic acid in water (ASTM Type I) at a flow rate of 300 nl/min operated on an EASY-nLC 1200. Measurements were performed on an Orbitrap Elite mass spectrometer performing one full MS scan in a range from 300 to 1700 m/z followed by a data-dependent MS/MS scan of the 20 most intense ions, a dynamic exclusion repeat count of 1, and repeat exclusion duration of 30 s.

Metagenomic and metatranscriptomic data analysis

For each individual time point, metagenomic (MG) and metatranscriptomic (MT) data were processed and co-assembled using the Integrated Meta-omic Pipeline (IMP) [49] which includes steps for the trimming and quality filtering of the reads, the filtering of rRNA from the MT data, and the removal of human reads after mapping against the human genome (hg38). Pre-processed DNA and RNA reads were co-assembled using the IMP-based iterative co-assembly using MEGAHIT 1.0.3 [50]. After co-assembly, prediction and annotation of open-reading frames (ORFs) were performed using IMP and followed by binning and then taxonomic annotation at both the contig and bin level. MG and MT read counts for the predicted genes obtained using featureCounts [51] were linked to the different annotation sources (KEGG [52], Pfam [53], Resfams [54], dbCAN [55], Cas [56], and DEG [57], as well as to taxonomy (mOTUs 2.5.1 [58] and Kraken2 using the maxikraken2_1903_140GB database [59]). Kraken2 annotations were used to generate read count matrices for each taxonomic rank (phylum, class, order, family, genus, and species) by summing up reads at the respective levels.

Identification of variants

IMP produced the mapping of the processed DNA and RNA reads against the final co-assembled contigs with the Burrows-Wheeler Aligner tool (BWA 0.7.17) [60] using the BWA-MEM algorithm with default parameters. Additionally for each individual, the oral DNA reads from all available visits were mapped against the gut contigs produced from all available visits with the same parameters.

All alignment files per sample were used to call variants using bcftools 1.9 [61, 62]. Bcftools *mpileup* was run on the gut contigs as reference FASTA file with default parameters except for the *--max-depth* being set to 1000 to increase variant calling certainty. Called variants were filtered based on their quality and read depth with minimum values set to 20 and 10, respectively and indels were excluded. Subsequently, in order to reinforce confidence in the variant calling, variants were kept for downstream analysis, only if they fitted the following criteria: (i) positive allelic depths on both the forward and reverse strands for the corresponding gut and oral DNA reads, and (ii) presence of an alternative allele (genotype = 1 in the vcf file) at the oral DNA reads and the gut RNA read levels. These criteria ensured that the variants were resolved in both the gut and oral samples at both the DNA and RNA levels.

Because we have different assemblies, we obtained different mappings and different variants. In order to perform a comparison between samples, the reads containing the variants were extracted from the mapping files and taxonomically annotated using Kraken2. For metaproteomics, missense variants (variant that leads to a different amino acid) were identified using an in-house script [46] and the generated ORFs containing variants were added to the metaproteomic database (see below).

Metaproteomic data analysis

As the mass spectrometry analysis of the protein fraction was performed at different facilities for families M01-04 and families M05, M06, M08, and M11, certain parts of the preprocessing workflow and analyses had to be tailored to the data, as mentioned below.

Raw files were converted to mzML format using ThermoRawFileParser [63] and to ms2 format using ProteoWizard's msconvert [64]. The files for families M01–04 were filtered for the top 300 most intense spectra, the files for the other families for the top 150 most intense spectra to optimise protein identifications.

For each sample, microbial protein sequence databases were constructed from the Prokka [65] predicted protein sequences of the IMP co-assemblies and supplemented with variant protein sequences (missense variants) identified in both the oral cavity and the gut, during the variant calling step. This was done in order to consider only the variant sequences originating from the oral cavity that could also be found in the gut. If no database was available for a single sample, all databases available from the individual were concatenated. If an individual had no database, all databases from the individual's family were concatenated. In addition, the human RefSeq protein sequences (release 92), a collection of plant storage proteins that might be present due to food intake as well as the cRAP contaminant database (release 04/03/2019) were added. The databases were then filtered according to size (60–40,000 residues) to eliminate noise from very large or small proteins that can be erroneously produced during the ORF prediction step. Duplicate sequences were removed by sequence using SeqKit [66].

Concatenated target-decoy databases were built using Sipros Ensembles *sipros_prepare_protein_database.py*. Using Sipros Ensemble [67], each sample was searched against the prepared database for that sample. Identifications were filtered to a protein FDR of 1%.

After the search, human and microbial protein identifications were treated separately. Human proteins/protein groups that ended up having identical protein identifiers after processing the database identifiers in the output were collapsed and their spectral counts summed up. The same was done for the microbial proteins but gene identifiers were replaced by the corresponding annotation identifiers from the respective source (e.g. KEGG, Pfam. (see above)).

Diversity analysis

Raw read counts per taxon for each sample were transformed from absolute counts to relative abundances by dividing each value by samples total taxon read counts. The richness as a total number of detected species after filtering was recorded as well as alpha diversity using the Simpson index [68]. Beta diversity was analysed using Bray-Curtis as a distance measure with hierarchical clustering, distance-based redundancy analysis (dbRDA), and nonmetric multi-dimensional scaling (NMDS). Significance tests between groups were carried out using the Mann–Whitney–Wilcoxon test (MWW) or analysis of variance (ANOVA, dbRDA formula: *species~condition+family*). Analyses were performed in R using the *picante* [69] and *vegan* [70] packages.

Statistical analyses

An initial screening was performed based on MG and MT sequencing and assembly statistics, principal component analysis and hierarchical clustering on gene abundances to highlight potential outliers. Samples whose sequencing and assembly statistics consistently appeared outside ± 1.5× the interquartile range and clustered substantially differently compared to other samples from the same individual with hierarchical clustering were considered as outliers and removed from the dataset. Similarly, filtering was performed for the MP data with MS raw data quality and protein identification rate. After quality control, several individuals were removed because of their high variability due to either a very young age (age under 4 years old for M08–04 and M11–03) or a comorbidity that was not present in the rest of the dataset (T2DM for M11–05 and M11–06).

After taxonomic and functional analysis, gene/taxa read count and protein spectral count matrices were generated for differential abundance and expression analysis

scale

using the DESeq2 R package [71]. As the sampling visits for each individual are not independent, the median value for each gene/protein of the available visits for each individual was computed to obtain a matrix with one representative value per gene/protein per individual. Additionally, genes in read count matrices were removed if they did not have at least 20 reads in 25% of all the individuals, ensuring sufficient representation of the gene in the sample set for downstream statistical analyses. Proteins in the spectral count matrices were removed if they did not have at least 10 spectra in 25% of all the individuals. Finally, family membership was set as confounder for the DESeq2 the differential analyses.

Correlation analyses were performed on the same filtered matrices and combined depending on what correlations were tested. Spearman's rank correlation coefficients were calculated with two-sided significance tests corrected for multiple testing using the Benjamini-Hochberg method. For the correlations between transcripts and differentially active taxa in the oral cavity a significance threshold of 0.001 and a correlation threshold of 0.7 were applied and the analysis was performed with the *rcorr* function of the *Hmisc* R package (https://github.com/harrelfe/Hmisc). All other correlation analyses were performed with a significance threshold of 0.05 using the *rstatix* R package (https://github.com/kassambara/rstatix).

Results and discussions

Study description

In this study, we performed a multi-omic oral and gut microbiome study of eight families with at least two T1DM cases per family (Fig. 1A). This expanded on previous studies focusing on a subset of the data [41, 46]. The present work additionally includes metagenomic (MG) and metatranscriptomic (MT) analyses of the oral cavity for all participants. In total, we analysed 84 stool and 76 saliva samples from 35 individuals coming from multiple visits. We generated MG data for 84 stool and 74 saliva, MT data for 64 stool and 71 saliva, and MP data for 71 stool sample (Table 1). Of the 35 individuals, 17 were T1DM patients and 18 were healthy family members (Fig. 1A). In total, 653.4 Gbp of DNA sequencing data, 870.6 Gbps RNA sequencing data, and 13,833,325 fragment ion spectra were acquired.

Over all samples, the DNA and RNA sequencing data per sample amounted to on average 4.2 ± 0.9 Gbp for MG and 6.3 ± 1.6 Gbp for MT. While the gut data consisted of 4.2 ± 0.8 Gbp of MG and 5.6 ± 1.1 Gbp MT sequencing data, the oral data represented 4.2 ± 0.9 Gbp of MG and 7.0 ± 1.6 Gbp MT sequencing data. For the stool samples, on average 95,000 ± 59,000 MS2 scans were performed and 4500 ± 3400 proteins identified. For samples from families 01–04 on average 63,000 ± 4700 fragment ion scans were obtained. The database searches resulted in 1500 ± 300 proteins on average. A mean of 203,000 ± 11,800 fragment ion scans were obtained for samples from families 05, 06, 08, and 11 and 8000 ± 1600 proteins could be identified. For detailed statistics see Supplementary Table 1. In the present study, we combined information from three omes in order to identify and follow strain-variants across the two body sites. To be able to do so, the overlap among the different omes had to be maximised to preserve all their sample specificity. Thus, the complete set of contigs from sample-specific assemblies were used rather than metagenome-assembled genomes that would have only covered a subset of all the multi-omic data (Fig. 1B).

Overall microbial community structure does not differ significantly between T1DM and healthy controls

We compared the community structures of both body sites between T1DM patients and controls using the MG data. Overall, the number of total species detected in the gut varied more in healthy individuals, but no significant differences in richness (MWW: *p* val 0.72, Supplementary Fig. 1A) nor in Simpson's index of diversity were observed (MWW: *p* val 0.53, Supplementary Fig. 1B). Beta diversity differed significantly according to family membership but not between T1DM patients and controls (ANOVA on dbRDA; *p* vals 0.001 (family), 0.11 (condition); R^2 0.49; Supplementary Fig. 1C).

The oral microbiota did not differ significantly in species richness (MWW: *p* val 0.48, Supplementary Fig. 1A) nor in their Simpson's Index of Diversity (MWW: *p* val 0.90, Supplementary Fig. 1B). The beta diversity, as in the gut, showed no significant difference for T1DM but for family membership (ANOVA on dbRDA, *p* vals 0.5 (condition), 0.003 (family); R^2 0.37; Supplementary Fig. 1C). Thereby, for both body sites, no evidence was found that suggested a significant effect of T1DM on the overall microbiota community diversity. As shown before, observable differences in oral community composition may instead be related to family membership [46].

The acidification of the oral cavity in T1DM impacts specific taxa and destabilises the equilibrium between *Streptococcus* species

Streptococcus species are the primary colonisers of the oral cavity and are key players in oral homeostasis and disease [72]. In healthy subjects, there is a balance between the abundance of opportunistic pathogens (e.g. *S. mutans* or *S. pneumoniae*) and non-pathogenic commensal species (e.g. *S. salivarius*, *S. parasanguinis*, or *S. mitis)* which compete with each other via different

scale

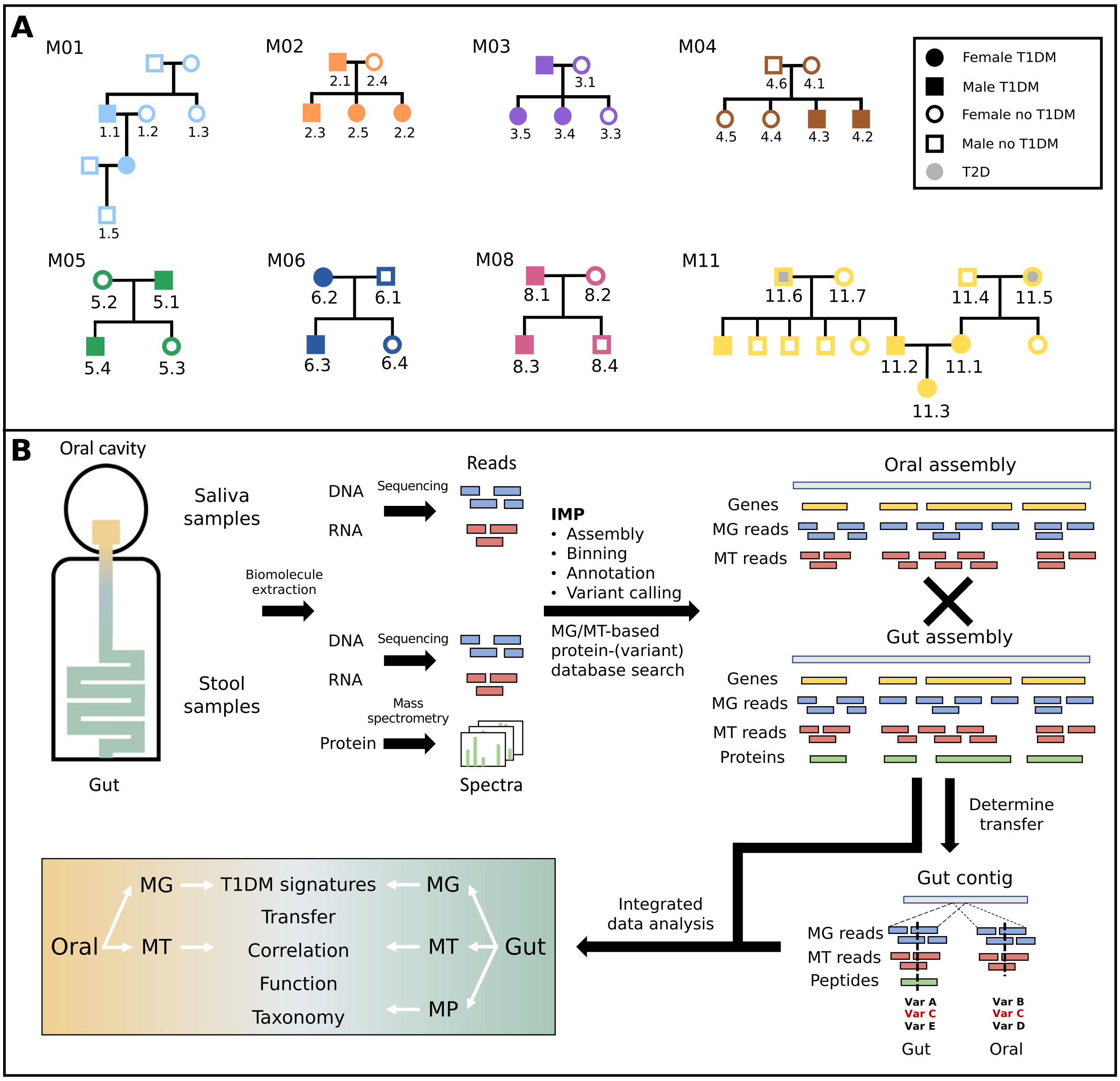


Fig. 1 Description of the cohort and overview of the study workflow. The upper panel (**A**) shows the different individuals with family membership as well as disease status in the cohort. The lower panel (**B**) describes the integrated multi-omics analysis workflow to process, integrate and analyse metagenomic (MG), metatranscriptomic (MT), and metaproteomic (MP) data from saliva and stool samples

mechanisms such as acid or base production, or secretion of bacteriocins [72–75].

In our study, the abundance of several members of the genus *Streptococcus* varied in the oral cavity of T1DM patients compared to controls. In particular at the MG level, we observed high variability among *Streptococcus* species (Fig. 2). Such variability is in agreement with previous findings whereby the numbers of different *Streptococcus* species were found to be increased or decreased in T1DM depending on the study [76, 77]. For example, a 16S rRNA gene-based study of both body sites observed an increase in the abundance of the genus *Streptococcus* in the mouth but a decrease in the gut of T1DM patients [45].

We observed an increased abundance of the acid-tolerant but non-pathogenic *Streptococcus parasanguinis* and closely related *Streptococcus* HMSC073D05 ($\log_2$ fold changes 3.5 and 3.4, respectively; adj. p val < 0.05). In contrast, the abundance of the commensal and

scale

Table 1 Overview of the multi-omics study data

	Total samples	MG samples	MG sequencing data [Gbp]	MG read recruitment [%]	MT samples	MT sequencing data [Gbp]	MT read recruitment [%]	MP samples	MP MS2 scans [k]	MP protein groups [k]
Stool	84	84	4.2 ± 0.8	95.3 ± 2.5	64	5.6 ± 1.1	56.8 ± 9.8	71	94.7 ± 59.4	4.5 ± 3.5
Oral	76	74	4.2 ± 0.9	83.5 ± 12.1	74	7.0 ± 1.6	62.7 ± 16.7	–	–	–

Number of samples used in total and per ome, average values of sequencing and mass spectrometry (MS) raw data and key metrics from saliva and stool samples. The number of samples per ome and location pertains to the number of available biomolecule samples available from the individuals. Sequencing data acquired are shown in Giga base pairs (Gbp) and percentage of recruitment of sequencing reads to assemblies, fragment ion spectra acquired (MS2 scans), and protein groups identified. All values are shown with standard deviation

MG metagenomics, *MT* metatranscriptomics, *MP* metaproteomics, *k* kilo, *Gbp* Giga base pairs

scale

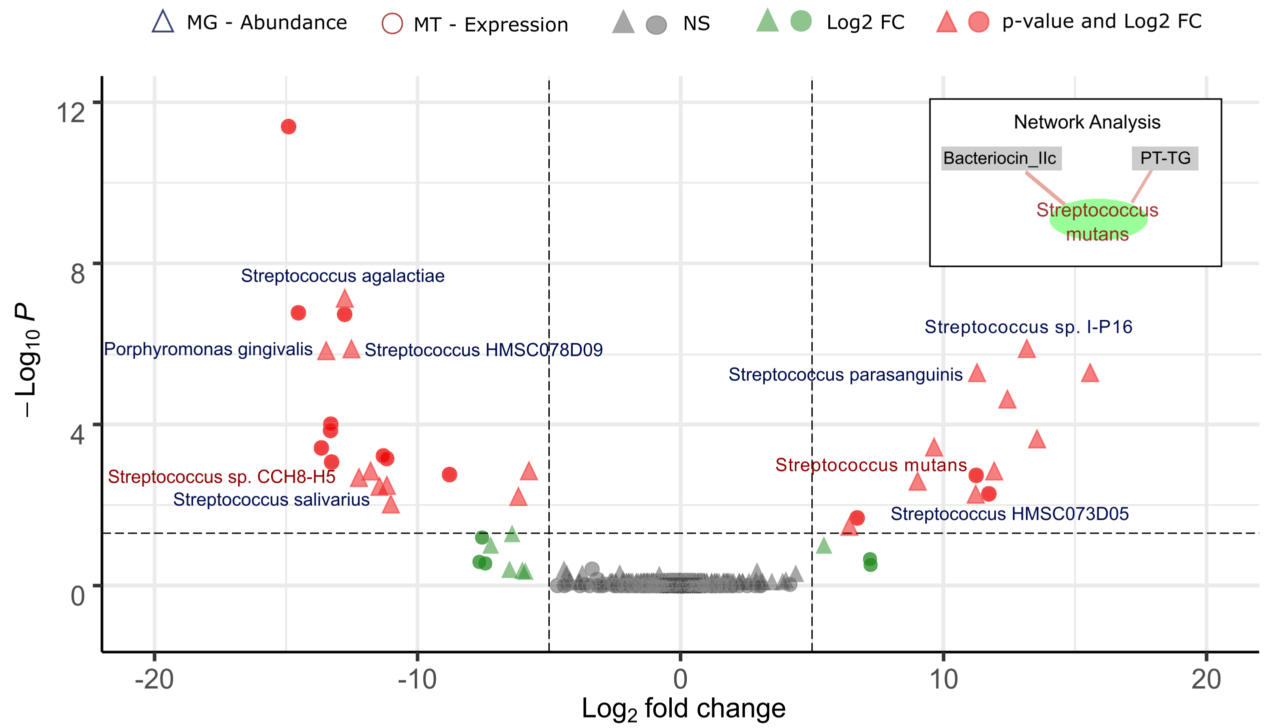


Fig. 2 Taxon-resolved differential abundance and gene expression in the oral microbiome in T1DM. The differences in abundance (triangles) and expression (circle) in T1DM versus healthy individuals using metagenomic and metatranscriptomic data, respectively, are shown on the volcano plot. A minimum $\log_2$ fold change of 5 (dashed vertical lines) and an adjusted *p* value of 0.01 (dashed horizontal line) were required (red dots). Taxa that satisfy the fold-change threshold but not the adjusted *p* value threshold are displayed in green. A subset of Supplementary Fig. 2 is shown in the insert in the upper-right and highlights the correlation between *S. mutans* activity and the expression of a target-specific bacteriocin

acid-intolerant *Streptococcus salivarius* was found to be decreased in T1DM ($\log_2$ fold change − 3.5; adj. *p* val < 0.05) [78]. Additionally, we observed a decreased abundance of *Porphyromonas gingivalis* in the cavity of T1DM patients. *P. gingivalis* is usually associated with a dysbiotic state but is also known to be unable to grow in acidic conditions [79]. Taken together, these results indicate a microbial profile corresponding to an acidified cavity in the case of T1DM patients [42–44, 80].

Further evidence was provided by the metatranscriptomic data, which showed a significantly increased activity of the pathogenic *Streptococcus mutans* [81] ($\log_2$ fold change 11.3; adj. *p* val < 0.05), while other *Streptococci*, notably *S. salivarius/S.* sp. CCH8-H5 ($\log_2$ fold change − 13.3 at adj. *p* val < 0.05) were less active (Fig. 2). *S. mutans* is a common pathogen of the oral cavity associated with periodontal diseases and known for its acid-tolerance and acidogenicity, which leads to further microbial acidification of the oral cavity in T1D patients [82, 83].

In order to better understand the underlying patterns in the oral microbiomes, we looked at correlations of the expressed genes with the taxa that were found to be differentially active. We observed significant positive correlations (rho > 0.7 at *p* value < 0.001) between *S. mutans* and two specific expressed transcripts related to bacterial competition among closely related species, namely bacteriocin IIc and pre-toxin TG, which are the constituent domains of uberolysin (Fig. 2—network analysis and Supplementary Fig. 2). This peptidic toxin is a circular bacteriocin characterised in the genus *Streptococcus* and has a broad spectrum of inhibitory activity, which includes most streptococci with the notable exception of *S. rattus* and *S. mutans* [84, 85]. The corresponding gene expression was not found to be linked with a particular species. However, the fact that *S. mutans* is resistant to the toxin and the observation that *S. mutans* is strongly correlated with both transcripts for this toxin, supports our hypothesis that *S. mutans* is responsible for the expression of the bacteriocin. The acidified oral cavity of T1DM patients, originally due to the host pathophysiology [42–44], according to our data, leads to the decreased abundance of acid-intolerant bacteria and favours the growth of acid-tolerant pathogenic *S. mutans*, which then further acidifies the environment and outcompetes the commensal *S. salivarius* by expressing a target-specific bacteriocin.

scale

Streptococcus salivarius' abundance decreases in the gut favouring an inflamed environment and an enterobacterial bloom

The differential abundance analysis of the gut-derived multi-omic data showed few differences between conditions. The lower abundance of *S. salivarius* in the gut follows the trend we observed in the oral cavity (Supplementary Table 2). *S. salivarius* colonises the intestine of adults and contributes to gut homeostasis by anti-inflammatory effects as well as by preventing the bloom of pathogens [86–88]. Previous studies have shown that a *S. salivarius* strain isolated from the oral cavity was able to prevent inflammatory responses both in vitro and in vivo by significantly reducing the activation of NF-κB and IL-8 secretion in intestinal epithelial and immune cell lines [86, 89, 90]. Therefore, a decrease of *S. salivarius* abundance may culminate in a more inflamed gut environment.

We also observed an increased abundance in the *Escherichia coli* (*Enterobacteria*) in the gut (Supplementary Table 2). *Enterobacteria* are among the most commonly overgrowing potential pathobionts whose expansion is associated with many diseases and, in particular, inflammation [91].

By investigating gene expression in the gut, we found multiple differentially expressed genes in T1DM in comparison to healthy controls (Fig. 3). Strikingly, a majority of the overexpressed genes are associated with *Enterobacteria* indicating a strong activity of this group in T1DM patients. They are usually found in low abundance in the gut in close proximity to the mucosal epithelium due to their facultative anaerobic metabolism [92]. *Enterobacteria* are also well known to have their growth favoured in many conditions involving inflammation [93]. The identified overexpressed genes contribute to bacterial virulence, oxidative stress response, cell motility and biofilm formation, and general replication and growth. Notably, an upregulation of a catalase-peroxidase was identified, an enzyme that detoxifies reactive oxygen intermediates such as H_2O_2 and, thus, is involved in protection against oxidative stress produced by the host. Enzymes associated with biofilm formation (YliH) were also overexpressed. Finally, OmpA-like transmembrane domain was identified as well the protein HokC/D, which corresponds to the *E. coli* toxin-antitoxin system that ensures the transmission of the associated plasmid.

There are multiple possible mechanisms of inflammation-driven blooms of *Enterobacteria* in the gut. One of them relies on the inflammatory host response that

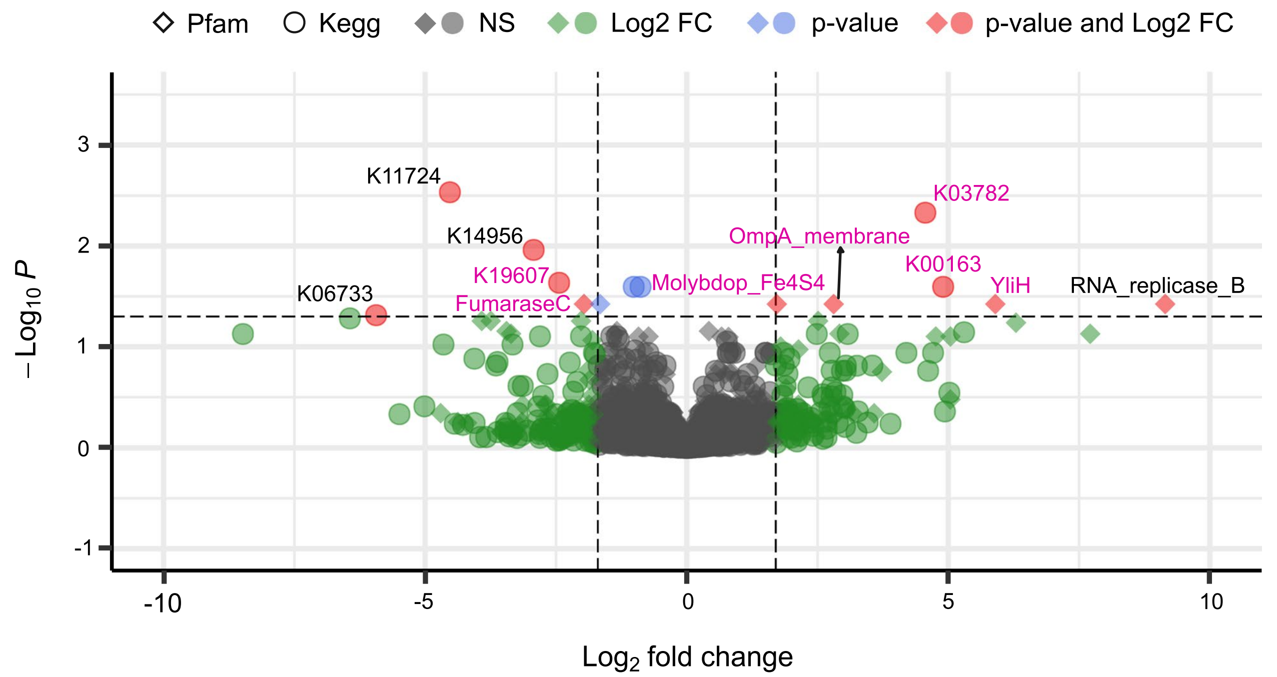


Fig. 3 Differential gene expression analysis within the gut in T1DM. Difference in expression using metatranscriptomic data is shown on the volcano plot. A minimum $\log_2$ fold change of 2 (dashed vertical lines) and adjusted *p* value of 0.05 (dashed horizontal line) were required (red dot). Functions that satisfy only the fold change or the adjusted *p* value threshold are displayed in green and blue, respectively. Diamonds and circles respectively indicate complementary annotations from both the Pfam and KEGG databases. Genes associated with *Enterobacteria* are marked in pink

scale

produces a potent antimicrobial agent (peroxynitrite) which is quickly converted to nitrate and can then be used for bacterial growth through nitrate respiration [93]. Since the genes encoding nitrate reductase in the gut are mostly encoded by *Enterobacteria*, this nitrate-rich environment provides a growth advantage for *Enterobacteria* such as *E. coli.* In addition to the genes involved in oxidative stress, we also found the molybdopterin oxidoreductase 4Fe-4S domain to be overexpressed in T1DM (Fig. 3 and Supplementary Table 3). This domain is found in a number of reductase/dehydrogenase families and notably the respiratory nitrate reductase in *E. coli* which further supports our hypothesis of inflamed gut in the context of T1DM. Increased abundance of *Enterobacteria* in T1DM has been partially observed before but the signal was not necessarily clear [94] or was associated with confounding factors like antibiotic-induced acceleration of T1DM [95] and no functional evidence were found.

Additionally, we looked at the effect of T1DM on the abundance of human proteins in the gut. We hypothesised that inflammation of the gut would lead to higher abundances of proteins involved in the host immune response. Interestingly, we mostly found evidence of exocrine pancreatic insufficiency with several types of proteases, such as pancreatic carboxypeptidases, elastases or trypsin-related enzymes, being less abundant in T1DM (Fig. 4 and Supplementary Table 4) which can be associated with T1DM [96]. One protein involved in the host immune response, the polymeric immunoglobulin receptor (pIgR), was found at elevated levels in T1DM ($\log_2$ fold change 0.42 at p val < 0.05) (Fig. 4 and Supplementary Table 4). pIgR is a transmembrane protein expressed by epithelial cells and responsible for the transcytosis of the secreted polymeric IgA produced in the mucosa by plasma cells to the gut lumen [97, 98]. Binding of polymeric IgA to the microbial surface protects the intestinal mucosa by preventing attachment to the epithelial cells, thus inhibiting infection and colonisation. When looking at differentially expressed proteins taking all visits as independent samples into account (see "Methods" section), we found similar proteins as when using median information but also several additional proteins associated with the host immune response and inflammation to be more expressed in T1DM (Supplementary Fig. 3 and

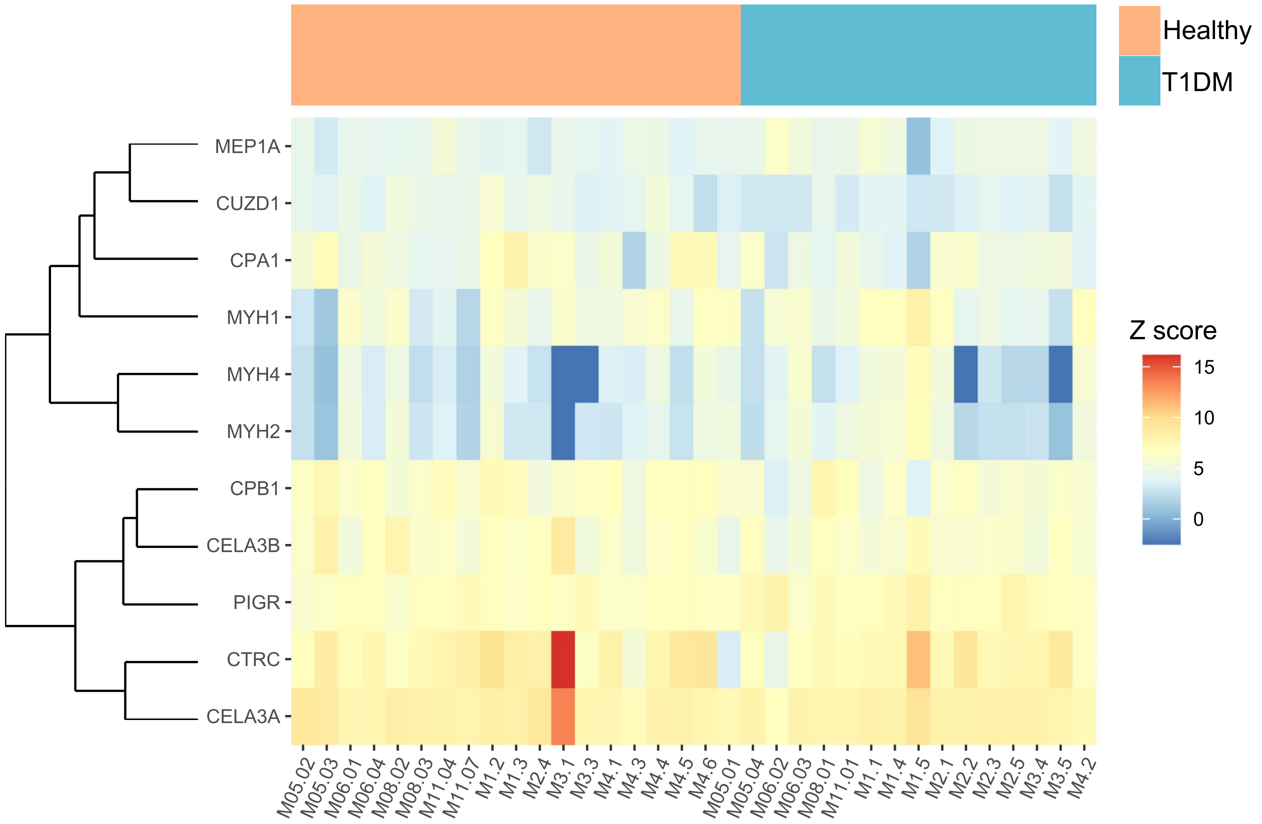


Fig. 4 Human proteome differences in T1DM. Heatmap displaying the relative abundances of human proteins with the highest significance in a differential analysis of T1DM versus healthy individuals (unadjusted p value < 0.05). The samples are ordered by conditions. Healthy individuals and T1DM patients are respectively shown in orange and blue boxes

scale

Supplementary Table 5). While that approach is statically less robust (see "Methods" section), it allows to observe additional trends in the dataset. Notably, we found higher levels of the lipocalin 2 enzyme (LCN2) ($\log_2$ fold change 0.37 at p val < 0.05) which is a typical biomarker in human inflammatory disease [99] and has been associated with metabolic disorders such as obesity and diabetes [100–102]. The analysis also confirmed the higher expression of the lactotransferrin (LTF) ($\log_2$ fold change 0.76 at p val < 0.05), which was already found in our previous study [46]. LTF plays a role in innate immunity and insulin function [103, 104] and its antimicrobial activity can influence the gastrointestinal microbiota [105].

Multi-omics integration highlights the transfer and the activity of bacteria from the oral cavity to the gut

Since a lower abundance of *S. salivarius* was found in both the oral cavity and the gut, we sought to explore the transmission between both extremities of the gastrointestinal tract and assess the levels of transfer in our cohort. To do so, we identified and followed genomic variations with read support from both the oral cavity and the gut (see "Methods"). In contrast to a previous study that only looked at the transmission using MG-based strain-variants [41], we additionally took advantage of the MT and/or MP data to identify not only transferred but also functionally active strain-variants. Furthermore, while MG and MT analyses are based on sequencing, metaproteomics provides an independent layer of information based on peptides and mass-spectrometry analyses. This provides the opportunity to strongly validate identified

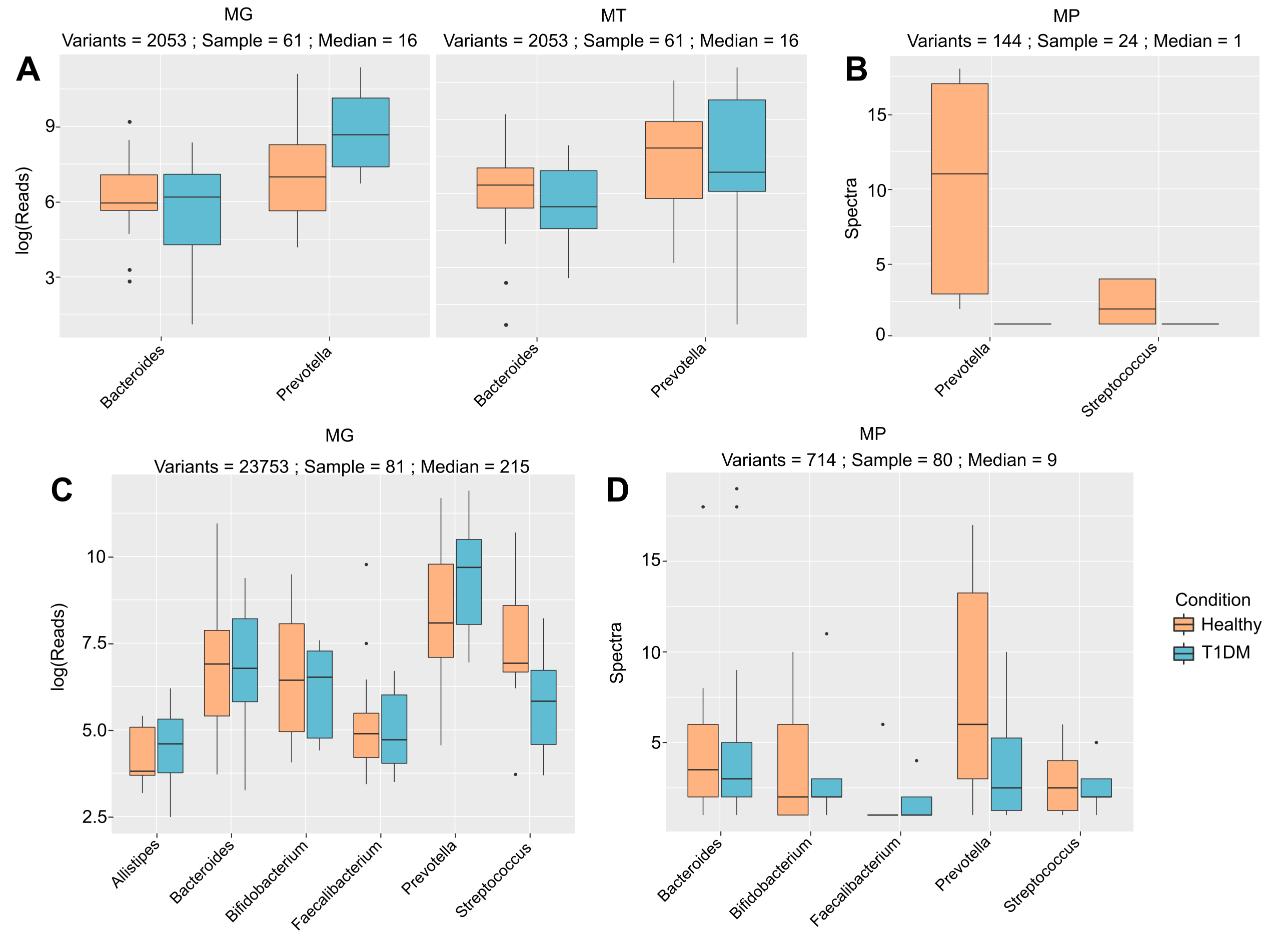


Fig. 5 Identified variants of genera across multiple omes. The figure indicates the distribution of reads for metagenomic (MG) and metatranscriptomic (MT) abundance, and spectra for metaproteomic (MP) abundance for each set of variants associated with a taxa. The numbers on top of each box indicate the number of identified variants, the number of samples in which variants have been identified and the median number of variants per sample. **A** and **B** correspond to the MG-MT supported variants while **C** and **D** show the MG-only supported variants. Comparisons of distributions were also performed and are represented by a light orange (healthy controls) and a light blue box (T1DM patients)

scale

transferred missense variants by identifying the translated protein with the variant amino acid sequence. Using first all genomic variants (synonymous and missense) with read support from both the oral cavity and the gut, we identified the genera *Prevotella* and *Bacteroides* to be transferred and active at the MT level in the gut in the majority of our cohort (Fig. 5A). The genus *Prevotella* is relatively common and abundant in the oral cavity but less prevalent in the gut. Finding it to be transferred and active is thus not surprising. In contrast, while the genus *Bacteroides* is strongly abundant in the gut, it is rarely identified in the oral cavity [9, 106]. Indeed, in our study, the signal observed from *Bacteroides* mostly came from a few particular individuals and was not representative of the entire cohort.

Remarkably, we identified several peptides supporting strain-variants at the MP level (Fig. 5B), showing that we could follow, and thus validate, variants across all three omic layers. Whilst the number of variant-supporting peptides is relatively low (due notably to the typical lower depth of MP or expected lower abundance of variant peptides), their identification confirms that the taxa we find to be transferred from the oral cavity are also active in the gut. Strain-variants belonging to the genus *Bacteroides* is not identified anymore at the variant peptide level, which can be explained by the low number of samples in which *Bacteroides* was identified. More surprising is the absence of the *Streptococcus* genus using MT-supported variants but its presence at the MP level. This indicates that the representation of strain-variants belonging to the genus *Streptococcus* was too low at the MT but not at the MP level to be detected over their respective threshold (see "Methods" section and Supplementary Table 1). Additionally, *Streptococci* are known to inhabit the upper part (small intestine) of the gut rather than the lower part (colon) [107, 108]. As RNA transcripts are less stable than proteins, it is not surprising that only peptides are identifiable from taxa active in the upper gut. We thus hypothesised that the applied strict MT read abundance threshold might be too stringent to identify transferred bacteria active in the upper part of the gut and that MP support would be more appropriate. To test this, we used missense variants with only MG read support and performed the metaproteomic search including the new protein variants. We distinguish variants supported by MG from the oral cavity and MG and MT from the gut (referred to as MG-MT supported variants) and variants supported only by MG from the oral cavity and MG from the gut (referred to as MG-only supported variants). Both types of variants can be further supported at the MP level (Fig. 5B, D).

By applying only the MG support criterion, around 10 times more variants across 81 samples were found and additional genera including *Alistipes*, *Bifidobacterium*, and *Faecalibacterium* were identified as transferred. With the exception of *Faecalibacterium*, all those taxa are commonly found at both body sites [9]. As hypothesised, strain-variants belonging to the genus *Streptococcus* were now found at the MG level (Fig. 5C). Adding the MP layer notably confirmed the presence and the activity of the *Streptococcus* strain-variants while those from the genera *Alistipes* and *Faecalibacterium* (initially not found by MG-MT variants) are not found (Fig. 5D). Metaproteomics thus essentially supports and validates the variants detected via the others omes, either due to the higher stability of proteins or to metaproteomics' different and independent technology (e.g. it does not suffer from sequencing errors). Furthermore, as proteins are immunogenic, using metaproteomics to detect strain-variant peptides adds a valuable layer of information as proteins from the oral cavity may fuel inflammation in the large intestine.

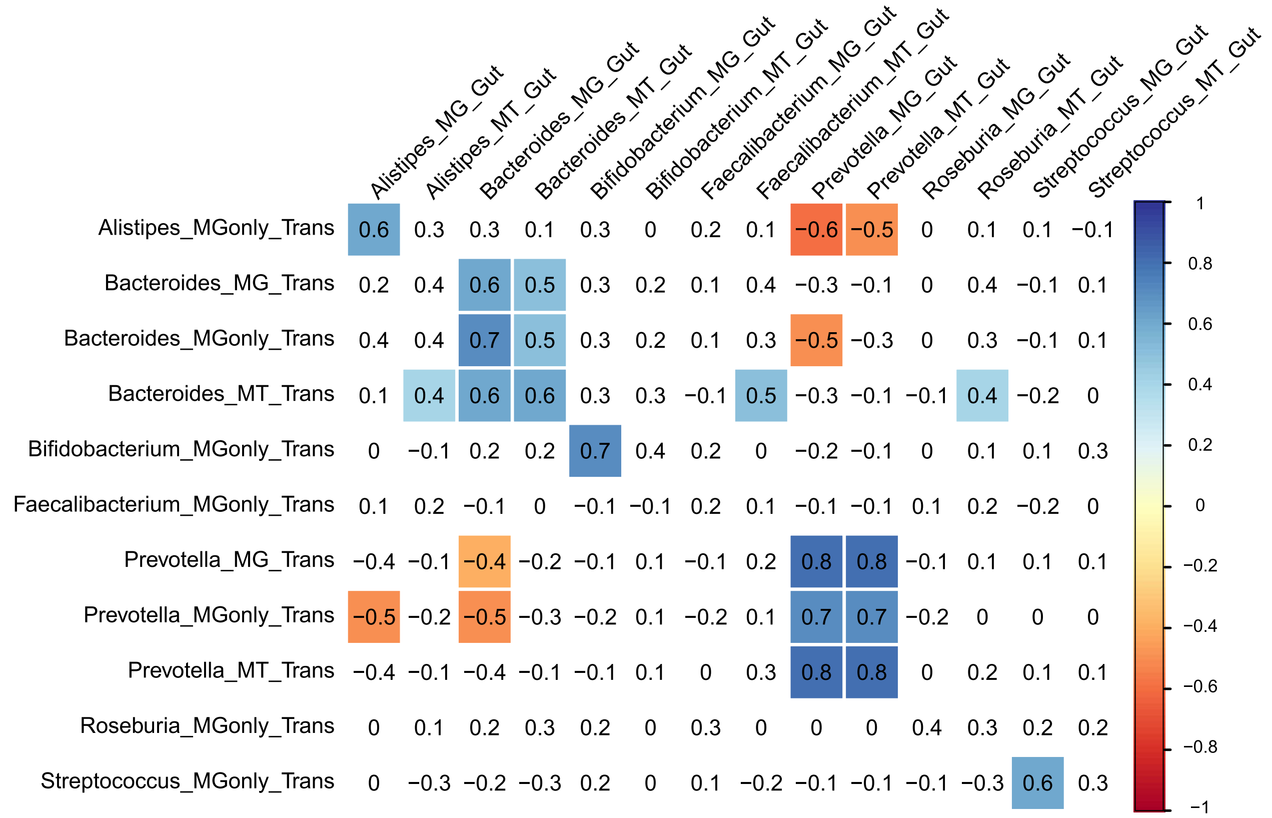
Streptococcus is less transmitted in T1DM in comparison to healthy controls

Being able to identify and follow variants across all omic layers and both body sites allowed us to assess the level of transfer of the different identified taxa. *Streptococcus salivarius* was found to be less abundant and less active in both the oral cavity and the gut in TIDM. While the difference is not significant, a similar trend can be observed at the transfer level for the *Streptococcus* genus. Not only does *Streptococcus* seem less transferred at the MG-only level (Fig. 5C), this trend seems to be further supported by lower amount of peptides, and thus a lower activity, associated to *Streptococcus* at the metaproteomic level using both the MG-MT supported variants and the MG-only supported variants (Fig. 5B, D). This suggests that the lower abundance of *S. salivarius* in the oral cavity and in the gut may indeed be connected. However, the lack of taxonomic resolution due to the method employed prevents strong conclusions. Further analyses should use a common assembly for all samples and be fully resolved at the species level to validate our findings.

Transmission levels strongly correlate with taxa abundances in the gut but not in the oral cavity

Correlation analyses between the MG and MT levels of the transferred bacteria and their abundance in the oral cavity and in the gut were performed in order to verify if the taxa abundances at both extremities of the gastrointestinal tract were associated. Strong positive correlations ($r_s = 0.6–0.7$ at p value < 0.05) were found between the abundance (MG and MG_Only) of the transferred bacteria and their abundance in the gut, which indicates that the levels of transfer indeed influences the final

scale



	Alistipes_MG_Gut	Alistipes_MT_Gut	Bacteroides_MG_Gut	Bacteroides_MT_Gut	Bifidobacterium_MG_Gut	Bifidobacterium_MT_Gut	Faecalibacterium_MG_Gut	Faecalibacterium_MT_Gut	Prevotella_MG_Gut	Prevotella_MT_Gut	Roseburia_MG_Gut	Roseburia_MT_Gut	Streptococcus_MG_Gut	Streptococcus_MT_Gut
Alistipes_MGonly_Trans	0.6	0.3	0.3	0.1	0.3	0	0.2	0.1	−0.6	−0.5	0	0.1	0.1	−0.1
Bacteroides_MG_Trans	0.2	0.4	0.6	0.5	0.3	0.2	0.1	0.4	−0.3	−0.1	0	0.4	−0.1	0.1
Bacteroides_MGonly_Trans	0.4	0.4	0.7	0.5	0.3	0.2	0.1	0.3	−0.5	−0.3	0	0.3	−0.1	0.1
Bacteroides_MT_Trans	0.1	0.4	0.6	0.6	0.3	0.3	−0.1	0.5	−0.3	−0.1	−0.1	0.4	−0.2	0
Bifidobacterium_MGonly_Trans	0	−0.1	0.2	0.2	0.7	0.4	0.2	0	−0.2	−0.1	0	0.1	0.1	0.3
Faecalibacterium_MGonly_Trans	0.1	0.2	−0.1	0	−0.1	−0.1	0.2	0.1	−0.1	−0.1	0.1	0.2	−0.2	0
Prevotella_MG_Trans	−0.4	−0.1	−0.4	−0.2	−0.1	0.1	−0.1	0.2	0.8	0.8	−0.1	0.1	0.1	0.1
Prevotella_MGonly_Trans	−0.5	−0.2	−0.5	−0.3	−0.2	0.1	−0.2	0.1	0.7	0.7	−0.2	0	0	0
Prevotella_MT_Trans	−0.4	−0.1	−0.4	−0.1	−0.1	0.1	0	0.3	0.8	0.8	0	0.2	0.1	0.1
Roseburia_MGonly_Trans	0	0.1	0.2	0.3	0.2	0	0.3	0	0	0	0.4	0.3	0.2	0.2
Streptococcus_MGonly_Trans	0	−0.3	−0.2	−0.3	0.2	0	0.1	−0.2	−0.1	−0.1	−0.1	−0.3	0.6	0.3

Fig. 6 Correlations between the abundances of transferred taxa in comparison to the abundance in the gut. The figure shows the correlation between the transfer and the gut abundances. Abundances of taxa with either MG or MT labels correspond to the abundances of supported variants at the metagenomic and metatranscriptomic levels. MGonly is used if variants were supported with MG reads only and not on the MT level. Colored values indicate positive (blue) or negative (red) significant correlations (adj. *p* value < 0.05). Values with white background indicate non-significant correlations

abundance of the taxa in the gut (Fig. 6). The activities (MT) were also positively correlated but at lower values ($r_s = 0.4–0.5$ at p value < 0.05). Interestingly, no correlations were found between the oral MG abundance of the taxa and their level of transfer (Supplementary Fig. 4), which is consistent with the correlations found in our previous study [41]. This would suggest that the transfer rate does not simply depend on the original abundance of the taxa in the oral cavity, but rather is driven by other parameters. For example, the host physiology of the oral cavity (saliva flow-rate, glucose concentration, pH) might affect the levels of transmission along the gastrointestinal tract as well as the microbial physiology (e.g. low pH and bile acids tolerance). We therefore looked at correlations between the level of transfer and the available metadata but no strong significant correlations were found (Supplementary Fig. 5, Supplementary Data 1).

Conclusions

In this study, we looked at the microbiota of two important body sites at both extremes of the gastrointestinal tract, the oral cavity and the gut, and identified differences in composition, function, and transfer of bacterial taxa in a case study of familial T1DM.

In the oral cavity of T1DM patients, the abundances of different taxa strongly resembled an acidified cavity. Notably, we found a lower abundance and activity of the commensal acid-intolerant *S. salivarius* and a higher activity of the acid-tolerant pathogenic *S. mutans*, which additionally correlated with the expression of a bacteriocin, highlighting competition between the two *Streptococci* species (Fig. 2).

In the gut, we observed lower abundance of *S. salivarius* and higher abundance of *E. coli* as well as an overall increased expression of genes involved in bacterial virulence and oxidative stress response related to the *Enterobacteriaceae* family (Fig. 3). Besides the increased abundance and activity of *Enterobacteria*, we found further evidence of gut inflammation in T1DM through the overexpression of several human proteins involved either in the host immune response or inflammation (Fig. 4 and Supplementary Fig. 3).

The multi-omic data for both body sites enabled us for the first time to trace the variants and taxa across all three omic layers and thus to identify specific taxa that were both transmitted along the gastrointestinal tract, and active in the gut. This strengthened the identification of transmitted variants and brought additional evidence on actual gut colonisation by oral bacteria. We found multiple genera to be transmitted and we have highlighted the importance of using functional omic support to identify taxa active in the gut (Fig. 5). We also discussed the limitations inherent to metatranscriptomics and highlighted how metaproteomics can be advantageously used to validate identified variants and explore the upper part of the gut.

By contextualising the information concerning oral to gut transfer in T1DM, we notably found a trend of lower levels of transmission of *Streptococcus* in T1DM patients, thereby reinforcing the notion that the lower abundance of *S. salivarius* in the oral cavity and the gut are indeed connected and both in relation to T1DM (Fig. 5B, D). However, correlations between the levels of transmission of taxa and their abundance at both body sites showed strong correlations with the gut but not with the oral cavity (Fig. 6 and Supplementary Fig. 4). As the physiology of the oral cavity is altered in T1DM patients [42–44], we would hypothesise that some of those factors (e.g. saliva flow-rate, glucose concentration, pH) might have a stronger influence on the transmission rate of oral microbes along the gastrointestinal tract than just their initial abundances. A follow-up study could combine different metadata measurements of the oral cavity together with the newly developed strain-variant methodology and assess if any physiological parameter influences the abundance of particular variants and their transmission rate along the gastrointestinal tract.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-022-01435-4.

Additional file 1: Supplementary Data 1. Metadata.

Additional file 2: Supplementary Table 1. Summary statistics of metrics extracted from IMP.

Additional file 3: Supplementary Figure 1. Oral and gut community structure analysis. Box plots of species richness (a) and alpha diversity (b) between controls and T1DM patients. Hierarchical clustering based on Bray-Curtis distance (c). NMDS ordination based on Bray-Curtis distance (d). All analyses are based on metagenomic data. **Supplementary Figure 2.** Correlation network of gene transcripts from all samples and differentially active taxa in the oral cavity. The figure corresponds to a subset of the correlation analysis where only the differentially abundant taxa and correlation over 0.7 are plotted. Green and red nodes indicate if the taxon is up- or down-regulated in the oral cavity. Annotations are based on the Pfam database. **Supplementary Figure 3.** Metaproteomic differences in T1DM at the visit level. Heatmap displaying the relative abundances of human proteins at the visit level with the highest significance in a differential analysis of T1DM. The samples are ordered by conditions. Healthy individuals and T1DM patients are respectively shown in orange and blue boxes. **Supplementary Figure 4.** Correlations between the abundances of transferred taxa in comparison to their abundance in the oral cavity. The figure shows the correlation between the transfer and the oral cavity. The labels with MG and MT correspond to the abundances at the metagenomic and metatranscriptomic levels for the MG-MT supported variants. MG_only is used for the MG abundances of variants supported by MG reads only and not on the MT level. Colored squares indicate a positive (blue) or negative (red) significant correlations (*p*val < 0.05). White squares indicate non-significant correlations. **Supplementary Figure 5.** Correlation analyses of transferred taxa and metadata. The figure shows the correlation between the transfer and the available metadata. The labels with MG and MT correspond to the abundances at the metagenomic and metatranscriptomic levels for the MG-MT supported variants. MG_only is used for the MG abundances of variants supported by MG reads only and not on the MT level. Colored squares indicate a positive (blue) or negative (red) significant correlations (*p*-val < 0.05). White squares indicate non-significant correlations.

Additional file 4: Supplementary Table 2. Eubacterium siraeum DSM 15702.

Additional file 5: Supplementary Table 3. Genera.

Additional file 6: Supplementary Table 4. Differentially expressed human proteins using median data.

Additional file 7: Supplementary Table 5. Differentially expressed human proteins using all individuals and visits.

Acknowledgements
Not applicable.

Authors' contributions
B.J.K. and O.H contributed equally as the main authors of the manuscript; doing the data curation, analysis and visualisation, and writing the manuscript. L.A.L. and A.H-B. carried out the sample collection and the biomolecular extractions. R.H. performed the sequencing. D.B. and O.H. provided the metaproteomic data. P.Q. and P.M participated in the data processing and analysis. C.M-G., A.H-B, T.S.B.D.S., and P.M. helped with the data interpretation. T.S.B.D.S., M.R.H. C.d.B., and P.B. P.W. conceived the study and participated in its design. PM and P.W. coordinated the study. B.J.K, O.H., A.H-B., P.M., and P.W. wrote the manuscript. All authors read and approved the final manuscript.

Funding
We would like to thank the European Molecular Biology Laboratory (EMBL) for its support. This work was supported by the Luxembourg National Research Fund (FNR) under grants CORE/15/BM/10404093, CORE/19/BM/13684739, PRIDE/11823097, and the European Research Council (ERC-CoG 863664) to P.W.

Availability of data and materials
Metagenomic and metatranscriptomic sequencing reads can be accessed from NCBI BioProject PRJNA289586. All mass spectrometry proteomics data and results were deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE [109] partner repository with the data set identifier PXD031579. All custom scripts are available at https://git-r3lab.uni.lu/ESB/must2.

Declarations

Ethics approval and consent to participate
This study was approved by the Comité d'Ethique de Recherche (CNER; reference no. 201110/05) and the National Commission for Data Protection in Luxembourg. Written informed consent was obtained from all subjects enrolled in the study.

scale

Consent for publication

Not applicable.

Competing interests

All authors declare that they have no competing interests.

Author details

[1]Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg. [2]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [3]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA. [4]Institute of Microbiology, University of Greifswald, Greifswald, Germany. [5]Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands. [6]Clinique Pédiatrique, Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg. [7]Max Delbrück Centre for Molecular Medicine, Berlin, Germany. [8]Yonsei Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea. [9]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. [10]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Belvaux, Luxembourg.

Received: 30 March 2022 Accepted: 4 October 2022
Published online: 28 December 2022

References

1. Gilbert JA, et al. Current understanding of the human microbiome. Nat Med. 2018;24:392–400.
2. Karczewski J, Poniedziałek B, Adamski Z, Rzymski P. The effects of the microbiota on the host immune system. Autoimmunity. 2014;47:494–504.
3. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. Cell Res. 2020;30:492–506.
4. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8:1784.
5. Gilbert JA, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature. 2016;535:94–103.
6. Wen L, et al. Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature. 2008;455:1109–13.
7. Hooper LV, Littman DR, Macpherson AJ. Interactions between the microbiota and the immune system. Science. 2012;336:1268–73.
8. Honda K, Littman DR. The microbiota in adaptive immune homeostasis and disease. Nature. 2016;535:75–84.
9. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
10. Lloyd-Price J, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
11. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020;18:491–506.
12. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14:347–55.
13. Denef VJ. Peering into the genetic makeup of natural microbial populations using metagenomics. In: Polz MF, Rajora OP, editors. Population genomics: microorganisms. New York City: Springer; 2019. p. 49–75.
14. Zojer M, et al. Variant profiling of evolving prokaryotic populations. PeerJ. 2017;5:e2997.
15. Wang Z, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011;472:57–63.
16. Frank DN, et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. 2007;104:13780–5.
17. Turnbaugh PJ, et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444:1027–31.
18. Garrett WS. Cancer and the microbiota. Science. 2015;348:80–6.
19. Spielman LJ, Gibson DL, Klegeris A. Unhealthy gut, unhealthy brain: the role of the intestinal microbiota in neurodegenerative diseases. Neurochem Int. 2018;120:149–63.
20. Zhang X, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med. 2015;21:895–905.
21. Paun A, Yau C, Danska JS. The influence of the microbiome on type 1 diabetes. J Immunol. 2017;198:590–5.
22. Sharma S, Tripathi P. Gut microbiome and type 2 diabetes: where we are and where to go? J Nutr Biochem. 2019;63:101–8.
23. Diaz-Valencia PA, Bougnères P, Valleron A-J. Global epidemiology of type 1 diabetes in young adults and adults: a systematic review. BMC Public Health. 2015;15:255.
24. Dabelea D. The accelerating epidemic of childhood diabetes. Lancet. 2009;373:1999–2000.
25. Patterson CC, et al. Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study. Lancet. 2009;373:2027–33.
26. Mobasseri M, et al. Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis. Health Promot Perspect. 2020;10:98–115.
27. Rewers M, Ludvigsson J. Environmental risk factors for type 1 diabetes. Lancet. 2016;387:2340–8.
28. Rooks MG, Garrett WS. Gut microbiota, metabolites and host immunity. Nat Rev Immunol. 2016;16:341–52.
29. Brown CT, et al. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. PLoS ONE. 2011;6:e25792.
30. Alkanani AK, et al. Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. Diabetes. 2015;64:3510–20.
31. Giongo A, et al. Toward defining the autoimmune microbiome for type 1 diabetes. ISME J. 2011;5:82–91.
32. Jamshidi P, et al. Is there any association between gut microbiota and type 1 diabetes? A systematic review. Gut Pathog. 2019;11:49.
33. Xiao J, Fiscella KA, Gill SR. Oral microbiome: possible harbinger for children's health. Int J Oral Sci. 2020;12:12.
34. Hajishengallis G. Periodontitis: from microbial immune subversion to systemic inflammation. Nat Rev Immunol. 2015;15:30–44.
35. Cullinan MP, Seymour GJ. Periodontal disease and systemic illness: will the evidence ever be enough? Periodontol 2000. 2013;62:271–86.
36. Song I-S, et al. Severe periodontitis is associated with insulin resistance in non-abdominal obese adults. J Clin Endocrinol Metab. 2016;101:4251–9.
37. Borgnakke WS, Ylöstalo PV, Taylor GW, Genco RJ. Effect of periodontal disease on diabetes: systematic review of epidemiologic observational evidence. J Periodontol. 2013;84:S135–52.
38. Segata N, et al. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol. 2012;13:R42.
39. Martinsen TC, Bergh K, Waldum HL. Gastric juice: a barrier against infectious diseases. Basic Clin Pharmacol Toxicol. 2005;96:94–102.
40. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature. 2014;509:357–60.
41. Schmidt TS, et al. Extensive transmission of microbes along the gastrointestinal tract. Elife. 2019;8:e42693.
42. Naing C, Mak JW. Salivary glucose in monitoring glycaemia in patients with type 1 diabetes mellitus: a systematic review. J Diabetes Metab Disord. 2017;16:2.
43. Seethalakshmi C, Reddy RCJ, Asifa N, Prabhu S. Correlation of salivary pH, incidence of dental caries and periodontal status in diabetes mellitus patients: a cross-sectional study. J Clin Diagn Res. 2016;10:ZC12–4.
44. Gandara BK, Morton TH. Non-periodontal oral manifestations of diabetes: a framework for medical care providers. Diabetes Spectr. 2011;24:199–205.
45. de Groot PF, et al. Distinct fecal and oral microbiota composition in human type 1 diabetes, an observational study. PLoS ONE. 2017;12:e0188475.
46. Heintz-Buschart A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2:16180.
47. Roume H, et al. A biomolecular isolation framework for eco-systems biology. ISME J. 2013;7:110–21.
48. Kroniger T, et al. Proteome analysis of the Gram-positive fish pathogen Renibacterium salmoninarum reveals putative role of membrane

scale

vesicles in virulence. Res Square. 2021. https://doi.org/10.21203/rs.3.rs-744942/v1.

49. Narayanasamy S, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016;17:260.
50. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
51. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
52. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
53. El-Gebali S, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.
54. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015;9:207–16.
55. Zhang H, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:W95–101.
56. Burstein D, et al. New CRISPR-Cas systems from uncultivated microbes. Nature. 2017;542:237–41.
57. Luo H, et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res. 2021;49:D677–86.
58. Milanese A, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10:1014.
59. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.
60. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
61. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011;27:1157–8.
62. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
63. Hulstaert N, et al. ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. J Proteome Res. 2020;19:537–42.
64. Chambers MC, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30:918–20.
65. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
66. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
67. Guo X, et al. Sipros Ensemble improves database searching and filtering for complex metaproteomics. Bioinformatics. 2018;34:795–802.
68. Simpson EH. Measurement of diversity. Nature. 1949;163:688.
69. Kembel SW, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26:1463–4.
70. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–30.
71. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
72. Abranches J, et al. Biology of oral Streptococci. Microbiol Spectr. 2018;6(5):6–5.
73. Nes IF, Diep DB, Holo H. Bacteriocin diversity in Streptococcus and Enterococcus. J Bacteriol. 2007;189:1189–98.
74. Mignolet J, et al. Circuitry rewiring directly couples competence to predation in the gut dweller Streptococcus salivarius. Cell Rep. 2018;22:1627–38.
75. Hibbing ME, Fuqua C, Parsek MR, Peterson SB. Bacterial competition: surviving and thriving in the microbial jungle. Nat Rev Microbiol. 2010;8:15–25.
76. Dedrick S, et al. The role of gut microbiota and environmental factors in type 1 diabetes pathogenesis. Front Endocrinol. 2020;11:78.
77. Babatzia A, et al. Clinical and microbial oral health status in children and adolescents with type 1 diabetes mellitus. Int Dent J. 2020;70:136–44.
78. Garnett JA, et al. Structural insight into the role of Streptococcus parasanguinis Fap1 within oral biofilm formation. Biochem Biophys Res Commun. 2012;417:421–6.
79. Takahashi N, Saito K, Schachtele CF, Yamada T. Acid tolerance and acid-neutralizing activity of Porphyromonas gingivalis, Prevotella intermedia and Fusobacterium nucleatum. Oral Microbiol Immunol. 1997;12:323–8.
80. Takahashi N. Oral microbiome metabolism: from 'who are they?' to 'what are they doing?' J Dent Res. 2015;94:1628–37.
81. Lemos JA, et al. The biology of Streptococcus mutans. Microbiol Spectr. 2019;7. https://doi.org/10.1128/microbiolspec.GPP3-0051-2018.
82. Matsui R, Cvitkovitch D. Acid tolerance mechanisms utilized by Streptococcus mutans. Future Microbiol. 2010;5:403–17.
83. Liu Y-L, Nascimento M, Burne RA. Progress toward understanding the contribution of alkali generation in dental biofilms to inhibition of dental caries. Int J Oral Sci. 2012;4:135–40.
84. Wirawan RE, Swanson KM, Kleffmann T, Jack RW, Tagg JR. Uberolysin: a novel cyclic bacteriocin produced by Streptococcus uberis. Microbiology. 2007;153:1619–30.
85. Gabrielsen C, Brede DA, Nes IF, Diep DB. Circular bacteriocins: biosynthesis and mode of action. Appl Environ Microbiol. 2014;80:6854–62.
86. Kaci G, et al. Anti-inflammatory properties of Streptococcus salivarius, a commensal bacterium of the oral cavity and digestive tract. Appl Environ Microbiol. 2014;80:928–34.
87. Villmones HC, et al. Species level description of the human ileal bacterial microbiota. Sci Rep. 2018;8:4736.
88. Couvigny B, et al. Commensal Streptococcus salivarius modulates PPARγ transcriptional activity in human intestinal epithelial cells. PLoS ONE. 2015;10:e0125371.
89. Cosseau C, et al. The commensal Streptococcus salivarius K12 down-regulates the innate immune responses of human epithelial cells and promotes host-microbe homeostasis. Infect Immun. 2008;76:4163–75.
90. Kaci G, et al. Inhibition of the NF-kappaB pathway in human intestinal epithelial cells by commensal Streptococcus salivarius. Appl Environ Microbiol. 2011;77:4681–4.
91. Winter SE, Bäumler AJ. Dysbiosis in the inflamed intestine: chance favors the prepared microbe. Gut Microbes. 2014;5:71–3.
92. Brenner DJ, Farmer JJ III. Enterobacteriaceae. In: Bergey's manual of systematics of archaea and bacteria. 2015. p. 1–24. https://doi.org/10.1002/9781118960608.fbm00222.
93. Zeng MY, Inohara N, Nuñez G. Mechanisms of inflammation-driven bacterial dysbiosis in the gut. Mucosal Immunol. 2017;10:18–26.
94. Soyucen E, et al. Differences in the gut microbiota of healthy children and those with type 1 diabetes. Pediatr Int. 2014;56:336–43.
95. Zhang X-S, et al. Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. Elife. 2018;7:e37816.
96. Campbell-Thompson M, Rodriguez-Calvo T, Battaglia M. Abnormalities of the exocrine pancreas in type 1 diabetes. Curr Diab Rep. 2015;15:79.
97. Kaetzel CS. The polymeric immunoglobulin receptor: bridging innate and adaptive immune responses at mucosal surfaces. Immunol Rev. 2005;206:83–99.
98. Kaetzel CS, Robinson JK, Chintalacharuvu KR, Vaerman JP, Lamm ME. The polymeric immunoglobulin receptor (secretory component) mediates transport of immune complexes across epithelial cells: a local defense function for IgA. Proc Natl Acad Sci U S A. 1991;88:8796–800.
99. Moschen AR, Adolph TE, Gerner RR, Wieser V, Tilg H. Lipocalin-2: a master mediator of intestinal and metabolic inflammation. Trends Endocrinol Metab. 2017;28:388–97.
100. Guo H, et al. Lipocalin 2, a regulator of retinoid homeostasis and retinoid-mediated thermogenic activation in adipose tissue. J Biol Chem. 2016;291:11216–29.
101. Bhusal A, Rahman MH, Lee I-K, Suk K. Role of hippocampal lipocalin-2 in experimental diabetic encephalopathy. Front Endocrinol. 2019;10:25.
102. Arellano-Buendía AS, et al. Urinary excretion of neutrophil gelatinase-associated lipocalin in diabetic rats. Oxid Med Cell Longev. 2014;2014:961326.
103. Legrand D, et al. Lactoferrin structure and functions. Adv Exp Med Biol. 2008;606:163–94.
104. Akiyama Y, et al. A lactoferrin-receptor, intelectin 1, affects uptake, sub-cellular localization and release of immunochemically detectable lactoferrin by intestinal epithelial Caco-2 cells. J Biochem. 2013;154:437–48.
105. Bertuccini L, et al. Lactoferrin prevents invasion and inflammatory response following E. coli strain LF82 infection in experimental model of Crohn's disease. Dig Liver Dis. 2014;46:496–504.

scale

106. Dewhirst FE, et al. The human oral microbiome. J Bacteriol. 2010;192:5002–17.

107. Zoetendal EG, et al. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. ISME J. 2012;6:1415–26.

108. Friedman ES, et al. Microbes vs. chemistry in the origin of the anaerobic gut lumen. Proc Natl Acad Sci U S A. 2018;115:4170–5.

109. Perez-Riverol Y, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47:D442–50.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

6 Perspectives and Conclusions

6.1 Meta-omics technologies

Meta-omics technologies have experienced rapid advances. However, there is a critical need to address the challenges arising from the high complexity of input from samples of many microbial communities. The vast amounts of biomolecules with orders of magnitudes of difference in abundance, originating from a large array of (unknown) organisms, make their recording, attribution of origin, description of function and precise quantification difficult.

6.1.1 Experimental approaches

Ensuring the best sampling and storage methods is essential to ensure the sample is intact and stays in the exact condition as it was at the point of sampling, preserving information from low abundant and/or sensitive organisms. Fast flash-freezing in an adequate medium, for example, will allow long-term storage of various biomolecules [231]. Better sampling strategies, which ensure sample integrity while being affordable and easy to employ, both crucial when conducting large scale studies, will help minimize information loss early on, which will be irrecoverable at later stages of sample processing and data analysis.

There is also a need for advanced high-throughput sample preparation strategies. Pre-separating members of a sampled community, thereby reducing complexity will not only aid in recovering more information during sequencing and mass spectrometry by allowing the recovery of signals of organisms with low abundance or activity as well as enabling the capture of, for example, large and/or fragmented genomes, but also provide valuable presorting of information which would be difficult to separate during computational analyses [232].

Furthermore, deeper sequencing is becoming increasingly feasible as costs continue to decline, further improving the complete recovery of sequences in a sample and ensuring deep coverage of genomes, which is valuable for various computational analysis strategies such as genome binning, variant calling and assembly.

Another significant development will be the shift from short-read towards long-read sequencing technologies. They provide much longer reads that span hard-to-assemble regions, leading to much more contiguous assemblies and consequently more complete MAGs and better annotation of genomes with consequently deeper understanding of genome structure and regulation.

Lastly, employing nanopore technologies for direct protein/peptide sequencing, would allow for substantial improvements in their identification, while also making the study of post-translational modifications much more feasible [233]. In the meantime, improvement in mass spectrometry-based approaches with higher resolution and faster mass spectrometers, as well as better chromatography approaches, are also underway. For example, routinely applying automated, multi-phase separation steps is becoming more common, allowing more spectra with higher resolution to be recovered from complex samples [234].

6.1.2 Computational approaches

The effective utilization of state-of-the-art data for reference-based methods and the development of new generative approaches represent areas where significant advancements can be achieved. For all omics technologies, making use of next-generation 'AI' tools, such as various deep learning methods, might allow substantial improvements in describing and interpreting data from microbial communities.

A key strategy lies in utilizing complex combinations of informative features, which are likely of great importance in understanding a community's or organism's biology, but are not always obvious to humans [64, 89, 235]. For example, this could include recovering a larger portion of a metagenome in the form of complete MAGs, based on information from selected k-mers of possibly large size and features from combinations of sequence subsets at distant positions in the genome, instead of applying the most common feature set of tetramer frequencies combined with average read coverage depth of sequences. There are also likely untapped features in read depth of coverage information of metagenomic contigs/scaffolds that could provide organism-specific signatures, improving the separation of closely related organisms.

Protein functional annotation could be enhanced through combined guilt-by-association-based strategies informed by complex features from synteny and/or tertiary/quaternary structure. This could potentially inform about function even when no references with similar sequences are available [103, 236, 237].

6.1.3 Multi-omics integration

The integration of multiple 'omics' disciplines in studying host-associated microbial communities has proven to be highly effective. The study presented in this thesis demonstrates that the combination of different omics offers more insights than the mere sum of their parts. It allowed for tracking gene/protein variants on multiple levels to infer organism body site transfer, connecting organism abundance with activity on transcript and protein level and showed disease-relevant host activities.

There are various other ways to make use of this kind of data, most of which are rarely used in meta-omics studies or confined to study isolates of single organisms. The correlation ratios between omes can provide insights into gene expression regulation. For instance, a high amount of a protein of unknown function in the metaproteome negatively correlating with the expression of a transcript in the metatranscriptome could hint at a transcription regulator. This provides the basis for further investigation into the protein's function. Similarly, ratios of transcripts to genes can be used to analyze gene expression regulation [238, 239, 240].

The metatranscriptome and/or metaproteome can be used to find and annotate genes or regulatory regions/motifs on genomes with non-standard structures [241, 242, 243]. This approach can be particularly valuable for investigating novel genes, regulatory elements, or unusual codon usage of unculturable organisms under native conditions by matching genome sequence, translation (mRNAs, tRNAs), and proteins. For example, mapping proteins/peptides back to unannotated regions of a MAG could reveal an unknown open reading frame (ORF), and transcripts mapping to the same region could help confirm the finding and elucidate regulatory elements [244]. Integrating additional omes, such as metabolomes, enables the construction of a full network of potential and realized functions of a community. This can be utilized, for example, to elucidate the multi-organism biochemical pathway to produce a complex vitamin, where information on which organisms have the capability to perform certain steps, which ones actively express these capabilities, and which activities are actually taking place [240].

When studying host-associated communities, it is as essential as it is challenging to integrate information from the host. Host genetics, to a certain extent, are employed to identify correlations between genes or sequence variants and microbiota or disease phenotypes [245]. Also, identifying biomolecules originating from the host interacting with microbiota is needed since it can inform about the status of host-microbiota homeostasis by indicating inflammation or other metabolic activities potentially relevant to, for example, a disease phenotype. In the study presented in this thesis, T1DM cases' expected phenotype of pancreatic exocrine insufficiency could be validated by detecting a significantly lower levels of pancreatic proteins. The metaproteome also provided evidence of inflammation in disease cases, linking host status to intestinal microbiota.

To integrate all these aspects into a cohesive workflow, pipelines that can utilize multiple omes to benefit from the synergistic information are vital to make multi-meta-omics analysis more feasible on a large scale and reproducible by providing a stable framework for analyses. An example of such a pipeline is the Integrated Meta-omics Pipeline (IMP), which integrates metagenomics, -transcriptomics, and -proteomics data [246]. It facilitates automated hybrid assembly of metagenomics and metatranscriptomics reads and, in its recent versions, uses Mantis for consensus functional annotation generation using state-of-the-art databases. Additionally, a high-performance binning module was implemented, including advanced binners such as binny, and a multi-step ensemble refining strategy. This combination of several strategies integrating multiple omes yields assemblies with high contiguity, which aids in recovering larger percentages of the assemblies as MAGs of high quality, which have high-resolution annotations, all of which are essential in dealing with the high volumes of omics data required to investigate microbial communities in a holistic manner.

Further improvement of the integration of multi-meta-omic data in this framework presents significant opportunities for advancement. Firstly, assembling long and short read data of metagenomes and transcriptomes together in a multi-level hybrid assembly approach, would likely yield much more contiguous assemblies with fewer errors.

In general, enhancing the data richness for downstream steps by initially separating reads into groups at the lowest taxonomic level where a majority of reads are still annotated, will have various benefits. It can improve assembly quality by reducing complexity, leading to more contiguous assemblies with fewer chimeric contigs mixing reads from different organisms. In an optimal scenario, the separation is so refined that isolate assemblers, typically not feasible for use with metagenomic data, can be used on the separated read groups. This would result in higher purity and contiguity of contigs, since those assemblers usually provide higher contiguity.Alternatively, there's the possibility of working on the level of assemblies. The contigs could, instead of the reads, be categorized into subgroups based on their taxonomy.

Regardless of the approach, pre-separation significantly benefits subsequent steps. It accelerates the binning process and reduces the likelihood of contamination by decreasing the chances of including chimeric contigs of different organisms that share read depth or k-mer frequency signals in the same MAG. Additionally, functional annotation techniques can utilize taxonomy-specific information to increase quality and coverage. This could involve using taxon-specific hidden Markov models (HMMs) for greater specificity or dynamically adjusting codon table selection to accurately identify the start and end positions of open reading frames (ORFs).

Moreover, the integration of often underutilized information in the context of multi-meta-omics will become more straightforward on pre-separated reads or contigs. Usually analyzed individually, the viral and eukaryotic components of host-associated microbial communities are rarely integrated together with prokaryotic data. Incorporating these elements could be critical in closing gaps in our understanding of microbiota interactions since these components are as integral to the evolutionary history and shaping of microbial communities' genotypes and phenotypes as the more commonly studied bacteria and archaea.

Finally, the task of modeling the interactions between all these elements across different omes remains a challenge. While some methods attempt this, usually for multi-omics of isolate samples, the complexity of interactions makes it difficult. It's unlikely that a single, simple feature set, such as the relative abundance of a taxon or gene of known function, can sufficiently explain a phenotype. Additionally, the low overlap between omes often provides insufficient data to calculate informative statistics with the required confidence [247]. This complexity underscores the need for more advanced methods to accurately interpret and utilize multi-meta-omic information, which will probably require various of the aforementioned improvements of experimental and computational omics methods.

6.1.4 Conclusion

In the future, making use of the discussed experimental and computational approaches in an integrative manner might allow to finally get a comprehensive view of microbial communities, filling the massive gaps that still remain in the knowledge of composition, function and interaction. This is especially relevant in understanding the key roles host-associated communities seem to play in many chronic and/or auto-immune diseases, where, at this time, only basic associations and mechanisms involving few microbial actors are thought to be understood and supporting findings well reproducible.

In the end, while progress is rapid, an extraordinarily large amount of work remains to be done on all stages of investigating microbial communities in different environments ranging from oceans to the human intestine, and thus deepen humanity's understanding of life on Earth and learn how to benefit from this knowledge.

Bibliography

[1] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, The Earth Microbiome Project Consortium, Jose L. Agosto Rivera, Lisa Al-Moosawi, John Alverdy, Katherine R. Amato, Jason Andras, Largus T. Angenent, Dionysios A. Antonopoulos, Amy Apprill, David Armitage, Kate Ballantine, Jirˇí Bárta, Julia K. Baum, Allison Berry, Ashish Bhatnagar, Monica Bhatnagar, Jennifer F. Biddle, Lucie Bittner, Bazartseren Boldgiv, Eric Bottos, Donal M. Boyer, Josephine Braun, William Brazelton, Francis Q. Brearley, Alexandra H. Campbell, J. Gregory Caporaso, Cesar Cardona, JoLynn Carroll, S. Craig Cary, Brenda B. Casper, Trevor C. Charles, Haiyan Chu, Danielle C. Claar, Robert G. Clark, Jonathan B. Clayton, Jose C. Clemente, Alyssa Cochran, Maureen L. Coleman, Gavin Collins, Rita R. Colwell, Mónica Contreras, Benjamin B. Crary, Simon Creer, Daniel A. Cristol, Byron C. Crump, Duoying Cui, Sarah E. Daly, Liliana Davalos, Russell D. Dawson, Jennifer Defazio, Frédéric Delsuc, Hebe M. Dionisi, Maria Gloria Dominguez-Bello, Robin Dowell, Eric A. Dubinsky, Peter O. Dunn, Danilo Ercolini, Robert E. Espinoza, Vanessa Ezenwa, Nathalie Fenner, Helen S. Findlay, Irma D. Fleming, Vincenzo Fogliano, Anna Forsman, Chris Freeman, Elliot S. Friedman, Giancarlo Galindo, Liza Garcia, Maria Alexandra Garcia-Amado, David Garshelis, Robin B. Gasser, Gunnar Gerdts, Molly K. Gibson, Isaac Gifford, Ryan T. Gill, Tugrul Giray, Antje Gittel, Peter Golyshin, Donglai Gong, Hans-Peter Grossart, Kristina Guyton, Sarah-Jane Haig, Vanessa Hale, Ross Stephen Hall, Steven J. Hallam, Kim M. Handley, Nur A. Hasan, Shane R. Haydon, Jonathan E. Hickman, Glida Hidalgo, Kirsten S. Hofmockel, Jeff Hooker, Stefan Hulth, Jenni Hultman, Embriette Hyde, Juan Diego Ibáñez-Álamo, Julie D. Jastrow, Aaron R. Jex, L. Scott Johnson, Eric R. Johnston, Stephen Joseph, Stephanie D. Jurburg, Diogo Jurelevicius, Anders Karlsson, Roger Karlsson, Seth Kauppinen, Colleen T. E. Kellogg, Suzanne J. Kennedy, Lee J. Kerkhof, Gary M. King, George W. Kling, Anson V. Koehler, Monika Krezalek, Jordan Kueneman, Regina Lamendella, Emily M. Landon, Kelly Lane-deGraaf, Julie LaRoche, Peter Larsen, Bonnie Laverock, Simon Lax, Miguel Lentino, Iris I. Levin, Pierre Liancourt, Wenju Liang, Alexandra M. Linz, David A. Lipson, Yongqin Liu, Manuel E. Lladser, Mariana Lozada, Catherine M. Spirito, Walter P. MacCormack, Aurora MacRae-Crerar, Magda Magris, Antonio M. Martín-Platero, Manuel Martín-Vivaldi, L. Margarita Martínez, Manuel Martínez-Bueno, Ezequiel M. Marzinelli, Olivia U. Mason, Gregory D. Mayer, Jamie M. McDevitt-Irwin, James E. McDonald, Krista L. McGuire, Kather-

ine D. McMahon, Ryan McMinds, Mónica Medina, Joseph R. Mendelson, Jessica L. Metcalf, Folker Meyer, Fabian Michelangeli, Kim Miller, David A. Mills, Jeremiah Minich, Stefano Mocali, Lucas Moitinho-Silva, Anni Moore, Rachael M. Morgan-Kiss, Paul Munroe, David Myrold, Josh D. Neufeld, Yingying Ni, Graeme W. Nicol, Shaun Nielsen, Jozef I. Nissimov, Kefeng Niu, Matthew J. Nolan, Karen Noyce, Sarah L. O'Brien, Noriko Okamoto, Ludovic Orlando, Yadira Ortiz Castellano, Olayinka Osuolale, Wyatt Oswald, Jacob Parnell, Juan M. Peralta-Sánchez, Peter Petraitis, Catherine Pfister, Elizabeth Pilon-Smits, Paola Piombino, Stephen B. Pointing, F. Joseph Pollock, Caitlin Potter, Bharath Prithiviraj, Christopher Quince, Asha Rani, Ravi Ranjan, Subramanya Rao, Andrew P. Rees, Miles Richardson, Ulf Riebesell, Carol Robinson, Karl J. Rockne, Selena Marie Rodriguezl, Forest Rohwer, Wayne Roundstone, Rebecca J. Safran, Naseer Sangwan, Virginia Sanz, Matthew Schrenk, Mark D. Schrenzel, Nicole M. Scott, Rita L. Seger, Andaine Seguin-Orlando, Lucy Seldin, Lauren M. Seyler, Baddr Shakhsheer, Gabriela M. Sheets, Congcong Shen, Yu Shi, Hakdong Shin, Benjamin D. Shogan, Dave Shutler, Jeffrey Siegel, Steve Simmons, Sara Sjöling, Daniel P. Smith, Juan J. Soler, Martin Sperling, Peter D. Steinberg, Brent Stephens, Melita A. Stevens, Safiyh Taghavi, Vera Tai, Karen Tait, Chia L. Tan, Neslihan Tas¸, D. Lee Taylor, Torsten Thomas, Ina Timling, Benjamin L. Turner, Tim Urich, Luke K. Ursell, Daniel Van Der Lelie, William Van Treuren, Lukas Van Zwieten, Daniela Vargas-Robles, Rebecca Vega Thurber, Paola Vitaglione, Donald A. Walker, William A. Walters, Shi Wang, Tao Wang, Tom Weaver, Nicole S. Webster, Beck Wehrle, Pamela Weisenhorn, Sophie Weiss, Jeffrey J. Werner, Kristin West, Andrew Whitehead, Susan R. Whitehead, Linda A. Whittingham, Eske Willerslev, Allison E. Williams, Stephen A. Wood, Douglas C. Woodhams, Yeqin Yang, Jesse Zaneveld, Iratxe Zarraonaindia, Qikun Zhang, and Hongxia Zhao. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, November 2017.

[2] Justin P. Shaffer, Louis-Félix Nothias, Luke R. Thompson, Jon G. Sanders, Rodolfo A. Salido, Sneha P. Couvillion, Asker D. Brejnrod, Franck Lejzerowicz, Niina Haiminen, Shi Huang, Holly L. Lutz, Qiyun Zhu, Cameron Martino, James T. Morton, Smruthi Karthikeyan, Mélissa Nothias-Esposito, Kai Dührkop, Sebastian Böcker, Hyun Woo Kim, Alexander A. Aksenov, Wout Bittremieux, Jeremiah J. Minich, Clarisse Marotz, MacKenzie M. Bryant, Karenina Sanders, Tara Schwartz, Greg Humphrey, Yoshiki Vásquez-Baeza, Anupriya Tripathi, Laxmi Parida, Anna Paola Carrieri, Kristen L. Beck, Promi Das, Antonio González, Daniel McDonald, Joshua Ladau, Søren M. Karst, Mads Albertsen, Gail Ackermann, Jeff DeReus, Torsten Thomas, Daniel Petras, Ashley Shade, James Stegen, Se Jin Song, Thomas O. Metz, Austin D. Swafford, Pieter C. Dorrestein, Janet K. Jansson, Jack A. Gilbert, Rob Knight, the Earth Microbiome Project 500 (EMP500) Consortium, Lars T. Angenant, Alison M. Berry, Leonora S. Bittleston, Jennifer L. Bowen, Max Chavarría, Don A. Cowan, Dan Distel, Peter R. Girguis, Jaime Huerta-Cepas, Paul R. Jensen, Lingjing Jiang, Gary M. King, Anton Lavrinienko, Aurora MacRae-Crerar, Thulani P. Makhalanyane, Tapio Mappes, Ezequiel M. Marzinelli, Gregory Mayer, Katherine D. McMahon, Jessica L. Metcalf, Sou Miyake, Timothy A. Mousseau, Catalina Murillo-Cruz, David Myrold, Brian Palenik, Adrián A. Pinto-Tomás, Dorota L. Porazinska, Jean-Baptiste Ramond, Forest Rowher, Taniya RoyChowdhury, Stuart A. Sandin, Steven K. Schmidt, Henning Seedorf, Ashley Shade, J. Reuben Ship-

way, Jennifer E. Smith, James Stegen, Frank J. Stewart, Karen Tait, Torsten Thomas, Yael Tucker, Jana M. U'Ren, Phillip C. Watts, Nicole S. Webster, Jesse R. Zaneveld, and Shan Zhang. Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nature Microbiology*, 7(12):2128–2150, November 2022.

[3] Wen-Sheng Shu and Li-Nan Huang. Microbial diversity in extreme environments. *Nature Reviews Microbiology*, 20(4):219–235, April 2022.

[4] Paul G. Falkowski, Tom Fenchel, and Edward F. Delong. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039, May 2008.

[5] S. Louca, L. W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, September 2016.

[6] Eric A. Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C. Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, June 2015.

[7] Danping Zheng, Timur Liwinski, and Eran Elinav. Interaction between microbiota and immunity in health and disease. *Cell Research*, 30(6):492–506, June 2020.

[8] Katherine Donald and B. Brett Finlay. Early-life interactions between the microbiota and immune system: impact on immune system development and atopic disease. *Nature Reviews Immunology*, 23(11):735–748, November 2023.

[9] Fanette Fontaine, Sondra Turjeman, Karel Callens, and Omry Koren. The intersection of undernutrition, microbiome, and child development in the first years of life. *Nature Communications*, 14(1):3554, June 2023.

[10] Muhammad Saleem, Jie Hu, and Alexandre Jousset. More Than the Sum of Its Parts: Microbiome Biodiversity as a Driver of Plant Growth and Soil Health. *Annual Review of Ecology, Evolution, and Systematics*, 50(1):145–168, November 2019.

[11] Manuel Delgado-Baquerizo, Fernando T. Maestre, Peter B. Reich, Thomas C. Jeffries, Juan J. Gaitan, Daniel Encinar, Miguel Berdugo, Colin D. Campbell, and Brajesh K. Singh. Microbial diversity drives multifunctionality in terrestrial ecosystems. *Nature Communications*, 7(1):10541, January 2016.

[12] Alicia E. Graham and Rodrigo Ledesma-Amaro. The microbial food revolution. *Nature Communications*, 14(1):2231, April 2023.

[13] R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, and U. T. Bornscheuer. From nature to industry: Harnessing enzymes for biocatalysis. *Science*, 382(6673):eadh8615, November 2023.

[14] Alan L. Harvey, RuAngelie Edrada-Ebel, and Ronald J. Quinn. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, 14(2):111–129, February 2015.

[15] the International Natural Product Sciences Taskforce, Atanas G. Atanasov, Sergey B. Zotchev, Verena M. Dirsch, and Claudiu T. Supuran. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery*, 20(3):200–216, March 2021.

[16] Jan Steensels, Brigida Gallone, Karin Voordeckers, and Kevin J. Verstrepen. Domestication of Industrial Microbes. *Current Biology*, 29(10):R381–R393, May 2019.

[17] James M. Clomburg, Anna M. Crumbley, and Ramon Gonzalez. Industrial biomanufacturing: The future of chemical production. *Science*, 355(6320):aag0804, January 2017.

[18] Margaret McFall-Ngai, Michael G. Hadfield, Thomas C. G. Bosch, Hannah V. Carey, Tomislav Domazet-Lošo, Angela E. Douglas, Nicole Dubilier, Gerard Eberl, Tadashi Fukami, Scott F. Gilbert, Ute Hentschel, Nicole King, Staffan Kjelleberg, Andrew H. Knoll, Natacha Kremer, Sarkis K. Mazmanian, Jessica L. Metcalf, Kenneth Nealson, Naomi E. Pierce, John F. Rawls, Ann Reid, Edward G. Ruby, Mary Rumpho, Jon G. Sanders, Diethard Tautz, and Jennifer J. Wernegreen. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*, 110(9):3229–3236, February 2013.

[19] Matthieu E. Galvez, Woodward W. Fischer, Samuel L. Jaccard, and Timothy I. Eglinton. Materials and pathways of the organic carbon cycle through time. *Nature Geoscience*, 13(8):535–546, August 2020.

[20] Tadeo Sáez-Sandino, Pablo García-Palacios, Fernando T. Maestre, César Plaza, Emilio Guirado, Brajesh K. Singh, Juntao Wang, Concha Cano-Díaz, Nico Eisenhauer, Antonio Gallardo, and Manuel Delgado-Baquerizo. The soil microbiome governs the response of microbial respiration to warming across the globe. *Nature Climate Change*, 13(12):1382–1387, December 2023.

[21] Mary Ann Moran, Elizabeth B. Kujawinski, William F. Schroer, Shady A. Amin, Nicholas R. Bates, Erin M. Bertrand, Rogier Braakman, C. Titus Brown, Markus W. Covert, Scott C. Doney, Sonya T. Dyhrman, Arthur S. Edison, A. Murat Eren, Naomi M. Levine, Liang Li, Avena C. Ross, Mak A. Saito, Alyson E. Santoro, Daniel Segrè, Ashley Shade, Matthew B. Sullivan, and Assaf Vardi. Microbial metabolites in the marine carbon cycle. *Nature Microbiology*, 7(4):508–523, April 2022.

[22] Zhe Lyu, Nana Shao, Taiwo Akinyemi, and William B. Whitman. Methanogenesis. *Current Biology*, 28(13):R727–R732, July 2018.

[23] Jonathan Gropp, Qusheng Jin, and Itay Halevy. Controls on the isotopic composition of microbial methane. *Science Advances*, 8(14):eabm5713, April 2022.

[24] Marcel M. M. Kuypers, Hannah K. Marchant, and Boran Kartal. The microbial nitrogen-cycling network. *Nature Reviews Microbiology*, 16(5):263–276, May 2018.

[25] David Fowler, Mhairi Coyle, Ute Skiba, Mark A. Sutton, J. Neil Cape, Stefan Reis, Lucy J. Sheppard, Alan Jenkins, Bruna Grizzetti, James N. Galloway, Peter Vitousek, Allison Leach,

Alexander F. Bouwman, Klaus Butterbach-Bahl, Frank Dentener, David Stevenson, Marcus Amann, and Maren Voss. The global nitrogen cycle in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1621):20130164, July 2013.

[26] Alexander J. Probst. ‘Omics Technologies. In Muriel Gargaud, William M. Irvine, Ricardo Amils, Henderson James Cleaves, Daniele Pinti, José Cernicharo Quintanilla, and Michel Viso, editors, *Encyclopedia of Astrobiology*, pages 1–2. Springer Berlin Heidelberg, Berlin, Heidelberg, 2020.

[27] Xiaofeng Dai and Li Shen. Advances and Trends in Omics Technology Development. *Frontiers in Medicine*, 9:911861, July 2022.

[28] Xu Zhang, Leyuan Li, James Butcher, Alain Stintzi, and Daniel Figeys. Advancing functional and translational microbiome research using meta-omics approaches. *Microbiome*, 7(1):154, December 2019.

[29] Raphaël Rodriguez and Yamuna Krishnan. The chemistry of next-generation sequencing. *Nature Biotechnology*, 41(12):1709–1715, December 2023.

[30] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, November 2021.

[31] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):5029, November 2019.

[32] Francesco Durazzi, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Remondini, and Alessandra De Cesare. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1):3030, February 2021.

[33] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. De Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G. S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso,

Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, April 2022.

[34] Eli L. Moss, Dylan G. Maghini, and Ami S. Bhatt. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, 38(6):701–707, June 2020.

[35] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, May 2017.

[36] Chao Yang, Debajyoti Chowdhury, Zhenmiao Zhang, William K. Cheung, Aiping Lu, Zhaoxiang Bian, and Lu Zhang. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314, 2021.

[37] Mikhail Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith, and Pavel A. Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, November 2020.

[38] Gaëtan Benoit, Sébastien Raguideau, Robert James, Adam M. Phillippy, Rayan Chikhi, and Christopher Quince. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology*, January 2024.

[39] Alexander J. Westermann and Jörg Vogel. Cross-species RNA-seq for deciphering host–microbe interactions. *Nature Reviews Genetics*, 22(6):361–378, June 2021.

[40] Elena Bushmanova, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8(9):giz100, September 2019.

[41] Manuel Kleiner. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems*, 4(3):e00115–19, June 2019.

[42] Nobuaki Miura and Shujiro Okuda. Current progress and critical challenges to overcome in the bioinformatics of mass spectrometry-based metaproteomics. *Computational and Structural Biotechnology Journal*, 21:1140–1150, 2023.

[43] Tim Van Den Bossche, Benoit J. Kunath, Kay Schallert, Stephanie S. Schäpe, Paul E. Abraham, Jean Armengaud, Magnus Ø. Arntzen, Ariane Bassignani, Dirk Benndorf, Stephan Fuchs, Richard J. Giannone, Timothy J. Griffin, Live H. Hagen, Rashi Halder, Céline Henry,

Robert L. Hettich, Robert Heyer, Pratik Jagtap, Nico Jehmlich, Marlene Jensen, Catherine Juste, Manuel Kleiner, Olivier Langella, Theresa Lehmann, Emma Leith, Patrick May, Bart Mesuere, Guylaine Miotello, Samantha L. Peters, Olivier Pible, Pedro T. Queiros, Udo Reichl, Bernhard Y. Renard, Henning Schiebenhoefer, Alexander Sczyrba, Alessandro Tanca, Kathrin Trappe, Jean-Pierre Trezzi, Sergio Uzzau, Pieter Verschaffelt, Martin Von Bergen, Paul Wilmes, Maximilian Wolf, Lennart Martens, and Thilo Muth. Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nature Communications*, 12(1):7305, December 2021.

[44] Jinzhi Zhao, Yi Yang, Hua Xu, Jianxujie Zheng, Chengpin Shen, Tian Chen, Tao Wang, Bing Wang, Jia Yi, Dan Zhao, Enhui Wu, Qin Qin, Li Xia, and Liang Qiao. Data-independent acquisition boosts quantitative metaproteomics for deep characterization of gut microbiota. *npj Biofilms and Microbiomes*, 9(1):4, January 2023.

[45] David Gómez-Varela, Feng Xian, Sabrina Grundtner, Julia Regina Sondermann, Giacomo Carta, and Manuela Schmidt. Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Frontiers in Microbiology*, 14:1258703, October 2023.

[46] Harriett Fuller, Yiwen Zhu, Jayna Nicholas, Haley A. Chatelaine, Emily M. Drzymalla, Afrand K. Sarvestani, Sachelly Julián-Serrano, Usman A. Tahir, Nasa Sinnott-Armstrong, Laura M. Raffield, Ali Rahnavard, Xinwei Hua, Katherine H. Shutta, and Burcu F. Darst. Metabolomic epidemiology offers insights into disease aetiology. *Nature Metabolism*, 5(10):1656–1672, October 2023.

[47] 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.

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

[49] Fernando Meyer, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, Mohammed Alser, Dmitry Antipov, Francesco Beghini, Denis

Bertrand, Jaqueline J. Brito, C. Titus Brown, Jan Buchmann, Aydin Buluç, Bo Chen, Rayan Chikhi, Philip T. L. C. Clausen, Alexandru Cristian, Piotr Wojciech Dabrowski, Aaron E. Darling, Rob Egan, Eleazar Eskin, Evangelos Georganas, Eugene Goltsman, Melissa A. Gray, Lars Hestbjerg Hansen, Steven Hofmeyr, Pingqin Huang, Luiz Irber, Huijue Jia, Tue Sparholt Jørgensen, Silas D. Kieser, Terje Klemetsen, Axel Kola, Mikhail Kolmogorov, Anton Korobeynikov, Jason Kwan, Nathan LaPierre, Claire Lemaitre, Chenhao Li, Antoine Limasset, Fabio Malcher-Miranda, Serghei Mangul, Vanessa R. Marcelino, Camille Marchet, Pierre Marijon, Dmitry Meleshko, Daniel R. Mende, Alessio Milanese, Niranjan Nagarajan, Jakob Nissen, Sergey Nurk, Leonid Oliker, Lucas Paoli, Pierre Peterlongo, Vitor C. Piro, Jacob S. Porter, Simon Rasmussen, Evan R. Rees, Knut Reinert, Bernhard Renard, Espen Mikal Robertsen, Gail L. Rosen, Hans-Joachim Ruscheweyh, Varuni Sarwal, Nicola Segata, Enrico Seiler, Lizhen Shi, Fengzhu Sun, Shinichi Sunagawa, Søren Johannes Sørensen, Ashleigh Thomas, Chengxuan Tong, Mirko Trajkovski, Julien Tremblay, Gherman Uritskiy, Riccardo Vicedomini, Zhengyang Wang, Ziye Wang, Zhong Wang, Andrew Warren, Nils Peder Willassen, Katherine Yelick, Ronghui You, Georg Zeller, Zhengqiao Zhao, Shanfeng Zhu, Jie Zhu, Ruben Garrido-Oter, Petra Gastmeier, Stephane Hacquard, Susanne Häußler, Ariane Khaledi, Friederike Maechler, Fantin Mesny, Simona Radutoiu, Paul Schulze-Lefert, Nathiana Smit, Till Strowig, Andreas Bremges, Alexander Sczyrba, and Alice Carolyn McHardy. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods*, 19(4):429–440, April 2022.

[50] 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.

[51] Yu-Wei Wu, Blake A. Simmons, and Steven W. Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, February 2016.

[52] Dongwan D. Kang, Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, July 2019.

[53] 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.

[54] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Jan Van Den Bussche, and Victor Vianu, editors, *Database Theory — ICDT 2001*, volume 1973, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science.

[55] Taylor M. Royalty and Andrew D. Steen. Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes. *mSystems*, 4(5):e00384–19, October 2019.

[56] Eugene V. Koonin and Yuri I. Wolf. Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, 11(7):487–498, July 2010.

[57] 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.

[58] Mosè Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simão, and Evgeny M Zdobnov. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, September 2021.

[59] Alex Chklovski, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8):1203–1212, August 2023.

[60] Tom O. Delmont and A. Murat Eren. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, 4:e1839, March 2016.

[61] Oskar Hickl, Pedro Queirós, Paul Wilmes, Patrick May, and Anna Heintz-Buschart. *binny* : an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Briefings in Bioinformatics*, 23(6):bbac431, November 2022.

[62] Shaojun Pan, Chengkai Zhu, Xing-Ming Zhao, and Luis Pedro Coelho. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications*, 13(1):2326, April 2022.

[63] Cong-Cong Liu, Shan-Shan Dong, Jia-Bin Chen, Chen Wang, Pan Ning, Yan Guo, and Tie-Lin Yang. MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome*, 10(1):46, March 2022.

[64] Nguyen Quoc Khanh Le, Quang-Thai Ho, Trinh-Trung-Duong Nguyen, and Yu-Yen Ou. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Briefings in Bioinformatics*, 22(5):bbab005, September 2021.

[65] Stephen Woloszynek, Zhengqiao Zhao, Jian Chen, and Gail L. Rosen. 16S rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses. *PLOS Computational Biology*, 15(2):e1006721, February 2019.

[66] L. V. Alteio, F. Schulz, R. Seshadri, N. Varghese, W. Rodriguez-Reillo, E. Ryan, D. Goudeau, S. A. Eichorst, R. R. Malmstrom, R. M. Bowers, L. A. Katz, J. L. Blanchard, and T. Woyke.

Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems*, 5(2):e00768–19, April 2020.

[67] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. 2023. Publisher: arXiv Version Number: 1.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 2017. Publisher: arXiv Version Number: 7.

[69] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec

Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. 2023. Publisher: arXiv Version Number: 4.

[70] Francesco Asnicar, Andrew Maltez Thomas, Andrea Passerini, Levi Waldron, and Nicola Segata. Machine learning for microbiologists. *Nature Reviews Microbiology*, November 2023.

[71] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20(4):1125–1136, July 2019.

[72] Marnix H. Medema, Tristan De Rond, and Bradley S. Moore. Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics*, 22(9):553–571, September 2021.

[73] S. Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, January 2012.

[74] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, November 2012.

[75] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, January 2022.

[76] Philip Hugenholtz, Maria Chuvochina, Aharon Oren, Donovan H. Parks, and Rochelle M. Soo. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *The ISME Journal*, 15(7):1879–1892, July 2021.

[77] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[78] Anupam Gautam, Wenhuan Zeng, and Daniel H Huson. MeganServer: facilitating interactive access to metagenomic data on a server. *Bioinformatics*, 39(3):btad105, March 2023.

[79] Benjamin Buchfink, Chao Xie, and Daniel H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, January 2015.

[80] J. Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, April 2004.

[81] Yang Liu, Kira S. Makarova, Wen-Cong Huang, Yuri I. Wolf, Anastasia N. Nikolskaya, Xinxu Zhang, Mingwei Cai, Cui-Jing Zhang, Wei Xu, Zhuhua Luo, Lei Cheng, Eugene V. Koonin, and Meng Li. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature*, 593(7860):553–557, May 2021.

[82] Brian D. Ondov, Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20(1):232, December 2019.

[83] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1):257, November 2019.

[84] XiaoFei Zhao. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics*, 35(4):671–673, February 2019.

[85] Vitor C Piro, Temesgen H Dadi, Enrico Seiler, Knut Reinert, and Bernhard Y Renard. ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, 36(Supplement_1):i12–i20, July 2020.

[86] Hans-Joachim Ruscheweyh, Alessio Milanese, Lucas Paoli, Nicolai Karcher, Quentin Clayssen, Marisa Isabell Keller, Jakob Wirbel, Peer Bork, Daniel R. Mende, Georg Zeller, and Shinichi Sunagawa. Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome*, 10(1):212, December 2022.

[87] Aitor Blanco-Míguez, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, Leonard Dubois, Kun D. Huang, Andrew Maltez Thomas, William A. Nickols, Gianmarco Piccinno, Elisa Piperni, Michal Punčochář, Mireia Valles-Colomer, Adrian Tett, Francesca Giordano, Richard Davies, Jonathan Wolf, Sarah E. Berry, Tim D. Spector, Eric A. Franzosa, Edoardo Pasolli, Francesco Asnicar, Curtis Huttenhower, and Nicola Segata. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, 41(11):1633–1644, November 2023.

[88] Qiaoxing Liang, Paul W Bible, Yu Liu, Bin Zou, and Lai Wei. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics*, 2(1):lqaa009, March 2020.

[89] Florian Mock, Fleming Kretschmer, Anton Kriese, Sebastian Böcker, and Manja Marz. Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks. *Proceedings of the National Academy of Sciences*, 119(35):e2122636119, August 2022.

[90] Marie Touchon and Eduardo P.C. Rocha. Coevolution of the Organization and Structure of Prokaryotic Genomes. *Cold Spring Harbor Perspectives in Biology*, 8(1):a018168, January 2016.

[91] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M Yunes, Ameet S Talwalkar, Susanna Repo, Michael L Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W A Buchan, Kevin Bryson, David T Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaßner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas A Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N Wass, Michael J E Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis A I Kourmpetis, Aalt D J Van Dijk, Cajo J F Ter Braak, Yuanpeng Zhou, Qingtian Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C Babbitt, Steven E Brenner, Christine Orengo, Burkhard Rost, Sean D Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, March 2013.

[92] Abigail L. LaBella, Dana A. Opulente, Jacob L. Steenwyk, Chris Todd Hittinger, and Antonis Rokas. Variation and selection on codon usage bias across an entire subphylum. *PLOS Genetics*, 15(7):e1008304, July 2019.

[93] Doug Hyatt, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, March 2010.

[94] Alexandre Lomsadze, Karl Gemayel, Shiyuyun Tang, and Mark Borodovsky. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Research*, 28(7):1079–1089, July 2018.

[95] Felix Van Der Jeugt, Peter Dawyndt, and Bart Mesuere. FragGeneScanRs: faster gene prediction for short reads. *BMC Bioinformatics*, 23(1):198, December 2022.

[96] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W. Ussery. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108, May 2007.

[97] D. Laslett. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16, January 2004.

[98] Eric P. Nawrocki and Sean R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, November 2013.

[99] Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, and Alex Bateman. InterPro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, January 2023.

[100] Ana Hernández-Plaza, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, Thomas Rattei, Ivica Letunic, Lars J Jensen, Peer Bork, Christian von Mering, and Jaime Huerta-Cepas. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, 51(D1):D389–D394, January 2023.

[101] The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, Paul Denny, Tunca Dogan, ThankGod Ebenezer, Jun Fan, Penelope Garmiri, Leonardo Jose Da Costa Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Marija Lugaric, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Alok Mishra, Katie Moulang, Andrew Nightingale, Sangya Pundir, Guoying Qi, Shriya Raj, Pedro Raposo, Daniel L Rice, Rabie Saidi, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard De Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Venkatesh Muthukrishnan, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess,

Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, and Jian Zhang. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023.

[102] 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.

[103] Álvaro Rodríguez Del Río, Joaquín Giner-Lamia, Carlos P. Cantalapiedra, Jorge Botas, Ziqi Deng, Ana Hernández-Plaza, Martí Munar-Palmer, Saray Santamaría-Hernando, José J. Rodríguez-Herva, Hans-Joachim Ruscheweyh, Lucas Paoli, Thomas S. B. Schmidt, Shinichi Sunagawa, Peer Bork, Emilia López-Solanilla, Luis Pedro Coelho, and Jaime Huerta-Cepas. Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature*, December 2023.

[104] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[105] Pedro Queirós, Francesco Delogu, Oskar Hickl, Patrick May, and Paul Wilmes. Mantis: flexible and consensus-driven genome annotation. *GigaScience*, 10(6):giab042, June 2021.

[106] Andrew L. Kau, Philip P. Ahern, Nicholas W. Griffin, Andrew L. Goodman, and Jeffrey I. Gordon. Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351):327–336, June 2011.

[107] Karen Delbaere, Inez Roegiers, Auriane Bron, Claude Durif, Tom Van de Wiele, Stéphanie Blanquet-Diot, and Ludovica Marinelli. The small intestine: dining table of host–microbiota meetings. *FEMS Microbiology Reviews*, 47(3):fuad022, May 2023.

[108] J Magarian Blander, Randy S Longman, Iliyan D Iliev, Gregory F Sonnenberg, and David Artis. Regulation of inflammation by microbiota interactions with the host. *Nature Immunology*, 18(8):851–860, August 2017.

[109] June L. Round and Sarkis K. Mazmanian. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5):313–323, May 2009.

[110] Valentina Tremaroli and Fredrik Bäckhed. Functional interactions between the gut microbiota and host metabolism. *Nature*, 489(7415):242–249, September 2012.

[111] Kaitlyn Oliphant and Emma Allen-Vercoe. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome*, 7(1):91, December 2019.

[112] Erwin G Zoetendal, Jeroen Raes, Bartholomeus Van Den Bogert, Manimozhiyan Arumugam, Carien Cgm Booijink, Freddy J Troost, Peer Bork, Michiel Wels, Willem M De Vos, and Michiel Kleerebezem. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *The ISME Journal*, 6(7):1415–1426, July 2012.

[113] Stephanie C. Ganal-Vonarburg, Mathias W. Hornef, and Andrew J. Macpherson. Microbial–host molecular exchange and its functional consequences in early mammalian life. *Science*, 368(6491):604–607, May 2020.

[114] Lluis Quintana-Murci. Human Immunology through the Lens of Evolutionary Genetics. *Cell*, 177(1):184–199, March 2019.

[115] Camille Martin-Gallausiaux, Ludovica Marinelli, Hervé M. Blottière, Pierre Larraufie, and Nicolas Lapaque. SCFA: mechanisms and functional importance in the gut. *Proceedings of the Nutrition Society*, 80(1):37–49, February 2021.

[116] N. Takahashi. Oral Microbiome Metabolism: From "Who Are They?" to "What Are They Doing?". *Journal of Dental Research*, 94(12):1628–1637, December 2015.

[117] Harry J. Flint, Edward A. Bayer, Marco T. Rincon, Raphael Lamed, and Bryan A. White. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology*, 6(2):121–131, February 2008.

[118] Abdessamad El Kaoutari, Fabrice Armougom, Jeffrey I. Gordon, Didier Raoult, and Bernard Henrissat. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7):497–504, July 2013.

[119] Douglas J. Morrison and Tom Preston. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*, 7(3):189–200, May 2016.

[120] Dallas R. Donohoe, Nikhil Garge, Xinxin Zhang, Wei Sun, Thomas M. O'Connell, Maureen K. Bunger, and Scott J. Bultman. The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metabolism*, 13(5):517–526, May 2011.

[121] H. M. Hamer, D. Jonkers, K. Venema, S. Vanhoutvin, F. J. Troost, and R.-J. Brummer. Review article: the role of butyrate on colonic function. *Alimentary Pharmacology & Therapeutics*, 27(2):104–119, January 2008.

[122] Roderick H. Dashwood, Melinda C. Myzak, and Emily Ho. Dietary HDAC inhibitors: time to rethink weak ligands in cancer chemoprevention? *Carcinogenesis*, 27(2):344–349, February 2006.

[123] GuoYan Wang, SenLin Qin, Lei Chen, HuiJun Geng, YiNing Zheng, Chao Xia, JunHu Yao, and Lu Deng. Butyrate dictates ferroptosis sensitivity through FFAR2-mTOR signaling. *Cell Death & Disease*, 14(4):292, April 2023.

[124] Petra Louis, Georgina L. Hold, and Harry J. Flint. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, 12(10):661–672, October 2014.

[125] Caleb J. Kelly, Leon Zheng, Eric L. Campbell, Bejan Saeedi, Carsten C. Scholz, Amanda J. Bayless, Kelly E. Wilson, Louise E. Glover, Douglas J. Kominsky, Aaron Magnuson, Tiffany L. Weir, Stefan F. Ehrentraut, Christina Pickel, Kristine A. Kuhn, Jordi M. Lanis, Vu Nguyen, Cormac T. Taylor, and Sean P. Colgan. Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function. *Cell Host & Microbe*, 17(5):662–671, May 2015.

[126] Leon Zheng, Caleb J. Kelly, Kayla D. Battista, Rachel Schaefer, Jordi M. Lanis, Erica E. Alexeev, Ruth X. Wang, Joseph C. Onyiah, Douglas J. Kominsky, and Sean P. Colgan. Microbial-Derived Butyrate Promotes Epithelial Barrier Function through IL-10 Receptor–Dependent Repression of Claudin-2. *The Journal of Immunology*, 199(8):2976–2984, October 2017.

[127] Wenchao Wei, Chi Chun Wong, Zhongjun Jia, Weixin Liu, Changan Liu, Fenfen Ji, Yasi Pan, Feixue Wang, Guoping Wang, Liuyang Zhao, Eagle S. H. Chu, Xiang Zhang, Joseph J. Y. Sung, and Jun Yu. Parabacteroides distasonis uses dietary inulin to suppress NASH via its metabolite pentadecanoic acid. *Nature Microbiology*, 8(8):1534–1548, June 2023.

[128] Luying Peng, Zhong-Rong Li, Robert S. Green, Ian R. Holzmanr, and Jing Lin. Butyrate Enhances the Intestinal Barrier by Facilitating Tight Junction Assembly via Activation of AMP-Activated Protein Kinase in Caco-2 Cell Monolayers. *The Journal of Nutrition*, 139(9):1619–1625, September 2009.

[129] Scott J. Bultman. Bacterial butyrate prevents atherosclerosis. *Nature Microbiology*, 3(12):1332–1333, November 2018.

[130] Kazuyuki Kasahara, Kimberly A. Krautkramer, Elin Org, Kymberleigh A. Romano, Robert L. Kerby, Eugenio I. Vivas, Margarete Mehrabian, John M. Denu, Fredrik Bäckhed, Aldons J. Lusis, and Federico E. Rey. Interactions between Roseburia intestinalis and diet modulate atherogenesis in a murine model. *Nature Microbiology*, 3(12):1461–1471, November 2018.

[131] Ana Nogal, Ana M. Valdes, and Cristina Menni. The role of short-chain fatty acids in the interplay between gut microbiota and diet in cardio-metabolic health. *Gut Microbes*, 13(1):1897212, January 2021.

[132] Nagendra Singh, Ashish Gurav, Sathish Sivaprakasam, Evan Brady, Ravi Padia, Huidong Shi, Muthusamy Thangaraju, Puttur D. Prasad, Santhakumar Manicassamy, David H. Munn, Jeffrey R. Lee, Stefan Offermanns, and Vadivel Ganapathy. Activation of Gpr109a, Receptor for Niacin and the Commensal Metabolite Butyrate, Suppresses Colonic Inflammation and Carcinogenesis. *Immunity*, 40(1):128–139, January 2014.

[133] Ziying Zhang, Haosheng Tang, Peng Chen, Hui Xie, and Yongguang Tao. Demystifying the manipulation of host immunity, metabolism, and extraintestinal tumors by the gut microbiome. *Signal Transduction and Targeted Therapy*, 4(1):41, October 2019.

[134] Izhak Levi, Michael Gurevich, Gal Perlman, David Magalashvili, Shay Menascu, Noam Bar, Anastasia Godneva, Liron Zahavi, Danyel Chermon, Noa Kosower, Bat Chen Wolf, Gal Malka, Maya Lotan-Pompan, Adina Weinberger, Erez Yirmiya, Daphna Rothschild, Sigal Leviatan, Avishag Tsur, Maria Didkin, Sapir Dreyer, Hen Eizikovitz, Yamit Titngi, Sue Mayost, Polina Sonis, Mark Dolev, Yael Stern, Anat Achiron, and Eran Segal. Potential role of indolelactate and butyrate in multiple sclerosis revealed by integrated microbiome-metabolome analysis. *Cell Reports Medicine*, 2(4):100246, April 2021.

[135] Pamela V. Chang, Liming Hao, Stefan Offermanns, and Ruslan Medzhitov. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proceedings of the National Academy of Sciences*, 111(6):2247–2252, February 2014.

[136] Ankita Sarkar, Priya Mitra, Abhishake Lahiri, Tanusree Das, Jit Sarkar, Sandip Paul, and Partha Chakrabarti. Butyrate limits inflammatory macrophage niche in NASH. *Cell Death & Disease*, 14(5):332, May 2023.

[137] Maik Luu, Katharina Weigand, Fatana Wedi, Carina Breidenbend, Hanna Leister, Sabine Pautz, Till Adhikary, and Alexander Visekruna. Regulation of the effector function of CD8+ T cells by gut microbiota-derived metabolite butyrate. *Scientific Reports*, 8(1):14430, September 2018.

[138] Jian Ji, Dingming Shu, Mingzhu Zheng, Jie Wang, Chenglong Luo, Yan Wang, Fuyou Guo, Xian Zou, Xiaohui Lv, Ying Li, Tianfei Liu, and Hao Qu. Microbial metabolite butyrate facilitates M2 macrophage polarization and function. *Scientific Reports*, 6(1):24838, April 2016.

[139] Lesley Hoyles, Tom Snelling, Umm-Kulthum Umlai, Jeremy K. Nicholson, Simon R. Carding, Robert C. Glen, and Simon McArthur. Microbiome–host systems interactions: protective effects of propionate upon the blood–brain barrier. *Microbiome*, 6(1):55, December 2018.

[140] Simone Di Giovanni, Elisabeth Serger, Jessica Chadwick, Lucia Luengo, Guiping Kong, Luming Zhou, Greg Crawford, Matt Danzi, Antonis Myridakis, Alexander Brandis, Adesola Bello, Francesco De Virgiliis, Marc-Emmanuel Dumas, Jessica Strid, and Dylan Dodd. The intermittent fasting-dependent gut microbial metabolite indole-3 propionate promotes nerve regeneration and recovery after injury. preprint, In Review, December 2020.

[141] Jianlong Yan, Yanbin Pan, Wenming Shao, Caiping Wang, Rongning Wang, Yaqiong He, Min Zhang, Yongshun Wang, Tangzhiming Li, Zhefeng Wang, Wenxing Liu, Zhenmin Wang, Xin Sun, and Shaohong Dong. Beneficial effect of the short-chain fatty acid propionate on vascular calcification through intestinal microbiota remodelling. *Microbiome*, 10(1):195, November 2022.

[142] Andrew J. Brown, Susan M. Goldsworthy, Ashley A. Barnes, Michelle M. Eilert, Lili Tcheang, Dion Daniels, Alison I. Muir, Mark J. Wigglesworth, Ian Kinghorn, Neil J. Fraser, Nicholas B. Pike, Jay C. Strum, Klaudia M. Steplewski, Paul R. Murdock, Julie C. Holder, Fiona H.

Marshall, Philip G. Szekeres, Shelagh Wilson, Diane M. Ignar, Steve M. Foord, Alan Wise, and Simon J. Dowell. The Orphan G Protein-coupled Receptors GPR41 and GPR43 Are Activated by Propionate and Other Short Chain Carboxylic Acids. *Journal of Biological Chemistry*, 278(13):11312–11319, March 2003.

[143] Trond Ulven. Short-chain free fatty acid receptors FFA2/GPR43 and FFA3/GPR41 as new potential therapeutic targets. *Frontiers in Endocrinology*, 3, 2012.

[144] Meng Li, Betty C.A.M. Van Esch, Gerry T.M. Wagenaar, Johan Garssen, Gert Folkerts, and Paul A.J. Henricks. Pro- and anti-inflammatory effects of short chain fatty acids on immune and endothelial cells. *European Journal of Pharmacology*, 831:52–59, July 2018.

[145] E. N. Bergman. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiological Reviews*, 70(2):567–590, April 1990.

[146] Johanne G. Bloemen, Koen Venema, Marcel C. Van De Poll, Steven W. Olde Damink, Wim A. Buurman, and Cornelis H. Dejong. Short chain fatty acids exchange across the gut and liver in humans measured at surgery. *Clinical Nutrition*, 28(6):657–661, December 2009.

[147] Hiromi Yamashita, Takao Kaneyuki, and Kunio Tagawa. Production of acetate in the liver and its utilization in peripheral tissues. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1532(1-2):79–87, May 2001.

[148] C. D. Seufert, W. Mewes, and H. D. Soeling. Effect of long-term starvation on acetate and ketone body metabolism in obese patients. *European Journal of Clinical Investigation*, 14(2):163–170, April 1984.

[149] John R. Moffett, Narayanan Puthillathu, Ranjini Vengilote, Diane M. Jaworski, and Aryan M. Namboodiri. Acetate Revisited: A Key Biomolecule at the Nexus of Metabolism, Epigenetics, and Oncogenesis – Part 2: Acetate and ACSS2 in Health and Disease. *Frontiers in Physiology*, 11:580171, November 2020.

[150] Amber Luong, Voe C. Hannah, Michael S. Brown, and Joseph L. Goldstein. Molecular Characterization of Human Acetyl-CoA Synthetase, an Enzyme Regulated by Sterol Regulatory Element-binding Proteins. *Journal of Biological Chemistry*, 275(34):26458–26466, August 2000.

[151] Zhiguang Huang, Menglu Zhang, Abigail A. Plec, Sandi Jo Estill, Ling Cai, Joyce J. Repa, Steven L. McKnight, and Benjamin P. Tu. ACSS2 promotes systemic fat storage and utilization through selective regulation of genes involved in lipid metabolism. *Proceedings of the National Academy of Sciences*, 115(40), October 2018.

[152] Gary Frost, Michelle L. Sleeth, Meliz Sahuri-Arisoylu, Blanca Lizarbe, Sebastian Cerdan, Leigh Brody, Jelena Anastasovska, Samar Ghourab, Mohammed Hankir, Shuai Zhang, David Carling, Jonathan R. Swann, Glenn Gibson, Alexander Viardot, Douglas Morrison, E Louise Thomas, and Jimmy D. Bell. The short-chain fatty acid acetate reduces appetite via a central homeostatic mechanism. *Nature Communications*, 5(1):3611, April 2014.

[153] Patrick M. Smith, Michael R. Howitt, Nicolai Panikov, Monia Michaud, Carey Ann Gallini, Mohammad Bohlooly-Y, Jonathan N. Glickman, and Wendy S. Garrett. The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic T $_{\text{reg}}$ Cell Homeostasis. *Science*, 341(6145):569–573, August 2013.

[154] Saioa Márquez, José Javier Fernández, Cristina Mancebo, Carmen Herrero-Sánchez, Sara Alonso, Tito A. Sandoval, Macarena Rodríguez Prados, Juan R. Cubillos-Ruiz, Olimpio Montero, Nieves Fernández, and Mariano Sánchez Crespo. Tricarboxylic Acid Cycle Activity and Remodeling of Glycerophosphocholine Lipids Support Cytokine Induction in Response to Fungal Patterns. *Cell Reports*, 27(2):525–536.e4, April 2019.

[155] Robert E. Steinert, Yuan-Kun Lee, and Wilbert Sybesma. Vitamins for the Gut Microbiome. *Trends in Molecular Medicine*, 26(2):137–140, February 2020.

[156] Hamid M Said and Zainab M Mohammed. Intestinal absorption of water-soluble vitamins: an update:. *Current Opinion in Gastroenterology*, 22(2):140–146, March 2006.

[157] João Carlos Gomes-Neto and June L. Round. Gut microbiota: a new way to take your vitamins. *Nature Reviews Gastroenterology & Hepatology*, 15(9):521–522, September 2018.

[158] John Conly and K E Stein. The absorption and bioactivity of bacterially synthesized menaquinones. *Clinical and Investigative Medicine*, 1993.

[159] J Stenflo and J W Suttie. Vitamin K-Dependent Formation of -Carboxyglutamic Acid. *Annual Review of Biochemistry*, 46(1):157–172, June 1977.

[160] Martin J. Shearer and Toshio Okano. Key Pathways and Regulators of Vitamin K Function and Intermediary Metabolism. *Annual Review of Nutrition*, 38(1):127–151, August 2018.

[161] J. W. Suttie. The Importance of Menaquinones in Human Nutrition. *Annual Review of Nutrition*, 15(1):399–417, July 1995.

[162] Fresia Fernandez and Matthew D. Collins. Vitamin K composition of anaerobic gut bacteria. *FEMS Microbiology Letters*, 41(2):175–180, April 1987.

[163] D. B. McCormick. Two interconnected B vitamins: riboflavin and pyridoxine. *Physiological Reviews*, 69(4):1170–1198, October 1989.

[164] Victor A. Najjar. THE BIOSYNTHESIS OF RIBOFLAVIN IN MAN. *Journal of the American Medical Association*, 126(6):357, October 1944.

[165] Zainab M. Said, Veedamali S. Subramanian, Nosratola D. Vaziri, and Hamid M. Said. Pyridoxine uptake by colonocytes: a specific and regulated carrier-mediated process. *American Journal of Physiology-Cell Physiology*, 294(5):C1192–C1197, May 2008.

[166] Gregory S. Ducker and Joshua D. Rabinowitz. One-Carbon Metabolism in Health and Disease. *Cell Metabolism*, 25(1):27–42, January 2017.

[167] P. K. Dudeja, S. A. Torania, and H. M. Said. Evidence for the existence of a carrier-mediated folate uptake mechanism in human colonic luminal membranes. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 272(6):G1408–G1415, June 1997.

[168] E Camilo, J Zimmerman, Jb Mason, B Golner, R Russell, J Selhub, and Ih Rosenberg. Folate synthesized by bacteria in the human upper small intestine is assimilated by the host. *Gastroenterology*, 110(4):991–998, April 1996.

[169] Alanna Lakoff, Zia Fazili, Susanne Aufreiter, Christine M Pfeiffer, Bairbie Connolly, Jesse F Gregory, Paul B Pencharz, and Deborah L O'Connor. Folate is absorbed across the human colon: evidence by using enteric-coated caplets containing 13C-labeled [6S]-5-formyltetrahydrofolate. *The American Journal of Clinical Nutrition*, 100(5):1278–1286, November 2014.

[170] Salih J. Wakil and David M. Gibson. Studies on the mechanism of fatty acid synthesis VIII. The participation of protein-bound biotin in the biosynthesis of fatty acids. *Biochimica et Biophysica Acta*, 41(1):122–129, June 1960.

[171] F. Lynen, J. Knappe, E. Lorch, G. Jütting, and E. Ringelmann. Die biochemische Funktion des Biotins. *Angewandte Chemie*, 71(15-16):481–486, August 1959.

[172] Joel Moss and M. Daniel Lane. The Biotin-Dependent Enzymes. In Alton Meister, editor, *Advances in Enzymology - and Related Areas of Molecular Biology*, volume 35, pages 321–442. Wiley, 1 edition, January 1971.

[173] Hamid M. Said, Alvaro Ortiz, Eric McCloud, David Dyer, Mary Pat Moyer, and Stanley Rubin. Biotin uptake by human colonic epithelial NCM460 cells: a carrier-mediated process shared with pantothenic acid. *American Journal of Physiology-Cell Physiology*, 275(5):C1365–C1371, November 1998.

[174] Jack C. Reidling, Svetlana M. Nabokina, and Hamid M. Said. Molecular mechanisms involved in the adaptive regulation of human intestinal biotin uptake: a study of the hSMVT system. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 292(1):G275–G281, January 2007.

[175] Sachiko Iinuma. SYNTHESIS OF RIBOFLAVIN BY INTESTINAL BACTERIA. *THE JOURNAL OF VITAMINOLOGY*, 1(2):6–13, 1955.

[176] Hamid M. Said, Alvaro Ortiz, Mary Pat Moyer, and Norimoto Yanagawa. Riboflavin uptake by human-derived colonic epithelial NCM460 cells. *American Journal of Physiology-Cell Physiology*, 278(2):C270–C276, February 2000.

[177] Van T. Pham, Sophie Fehlbaum, Nicole Seifert, Nathalie Richard, Maaike J. Bruins, Wilbert Sybesma, Ateequr Rehman, and Robert E. Steinert. Effects of colon-targeted vitamins on the composition and metabolic activity of the human gut microbiome– a pilot study. *Gut Microbes*, 13(1):1875774, January 2021.

[178] Margaret F. Romine, Dmitry A. Rodionov, Yukari Maezato, Lindsey N. Anderson, Premchendar Nandhikonda, Irina A. Rodionova, Alexandre Carre, Xiaoqing Li, Chengdong Xu, Therese R. W. Clauss, Young-Mo Kim, Thomas O. Metz, and Aaron T. Wright. Elucidation of roles for vitamin B $_{12}$ in regulation of folate, ubiquinone, and methionine metabolism. *Proceedings of the National Academy of Sciences*, 114(7), February 2017.

[179] Olga M. Sokolovskaya, Amanda N. Shelton, and Michiko E. Taga. Sharing vitamins: Cobamides unveil microbial interactions. *Science*, 369(6499):eaba0165, July 2020.

[180] Les Dethlefsen, Margaret McFall-Ngai, and David A. Relman. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449(7164):811–818, October 2007.

[181] Ruth E. Ley, Micah Hamady, Catherine Lozupone, Peter J. Turnbaugh, Rob Roy Ramey, J. Stephen Bircher, Michael L. Schlegel, Tammy A. Tucker, Mark D. Schrenzel, Rob Knight, and Jeffrey I. Gordon. Evolution of Mammals and Their Gut Microbes. *Science*, 320(5883):1647–1651, June 2008.

[182] Qing Zhao and Charles O. Elson. Adaptive immune education by gut microbiota antigens. *Immunology*, 154(1):28–37, May 2018.

[183] Gustavo Caballero-Flores, Joseph M. Pickard, and Gabriel Núñez. Microbiota-mediated colonization resistance: mechanisms and regulation. *Nature Reviews Microbiology*, 21(6):347–360, June 2023.

[184] Richard J. Lamont, Hyun Koo, and George Hajishengallis. The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology*, 16(12):745–759, December 2018.

[185] Simon Heilbronner, Bernhard Krismer, Heike Brötz-Oesterhelt, and Andreas Peschel. The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*, 19(11):726–739, November 2021.

[186] Carlos Asensio, José C. Pérez-Díaz, Mary Carmen Martínez, and Fernando Baquero. A new family of low molecular weight antibiotics from enterobacteria. *Biochemical and Biophysical Research Communications*, 69(1):7–14, March 1976.

[187] Alan M. O'Neill, Teruaki Nakatsuji, Asumi Hayachi, Michael R. Williams, Robert H. Mills, David J. Gonzalez, and Richard L. Gallo. Identification of a Human Skin Commensal Bacterium that Selectively Kills Cutibacterium acnes. *Journal of Investigative Dermatology*, 140(8):1619–1628.e2, August 2020.

[188] Antonio Maldonado-Barragán, Belén Caballero-Guerrero, Virginia Martín, José Luis Ruiz-Barba, and Juan Miguel Rodríguez. Purification and genetic characterization of gassericin E, a novel co-culture inducible bacteriocin from Lactobacillus gasseri EV1461 isolated from the vagina of a healthy woman. *BMC Microbiology*, 16(1):37, December 2016.

[189] Katharina Bitschar, Birgit Sauer, Jule Focken, Hanna Dehmer, Sonja Moos, Martin Konnerth, Nadine A. Schilling, Stephanie Grond, Hubert Kalbacher, Florian C. Kurschus, Friedrich Götz, Bernhard Krismer, Andreas Peschel, and Birgit Schittek. Lugdunin amplifies innate immune responses in the skin in synergy with host- and microbiota-derived factors. *Nature Communications*, 10(1):2730, June 2019.

[190] Michael S. Gilmore, Don B. Clewell, Yasuyoshi Ike, and Nathan Shankar, editors. *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection*. Massachusetts Eye and Ear Infirmary, Boston, 2014.

[191] Sophie Tronnet, Pauline Floch, Laetitia Lucarelli, Deborah Gaillard, Patricia Martin, Matteo Serino, and Eric Oswald. The Genotoxin Colibactin Shapes Gut Microbiota in Mice. *mSphere*, 5(4):e00589–20, August 2020.

[192] Jun Hu, Libao Ma, Yangfan Nie, Jianwei Chen, Wenyong Zheng, Xinkai Wang, Chunlin Xie, Zilong Zheng, Zhichang Wang, Tao Yang, Min Shi, Lingli Chen, Qiliang Hou, Yaorong Niu, Xiaofan Xu, Yuhua Zhu, Yong Zhang, Hong Wei, and Xianghua Yan. A Microbiota-Derived Bacteriocin Targets the Host to Confer Diarrhea Resistance in Early-Weaned Piglets. *Cell Host & Microbe*, 24(6):817–832.e8, December 2018.

[193] Gaëlle Vassiliadis, Delphine Destoumieux-Garzón, Carine Lombard, Sylvie Rebuffat, and Jean Peduzzi. Isolation and Characterization of Two Members of the Siderophore-Microcin Family, Microcins M and H47. *Antimicrobial Agents and Chemotherapy*, 54(1):288–297, January 2010.

[194] Thomas Ulas, Sabine Pirr, Beate Fehlhaber, Marie S Bickes, Torsten G Loof, Thomas Vogl, Lara Mellinger, Anna S Heinemann, Johanna Burgmann, Jennifer Schöning, Sabine Schreek, Sandra Pfeifer, Friederike Reuner, Lena Völlger, Martin Stanulla, Maren Von Köckritz-Blickwede, Shirin Glander, Katarzyna Barczyk-Kahlert, Constantin S Von Kaisenberg, Judith Friesenhagen, Lena Fischer-Riepe, Stefanie Zenker, Joachim L Schultze, Johannes Roth, and Dorothee Viemann. S100-alarmin-induced innate immune programming protects newborn infants from sepsis. *Nature Immunology*, 18(6):622–632, June 2017.

[195] Deepshika Ramanan, Alvin Pratama, Yangyang Zhu, Olivia Venezia, Martina Sassone-Corsi, Kaitavjeet Chowdhary, Silvia Galván-Peña, Esen Sefik, Chrysothemis Brown, Adélaïde Gélineau, Diane Mathis, and Christophe Benoist. Regulatory T cells in the face of the intestinal microbiota. *Nature Reviews Immunology*, 23(11):749–762, November 2023.

[196] Catalina Cosovanu and Christian Neumann. The Many Functions of Foxp3+ Regulatory T Cells in the Intestine. *Frontiers in Immunology*, 11:600973, October 2020.

[197] B-H Yang, S Hagemann, P Mamareli, U Lauer, U Hoffmann, M Beckstette, L Föhse, I Prinz, J Pezoldt, S Suerbaum, T Sparwasser, A Hamann, S Floess, J Huehn, and M Lochner. Foxp3+ T cells expressing RORt represent a stable regulatory T-cell effector lineage with enhanced suppressive capacity during intestinal inflammation. *Mucosal Immunology*, 9(2):444–457, March 2016.

[198] Yuri P. Rubtsov, Jeffrey P. Rasmussen, Emil Y. Chi, Jason Fontenot, Luca Castelli, Xin Ye, Piper Treuting, Lisa Siewe, Axel Roers, William R. Henderson, Werner Muller, and Alexander Y. Rudensky. Regulatory T Cell-Derived Interleukin-10 Limits Inflammation at Environmental Interfaces. *Immunity*, 28(4):546–558, April 2008.

[199] Christian Neumann, Jonas Blume, Urmi Roy, Peggy P. Teh, Ajithkumar Vasanthakumar, Alexander Beller, Yang Liao, Frederik Heinrich, Teresita L. Arenzana, Jason A. Hackney, Celine Eidenschenk, Eric J. C. Gálvez, Christina Stehle, Gitta A. Heinz, Patrick Maschmeyer, Tom Sidwell, Yifang Hu, Derk Amsen, Chiara Romagnani, Hyun-Dong Chang, Andrey Kruglov, Mir-Farzin Mashreghi, Wei Shi, Till Strowig, Sascha Rutz, Axel Kallies, and Alexander Scheffold. c-Maf-dependent Treg cell control of intestinal TH17 cells and IgA establishes host–microbiota homeostasis. *Nature Immunology*, 20(4):471–481, April 2019.

[200] Chris Schiering, Thomas Krausgruber, Agnieszka Chomka, Anja Fröhlich, Krista Adelmann, Elizabeth A. Wohlfert, Johanna Pott, Thibault Griseri, Julia Bollrath, Ahmed N. Hegazy, Oliver J. Harrison, Benjamin M. J. Owens, Max Löhning, Yasmine Belkaid, Padraic G. Fallon, and Fiona Powrie. The alarmin IL-33 promotes regulatory T-cell function in the intestine. *Nature*, 513(7519):564–568, September 2014.

[201] Moshe Biton, Adam L. Haber, Noga Rogel, Grace Burgin, Semir Beyaz, Alexandra Schnell, Orr Ashenberg, Chien-Wen Su, Christopher Smillie, Karthik Shekhar, Zuojia Chen, Chuan Wu, Jose Ordovas-Montanes, David Alvarez, Rebecca H. Herbst, Mei Zhang, Itay Tirosh, Danielle Dionne, Lan T. Nguyen, Michael E. Xifaras, Alex K. Shalek, Ulrich H. Von Andrian, Daniel B. Graham, Orit Rozenblatt-Rosen, Hai Ning Shi, Vijay Kuchroo, Omer H. Yilmaz, Aviv Regev, and Ramnik J. Xavier. T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. *Cell*, 175(5):1307–1320.e22, November 2018.

[202] Saiyu Hang, Donggi Paik, Lina Yao, Eunha Kim, Jamma Trinath, Jingping Lu, Soyoung Ha, Brandon N. Nelson, Samantha P. Kelly, Lin Wu, Ye Zheng, Randy S. Longman, Fraydoon Rastinejad, A. Sloan Devlin, Michael R. Krout, Michael A. Fischbach, Dan R. Littman, and Jun R. Huh. Bile acid metabolites control TH17 and Treg cell differentiation. *Nature*, 576(7785):143–148, December 2019.

[203] Clarissa Campbell, Peter T. McKenney, Daniel Konstantinovsky, Olga I. Isaeva, Michail Schizas, Jacob Verter, Cheryl Mai, Wen-Bing Jin, Chun-Jun Guo, Sara Violante, Ruben J. Ramos, Justin R. Cross, Krishna Kadaveru, John Hambor, and Alexander Y. Rudensky. Bacterial metabolism of bile acids promotes generation of peripheral regulatory T cells. *Nature*, 581(7809):475–479, May 2020.

[204] Xinyang Song, Ximei Sun, Sungwhan F. Oh, Meng Wu, Yanbo Zhang, Wen Zheng, Naama Geva-Zatorsky, Ray Jupp, Diane Mathis, Christophe Benoist, and Dennis L. Kasper. Microbial bile acid metabolites modulate gut ROR+ regulatory T cell homeostasis. *Nature*, 577(7790):410–415, January 2020.

[205] Esen Sefik, Naama Geva-Zatorsky, Sungwhan Oh, Liza Konnikova, David Zemmour, Abigail Manson McGuire, Dalia Burzyn, Adriana Ortiz-Lopez, Mercedes Lobera, Jianfei Yang,

Shomir Ghosh, Ashlee Earl, Scott B. Snapper, Ray Jupp, Dennis Kasper, Diane Mathis, and Christophe Benoist. Individual intestinal symbionts induce a distinct population of ROR $^{+}$ regulatory T cells. *Science*, 349(6251):993–997, August 2015.

[206] Susan Jackson, Zina Moldoveanu, and Jiri Mestecky. Collection and Processing of Human Mucosal Secretions. In *Mucosal Immunology*, pages 2345–2353. Elsevier, 2015.

[207] A Ferguson, K A Humphreys, and N M Croft. Technical Report: results of immunological tests on faecal extracts are likely to be extremely misleading. *Clinical and Experimental Immunology*, 99(1):70–75, June 2008.

[208] N. Lycke, L. Eriksen, and J. Holmgren. Protection against Cholera Toxin after Oral Immunization is Thymus-Dependent and Associated with Intestinal Production of Neutralizing IgA Antitoxin. *Scandinavian Journal of Immunology*, 25(4):413–419, April 1987.

[209] Takashi Obata, Yoshiyuki Goto, Jun Kunisawa, Shintaro Sato, Mitsuo Sakamoto, Hiromi Setoyama, Takahiro Matsuki, Kazuhiko Nonaka, Naoko Shibata, Masashi Gohda, Yuki Kagiyama, Tomonori Nochi, Yoshikazu Yuki, Yoshiko Fukuyama, Akira Mukai, Shinichiro Shinzaki, Kohtaro Fujihashi, Chihiro Sasakawa, Hideki Iijima, Masatoshi Goto, Yoshinori Umesaki, Yoshimi Benno, and Hiroshi Kiyono. Indigenous opportunistic bacteria inhabit mammalian gut-associated lymphoid tissues and share a mucosal antibody-mediated symbiosis. *Proceedings of the National Academy of Sciences*, 107(16):7419–7424, April 2010.

[210] Jun Kunisawa and Hiroshi Kiyono. Alcaligenes is Commensal Bacteria Habituating in the Gut-Associated Lymphoid Tissue for the Regulation of Intestinal IgA Responses. *Frontiers in Immunology*, 3, 2012.

[211] Naoko Shibata, Jun Kunisawa, Koji Hosomi, Yukari Fujimoto, Keisuke Mizote, Naohiro Kitayama, Atsushi Shimoyama, Hitomi Mimuro, Shintaro Sato, Natsuko Kishishita, Ken J Ishii, Koichi Fukase, and Hiroshi Kiyono. Lymphoid tissue-resident Alcaligenes LPS induces IgA production without excessive inflammatory responses via weak TLR4 agonist activity. *Mucosal Immunology*, 11(3):693–702, May 2018.

[212] Andrea J. Wolf and David M. Underhill. Peptidoglycan recognition by the innate immune system. *Nature Reviews Immunology*, 18(4):243–254, April 2018.

[213] Hiutung Chu and Sarkis K Mazmanian. Innate immune recognition of the microbiota promotes host-microbial symbiosis. *Nature Immunology*, 14(7):668–675, July 2013.

[214] Thomas B Clarke, Kimberly M Davis, Elena S Lysenko, Alice Y Zhou, Yimin Yu, and Jeffrey N Weiser. Recognition of peptidoglycan from the microbiota by Nod1 enhances systemic innate immunity. *Nature Medicine*, 16(2):228–231, February 2010.

[215] Signe Altmäe, Jason M. Franasiak, and Reet Mändar. The seminal microbiome in health and disease. *Nature Reviews Urology*, 16(12):703–721, December 2019.

[216] Maayan Levy, Aleksandra A. Kolodziejczyk, Christoph A. Thaiss, and Eran Elinav. Dysbiosis and the immune system. *Nature Reviews Immunology*, 17(4):219–232, April 2017.

[217] Felix Sommer, Malte Christoph Rühlemann, Corinna Bang, Marc Höppner, Ateequr Rehman, Christoph Kaleta, Phillippe Schmitt-Kopplin, Astrid Dempfle, Stephan Weidinger, Eva Ellinghaus, Susanne Krauss-Etschmann, Dirk Schmidt-Arras, Konrad Aden, Dominik Schulte, David Ellinghaus, Stefan Schreiber, Andreas Tholey, Jan Rupp, Matthias Laudes, John F Baines, Philip Rosenstiel, and Andre Franke. Microbiomarkers in inflammatory bowel diseases: caveats come with caviar. *Gut*, 66(10):1734–1738, October 2017.

[218] Aonghus Lavelle and Harry Sokol. Gut microbiota-derived metabolites as key actors in inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology*, 17(4):223–237, April 2020.

[219] IBDMDB Investigators, 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.

[220] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe*, 15(3):382–392, March 2014.

[221] Eric A. Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J. Haiser, Stefan Reinker, Tommi Vatanen, A. Brantley Hall, Himel Mallick, Lauren J. McIver, Jenny S. Sauk, Robin G. Wilson, Betsy W. Stevens, Justin M. Scott, Kerry Pierce, Amy A. Deik, Kevin Bullock, Floris Imhann, Jeffrey A. Porter, Alexandra Zhernakova, Jingyuan Fu, Rinse K. Weersma, Cisca Wijmenga, Clary B. Clish, Hera Vlamakis, Curtis Huttenhower, and Ramnik J. Xavier. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology*, 4(2):293–305, December 2018.

[222] Henri Duboc, Sylvie Rajca, Dominique Rainteau, David Benarous, Marie-Anne Maubert, Elodie Quervain, Ginette Thomas, Véronique Barbu, Lydie Humbert, Guillaume Despras, Chantal Bridonneau, Fabien Dumetz, Jean-Pierre Grill, Joëlle Masliah, Laurent Beaugerie, Jacques Cosnes, Olivier Chazouillères, Raoul Poupon, Claude Wolf, Jean-Maurice Mallet, Philippe Langella, Germain Trugnan, Harry Sokol, and Philippe Seksik. Connecting dys-

biosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. *Gut*, 62(4):531–539, April 2013.

[223] Douwe F. De Wit, Nordin M. J. Hanssen, Koen Wortelboer, Hilde Herrema, Elena Rampanelli, and Max Nieuwdorp. Evidence for the contribution of the gut microbiome to obesity and its reversal. *Science Translational Medicine*, 15(723):eadg2773, November 2023.

[224] Mireia Valles-Colomer, Cristina Menni, Sarah E. Berry, Ana M. Valdes, Tim D. Spector, and Nicola Segata. Cardiometabolic health, diet and the gut microbiome: a meta-omics perspective. *Nature Medicine*, 29(3):551–561, March 2023.

[225] Yong Fan and Oluf Pedersen. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, 19(1):55–71, January 2021.

[226] Pari Mokhtari, Julie Metos, and Pon Velayutham Anandh Babu. Impact of type 1 diabetes on the composition and functional potential of gut microbiome in children and adolescents: possible mechanisms, current knowledge, and challenges. *Gut Microbes*, 13(1):1926841, January 2021.

[227] B. J. Kunath, O. Hickl, P. Queirós, C. Martin-Gallausiaux, L. A. Lebrun, R. Halder, C. C. Laczny, T. S. B. Schmidt, M. R. Hayward, D. Becher, A. Heintz-Buschart, C. De Beaufort, P. Bork, P. May, and P. Wilmes. Alterations of oral microbiota and impact on the gut microbiome in type 1 diabetes mellitus revealed by integrated multi-omic analyses. *Microbiome*, 10(1):243, December 2022.

[228] Thomas C Fung, Christine A Olson, and Elaine Y Hsiao. Interactions between the microbiota, immune and nervous systems in health and disease. *Nature Neuroscience*, 20(2):145–155, February 2017.

[229] Shokufeh Ghasemian Sorboni, Hanieh Shakeri Moghaddam, Reza Jafarzadeh-Esfehani, and Saman Soleimanpour. A Comprehensive Review on the Role of the Gut Microbiome in Human Neurological Disorders. *Clinical Microbiology Reviews*, 35(1):e00338–20, January 2022.

[230] Ai Huey Tan, Shen Yang Lim, and Anthony E. Lang. The microbiome–gut–brain axis in Parkinson disease — from basic research to the clinic. *Nature Reviews Neurology*, 18(8):476–495, August 2022.

[231] Oskar Hickl, Anna Heintz-Buschart, Anke Trautwein-Schult, Rajna Hercog, Peer Bork, Paul Wilmes, and Dörte Becher. Sample Preservation and Storage Significantly Impact Taxonomic and Functional Profiles in Metaproteomics Studies of the Human Gut Microbiome. *Microorganisms*, 7(9):367, September 2019.

[232] Peifeng Ji, Yanming Zhang, Jinfeng Wang, and Fangqing Zhao. MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nature Communications*, 8(1):14306, January 2017.

[233] Arunima Singh. Nanopores for sequencing proteins. *Nature Methods*, 20(12):1870–1870, December 2023.

[234] Matthew R. McIlvin and Mak A. Saito. Online Nanoflow Two-Dimension Comprehensive Active Modulation Reversed Phase–Reversed Phase Liquid Chromatography High-Resolution Mass Spectrometry for Metaproteomics of Environmental and Microbiome Samples. *Journal of Proteome Research*, 20(9):4589–4597, September 2021.

[235] Mingon Kang, Euiseong Ko, and Tesfaye B Mersha. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1):bbab454, January 2022.

[236] Dennis Svedberg, Rahel R. Winiger, Alexandra Berg, Himanshu Sharma, Christian Tellgren-Roth, Bettina A. Debrunner-Vossbrinck, Charles R. Vossbrinck, and Jonas Barandun. Functional annotation of a divergent genome using sequence and structure-based similarity. *BMC Genomics*, 25(1):6, January 2024.

[237] Michel Van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, May 2023.

[238] Feng Ju, Karin Beck, Xiaole Yin, Andreas Maccagnan, Christa S McArdell, Heinz P Singer, David R Johnson, Tong Zhang, and Helmut Bürgmann. Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes. *The ISME Journal*, 13(2):346–360, February 2019.

[239] Maria Muñoz-Benavent, Felix Hartkopf, Tim Van Den Bossche, Vitor C Piro, Carlos García-Ferris, Amparo Latorre, Bernhard Y Renard, and Thilo Muth. gNOMO: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms. *NAR Genomics and Bioinformatics*, 2(3):lqaa058, September 2020.

[240] 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, September 2020.

[241] Sung-Huan Yu, Jörg Vogel, and Konrad U Förstner. ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*, 7(9), September 2018.

[242] Felix Grünberger, Robert Reichelt, Boyke Bunk, Cathrin Spröer, Jörg Overmann, Reinhard Rachel, Dina Grohmann, and Winfried Hausner. Next Generation DNA-Seq and Differential RNA-Seq Allow Re-annotation of the Pyrococcus furiosus DSM 3638 Genome and Provide Insights Into Archaeal Antisense Transcription. *Frontiers in Microbiology*, 10:1603, July 2019.

[243] Daniel Ryan, Laura Jenniches, Sarah Reichardt, Lars Barquist, and Alexander J. Westermann. A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe Bacteroides thetaiotaomicron. *Nature Communications*, 11(1):3557, July 2020.

[244] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, November 2014.

[245] Antton Alberdi, Sandra B. Andersen, Morten T. Limborg, Robert R. Dunn, and M. Thomas P. Gilbert. Disentangling host–microbiota complexity through hologenomics. *Nature Reviews Genetics*, 23(5):281–297, May 2022.

[246] 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.

[247] Ashwin Chetty and Ran Blekhman. Multi-omic approaches for host-microbiome data integration. *Gut Microbes*, 16(1):2297860, December 2024.