UNIVERSITÉ DE GENÈVE FACULTÉ DES SCIENCES

Section de médecine fondamentale
Département de physiologie cellulaire et métabolisme
Département de Génétique et Évolution

Professeur Mirko Trajkovski
Professeur Evgeny M. Zdobnov

Computational approaches for a healthier microbiome

THÈSE

présentée aux Facultés de médecine et des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences en sciences de la vie,
mention Sciences biomédicales

par

Silas Daniel Kieser

de

Lenzburg (Aargau)

Thèse N° nnnn

GENÈVE

Nom de l'Atelier d'Impression

2021

Table of contents

Terms & Acronyms

Microbe	Microorganism, organisms that are very small. Most microbes are prokaryotes but also some small eukaryotes are counted as microbes[1].
Prokaryotes	Unicellular organisms that don't contain a nucleus. Bacteria and Archaea.
Eukaryotes	All organisms that contain a nucleus. All animals, plants, fungi, but also protists
Genome	All genetic material of a cell.
Contig	A continuous DNA sequence, often the result of a assembly of multiple reads
Sequencing	The process of reading DNA, often in fractions
Read	A (short) DNA sequence, the output of a sequencer
Assembly	The process of putting small DNA fragments together to create longer ones
16S sequencing	Sequencing (parts of) the 16S that allows the identification of microbes
16S	The RNA of the small ribosomal subunit in prokaryotes
Amplicon sequencing	I use is as synonym for 16S sequencing
Primers	Small DNA or used to initiate PCR amplification
PCR	Polymerase chain reaction, a standard labaratory method to amplify even small amounts of DNA fragments up to 10kbp in length. Invented in 1983.
Metagenome	Collection of all genomes from an environment. See Box.
Metagenomics	The study of metagenomes
MAG	Metagenome assembled genome
kbp	Killo base pairs = 1000 bp, unit for measurnign the lenght of DNA
CLR	Centered log ratio

[1] The definition is not very clear as discussed in "What Counts as a Microbe?" on asm.org

Introduction

> *What is the role, in the overall scheme of creation, of some of these little beings who are the agents of fermentation, the agents of putrefaction, of disorganization of everything that life has had on the surface of the globe? This role is immense, marvelous, really moving. Maybe one day, I will be given the opportunity to explain some of these results.*
>
> — Louis Pasteur[1]

Ex nihilo nihil fit (From nothing, nothing comes) is a fundamental philosophical principle. There might still be a debate in metaphysics about whether the universe could arise from nothing and what that *nothing* would be. In biology, the idea of spontaneous creation was eliminated from the discipline by Louis Pasteur in the middle of the 19th century. He not only showed that a closed, sterile system remains sterile but at the same time that microbes are all around us. He became one of the founders of Microbiology and postulated that germs are the cause of many diseases. The change in mentality led to the identification of many diseases' pathogens. Lives could be saved by simple measures such as hand sanitation, which was not a common practice for clinicians at that time.

Even if it was possible to see microbes since Antonie van Leeuwenhoek discovered the microscope, a large part of the microbial diversity was still hidden until 1950 when Robert. E. Hungate developed his technique to cultivate anaerobic microbes (Hungate, 1944). Even with the ability to culture anaerobic microorganisms, the vast majority of microbes went unnoticed. In 1985, it was estimated that less than 1% of the microorganisms found in an environmental sample could be cultured on plates. This fact came to be known as the "*the great plate count*

[1] René Vallery-Radot (1902). *The life of Pasteur*. New York: Phillips McClure, p. 142

anomaly" (Staley & Konopka, 1985) and spurred the interest in sequencing microbes directly from the environment.

1.1 The history of microbiome sequencing

> *[Sequencing] ... has been at the center of all my research since 1943, both because of its intrinsic fascination and my conviction that a knowledge of sequences could contribute much to our understanding of living matter*
>
> — Frederick Sanger[2]

In 1977, Frederick Sanger started to develop his sequencing method (Sanger et al., 1977), which kicked off a revolution in sequencing. Even before that, researchers studied microorganisms by sequencing short fragments of RNA (Heather & Chain, 2016). So determined Carl. R. Woese & George. E. Fox 20 years before the first microbial genome was sequenced, that living organisms consist not only of Eukaryotes[3] and bacteria but that there is a third domain: the Archaea (Woese & Fox, 1977). The discovery of a whole new domain of life was revolutionary for its time and was met with a lot of criticism (Goldenfeld & Pace, 2013). It showed that life evolved not in a linear fashion from simple to more complex life, but that there are deep branches in the *tree of life* going all the way down to its root. "*The 1977 paper is one of the most influential in microbiology and arguably, all of biology. It ranks with the works of Watson and Crick and Darwin, providing an evolutionary framework for the incredible diversity of the microbial world*" (Nair, 2012).

More practically, Carl. R. Woese & George. E. Fox showed convincingly that (microbial) life could be analyzed by sequencing. They differentiated the organisms based on the small ribosomal unit, whose RNA is highly abundant in cells. The corresponding gene, also known as the 16S gene (18S in Eukaryotes), is present in all living organisms. The functional constraints on the ribosome keep the sequence of the rRNA-gene comparable between all domains of life. The publication from 1977 consists mainly of one table that shows how the 16S

[2] Frederick Sanger, Biographical, NobelPrize.org

[3] Organisms that have a nucleus: animals, plants, fungi, protists...

(18S) sequence from representatives of the three domains of life are more similar within the domain than between the domains.

However, the sequencing of the small ribosomal unit still required the cultivation of the species under investigation, which was difficult, especially for the new domain of archaea. The solution to this problem came almost ten years later. During this time, the polymerase chain reaction (PCR) was invented, which allowed the amplification of small amounts of DNA, and the Sanger sequencing became widely available. In 1985, Norman R. Pace and co-workers developed a technique to rapidly sequence the 16S gene directly from the environment (Lane et al., 1985). They propose *universal primers* for amplification of the 16S gene. Norman Pece came to be named "The man who blew the door off the microbial world" (YONG, 2017), as his technique allowed for the first time to see the full diversity of microbes in a sample. Again 10 years later, the first time the human gut microbiome was sequenced (Wilson & Blitchington, 1996), see also box.

1.1.1 16S amplicon sequencing

Today amplicon sequencing of the 16S gene is a routine experiment: First, DNA is extracted from a sample, then PCR is used to amplify a part of the 16S gene. Not all regions of the 16S genes are equally conserved between taxa. Conserved regions are interspaced by 9 variable regions (Fig. 1.1). These variable regions belong to loops in the ribosome with little constraints and are therefore free to evolve. The *universal primers* target the flanking conserved regions around one or multiple variable regions, which then are sequenced.

As with almost everything in biology, there are no rules that apply without exception, and so not the *universal* primer sites of the 16S not universal. Different versions of primer(-mixtures) are proposed for different target microbiomes (For example Sim et al., 2012). Also, not all variable regions are suited to discriminating all taxa, and PCR bias can artificially increase specific species. Therefore, it is generally accepted that 16S amplicon sequencing, based on one or two variable regions, can classify organisms down to the genus level but is not suited to classify at species level robustly (See section 5.2.1 for an in-depth discussion). Despite all these drawbacks, 16S sequencing is the most used technique to analyze microbiomes at a low cost. Many of these limitations are addressed by

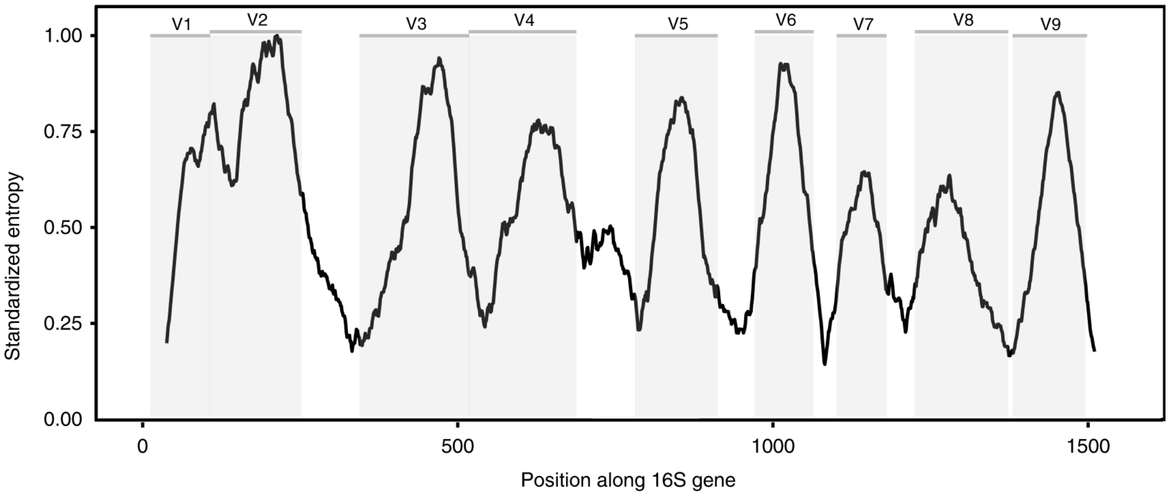


Fig. 1.1 Variability across the 16S gene based on the alignment of a single representative sequence for each known species present in the Greengenes database. Sequences were aligned against a single reference 16S gene for Escherichia coli K-12 MG1655. Gray panels depict variable regions defined by commonly used primer-binding sites.
Adapted from Johnson et al., 2019

long-read sequencing, which allows the sequencing of all nine variable regions (Karst et al., 2018).

The human microbiome

The ubiquity of microbial life was confirmed over and over again. It is difficult to imagine a place on earth that is not populated with microbial life, and it is even speculated that Mars is the home of microbes brought to earth. Interestingly, the human gut is among the most densely populated habitats for microorganisms in the world (Whitman et al., 1998). Also, the gut is densely covered with immune cells from within (Vijay-Kumar et al., 2014), indicating that the gut is an essential point in host-microbe interaction. Bacteria are the most numerous domain of life in the human gut microbiota (Guarner & Malagelada, 2003). Archaea, fungi, and protists belong to the microorganism living in the human gut. Not surprisingly, phages and other viruses outnumber the living organisms and play the critical role of predator in this environment (Fernández et al., 2018).

1.1.2 Metagenomics

Even with the improved techniques, the 16S rDNA sequencing allows only to identify the taxonomy of a prokaryote. If the microbe is not present in a database, one can match the sequence to the closest relative, but no more information can be gained from the small ribosomal unit. What can one do if one wants to know more about an unknown species of prokaryote? Ideally, one wants to culture the microbe, but this might not be easy, especially as we do not know in which medium we would need to cultivate the prokaryote.

Stein et al. faced the same challenge: They wanted to know more about an uncultured clade of marine Archaea. They knew from 16S marker surveys that this clade abundant in the surface water of the Hawaiian ocean. Stein et al. was the first to use a metagenomic sequencing approach. They extracted DNA fragments from the ocean water and cloned them into a library of E. coli. Clones that contained the 16S signature of the clade of interest were selected for sequencing. Sequencing was labor-intensive and involved digesting the fragment of interest with restriction enzymes and cloning sub-fragments into plasmids before sequencing the sub-fragments in steps of less than 1kpb in length. Using a technique called primer-walking, Stein et al. were able to, in the best case, rebuild a 40kb-fragment around the 16S gene.[4] Even though this fragment did not contain any genes with novel functions, their study laid the groundwork for a new era of environmental sequencing of uncultured species.

[4] It is worth noting the progress in sequencing technology. DNA fragments of 40kpb and longer can be sequenced today as a single molecule using a long-read sequencer (Dijk et al., 2018)

Origin of the term metagenome

The term metagenome was first used by Handelsman et al. in their work about the soil metagenome (Handelsman et al., 1998). They did not use sequencing but rather extracted long DNA fragments from the soil and cloned them into *E. coli* cells. Testing this *E. coli*, library for new natural products, accelerated the search and identification of biosynthetic gene clusters. As the DNA fragments did not come from specific microbes, Handelsman et al. introduced the term *metagenome* to describe the ensemble of genomes and biosynthetic machinery of the soil microflora. The term is sometimes used as a synonym for *microbiome*, as both describe the collection of all microbes in an environment.

Shotgun genomics

Primer walking is inherently slow as one has to finish sequencing one step to design a primer for the next. **shotgun sequencing** was invented in order to circumvent this constraint (Staden, 1979; Anderson, 1981). DNA is shredded into random (overlapping) fragments that are assembled with the aid of a computer. Because of computational limitations, shotgun metagenomics was initially limited to DNA fragments of 50kb. It allowed the sequencing of viruses, but to sequence larger genomes, the genome has to be split into fragments of this size, amplified in bacteria or yeast, before being sequenced with the shotgun approach. This hierarchical shotgun strategy was used for most of the genomes sequenced in the context of the human genome project.

The team of Craig Venter leveraged shotgun metagenomics to assemble whole genomes. They sequenced the first bacterial genome in 1995 (*Haemophilus Influenzae*, 1.8 Mbp, Fleischmann et al., 1995). They proposed to use *whole genome shotgun sequencing* for the human genome (J. Craig Venter, Smith, et al., 1996), but the proposition was not received. There was still doubt if the shotgun approach can scale to large genomes of eukaryotes with many repetitive sequences (J Craig Venter, 2006). The team of Craig venter sequenced the genome of *Drosophila* (Myers et al., 2000) and finally *Homo sapiens* (J. Craig Venter, Adams, et al., 2001) as part of a private initiative.

Shotgun meta-genomics

After having sequenced the human genome, Craig Venter went on to sequence the marine metagenome. The goal was to sequence samples from all the world's oceans, but already for the first samples, the Caldera assembler failed (J Craig Venter, 2004). Only a quarter of the reads could be assembled into well-covered contigs. The diversity of the microbes made it too difficult to assemble their genomes. If genomes cannot be assembled, it is still possible to perform analyses about the functional and taxonomic composition of a metagenome based on genes predicted on the contigs or the reads themselves. In this way, J Craig Venter estimated that they sequenced DNA from almost 2000 different species, including 148 types of bacteria never seen before.

Such gene-centric metagenome analyses became wildly popular at the beginning of large-scale metagenome studies (McMahon, 2015). Significant collaborative efforts were undertaken to construct reference gene-catalogs for specific microbiomes, for example, the mentioned *Global Ocean Sampling Expedition*, the Human microbiome project, and the MetaHIT (METAgenomics of the Human Intestinal Tract) project (Yooseph et al., 2007; Turnbaugh et al., 2007; Qin et al., 2010). These efforts uncovered millions of new genes for different metagenomes. Even if the fraction of annotatable genes is small, these catalogs enabled the comparative analysis of metagenomes based on function and even the inference of metabolic pathways (Tringe, 2005). Metagenomic tools developed during this time, e.g., HUMAnN2 (Abubucker et al., 2012) are still prevalent.

However, gene-centric annotations approaches have several limitations because they treat the whole microbiome as one entity: First, the quantification of functions is based on reads mapping to genes that are even more subject to variation than genomes, e.g., through ongoing genome duplication or gene multiplicity. Second, the taxonomic and the functional annotation are not linked. It is not easy to see which species is responsible for which function. Therefore, it becomes difficult to relate taxonomic changes to changes in functional abundances. Third, metabolic pathways are reconstructed for the whole metagenome instead of individual genomes, which obfuscate any metabolic mutualism or competition between different species by considering them as the same entity.

> *Genes are expressed within cells, not in a homogenized cytoplasmic soup. It matters a lot whether a complete metabolic pathway is found in a genome versus distributed across multiple distinct genomes.*
>
> — McMahon, 2015

A genome-centered approach to metagenomic data can overcome these limitations. In the following section, we will look at the technological and algorithmic advancements that made this change possible.

1.2 Genome-resolved metagenomics

1.2.1 Metagenome assembly

The assembly of metagenomes is still a challenge today. One of the significant advances in terms of algorithms was the introduction of graph-based assemblers. Before assembly, algorithms compared each read which each other, successively merged reads with the most significant overlap to creating a consensus sequence. Overlap-layout consensus (OLC) algorithms have difficulty finding a consensus if the reads have multiple overlaps due to repeated regions. Nor do they scale well to large sets of reads, as each read has to be compared to each other, making the computational burden increase exponentially with the number of reads.

> *Children like puzzles, and they usually assemble them by trying all possible pairs of pieces and putting together pieces that match. Biologists assemble genomes in a surprisingly similar way, the main difference being that the number of pieces is more significant.*
>
> — Pevzner et al., 2001

Myers, who helped to assemble the *Drosophila* and the human genome, saw the limitations of their assembly algorithm and proposed a new idea. The idea is to use a (de Bruijn) graph to represent the sequenced reads and framed the assembly as a problem to find the optimal path in this graph. The sequenced reads are split into short sub-sequences of a fixed size, the so-called k*-mers*. These represent the graph's nodes. Two k-mersare connected in the chart if they are adjacent in at least one sequencing read. The reads represent short paths connecting this node. The assembly consists of applying several graph-simplification algorithms to find paths that are as long as possible. In theory, the optimal path can be found in linear-time (Pevzner et al., 2001), but because the coverage is never homogenous, it is rarely possible to find a complete genome. The choice of the k-mer-lenght is crucial: short k-merscreate a well-connected graph, with the risk of having too many overlaps that make it challenging to resolve repeats. Long k-merscan better resolve repeats but decrease the chance of overlaps between the sequencing reads. Algorithms based on the de Bruijn graph reduce the computational costs drastically because the chart stores the sequencing data more efficiently. The graph's memory requirements can be even more reduced by removing sequencing errors before or during the building of the graph. Sequencing errors introduce false k-mernodes and lead to spurious branches. Graph-based assembly algorithms became popular after the human genome was sequenced, and (**next-generation sequencing**) came available that allows much cheaper and parallelized sequencing albeit with a limitation of the read-length.

Even though prokaryote genomes are relatively simple to assemble, the assembly of a complex mixture thereof, a metagenome, is challenging. Overlaps within a genome and between different genomes are formed at regions that are similar, such as the conserved genes for the ribosome. Many bacterial species in a microbiome are represented by a mixture of strains with a conserved core genome and additional variable regions (Kashtan et al., 2014). The coverage over the different genomes is very uneven due to the overlaps, amplification bias, and the difference in the abundance of the organisms in the original sample. It can become challenging to distinguish low-abundant strain variation from sequencing errors.

Modern metagenome assemblers such as MegaHit (Li et al., 2015) and MetaSPAdes (Nurk et al., 2017) deal with these challenges by different modifications to a standard-genome assembly algorithm. Both use multiple graphs with different k-merlengths to benefit from the advantages of both small and large k-mers. Smal k-mersallow assembling better regions with low coverage, whereas longer k-mersallow disentangling repeats and inter-genome overlaps. The graph for one k-merlength is used to build the next larger k-mergraph. The tools fall on different sides of the tradeoff between correcting sequencing errors and overwriting strain variation. MegaHit is sensitive to strain variation. There are no specific error-correction steps other than removing k-mersbelow a threshold from the graph. Discarded k-merscan be recovered if they fit in the assembly graph. MetaSpades, on the other hand, corrects the sequencing reads before the graph generation. The tool uses stringent graph simplification algorithms that ignore strain-specific features of rare strains to reconstruct a consensus backbone of a strain mixture. Interestingly, the aglomerating approach of metaSPAdes gives the best results as an independent benchmark of metagenome assemblers from 2017 showed(Vollmers et al., 2017). The tool produces the longest scaffolds, even from highly complex metagenomes. MetaSPAdes is often used for large-scale metagenome assembly projects unless the memory requirements are too high (Pasolli et al., n.d.; Almeida et al., 2019; Nayfach et al., 2019). In which case, often, MegaHit is used instead. MegaHit is the most efficient tool, according to the same benchmark, and often produces the largest assembly even if it is more fractionized (Vollmers et al., 2017). Both assemblers rarely produce complete genomes, and therefore, it is necessary to cluster the resulting contigs into bins that could be thought of as genomes, a process called *binning*.

1.2.2 Recovering genomes from metagenomes

The beginning of binning

Even before performant graph-based assemblers were available, researchers attempted to recover genomes from metagenomes. Tyson et al., were the first that succeeded. In their landmark paper (Tyson et al., 2004), they focus on a microbiome with relatively low diversity from an acid mine drainage. The microbes living in these harsh conditions (pH 0.83) oxidize iron and exacerbate the pollution of the mine outflows. The microbiome that consists mainly of different bacterial species grows as a biofilm. To understand the metabolic interaction between the different uncultured species, Tyson et al. attempted to reconstruct genomes directly from the metagenome sample.

To do so, they performed an unprecedentedly deep metagenomic sequencing (76.2 Mbp) that should cover the genomes up to 10 fold. The modified overlap-based assembly algorithm prioritizes continuity of the assembly over the accuracy, which is in line with the modifications in modern algorithms as described

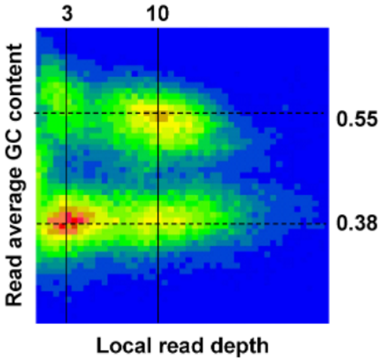


Fig. 1.2 Distribution of GC content versus local read depth (coverage) of a deeply sequenced metagenome from an acid mine drainage. Four peaks are visible at the intersections of the grid lines. Source: (Tyson et al., 2004)

above. The shift had its effect; over 85 % of the reads could be assembled into scaffolds longer than 2 kbp. Tyson et al. calculated the average GC-content (fraction of guanine + cytosine) and coverage for each contig. When they looked at their distribution, four peaks emerged (Fig. 1.2). Could this be four different genomes? If so, are they complete? The contigs of two clusters showed similarity to an earlier sequenced genome of an archaeon, and the sum of the contig's length matched the genome size of the sequenced genome. The two other clusters contained a 16S gene corresponding to the dominant bacterial genus *Leptospirillum*, which had no sequenced genome for the entire phylum.

How would one show the completeness of a genome without related reference genomes? Tyson et al. did so by verifying if all genes for the essential transfer RNA synthetases were present in the cluster. Based on this estimation, one *Leptospirillum* species was complete and one nearly. Interestingly, the lower abundant species were the only species to fix nitrogen and supply it to the community. This result could not be obtained other than culturing or genome-resolved metagenomics. Also, because they used such a deep sequencing, they could investigate the strain distribution. The *Leptospirillum* species existed as a unique strain, whereas the archaea species was a mosaic of free recombining cells that originated from three ancestral strains. This initial success of recovering two nearly complete genomes laid the stepping stone for further studies.

Tyson et al. used simple thresholds to define *bins* which they showed to be reasonable approximations of genomes. Even though today, much more sophisticated algorithms are used to segregate contigs into genomes, the term **binning** stuck. Most algorithms work under the paradigm that one contig belongs only to one bin.

How do we bin contigs into genomes?

Since the first genomes were recovered from metagenomes, a plethora of algorithms was developed for metagenomic binning. Some rely on visual inspection and manual curation (Zhu et al., 2018; Eren, Esen, et al., 2015), while others are fully automated. The explosion of developed algorithms and the advancement of metagenome assembly made it difficult to compare the different tools on even ground. The *critical assessment of metagenome interpretation* (CAMI) challenge was initiated to benchmark various metagenomic tools on newly sequenced and realistic datasets, but the benchmark could not keep up with the development of new algorithms.

The binning algorithms become more and more sophisticated, but the idea says the same: Grouping contigs into genomes based on two sets of features, **sequence features** and **abundance features**: Many ways exist to extract features from the DNA sequence of a contig (or a read). It is possible to predict genes and map them or their translated protein to a database to find similarities with existing genes. Genes that are generally found only once in a genome can be used to constrain the clustering or estimate the number of genomes present in the sample (See more in sec. 1.2.3 and 5.1.1).

More directly, the raw **sequence composition**, measured as frequencies of k-mers, can be used as features. Even before sequencing was available, it was recognized that the frequencies of the four DNA bases (A, T, G, C) and their combination follows non-random patterns (Josse et al., 1961). It was shown that the sequence composition is more similar between genomes of the same taxon than from a different (Nussinov, 1980) and that it can be used to classify genomes or contigs into a taxonomy (Sandberg et al., 2001). The GC-content is the sum of guanine (G) and cytosine (C) frequencies ($k = 1$). While the specificity of the genome signature increases with the length, the number of possibilities increases exponentially with k. Today, most of the time, frequencies of **tetra-nucleotides**, k-mersof size 4, were used to measure the sequence composition. The tetra-nucleotide frequency capture more of the particularities of the contig sequences as they not only incorporate the GC-content but also codon-bias.

The feature of the abundance is, most of the time, the average coverage of a contig. However, instead of using only the coverage information from one sample, contigs can be clustered by the abundance from different samples. An idea that was first developed by Albertsen et al. who used two different extraction protocols of the same microbiome sample to get two different abundance values of the same strains. Using this method, the researchers could recover 31 genomes from a bioreactor, including from low abundant species.

Binning based on differential abundance

Binning based on differential abundance requires one assembly with coverage information from different samples. To achieve this, usually, all samples are **co-assembled**, and the reads from all samples are mapped to the same assembly. Because co-assembly combines the data from multiple samples, it allows the assembly of contigs from more low-abundant genomes. However, co-assembly also cumulates the (strain-)diversity from numerous samples and multiplies the challenges associated with metagenome assembly. It is also possible to assemble samples separately and merge the assemblies, which reduces the computational burden for the co-assembly. However, it looses also the advantage of assembling low-abundant genomes, and it still contains inter-sample chimeras.

Alternatively, one can map the reads from each sample to all assemblies of a project. In this way, one keeps the advantages of both single-sample assembly and binning based on differential abundance. However, the downside of **cross-mapping** is that it requires $N_{samples}^2$ mappings and is therefore not salable.

An interesting new approach, implemented by Nissen et al., is based on single-sample assembly but **co-binning**. Each sample is assembled separately, and the assemblies are concatenated. As in co-assembly, the reads from all samples are mapped to the same assembly. Because the combined assembly contains multiple times the same genome, the multiple mapping sites of reads have to be considered. Binning is performed using a deep variational autoencoder, which autonomously learns how to weight the sequence and the abundance feature. After clustering, the contigs from each sample are separated to create sample-specific bins for each cluster. This method pays much attention to strain variation by assembling samples separately and splitting the clusters in sample-specific bins. It allows disentangling strains up to 99.5%, according to the authors.

This idea of using the **differencial abundance** of contigs in multiple samples was leveraged by Nielsen et al. from the MetaHIT consortium to produce the first binning of a complex microbiome (Nielsen et al., 2014). As the metagenome assemblers were still not at their best, the researchers based their analysis on genes quantified in 396 human gut metagenome samples. The correlation of these abundance values allowed them to segregate the genes into clusters. Some of these clusters had the size of bacterial genomes, whereas others could be associated with phages. As the method is based on genes, the resulting clusters do not contain the genome sequence, but the segregation of the genes can assist assembly. By re-assembling all the reads associated with a specific cluster, the authors could reconstruct 360 genomes.

The year after, Christopher T. Brown et al. published eight complete and 789 draft genomes from a group of tiny bacteria (< 0.2 µm). These bacteria, which were only known by environmental sequencing, have shrunken genomes and lack many genes and pathways, thought to be essential (Christopher T. Brown et al., 2015). Their 16S gene is different from known bacteria, making more than half of these organisms undetectable by 16S sequencing. They were first thought to represent multiple phyla and are therefore called the candidate phylum radiation (CPR). However, subsequent phylogenetic analysis showed that they are part of the phylum *Patescibacteria* (Donovan H Parks et al., 2018). This example shows again how genome-resolved metagenomics allows the studying of previously undetected organisms.

At the beginning of my Ph.D. Tyson and his collaborators re-analyzed many public metagenomes and recovered nearly 8'000 genomes (Donovan H. Parks, Rinke, et al., 2017). Their publication marked the beginning of a new area of large-scale (re-)assembly in metagenomics. Other studies were soon to follow that tried to reconstruct large sets of genomes from metagenomes.

1.2.3 What does it mean to be complete?

Metagenome-assembled genomes, short MAGs, are only a bunch of contigs clustered together. Do they correspond to real genomes? Are they as good as genome sequencing from isolates? Comparing a MAG to a close reference genome can help to validate the binning. If the MAG does not have a related reference genome, the MAG quality is commonly estimated based on marker genes. Similar to Tyson et al. one can search for genes found in practically all microbial genomes, such as t- and rRNA genes and their associated proteins. The **completeness** of a genome is estimated as the fraction of marker genes present divided by their expected number (See equation below). Similarly, marker genes present only once in practically all microbial genes are used to estimate **contamination** of MAGs.

$$\text{Completeness} = \frac{\text{present}}{\text{expected}} \text{ marker genes}$$

$$\text{Contamination} = \frac{\text{duplicated}}{\text{expected}} \text{ marker genes}$$

In practice, the same set of genes is used to estimate completeness and contamination, even if not required. The set of marker genes can be adapted to the novelty of a MAG. For example, a MAG is first assessed based on universal marker genes and then is placed in a phylogenetic tree. This approximative taxonomy is used to identify the closest clade for which a marker gene set is available, and the MAGs quality is assessed with more detail. There are approximately 50 genes that are single-copy and present in all bacteria and archaea. Still, over 100 phylum-specific marker genes can be defined, and even more for lower taxonomic levels. The tool `checkM` is the most used tool that performs this phylogenetic-specific quality assessment, but its database and phylogenetic tree are not updated since 2015. During my Ph.D., I contributed to adapt `BUSCO` (Seppey et al., 2019) to perform the same task. `BUSCO` is based on the marker genes of the regularly updated OrthoDB (Kriventseva et al., 2019) and assesses the quality not only of prokaryotes but also eukaryotes.

Commonly the completeness and contamination estimates are combined in a single quality score with five times more weight on the contamination (See equation below). Genomes below a quality score of 50% are regarded as low-quality, and genomes with >90% are counted as high-quality genomes. The *Minimum information about a metagenome-assembled genome* (MIMAG) criteria additional expect the presence of the rRNA genes and at least 18 tRNAs (Bowers et al., 2017). However, these genes are complicated to assemble from metagenomes and are usually not counted as a requirement for high quality or near-complete MAG (Almeida et al., 2019; Gruber-Vodicka et al., 2020).

$$\text{Quality score} = \text{Completeness} - 5 \times \text{Contamination}$$

Most genomes recovered from metagenomes do not reach the maximal quality score. There can be biological reasons for this, as it is the case for the tiny bacteria of the candidate phylum radiation (sec. 1.2.2). These bacteria systematically miss subsets of the marker genes present in 90% of all bacteria (Eren & Delmont, 2017). However, this group is only the exception that confirms the rule. Most MAGs are not complete. Even MAGs that are of high quality might contain assembly errors, strain chimeras and are often much more fractonized (measured by the $N50$-metric) than isolated genomes (Chen et al., 2020). The isolation of a species followed by sequencing its genome is the optimal approach, even if this does not guarantee a complete genome nor a genome without contamination. As described in chapter 3, we found low-quality isolated genomes that are part of official culture databases. Some had even contamination of 100 %, meaning they consist of an isolate of two strains.

1.2.4 Who is doing what?

Regardless if a genome is assembled from an isolate or recovered from a metagenome, the bare DNA sequence is only an intermediate step. It is more interesting to annotate a genome with its taxonomy and the functional potential of its genes. Genes are predicted by identifying gene start sites and stop codons. The coding sequence in between is then translated to the corresponding protein. The predicted protein sequences are mapped to a database of functions to annotate them. Proteins can have wildly divergent sequences and still exhibit the same function. To optimize speed and sensitivity, often *hidden Markov models* (HMMs) are used to annotate genes with functions. The HMMs are constructed from the alignment of proteins with the same function and encode all of the similarities and variabilities of the proteins in a simple numerical matrix.

An optional but beneficial step for functional annotation is integrating the protein annotations into a higher organization order. We often use the term **pathway** to describe a collection of functions that represent a well-characterized segment of the molecular machinery of a cell. For instance, a metabolic pathway describes a group of enzymes used to produce a metabolite or its degradation. Other pathways may be implicated in the cell-to-cell communication (quorum sensing) or the adaptation to an environment, for example, sporulation or biofilm formation. Some pathways are encoded by physically clustered groups of genes, for example, **biosynthetic gene clusters** (BGCs), genes that together encode a biosynthetic pathway (Medema et al., 2015), or operons, commonly regulated adjacent genes. Nonetheless, in general, pathways are only human-made schematic representations without genetic correspondence.

The most used database for functional annotation is KEGG (Kyoto Encyclopedia of Genes and Genomes, Kanehisa & Goto, 2000). The organization of the KEGG database is based on KEGG-orthologs (KOs), curated groups of genes that have the same function. Metabolic KOs correspond most of the time to an enzyme from the Enzyme Commission (EC number). KOs are organized into modules (segments of pathways) and large overarching pathway maps. Analogous pathways can also be found in the MetaCyc database (Caspi et al., 2016).

The crucial step of the pathway integration is deciding if a set of genes are sufficient to mark the presence of a pathway in a genome or not. The challenge arises because functions can be part of multiple pathways and that often various versions of pathways exist. Therefore the presence of one or two functions does not mean that a pathway is present. On the other hand, a missing annotation might be a false negative, especially in MAGs, which are rarely complete. To infer the presence of pathways, it is necessary to model them to a certain extent. It is even possible to create metabolic models for the whole organism. These can then be used for inferring metabolic symbiosis between members in a microbiome (Machado, Andrejev, et al., 2018; Belcour et al., 2020; Machado, Maistrenko, et al., 2021). Nevertheless, less than half of genes can be annotated for most species. Even less can be integrated into pathways (Richardson et al., 2019), which lets substantial room for identifying new gene functions through metagenomic studies.

Taxonomic annotation of Genomes

Taxonomic annotation is also based on genes. Usually, conserver marker genes are predicted and aligned to a reference taxonomy, which allows to place the genome into a phylogenetic tree. On the other hand, for accurate species attribution, the calculation of the **average nucleotide identity** (ANI) is often preferred (Donovan H. Parks, Chuvochina, et al., 2020). However, because calculating the ANI requires pairwise alignments of genomes, it scales quadratically with the number of genomes. To solve this computational problem, efficient tools have to be used to deal with the flood of genomes recovered from metagenomes. One essential advancement is the implementation of a MinHash algorithm for genome comparison. So, allows the tool `mash` and other implementations, the fast comparisons of millions of genomes in low memory (Ondov et al., 2016).

Still, the best algorithms for taxonomic annotation are useless without a robust taxonomy, which was a critical bottleneck until recently. ”*Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification.* ” (Donovan H Parks et al., 2018). Philip Hugenholtz, which was already a forerunner in systematizing prokaryote taxonomy based on the 16S gene (McDonald et al., 2012), developed the **genome taxonomy database** (GTDB), that establishes a robust genome-based taxonomy (Donovan H Parks et al., 2018; Donovan H. Parks, Chuvochina, et al., 2020). It includes all cultured genomes but also many genomes recovered from metagenomes. Not very surprising, most species in GTDB come only from metagenome-assembled genomes[5].

1.3 Goal of the thesis

The goal of my P.h.D thesis is to enable the analysis of the mouse gut metagenome.

Mice are the most used model organism to study the impact of the microbiome on its host. Several factors make the mouse a good model: The availability of samples from different parts of the gastrointestinal tract, treatment options, controlled diet, and housing environment, defined genetic background, and ethical considerations. However, the mouse gut microbiota has been poorly characterized, and only a fraction of the diversity observed by 16S rDNA sequencing is represented by genomes in public databases (Lagkouvardos et al., 2016). The majority of mouse microbiome studies are performed by sequencing the variable regions of the 16S gene. While this technique has allowed a general overview of the microbiota down to the genus level, it is not suited for

[5] gtdb.ecogenomic.org – Stats

identifying species for most organisms (Johnson et al., 2019). Different species from the same genus and even subspecies from the same species can exert distinct functions (Costea et al., 2017), stressing the importance of annotating the microbiome content at the lowest taxonomic level.

Shotgun metagenomics allows studying the full microbiota diversity of an environment, including uncultured microorganisms, viruses, and plasmids. However, its interpretation is limited by the availability of reference genomes. There is a lack of reference genomes from the members of the mouse microbiome, as is apparent from the low mapping rate (Fig. 1.3). We could classify less than 20% using commonly used workflows to analyze the (human) metagenome. Even with a mouse-specific gene-catalog (Xiao et al., 2015), we could annotate about half of the reads, which corresponds to what the authors of the gene catalog have stated. It is important to note that while the gene-catalog is an essential reference for genes and functions, it has only a limited taxonomic annotation and does not contain genomes, which hinders the linking of function to species.

To solve the lack of reference genomes for the mouse microbiome, I turned to algorithms that make it possible to reconstruct genomes from metagenomes. I implemented these algorithms in an efficient pipeline that allows users to go from the raw reads to reconstructed and annotated genomes. In chapter 2, I describe the development of this pipeline called `metagenome-atlas`. We took the challenge to create a comprehensive collection of reference genomes from the mouse microbiome by assembling all public mouse metagenomes and our own. This effort I document in chapter 3. We demonstrated how reconstructed genomes could be used to analyze metagenomes of mice and relate them to the host's health in section 3.1 and chapter 4. Finally, we annotated the microbiome's functional potential and predicted changes in metabolites, which were confirmed by targeted-metabolomics and have the potential to improve osteoporosis (ch. 4).

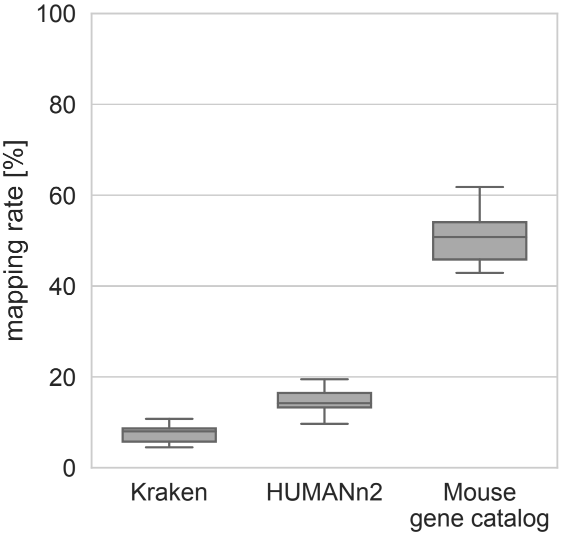


Fig. 1.3 Benchmark of metagenomics tools to classify reads from a mouse gut metagenome at the beginning of my Ph.D.: The faction of reads classified by HUMAnN2 (Franzosa et al., 2018), Kraken2 (Derrick E. Wood et al., 2019) with a database of all bacteria and virus genomes from RefSeq or the fraction or reds mapping to the mouse gene catalog (Xiao et al., 2015). Reads were mapped using bbmap, with a minimum identity of 0.9. To see the improvement achieved see Ch. 3 Fig. 3B.

Recovering genomes from metagenomes made easy

BMC Bioinformatics 21, 2020

Title: ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data

Authors: Silas Kieser, Joseph Brown, Evgeny M. Zdobnov, Mirko Trajkovski & Lee Ann McCue

Status: Published in *BMC Bioinformatics* 21, 2020

This manuscript describes the key method used in the following parts of my thesis. It describes the tool metagenome-atlas, a state-of-the-art pipeline for the analysis of metagenome data. Atlas allows to recover genomes from short metagenomic reads, it also quantifies the abundance of this species and annotates the taxonomy and the functional potential of their genes. The pipeline is easy-to-use and handles all steps from quality control, assembly, binning, to annotation and quantification. It downloads all the required software tools and databases on the fly.

Contribution statement

The pipeline was initially developed by Joseph Brown in the Pacific Northwest National Laboratory. In May 2018, I asked to contribute to the opensource pipeline and was accepted in the developer team. Since then, I became the main contributor and lead developer as can be seen from the contributor statistics on the GitHub repository. I maintained the tool and hellped to it's popularization with tutorials, classes, and workshops. I wrote the manuscript together with the other co-authors.

The comprehensive mouse gut metagenome catalog (CMGM)

Title: Comprehensive mouse gut metagenome catalog reveals major difference to the human counterpart

Authors: Silas Kieser, Evgeny M Zdobnov & Mirko Trajkovski

Status: Submited. Here we show the original version of this articles, a revised version is available as pre-print from *bioRxiv*, 2021

Mouse is the most used model for studying the impact of microbiota on its host, but the repertoire of species from the mouse gut microbiome remains largely unknown. We took on the challenge to create a comprehensive catalog of genomes from the mouse gut. We predicted bacterial and viral genomes from over a thousand public mouse metagenomes as well as our own. We also included reference genomes isolated from the mouse microbiome. We compared the resulting catalog of metagenome-assembled genomes and reference genomes to the Unified catalog of genomes from the human gut and uncovered major differences in the species composition. Our catalog increases our knowledge of the mouse microbiota gene repertoire by ten-fold and allows comprehensive analysis of the mouse gut microbiome at an unprecedented depth.

Contribution statement

I wrote the code, analysed, and interpreted the data, and generated the figures. E.Z. and M.T. guided the project and supervised the work. All authors conceptualized the study and wrote the paper.

3.1 CMGM enables comparative analysis of mouse metagenomes by relating functional changes to driver species

This section shows how the comprehensive mouse gut metagenome catalog (CMGM) can be used to analyze mouse microibiome data on both a functional and a species level and to relate one to the other.

This analysis was is part of the revised manuscript available from *bioRxiv*, 2021 .

The code for this analysis is available from github.com/SilasK/CMGM

To illustrate how this catalog allows discovering compelling biological insights, we analyzed the metagenome from mice exposed to cold ambient air temperature. Cold exposure is a stimulus that activates the classical brown fat and promotes beige cell development within the subcutaneous white adipose tissue[1–3]. As such, it is an extensively used intervention for enhancing thermogenic and mitochondrial activity in adipose tissues, leading to decreased adipose tissue amount and improved glycemic status. We[4], and others[5] showed that cold exposure leads to a marked shift of the microbiota composition observed by 16S

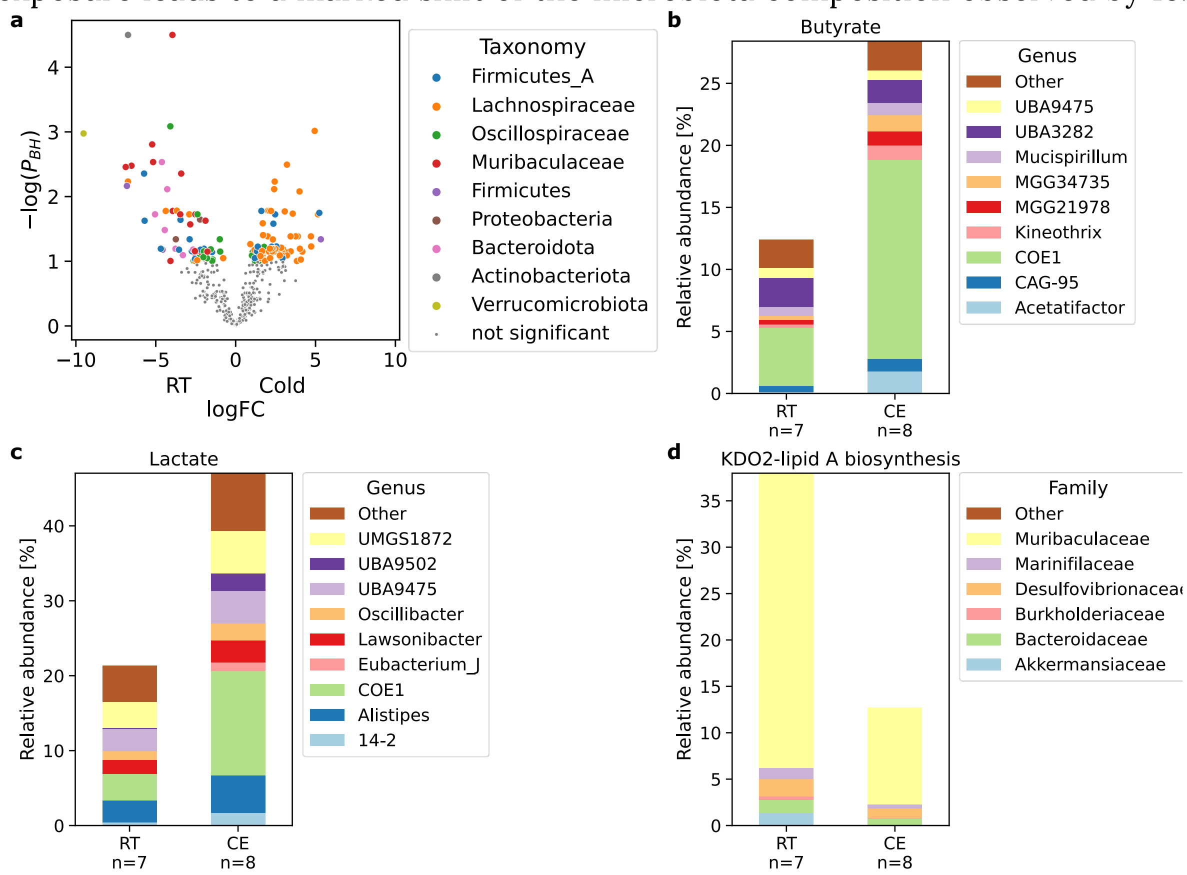


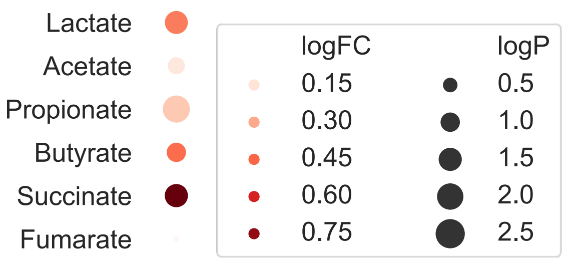
Figure 1 | CMGM links functional changes to driver species

a, Volcano plot of species changes in mouse cecal microbiota upon cold exposure. Significantly changed species are colored by their phylum. P_{BH}: P-value corrected for multiple testing using Benjamini-Hochberg procedure. b-d, Bar plots of significantly changed pathways in mouse cecal microbiota upon cold exposure. The contribution to the relative abundance of each module is partitioned by genus B+C and family D.

CE: Cold exposure, RT: Room temperature control analysis, which is in itself sufficient to improve the insulin sensitivity, induce tolerance to cold, increase the energy expenditure and lower the fat content– an effect in part mediated by activation of the brown fat[4,5] and browning of the white fat depots in the cold microbiota-transplanted mice[4,6–9]. These results indicate an existence of a microbiota-fat signaling axis; however, the signaling cascades mediating this process remain poorly understood.

As noticed previously[4], here we confirmed that *Akkermansia muciniphila*, the only representative of the phylum *Verrucomicrobiota* was eliminated by cold exposure (Fig. 1A). The species NM07-P-09 sp004793665 (the most abundant species from the phylum *Actinobacteriota*) and the *Muribaculaceae* species UBA7173 sp002491305 were even more significantly decreased (P_{BH} < 1e-4, Fig. 1A). We found that cold exposure leads to an increase of the family *Lachnospiraceae* and a decrease of the families *Muribaculaceae,* and *Oscillospiraceae.*

On a functional level, cold exposure led to a doubling of butyrate and lactate production. These changes were mainly due to the increase of the family *Lachnospiraceae,* specifically the increase of the uncultured genus *COE1* (Fig. 1B, C). To address whether these uncovered metagenomic changes are indeed reflected in differences of the actual metabolite levels, we looked at the germ-free mice transplanted with microbiota from the cold-exposed mice or from their RT-kept controls. Transplantation of the cold-adapted microbiota led to an increase in the production of butyrate, lactate, propionate, and succinate in the recipients' cecum compared to germ-free mice inoculated with microbiota from control mice (Fig. 2). Interestingly, the increased lactate was also measured in the cecum and serum of mice with an intermittent fasting feeding regime[10], which has been shown to induce browning via the induction of the Vascular endothelial growth factor[11]. Similarly, succinate is linked to the increase of thermogenesis[12]. We found a decrease of the prokaryotic succinate dehydrogenase, which metabolizes succinate to fumarate, suggesting a mechanistic link between the cold-induced microbiota changes and the adipose tissue browning.



1

Figure 2| Metabolite changes by cold-adapted microbiome
Dot-plot of metabolite changes in ceca of germ-free mice transplanted with cold-adapted microbiota compared to RT-microbiota transplanted controls (Data from Ref[4]).

We also observed a decrease in Lipopolysaccharide (LPS) synthesis, both in an LpxL-LpxM–dependent and –independent way, primarily attributed to the cold-induced reduction of *Muribaculaceae* (Fig. 1D). LPS administration leads to reduced core body temperature and heat release, correlated with mitochondrial dysfunction[13]. In contrast, genetic deletion of the LPS receptor, the toll-like receptor 4 (TLR4), leads to resistance against high caloric diet-induced obesity, improved glucose tolerance and insulin sensitivity, and adipose tissue browning[14]. These findings suggest an additional possible link between the cold-induced microbiota changes and adipose tissues both at mechanistic and bacterial level, contributing to improved insulin sensitivity and browning of the white fat

This example illustrates the CMGM catalog's usability as a reference for metagenomic studies, enabling discovering precise and comprehensive changes of species and the related function induced by a treatment or a disease. The CMGM sets the ground for reanalysis of the existing datasets for uncovering species and bacterial functions that are involved or altered by the condition of interest.

Methods

Sequencing of metagenomic data of mice

The mouse experiments were approved by the Swiss federal and Geneva cantonal authorities for animal experimentation (Office Vétérinaire Fédéral and Commission Cantonale pour les Expériences sur les animaux de Genève). Animals were on C57Bl/6J background, commercially available through Charles River, France. The mice experiment is detailed in[4]. Paired-end metagenomic libraries were prepared from 100 ng DNA using TruSeq Nano DNA Library Prep Kit (Illumina) and size selected at about 350 bp. The pooled indexed library was sequenced in a HiSeq4000 instrument at the iGE3 facility (University of Geneva).

Quantification

We used BBsplit (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/) with the parameters' ambiguous2=best minid=0.9' to map metagenomic reads to the reference genomes with 90 identity. For most quantification, the mapped reads per genome were summed, and the centered log-ratio (CLR) was calculated using the sci-kit bio package (http://scikit-bio.org/) after imputing zeros using a multiplicative replacement approach. We used the two-sided Welch test and Benjamini-Hochberg correction to estimate the significance of changes in clr-transformed genome abundance between experimental groups. We estimated the genome coverage as the median of coverage over 1000bp blocks, to calculate relative abundance.

Functional annotation

The species representatives of both the CMGM and the UHGG were annotated using DRAM[15]. A Kegg-module is inferred to be present if ¾ of all the steps were present in a genome. As there are no modules for short chain fatty acids in Kegg we created custom modules (see the 'Code' section). The step-coverage was calculated with DRAM for all Kegg-modules. The metagenome-side abundance of functional modules was calculated as the sum of the relative abundances of all genomes containing a module. We used the two-sided Welch test and Benjamini-Hochberg correction to estimate the significance of changes in module abundance between experimental groups.

Code Availability

The code for the analysis of the cold-exposed microbiota is available from https://github.com/SilasK/CMGM

References

1. Cannon, B. & Nedergaard, J. Cell Metabolism Previews What Ignites UCP1? *Cell Metab.* **26**, 697–698 (2017).

2. Chechi, K., Carpentier, A. C. & Richard, D. Understanding the brown adipocyte as a contributor to energy homeostasis. *Trends Endocrinol. Metab.* **24**, 408–420 (2013).

3. Stojanović, O., Kieser, S. & Trajkovski, M. Common traits between the beige fat-inducing stimuli. *Curr. Opin. Cell Biol.* **55**, 67–73 (2018).

4. Chevalier, C. *et al.* Gut Microbiota Orchestrates Energy Homeostasis during Cold. *Cell* **163**, 1360–1374 (2015).

5. Ziętak, M. *et al.* Altered Microbiota Contributes to Reduced Diet-Induced Obesity upon Cold Exposure. *Cell Metab.* **23**, 1216–1223 (2016).

6. Guerra, C., Koza, R. A., Yamashita, H., Walsh, K. & Kozak, L. P. Emergence of brown adipocytes in white fat in mice is under genetic control. Effects on body weight and adiposity. *J. Clin. Invest.* **102**, 412–420 (1998).

7. Kopecky, J., Clarke, G., Enerbäck, S., Spiegelman, B. & Kozak, L. P. Expression of the mitochondrial uncoupling protein gene from the aP2 gene promoter prevents genetic obesity. *J. Clin. Invest.* **96**, 2914–2923 (1995).

8. Ghorbani, M., Claus, T. H. & Himms-Hagen, J. Hypertrophy of brown adipocytes in brown and white adipose tissues and reversal of diet-induced obesity in rats treated with a β3-adrenoceptor agonist. *Biochem. Pharmacol.* **54**, 121–131 (1997).

9. Cypess, A. M. *et al.* Activation of Human Brown Adipose Tissue by a β3-Adrenergic Receptor Agonist. *Cell Metab.* **21**, 33–38 (2015).

10. Li, G. *et al.* Intermittent Fasting Promotes White Adipose Browning and Decreases Obesity by Shaping the Gut Microbiota. *Cell Metab.* **26**, 672-685.e4 (2017).

11. Kim, K.-H. *et al.* Intermittent fasting promotes adipose thermogenesis and metabolic homeostasis via VEGF-mediated alternative activation of macrophage. *Cell Res.* **27**, 1309–1326 (2017).

12. Mills, E. L. *et al.* Accumulation of succinate controls activation of adipose tissue thermogenesis. *Nature* **560**, 102–106 (2018).

13. Okla, M. *et al.* Activation of Toll-like Receptor 4 (TLR4) Attenuates Adaptive Thermogenesis via Endoplasmic Reticulum Stress. *J. Biol. Chem.* **290**, 26476–26490 (2015).

14. Fabbiano, S. *et al.* Functional Gut Microbiota Remodeling Contributes to the Caloric Restriction-Induced Metabolic Improvements. *Cell Metab.* **0**, (2018).

15. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

Identifying the functional changes in the microbiome

Title: Warmth Prevents Bone Loss Through the Gut Microbiota

Authors: Claire Chevalier, Silas Kieser, Melis Çolakoğlu, Noushin Hadadi, Julia Brun, Dorothée Rigo, Nicolas Suárez-Zamorano, Martina Spiljar, Salvatore Fabbiano, Björn Busse, Julijana Ivanišević, Andrew Macpherson, Nicolas Bonnet & Mirko Trajkovski

Status: Published in *Cell Metabolism* 32, 2020

In this study is the prime example how metagneome-atlas can be used to analyze the mouse metagenome. My collegue, Claire Chevalier, analyzed the effect of ambiant temperature on the metabolism of mice and the changes in their gut microbiome. When mice are exposed to warm ambient temperature for a prolonged time (34°C, 8weeks), they enlarge the tail and ears, whcih creates more surface for heat-disipation. The warm temperature not only has a remodeling effect on the bone but also on the gut microbiome. The transplantation of the gut microbiota of warm exposed mice induces similar changes on the bone of the recipient mice and can prevent osteoporosis. Epidimological analysis shows a significant correlation between the incidence of hip-fracture and the average temperature in 60 countries.

In order to identify the mechanism how the gut microbiome infulences the bone strength of the host, we used metagenomics and predicted the functional potential of the gut microbiome. Polyamin-synthesis was one of the most significant increased pathways in the gut microbiome upon warm-exposure (Fig. 7, S6). This prediction was confirmed by targeted-metabolomics and the effect of polyamine on the host was further corroborated by *in vivo* suplementation and inihibition of their synthesis. The genome-resolved metagenomics not only allowed to identify an important pathway how the gut microbiome influences the host health, but also to identify the driver species (Fig. S6).

Contribution statement

This project was lead by my collegue Claire Chevalier. Like most other lab mebers, I helped at the end of the mouse experiment for the scrifice of the mice. Together with her I extracted the DNA and prepared the library for the 16S amplicon sequencing and shot-gun metagenomics. I analyzed the metagenome data and implmented a functional prediction specific for this project. Together with her we interpreted this data. I also (re-) analyzed the 16S data and made the related figures. I performed the correlation analysis of the human epidemiological data.

Discussion

5.1 Metagenomics in a post-assembly area

The goal of my P.h.D. was to enable the analysis of metagenomes with a lack of reference genomes to study the microbiome's influence on the host's health. The core methodology I implemented is the reconstruction of genomes from metagenomes. Now, at the end of my P.h.D. I wonder what the future of this methodology will be?

With the accumulation of large-scale studies that recovered genomes from metagenomes, the lack of reference genomes for major host-associated microbiomes diminishes. So have, during my thesis, three groups assembling and processing almost all publicly available human gut metagenomes (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., n.d.). Collections of genomes for the gut microbiome of farm animals were also released during my P.h.D. (Stewart et al., 2019; Glendinning et al., 2020). Finally, in chapter 3, I document my effort to reconstruct a comprehensive set of genomes from the mouse gut metagenome. Each study recovers hundreds of species, and rarefaction analyses show that the saturation is approaching.

Surely, I do not want to ignore the challenges in non-gut metagenomes. For example skin, and lung microbiomes yield low-biomass samples and are therefore difficult to assemble. Also, soil and ocean microbiomes are more complex than the typical gut microbiome, and therefore harder to assemble[1]. Nevertheless, I would say that we now have reference genomes for the majority of species in many metagenomes.

[1] Even though, for the ocean microbiome, a significant improvement was made recently (Lucas Paoli et al., 2021).

If this is the case, we are about to enter a *post-assembly area*, where the assembly of metagenomes is no longer necessary, and microbiomes can be profiled directly. I would say, having comprehensive catalogs of genomes for many microbiomes is a milestone in the analysis of microbiomes[2]. Before, we had to rely on taxonomic databases of marker genes, such as 16S or mOTUs (Milanese et al., 2019) or the functional annotations of gene catalogs. Now we have the integration of both in genome collections.

Where do we go from here? I think there is still much room to improve the quality of the genomes recovered from metagenomes. Also, the reliance solely on marker genes for the quality estimation of genes might blind us to other sources of errors during the genome assembly and binning. Finally, more important to improve the current genome references is to use them effectively. In section 5.2 I discuss what I have learned about the statistical analysis of metagenomes.

5.1.1 Recover complete genomes from metagenomes

The genomes reconstructed from metagenomes are of variable quality. While for highly prevalent species, multiple genome reconstructions are available, and one can choose the best one as a representative, for rare species are often represented only by one medium-quality MAG. Incomplete or ”*composite metagenome-assembled genomes reduce the quality of public genome repositories*” (Shaiber & Eren, 2019). Moreover, even what is called a high-quality genome is mostly an estimation based on marker genes (See section 1.2.3). The quality estimation is dependent on the marker gene set used. Therefore a bias in the marker gene set induces a bias in the genome estimation. More fundamentally, I think genome quality estimation is overused. Metagenomic binners are evaluated on the quality score of their genome predictions. Marker genes are even used during the binning by some algorithms or by tools that combine and consolidate the results of multiple binners (Sieber et al., 2018). I fear this to be an example of **Goodhart's Law**: ”*When a measure becomes a target, it ceases to be a good measure.*” (Strathern, 1997).

The fundamental problem of estimating the quality of a genome solely by assessing the presence and duplication of marker genes is that this approach is

[2] It is actually included in Nature's ”Milestones in human microbiota research” (2019)

entirely blind to contigs that do not contain marker genes. A MAG may have many contigs from a wholly different species without affecting the contamination estimation. Similarly, a ”complete” genome might still be missing genome content that is not assessed by marker genes.

New tools are developped that claim to purify a MAG of this unassesed contamination (`MAGpurify` (Nayfach et al., 2019), `GUNC` (Orakov et al., 2020) and `conterminator` (Steinegger & Steven L. Salzberg, 2020)) or to search for additional contigs that were missed (`Spacegraphcats` (C. Titus Brown et al., 2020) and `GraphBin` (Mallawaarachchi et al., 2020)). However, often it is only through manual curation to achieve an accurate and complete genome from metagenomes (Chen et al., 2020).

The optimal MAG would be one that is assembled in one continuous sequence. For now, this only rarely happens. It is also important to note that most large-scale efforts use single-sample assembly, as this approach is the most scalable. Binning methods that use differential abundance in a targeted way (See box in sec 1.2.2) are promising ways to improve the continuity and quality of MAGs.

5.2 Measuring the microbiome

A fundamental step of almost every endeavor in science is measuring. We measure to compare our measurements between different experiments, samples, or even studies and draw generalizable conclusions from them. Measuring in the context of sequencing-based approaches often comes down to mapping (short-)reads to a reference (-genome) and counting the mapped reads.

Nevertheless, measuring is not straightforward, as we will see, both with 16S amplicon sequencing and metagenomics. One has to define appropriate units for the quantification of the microbiome. For this, one has to consider a *tradeoff between specificity and comparability*. Finally, it is crucial to heed the **compositional nature** of microbiome data for its interpretation.

5.2.1 What to measure?

Defining units for 16S amplicon sequencing

There is a long-standing discussion on how to define units for 16S amplicon sequencing. In a typical amplicon study, most sequences are only remotely similar to annotated species. One option is to map a sequenced amplicon to the closest matching sequence in a database of curated 16S genes, like SILVA (Quast et al., 2012). This **closed-reference** approach has the advantage that the annotated units can be compared between studies (At least the one that uses the same version of the same database). However, the mapping to the database might be ambiguous, and many new variants specific to the study might be missed. It is often much more informative to cluster all the sequenced amplicons for one study and use them as measuring units. As these clusters do not necessarily correspond to biological species, they are called **operational taxonomic units** (OTUs). This ***de novo***-clustering approach has the disadvantage that the OTUs are specific to one study, limiting the inter-study comparison.

In 1994, E Stackebrandt & Brett M. Goebel proposed 97% similarity as a species threshold for the full-length 16S gene. This threshold was adopted for amplicon sequencing (Patrick D. Schloss & Handelsman, 2005), even though this technique is based on a much smaller gene fraction. More recent analyses of the correspondence between the 16S gene and species found that the 99% would be a better threshold for the full-length 16S gene and that amplicons of the 16S gene are not suited to achieve consistent species resolution (Edgar, 2018; Johnson et al., 2019). Today a 100%-threshold is commonly used. In order to distinguish biological variation from sequencing errors, the reads have to be denoised. This denoising step is implemented, for example, in the tool `DADA2` (Benjamin J Callahan et al., 2016) in an efficient way. The 100% clusters are also referred to as **amplicon sequence variants** (ASVs). Using ASVs combines the advantages of the *de novo*- and the closed-reference approach. ASVs cover all sequence variance within a dataset and are unique and consistent, allowing them to be compared across different datasets (Benjamin J. Callahan et al., 2017).

Critics of 100% ASVs point out that a genome may be split into multiple ASVs if it contains multiple copies of the 16S gene that are sufficiently diverged (Patrick D Schloss, 2021). Re-clustering of ASVs based on correlation across different datasets could mitigate this problem. In conclusion, the ASVs, are the most specific and comparable unit to measure a microbiome even though these units do not correspond precisely to species.

Defining units for metagenomics

Whereas in 16S amplicon sequencing, the denoised reads or ASVs *are* the unit for quantification, for shot-gun metagenomics reference units needed to be defined to which the reads can be mapped. Usually, the taxonomic unit of species is used for metagenome profiling, which permits the results to be compared among studies and previous knowledge.

The main problem for metagenome quantification was, until recently, the lack of reference genomes. In absence of good references the `mOTUs` was invaluable. This tool creates taxonomic units based on 40 universal single-copy marker genes (Sunagawa et al., 2013). Because this tool is based on universal marker genes, it allows targeting virtually all (prokaryote) genomes in metagenomes and, therefore, precise estimation of their relative abundance. It is possible to create mOTUs even for low-abundant and unknown species by assembling these genes from metagenomes and linking them through the correlation of their coverage across many samples.

As described above, genome-resolved metagenomics allows the generation of reference genomes for metagenomics. Now, the problem becomes how to map metagenomic reads efficiently to the ever-growing reference databases. In a breakthrough publication in 2014, Derrick E Wood & Steven L Salzberg showed that it is possible to circumvent the time-consuming step of precise alignment and directly assign reads to taxonomy utilizing exact k-mer-matches. The tool called `Kraken` was used to show that classifications, which previously took hours, were tractable in minutes. Notwithstanding, most k-mersare not specific to a species, which leads to the classification of many reads at higher taxonomic levels. It is shown that this problem is only aggravated with the inclusion of more and more reference genomes (Nasko et al., 2018). As an alternative to decomposing the whole genome into more or less specific k-mers, it makes sense to search for genome regions, usually genes, specific for a species. The classical tool `metaphlan` is based on this approach (Segata, Waldron, et al., 2012) and the idea was reused in a quantification tool based non the collection of recently recovered human MAGs (Nayfach et al., 2019). Mapping only to a subset of specific genes accelerates the profiling while keeping the ambiguity low.

However, all these tools depend on closed (species) databases. Insofar they have the same disadvantages as 16S OTUs that are defined based on a *closed-reference*. Because the reference usually does not contain the exact strain(s) present in the sample, sub-optimal mapping is expected. The presence of multiple strains in a microbiome additionally complexifies quantification (See example in section 1.2.2). Finally, while the quantification on species levels is ideal for comparison with other studies, much of the sub-species diversity is ignored, showing the tradeoff between comparability and specificity again.

The problem can, naturally, be solved by the creation of dataset-specific references through genome-resolved metagenomics. Ideally, one would recover the genome of each strain in a sample. Limitations of the assembly and binning of low abundant genomes make this impossible. Assembly needs minimal coverage to correctly assemble a genome[3] and strains that are more than 98% similar are merged in the same assembly (Fritz et al., 2019).

Combining the genomes recovered from multiple samples allows complementing the catalog of genomes that serve as a reference. For instance, a strain that is low in abundance might have a nearly identical strain that is abundant in another sample. Therefore, collecting the MAGs recovered from all samples of a study allows quantifying low-abundant strains in samples where they are unrecoverable. On the other hand, it does not make sense to include every genome in the reference. The redundancy can create ambiguous results during the quantification as reads will have multiple mapping sites. Therefore MAGs are usually **de-replicated** before they are used as a reference for quantification. Genomes are usually clustered based on their ANI, and the genome with the highest quality is chosen as representative. As an alternative, one might choose the medoid, the genome that represents the cluster the best. Commonly the threshold of 95% average nucleotide identity is used for de-replication.

This is the strategy implemented in `metagenome-atlas` (Ch. 2). Each sample is assembled and binned separately (optionally using differential abundance by mapping the reads from other samples to the assembly). The bins are then de-replicated, and the best genome is chosen for each cluster. The threshold is set

[3] The assembly of a single prokaryote genome needed eight-fold coverage to be entirely assembled in a simulated experiment. (Fritz et al., 2019)

to the common species threshold but can be set lower to capture sub-species variation. This approach creates a *dataset-specific* reference that allows consistent quantification of all the samples. Low-abundant species can still be quantified if they are recovered in at least one sample. The annotation of the species representatives with a standard taxonomy allows comparison beyond the study. Though this method is resource-intensive, it optimizes specificity and comparability, representing similar advantages as ASVs over OTUs.

5.2.2 How *not* to interpret microbiome data?

Regardless of the sequencing technology and workflow used for quantifying a microbiome, the typical output is an abundance table, a table with non-negative values, integers most of the time, also known as *counts*. We obtain counts as the output of most amplicon or metagenomics analysis which might tempt us to interpret them as counting organisms or that the counts somehow relate to the number of organisms in the sample.

Most of the time, we only sequence a sample of the (biological) sample of the microbiome, which distorts the data in many ways. To begin with, the number of reads obtained from different samples can vary in the orders of magnitude. A common practice to account for this difference in **sequencing depth** is to transform the counts into **relative abundance** by dividing by the total number of reads per sample. Relative abundance can be serviceable for visualizing the microbiome but is less so for their analysis. Relative abundance profiles of microbiomes give the impression that each quantified species can be analyzed independently from the other, but that is not the case. Due to the sequencers constraining the reads to a fixed total number, the data is not only relative (to the total number of reads) but also **compositional**. Compositional means that the observed abundance of any given species is dependent on the observed abundance of all other species. For example, the decrease of an abundant member of a microbial community leads to an apparent increase of lower abundant members, which can be even significant.

We had this situation when analyzing the microbiota of children with diarrhea (Kieser, Sarker, et al., 2018). We were surprised by an increase of *Streptococcus* irrespective of the pathogen causing diarrhea, which we related to the flushing out of the majority of the colonic microbiome, which makes *Streptococcus* only

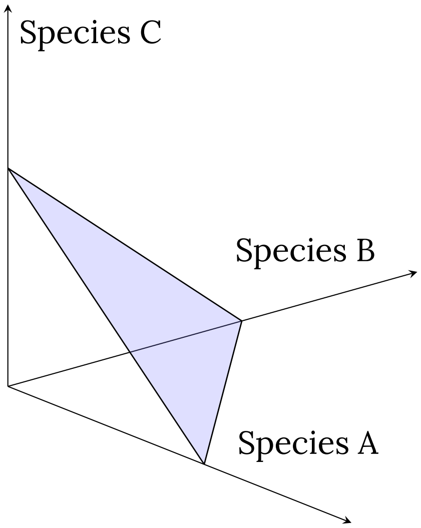


Fig. 5.1 The graph represents the space of possible values of a microbiome quantification can take. The entire space for absolute quantifications of a microbiome and the shaded triangle for compositional quantification such as sequencing.

appear to increase. Diarrhea is an obvious example, but the compositionality leads to a *negaitve correlation* between all species that affects the interpretation of microbiome data in more subtle ways.

An interesting way to visualize how compositional data is different from what is expected is to visualize the space values can take (Fig 5.1). Suppose we measure the composition of an example microbiome consisting of three species (A, B, C) with random abundance. In that case, we can represent the measurement as a point in a three-dimensional space. When counting the actual number of an organism, the point can be anywhere in the (non-negative) space. However, when sequenced, the possible values for our measurement are constrained to the **simplex** (shaded triangle in figure 5.1). This area, called a simplex, becomes a hyper-tetraether for higher numbers of species. Analyzing the data as if the whole space are possible values when they are not leads obviously to biased results.

It is important to note that microbiomes can be quantified in absolute terms if a measure for the number of cells is available. In a landmark study Vandeputte et al. use flow cytometric enumeration to quantify the microbial load and multiply it with the relative abundance based on amplicon sequencing (Vandeputte et al., 2017). The authors show convincingly how the number of microorganisms can vary by up to ten-fold in microbiome samples of healthy individuals. Microbiome load can be a fundamental driver of the microbiome alterations in Crohn's disease. Flow cytometric enumeration needs fresh stool samples, which is difficult to obtain for larger study cohorts. However, it was shown that quantitative PCR of the 16S gene could also reliably be used to transform the relative abundance of amplicon sequencing into absolute counts (Tettamanti Boshier et al., 2020).

The fact that microbiome data is compositional is still not widely appreciated. At the beginning of my P.h.D., Gloor et al. published an article entitled "Microbiome Datasets Are Compositional: And This Is Not Optional", where they show how the compositionality of microbiome data affects all areas of data interpretation from ordination, clustering, network analysis to differential (relative) abundance determination. They highlight fatal issues with common approaches for analyzing microbiome data that do not account for its compositionality. For instance, standard statistical tests or even microbiome-specific tools such as LEfSe (Segata, Izard, et al., 2011) give biased results. Similarly, most metrics for microbiome data, such as the UniFrac distance, also suffer from this problem. In the same article, the authors also show that there are tools that explicitly take the compositionality of the data into account and can make the interpretation of microbiome data more robust.

5.2.3 Compositional data analysis of microbiome data

The core idea of compositional data analysis (CoDa) is to analyze (microbiome) data not as independent abundances but rather as *ratios* between species to describe a sample. The ratios are the same whether the data are counts or proportions. They are also not affected by differences in sequencing depths nor unmapped reads. The idea of compositional data analysis goes back to John Aitchison, who formalized the analysis of data that consists of proportions. He proposed to transform compositional data using the geometric mean. This transformation is called **centered log-ratio** (CLR) and is defined as

$$CLR(\vec{x}) = \log(\vec{x}) - \text{mean}(\log(\vec{x}))$$

For each sample vector $\vec{x}$ that contains either counts or relative abundance. Taking the logarithm of the species abundance transforms them to ratio to the geometric mean of the abundances or the mean of the log-transformed abundances. Because the logarithm of zero is undefined, one needs to impute or estimate the zero values. Fortunately, there are many acceptable ways to deal with the zeros in sequencing data. The most basic approach is to impute a small numeric value; 0.65 is commonly used. Multiplicative replacement adjusts these imputed values to preserve the relative multivariate structure of the data (J. A. Martín-Fernández et al., 2003; Palarea-Albaladejo & Josep Antoni Martín-Fernández, 2015). Counts can be modeled as a probability distribution, where zero is a possible outcome (Fernandes et al., 2014) or estimated using matrix completion Kuczynski et al. Noteworthy, zeros are a minor problem in metagenomics data compared to amplicon data because of spurious mapping of reads to genomes, which makes zeros in count tables very unlikely.

Ordination and multivariate statistics

The CLR.transformation makes the data symmetric and *linearly related*, so that they can be easily be interpreted by traditional statistics and machine-learning algorithms. What is more, the simple difference between two data points becomes a meaningful □-diversity metric, sometimes named **Aitchison distance**. The Aitchison distance is a proper linear distance, making it ideal for multivariate analysis, clustering, and ordination. Ordination refers to the representation of all samples of a study in a single plot. It is usually the first step of an (exploratory) microbiome analysis. The highly dimensional data needs to be transformed into a low-dimensional, preferably two-dimensional, space. However, the dimensional reduction should keep as much as possible of the variability to visually or statistically discriminate between groups and identify outliers. As the typical distance metrics, such as Jensen-Shannon, Bray-Curtis, and UniFrac, are not linear distances, this has to be achieved using a *principal coordinate analysis* (PCoA), which not only takes time to calculate but is also sensitive to inclusion or exclusion of samples (Wong et al., 2016). PCoA, based on standard metrics, discriminates mainly on the most abundant members of a community, which might not affect the most discriminatory species between groups (Gloor et al., 2017; Kuczynski et al., 2010a). Another major problem for the ordination of microbiome datasets is sparsity, the fact that most species are absent from

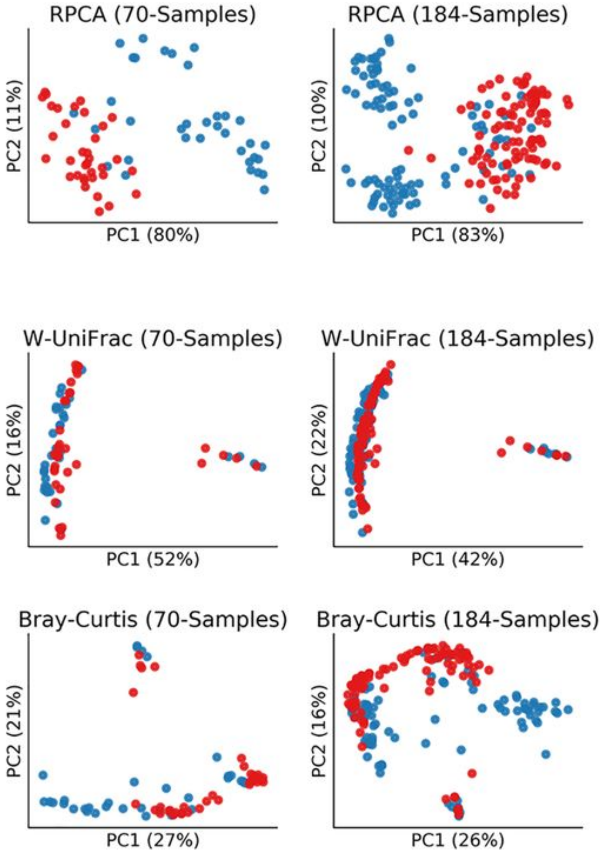


Fig. 5.2 Ordination plots of a dataset Robust Aitchison PCA (RAPCA), PCA based on Aitchison distance with zero-imputation based on matrix completion (Kuczynski et al., 2010a), compared to PCoAs based on the weighted unifrac or the Bray-Curtis distance. RAPCA is able to discriminate bwetewen the two groups (red and blue) even with lower samples (left column). Source (Kuczynski et al., 2010a).

most samples. Sparsity can lead to a distortion of gradients in the ordination plot using standard distant metrics, often referred to as the Horseshoe-effect (Kuczynski et al., 2010b; Morton et al., 2017).

In contrast, the Aitchison distance can be used in a **principal component analysis** (PCA), which is not only swift to calculate, but also robust to sparsity (Wong et al., 2016; Morton et al., 2017), meaning that the exploratory analysis is reproducible even if additional samples are included or outliers excluded (Figure 5.2).Both principal *component* analysis and principal *coordinate* analysis identify the most important components of the data. In addition, PCA identifies which species contribute to which extent to the components; these values are referred to as the **feature loadings**. They allow the direct identification of species that contribute the most to the differences in the datasets, for example, which species contribute the most to differences between clusters separated by the PCA. It is possible to plot the feature loadings on the same plot as the PCA-transformed data, enhancing exploratory data analysis efficiently.

Differencial abundance analysis

CLR transformed abundances can directly be used for differential abundance analysis. Even if the plot of CLR values might seem unalike a plot with relative abundance, its interpretation is intuitive (Figure 5.3). Higher CLR values represent a higher abundance of a species in a sample, whereas negative values represent low abundances. The difference between the mean of the two groups can be intuitively interpreted as the log-fold-change between the two groups. The CLR-transformed values are approximatively normally distributed, which renders the use of the parametric tests justified, which have more power than assessing the significance of relative abundance values using the non-parametrical tests.

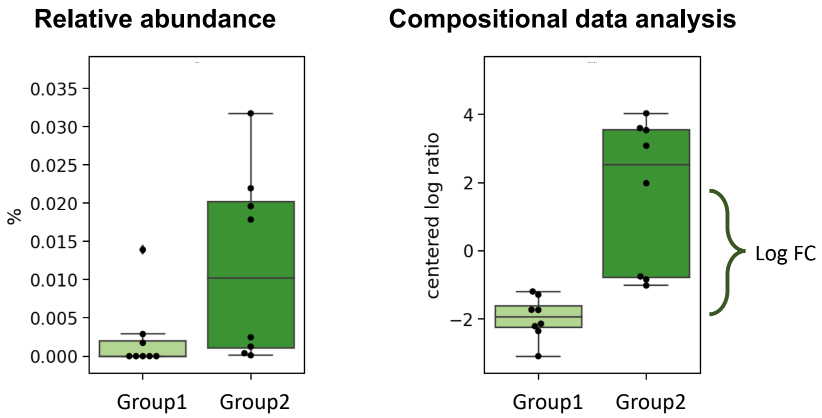


Fig. 5.3 The univariate comparison of the abundance of a species both as relative abundance and under the compositional data analysis paradigm. The significance of the relative abundance values is assessed using a Mann–Whitney U test, whereas a Welch test was used for the centered log-ratios (CLR). The difference between the average CLR of the groups is the log fold change.

Differential abundance analysis is often the final goal of microbiome analysis. After removing outliers and checking the data for inhomogeneities, one would like to identify which microbes are most significantly different between the two groups.

It is crucial to keep in mind that the CLR transformation is not an actual normalization (Quinn, Erb, et al., 2018). It does not remove the constraint and the correlation bias between the different species' abundances in a sample. The CLR-transformed data can give the impression that they refer to single species. However, the transformed data refer to the ratios of the species abundance to the geometric mean of the abundance. The geometric mean can change with the inclusion or removal of species. Therefore, slight variations of the CLR are proposed to mitigate this dependence by taking the geometric mean of all the values in the interquartile range (Wu et al., 2017).

Ratio based biomarker discovery

A more general approach is to calculate ratios between species or taxa directly. For instance, the ratio between the phyla *Bacteroidetes* and *Firmicutes* or between the genera *Prevotella* and *Bacteroides* are well known exampled used in the field. Even when their associations with specific host phenotypes is put into question (Magne et al., 2020), these ratios represent an example of robust biomarkers that can be used to characterize microbiomes even across studies[4]. Instead of calculating all ratios between different taxa and manually testing their association with the metadata (e.g., groups), more sophisticated methods are available that search more efficiently and control the false discovery rate.

For instance, the tool `phylofactor` generalizes this idea by looking for searches for the optimal split in a phylogenetic tree to create a ratio, or balance, with the strongest association with the metadata (Washburne et al., 2019). It can identify broad clades of species that account for maximum variation or more specific clades that are most significantly associated with the metadata. This framework can build models based on multiple metadata variables. Phylofactor can identify differences between clades anywhere in the phylogenetic tree and is not limited to the annotated taxonomic levels. For instance, it can be used on taxonomically unannotated OTUs or genomes.

[4] However it is important to standardize the extraction protocol, which is an important confounding factor for the Bacteroidetes-Firmicutes-ratio (Magne et al., 2020).

However, like all phylogenetic analyses, it is also constrained by the tree. Tools like `codacore` (Gordon-Rodriguez et al., 2021) can identify ratios of any combinations of species associated with the metadata. Species can be either summed, similarly as one would add the abundance of all species that exhibit the same pathway, or multiplied to capture multiplicative interactions. The more species are combined to create a ratio, the better the ratio is associated with the metadata. However, to limit over-fitting and to make the ratio more interpretable fewer species are desired. A parameter in `codacore` allows tuning how many species should be included to create a ratio resulting in easy-interpretable and robust biomarkers.

This novel machine learning framework termed **ratio based biomarker** analysis (Quinn, Gordon-Rodriguez, et al., 2021), fully accounts for the biases in (microbiome) sequencing data and allows the integration of multiple omics data. This framework presents many exciting opportunities for thorough measuring and analysis of microbiome data, especially in a post-assembly area.

References

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, Jeremy Zucker, Mathangi Thiagarajan, Bernard Henrissat, Owen White, Scott T. Kelley, Barbara Methé, Patrick D. Schloss, Dirk Gevers, Makedonka Mitreva & Curtis Huttenhower (2012). "Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome". In: *PLoS Computational Biology* 8.6. Ed. by Jonathan A. Eisen, e1002358. DOI: 10.1371/journal.pcbi.1002358.

Aitchison, John (1982). "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160. DOI: 10.1111/j.2517-6161.1982.tb01195.x.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson & Per H Nielsen (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes". In: *Nature Biotechnology* 31.6, pp. 533–538. DOI: 10.1038/nbt.2579.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley & Robert D. Finn (2019). "A new genomic blueprint of the human gut microbiota". In: *Nature* 568.7753, pp. 499–504. DOI: 10.1038/s41586-019-0965-1.

Anderson, Stephen (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments". In: *Nucleic Acids Research* 9.13, pp. 3015–3027. DOI: 10.1093/nar/9.13.3015.

Belcour, Arnaud, Clémence Frioux, Méziane Aite, Anthony Bretaudeau, Falk Hildebrand & Anne Siegel (2020). "Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species". In: *eLife* 9. DOI: 10.7554/eLife.61968.

Bowers, Robert M et al. (2017). "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea". In: *Nature Biotechnology* 35.8, pp. 725–731. DOI: 10.1038/nbt.3893.

Brown, C. Titus, Dominik Moritz, Michael P. O'Brien, Felix Reidl, Taylor Reiter & Blair D. Sullivan (2020). "Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity". In: *Genome Biology* 21.1, p. 164. DOI: 10.1186/s13059-020-02066-4.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H.

Williams & Jillian F. Banfield (2015). “Unusual biology across a group comprising more than 15% of domain Bacteria”. In: *Nature* 523.7559, pp. 208–211. DOI: 10.1038/nature14486.

Callahan, Benjamin J., Paul J. McMurdie & Susan P. Holmes (2017). “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis”. In: *ISME Journal* 11.12, pp. 2639–2643. DOI: 10.1038/ismej.2017.119.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes (2016). “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature Methods* 13.7, pp. 581–583. DOI: 10.1038/nmeth.3869.

Caspi, Ron, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver & Peter D. Karp (2016). “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases”. In: *Nucleic Acids Research* 44.D1, pp. D471–D480. DOI: 10.1093/nar/gkv1164.

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren & Jillian F. Banfield (2020). “Accurate and complete genomes from metagenomes”. In: *Genome Research* 30.3, pp. 315–333. DOI: 10.1101/gr.258640.119.

Chevalier, Claire, Silas Kieser, Melis Çolakoğlu, Noushin Hadadi, Julia Brun, Dorothée Rigo, Nicolas Suárez-Zamorano, Martina Spiljar, Salvatore Fabbiano, Björn Busse, Julijana Ivanišević, Andrew Macpherson, Nicolas Bonnet & Mirko Trajkovski (2020). “Warmth Prevents Bone Loss Through the Gut Microbiota”. In: *Cell Metabolism* 32.4, 575–590.e7. DOI: 10.1016/j.cmet.2020.08.012.

Costea, Paul I, Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, Almagul Kushugulova, Georg Zeller & Peer Bork (2017). “Subspecies in the global human gut microbiome”. In: *Molecular Systems Biology* 13.12, p. 960. DOI: 10.15252/msb.20177589.

Dijk, Erwin L. van, Yan Jaszczyszyn, Delphine Naquin & Claude Thermes (2018). “The Third Revolution in Sequencing Technology”. In: *Trends in Genetics* 34.9, pp. 666–681. DOI: 10.1016/j.tig.2018.05.008.

Edgar, Robert C (2018). “Updating the 97% identity threshold for 16S ribosomal RNA OTUs”. In: *Bioinformatics* 34.14. Ed. by Alfonso Valencia, pp. 2371–2375. DOI: 10.1093/bioinformatics/bty113.

Eren, A. Murat & Tom O. Delmont (2017). *Predicting CPR genomes in metagenomic bins – Meren Lab.*

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin & Tom O. Delmont (2015). “Anvi’o: an advanced analysis and visualization platform for 'omics data". In: *PeerJ* 3, e1319. DOI: 10.7717/peerj.1319.

Fernandes, Andrew D, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell & Gregory B Gloor (2014). "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis." In: *Microbiome* 2, p. 15. DOI: 10.1186/2049-2618-2-15.

Fernández, Lucía, Ana Rodríguez & Pilar García (2018). "Phage or foe: an insight into the impact of viral predation on microbial communities". In: *The ISME Journal* 12.5, pp. 1171–1179. DOI: 10.1038/s41396-018-0049-5.

Fleischmann, Robert D. et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223, pp. 496–512. DOI: 10.1126/science.7542800.

Franzosa, Eric A, Lauren J. McIver, Gholamali Rahnavard, Luke R Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, Rob Knight, J Gregory Caporaso, Nicola Segata & Curtis Huttenhower (2018). "Species-level functional profiling of metagenomes and metatranscriptomes". In: *Nature Methods* 15.11, pp. 962–968. DOI: 10.1038/s41592-018-0176-y.

Fritz, Adrian, Peter Hofmann, Stephan Majda, Eik Dahms, Johannes Dröge, Jessika Fiedler, Till R. Lesker, Peter Belmann, Matthew Z. Demaere, Aaron E. Darling, Alexander Sczyrba, Andreas Bremges & Alice C. McHardy (2019). "CAMISIM: Simulating metagenomes and microbial communities". In: *Microbiome* 7.1, p. 17. DOI: 10.1186/s40168-019-0633-6.

Glendinning, Laura, Robert D. Stewart, Mark J. Pallen, Kellie A. Watson & Mick Watson (2020). "Assembly of hundreds of novel bacterial genomes from the chicken caecum". In: *Genome Biology* 21.1, p. 34. DOI: 10.1186/s13059-020-1947-1.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn & Juan J. Egozcue (2017). "Microbiome Datasets Are Compositional: And This Is Not Optional". In: *Frontiers in Microbiology* 8, p. 2224. DOI: 10.3389/fmicb.2017.02224.

Goldenfeld, Nigel & Norman R. Pace (2013). *Carl R. Woese (1928-2012)*. DOI: 10.1126/science.1235219.

Gordon-Rodriguez, Elliott, Thomas P Quinn & John P Cunningham (2021). "Learning Sparse Log-Ratios for High-Throughput Sequencing Data". In: *bioRxiv*. DOI: 10.1101/2021.02.11.430695.

Gruber-Vodicka, Harald R., Brandon K. B. Seah & Elmar Pruesse (2020). "phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes". In: *mSystems* 5.5. Ed. by Mani Arumugam. DOI: 10.1128/mSystems.00920-20.

Guarner, Francisco & Juan-R Malagelada (2003). "Gut flora in health and disease". In: *The Lancet* 361.9356, pp. 512–519. DOI: 10.1016/S0140-6736(03)12489-0.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy & Robert M. Goodman (1998). “Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products”. In: *Chemistry & Biology* 5.10, R245–R249. DOI: 10.1016/S1074-5521(98)90108-9.

Heather, James M. & Benjamin Chain (2016). *The sequence of sequencers: The history of sequencing DNA*. DOI: 10.1016/j.ygeno.2015.11.003.

Hungate, R E (1944). “Studies on Cellulose Fermentation: I. The Culture and Physiology of an Anaerobic Cellulose-digesting Bacterium.” In: *Journal of bacteriology* 48.5, pp. 499–513. DOI: 10.1128/JB.48.5.499-513.1944.

Johnson, Jethro S., 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 & George M. Weinstock (2019). “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis”. In: *Nature Communications* 10.1, p. 5029. DOI: 10.1038/s41467-019-13036-1.

Josse, John, A D Kaiser & Arthur Kornberg (1961). *Enzymatic Synthesis of Deoxyribonucleic Acid VIII. FREQUENCIES OF NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACID*. Tech. rep. 3, pp. 864–875. DOI: 10.1016/S0021-9258(18)64321-2.

Kanehisa, M & S Goto (2000). “KEGG: kyoto encyclopedia of genes and genomes.” In: *Nucleic acids research* 28.1, pp. 27–30.

Karst, Søren M, Morten S Dueholm, Simon J McIlroy, Rasmus H Kirkegaard, Per H Nielsen & Mads Albertsen (2018). “Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias”. In: *Nature Biotechnology* 36.2, pp. 190–195. DOI: 10.1038/nbt.4045.

Kashtan, N., S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, M. J. Follows, R. Stepanauskas & S. W. Chisholm (2014). “Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus”. In: *Science* 344.6182, pp. 416–420. DOI: 10.1126/science.1248575.

Kieser, Silas, Joseph Brown, Evgeny M. Zdobnov, Mirko Trajkovski & Lee Ann McCue (2020). “ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data”. In: *BMC Bioinformatics* 21.1, p. 257. DOI: 10.1186/s12859-020-03585-4.

Kieser, Silas, Shafiqul A. Sarker, Olga Sakwinska, Francis Foata, Shamima Sultana, Zeenat Khan, Shoheb Islam, Nadine Porta, Séverine Combremont, Bertrand Betrisey, Coralie Fournier, Aline Charpagne, Patrick Descombes, Annick Mercenier, Bernard Berger & Harald Brüssow (2018). “Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology”. In: *Environmental Microbiology* 20.6, pp. 2256–2269. DOI: 10.1111/1462-2920.14274.

Kieser, Silas, Evgeny M Zdobnov & Mirko Trajkovski (2021). “Comprehensive mouse gut metagenome catalog reveals major difference to the human counterpart". In: *bioRxiv*, p. 2021.03.18.435958. DOI: 10.1101/2021.03.18.435958.

Kriventseva, Evgenia V, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A Simão & Evgeny M Zdobnov (2019). "OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs". In: *Nucleic Acids Research* 47.D1, pp. D807–D811. DOI: 10.1093/nar/gky1053.

Kuczynski, Justin, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Noah Fierer & Rob Knight (2010a). "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns". In: *Nature Methods* 7.10. Ed. by Josh D. Neufeld, pp. 813–819. DOI: 10.1038/nmeth.1499.

– (2010b). "Microbial community resemblance methods differ in their ability to detect biologically relevant patterns". In: *Nature Methods* 7.10, pp. 813–819. DOI: 10.1038/nmeth.1499.

Lagkouvardos, Ilias et al. (2016). "The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota". In: *Nature Microbiology* 1.10, p. 16131. DOI: 10.1038/nmicrobiol.2016.131.

Lane, D J, B Pace, G J Olsen, D A Stahl, M L Sogin & N R Pace (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses." In: *Proceedings of the National Academy of Sciences of the United States of America* 82.20, pp. 6955–9. DOI: 10.1073/pnas.82.20.6955.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane & Tak-Wah Lam (2015). "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph". In: *Bioinformatics* 31.10, pp. 1674–1676. DOI: 10.1093/bioinformatics/btv033.

Lucas Paoli et al. (2021). "Uncharted biosynthetic potential of the ocean microbiome". In: *bioRxiv*. DOI: 10.1101/2021.03.24.436479.

Machado, Daniel, Sergej Andrejev, Melanie Tramontano & Kiran Raosaheb Patil (2018). "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities". In: *Nucleic Acids Research* 46.15, pp. 7542–7553. DOI: 10.1093/nar/gky537.

Machado, Daniel, Oleksandr M. Maistrenko, Sergej Andrejev, Yongkyu Kim, Peer Bork, Kaustubh R. Patil & Kiran R. Patil (2021). "Polarization of microbial communities between competitive and cooperative metabolism". In: *Nature Ecology & Evolution* 5.2, pp. 195–203. DOI: 10.1038/s41559-020-01353-4.

Magne, Fabien, Martin Gotteland, Lea Gauthier, Alejandra Zazueta, Susana Pesoa, Paola Navarrete & Ramadass Balamurugan (2020). "The Firmicutes / Bacteroidetes Ratio: A Relevant Marker of Gut Dysbiosis in Obese Patients?" In: *Nutrients* 12.5. DOI: 10.3390/nu12051474.

Mallawaarachchi, Vijini, Anuradha Wickramarachchi & Yu Lin (2020). "GraphBin: refined binning of metagenomic contigs using assembly graphs". In:

Bioinformatics 36.11. Ed. by Alfonso Valencia, pp. 3307–3313. DOI: 10.1093/bioinformatics/btaa180.

Martín-Fernández, J. A., C. Barceló-Vidal & V. Pawlowsky-Glahn (2003). "Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation". In: *Mathematical Geology* 35.3, pp. 253–278. DOI: 10.1023/A:1023866030544.

McDonald, Daniel, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight & Philip Hugenholtz (2012). "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea". In: *The ISME Journal* 6.3, pp. 610–618. DOI: 10.1038/ismej.2011.139.

McMahon, Katherine (2015). "'Metagenomics 2.0'". In: *Environmental Microbiology Reports* 7.1, pp. 38–39. DOI: 10.1111/1758-2229.12253.

Medema, Marnix H. et al. (2015). "Minimum Information about a Biosynthetic Gene cluster". In: *Nature Chemical Biology* 11.9, pp. 625–631. DOI: 10.1038/nchembio.1890.

Milanese, Alessio, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, Renato Alves, Paul I Costea, Luis Pedro Coelho, Thomas S. B. Schmidt, Alexandre Almeida, Alex L Mitchell, Robert D. Finn, Jaime Huerta-Cepas, Peer Bork, Georg Zeller & Shinichi Sunagawa (2019). "Microbial abundance, activity and population genomic profiling with mOTUs2". In: *Nature Communications* 10.1, p. 1014. DOI: 10.1038/s41467-019-08844-4.

Morton, James T., Liam Toran, Anna Edlund, Jessica L. Metcalf, Christian Lauber & Rob Knight (2017). "Uncovering the Horseshoe Effect in Microbial Analyses". In: *mSystems* 2.1, pp. 166–182. DOI: 10.1128/msystems.00166-16.

Myers, Eugene W. (1995). "Toward Simplifying and Accurately Formulating Fragment Assembly". In: *Journal of Computational Biology* 2.2, pp. 275–290. DOI: 10.1089/cmb.1995.2.275.

Myers, Eugene W. et al. (2000). "A Whole-Genome Assembly of Drosophila". In: *Science* 287.5461, pp. 2196–2204. DOI: 10.1126/science.287.5461.2196.

Nair, Prashant (2012). "Woese and Fox: Life, rearranged." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.4, pp. 1019–21. DOI: 10.1073/pnas.1120749109.

Nasko, Daniel J., Sergey Koren, Adam M. Phillippy & Todd J. Treangen (2018). "RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification". In: *Genome Biology* 19.1, pp. 1–10. DOI: 10.1186/s13059-018-1554-6.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard & Nikos C. Kyrpides (2019). "New insights from uncultivated genomes of the global human gut microbiome". In: *Nature* 568.7753, pp. 505–510. DOI: 10.1038/s41586-019-1058-x.

Nielsen, Henrik Bjørn et al. (2014). "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes". In: *Nature Biotechnology* 32.8, pp. 822–828. DOI: 10.1038/nbt.2939.

Nissen, Jakob Nybo, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, Henrik Bjørn Nielsen, Thomas Nordahl Petersen, Ole Winther & Simon Rasmussen (2021). "Improved metagenome binning and assembly using deep variational autoencoders". In: *Nature Biotechnology*. DOI: 10.1038/s41587-020-00777-4.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov & Pavel A. Pevzner (2017). "metaSPAdes: a new versatile metagenomic assembler". In: *Genome Research* 27.5, pp. 824–834. DOI: 10.1101/gr.213959.116.

Nussinov, Ruth (1980). "Some rules in the ordering of nucleotides in the DNA". In: *Nucleic Acids Research* 8.19, pp. 4545–4562. DOI: 10.1093/nar/8.19.4545.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren & Adam M. Phillippy (2016). "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1, p. 132. DOI: 10.1186/s13059-016-0997-x.

Orakov, Askarbek, Anthony Fullam, Luis Pedro Coelho, Supriya Khedkar, Damian Szklarczyk, Daniel R. Mende, Thomas S.B. Schmidt & Peer Bork (2020). *GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes*. DOI: 10.1101/2020.12.16.422776.

Palarea-Albaladejo, Javier & Josep Antoni Martín-Fernández (2015). "ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach". In: *Chemometrics and Intelligent Laboratory Systems* 143, pp. 85–96. DOI: 10.1016/j.chemolab.2015.02.019.

Parks, Donovan H., Maria Chuvochina, Pierre-alain Chaumeil, Christian Rinke, Aaron J. Mussig & Philip Hugenholtz (2020). "A complete domain-to-species taxonomy for Bacteria and Archaea". In: *Nature Biotechnology* 38.9, pp. 1079–1086. DOI: 10.1038/s41587-020-0501-8.

Parks, Donovan H, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz (2018). "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". In: *Nature Biotechnology* 36.10, pp. 996–1004. DOI: 10.1038/nbt.4229.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz & Gene W. Tyson (2017). "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life". In: *Nature Microbiology* 2.11, pp. 1533–1542. DOI: 10.1038/s41564-017-0012-7.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, Maria Carmen Collado, Benjamin L Rice, Casey DuLong, Xochitl C Morgan, Christopher D Golden, Christopher Quince, Curtis Huttenhower & Nicola Segata (n.d.). “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle”. In: *Cell* 176.3, 649–662.e20. DOI: 10.1016/j.cell.2019.01.001.

Pevzner, P. A., H. Tang & M. S. Waterman (2001). “An Eulerian path approach to DNA fragment assembly”. In: *Proceedings of the National Academy of Sciences* 98.17, pp. 9748–9753. DOI: 10.1073/pnas.171285098.

Qin, Junjie et al. (2010). “A human gut microbial gene catalogue established by metagenomic sequencing”. In: *Nature* 464.7285, pp. 59–65. DOI: 10.1038/nature08821.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies & Frank Oliver Glöckner (2012). “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools”. In: *Nucleic Acids Research* 41.D1, pp. D590–D596. DOI: 10.1093/nar/gks1219.

Quinn, Thomas P, Ionas Erb, Mark F Richardson & Tamsyn M Crowley (2018). “Understanding sequencing data as compositions: an outlook and review”. In: *Bioinformatics* 34.16. Ed. by Jonathan Wren, pp. 2870–2878. DOI: 10.1093/bioinformatics/bty175.

Quinn, Thomas P, Elliott Gordon-Rodriguez & Ionas Erb (2021). “A Critique of Differential Abundance Analysis, and Advocacy for an Alternative”. In:

Richardson, Lorna J, Neil D Rawlings, Gustavo A Salazar, Alexandre Almeida, David R Haft, Gregory Ducq, Granger G Sutton & Robert D Finn (2019). “Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes”. In: *Nucleic Acids Research* 47.D1, pp. D564–D572. DOI: 10.1093/nar/gky1013.

Sandberg, Rickard, Gösta Winberg, Carl Ivar Brändén, Alexander Kaske, Ingemar Ernberg & Joakim Cöster (2001). “Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier”. In: *Genome Research* 11.8, pp. 1404–1409. DOI: 10.1101/gr.186401.

Sanger, F., S. Nicklen & A. R. Coulson (1977). “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

Schloss, Patrick D (2021). “ASVs artificially split bacterial genomes Observation Format”. In: *bioRxiv*, p. 2021.02.26.433139. DOI: 10.1101/2021.02.26.433139.

Schloss, Patrick D. & Jo Handelsman (2005). “Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness”. In: *Applied and Environmental Microbiology* 71.3, pp. 1501–1506. DOI: 10.1128/AEM.71.3.1501-1506.2005.

Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett & Curtis Huttenhower (2011). “Metagenomic biomarker discovery and explanation". In: *Genome Biology* 12.6, R60. DOI: 10.1186/gb-2011-12-6-r60.

Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson & Curtis Huttenhower (2012). "Metagenomic microbial community profiling using unique clade-specific marker genes". In: *Nature Methods* 9.8, pp. 811–814. DOI: 10.1038/nmeth.2066.

Seppey, Mathieu, Mosè Manni & Evgeny M. Zdobnov (2019). "BUSCO: Assessing Genome Assembly and Annotation Completeness". In: *Gene Prediction*, pp. 227–245. DOI: 10.1007/978-1-4939-9173-0_14.

Shaiber, Alon & A. Murat Eren (2019). "Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories". In: *mBio* 10.3. Ed. by David A. Relman. DOI: 10.1128/mBio.00725-19.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe & Jillian F. Banfield (2018). "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy". In: *Nature Microbiology* 3.7, pp. 836–843. DOI: 10.1038/s41564-018-0171-1.

Sim, Kathleen, Michael J. Cox, Harm Wopereis, Rocio Martin, Jan Knol, Ming-Shi Li, William O. C. M. Cookson, Miriam F. Moffatt & J. Simon Kroll (2012). "Improved Detection of Bifidobacteria with Optimised 16S rRNA-Gene Based Pyrosequencing". In: *PLoS ONE* 7.3. Ed. by Niyaz Ahmed, e32543. DOI: 10.1371/journal.pone.0032543.

Stackebrandt, E & Brett M. Goebel (1994). "Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology". In: *International Journal of Systematic and Evolutionary Microbiology* 44.4, pp. 846–849. DOI: 10.1099/00207713-44-4-846.

Staden, R. (1979). "A strategy of DNA sequencing employing computer programs". In: *Nucleic Acids Research* 6.7, pp. 2601–2610. DOI: 10.1093/nar/6.7.2601.

Staley, J T & A Konopka (1985). "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats". In: *Annual Review of Microbiology* 39.1, pp. 321–346. DOI: 10.1146/annurev.mi.39.100185.001541.

Stein, J L, T L Marsh, K Y Wu, H Shizuya & E F DeLong (1996). "Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon." In: *Journal of bacteriology* 178.3, pp. 591–599. DOI: 10.1128/JB.178.3.591-599.1996.

Steinegger, Martin & Steven L. Salzberg (2020). "Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank". In: *Genome Biology* 21.1, p. 115. DOI: 10.1186/s13059-020-02023-1.

Stewart, Robert D., Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe & Mick Watson (2019). "Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery". In: *Nature Biotechnology* 37.8, pp. 953–961. DOI: 10.1038/s41587-019-0202-3.

Strathern, Marilyn (1997). "'Improving ratings': audit in the British University system". In: *European Review* 5.03, p. 305. DOI: 10.1017/S1062798700002660.

Sunagawa, Shinichi et al. (2013). "Metagenomic species profiling using universal phylogenetic marker genes". In: *Nature Methods* 10.12, pp. 1196–1199. DOI: 10.1038/nmeth.2693.

Tettamanti Boshier, Florencia A., Sujatha Srinivasan, Anthony Lopez, Noah G. Hoffman, Sean Proll, David N. Fredricks & Joshua T. Schiffer (2020). "Complementing 16S rRNA Gene Amplicon Sequencing with Total Bacterial Load To Infer Absolute Species Concentrations in the Vaginal Microbiome". In: *mSystems* 5.2. Ed. by J. Gregory Caporaso. DOI: 10.1128/mSystems.00777-19.

Tringe, S. G. (2005). "Comparative Metagenomics of Microbial Communities". In: *Science* 308.5721, pp. 554–557. DOI: 10.1126/science.1107851.

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon (2007). "The Human Microbiome Project". In: *Nature* 449.7164, pp. 804–810. DOI: 10.1038/nature06244.

Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar & Jillian F. Banfield (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment". In: *Nature* 428.6978, pp. 37–43. DOI: 10.1038/nature02340.

Vallery-Radot, René (1902). *The life of Pasteur*. New York: Phillips McClure, p. 142.

Vandeputte, Doris, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, Raul Y Tito, Lindsey De Commer, Youssef Darzi, Séverine Vermeire, Gwen Falony & Jeroen Raes (2017). "Quantitative microbiome profiling links gut community variation to microbial load". In: *Nature* 551.7681, pp. 507–511. DOI: 10.1038/nature24460.

Venter, J Craig (2004). "Environmental Genome Shotgun Sequencing of the Sargasso Sea". In: *Science* 304.5667, pp. 66–74. DOI: 10.1126/science.1093857.

– (2006). "Shotgunning the Human Genome: A Personal View". In: *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd. DOI: 10.1038/npg.els.0005850.

Venter, J. Craig, Mark D. Adams, et al. (2001). "The Sequence of the Human Genome". In: *Science* 291.5507, pp. 1304–1351. DOI: 10.1126/science.1058040.

Venter, J. Craig, Hamilton O. Smith & Leroy Hood (1996). "A new strategy for genome sequencing". In: *Nature* 381.6581, pp. 364–366. DOI: 10.1038/381364a0.

Vijay-Kumar, Matam, Benoit Chassaing, Manish Kumar, MarkT Baker & Vishal Singh (2014). "Mammalian gut immunity". In: *Biomedical Journal* 37.5, p. 246. DOI: 10.4103/2319-4170.130922.

Vollmers, John, Sandra Wiegand & Anne-Kristin Kaster (2017). “Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist’s Perspective - Not Only Size Matters!” In: *PLOS ONE* 12.1, e0169662. DOI: 10.1371/journal.pone.0169662.

Washburne, Alex D., Justin D. Silverman, James T. Morton, Daniel J. Becker, Daniel Crowley, Sayan Mukherjee, Lawrence A. David & Raina K. Plowright (2019). “Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data”. In: *Ecological Monographs* 89.2, e01353. DOI: 10.1002/ecm.1353.

Whitman, W. B., D. C. Coleman & W. J. Wiebe (1998). “Prokaryotes: The unseen majority”. In: *Proceedings of the National Academy of Sciences* 95.12, pp. 6578–6583. DOI: 10.1073/pnas.95.12.6578.

Wilson, K H & R B Blitchington (1996). “Human colonic biota studied by ribosomal DNA sequence analysis.” In: *Applied and environmental microbiology* 62.7, pp. 2273–8. DOI: 10.1128/AEM.62.7.2273-2278.1996.

Woese, Carl. R. & George. E. Fox (1977). “Phylogenetic structure of the prokaryotic domain: The primary kingdoms”. In: *Proceedings of the National Academy of Sciences* 74.11, pp. 5088–5090. DOI: 10.1073/pnas.74.11.5088.

Wong, Ruth G., Jia R. Wu & Gregory B. Gloor (2016). “Expanding the UniFrac Toolbox”. In: *PLOS ONE* 11.9. Ed. by Gabriel Moreno-Hagelsieb, e0161196. DOI: 10.1371/journal.pone.0161196.

Wood, Derrick E., Jennifer Lu & Ben Langmead (2019). “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1, p. 257. DOI: 10.1186/s13059-019-1891-0.

Wood, Derrick E & Steven L Salzberg (2014). “Kraken: ultrafast metagenomic sequence classification using exact alignments”. In: *Genome Biology* 15.3, R46. DOI: 10.1186/gb-2014-15-3-r46.

Wu, Jia R., Jean M. Macklaim, Briana L. Genge & Gregory B. Gloor (2017). “Finding the centre: corrections for asymmetry in high-throughput sequencing datasets”. In:

Xiao, Liang et al. (2015). “A catalog of the mouse gut metagenome”. In: *Nature Biotechnology* 33.10, pp. 1103–1108. DOI: 10.1038/nbt.3353.

YONG, ED (2017). *Norm Pace Blew The Door Off The Microbial World.*

Yooseph, Shibu et al. (2007). “The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families”. In: *PLoS Biology* 5.3. Ed. by Sean Eddy, e16. DOI: 10.1371/journal.pbio.0050016.

Zhu, Qiyun, Christopher L. Dupont, Marcus B. Jones, Kevin M. Pham, Zhi-Dong Jiang, Herbert L. DuPont & Sarah K. Highlander (2018). “Visualization-assisted binning of metagenome assemblies reveals potential new pathogenic profiles in idiopathic travelers’ diarrhea”. In: *Microbiome* 6.1, p. 201. DOI: 10.1186/s40168-018-0579-0.