Analysis and assembly methods for microbiome sequencing data

Marcus W. Fedarko[1]

[1]Department of Computer Science and Engineering, University of California San Diego

ABSTRACT

Recent and frequent advances in DNA sequencing technologies and algorithms have enabled the study of microbiomes, populations of microorganisms in environments ranging from the human gut to the Earth's oceans, at higher and higher degrees of resolution. As the quantity and quality of available sequencing data has increased, biologists have become able to ask increasingly targeted questions about the nature of individual microbial samples. Here I provide a broad survey of the study of microbiomes, with a focus on sequencing-based approaches and metagenome sequence assembly. I then discuss next steps for the field, leading to the goal of assembling individual strains of genomes from microbiome sequencing data.

1. INTRODUCTION

1.1 A brief history of microbiome research

Although sequencing technologies have accelerated the study of microbes and microbial communities in recent decades, knowledge of the ubiquity and importance of microbes is centuries old. The field of microbiology dates back to Antonie van Leeuwenhoek's seventeenth century C.E. observation, using microscopes, of "little Animals ... observed in Rain- Well- Sea- and Snow water; as also in water wherein Pepper had lain infused" [67, 65, 108]. The idea that microbes may have an impact in human health is even older: Marcus Terentius Varro wrote in the first century B.C.E. about "certain minute creatures which cannot be seen by the eyes, which float in the air and enter the body through the mouth and nose and there cause serious diseases" [125]. Finally, "yellow soup" containing human fecal matter has a long history of use in traditional Chinese medicine, dating back to a case record written by Hong Ge in the fourth century C.E. [39, 93].

The "yellow soup" example bears close resemblance to what has since been termed *fecal microbiota transplantation* (FMT), in which fecal matter from a donor individual is transplanted to a sick individual's gastrointestinal tract [9, 39]. Centuries later, there is consensus today that FMTs (when performed properly) are effective at treating *Clostridium difficile* infection (CDI) [9, 93]. However, the scope of FMTs' ability to benefit human health is controversial [114]. There is disagreement over exactly which health conditions FMTs can impact [129], and there has been caution against immediate reliance on FMTs as a treatment method for CDI [123]. In any case, the effectiveness of FMTs for treating CDI nonetheless illustrates the potential for microbiome research to improve human health.

1.2 The promise of microbiome research

This potential, along with the advent of high-throughput DNA sequencing [36], has fueled many studies in recent years on how the gut microbiomes of individuals from across the world differ [74], how the gut microbiomes of individuals change over time [17, 23], and relationships between the host gut microbiome and host health [115, 9, 117]. Research has also focused on the microbiomes of other body sites besides the gut: for example, the human skin and oral microbiomes [23]. The *taxonomic compositions* (e.g. "which specific microbes are present in these samples?") of these and other body sites' microbiomes have also been linked with relevant host diseases; for example, variations in skin microbiomes' compositions have been linked to atopic dermatitis [81], and variations in oral microbiomes' compositions have been linked to dental caries [8].

The study of microbiomes has many societal applications besides human health. Detailed knowledge of the particular microbes present within a sample can aid in food safety [94, 30]; biodefense [97, 79]; environmental surveying [27]; and in the study of antibiotic resistance [68, 142, 119], to name a few examples.

1.3 Reckoning with the "hype and hyperbole" of microbiome research

In spite of this potential in the study of microbiomes—for human health and for various other applications—the field has struggled with the ability to reproduce many of its own results [108], and the influx of microbiome studies seen in recent years has not resulted in a correspondingly large amount of clear, reproducible relationships between microbiome taxonomic composition and host disease state [129]. Obvious cases where individual microbes have been reproducibly implicated in certain host diseases, such as *Clostridium difficile* and CDI or *Helicobacter pylori* and various stomach-based diseases [129, 114, 93], are the exception rather than the norm. Counterintuitively, positive results—seemingly novel, compelling findings that link taxonomic composition, diversity measurements, or other microbiome characteristics to the conditions being studied—nonetheless abound in microbiome research: the field, some argue, is plagued by "hype and hyperbole" [129].

There are various reasons for these murky results. Many of these reasons are not especially specific to microbiome re-

search, and cause similar trouble for other areas of science: a few of these classic reasons include limited sample sizes [129], institutional biases against the publication of null results [129, 46], and researchers hypothesizing after a study's data has already been collected without reporting so, also known as "HARKing" or "SHARKing" [50].

These already-challenging problems are complemented by a host of problems specific to microbiome research and related areas of bioinformatics. One of the most obvious of these is uncertainty on how microbiome sequencing data should be generated and computationally analyzed in the first place, and uncertainty on the advantages and disadvantages of the different available methods. Many of these approaches differ widely in their *resolution*, the amount of detail generated about the specific microbes in a microbiome; this is an important factor to consider in study design and data interpretation [45, 57].

Throughout this report I will focus on the goal of improving the resolution with which researchers can study microbiomes. The ability to study individual *strains* of the microbes in a microbiome—that is, all unique microbial genomes present in a sample, rather than aggregate information about species, genera, etc.—will help in understanding the biological functions of these microbes [127, 113]. This is because different strains of the same species of microbe can have drastically different phenotypes [127, 103, 113], which may explain some of the aforementioned contradictory results endemic to microbiome research. In the next section I examine one particularly common class of methods used for studying microbiomes, with an eye on these methods' resolution.

2. CULTURE-INDEPENDENT METHODS FOR SEQUENCING MICROBIOMES

Conventional methods of studying microbiomes generally involve the isolation and culturing of individual microbes from a sample. However, researchers thus far have been unable to culture the vast majority (estimated at over 99%) of microbes [51]. Furthermore, the process of culturing can result in unintended changes to the microbe being cultured due to adaptive evolution [128]. These problems complicate the study of essentially all microbiomes.

Culture-independent methods, in which the microbes in a sample are studied by sequencing their DNA *en masse* rather than by attempting direct culturing, can circumvent these problems [51, 95, 30, 128]. Although not all microbes are easily culturable, all have genomes: and by extracting DNA from a sample and sequencing some or all of this DNA, researchers can study the microbes present in a sample without many of the limitations of traditional culture-based methods. Culture-independent methods open the door for identifying a larger number of microbes present in a sample than would be available through culturing, and also provide the ability to process many samples rapidly using high-throughput DNA sequencing techniques. However, these methods also introduce many biological and computational challenges—and have their own biases and shortcomings—that need to be overcome and acknowledged in order to use them effectively.

Various culture-independent methods have been introduced over the years; early methods include terminal restriction fragment length polymorphism, denaturing gradient gel electrophoresis, and fluorescence *in situ* hybridization [109, 44, 30]. Here, I focus on two more relatively recent methods which have largely superseded these older ones, at least from the perspective of studying microbes within microbiomes. The first of these modern, popular culture-independent methods is *marker gene sequencing* (also referred to as amplicon sequencing, metabarcoding, etc.), and the second is *metagenome sequencing* (also referred to as *metagenomics*, *shotgun metagenome sequencing*, *whole metagenome sequencing*, etc.) [75]. Both approaches generate sequencing data ("reads" of DNA) describing the microbial genomes in a metagenome, and standard pipelines for processing both types of data often produce abundance table(s) of samples by features [38]. However, beyond these surface-level similarities, the details, potential uses, and limitations of these data vary substantially between methods [38].

2.1 Marker gene sequencing

In 1977, Carl Woese and George Fox announced the discovery of the archaea, a distinct group of microbial life [135]. Woese and Fox made this discovery using analysis of ribosomal RNA ("rRNA") gene sequences, which are shared by the genomes of bacteria, archaea, and eukaryotes: the subtle differences between these gene sequences in different organisms' genomes made it possible for Woese and Fox to propose an evolutionary history in which the already-accepted bacterial and eukaryotic domains of life were complemented by a third domain, the archaea [135]. Despite initial confusion and backlash from the scientific community, in part due to articles from the general press featuring dubious headlines like "Martianlike Bugs May Be Oldest Life" [102], the truth bore out: today the existence of the archaea is widely accepted, and—in a related turn of events—certain rRNA gene sequences are commonly used as *marker genes* to study the microbes present in a microbiome [38, 137].

Useful marker genes, including certain rRNA genes, tend to have a specific structure that includes *hypervariable* (high mutation rate) region(s) surrounded by *conserved* (low mutation rate) regions [111]. The conserved regions in the gene enable the design of primers for use in a polymerase chain reaction (PCR) procedure, which can be used to selectively amplify specific region(s) of the marker gene sequence for the genomes present within a sample that contain this gene [84, 137]. The primers can be designed to "target" these conserved regions, which—since they have low mutation rates—should be similar or identical across the many copies of the gene observed in a population of microbes. The hypervariable regions inside the gene sequence, on the other hand, are useful for the opposite reason: the high mutation rates in these regions make it possible to compare the amplified gene sequences from different microbes, which usually show obvious differences in these hypervariable regions for evolutionarily distant microbes. In essence, marker genes are useful for studying the taxonomic composition of microbiomes for the same reasons they were useful to Woese and Fox in studying the evolutionary history of life.

The 16S rRNA gene is arguably the most common marker gene in use today. This gene is specific to bacteria and archaea [138] and contains nine hypervariable regions [137,

126]; different studies have sequenced some or all of these regions [126, 22]. Whether using the 16S rRNA gene or another marker gene [138], marker gene sequencing enables sequencing the genetic material of the microbes in a sample without having to sequence the full genomes of these microbes. For context, the full 16S rRNA gene is roughly 1,550 nucleotides long [28], whereas bacterial genomes generally range from 100,000 to 15,000,000 nucleotides in length [91]: sequencing some or all of the 16S rRNA gene thus represents a drastic reduction in the amount of sequencing necessary to perform in order to profile a single microbe, let alone thousands. In this way, marker gene sequencing provides a broad taxonomic overview of the types of microbes present in a sample [38].

Although marker gene sequencing is relatively convenient and cost-effective, its output information is inherently incomplete. Many marker genes are only specific to certain types of microbes: although rRNA genes of some kind are shared by bacteria, archaea, and eukaryotes [135], the 16S rRNA gene is limited to bacteria and archaea and the 18S rRNA gene is limited to eukaryotes [138]. Furthermore, despite being eukaryotes, fungi are often profiled using other markers such as the internal transcribed spacer (ITS) region: this is because fungi have fewer hypervariable regions in the 18S rRNA gene than other eukaryotes, making the 18S rRNA gene less useful as a marker for fungi [38, 111].[1] The results of a single marker gene sequencing study are thus limited, as compared to metagenome sequencing studies which can capture information about microbes from all domains of life without reliance on shared marker regions (see section 2.2).

The limitations of marker gene sequencing go beyond specificity to certain types of microbes. Many of these are inherent limitations that have been well-known since the early days of marker gene sequencing [109]. Briefly, some organisms may have multiple copies of a marker gene sequence [56]; errors in the PCR process can introduce "chimeric" gene sequences [52]; and there is ongoing disagreement as to the best-practice way to correct sequencing errors in the raw read data generated by marker gene sequencing [21, 110, 57].[2]

From the perspective of studying strain-level genomes of the microbes within a microbiome, as this report has set out to do, perhaps the most damning limitation of marker gene sequencing is its lack of resolution. The marker gene sequences of different strains of the same species, or even of different species within the same genus, are often indistinguishable [69]; it is thus standard practice to ignore taxonomic classifications more specific than the genus level when working solely with 16S rRNA marker gene sequencing data [34].[3] Notably, this is not a problem for many uses of marker gene sequencing; the method is useful for profiling the taxonomic compositions of microbiomes at a broad level, and marker gene sequencing projects are currently far less expensive and require far less computational effort than comparable metagenome sequencing projects [109].

Regardless, since the goal of this report is studying microbial genomes in detail, I turn next to metagenome sequencing. This is another culture-independent method that can provide more detailed resolution of the microbes in a microbiome.

2.2 Metagenome sequencing

In metagenome sequencing, all DNA in a sample is randomly fragmented and then sequenced [104]. This is in contrast to marker gene sequencing, in which only the amplified DNA sequences in a sample (the marker gene, or certain region(s) of the marker gene) are sequenced.

Although many software pipelines and best-practice recommendations for processing marker gene sequencing data have, to an extent, been refined over the years [109, 57], metagenome sequencing research is somewhat less standardized [57]. This is due in part to the many choices available to researchers analyzing metagenome sequencing data, as well as the often wildly different characteristics of different metagenome sequencing datasets.

Since the reads produced by metagenome sequencing are comprised of DNA from throughout all parts of the genomes in a sample, not just from specific marker regions, metagenome sequencing provides researchers the opportunity to identify genes, promoters, operons, and other genomic elements from the sequencing data [54, 118]. Metagenome sequencing therefore enables *functional annotation* of a microbiome [136, 109].

The exact definition of the term "function" is controversial in biology [49]. In the context of metagenomes, the term "functional annotation" generally refers to an attempt to study the practical effects of genes, operons, metabolic pathways, and other potentially-but-not-necessarily "active" biological units in a sample [45, 136, 38]. This is often done by comparing sequences to a reference database with information on function in known organisms [38]; recent years have also seen the development of metabolic modelling techniques that attempt to provide a more realistic estimate of functional activity in a microbiome [45, 37].

In any case, functional annotation can provide important context about metagenome sequencing data—for example, although taxonomic compositions vary noticeably between microbiomes from different human body sites, functional annotation shows that metabolic pathway relative abundances seem very stable across these body sites' microbiomes [53].

[1] The ITS region is not technically a single gene, so calling it a "marker gene" is a misnomer [38]. However, the ITS region is nonetheless useful as a marker for fungi [111], and is used in essentially the same way as "true" marker genes [138].

[2] There are of course additional limitations (and benefits) of marker gene sequencing besides these. Reference [38] provides a thorough overview of the differences between marker gene and metagenome sequencing.

[3] An example taxonomic classification for a given 16S rRNA gene sequence or region thereof might look something like `k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus; s__brevis` [18]. In some cases (depending on the sequencing technology, the marker gene, the taxonomic classification method, etc.) a researcher may be confident enough to say that this gene sequence originates from *Lactobacillus brevis*, but often it is safer to simply say that it is from the genus *Lactobacillus*.

This result implies that, although two microbiomes might be populated by very different types of microbes, the overall functional activities of both microbiomes may be nonetheless similar.

In spite of this promise, many functional annotation methods—including the method [1] used to determine the aforementioned result in [53]—have been criticized as so-called *bag-of-genes* approaches that assume unrealistically that all genes in a metagenome act independently and not as units of individual genomes [131, 45]. Functional annotation can thus be improved by attempting to reconstruct the microbial genomes present in a metagenomic sample, and then performing functional annotation that takes these genomes into account, rather than just considering raw reads [45].

The process of reconstructing a genome, or at least longer fragments of a genome, from the reads produced by a sequencing machine is referred to as *assembly*. Entire microbial genomes are rarely recovered completely in individual reads of DNA: as is described in section 3, a large amount of computational work goes into the process of assembling longer sequences from raw reads [86].[4]

Once reads have been generated (and optionally processed using various quality control methods), metagenome sequencing studies can make use of *assembly-based* or *read-based* computational methods to study the taxonomic and functional characteristics of a sample [104]. Although performing assembly is not a prerequisite for many methods, longer DNA sequences generally contain more information, so beginning an analysis with assembly is desirable if possible [136, 57].

I note that, to an extent, functional information is possible to predict using marker gene sequencing data [66]: however, the greater resolution afforded by metagenome sequencing generally makes metagenome functional annotations far more valuable [45], and predictive methods for marker gene sequencing data have been shown to perform poorly on samples from less-well-studied microbiomes [122].

Compared to marker gene sequencing, metagenome sequencing affords researchers the theoretical ability to assemble individual microbial genomes from a microbiome, and in so doing study and compare microbiomes in detail. In order to come closer to this goal, I next focus on the process of metagenome assembly.

3. (META)GENOME ASSEMBLY

Unfortunately for this report's goal, metagenome assembly is a far more difficult problem than *genome assembly*. Genome assembly (also known as single-genome assembly) involves assembling the sequence of a single genome that has been isolated and cultured [136, 58]; metagenome assembly, in which the input is many similar and different genomes from an uncultured sample, presents many unique problems [87, 90, 58].

[4]That being said, the arrival of sequencing technologies that can generate increasingly long reads (as discussed in section 3.1) brings the field closer and closer to this goal, especially for relatively short genomes [60].

In order to understand the unique challenges posed by metagenome assembly, I will examine the general structure of the assembly problem. I begin by discussing the input sequencing data, and how different types of sequencing data can complicate or simplify assembly; I then discuss the general outputs of assembly, with a focus on how these outputs are often used when performing metagenome assembly. Finally, I conclude this section by examining various types of assembly methods.

3.1 Assembly input: reads

The output of a DNA sequencing machine, and the main input to an assembly tool, is a collection of *reads*.[5,6] Each read can be thought of as a string from the alphabet $\{A, C, G, T\}$, corresponding respectively to the four nucleotides adenine, cytosine, guanine, and thymine.[7] Reads are often accompanied by corresponding numeric values indicating *quality scores*, which generally indicate the confidence of the sequencing machine that a given nucleotide at a given position is correct [29].

Depending on the DNA sequencing machine and the way it is being used, many attributes of these reads may change. The number of reads, the approximate lengths of these reads, and the expected *error rates* of these reads can all vary substantially between technologies [5, 60]. For the sake of clarity, I take a brief detour and examine some of these technologies—and their differences' impacts on assembly—below.

3.1.1 A brief overview of sequencing technologies, and the reads they generate

The 1970s saw the development of multiple early DNA sequencing techniques [105, 73, 106, 12]. The most famous of these, known as *Sanger sequencing* [112], was published in 1977 [106]. Sanger sequencing was originally a labor-intensive technique: researchers would need to manually read off nucleotides from a gel in "reads" of at most ∼80 nucleotides [106, 112]. In the years since Sanger sequencing's initial development, the process has been automated and improved to the point where it can produce reads of at most ∼1,000 nucleotides with error rates as low as 0.001% [116]. However, despite these improvements the method remains expensive and relatively low-throughput [130]. Many uses of Sanger sequencing today are therefore constrained to confirming other sequencing technologies' results in applications where accuracy is especially important [13, 83].

Many sequencing technologies have been developed since Sanger sequencing [36]. To understand the types of reads used as input to modern metagenome assembly, I focus in

[5]Herein I refer to assembly tools as just *assemblers*.

[6]As briefly mentioned in section 2.2, some workflows suggest an initial "quality control" step in which reads are filtered, trimmed, or otherwise processed to address undesirable sequencing artifacts [141, 104]. Since some assemblers do not require this step (and/or perform this step automatically), for example [59] and [58], I omit a detailed description of quality control approaches in this report.

[7]This alphabet is occasionally extended to include other symbols when the nucleotide at a given position is ambiguous: see [32]. An easy-to-access table showing this extended alphabet is available as of writing at `https://en.wikipedia.org/wiki/Nucleic_acid_notation`.

the remainder of this section on three broad classes of technologies that are all commonly in use today.

The first of these technologies are referred to as *short-read* technologies. These are exemplified today by the Illumina MiSeq and HiSeq platforms, which use "sequencing-by-synthesis" technology [14] to produce reads of up to ∼150 nucleotides [101]. These reads boast a relatively low error rate of less than 1% [101, 57]; the combination of short read length and high accuracy means that these technologies are often used in marker gene sequencing (see section 2.1) [124, 57].

The second of these technologies can be generally described as *long, error-prone reads* [59]. These are produced by the Oxford Nanopore and PacBio single-molecule real time (SMRT) technologies [132]. These reads can span over 10,000 nucleotides, which is extremely useful for assembly (as discussed in section 3.1.2.1). However, these reads have much higher error rates than short reads, with error rates of around 10–25% [132, 101]. This can complicate the process of separating "real" variations in the data from sequencing errors (as discussed in section 3.1.2.2).

The third and final of these technologies are *long and accurate* reads, also referred to as simply *HiFi* reads [132]. These are produced by PacBio circular consensus sequencing technology, and represent an improvement of the SMRT technology mentioned above [132]. In essence, HiFi reads offer the best of both worlds from short and long reads: they can span over 10,000 nucleotides and have error rates of less than 1% [132, 16], although they are somewhat shorter than some long, error-prone reads [16]. It is worth noting that the low error rate for HiFi reads only applies to errors in base calling (i.e. in "point" mutations); HiFi reads have a notably higher error rate when sequencing insertions and deletions [132]. That said, HiFi reads are a recent development, and work has already been done on accounting for these structural errors after sequencing—for example using machine learning approaches [63].

This three-category description of sequencing (short reads, long and error-prone reads, HiFi reads) omits some details. It is possible to combine multiple sequencing technologies in a single study; so-called "hybrid" assemblers have been developed to exploit this sort of data [15]. Additionally, reads can be more complex than individual linear strings of DNA: certain sequencers can produce so-called *paired-end reads* consisting of two reads that are separated by a roughly known length [101, 76]. Paired-end reads can be used to improve assembly by helping to resolve repeats (see section 3.1.2.1) [76, 31].

3.1.2 Sequencing technologies' impacts on assembly

The sequencing technologies I have examined so far vary widely in their read lengths and error rates. Here I examine the impact of these two factors on assembly (with the caveat that many other factors besides these two can have large impacts on the process, for example coverage [also discussed in section 3.3.3] [86] and contamination [35]).

3.1.2.1 Read lengths and repeat resolution

The key challenge in assembly, whether in single-genome or metagenome assembly, is accounting for *repeats* [47]. A repeat is defined as an "identical, or nearly identical, [stretch] of DNA" [76]. Repeats complicate assembly because they introduce ambiguity as to the full structure of a genome, or of genomes: assembling a genome with many repeats is comparable to putting together a jigsaw puzzle containing many "blue sky" pieces [72].

Repeats in an assembly become problematic when they exceed the read length [47]. As I will discuss later in section 3.3.3, this is a reason why short reads have historically failed to provide contiguous metagenome assemblies [16, 72]. Assembly projects, and in particular metagenome assembly projects, thus benefit tremendously from longer read lengths that can span repeats [72].

3.1.2.2 Error rates and variant calling

As may be expected, assembly projects also benefit from lower read error rates. Depending on the assembly algorithm in use (see section 3.3), the assembly process will often involve the observation (e.g. after sequence alignment) of multiple overlapping reads at a given position in a partially-reconstructed genome [62, 89]. In many cases, these reads' overlapping positions will be similar but not identical. The assembler must then determine whether each non-unanimous position is the result of sequencing error or of real variation [80].

Lower error rates permit more confident classification of these sorts of positions as "real" variations or as errors. Lower error rates also enable the detection of rarer variations [134], which in the context of microbiome sequencing data can indicate rare strains of a microbe seen in a microbiome [89, 16]. The process of distinguishing variations from errors is referred to as *variant calling*; although this is a well-studied problem when comparing sequencing data to a single reference sequence [33], dealing with variations is a challenging problem in the context of *de novo* assembly (see section 3.3.1) [80].

How an assembler handles these sorts of ambiguities—whether the assembler treats an ambiguous position as an error or as a legitimate variation—is reflected in the output data structures produced by an assembler. I focus next on these structures and their interpretation.

3.2 Assembly outputs

Although many different assembly algorithms exist, assemblers tend to produce mostly similar output. That being said, understanding how to interpret these outputs is critical—especially when performing assembly on a complex microbiome sequencing dataset, when the best possible assembly given the limitations of the data will usually be imperfect. Here I discuss the outputs produced by assembly and how these outputs are often used in downstream applications.

3.2.1 Contigs

The main output of an assembler is a collection of *contigs* representing fragments of DNA joined together from reads [120]. Similarly to reads, contigs can be thought of as strings from the alphabet $\{A, C, G, T\}$. In the ideal assembly project, one contig would be recovered for each unique molecule of DNA present in a sample; in practice, there will

usually be more contigs present due to difficulties encountered in assembly, for example due to repeats (see section 3.1.2.1) [133].

Many metrics exist to evaluate the relative "quality" of a set of contigs produced by an assembly [20]; arguably the most common of these is the N50 metric, which measures the length of the contigs in an assembly [40]. Although the N50 and other metrics can be useful when considered carefully, they are vulnerable to Strathern's generalization of Goodhart's law: namely, that "When a measure becomes a target, it ceases to be a good measure" [121, 86]. For example, an antagonistic assembler could naïvely concatenate all reads together to achieve a seemingly good N50 metric [82].

After metagenome assembly, in particular, contigs are often combined—or *binned*—into groups of contigs that putatively originate from the same genome [4, 57, 19]. These bins are referred to as *metagenome-assembled genomes* (*MAGs*). MAGs are then often evaluated using methods such as CheckM [96] which make use of known single-copy genes [19]. CheckM estimates *completeness* ("how much of the full genome have we recovered?") and *contamination* ("to what degree are other genomes mixed with this MAG?") of bacterial and archaeal MAGs [96, 2, 35]. Guidelines have been proposed for how to interpret these statistics for a given MAG [19], although—as noted in its original publication—CheckM runs into problems when working with less-well-studied or non-bacterial / archaeal genomes [96]. To an extent, then, the same caveats as with the N50 and other assembly metrics also apply to CheckM's estimates of completeness and contamination.

In addition to MAGs produced by binning, recent metagenome assembly projects have seen the advent of entire contigs being classified by CheckM as high-completeness and low-contamination—this is an encouraging sign that improvements to sequencing technologies and algorithms are lessening the difficulties of metagenome assembly, enabling the automatic recovery of complete or nearly-complete microbial genomes from microbiome sequencing data [58, 16].

Although these immediate "MAG-quality" contigs are the ideal outcome of a metagenome assembly project, most studies are not so lucky: after performing assembly, the vast majority of genomes are usually still split up into multiple contigs. To get context on how these contigs relate to each other, it is therefore often useful to consider additional assembly outputs besides the contigs alone.

3.2.2 Assembly graph

As will be discussed later in section 3.3.2, many assembly algorithms model the assembly problem as a graph traversal. In addition to contigs, many assemblers thus also output an *assembly graph* representing the connections between contigs in the data [133].

Due to assembler differences regarding the exact graph structures in use, providing a universal formal definition of an assembly graph is challenging. Regardless of their details, these graphs are usually visualized in practice as directed or undirected graphs $G = (V, E)$, where nodes $\in V$ correspond to contigs and edges $\in E$ connect overlapping contigs [58, 133, 42].

Assembly graphs can be useful in representing diversity in a metagenomic dataset [48]. For example, variations between otherwise similar sequences might result in a *bubble* structure—in which a single path in the graph splits into multiple (usually parallel) paths, after which these paths converge back to a single path [80].[8]

Recent years have accordingly seen a rise in the visualization of assembly graphs [133, 42], as well as in algorithms that make use of the assembly graph to identify further information about the strain-level genomes present in a dataset [103, 48, 61].

3.3 Assembly methods

The assembly process has been compared humorously to sausage-making [26]. Having examined the ingredients of the proverbial sausage (section 3.1) and what researchers can expect to get out of the process (section 3.2), I turn finally to the sausage-making process itself and the algorithms involved therein.

It is convenient to think of assemblers in terms of dichotomies: this section describes three such dichotomies that help classify assemblers. Although these are mostly false dichotomies (there is no reason a project could not use both *de novo* and reference-based assembly approaches, for example) these distinctions are nonetheless convenient for thinking about assembly methods. The third and final dichotomy considered examines some of the factors and shared struggles that distinguish specialized metagenome assemblers from traditional single-genome assemblers.

3.3.1 de novo versus reference-based assemblers

An important distinction to make in the context of assembly, whether assembling single-genome or metagenome data, is whether *de novo* or *reference-based* approaches are being used. *de novo* assembly methods use only the available sequencing data; reference-based methods additionally make use of reference sequences of some sort, for example already-assembled genomes stored in a reference database like RefSeq [92, 80]. If sufficient reference data exists for the genome(s) in a sample, reference-based methods can drastically simplify the assembly process—since the problem of assembling reads into genome(s) can be reduced to, or at least aided by, the well-studied problem of aligning reads to reference genome(s) [107, 100].

Many *de novo* and reference-based methods alike exist for assembling single-genome sequencing data [80]. In the specific context of metagenome assembly, however, *de novo* methods (e.g. [58, 90, 87]) are far more commonly used than reference-based methods (e.g. [24, 71, 7]). Some of this discrepancy can likely be attributed to a hesitance from researchers to rely on reference databases [87]. Variations in these databases can have difficult-to-interpret biases on

[8] See Figure 4a in [58] for an example of one connected component of an assembly graph. This component, which corresponds to a bacterial genome classified in the class *Clostridia*, contains many bubbles that indicate strain-level diversity.

analyses that make use of them [88, 71], and many microbiomes may be expected to contain previously-unstudied microbial genomes that might benefit from an analysis unbiased by reference databases [6, 71]. Notably, some reference-based metagenome assemblers make use of both *de novo* and reference-based methods [24, 71, 6].

As the quality of reference databases continues to improve, reference-based metagenome assembly methods' performance will also improve [24], likely resulting in a rise in the popularity of reference-based (or both *de novo* and reference-based) metagenome assemblers. For the sake of simplicity, the remaining two dichotomies to be discussed in this section will focus on *de novo* assembly methods.

3.3.2 Overlap graph versus de Bruijn graph assemblers

Given a collection of input reads, most *de novo* assemblers construct a graph of some sort from these reads and attempt to find paths or cycles of some sort within this graph [80]. However, the details of this approach vary between assemblers. The two most common types of graph used are *overlap graphs* and *de Bruijn graphs* [77, 107]. These methods have been compared in the literature numerous times [77, 31, 70, 80, 107, 99], so I will limit my discussion of these approaches to a very basic examination of their graph structures and algorithms used.

The overlap graph approach, also known as *overlap-layout-consensus*, saw use in the earliest genome assembly algorithms [120, 85, 70, 6]. Generally speaking, an overlap graph is a directed graph where reads are represented as nodes, and an edge from (v_1, v_2) indicates that read v_1 overlaps with read v_2 (with some leeway in how "overlap" between two sequences is defined) [77].

de Bruijn graph approaches to assembly [55, 99, 98, 139, 76, 10, 31] involve the use of a subtly different graph.[9] Given some integer $k \geq 2$, a simple de Bruijn graph can be constructed as a directed graph where all unique strings of length $k-1$ present in the reads are represented as nodes in the graph. Edges correspond to strings of length k, also known as k-mers: edges are added between two nodes (v_1, v_2) if there exists a k-mer in the reads whose first $k-1$ characters are v_1 and whose last $k-1$ characters are v_2 [55].

Often, the problem of assembly is reformulated as the problem of finding paths (or cycles) through either graph structure [77, 31]. The interest is thus usually in finding a Hamiltonian path/cycle through the overlap graph or finding an Eulerian path/cycle through the de Bruijn graph [77, 31]. The Hamiltonian Path problem is NP-complete, whereas the Eulerian Path problem is solvable in linear time: this has motivated the adoption of de Bruijn graph-based algorithms for assembly [99, 77].

In spite of this apparent gulf in efficiency, some modern assemblers still use overlap graphs [62, 25]. Reasons for this may include overlap graphs being subjectively easier to understand [70] and the recent argument that graph traversals are less important than the process of finding contigs, which is achievable in polynomial time regardless of the graph structure in use [77].

I note that this brief discussion has necessarily left out many relevant details from the past few decades—for example, various optimizations [85, 62] and error correction methods [99, 78] that have been developed to improve these approaches.

3.3.3 Single-genome versus metagenome assemblers

Recent years have seen a rise in specialized metagenome assemblers. These are often published as extensions to existing assemblers: for example, metaFlye [58] and Flye [59], metaSPAdes [90] and SPAdes [10], or MetaVelvet [87] and Velvet [139]. These metagenome assemblers use a variety of techniques to account for the difficult properties of metagenome sequencing data, as compared with ordinary genome sequencing data. Here I examine some of these difficult properties as well as the ways in which metagenome assemblers attempt to handle them.

One well-documented characteristic of metagenome sequencing data is *uneven coverage* [90, 58]. Different microbes are often present at wildly different abundances in a microbiome [80], meaning that the *coverage*—or number of reads "covering" a given position in a given genome [3]—will often vary substantially throughout a metagenome sequencing dataset. In the worst case scenario, low-abundance genomes may be only partially represented in the reads: these genomes will thus be impossible to completely assemble from the sequencing data alone [103].

Even if all genomes present in a given microbiome are completely represented in the corresponding reads (albeit still at uneven coverages), these coverage differences can pose problems for certain assemblers. For example, in the Flye assembler [59], rarer microbes can go unassembled due to the way in which Flye identifies k-mers that seem "solid," or high-frequency [58]. The metagenome-specific exension of Flye, metaFlye, addresses this problem by adding a specific "k-mer selection mode" that accounts for uneven coverages [58]. In essence, this solution represents a concession that metagenome sequencing data will not have coverages as uniform as those seen in most single-genome sequencing assemblies. As the metaFlye manuscript demonstrates, using this new k-mer selection mode increases the process' accuracy for metagenome data but decreases the process' accuracy for single-genome data [58].

Another problem inherent to metagenome assembly is dealing with repeats. I have already discussed the problematic effects that repeats can have on assembly in general (see section 3.1.2.1). Repeats are a particular problem in metagenome assembly: generally speaking, evolutionarily similar microbes in a microbiome will have similar genomes [90], and these similar genomic regions constitute *intergenomic repeats* that will complicate metagenome assembly [72, 90]. For example, although marker genes are a useful prerequisite for marker gene sequencing (see section 2.1), their conserved regions are effectively repeats since multiple genomes will contain these regions [140]—so, counterintu-

[9] Many different variations of de Bruijn graphs have been applied to assembly over the years, including [55, 99, 98, 10]. For the sake of simplicity, the description given here resembles the description of the "spectrum graph" given in [55], or of a de Bruijn graph given in [31].

itively, marker genes actually tend to cause problems for metagenome assembly [90].

Similarly to how the effects of uneven coverage can in theory be mitigated by increasing coverage, i.e. generating more reads from a sample [103], the problem of intergenomic repeats can in theory be mitigated by increasing read lengths [72]. However, this is an unhelpful observation when only short-read data are available, or when dealing with extremely long repeats. The short-read metagenome assemblers metaSPAdes [90] and MetaVelvet [87] both attempt to address this problem using heuristic methods that, interestingly, *exploit* expected coverage variations between strains in a microbiome in order to resolve repeats. So, similarly to how marker genes are helpful for marker gene sequencing but harmful for metagenome assembly, uneven coverages can cause problems for some aspects of metagenome assembly [58] but prove useful for other aspects [90, 87].

The strains of microbes present in a given microbiome will usually have many similar regions to related strains (intergenomic repeats) along with some subtle differences. As will be discussed in section 4, the problem of completely assembling these so-called *strain mixtures* resembles the problem of assembling haplotypes from a diploid genome [90]. As I have discussed previously (section 3.2), an assembler would ideally automatically assemble each unique strain into its own contig—however, it should be clear by now that intergenomic repeats and/or limited read lengths will usually complicate this process.

Viewed from a high-level perspective, different metagenome assemblers attempt to address the problem of assembling strain mixtures in different ways. metaSPAdes, for example, focuses first on generating a "consensus backbone" of related strains in a microbiome and then identifying individual strains in this backbone after the fact [90]. By default, metaFlye uses a similar "strain-suppression" mode that collapses certain assembly graph structures indicative of strain variation (e.g. bubbles, as discussed in section 3.2.2) [58, 48]. However, metaFlye also supports an alternative "strain" mode that preserves these structures: this option reduces the contiguity of the assembly, but provides a more conservative view of the potential strains in a microbiome [58]. Finally, MetaVelvet constructs a combined ("mixed") assembly graph of the input sequencing data and then attempts to decompose this graph into individual graphs for each strain in the microbiome based on expected coverage variations between strains—with the goal of transforming the metagenome assembly problem into multiple easier-to-solve single-genome assembly problems, while mitigating the effects of intergenomic repeats as discussed above [87].

Although all of these methods have proved useful for metagenome assembly, none of them are perfect. Metagenome assembly, and in particular the *de novo* assembly of strains within a metagenome, remains a fundamentally challenging problem [89].

This inherent difficulty of metagenome assembly means that this report's quest to study individual strains of microbial genomes in detail may not be immediately solvable by a single run of an assembler. However, recent years' improvements in sequencing technologies and assembly methods alike show that the field of bioinformatics is closer to this goal than ever before. I conclude this report with a discussion of the next steps on the path to strain-level metagenome assembly, and some of the major obstacles that will lie in the way.

4. FUTURE WORK: SOLVING THE STRAIN SEPARATION PROBLEM

Metagenome sequencing provides researchers with the theoretical opportunity to assemble individual microbial strains' genomes from a sample. Capitalizing on this opportunity will require making use of improvements to sequencing technologies and assembly algorithms alike.

Contigs produced by most metagenome assemblers represent chimeras of closely-related sequences from different genomes, where variations have been collapsed in order to improve assembly contiguity [80, 90, 58, 89]. Certain assemblers (for example, metaFlye's "strain" mode [58] as mentioned in section 3.3.3) can occasionally avoid this problem by sacrificing contiguity, but—in general, as of writing—substantial work is required after metagenome assembly to resolve strain-level genomes [89].

This task has been labelled the *strain separation problem* [127]. Fortunately, the strain separation problem is essentially a special case of the *haplotype assembly problem*, a well-studied problem in bioinformatics [64, 11, 41, 89]; unfortunately, the haplotype assembly problem is NP-hard [64], although various heuristic methods have been developed for it [64, 11, 41].

The haplotype assembly problem was first formulated in the context of sequencing diploid (e.g. human) genomes: humans usually have two copies of each chromosome, and assembling the full sequences of both chromosome copies (rather than a single sequence representing a chimera of both chromosome copies) is biologically useful [64]. Since most positions in a diploid genome are *homozygous* (i.e. the same between both chromosome copies), performing haplotype assembly in a diploid genome involves finding reads that span *heterozygous* (i.e. differing between chromosome copies) positions [41], and "assembling" these reads to determine the two distinct chromosome sequences. In this way, the haplotype assembly problem—as with the ordinary assembly problems we've discussed (section 3.1.2.1)—benefits from long read lengths that can connect multiple heterozygous positions, as well as from low error rates that can provide confidence that what a read says about these positions is accurate (section 3.1.2.2) [41].

The strain separation problem poses additional challenges in comparison to the original haplotype assembly problem, since—unlike the number of chromosome copies most humans have—the number of strains present in a microbiome is rarely known *a priori* [89]. For example, one recently developed tool for attempting strain separation requires that this number is specified as a parameter [127]. The other problems unique to metagenome assembly, such as uneven coverage and intergenomic repeats, will also complicate the strain separation problem.

That said, thanks to the rise of long-read sequencing [132] and long-read metagenome assembly algorithms [58], recent years have seen the development of early attempts at solving the strain separation problem [89, 127, 16]. HiFi reads in particular bode well for their ability to address this problem: these reads' low error rates will help in the identification of positions that vary between strains, and these reads' high lengths will help connect these positions [89, 16]. A recent preprint that used HiFi reads (in conjunction with additional sequencing data) to generate 428 complete[10] MAGs from a single microbiome is proof that these approaches are bringing the field closer to the automatic assembly of strain-level genomes [16].

Microbiomes are complex, vaguely understood environments, even centuries after their discovery. Assembling microbial genomes—and, more importantly, studying the biological functions of these genomes' communities—remain challenging problems. These problems are difficult due both to the technical challenges inherent to metagenome assembly [136, 89] and due to challenges in data interpretation and study design that are a hallmark of nascent scientific fields [129]. The various methods discussed throughout this report—marker gene sequencing, metagenome sequencing, and metagenome assembly—have already proven to be useful techniques for studying different aspects of microbiome sequencing data. As these methods continue to be refined, I believe that microbiome research will continue to progress toward its true potential.

5. ACKNOWLEDGEMENTS

I thank Prof. Pavel Pevzner, and the members of the Pevzner Lab, for their advice and support throughout the past year. I also thank the members of the Knight Lab and Center for Microbiome Innovation for their advice and support throughout my time at UC San Diego. I thank Lisa Marotz for bringing the history of FMTs to my attention in 2018. I thank Yoshiki Vázquez-Baeza for suggesting I consult the Human Microbiome Project paper [53]'s discussion of functional "stability" in 2020. I thank Prof. Pevzner for mentioning to me in 2020 (while I was taking CSE 282) that marker genes tend to cause problems in metagenome assemblies, as I discuss in section 3.3.3.

I note that my citation of [125] in section 1.1 was based on reading the Wikipedia article on microorganisms. Similarly, my citation of [49] in section 2.2 was based on reading the Wikipedia article on functional genomics, and my citation of [121] in section 3.2.1 was based on reading the Wikipedia article on Goodhart's Law.

5.1 Funding

I started my PhD at UC San Diego in Fall 2018. During the Fall 2018 and Winter 2019 quarters I was funded by a standard first-year CSE departmental fellowship; during the Spring 2019 and Summer 2019 quarters I was funded by the Joint University Microelectronics Program (JUMP)'s Center for Research on Intelligent Storage and Processing-in-memory (CRISP); from the Fall 2019 quarter throughout the Fall 2020 quarter I was funded by IBM Research AI through the AI Horizons Network and the UC San Diego Center for Microbiome Innovation; in the Winter 2021 quarter I was funded as a teaching assistant for CSE 282; and since the Spring 2021 quarter I have been funded as a graduate student researcher by Prof. Pevzner's grants.

(Finally, I acknowledge that I copied some of the wordings in the last paragraph's funding information from the wordings we used in the funding section of [43].)

[10] As estimated by CheckM; see section 3.2.1.

6. REFERENCES

[1] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS computational biology*, 8(6):e1002358, 2012.

[2] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, 31(6):533–538, 2013.

[3] C. Alex Buerkle and Z. Gompert. Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, 22(11):3028–3035, June 2013.

[4] J. Alneberg, B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11):1144–1146, 2014.

[5] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):1–16, 2020.

[6] M. Ayling, M. D. Clark, and R. M. Leggett. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, 21(2):584–594, 2020.

[7] J. A. Baaijens, A. Z. El Aabidine, E. Rivals, and A. Schönhuth. de novo assembly of viral quasispecies using overlap graphs. *Genome research*, 27(5):835–848, 2017.

[8] J. L. Baker, J. T. Morton, M. Dinis, R. Alvarez, N. C. Tran, R. Knight, and A. Edlund. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome research*, 31(1):64–74, 2021.

[9] J. S. Bakken, T. Borody, L. J. Brandt, J. V. Brill, D. C. Demarco, M. A. Franzos, C. Kelly, A. Khoruts, T. Louie, L. P. Martinelli, et al. Treating Clostridium difficile infection with fecal microbiota transplantation. *Clinical Gastroenterology and Hepatology*, 9(12):1044–1049, 2011.

[10] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[11] V. Bansal and V. Bafna. Hapcut: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, 24(16):i153–i159, 2008.

[12] W. M. Barnes. DNA sequencing by partial ribosubstitution. *Journal of molecular biology*, 119(1):83–99, 1978.

[13] L. M. Baudhuin, S. A. Lagerstedt, E. W. Klee, N. Fadra, D. Oglesbee, and M. J. Ferber. Confirming variants in next-generation sequencing panel testing by Sanger sequencing. *The Journal of Molecular Diagnostics*, 17(4):456–461, 2015.

[14] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59, 2008.

[15] D. Bertrand, J. Shaw, M. Kalathiyappan, A. H. Q. Ng, M. S. Kumar, C. Li, M. Dvornicic, J. P. Soldo, J. Y. Koh, C. Tong, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology*, 37(8):937–944, 2019.

[16] D. M. Bickhart, M. Kolmogorov, E. Tseng, D. Portik, A. Korobeynikov, I. Tolstoganov, G. Uritskiy, I. Liachko, S. T. Sullivan, S. B. Shin, A. Zorea, V. P. Andreu, K. Panke-Buisse, M. H. Medema, I. Mizrahi, P. A. Pevzner, and T. P. Smith. Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. *bioRxiv*, 2021.

[17] N. A. Bokulich, J. Chung, T. Battaglia, N. Henderson, M. Jay, H. Li, A. D. Lieber, F. Wu, G. I. Perez-Perez, Y. Chen, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science translational medicine*, 8(343):343ra82–343ra82, 2016.

[18] N. A. Bokulich, B. D. Kaehler, J. R. Rideout, M. Dillon, E. Bolyen, R. Knight, G. A. Huttley, and J. G. Caporaso. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, 6(1):1–17, 2018.

[19] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology*, 35(8):725–731, 2017.

[20] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):2047–217X, 2013.

[21] B. J. Callahan, P. J. McMurdie, and S. P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.

[22] B. J. Callahan, J. Wong, C. Heiner, S. Oh, C. M. Theriot, A. S. Gulati, S. K. McGill, and M. K. Dougherty. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic acids research*, 47(18):e103–e103, 2019.

[23] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):1–8, 2011.

[24] V. Cepeda, B. Liu, M. Almeida, C. M. Hill, S. Koren, T. J. Treangen, and M. Pop. MetaCompass: reference-guided assembly of metagenomes. *bioRxiv*, page 212506, 2017.

[25] H. Cheng, G. T. Concepcion, X. Feng, H. Zhang, and H. Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170–175, 2021.

[26] R. Chikhi. Question: is de novo genome assembly a solved problem with long reads, yet? `http://rayan.chikhi.name/pdf/2019-july-19-cgsi.pdf`, 2019.

[27] J. Chopyk, D. J. Nasko, S. Allard, M. T. Callahan, A. Bui, A. M. C. Ferelli, S. Chattopadhyay, E. F. Mongodin, M. Pop, S. A. Micallef, et al. Metagenomic analysis of bacterial and viral assemblages from a freshwater creek and irrigated field reveals temporal and spatial dynamics. *Science of the Total Environment*, 706:135395, 2020.

[28] J. E. Clarridge III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, 17(4):840–862, 2004.

[29] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

[30] L. Cocolin, V. Alessandria, P. Dolci, R. Gorra, and K. Rantsiou. Culture independent methods to assess the diversity and dynamics of microbiota during food fermentation. *International journal of food microbiology*, 167(1):29–43, 2013.

[31] P. Compeau and P. Pevzner. *Bioinformatics Algorithms: an active learning approach*, volume 1. Active Learning Publishers La Jolla, California, 2015. 2nd Edition.

[32] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations. *Nucleic acids research*, 13(9):3021, 1985.

[33] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[34] J. W. Debelius. `https://forum.qiime2.org/t/18934/8`, 2021. QIIME 2 discussion forum.

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

[36] J. M. Di Bella, Y. Bao, G. B. Gloor, J. P. Burton, and G. Reid. High throughput sequencing methods and analysis for microbiome research. *Journal of microbiological methods*, 95(3):401–414, 2013.

[37] C. Diener, S. M. Gibbons, and O. Resendis-Antonio. MICOM: metagenome-scale modeling to infer metabolic interactions in the gut microbiota. *mSystems*, 5(1):e00606–19, 2020.

[38] G. M. Douglas and M. G. I. Langille. A primer and discussion on DNA-based microbiome data and related bioinformatics analyses. *OSF Preprints*, Feb 2021.

[39] H. Du, T.-t. Kuang, S. Qiu, T. Xu, C.-L. G. Huan, G. Fan, and Y. Zhang. Fecal medicines used in traditional medical system of china: a systematic review of their names, original species, traditional uses, and modern investigations. *Chinese medicine*, 14(1):1–16, 2019.

[40] D. Earl, K. Bradnam, J. S. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, 21(12):2224–2241, 2011.

[41] P. Edge, V. Bafna, and V. Bansal. Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome research*, 27(5):801–812, 2017.

[42] M. Fedarko, J. Ghurye, T. Treangen, and M. Pop. MetagenomeScope: Web-based hierarchical visualization of metagenome assembly graphs. In *Proceedings of the 25th International Symposium on Graph Drawing and Network Visualization*. Springer, 2018.

[43] M. W. Fedarko, C. Martino, J. T. Morton, A. González, G. Rahman, C. A. Marotz, J. J. Minich, E. E. Allen, and R. Knight. Visualizing 'omic feature rankings and log-ratios using Qurro. *NAR Genomics and Bioinformatics*, 2(2), 04 2020. lqaa023.

[44] M. H. Fraher, P. W. O'toole, and E. M. Quigley. Techniques used to characterize the gut microbiota: a guide for the clinician. *Nature reviews Gastroenterology & hepatology*, 9(6):312–322, 2012.

[45] C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, 18:1722–1734, 2020.

[46] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348, 2013.

[47] J. Ghurye and M. Pop. Better identification of repeats in metagenomic scaffolding. In *International Workshop on Algorithms in Bioinformatics*, pages 174–184. Springer, 2016.

[48] J. Ghurye, T. Treangen, M. Fedarko, W. J. Hervey, and M. Pop. MetaCarvel: linking assembly graph motifs to biological variants. *Genome biology*, 20(1):1–14, 2019.

[49] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3):578–590, 2013.

[50] J. R. Hollenbeck and P. M. Wright. Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data, 2017.

[51] P. Hugenholtz, B. M. Goebel, and N. R. Pace. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, 180(18):4765–4774, 1998.

[52] P. Hugenholtz and T. Huber. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International journal of systematic and evolutionary microbiology*, 53(1):289–293, 2003.

[53] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207, 2012.

[54] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1–11, 2010.

[55] R. M. Idury and M. S. Waterman. A new algorithm for DNA sequence assembly. *Journal of computational biology*, 2(2):291–306, 1995.

[56] S. W. Kembel, M. Wu, J. A. Eisen, and J. L. Green. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLOS computational biology*, 8(10):e1002743, 2012.

[57] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

[58] M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. Smith, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17(11):1103–1110, 2020.

[59] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546, 2019.

[60] S. Koren and A. M. Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, 23:110–120, 2015.

[61] S. Koren, T. J. Treangen, and M. Pop. Bambus 2: scaffolding metagenomes. *Bioinformatics*, 27(21):2964–2971, 2011.

[62] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

[63] A. Lal, M. Brown, R. Mohan, J. Daw, J. Drake, and J. Israeli. Improving long-read consensus sequencing accuracy with deep learning. *bioRxiv*, 2021.

[64] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz. Snps problems, complexity, and algorithms. In *European symposium on algorithms*, pages 182–193. Springer, 2001.

[65] N. Lane. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666):20140344, 2015.

[66] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology*, 31(9):814–821, 2013.

[67] A. V. Leeuwenhoek. Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: concerning little animals by him observed in rain-well-sea-and snow water; as also in water wherein pepper had lain infused. *Philosophical Transactions of the Royal Society of London*, 12(133):821–831, 1677.

[68] S. A. Leyn, J. E. Zlamal, O. V. Kurnasov, X. Li, M. Elane, L. Myjak, M. Godzik, A. de Crecy, F. Garcia-Alcalde, M. Ebeling, et al. Experimental evolution in morbidostat reveals converging genomic trajectories on the path to triclosan resistance. *Microbial genomics*, 7(5), 2021.

[69] H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

[70] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37, 12 2011.

[71] Y.-Y. Lin, C.-H. Hsieh, J.-H. Chen, X. Lu, J.-H. Kao, P.-J. Chen, D.-S. Chen, and H.-Y. Wang. de novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline. *BMC bioinformatics*, 18(1):1–10, 2017.

[72] V. Marx. Long road to long-read assembly. *Nature Methods*, 18(2):125–129, 2021.

[73] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977.

[74] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, et al. American gut: an open platform for citizen science microbiome research. *mSystems*, 3(3):e00031–18, 2018.

[75] M. R. McLaren, A. D. Willis, and B. J. Callahan. Consistent and correctable bias in metagenomic sequencing experiments. *Elife*, 8:e46923, 2019.

[76] P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner. Paired de Bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, 18(11):1625–1634, 2011.

[77] P. Medvedev and M. Pop. What do Eulerian and Hamiltonian cycles have to do with genome assembly? *PLOS Computational Biology*, 17(5):e1008928, 2021.

[78] P. Medvedev, E. Scott, B. Kakaradov, and P. Pevzner. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, 27(13):i137–i141, 06 2011.

[79] J. S. Meisel, D. J. Nasko, B. Brubach, V. Cepeda-Espinoza, J. Chopyk, H. Corrada-Bravo, M. Fedarko, J. Ghurye, K. Javkar, N. D. Olson, et al. Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection: report from the Winter Mid-Atlantic Microbiome Meet-up, College Park, MD, January 10, 2018. *Microbiome*, 6(1):197, 2018.

[80] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[81] J. T. Morton, C. Marotz, A. Washburne, J. Silverman, L. S. Zaramela, A. Edlund, K. Zengler, and R. Knight. Establishing microbial composition measurement standards with reference frames. *Nature Communications*, 10(1):1–11, 2019.

[82] N. Moshiri and B. Behsaz. Toward a computational problem for genome sequencing. `https://www.youtube.com/watch?v=eJxP06h-QxE`, 2018.

[83] W. Mu, H.-M. Lu, J. Chen, S. Li, and A. M. Elliott. Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. *The Journal of molecular diagnostics*, 18(6):923–932, 2016.

[84] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology*, volume 51, pages 263–273. Cold Spring Harbor Laboratory Press, 1986.

[85] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, et al. A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196–2204, 2000.

[86] N. Nagarajan and M. Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, 2013.

[87] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155, 2012.

[88] D. J. Nasko, S. Koren, A. M. Phillippy, and T. J. Treangen. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome biology*, 19(1):1–10, 2018.

[89] S. M. Nicholls, W. Aubrey, K. De Grave, L. Schietgat, C. J. Creevey, and A. Clare. On the complexity of haplotyping a microbial community. *Bioinformatics*, 11 2020.

[90] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome research*, 27(5):824–834, 2017.

[91] H. Ochman and A. Caro-Quintero. Genome size and structure, bacterial. In R. M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 179–185. Academic Press, Oxford, 2016.

[92] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[93] R. Oprita, M. Bratu, B. Oprita, and B. Diaconescu. Fecal transplantation–the new, inexpensive, safe, and rapidly effective approach in the treatment of gastrointestinal tract diseases. *Journal of medicine and life*, 9(2):160, 2016.

[94] A. Ottesen, P. Ramachandran, E. Reed, J. R. White, N. Hasan, P. Subramanian, G. Ryan, K. Jarvis, C. Grim, N. Daquiqan, et al. Enrichment dynamics of listeria monocytogenes and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak. *BMC microbiology*, 16(1):1–11, 2016.

[95] N. R. Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740, 1997.

[96] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[97] V. Peddu, R. C. Shean, H. Xie, L. Shrestha, G. A. Perchetti, S. S. Minot, P. Roychoudhury, M.-L. Huang, A. Nalla, S. B. Reddy, et al. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. *Clinical chemistry*, 66(7):966–972, 2020.

[98] P. A. Pevzner, H. Tang, and G. Tesler. de novo repeat classification and fragment assembly. *Genome Research*, 14(9):1786–1796, 2004.

[99] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the national academy of sciences*, 98(17):9748–9753, 2001.

[100] M. Pop, A. Phillippy, A. L. Delcher, and S. L. Salzberg. Comparative genome assembly. *Briefings in Bioinformatics*, 5(3):237–248, 09 2004.

[101] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):1–13, 2012.

[102] D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

[103] C. Quince, S. Nurk, S. Raguideau, R. James, O. S. Soyer, J. K. Summers, A. Limasset, A. M. Eren, R. Chikhi, and A. E. Darling. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology*, 22(214), 2021.

[104] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844, 2017.

[105] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.

[106] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[107] M. C. Schatz, A. L. Delcher, and S. L. Salzberg. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173, 2010.

[108] P. Schloss. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mbio 9: e00525-18. *DOI*, 10:00525–18, 2018.

[109] P. D. Schloss. Reintroducing mothur: 10 years later. *Applied and environmental microbiology*, 86(2), 2020.

[110] P. D. Schloss. Amplicon sequence variants artificially split bacterial genomes into separate clusters. *bioRxiv*, 2021.

[111] C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, F. B. Consortium, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences*, 109(16):6241–6246, 2012.

[112] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18, 2008.

[113] N. Segata. On the road to strain-resolved comparative metagenomics. *mSystems*, 3(2), 2018.

[114] F. Shanahan. Separating the microbiome from the hyperbolome. *Genome medicine*, 7(1):1–3, 2015.

[115] G. Sharon, N. J. Cruz, D.-W. Kang, M. J. Gandal, B. Wang, Y.-M. Kim, E. M. Zink, C. P. Casey, B. C. Taylor, C. J. Lane, et al. Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell*, 177(6):1600–1618, 2019.

[116] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[117] L. Smarr, E. R. Hyde, D. McDonald, W. J. Sandborn, and R. Knight. Tracking human gut microbiome changes resulting from a colonoscopy. *Methods of information in medicine*, 56(06):442–447, 2017.

[118] V. Solovyevand and A. Salamov. Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*, pages 61–78, 2011.

[119] M. O. Sommer, G. M. Church, and G. Dantas. The human microbiome harbors a diverse reservoir of antibiotic resistance genes. *Virulence*, 1(4):299–303, 2010.

[120] R. Staden. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic acids research*, 8(16):3673–3694, 1980.

[121] M. Strathern. 'Improving ratings': audit in the British university system. *European review*, 5(3):305–321, 1997.

[122] S. Sun, R. B. Jones, and A. A. Fodor. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome*, 8(1):1–9, 2020.

[123] X. Tan and S. Johnson. Fecal microbiota transplantation (FMT) for C. difficile infection, just say 'No'. *Anaerobe*, 60:102092, 2019.

[124] L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017.

[125] M. T. Varro and M. P. Cato. *On Agriculture*, page 209. Harvard University Press, Cambridge, MA, 1934. Translated by W. D. Hooper and H. B. Ash.

[126] S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One*, 7(8):e42671, 2012.

[127] R. Vicedomini, C. Quince, A. E. Darling, and R. Chikhi. Automated strain separation in low-complexity metagenomes using long reads. *bioRxiv*, 2021.

[128] J. Vollmers, S. Wiegand, and A.-K. Kaster. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLOS one*, 12(1):e0169662, 2017.

[129] J. Walter, A. M. Armet, B. B. Finlay, and F. Shanahan. Establishing or exaggerating causality for the gut microbiome: lessons from human microbiota-associated rodents. *Cell*, 180(2):221–232, 2020.

[130] X. V. Wang, N. Blades, J. Ding, R. Sultana, and G. Parmigiani. Estimation of sequencing error rates in short reads. *BMC bioinformatics*, 13(1):1–12, 2012.

[131] T. Ward, J. Larson, J. Meulemans, B. Hillmann, J. Lynch, D. Sidiropoulos, J. R. Spear, G. Caporaso, R. Blekhman, R. Knight, et al. Bugbase predicts organism-level microbiome phenotypes. *BioRxiv*, page 133462, 2017.

[132] A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, 2019.

[133] R. R. Wick, M. B. Schultz, J. Zobel, and K. E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20):3350–3352, 2015.

[134] A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, and N. Nagarajan. Lofreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, 40(22):11189–11201, 2012.

[135] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[136] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLOS computational biology*, 6(2):e1000667, 2010.

[137] B. Yang, Y. Wang, and P.-Y. Qian. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, 17(1):1–8, 2016.

[138] H. Zafeiropoulos, H. Q. Viet, K. Vasileiadou, A. Potirakis, C. Arvanitidis, P. Topalis, C. Pavloudi, and E. Pafilis. PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, 9(3):giaa022, 2020.

[139] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[140] L. Zhang, X. Fang, H. Liao, Z. Zhang, X. Zhou, L. Han, Y. Chen, Q. Qiu, and S. C. Li. A comprehensive investigation of metagenome assembly by linked-read sequencing. *Microbiome*, 8(1):1–11, 2020.

[141] Q. Zhou, X. Su, and K. Ning. Assessment of quality control approaches for metagenomic data analysis. *Scientific reports*, 4(1):1–11, 2014.

[142] J. E. Zlamal, S. A. Leyn, M. Iyer, M. L. Elane, N. A. Wong, J. W. Wamsley, M. Vercruysse, F. Garcia-Alcalde, A. L. Osterman, and I. B. Zhulin. Shared and unique evolutionary trajectories to ciprofloxacin resistance in gram-negative bacterial pathogens. *mBio*, 12(3):e00987–21, 2021.