REVIEW

YJBM

Metagenomic Assembly: Overview, Challenges and Applications

Jay S. Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop*

Department of Computer Science and Center of Bioinformatics and Computational Biology, University of Maryland

Advances in sequencing technologies have led to the increased use of high throughput sequencing in characterizing the microbial communities associated with our bodies and our environment. Critical to the analysis of the resulting data are sequence assembly algorithms able to reconstruct genes and organisms from complex mixtures. Metagenomic assembly involves new computational challenges due to the specific characteristics of the metagenomic data. In this survey, we focus on major algorithmic approaches for genome and metagenome assembly, and discuss the new challenges and opportunities afforded by this new field. We also review several applications of metagenome assembly in addressing interesting biological problems.

INTRODUCTION

DNA sequencing has become an important tool in biological research. The cost of sequencing has been rapidly decreasing, leading to the use of sequencing technologies in a broad set of biological applications. In particular, sequencing has been used to characterize the microbial communities associated with human and animal bodies as well as with many environments within our world. The use of high throughput sequencing in the analysis of microbial communities has led to the creation of a new scientific field – metagenomics – the analysis of the combined genomes of organisms co-existing in a community. A critical step in such analyses is metagenomic assembly – the stitching together of individual DNA sequences into genes or organisms. Genome assembly algorithms have been an important component of efforts to characterize the genomes of single organisms and have been key to the modern genomic revolution. In the context of single organisms the genome assembly problem has been thoroughly studied and a number of effective strategies have been developed, strategies that underlie modern assembly tools. Metagenomic data, however, pose new challenges and create new scientific questions that still await an answer.

In this review, we will survey the key algorithmic paradigms underlying modern assembly tools. We will then discuss the specific challenges posed by metagenomic data and outline some of the strategies recently developed to address the complexities associated with these data. We will conclude with a discussion of specific biological findings that were made possible by the newly developed metagenomic assembly approaches.

GENOME ASSEMBLY OVERVIEW

Genome assembly [1] is the reconstruction of genomes from the smaller DNA segments called *reads* which are generated by a sequencing experiment. Various sequencing technologies have been developed in the past couple of decades. See Table 1 for a summary of various sequencing technologies along with their advantages and disadvantages. In many cases, reads are *pair ended* or *mate-paired*, which means that pairs of reads are sequenced from the same DNA fragment. The distance between the reads in each pair, and their relative orientation are approximately known. This information is used to resolve ambiguities caused by repetitive sequences during assembly [2] as well as to order and orient the assembled contigs – the fragments of the genome that could be

*To whom all correspondence should be addressed: Mihai Pop, Department of Computer Science and Center of Bioinformatics and Computational Biology, University of Maryland, Center for Bioinformatics and Computational Biology, Biomolecular Sciences Building. Rm. 3120F, College Park, MD 20742, Phone Number: 301-405-7245, Email address: mpop@umiacs.umd.edu.

†Abbreviations: NEC, necrotizing enterocolytis; NP, non-deterministic polynomial; OLC, Overlap-Layout-Consensus; OTU, operational taxonomic unit; RNA, ribonucleic acid; rRNA, ribosomal RNA.

Keywords: Metagenomics, Assembly, Microbiome

Table 1. Overview of current sequencing technologies.

Technology	Read Length	Accuracy	Time per run	Bases per run
Single Molecule Real-Time Sequencing (Pacific Biosciences)	10 kbp to 15 kbp	87% (Low)	30 minutes to 4 hours	5 – 10 Gb
Oxford Nanopore MinION Sequencing	5 kbp to 10 kbp	70% to 90% (Low)	1 to 2 days	500 Mb
Ion Semiconductor (Ion Torrent sequencing)	Up to 400 bp	98% (Medium)	2 hours	10Gb
Sequencing by synthesis (Illumina)	50-300bp	99.9% (High)	1 to 11 days	300 Gb
Sequencing by ligation (SOLiD sequencing)	75 bp	99.9% (High)	1 to 2 weeks	3 Gb
Pyrosequencing (454)	700 bp	98% (Medium)	24 hours	400 Mb
Chain termination sequencing (Sanger sequencing)	400 to 900 bp	99.9% (High)	20 mins to 3 hours	50 – 100 Kb

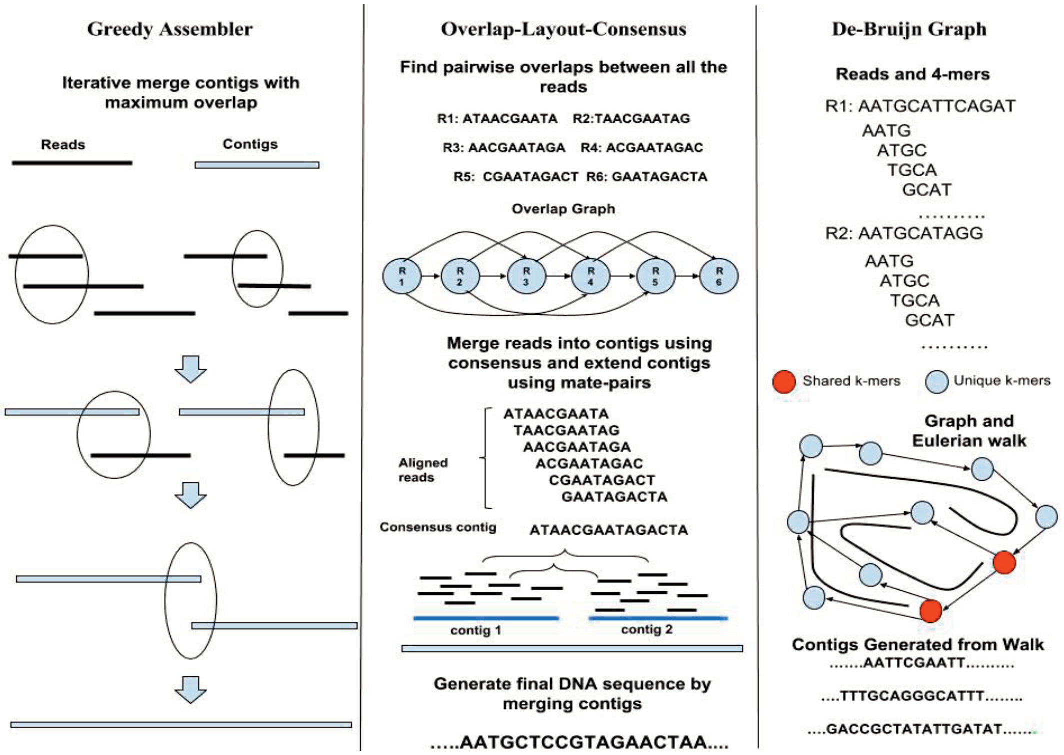

Figure 1: Overview of different de novo assembly paradigms. Schematic representation of the three main paradigms for genome assembly – Greedy, Overlap-Layout-Consensus, and de Bruijn. In Greedy assembler, reads with maximum overlaps are iteratively merged into contigs. In Overlap-Layout-Consensus approach, a graph is constructed by finding overlaps between all pairs of reads. This graph is further simplified and contigs are constructed by finding branch-less paths in the graph, and taking the consensus sequence of the overlapping reads implied by the corresponding paths. Contigs are further organized and extended using mate pair information. In de Bruijn graph assemblers, reads are chopped into short overlapping segments (k-mers) which are organized in a de Bruijn graph structure based on their co-occurrence across reads. The graph is simplified to remove artifacts due to sequencing errors, and branch-less paths are reported as contigs.

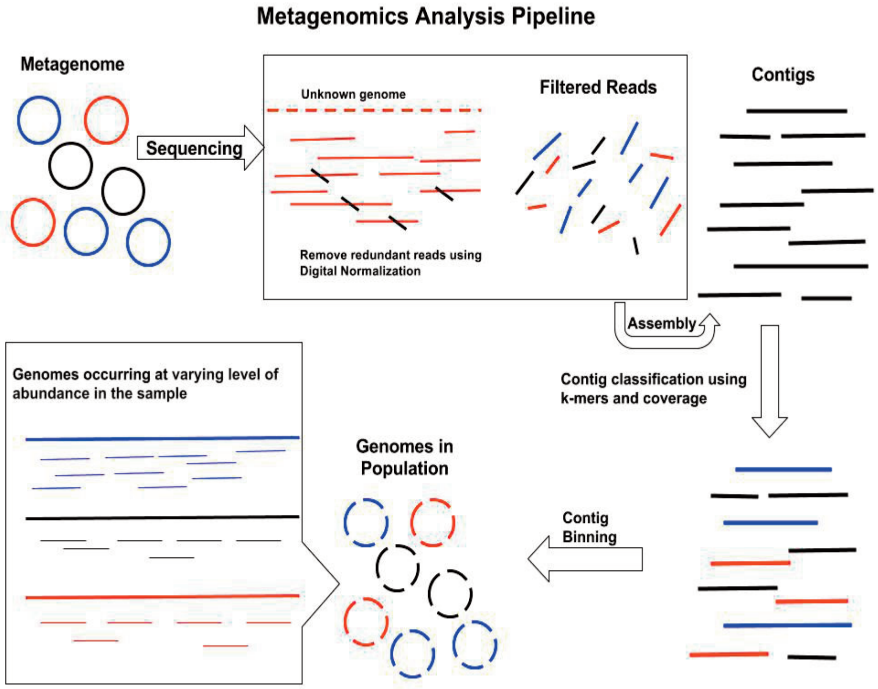
Metagenomics Analysis Pipeline
Metagenome
Sequencing
Unknown genome
Filtered Reads
Remove redundant reads using Digital Normalization
Contigs
Assembly
Contig classification using k-mers and coverage
Genomes occurring at varying level of abundance in the sample
Genomes in Population
Contig Binning

Figure 2: Metagenomic assembly pipeline. Multiple bacterial genomes within a community are represented as circles of different colors indicating multiple individuals form a same organism. Note the different levels of sequencing coverage for the individual organisms' genomes, due to the different abundance of the organisms in the original sample. After sequencing redundant reads can be removed through digital normalization, reducing the computational needs for assembly. The filtered reads are then assembled into contigs and they are classified using k-mers and coverage statistics. Contigs in each group are then binned to form draft genome sequences for organisms within the population.

stitched together from the set of reads [3]. Below, we detail these approaches.

Algorithms for Genome Assembly

In the following we will distinguish between *de novo* assembly – which involves reconstructing genomes directly from the read data, and comparative assembly – where the aim is to use the sequences of previously sequenced closely related organisms to guide the construction of a new genome. The general problem of *de novo* assembly is proved to be NP-Hard [4], which means that this problem cannot be solved efficiently. Due to the computational intractability, heuristic based methods have been devised to perform *de novo* assembly. The most widely used strategies (paradigms) are – greedy, overlap-layout-consensus (OLC†), and De Bruijn graph (See Figure 1).

Greedy

This is the most simple and intuitive method of assembly. In this method, individual reads are joined together into contigs in an iterative manner starting with the reads that overlap best and ending once no more reads or contigs can be merged. This approach is simple to implement and effective in many practical settings, and was used in several of the early genome assemblers such as TIGR [5], Phrap, VCAKE [6]. This simple greedy method, however, has some serious drawbacks. The choices made during merging of reads/contigs are locally optimal and do not consider global relationships between reads, As a result, the approach can get stuck or can result in incorrect assemblies within repetitive sequences.

Overlap-Layout-Consensus

This three step approach begins with a calculation pairwise overlaps between all pairs of reads. The overlaps are computed with a variant of a dynamic programming-based alignment algorithm, making assembly possible even if the reads contain errors. Using this information, an overlap graph is constructed where nodes are reads and edges denote overlaps between them. The layout stage consists of a simplification of the overlap graph to help identify a path that corresponds to the sequence of the genome. More precisely, a path through the overlap graph implies a 'layout' of the reads along the genome. In the consensus stage, layout is used to construct a multiple alignment of the reads and to infer the likely sequence of the genome. This assembly paradigm was used in a number of assemblers, including Celera Assembler [7], which was used to reconstruct the human genome, and Arachne [8] assembler used in many of the genome projects at the Broad Institute. The overlap-layout-consensus approach has also re-emerged recently as the primary paradigm used in assembling long reads with high error rates, such as those produced by the technologies from Pacific Biosciences and Oxford Nanopore.

De Bruijn Graph

The de Bruijn graph assembly paradigm focuses on relationship between substrings of fixed length k (k-mers) derived from the reads. The k-mers are organized in a graph structure where the nodes correspond to the k-1 prefixes and suffixes of k-mers, connected by edges that represent the k-mers. In this approach reads are not explicitly aligned to each other, rather their overlaps can be inferred

Table 2. Comparison of different *de novo* genome assembly methods. The columns in the table denote various assembly methods. The rows denote the parameters which are compared across these assembly methods. Prototypical assemblers are highlighted in each category. Assemblers marked with a * are not specifically designed for metagenomic applications.

	Greedy	OLC	De-Bruijn
Effect of repeats	✓	✓	✓
Effect of high depth of coverage	✓	✓	✗
Effect of sequencing errors	✗	✗	✓
Ease of implementation	✓	✗	✗
Assemblers	VCAKE*, phrap*, TIGR*	Celera Assembler*, Omega, SGA*	MetaVelvet, Meta-IDBA, Megahit, Meta-Ray, Meta-Spades

from the fact that they share k-mers. With this graph, assembly problem reduces to finding an Eulerian path – a path through the graph that visits each edge once. Unlike the Overlap-Layout-Consensus approach, the de Bruijn graph paradigm is affected by errors in the reads, errors which introduce false k-mers (false nodes and edges) in the graph. These errors must be eliminated prior to identifying an Eulerian path in the graph. All practical de Bruijn assemblers include a number of heuristic strategies for eliminating errors from the reads and the graph. This paradigm has become widely used after the introduction of high throughput and relatively low-error sequencing technologies, in part because it is easy to implement and efficient even in high depth of coverage settings. Some notable assemblers include: Velvet [9], SOAPdenovo [10], SOAPdenovo2 [11], ALLPATHS [12], and SPADES [13].

Comparative Assembly

The number of organisms whose genomes have been sequenced has been rapidly increasing. These genomes can be used to assist the assembly process through a strategy called *Reference Guided Assembly* or *Comparative Assembly*. Comparative assembly consists of two steps – first, all the reads are aligned against the reference genome; then a consensus sequence is generated by inferring the alignments. This approach is more effective than *de novo* assembly in resolving repeats and is thus able to get better results than *de novo* approaches especially at low depths of coverage. Long repeats are still a challenge as they lead to an ambiguous alignment of reads against the genome, though the use of mate-pair information can partly mitigate this issue and help identify the correct placement of reads. At the same time, the effectiveness of the comparative assembly approach depends on the availability of a closely related reference sequence. Differences between genome being assembled and the reference can lead to either errors in reconstruction or to a fragmented assembly. AMOScmp [14] comparative assembler attempts to identify such polymorphisms and rearrangements between genomes and breaks the assembly at these locations in order to avoid mis-assemblies.

A number of tools were developed to help augment or improve *de novo* assemblies with the help of reference genomes. OSLay [15], Projector 2 [16], ABACAS [17] and r2cat [18] simply use a reference sequences to identify the correct order and orientation of contigs from a *de novo* assembly. An extension of this approach was proposed by Husemann et al. [19] that leverages information from multiple related genomes, weighted by their evolutionary distance from the sequence being assembled. Scaffold_builder [20] also provides functionality to join together contigs that were left unassembled by the *de novo* approach, thereby helping improve the assembly through the use of a reference sequence. Finally, E-RGA [21] performs *de novo* and reference guided assembly independently first and then merges two assemblies later using a novel data structure called merge graph to avoid mis-assemblies and ambiguous overlaps.

Tradeoffs Between Different Assembly Methods

None of the methods described above is universally applicable, rather each method has specific strength and weaknesses depending on the characteristic of the data being assembled. The greedy method is easy to implement and is effective when the data contain no or only short repeats. The Overlap-Layout-Consensus approach is effective even at high error rates however its efficiency rapidly degrades with depth of coverage as it starts by computing overlaps. The de-Bruijn graph approach is computationally efficient even at high depths of coverage, however it is affected by errors in the data and is, thus, most appropriate for relatively clean datasets. Comparative assembly approaches are most effective when a sufficiently closed related sequence is available (Please refer to Table 2).

METAGENOMICS

Metagenomics is a fairly new research field focused on the analysis of sequencing data derived from mixtures of organisms. The assembly problem outlined above only become more complex as the goal is no longer to assemble a single genome, but to reconstruct the entire mixture (See Figure 2). Below we further detail these challenges and outline several of the approaches developed to address them.

Metagenomic Data

Metagenomic data consists of mixture of DNA from different organisms, and may comprise viral, bacterial, or eukaryotic organisms. The different organisms present in a mixture may have widely different levels of abundance, as well as different levels of relatedness with each other. These characteristics complicate the assembly process. As we described above, one of the main challenges to the assembly of single organisms is due to repetitive DNA segments within an organism's genome. For a single organism, assuming a uniform sequencing process, such repeats can be detected simply as anomalies in the depth of coverage (a two copy repeat would contain twice as many reads as expected). Due to the uneven (and unknown) representation of the different organisms within a metagenomic mixture, simple coverage statistics can no longer be used to detect the repeats. The confounding effect of repeats on the assembly process is further exacerbated by the fact that unrelated genomes may contain nearly-identical DNA (inter-genomic repeats) representing, for example, mobile genetic elements. At the other extreme, the multiple individuals from a same species may harbor small genetic differences (strain variants). The decision of whether such differences can be ignored when reconstructing the corresponding genome, or whether it is appropriate to reconstruct individual-specific genomes is not only computationally difficult but also ill-defined from a biological point of view. Furthermore, distinguishing true biological differences from sequencing errors becomes nearly impossible in a metagenomic setting.

A final challenge also arises from the uneven depth of sequencing coverage within a metagenomic mixture. Some organisms' genomes may be sequenced to high depths of coverage (often exceeding 1000-fold), situation that leads to high computational costs. In the Overlap-Layout-Consensus paradigm such high depths of coverage lead to a quadratic growth in the time necessary to compute overlaps (and in the number of overlaps that need to be processed), while in a de Bruijn graph setting, the higher depth of coverage amplifies the effect of errors on the assembly graph and may even stymie error correction algorithms (simply by chance multiple random errors can confirm each other).

Due to these complications, algorithms developed for single genome assembly cannot be applied directly to metagenomics data. Below we outline some approaches that have been developed in the community to deal with such challenges.

Depth Normalization and Error Correction

As outlined above, the high depth of sequencing coverage within abundant organisms in a sample impacts both the computational cost of the assembly process and also its accuracy as errors in the reads are hard to identify and correct. Brown et al. [22] proposed a strategy named *digital normalization* that aims to eliminate redundant reads within regions of high depth of coverage. This approach relies on k-mer frequencies to identify and remove reads from regions with high depth of coverage, thereby reducing the redundancy of the data. Within the reduced dataset sequencing errors are more easily detected and corrected, thereby allowing the subsequent assembly process to be both more efficient (in terms of time and memory use) and more accurate (see Figure 2).

Reducing Memory Requirements During Assembly

Most metagenomic assemblers developed to date (MetaVelvet [23], Meta-IDBA [24], MEGAHIT [25] and Ray [26]) use de Bruijn graph approach. The main assumption of this approach is that the reads contain few errors, or more precisely, that the errors can be easily corrected prior to assembly. As we mentioned above, even after filtering and error correction, many errors and polymorphisms remain in the data, causing an increase in the size of the resulting size of the de Bruijn graph. The size of the graph translates into the need for a larger memory size as the use of external memory would result in a loss of performance. Several approaches have been developed that allow storing and using the de Bruijn graphs in a lower memory footprint than the naïve solutions. One strategy involves the use of Bloom filters to partition the graph prior to assembly, leading to a large decrease in memory size [27]. Bloom filters are an inexact data-structure that trades off accuracy for memory size. To reduce the risk of false positives (nodes or edges not present in the real graph but reported by a Bloom-filter encoded de Bruijn graph), Chikhi et al. [28] introduced an extension to the approach that also compactly represents the information that may be incorrectly reported, allowing a more precise representation of the original information without losing the space efficiency. Salikhov et al. [29] further optimized graph representation by reducing storage by 30 percent to 40 percent by using a series of cascading Bloom filters.

Dealing with Genomic Variants

The approaches mentioned above address the memory requirements of assembly but not the confounding effect of genomic variants. Differences between closely related organisms can make it hard for assemblers to identify a consistent path through the assembly graph, leading to potentially fragmented assemblies. Many of the existing metagenomic assemblers try to address this issue by performing a more aggressive 'bubble popping' procedure – approach used to correct errors in the assembly of single organisms through the de Bruijn approach. Specifically, wherever parallel paths are found within the graph that differ by only a small amount, these paths are collapsed into one, allowing the assembly to reconstruct longer contiguous segments from the metagenome. Such an approach is employed, for example, by MetaVelvet [23] and Meta-IDBA [24].

Detecting and Reporting Genomic Variants Within the Assembly

Differences between closely related genomes are of potential interest to biologists, and approaches, such as those described above, which try to collapse such variants may, therefore, hide valuable information from the researchers. One of the first tools developed to find such variants after assembly is Strainer [30], a tool that analyzes the alignment pattern of reads against the reconstructed scaffold of assembled reads and provides researchers with a visualization of genetic variants found within the data. Bambus2 [31] includes a module that identifies patterns within the assembly graph that may indicate the presence of variants, approach that has been extended in Marygold [32] through the use of SPQR trees [33] – a graph data-structure that allows the efficient detection of complex 'bubble' structures within the assembly.

In the more specific case of viral metagenomic samples, where a reference sequences is available, a number of approaches have been developed to reconstruct the quasi-species structure of the data (the population of variants found within a sample). These approaches include ShoRah [34], Vispa [35] and QuRe [36], and all rely on combinatorial optimization approaches to identify a small number of genomic sequences that best explain the read data. A similar approach was also proposed in Genovo [37] in the context of full metagenomic assembly, and in EMIRGE [38] to reconstruct just the 16S rRNA gene from metagenomic mixtures. These latter approaches have substantial computational costs which limits their application to relatively small datasets.

Repeat Detection

As already mentioned, simple approaches for finding repeats based on depth of coverage anomalies are not effective within metagenomic data. An alternative approach involves the analysis of the graph structure itself, in order to find regions of the graph that appear to be 'tangled' by repeats. In Bambus2 [31] these regions are identified based on the concept of betweenness centrality [39] – a measure developed in the field of social network analysis to identify nodes in the graph that appear to have a central role (nodes traversed by many paths).

Identifying Specific Organisms within Metagenomic Samples

Even after applying the strategies outlined above, metagenomic assemblies are highly fragmented, consisting of small fragments of the genomes found in a sample. Linking together these fragments to obtain a partial reconstruction of individual fragments is challenging. A number of approaches have been developed for this purpose that leverage two complementary types of information – the DNA composition of the assembled contigs, and their depth of coverage. Sequences from the same organisms have long been shown to have a similar DNA composition (in terms of frequencies of 2-mers or 4-mers) [40,41], and this information can be used to group together contigs that have similar profiles [42]. Contigs from a same organisms can also be assumed to have similar sequencing depth within a sample, allowing them to be grouped together and even to separate out closely related sequences that may not be distinguishable by DNA composition alone [43].

The coverage approach can be further extended to leverage information from multiple samples containing a same organisms. Contigs with correlated abundance profiles can be assumed to come from a single organism. Approaches used to identify such correlations include clustering of data based on simple correlation metrics (such as Pearson or Spearman correlation of normalized abundance profiles) [44], the formulation of the problem as a under-constrained linear system of equations [45], and the combination of DNA composition measures and coverage information within a Bayesian framework as performed in CONCOCT [46]. Nielsen et al. [44] have demonstrated the power of such approaches by reconstructing 238 high quality genome sequences (as defined by the quality standards established by the Human Microbiome Project [47]) from 396 human gut samples sequenced as part of the MetaHIT project [48].

Metagenomic Analysis Pipelines

Assembly is just a small part of the data analysis process, and the increased use of metagenomic methods in biological research has led to the development of integrated pipelines for metagenomic analysis. Such pipelines include MetAmos [49] and MOCAT [50], which are stand-alone packages, as well as CloVR [51] – a framework that enables metagenomic analyses on cloud computing frameworks.

ASSEMBLY QUALITY, ASSEMBLY EVALUATION

It should be apparent by now that metagenomic assembly is a difficult computational problem. A largely overlooked analytical step is the validation of the resulting data. None of the algorithms described so far can be proven to correctly solve the assembly problem in a general setting, nor can one eliminate the possibility of errors introduced by programmers when implementing complex algorithmic techniques. Frequently, the quality of assemblies is evaluated through simple size statistics, such as the number and average sizes of the contigs generated. A measure developed in the context of the sequencing of the human genome, the N50 size (the weighted median contig size) is also often misused in a metagenomic context. The N50 size is the size of the largest contig *c* such as the sum of the sizes of contigs larger than *c* add up to the half of the correct genome size. In a metagenomic setting, the correct genome size is unknown, and therefore the N50 value is a meaningless measure. A better assessment of quality can be made by aligning metagenomic contigs to related genome sequences, as done by MetaQuast [52], or by exploring the internal consistency of the assembly (in terms of uniformity of depth of coverage and consistency of the placement of mate-pairs) as done in AMOSvalidate [53]. Recently, a number of tools have been developed that view assembly as a generative probabilistic process, allowing one to assign a likelihood to a genome assembly [54,55,56], approach that was also extended to a metagenomic context [57]. Such approaches cannot provide an absolute measure of assembly quality but can help rank multiple assemblies of the same dataset.

THE USE OF METAGENOMIC ASSEMBLY IN BIOLOGICAL APPLICATIONS

Below we highlight several examples of biological applications where metagenomic assembly approaches have been an important part of the biological results presented. These are just few from among the many other studies that have been and are being conducted, however a broader discussion of metagenomic analysis projects is beyond the scope of our paper.

Characterizing the Human-associated Microbiota

It has long been known that humans harbor complex microbial communities, but sequencing costs have prevented scientists from characterizing most of these microbes. The advent of inexpensive high throughput sequencing approaches has spurred a number of scientific efforts to better characterize the human-associate microbiota. The European project MetaHIT [48] focused on the characterization of the gut microbiota in healthy adults as well as in patients suffering from inflammatory bowel disease. Their initial publication surveyed 124 individuals through high throughput sequencing. The assembly of the resulting data reconstructed 3.3 million non-redundant gene sequences, most of which (99 percent) were derived from an estimated more than 1000 different bacterial species. Each individual was estimated to harbor an average of 160 microbial species. This initial study is only the beginning of understanding the true diversity of the human gut microbiota as evidenced by the continued discovery of new gene sequences in subsequent studies such as those by Li et al. [48] and Gevers et al. [47]. The NIH-led Human Microbiome Project [58,59] has further expanded this knowledge by adding data collected from the microbiota associated with other human body sites.

The gut microbiota is by far the best studied in humans, in no small part due to the ease of extracting samples from stool. The wealth of data collected from the gut microbiota have allowed scientists to address a number of interesting questions. Turnbaugh et al. [60] explored whether a core gut microbiota exist (a group of microbes present in all individuals) and found that while such a concept is hard to define at the organism level, the functions performed by the gut bacteria are highly conserved across people. The MetaHIT data revealed a non-random clustering of individuals in terms of their gut microbiota, leading to the proposal of a concept of 'enterotype' – semi-stable states within which a person's microbiota can exist [61]. This concept is controversial and has been debated in the scientific literature. Koren et al. [62] studied the effect of factors such as clustering methodology, distance metrics, OTU-picking approaches, sequencing depth, sequence data type and 16S rRNA region on detection of enterotypes and concluded that the concept of enterotype is not universal rather strongly tied to the methodology used to identify clustering within the data. Huse et al. [63], recently argued that the enterotypes are primarily defined by the most dominant organisms in a sample (commonly *Prevotella* or *Bacteroidetes* within human gut communities), rather than reflecting an actual "community state".

The study of the gut microbiota has also revealed the factors that influence its composition and diversity, such as diet [64,65,66], age [67,68], environment [69] and medication [70].

Premature Infant Gut Microbiome

A particularly fascinating research area is the study of the dynamic changes that occur in the human microbiota in the days and months after birth. This is not simply a matter of scientific curiosity, but also of important clinical relevance as premature infants frequently develop necrotizing enterocolitis (NEC) – a severe intestinal disease that can lead to death. The process of microbial colonization of the human gut begins at birth and continues throughout the first year of life until the gut microbiome reaches maturity. Sterile born babies acquire population of microbes through the birthing process either through the vaginal canal or from environmental introductions through cesarean delivery [71]. It is thought that in premature infants, aberrations during colonization may lead to illness or long-term health issues. Morowitz et al. [72] studied the gut microbiota within the first 3 weeks of life of a newborn baby, sequencing samples collected at four different times during this period. Their study revealed a shift in the microbial community from a community dominated by members of the Pseudomonas genus to a community dominated by organisms from the *Serratia* and *Citrobacter* genera. More importantly, however, a careful manual analysis of the assembled data revealed the presence within the developing gut microbiota of multiple *Citrobacter* strains. The relative abundance was shown to change across time, demonstrating the power of methods that explicitly take into account strain structure in the reconstruction of metagenomic data.

A recent study by Raveh-Sadka et al. [73] investigated a group of infants which developed NEC over a short period of time to find out which specific microbial strains were shared amongst co-hospitalized infants and whether the disease could be attributed to the single infectious agent. They also investigated strain level metabolic potential and population heterogeneity. Their study did not find any evidence for one common infective agent causing NEC and the dominant population of each bacterium acquired by each organism was genotypically distinct. This suggests the presence of barriers to the spread of bacteria among infants.

Global Ocean Microbiome

Microorganism in the ocean environment play important roles in various bio-geological processes. The recent advancements in metagenomics has enabled to study ocean microbial communities, their structural patterns and diversity [74]. The *Sorcerer II* Global Ocean Sampling sequenced and analyzed 6.3 Gb of DNA from surface water samples along the transect from the Northwest Atlantic to the Eastern Tropical Pacific [75]. Gene prediction within the assembly of the resulting data allowed scientists to essentially double the number of proteins available in public databases, demonstrating the power of metagenomic approaches in surveying previously uncultured organisms. Recently, the *Tara* Oceans expedition collected about 35,000 samples across multiple sea depths at global scale, in order to facilitate complete study of effect of environmental factors on ocean life [76]. While a large part of this study is focused on eukaryotic organisms, Sunagawa et al. [77] studied the bacterial microbiota of 248 samples. They generated 7.2 terabases of Illumina sequencing data, and used it to create a new annotated reference gene catalog for the ocean microbiome. Among the findings enabled by this catalog was the discovery that the vertical stratification of the composition of communities in the surface layer of ocean is mostly driven by temperature rather than geography or other environmental factors. Surprisingly, they also found that greater than 73 percent of the composition of the ocean microbiome is shared with the human gut microbiome, despite significant differences in these ecosystems. The studies of the ocean microbiome have also highlighted the broad geographical distribution of phylogenetically similar organisms, raising the question of whether specific genomic variants can be identified that correlate with or contribute to the geographical location of microbes.

CONCLUSION

The relatively recent development of inexpensive high throughput sequencing technologies has spurred efforts to characterize the microbial communities inhabiting the human body and the environment, leading to the development of a new field - metagenomics. The analysis of the resulting data has created the opportunity for developing new algorithms that account for the specific characteristics of metagenomic data. Here we have outlined the key challenges and opportunities created by this new field in the context of sequence assembly – the process used to reconstruct the genomes of organisms from DNA fragments. Despite advances in this field, further developments are still needed, particularly for the validation of the resulting assemblies in settings where a ground truth is not available. Also important is the development of new tools for uncovering and characterizing microbial communities at the strain level.

Repetitive sequences remain a challenge even for single genomes and their effect in metagenomic data is further amplified by the presence of cross-organismal repeats and uneven levels of representation of organisms within a sample. New sequencing technologies such as PacBio and Oxford Nanopore that provide long but error prone reads can overcome some of the challenges posed by repeats, however these approaches are still too expensive to be applied in a metagenomic data. Algorithms for long read assembly are still in the preliminary stage even for single genomes, and further algorithm and software development needs to take place before these technologies can be used effectively in a metagenomic setting.

In closing, we would like to note that metagenomics approaches are not the only tools available to researchers studying microbial communities. Techniques such as metatranscriptomics [78], metaproteomics [48] and metabolomics [79] have and are being developed to help provide a better understanding of the function microbes play in a community. Furthermore, targeted studies based on the 16S rRNA gene have already generated a wealth of data about microbial communities, primarily restricted to information about the taxonomic origin of organisms. Tremendous opportunities exist for the development of methods that combine all these different ways of interrogating microbial communities in order to provide a more complete understanding of the role these communities play in our world.

REFERENCES

1. Reich JG, Drabsch H, Diumler A. Nucleic Acids Research. 1984;12(13):5529–43.
2. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2013;13(1):36–46.

3. Kececioglu JD, Myers EW. Combinatiorial Algorithms For Dna Sequence Assembly. Algorithmica. 1995;13:7–51.
4. Medvedev P, Georgiou K, Myers G, Brudno M. Computability of Models for Sequence Assembly. Gene. 2007;4645:289–301.
5. Sutton GG, White O, Adams MD, Kerlavage AR. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Sci Technol. 1995;1(1):9–19.
6. Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics. 2007;23(21):2942–4.
7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.
8. Batzoglou S. {ARACHNE}: A Whole-Genome Shotgun Assembler. Genome Res. 2002;12(1):177–89.
9. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.
10. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660–6.
11. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):18.
12. Butler J, MacCallum I, Kleber M, Shlyakhter I a, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008;18(5):810–20.
13. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19(5):455–77.
14. Schatz MC, Phillippy AM, Sommer DD, Delcher AL, Puiu D, Narzisi G, et al. Hawkeye and AMOS: Visualizing and assessing the quality of genome assemblies. Brief Bioinform. 2013;14(2):213–24.
15. Richter DC, Schuster SC, Huson DH. OSLay: Optimal syntenic layout of unfinished assemblies. Bioinformatics. 2007;23(13):1573–9.
16. van Hijum SAFT, Zomer AL, Kuipers OP, Kok J. Projector 2: Contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. Nucleic Acids Res. 2005;33(SUPPL. 2):560–6.
17. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25(15):1968–9.
18. Husemann P, Stoye J. r2cat: Synteny plots and comparative assembly. Bioinformatics. 2009;26(4):570–1.
19. Husemann P, Stoye J. Phylogenetic comparative assembly. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2009;5724 LNBI:145–56.
20. Silva GG, Dutilh BE, Matthews TD, Elkins K, Schmieder R, Dinsdale E a, et al. Combining de novo and reference-guided assembly with scaffold_builder. Source Code Biol Med. 2013;8(1):23.
21. Vezzi F, Cattonaro F, Policriti A. e-RGA: enhanced Reference Guided Assembly of Complex Genomes. EMBnet.journal [Internet]. 2011;17(1):46–54. Available from: http://journal.embnet.org/index.php/embnetjournal/article/view/208/484
22. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv [Internet]. 2012;1203.4802(v2):1–18. Available from: http://arxiv.org/abs/1203.4802
23. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20).
24. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: A de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):94–101.
25. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2014;31(10):1674–6.
26. Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol. 2010;17(11):1519–33.
27. Pell J, Hintze a., Canino-Koning R, Howe a., Tiedje JM, Brown CT. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci. 2012;109(33):13272–7.
28. Chikhi R, Rizk G. Space-e cient and exact de Bruijn graph representation based on a Bloom filter. 2013;2:1–9.
29. Salikhov K, Sacomoto G, Kucherov G. Using cascading bloom filters to improve the memory usage for de Brujin graphs. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2013;8126 LNBI:364–76.
30. Eppley JM, Tyson GW, Getz WM, Banfield JF. Strainer: software for analysis of population variation in community genomic datasets. BMC Bioinformatics. 2007;8(1):398.
31. Koren S, Treangen TJ, Pop M. Bambus 2: Scaffolding metagenomes. Bioinformatics. 2011;27(21):2964–71.
32. Nijkamp JF, Pop M, Reinders MJT, De Ridder D. Exploring variation-aware contig graphs for (comparative) metagenomics using MARYGOLD. Bioinformatics. 2013;29(22):2826–34.
33. Mutzel CG. A Linear Time Implementation of {SPQR}-Trees. Proc 8th Int Symp Graph Draw. 2001;1984:70–90.
34. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011;12:119.
35. Astrovskaya I, Tork B, Mangul S, Westbrooks K, Măndoiu I, Balfe P, et al. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics. 2011;12 Suppl 6(Suppl 6):S1.
36. Prosperi MCF, Salemi M. QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics. 2012;28(1):132–3.
37. Laserson J, Jojic V, Koller D. Assembly for Metagenomes. J Comput Biol. 2011;18(3):429–43.
38. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 2011;12(5):R44.
39. Brandes U. A faster algorithm for betweenness centrality*. J Math Sociol. 2001;25(2):163–77.
40. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics. 2004;5:163.
41. Kariin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995;11(7):283–90.
42. Baker BJ, Banfield JF. Microbial communities in acid mine drainage. FEMS Microbiol Ecol. 2003;44(2):139–52.
43. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–8.
44. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822–8.
45. Carr R, Shen-Orr SS, Borenstein E. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun

Metagenomic Deconvolution. PLoS Comput Biol. 2013;9(10).

46. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–6.
47. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, et al. The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol. 2012;10(8):6–10.
48. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotech [Internet]. 2014;32(8):834-41
49. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14(1):R2.
50. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS One. 2012;7(10):1–6.
51. Angiuoli S V, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, et al. CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics. 2011;12:356.
52. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2015;32(7):1088-90.
53. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008;9(3):R55.
54. Clark SC, Egan R, Frazier PI, Wang Z. ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013;29(4):435–43.
55. Rahman A, Pachter L. CGAL: computing genome assembly likelihoods. Genome Biol. 2013;14(1):R8.
56. Ghodsi M, Hill CM, Astrovskaya I, Lin H, Sommer DD, Koren S, et al. De novo likelihood-based measures for comparing genome assembly. BMC Res Notes. 2011;6:334–51.
57. Christopher M. Hill, Irina Astrovskaya, Howard Huang, Sergey Koren, Atif Memon, Todd J. Treangen MP. De novo likelihood-based measures for comparing metagenomic assemblies. 2013 IEEE Int Conf Bioinforma Biomed. 2013;94-98.
58. Turnbaugh PJ, Ley RE, Hamady M, Fraser-liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007;449(7164):804–10.
59. Blottière HM, de Vos WM, Ehrlich SD, Doré J. Human intestinal metagenomics: State of the art and future. Curr Opin Microbiol. 2013;16(3):234–9.
60. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. J Physiol. 2009;587(Pt 17):4153–8.
61. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nat May. 2011;12(4737346):174–80.
62. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. PLoS Comput Biol. 2013;9(1).
63. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016;4:15.
64. Claesson J, Jeffery B, Conde S, Power E, O'Connor M, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature. 2012;488(7410):178.
65. Chewapreecha C. Your gut microbiota are what you eat. Nat Rev Microbiol. 2014;12(1):8.
66. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. 2014;505(7484):559–63.
67. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Ley RE, Sogin ML, et al. Human gut microbiome viewed across age and geography. Nature. 2009;457(740):222–7.
68. Biagi E, Candela M, Turroni S, Garagnani P, Franceschi C, Brigidi P. Ageing and gut microbes: Perspectives for health maintenance and longevity. Pharmacol Res. 2013;69(1):11–20.
69. Guo M, Huang K, Chen S, Qi X, He X, Cheng WH, et al. Combination of metagenomics and culture-based methods to study the interaction between ochratoxin a and gut microbiota. Toxicol Sci. 2014;141(1):314–23.
70. Xu D, Gao J, Gillilland M, Wu X, Song I, Kao JY, et al. Rifaximin alters intestinal bacteria and prevents stress-induced gut inflammation and visceral hyperalgesia in rats. Gastroenterology. 2014;146(2):1–18.
71. Josef Neu and Jona Rushing. Cesarean versus Vaginal Delivery: Long term infant outcomes and the Hygiene Hypothesis. NIH Public Access Author Manuscr. 2012;38(2):321–31.
72. Morowitz, M.J., Denef, V.J., Costello, E.K., Thomas, B.C., Poroyko, V., Relman, D.A. and Banfield JF. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc Natl Acad Sci. 2011;108(11):4512–4512.
73. Raveh-sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, et al. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. 2015;1–25.
74. Fuhrman JA. Microbial community structure and its functional implications. 2009;459(May).
75. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. 2007;5(3).
76. Karsenti E, Acinas SG, Bork P, Bowler C, Vargas C De, Raes J, et al. Community Page A Holistic Approach to Marine Eco-Systems Biology. 2011;9(10):7–11.
77. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. 2015;348(6237).
78. Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ Microbiol. 2009;11(6):1358–75.
79. Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, et al. Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One. 2011;6(3):1–9.