DEPARTMENT of COMPUTER SCIENCE & ENGINEERING
SOUTHEAST UNIVERSITY

Research Book on Metagenomic Assembly

A dissertation submitted to the Southeast University in partial fulfillment of the requirements for the degree of B. Sc. in Computer Science & Engineering

Submitted by	Supervised By
Atiq Israk Niloy 2016100000039 Batch 43 **Mohiuddin Ahammed Hemu** 2016000000129 Batch 42 Department of Computer Science & Engineering Southeast University	**Reza Shuvo** Lecturer Department of Computer Science & Engineering Southeast University **Mrinmoy Saha Roddur** Teaching Assistant University of Illinois at Urbana-Champaign

Declaration

It is declared that this Research is the official and original work that has not been submitted to any form of degree/diploma at any other institute of education center. Information that is gathered from the published work of other writers and has been acknowledged as a list of references in this work.

Atiq Israk Niloy
2016100000039
Batch 43
Department of Computer Science & Engineering
Southeast University

Mohiuddin Ahammed Hemu
2016000000129
Batch 42
Department of Computer Science & Engineering
Southeast University

Approval

The thesis titled “ ”has been submitted to the following respected members of the Board of examiners of the Faculty of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science & Engineering on October 2020 and has been accepted satisfactory.

Reza Shuvo
Lecturer and Supervisor
Department of Computer Science & Engineering
Southeast University

Mrinmoy Saha Roddur
Teaching Assistant
University of Illinois at Urbana-Champaign

Acknowledgement

First of all we would like to thank our honorable supervisors **Reza Shuvo**, Lecturer of Department of Computer Science at **Southeast University** and **Mrinmoy Saha Roddur** Teaching Assistant, University of Illinois at Urbana-Champaign. Their deep knowledge introduced us to the topic and guidance let us explore the recent ocean of Computer Science related research. Their valuable support helped us in completing our thesis within due time. Their interest and mentionable effort lead us to make a fruitful outcome of our Bachelor Degree thesis. In the end, we would like to appreciate and thank our respected parents, faculty members, classmates who were there to help us whenever we needed and as per as it was required during the research.

Abstract

Introduction

In the field of biological research, DNA sequencing tools have made revolutionary changes. By using these tools, sequencing DNA has become easier to get accurate results in less time. The sequencing cost is also being reduced continuously. The sequencing technologies are leading in the biological field for their efficiency and performance. Along with the microbiome societies, our world has been using these sequencing tools to feature the environment, human and animalistic bodies. A new scientific field has been created with the help of the analysis of microbial society. In the area of single genome assembly and revolution of modern genomics, the assembly algorithms for genomes have performed an important role. By investigating the problems of genome assembly, useful procedures have been generated in the condition of single organisms that carry modern assembly tools. Datasets related to Metagenomic Assembly are making new challenges and still require answers.

The Metagenomic Analysis was introduced to us by Venter *et al. [1]* by working on the Sargasso Sea's environmental genome analysis. One of the game-changing features was to extract diverse microbial organism genomes from the atmosphere that was not appropriately cultivated in the laboratory. The next-generation technologies have created several opportunities for metagenomic analysis of highly distinct microbial populations.

In this review, we will overview and compare the existing algorithms and tools for metagenomic Assembly. We will also discuss and review some of the challenges introduced by metagenomic data; discuss some of the assembly algorithms and tools used in modern times.

Chapter 1 : Genome Assembly

1.1 Overview

An organism's genome is the whole of its ancestral information encoded in its DNA (or, for some viruses, RNA). There are billions of base-pairs in a genome, which distinguishes among who we are. The Evolution of humans is the effect of changing the Genome Sequence slowly. Genome Sequencing is deciding the order of nucleotides of DNA or bases situated in a genome. Genome Assembly is taking reads(small pieces) of Sequences and reconstructing them to correct errors. The process is done by generating a sequence from reads. In past decades many sequencing technologies have been produced.

Contig came from the word "contiguous," which refers to a sequence of overlapping DNA sequences applied to make a material map that reproduces a single chromosome's unique DNA sequence or a chromosome range. A contig can also head to one of the DNA sequences used in making such a map.

A DNA fragment provides pairs of reads to be sequenced. Each pair has a distance between reads, and their relevant adjustment is also known. This information helps to solve the puzzle made by repetitive sequences during Assembly. [2] also to order the assembled contig that is stitched together from the set of reads[3]. There are area units, many terms employed in the context of ordination sequencing topics like Reads, Contigs, etc. In polymer sequencing, a browse is associate assumed order of try|nucleotide}s (or base pair probabilities) resembling all or a part of one polymer fragment. A typical sequencing analysis includes the ordination fragmentation of the ordination into countless molecules, that area unit size-selected and ligated to adapters.

1.2 Comparative Assembly

The number of organisms whose genomes are sequenced has been quickly increasing. These genomes can be accustomed to assist the assembly method through a technique referred to as Reference target-hunting Assembly or Comparative Assembly. Comparative Assembly consists of 2 steps – first, all the reads area units aligned against the reference genome; then, an accord sequence is generated by inferring the alignments. This approach is more practical than de novo assembly in breakdown repeats and is, therefore, able to get better results than First State Novo approaches, particularly at low depths of coverage. Long repeats area unit still a challenge as they result in AN ambiguous alignment of reads against the genome, tho' the employment of mate-pair data will partly mitigate this issue and facilitate determining the proper placement of reads.

At constant time, the effectiveness of the comparative assembly approach depends on the supply of a closely connected reference sequence. Variations between ordination being assembled and the reference will lead to either error in reconstruction or a fragmented assembly. AMOScmp [4] comparative program tries to spot such polymorphisms and rearrangements between genomes and breaks the Assembly at these locations so as to avoid mis assemblies.

A number of tools were developed to assist in augmenting or improve First State novo assemblies with the assistance of reference genomes. OSLay [5], Projector two [6], ABACAS [7] Moreover, r2cat [8] merely uses reference sequences to spot the correct order and orientation of contigs from a First State novo assembly. AN extension of this approach was projected by Husemann et al. [9] that leverages data from multiple connected genomes, weighted by their biological process distance from the sequence being assembled. Scaffold_builder [10] additionally provides practicality to hitch together contigs that were left unassembled by the First

State novo approach, thereby serving to improve the Assembly through the use of a reference sequence. Finally, e-RGA [11] performs First State novo and reference target-hunting Assembly severally first so merges two assemblies later employing a the novel organization referred to as a merge graph to avoid mis assemblies and ambiguous overlaps.

Chapter 2: Tradeoffs Between different Assembly ways

None of the ways delineated on top of is universally applicable; instead, every technique has specific strengths and weaknesses, looking at the characteristics of the information being assembled. The greedy technique is easy to implement Furthermore, it is useful once the knowledge contains no or entirely short repeats. The Overlap-Layout-Consensus(OLC) approach is practical even at high error rates; however, its efficiency speedily degrades with the depth of coverage as a result of it starting by computing overlaps. The de-Bruijn graph approach is computationally economical even at high depths of coverage; however, it is plagued by errors within the knowledge and is, thus, most applicable for comparatively clean datasets. Comparative assembly approaches square measure simplest once a sufficiently closed related sequence is accessible.

2.1 Single Genome Assembly:

There are advantages of assembling contigs into short reads. Longer ranges of a sequence contain more information. Researchers get to study whole genes or even gene clusters within a genome, and larger genetic variants and repeats are understood easily. Single Genome Assembly has the ability to remove most of the errors though it is an expensive process. Though individual reads being longer(500-100bp), the next-generation sequencing technologies has acquired the ability to produce more reads than that of first-generation technologies like Sanger Sequence technology.

2.2 Origin Assembly Pathways

OLC (Overlap Layout Consensus) approach was used in Sanger Sequence data to compare OLC's data compared to all other reads. Grouping overlaps together; they were formed into contigs (layout); after picking the most suitable nucleotides from the overlapping reads. By the time revolution, the next generation of technologies could produce shorter reads continuously when the number of reads was exponentially increasing. Nevertheless, the OLC required more computational time, which resulted in the inconsistency of the comparison limit as it needed to be compared millions or billions of reads.

The de Bruijn graph-based assembly approach was introduced to overcome the computational obstacle, which has grown widely in the field. A de Bruijn graph is a mathematical structure in which the vertex(node)of a graph shows a kmer(a string of nucleotides of length k). The nodes representing the graph made the last k-1 nucleotide of one node to overlap with the first k-1 nucleotide of another node. Each read is considered by decomposing the graph into individual overlapping kmers in turns, which creates nodes for new kmers and updates existing kmers coverage by adding vertices for remarkable transformations. Each node, apart from individual edges, is connected to each other. These nodes exist within one single line in an ideal case. It would be a trivial case to turn the graph into a theoretical contig by starting at one edge node and reaching the second edge by following labels.

Real data sets nevermore appear in such simple graphs for sure, and errors, heterozygosity, coverage differentials, repeats, and other structural variants are formed by complex branching structures. To Develop heuristics to simplify and navigate complex graphs that consist of millions of od k-mers to output contigs, and also approach to link the consist together innovates the genome assembly algorithms.

Popular genome assemblers like Velvet[12], AbySS, SOAP-denovo[13] have used utilized dBgn for these reasons. dBgn Assembly has some limitations. The standard of Assembly is going to be dramatically affected if an inappropriate k-mer is chosen while building a graph. Larger kmers provide more specificity and fewer loops than smaller kmers that cause more connected graphs. However, larger ones are more disconnected than gaps or errors included within the read data or coverage lacking within the genome.

An additional limitation is that the kmer size cannot practically exceed two but the read length to get a minimum of two edges. To ease this issue, kmers are getting used in various sizes by assemblers.

Chapter 3: METAGENOMICS

Metagenomics could be a reasonably new analysis field centered on the analysis of sequencing information derived from mixtures of organisms. The assembly downside made public on top of solely becomes a lot more complicated because the goal is not any longer to assemble one ordination, however, to reconstruct the whole mixture (See Figure 1). Below we tend to more detail these challenges and outline many of the approaches developed to handle them.

3.1 Metagenomic information

Metagenomic information consists of a mixture of desoxyribonucleic acid from different organisms, and will comprise infectious agent, bacterial, or eukaryotic organisms. The various organisms gift in a mixture could have broad completely different levels of abundance, as well as entirely different levels of connectedness with one another. These characteristics complicate the assembly method. As We delineate on top of, one of the biggest challenges to the Assembly of single organisms is thanks to repetitive desoxyribonucleic acid segments inside Associate in Nursing organism's ordination.

For one organism, assuming a consistent sequencing method, such repeats are detected merely as anomalies within the depth of coverage. Thanks to the uneven (and unknown) illustration of the various organisms inside a metagenomic mixture, detailed coverage statistics will no longer be wont to observe the repeats. The unsupportive result of repeats on the assembly method is more exacerbated by the actual fact that unrelated genomes could contain nearly-identical desoxyribonucleic acid (inter-genomic repeats) representing, as an example, mobile genetic parts.

At the opposite extreme, multiple people from the same species could harbor little genetic variations (strain variants). The choice of whether or not such variations is neglected once reconstructing the corresponding ordination, or whether or not it is appropriate to reconstruct individual-specific genomes is not solely computationally challenging however conjointly ill-defined from a biological purpose of reading.

What are more, characteristic true biological variations from sequencing errors become nearly not possible in a very metagenomic setting.

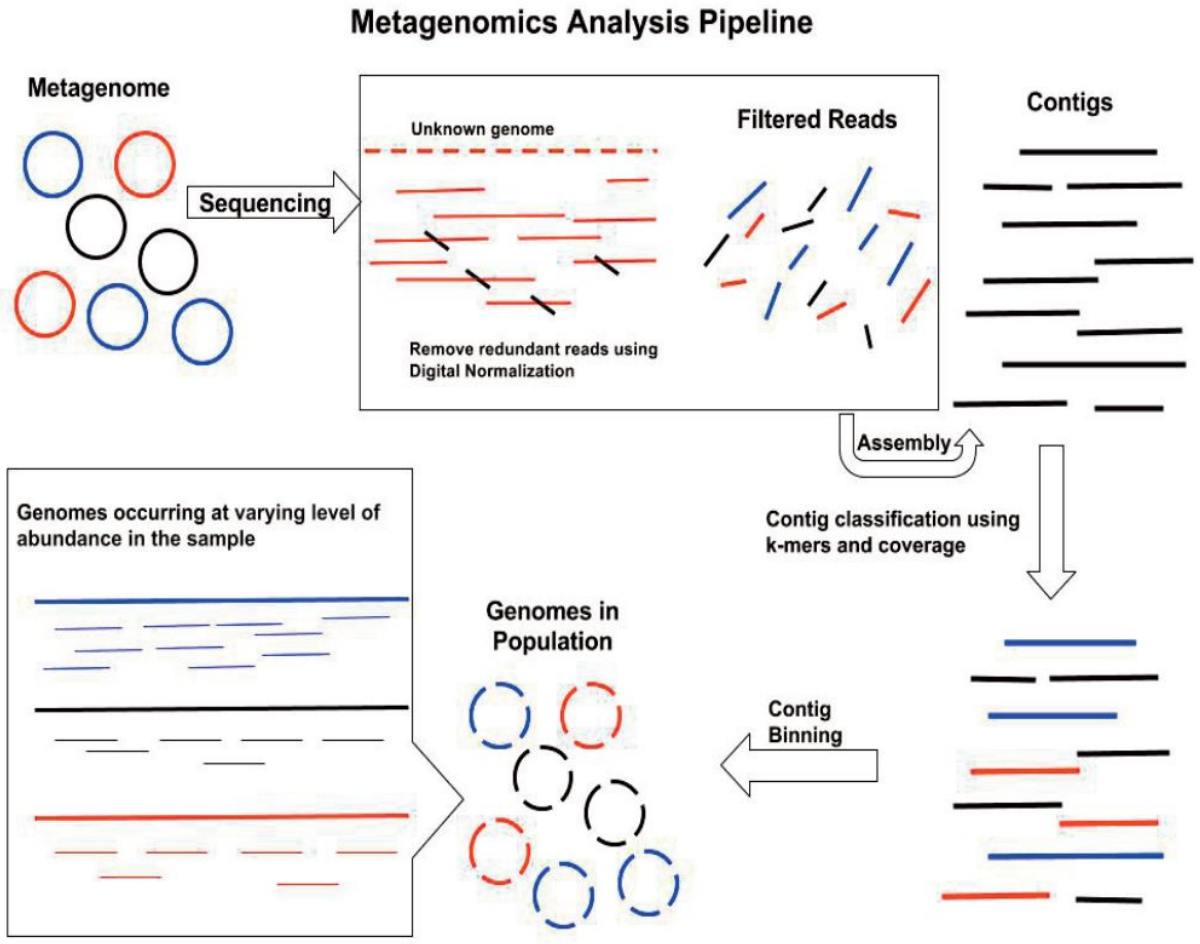

Figure 1 : Metagenomic Analysis Pipeline

A final challenge conjointly arises from the uneven depth of sequencing coverage inside a metagenomic mixture. Some organisms' genomes could also be sequenced to high depths of coverage (often exceptional 1000-fold), a situation that ends up in high process prices. within the Overlap-Layout-Consensus paradigm, such high depths of coverage lead to a quadratic growth within the time necessary to figure overlaps (and within the variety of overlaps that require to be processed), whereas in a very First State Bruijn graph setting, the higher depth of coverage amplifies the result of errors on the assembly graph and will even stymie error correction algorithms (entirely out of the blue multiple random errors will confirm every other).

Because of these complications, algorithms developed for single ordination assembly can not be applied on metagenomics information. Below we tend to define some approaches that are developed within the community to handle such challenges.

3.2 Depth standardization and Error Correction

As made public on top of, the high depth of sequencing coverage within abundant organisms in a very sample impacts each the process value of the assembly method and conjointly its accuracy as errors within the reads square measure onerous to spot and proper. Brown et al. [14] planned a method named digital normalization that aims to eliminate redundant reads within regions of the high depth of coverage. This approach relies on k-mer frequencies to spot and take away reads from regions with a high depth of coverage, as well as reducing the redundancy of the info. inside the reduced dataset sequencing errors square measure a lot of simply detected and corrected, thereby permitting the following assembly method to be both a lot of economical (in terms of your time and memory use) and more correct (see Figure 2).

3.3 Reducing Memory necessities throughout Assembly

Most metagenomic assemblers developed so far (Metavelvet [15], Meta-IDBA [16], smash hit [17] and Ray [21]) use the First State Bruijn graph approach. the most assumption of this approach is that the reads contain few errors or a lot of exactly, that the errors are simply corrected before Assembly. As we manage to mention on top of, even after filtering and error correction, several errors and polymorphisms stay within the information, inflicting a rise within the size of the ensuing size of the First State Bruijn graph. The size of the graph interprets into the requirement for a bigger memory size because the use of external memory would end in a loss of performance. many approaches are developed that enable storing and mistreatment of the First State Bruijn graphs in a very lower memory footprint than the naïve solutions.

One the strategy involves the utilization of Bloom filters to partition the graph before Assembly, resulting in an outsized decrease in memory size [18]. Bloom filters unit Associate in Nursing in actual data-structure, which trades off memory size accuracy. To diminish the risk of false positives (nodes or edges do not gift at intervals the real graph, however reportable by a Bloom-filter encoded First State Bruijn graph), Chikhi et al. [19] introduced Associate in Nursing extension
to the approach that conjointly succinctly represents the knowledge that will be incorrectly reportable, permitting a lot of precise illustration of the initial info while not losing the area potency. Salikhov et al. [20] more optimized graph illustration by reducing storage by thirty percent to forty p.c by employing a series of cascading Bloom filters.

3.4 Dealing with Genomic Variants

The approaches mentioned on top of address the memory necessities of Assembly, however, not the unsupportive result of genomic variants. variations between closely related organisms will build it onerous for assemblers to spot a standardized path through the assembly graph, leading Ghurye et al.: Metagenomic Assembly to doubtless fragmented assemblies. Several of the present metagenomic assemblers attempt to address this issue by performing a lot of aggressive 'bubble popping' procedure – approach wont to correct errors within the Assembly of single organisms through the First State Bruijn approach. Specifically, where parallel methods square measure found inside the graph that takes issue by solely a little quantity, these methods square measure folded into one, permitting the Assembly to reconstruct longer contiguous segments from the metagenome.

3.5 Assembly Memory Reduction Requirements

An approach is utilized, as an example, by Metavelvet [22] and Meta-IDBA [23]. Detecting and covering Genomic Variants inside the Assembly Differences between closely connected genomes square measure of potential interest to biologists, and approaches, such as those delineated on top of, that attempt to collapse such variants may, therefore, hide valuable info from the researchers. one among the primary tools developed to search out such variants when Assembly is filter [24], an analyzing tool that works on the alignment pattern of reads against the reconstructed

scaffold of assembled reads and provides researchers with an image of genetic variants found within the info.

Bambus2 [25], the tool that includes a module that identifies patterns inside the assembly graph that will indicate the presence of variants, the approach that has been extended in Marygold [26] through the utilization of SPQR trees [27] – a graph data-structure that permits the economic detection of complicated 'bubble' structures inside the Assembly. In the much specific case of infectious agent metagenomic samples, wherever a reference sequence is obtainable, a number of approaches are developed to reconstruct the quasi-species structure of the info (the population of variants found inside a sample). These approaches embrace vispa [29], ShoRah [28] and QuRe [30], and everyone trusts combinatorial optimization approaches to spot a little number of genomic sequences that best justify the browse data. The same approach was conjointly planned in Genovo [31] within the context of full metagenomic Assembly, and in eMIRGe [32] to reconstruct simply the 16S rRNA sequence from metagenomic mixtures. These latter approaches have substantial process prices that limit their application to comparatively little datasets.

Repeat Detection As already mentioned, straightforward approaches for locating repeats supported depth of coverage anomalies do not seem to be effective inside metagenomic information. an alternate approach involves the analysis of the graph structure itself in order to find regions of the graph that seem to be 'tangled' by repeats. In Bambus2 [25], these regions square measure known based on the conception of betweenness spatial relation [33] – a the measure developed within the field of social network analysis to identify nodes within the graph that seem to possess a central role (nodes traversed by several paths). Identifying Specific Organisms inside Metagenomic Samples even when applying the methods made public on top of, metagenomic assemblies square measure too fragmented, consisting of little fragments of the genomes found in the same sample.
Linking along these fragments to get a partial reconstruction of individual fragments is difficult. A number of approaches are developed for this purpose that leverage two complementary varieties of info – the

desoxyribonucleic acid composition of the assembled contigs, and their depth of coverage. Sequences from equivalent organisms have long been shown to possess the same desoxyribonucleic acid composition (in terms of frequencies of 2-mers or 4-mers) [34,35], and this info is wont to cluster along with contigs that have similar profiles [36]. Contigs from a same organisms may also be assumed to possess similar sequencing depth inside a sample, permitting them to be grouped along and even to separate closely connected sequences that will not be distinguishable by desoxyribonucleic acid composition alone [37]. The coverage approach is more extended to leverage info from multiple samples containing a same organisms. Contigs with correlate abundance profiles are assumed to return from one organism. Approaches wont to establish such correlations embrace clustering of knowledge supported straightforward correlation metrics (e.g., Pearson or Spearman normalized abundance profiles correlation) [38], the formulation of the matter as an equation of under-constrained linear system [39], and the combination of desoxyribonucleic acid composition measures and coverage info inside a theorem framework as performed in CONCOCT [40]. Nielsen et al. [38] have demonstrated the ability of such approaches by reconstructing 238 prime quality ordination sequences (as outlined by the standard standards established by the Human Microbiome Project [41]) from 396 human gut samples sequenced as a part of the MetaHIT project [42].

Metagenomic Analysis Pipelines Assembly is a little a part of the info analysis process, and therefore the accrued use of metagenomic ways in biological research has a light-emitting diode to the integrated pipeline event for analysis of metagenomics. Such stand-alone packages pipelines included like MetAmos [43] and MOCAT [44], which are stand-alone packages, likewise as ClovR [45] – a framework that allows metagenomic analyses on cloud computing frameworks.

Chapter 4 : Algorithms for Genome Assembly

The de-novo Assembly involves the corresponding sequenced data and the reconstructing genome that is directly found from the scanned information. Therefore to guide the development of a new sequence, previously sequenced data that is more likely to be connected to the organism is used. NP-Hard[4] is determined to be a common problem of the de-novo Assembly. That indicates, the problem can not be resolved effectively. Heuristic-based approaches have been formulated to perform the de-novo Assembly for computational obstinacy.

The algorithm that has widely been used are - De Bruijn, Greedy, Overlap Layout Consensus(OLC).

4.1 Greedy Algorithm

This is the most intuitive and straightforward method of Assembly. In this methodology, individual reads area unit joined along into contigs in an associate repetitious manner, beginning with the reads that overlap best and ending once no a lot of reads or contigs are often integrated. This approach is easy to implement and effective in several sensible settings and was utilized in many of the first order assemblers like TIGR[46], Phrap, vCAKe[47].
Although the greedy approach is simple, it has a few disadvantages.
While merging the contigs or reads, the algorithm makes an optimal choice locally rather than reflecting global relations within the reads. Thus the procedure can be abandoned, or it may result in false assemblies in the repetitive sequencing.

4.2 Overlap Layout Consensus:

By pairwise calculation of overlaps between all pairs of reads, this three-step approach begins. The overlaps area unit computed with a variant of a dynamic programming-based alignment algorithmic program, creating Assembly doable, albeit the reads contain errors. Mistreatment of this data, AN overlap graph is made wherever nodes area unit reads, and edges denote overlaps between them.

The layout stage consists of simplifying the overlap graph to a path that corresponds to the sequence of the ordination. A lot of exactly, a path through the overlap graph implies a 'layout' of the classification's reads.

In the consensus stage, the layout is employed to construct multiple alignments of the reads and to infer the probable sequence of the order. This gets together model was utilized in a few constructing agents, including Celera Assembler [48], which was utilized to remake the human genome, and Arachne [49] constructing agent utilized in a considerable lot of the genome ventures at the Broad Institute. The cover design agreement approach has likewise reappeared as of late as the essential worldview utilized in amassing long peruses with high mistake rates, for example, those delivered by the advances from Pacific Biosciences and Oxford Nanopore.

4.3 De Bruijn Graph:

The de Bruijn chart gets together worldview centers around the connection between substrings of fixed length (k-mers) from the peruses. The k-mers are sorted out in a diagram structure where the hubs relate to the k-1

prefixes and postfixes of k-mers, associated by edges that speak to the k-mers. In this methodology, peruses are not unequivocally lined up with one another; somewhat, their covers can be construed from how they share k-mers.

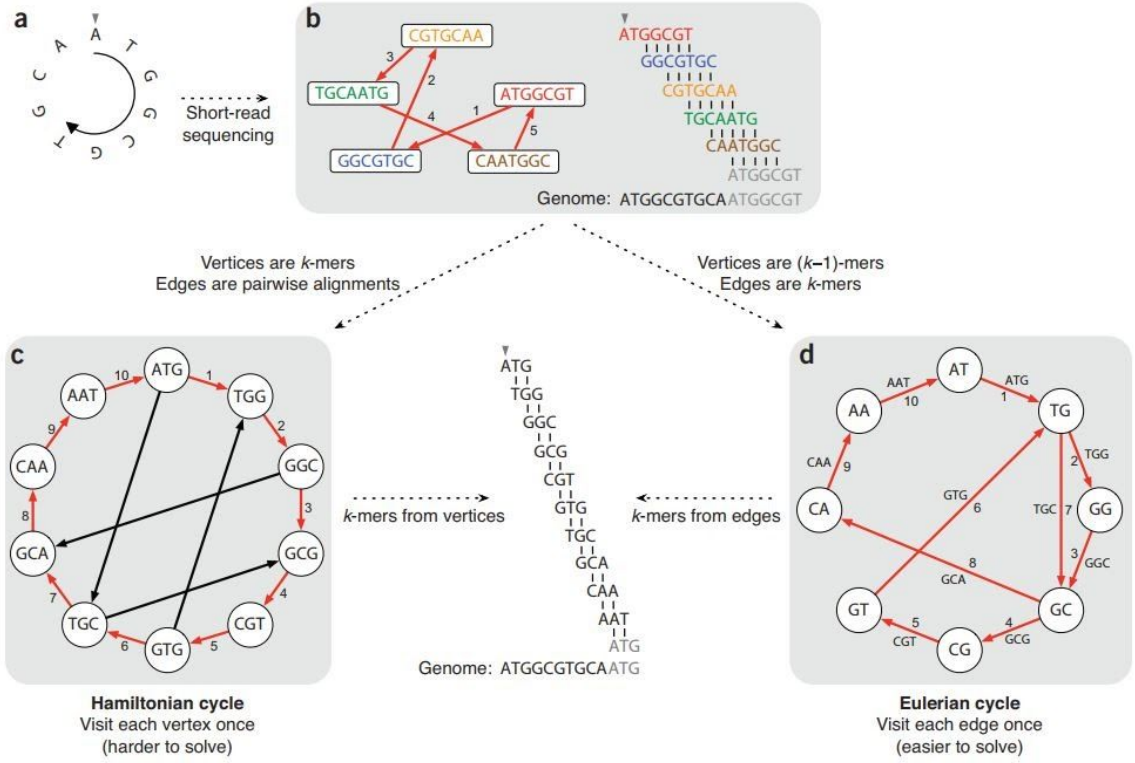

Figure 3: two methods for ordination assembly: from Hamiltonian cycles to Eulerian cycles. (a) associate degree example little circular ordination. (b) In ancient Sanger sequencing algorithms, reads were drawn as nodes throughout a graph, and edge drew alignments between reads. Walking on a Hamiltonian cycle by following the perimeters in numerical order permits one to reconstruct the circular ordination by combining serial reads' alignments. At the highest of the cycle, the sequence wraps around to the start of the ordination. The recurrent district of the sequence is colorless go in the alignment diagram. (c) associate degree alternate assembly technique initial splits read into all possible k-mers: with k = 3, ATCGAGT includes ATC, TCG, CGA, GAG, and AGT. These are following one

Hamiltonian cycle (indicated by red edges) permits one to reconstruct the ordination by forming associate degree alignment throughout that every serial k-mer (from serial nodes) is shifted by one position. This procedure recovers the ordination; however, it does not scale well to massive graphs. (d) fashionable short-read assembly algorithms construct a de Bruijn graph by representing all k-mer prefixes and suffixes as nodes, then forming edges representing k-mers having a unique prefix and suffix. As like, the k-mer edge ATG has prefix AT and suffix T.G. Finding associate degree Eulerian cycle permits one to reconstruct the ordination by forming associate degree alignment in which every serial k-mer (from serial edges) is shifted by one position. This generates an identical cyclic ordination sequence while not playacting the computationally dear task of finding a Hamiltonian cycle.

With this diagram, the get-together issue diminishes to finding an eulerian way – the way through the chart that visits each edge once. Not at all like the Overlap-Layout-Consensus approach, the de Bruijn diagram worldview is influenced by mistakes in the peruses, blunders which present bogus k-mers (bogus hubs and edges) in the chart. These mistakes must be dispensed with before recognizing an eulerian way in the diagram. All handy de Bruijn constructing agents incorporate various heuristic methodologies to remove blunders from the peruses and the diagram. This worldview has gotten generally utilized after the presentation of high throughput and moderately low-mistake sequencing advances, partially in light of the fact that it is anything but difficult to execute and effective even in the high profundity of inclusion settings.

Chapter 5 : Strategies of Algorithms

5.1 De Bruijn Assembly:

In 1946, the Dutch mathematician and scientist Nicolaas de Bruijn started finding a shortest circular 'superstring' which can be containing substring of all lengths within a given alphabet, which he named the 'superstring problem '12. If n kmers exist in an alphabet containing n symbols (for example, if an alphabet comprising A, T, G, and C is given), there will be 43 = 64 trinucleotides therein superstring. Nevertheless, If the alphabet contains only binary values, the 3-mers are going to be provided by all 3-digit binary values. Then the circular superstring may become something like 0110110001, which could contain all 3-mers. Moreover, it will be as short as possible. As we all know, it will contain each 3-mers exactly once.
However, the question appears that, is it possible to construct a superstring for all k-mers like that? Even for the k arbitrary values and alphabets?
De Bruijn has solved this issue by borrowing Euler's solution. There is a drag Called Bridges of Königsberg. It briefly explains that constructing a graph B (origin of which is de Bruijn graph) that every possible (k – 1)-mers assigned to a node; By connecting one (k–1)-mer by a directed edge to a second one if there is some k-mer whose prefix is that the former and whose the suffix is the latter. Edges of the de Bruijn graph symbolize all possible k-mers, and therefore an Eulerian cycle in B represents a shortest (cyclic) superstring that embraces each k-mer exactly once.

By marking that the indegree and outdegree of each node in B equals the dimensions of the alphabet, we will verify that B contains an Eulerian cycle. In turn, we will construct an Eulerian cycle using Euler's algorithm, accordingly solving the superstring problem. It

should now be manifest why the 'de Bruijn graph' construction described within the central theme, which does not use all probable k-mers as edges but preferably only those generated from our reads, is additionally named in honor of de Bruijn.

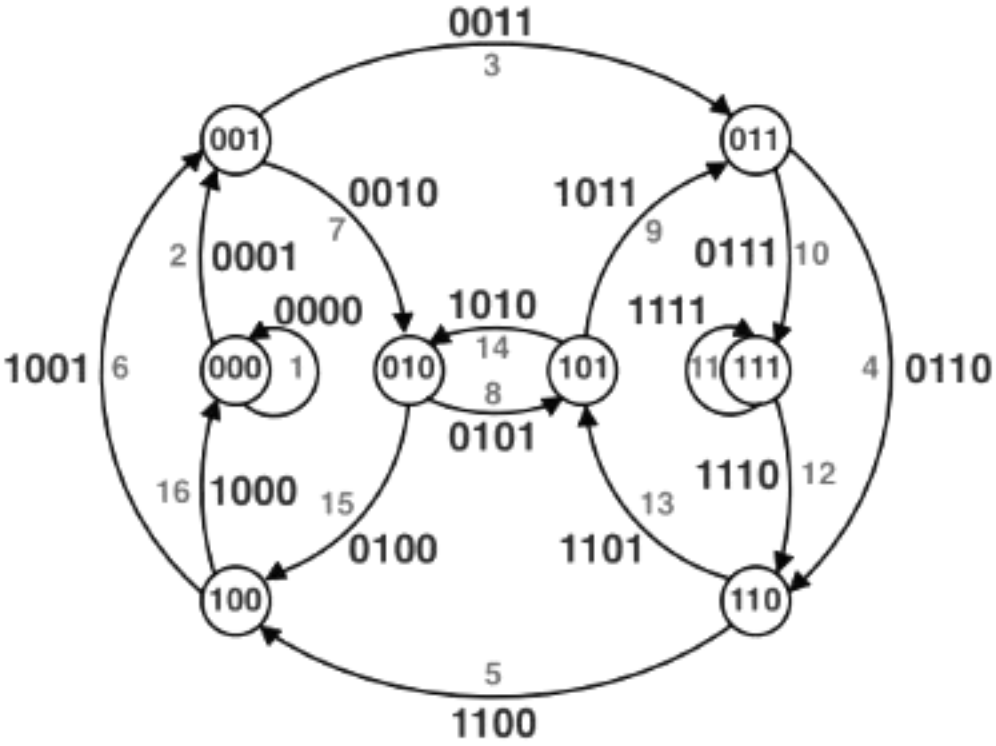

Figure 2 De Bruijn graph. The de Bruijn graph B for $k = 4$ and a two-character alphabet composed of the digits 0 and 1. This graph has an Eulerian cycle because each node has indegree and outdegree equal to 2. Following the blue numbered edges in order from 1 to 16 traces an Eulerian cycle **0**000, **0**001, **0**011, **0**110, **1**100, **1**001, **0**010, **0**101, **1**011, **0**111, **1**111, **1**110, **1**101, **1**010, **0**100, **1**000. Recording the first character (in boldface) of each edge label spells the cyclic superstring **0000110010111101**.

5.2 De Novo

As order assembly programs sew along associate degree organism's chromosomes from fragmented reads of desoxyribonucleic acid, they perform a number of the foremost complicated computations altogether of biology. Sanger sequencing, the primary thought sequencing technology, produces desoxyribonucleic acid fragments of up to one,000 base pairs; adjacent reads typically overlap by a few of hundred base pairs. This basically turns the haploid human order into a blank 30-million-piece puzzle, complicated by the fact that some items area unit missing altogether and a few items contain errors. To compensate, assemblers would like concerning eight copies of every piece of the order.

Short-read sequencing technologies have created the procedure challenge more durable. Next-generation sequencers will scan base pairs at a hundredth to a thousandth of the value of Sanger sequencing; however, the reads area unit a lot of shorter. With short-read sequencing technologies, the human-genome puzzle might contain a pair of or three billion items with a hundred copies of every piece.

Errors in Assembly occur for several reasons. Items area unit is usually incorrectly discarded as mistakes or repeats; other area units joined in the wrong places or orientations. Researchers are grappling with these styles of problems for a short while, says Adam Felsenfeld, director of the Large-Scale Sequencing program at the National Human order analysis Institute (NHGRI), in Bethesda, Maryland. "Very long, terribly high-quality reads can do wonders for assembly and fix several of those problems," he says. "But we tend to do not seem to be there, however."

To assemble an order, pc programs generally use knowledge consisting of single and paired reads. Single reads area unit merely the short sequenced fragments themselves; they will be joined up through overlapping regions into an infinite sequence referred to as a 'contig.' Repetitive sequences, polymorphisms, missing knowledge, and mistakes eventually limit the length of the contigs that assemblers will build.

Genome assembly stitches along with an order from short sequenced items of desoxyribonucleic acid. Credit: archangel Schatz, Cold Spring Harbor Paired reads generally area unit concerning an equivalent length as single reads; however, {they come they area unit available} from either finish of desoxyribonucleic acid fragments that are too long to be sequenced straight through. Looking at the preparation technique, that distance may be as short as two hundred base pairs or as giant as many tens of kilobases. Knowing that paired reads were generated from an equivalent piece of desoxyribonucleic acid will facilitate link contigs into 'scaffolds,' ordered assemblies of contigs with gaps in between. Paired-read knowledge also can indicate the scale of repetitive regions and the way apart contigs area unit.

Assessing quality is created harder as a result of sequencing technology changes thus quickly. In Jan of this year, Life Technologies launched new versions of its particle Torrent machines, which may supposedly sequence a person's order in a very day, for $1,000 in instrumentality and reagents. In Gregorian calendar month, Oxford Nanopore Technologies declared a technology that sequences tens of kilobases in continuous stretches, which might permit order assembly with rather more exactness and drastically less procedure work. Different corporations, like Pacific Biosciences, even have machines that manufacture long reads, and a minimum of some researchers area unit already combining knowledge varieties to reap the benefits of every.

Software engineers World Health Organization write assembly programs recognize they have to adapt. "Every time the information changes, it is a new drawback," says David Jaffe, World Health Organization works on order assembly strategies at the Broad Institute in Cambridge, Massachusetts. "Assemblers area unit continuously making an attempt to catch up to the information." after all, till technology has been out there for a short while, it is onerous to grasp what proportion researchers can use it. Cost, easy use, error rates, and dependableness area unit are onerous to

assess before a wider community gains much expertise with new procedures. Luckily, in-progress efforts for evaluating short-read assemblies ought to create innovations easier to judge and incorporate.

5.3 Usability

Table 3: Overview of the different *de novo* assembly tools evaluated in this study

Assembler	Version	MK	Setup	Usage	Runtime			Memory (GB)			Source	Year
					Min	Max	Median	Min	Max	Median		
Trans-ABySS	2.0.1	Yes	😊	😊	16 m	2 d 6 h 23 m	11 h 11 m	0.6	49.2	19.7	[9]	2010
Trinity	2.8.4	No	😊	😄	28 m	1 d 20 h 10 m	6 h 40 m	7.2	243.9	27.7	[10]	2011
Oases[a]	0.2.08	Yes	😊	😊	25 m	8 d 15 h 45 m	6 h 47 m	3.1	110.2	31.3	[11]	2012
SPAdes-sc[b]	3.13.0	Yes	😄	😄	16 m	7 h 52 m	2 h 26 m	5.0	37.4	25.3	[18]	2012
SPAdes-rna[b]	3.13.0	Yes[c]	😄	😄	11 m	7 h 24 m	2 h 17 m	5.0	44.2	19.5	[17]	2018
IDBA-Tran	1.1.1	Yes	😊	😊	7 m	8 h 49 m	2 h 44 m	0.6	29.1	9.6	[12]	2013
SOAPdenovo-Trans	1.03	No	😊	😕	1 m	1 h 48 m	24 m	2.1	45.6	26.4	[13]	2014
Bridger[d]	14-12-01	No	😊	😕	11 m	21 h 11 m	5 h 9 m	1.6	109.3	30.4	[14]	2015
BinPacker[d]	1.0	No	😊	😄	5 m	15 h 57 m	3 h 3 m	1.5	96.2	27.9	[15]	2016
Shannon	0.0.2	No	😕	😊	9 m	10 h 45 m	3 h 18 m	3.8	121.4	83.6	[16]	2016

We rated our experiences relating to the installation and value of every tool (Table 3). These experiences are also subjective, but we have a tendency to share them to allow inexperienced users thought of however difficult it is to put in and run every tool. a number of the tools rely on several dependencies or square measure harder to compile (Shannon, SOAPdenovo-Trans, Trans-ABySS), a minimum of on our take a look at system while not body permissions, whereas others may be put in only (SPAdes). moreover, some assemblers want additional parameter files for execution (SOAPdenovo-Trans) are circuitous to run (Oases, SOAPdenovo-Trans), want further preprocessing steps to be performed on the reads (IDBA-Tran assumes paired-end reads to be so as forward-reverse), or are just not terminating for all knowledge sets

(Bridger), whereas with others we had no issues and will execute them foursquare (Trinity, SPAdes, BinPacker). Bridger unsuccessful within the path search step for a few of the generated temporary files. Therefore, we have a tendency to performed the last step of Bridger by manually combining the transcript output. moreover, we have a tendency to had to begin Bridger two times for every knowledge set as a result of the tool crashed when once the primary begin however continued with the Assembly once started a second time on identical output folder (see execution commands in Files S3).

In the past, Oases and Trans-ABySS were continually circuitous to run as a result of the corresponding ordination assemblers Velvet [50] Furthermore, chasm [51] required to be dead initially with associate M.K. approach. These difficulties are somehow overcome by new wrapper scripts provided by the developers to execute mechanically the underlying ordination assemblers.

5.4 Computational potency

Because Diamond State novo transcriptome assembly will involve the analysis of enormous sequencing knowledge, process potency is a significant benchmark, particularly for deep sequencing comes and large sample sizes. Moreover, it has hugely counseled to run multiple assemblies with totally different tools and parameter settings (e.g., totally different k-mers); thus, computation time is a significant parameter to live for every tool. Table three summarizes the computational time and, therefore, the memory consumption of all knowledge sets and assemblers. Details are found in Electronic Supplement.

In observing, making an attempt to use de Bruijn graphs for practical information is not a simple procedure. We tend to describe some vital computational techniques that are devised to handle the possible challenges offered by mistakes and quirks in current sequencing technologies, moreover, on resolve the complexities created by repeat-rich

genomes. For example, the sharp observer will have noticed that the de Bruijn methodology for fragment assembly depends upon four hidden assumptions that do not hold for next-generation sequencing. We tend to take with no consideration that we are able to generate all k-mers gift within the ordering, that every one k-mers square measure error-free, that each k-mer seems at the most once within the ordering which the ordering consists of one circular body. For example, Illumina technology, which produces 100-nucleotide long reads, could miss some 100-mers gift within the ordering (even if the browser coverage is high) and the 100-mers that it will generate usually have errors.

Generating (nearly) all k-mers gifts within the ordering. 100-mers read generated by Illumina technology capture solely a tiny low fraction of 100-mers from the ordering (even for samples sequenced to high coverage), therefore violating de Bruijn graphs' critical assumption. However, if single breaks these reads into shorter k-mers, the ensuing k-mers usually represent nearly all k-mers from the ordering for sufficiently little k. as an example, de Bruijn graph-based assemblers could break each 100-nucleotide browse into forty-six overlapping 55-mers and additional assemble the ensuing 55-mers. Although some 100-mers occurring within the ordering are not generated as reads, this ‘Read breaking' procedure13 ensures that just about all 55-mers showing within the ordering square measure are detected. Within the example shown in Figure 3, the five reads do not account for all 7-mer substrings of the ordering. However, they are doing contain all 3-mers gift within the ordering, and this can be sufficient to reconstruct the ordering.

5.5 Handling errors in reads

Every error during a browse creates a 'bulge' within the de Bruijn graph (Supplementary Fig. 1), complicating Assembly. To make matters even worse, during ordering with inexact repeats (e.g., two regions differing by

one ester or little alternative variation), reads from the two repeat copies also will generate bulges within the de Bruijn graph. Associate in Nursing approach for 'error-correcting' reads, in which errors square measure resolved before even starting Assembly, was offered in 2001, and it is currently customarily applied. Associate in Nursing approach for removing bulges from de Bruijn graphs was outlined15 in 2004 and, with some variations, is employed in most existing short-read assemblers (e.g., EULER-SR16, Velvet17, ALLpATHS18, ABySS19, and SOApdenovo20). These and alternative recently developed tools introduced several new algorithmic and software package engineering concepts in assembly algorithms and sealed the manner toward collecting giant (e.g., mammalian) genomes with next-generation sequencing information (see refs. 21,22 for an in-depth comparison of those and alternative assemblers).

Handling desoxyribonucleic acid repeats. Imagine sequencing 3-base reads from the cyclic ordering, ATGCATGC. this could yield the four 3-mers: ATG, TGC, ground-controlled approach, and CAT. This Definition of de Bruijn graphs, however, would lead the U.S. to reconstruct the ordering as ATGC. The matter is that each of the 3-mers really happens double within the original ordering. Therefore, we are going to have to be compelled to modify ordering reconstruction, so we do not solely notice all k-mers occurring within the ordering. However, we tend to notice what number of times every such k-mer seems; additionally, that is named its 'k-mer multiplicity'. The great news is that we can and we will We square measure able to} still handle fragment assembly within the case once k-mer multiplicities are proverbial. We simply use the identical methodology to construct the de Bruijn graph, except that if the multiplicity of a k-mer is m, we are going to connect its prefix to its suffix victimization m directed edges (instead of merely one).

Extending the instance in Figure 3, if we tend to discover throughout browse generation that each of the four 3-mers GCG,TGC, CGT, and GTG has a multiplicity of 2, which each of the six 3-mers TGG, ATG, GGC, GCA, CAA, and AAT has a multiplicity of 1, we tend to produce the graph shown

in Supplementary Figure 2. what is more, the graph ensuing from adding multiplicity edges is balanced (and thus contains Associate in Nursing Eulerian cycle), as each the outdegree and indegree of a node (that represents a (k–1)-mer) may equal the number of times this (k–1)-mer seems within the ordering.

In observe, data concerning the multiplicities of k-mers within the ordering could also be troublesome to get with existing sequencing technologies. However, laptop scientists have found ways that to reconstruct the ordering, even once these information square measure unprocurable. One such technique involves 'paired reads.' Reads square measure is usually generated in pairs by sequencing each end of an extended fragment of desoxyribonucleic acid whose length can be calculable well. If one browse maps at or before the doorway to a repeat within the graph, and also the alternative maps at or once the exit, the read combine could also be wont to confirm the proper traversal through the graph. Handling multiple and linear chromosomes. We have mentioned examples during which the ordering consists of 1 circular body. If, instead, the body is linear, then we are going to have to be compelled to seek for Associate in Nursing Eulerian path, rather than Associate in Nursing Eulerian cycle; Associate in Nursing Eulerian path is not required to finish at the node wherever it begins. If there square measure multiple linear chromosomes, then we are going to have one path for everybody.

Euler's work will be custom-made to handle these complexities. Handling unsequenced regions. Regions that are not sequenced and sequencing errors could help additional break the chromosomes into contigs (a sequenced contiguous region of DNA) and gaps (unsequenced regions), with one path for every contig. Increasing the worth of k can tend to reduce the number of bulges and provides longer contigs in places with high coverage. However, some errors. Even so, it will additionally tend to interrupt contigs in regions that have low coverage. Serial contigs on a body could have overlaps of fewer than k nucleotides, or they may have gaps between them. The proper order and orientation of the contigs and also the approximate sizes of the gaps are set within the scaffolding section

of the Assembly. This section uses further data, together with paired reads, to see the order of contigs.

Chapter 6: Conclusion

The generally late advancement of cheap high throughput sequencing innovations has prodded endeavors to portray the microbial networks occupying the human body and the climate, promoting the advancement of another field - metagenomics. The examination of the subsequent information has made the open door for growing new calculations that represent the particular attributes of metagenomic information. Here we have illustrated the fundamental difficulties and openings made by this new field regarding grouping get together – the measure used to recreate the genomes of creatures from DNA sections. Notwithstanding progress in this field, further advancements are as yet required, especially for the approval of the subsequent gatherings in settings where ground truth is not accessible.

Likewise significant is the advancement of new instruments for revealing and portraying microbial networks at the strain level. Monotonous arrangements stay a test in any event, for single genomes, and their impact in metagenomic information is additionally enhanced by the presence of cross-organismal rehashes and lopsided degrees of the portrayal of life forms inside a test. New sequencing innovations, for example, PacBio and Oxford Nanopore that give long yet mistake inclined peruses, can beat a portion of the difficulties presented by rehashes; anyhow, these methodologies are still too costly to ever be applied in metagenomic information. Calculations for since quite a while ago read gathering are still in the starter stage in any event, for single genomes, and further calculation and programming improvement necessities to happen before these advances can be utilized successfully in a metagenomic perspective.

All things thought, we might want to take note of that metagenomics approaches are not by any means the only devices accessible to scientists studying microbial networks. Procedures, for example, metatranscriptomics [53], metaproteomics [52], and metabolomics [54] have and are being created to help give a superior comprehension of the capacity organisms play in a network. Moreover, directed examinations based on the 16S rRNA quality have just created an abundance of information about microbial networks, fundamentally confined to data about the ordered root of creatures. Gigantic open doors exist to improve strategies that join all these various methods of questioning microbial networks to give a complete comprehension of the job these networks play in our reality.

Chapter 7 : References

1. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science, 304, 66–74.

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. 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.

5. Richter DC, Schuster SC, Huson DH. et al. OSLay: Optimal syntenic layout of unfinished assemblies. Bioinformatics. 2007;23(13):1573–1579. [PubMed] [Google Scholar]

6. van Hijum SAFT, Zomer AL, Kuipers OP. et al. Projector 2: Contig mapping for efficient gapclosure of prokaryotic genome sequence assemblies. Nucleic Acids Res. 2005;33(2):560–566.

7. Assefa S, Keane TM, Otto TD. et al. ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25(15):1968–1969.

8. Husemann P, Stoye J. r2cat: Synteny plots and comparative assembly. Bioinformatics. 2009;26(4):570–571.

9. Husemann P, Stoye J. Phylogenetic comparative assembly. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). Bioinformatics. 2009;5724 LNBI:145–156.

10. Silva GG, Dutilh BE, Matthews TD. et al. Combining de novo and reference-guided assembly with scaffold_builder. Source Code Biol Med. 2013;8(1):23.

11. Vezzi F, Cattonaro F, Policriti A. et al. e-RGA: enhanced Reference Guided Assembly of ComplexGenomes.EMBnet.journal. 2011;17(1):46–54.http://journal.embnet.org/index.php/embnetjournal/article/view/208/484 .

12. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–829.

13. Xie Y, Wu G, Tang J. et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660–1666.

14. Brown CT, Howe A, Zhang Q. et al. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv. 2012;2:1–18. Available from: http://arxiv.org/abs/1203.4802 .

15. Namiki T, Hachiya T, Tanaka H. et al. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155.

16. Peng Y, Leung HCM, Yiu SM. et al. Meta-IDBA: A de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):94–101.

17. Li D, Liu CM, Luo R. et al. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2014;31(10):1674–1676.

18. Pell J, Hintze A, Canino-Koning R. et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci. 2012;109(33):13272–13277.

19. Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol. 2013;8:22.

20. 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) Bioinformatics. 2013;8126 LNBI:364–376.

21. Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of highthroughput sequencing technologies. J Comput Biol. 2010;17(11):1519–1533.

22. Namiki T, Hachiya T, Tanaka H. et al. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155.

23. Peng Y, Leung HCM, Yiu SM. et al. Meta-IDBA: A de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):94–101.

24. Eppley JM, Tyson GW, Getz WM. et al. Strainer: software for analysis of population variation in community genomic datasets. BMC Bioinformatics. 2007;8(1):398.

25. Koren S, Treangen TJ, Pop M. Bambus 2: Scaffolding metagenomes. Bioinformatics.2011;27(21):2964–2971.

26. Nijkamp JF, Pop M, Reinders MJT. et al. Exploring variation-aware contig graphs for (comparative) metagenomics using MARYGOLD. Bioinformatics. 2013;29(22):2826–2834.

27. Mutzel CG. A Linear Time Implementation of {SPQR}-Trees. Proc 8th Int Symp Graph Draw. 2001;1984:70–90.

28. Zagordi O, Bhattacharya A, Eriksson N. et al. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011;12:119.

29. Astrovskaya I, Tork B, Mangul S. et al. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics. 2011;12(6):S1.

30. Prosperi MCF, Salemi M. QuRe: Software for viral quasispecies reconstruction from nextgeneration sequencing data. Bioinformatics. 2012;28(1):132–133.

31. Laserson J, Jojic V, Koller D. Assembly for Metagenomes. J Comput Biol. 2011;18(3):429–443.

32. Miller CS, Baker BJ, Thomas BC. et al. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 2011;12(5):R44.

33. Brandes U. A faster algorithm for betweenness centrality*. J Math Sociol. 2001;25(2):163–177.

34. Teeling H, Waldmann J, Lombardot T. et al. 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.

35. Kariin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995;11(7):283–290.

36. Baker BJ, Banfield JF. Microbial communities in acid mine drainage. FEMS Microbiol Ecol. 2003;44(2):139–152.

37. Albertsen M, Hugenholtz P, Skarshewski A. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–538.

38. Nielsen HB, Almeida M, Juncker AS. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822–828.

39. Carr R, Shen-Orr SS, Borenstein E. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. PLoS Comput Biol. 2013;9(10):e1003292.

40. Alneberg J, Bjarnason BS, de Bruijn I. et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11(11):1144–1146.

41. Gevers D, Knight R, Petrosino JF. et al. The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. PLoS Biol. 2012;10(8):6–10.

42. Li J, Jia H, Cai X. et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotech. 2014;32(8):834–841.

43. Treangen TJ, Koren S, Sommer DD. et al. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14(1):R2.

44. Kultima JR, Sunagawa S, Li J. et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS One. 2012;7(10):1–6.

45. Angiuoli SV, Matalka M, Gussman A. et al. CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics. 2011;12:356.

46. Sutton GG, White O, Adams MD. et al. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Sci Technol. 1995;1(1):9–19.

47. Jeck WR, Reinhardt JA, Baltrus DA. et al. Extending assembly of short DNA sequences to handle error. Bioinformatics. 2007;23(21):2942–2944.

48. Venter JC, Adams MD, Myers EW. et al. The sequence of the human genome. Science. 2001;291(5507):1304–1351.

49. Batzoglou S. {ARACHNE}: A Whole-Genome Shotgun Assembler. Genome Res. 2002;12(1):177–189.

50. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.

51. Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. Genome Res 2009;19:1117–23

52. 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.

53. 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.

54. Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, JiménezHernández N, Pérez-Cobas Ae, et al. Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One. 2011;6(3):1–9.

[7][8][9][10]