


Exploring the Genomic Riches: Metagenomic Assembly and Taxonomic Profiling Algorithms

Ahmed Ali, Ahmed Ashraf,
Kareem Amr, Mohammad Makram, Ziad Elgayar
Department of Computer Engineering and Software Systems,
Faculty of Engineering, AinShams University, Cairo, Egypt
18P2517, 18P2981, 18P9093, 19P2645, e07818{@eng.asu.edu.eg}

Ashraf Abdelraouf 
Faculty of computer science,
Misr International University, Cairo, Egypt
ashraf.raouf@miuegypt.edu.eg

Abstract—In this paper, we explore the use of computing algorithms in short-read assembly and taxonomic profiling, which are two major phases in the field of metagenomics. Each phase is achieved through more than one computational algorithm. Greedy, Overlapping Layout Consensus (OLC), and De Bruijn graphs (DBGs) are proposed for short-read assembly. BLAST, Kraken, and MetaPhlAn are used in taxonomic profiling. Each algorithm is addressed in terms of theory, and comparison with other algorithms. Our study favors DBGs and BLAST algorithms as an efficient solution for the short-read assembly and taxonomic profiling phases respectively.

Index Terms—DBGs, BLAST, Metagenomics, microbiome

I. INTRODUCTION

In recent years, there have been advancements in DNA sequencing techniques used in biological research. This led to the spread of sequencing more biological data with lower cost. The Metagenomics field is a direct result of the advancements in DNA sequencing. These advancements enabled sequencing more microbial communities with higher accuracy. Next-generation sequencing technologies yield large amounts of short-reads, which have to be connected together to form genes or organisms. Assembly of single algorithm short-reads is thoroughly studied with a number of effective strategies used in this domain. On the other hand, Due to the mixed environment of microbial study, It is still a challenging task and a current area of research. In our study, we will review three different assembly algorithms: Greedy, OLC, and DBGs. After assembly of different genomes, we need to identify the taxonomic profile for each genome in our sample. Taxonomic profiling algorithms are used to match each assembled genome to a reference database. As for assembly algorithms, BLAST and Kraken algorithms are studied and compared.

II. METAGENOMICS ASSEMBLY

We define genome assembly [1] as the reconstruction of the whole genome from short-reads coming from Next-generation DNA sequencing technologies. Mostly, Pairs of the DNA are sequenced from the same DNA and the distance between reads in the pair are previously defined. Based on this information, we can order and align the previously assembled genome contigs.

Metagenomics assembly algorithms can be divided into two types: de novo assembly and comparative assembly. In de

novo, the genome is reconstructed from the short-read data without reference database (see Figure II). On the other hand, comparative assembly requires a previously constructed reference database to guide the assembly algorithm. Our research is focused on the de novo assembly algorithms, which are greedy, OLC, De Bruijn graphs (see Table I). De novo assembly is considered to be an NP-Hard problem due to its computational intractability[4].

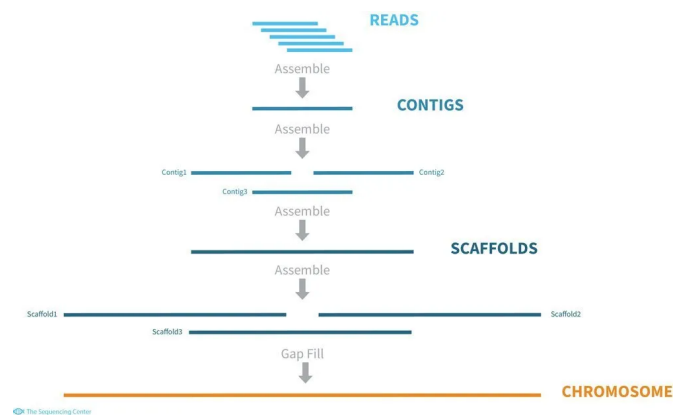


Fig. 1. De novo genome assembly from reads

A. Greedy

In this technique, short-reads with best overlaps are matched iteratively. The iteration ends with no more matching reads to be joined (see Figure 2). Early genome assemblers such as TIGT [2], VCKAE [3], has used this simple and intuitive algorithm. The main drawback of this simple greedy algorithm is finding locally optimal relationships between short-reads instead of globally optimal.

B. Overlap-Layout-Consensus

Each word of the term Overlap-layout-consensus- OLC - defines a phase of the algorithm. Firstly, pairwise overlap is calculated between pairs of short-reads using dynamic-programming based alignment algorithms such as Needleman Wunsch. Overlap graph is constructed using the previous information, where nodes and edges represent the short-reads and overlaps respectively. Simplification of the graph is made

to facilitate identifying a path for the genome sequence in the layout stage.

Each word of the term Overlap-layout-consensus- OLC - defines a phase of the algorithm. Firstly, pairwise overlap is calculated between pairs of short-reads using dynamic-programming based alignment algorithms such as Needleman Wunsch. Overlap graph is constructed using the previous information, where nodes and edges represent the short-reads and overlaps respectively (see Figure 2). Simplification of the graph is made to facilitate identifying a path for the genome sequence in the layout stage. Layouts from the previous stage are merged in one sequence based on majority voting in the consensus stage. Celera Assembler [4], which was used for human genome reconstruction, is built upon the OLC algorithm.

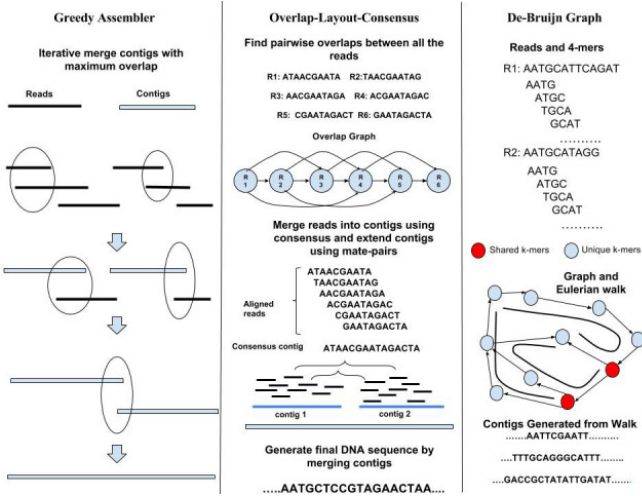


Fig. 2. Comparing the steps of the three different algorithms

C. De Bruijn Graph

Our last assembly algorithm is de Bruijn graph (DBGs), which is based on finding the relation between k-mers of the short-reads. As shown in (see Figure 3), graph structure is established with nodes as k-1 prefix and suffix of k-mers and connected by an edge, which is the k-mer itself. DBGs do not explicitly align the sequences, instead their overlap is based on the fact that they share k-mers. After the graph representation, the problem has been reduced to finding an Eulerian path - a path through all edges of the graph only once- in this graph. DBGs is prone to errors more than the other algorithms. Errors should be eliminated beforehand using a number of heuristic strategies. Velvet [5], SOAPdenovo [6], and SPAdes [7] are based on the DBGs algorithm.

In algorithm 1, to apply DBG algorithm on our sample we need to follow the steps in the Algorithm 1 and Algorithm 2 [8]. In Algorithm 1, Number of k-mers is calculated for the given DNA sequence. Our algorithm relies on non-cyclic k-mers calculations, which does not count terminal letters with length less than k. Hash table data structure is used to store the count of each unique k-mer in our sequence.

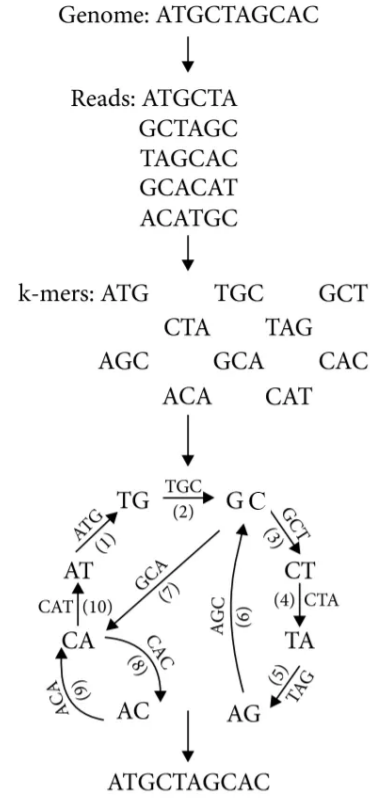


Fig. 3. General framework for De-bruijn assembly algorithm

Algorithm 1 Get K-mer Count from Sequence

Input : sequence, k=3, cyclic=True

Output: kmers

Function getKmerCount (sequence, k, cyclic)

```

kmers ← empty dictionary
for i ← 0 to length(sequence) do
    kmer ← sequence[i:i+k]
    length ← length(kmer)
    if cyclic then
        if length ≠ k then
            kmer ← kmer + sequence[: (k-length)]
        end
    end
    if length ≠ k then
        continue
    end
    if kmer ∈ kmers then
        kmers[kmer] ← kmers[kmer] + 1
    end
    else
        kmers[kmer] ← 1
    end
end
return kmers
end

```

TABLE I
COMPARISON BETWEEN TAXONOMIC PROFILING ALGORITHMS

Metric	Greedy	OLC	De-Bruijn
Effect of repeats	✓	✓	✓
Effect of high depth of coverage	✓	✓	✗
Effect of sequencing errors	✗	✗	✓
Ease of implementation	✓	✗	✗

In algorithm 2, after calculating the number of k-mers in our sequence, Edges, which represents the k-mers of the sequence are added to list by connecting nodes, which represents the k-1 mers of each unique k-mer.

D. COMPLEXITY ANALYSIS OF DBG

Before calculating the complexity of DBG, assumptions below were stated.

- 1) Our reads are error-free.
- 2) Each distinct k-mer has only one weighted edge.
- 3) G is the length of genome.
- 4) There is one node for each distinct k-1-mer.
- 5) There is only one edge for each k-mer.
- 6) Number of nodes and Edges are $O(\min(N, G))$

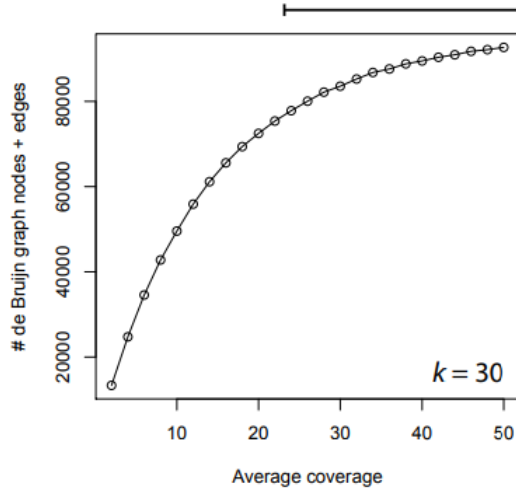


Fig. 4. Average coverage vs number of De-sbruijn graph nodes and edges

In our perfect world with previous assumptions, Graph space complexity can be $O(\min(N, G))$ only and when average coverage is high, $G \ll N$. So, the space complexity is $O(G)$, which is linear efficient space complexity. Compared to overlap graph in OLC, The DBG is more space efficient based on the previous results. The space complexity overlap graph is $O(N + n^2)$. (see Figure 4) for coverage analysis.

Algorithm 2 Get De Bruijn Edges from kmers

Input : kmers

Output: edges

```

Function getDebruijnEdgesFromKmers (kmers)
    edges  $\leftarrow$  empty set
    for  $k1 \leftarrow$  each kmer in kmers do
        for  $k2 \leftarrow$  each kmer in kmers do
            if  $k1 \neq k2$  then
                if  $k1[1:] = k2[:-1]$  then
                    | edges.add(( $k1[:-1]$ ,  $k2[:-1]$ ))
                end
                if  $k1[:-1] = k2[1:]$  then
                    | edges.add(( $k2[:-1]$ ,  $k1[:-1]$ ))
                end
            end
        end
    end
    return edges
end

```

E. Binning

It is an optional stage follows the sequencing stage, which meant to improve the efficiency of the assembly algorithms. These reads come from different organisms; the binning role is to cluster these reads into distinct clusters "bins". After that, in the assembly stage we can now search on each bin to reconstruct the genome, which is much easier than searching in hundreds of different reads (see Figure 5).

Binning can be done using various algorithms and methods

a) *sequence composition*: it makes use of the fact that genomes have conserved nucleotide composition (e.g a certain GC or abundance distribution of k-mers) and this will also be reflected in sequence fragments of the genomes.

b) *TETRA*: it is a statistical classifier, makes use of patterns of tetranucleotides in genomic fragments. tetramers are four consecutive nucleotide-based fragments, and since there are four possible nucleotides in DNA, there can be $4^4 = 256$ different tetramers. The way TETRA operates is by tabulating each tetramer's frequency for a certain sequence. These frequencies are then used to compute z-scores, which show how the tetramer differs from what would be predicted by looking at individual nucleotide compositions. To determine how similar the various sequences in the sample are to one another, the z-scores for each tetramer are assembled into a vector, and the vectors belonging to the various sequences are compared pair-wise. we expect that the most similar sequences to come in the same OTU.

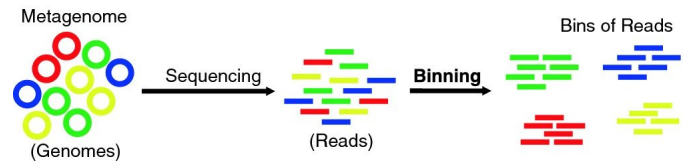


Fig. 5. Demonstration for the binning algorithm

III. TAXONOMIC PROFILLING

The identification and classification of microbial species within a given metagenomic dataset is made possible by taxonomic profiling, which is a crucial step in metagenomic analysis. In this procedure, DNA or protein sequences acquired from the understudied microbial populations are given taxonomic labels. A number of techniques have been developed that use various strategies to address the problems that this field is abundant with. These techniques can be roughly divided into machine learning-based techniques, composition-based techniques, and approaches based on sequence similarity.

The comparison of query sequences to a reference database of recognised taxonomic sequences is the foundation of sequence similarity-based methods. BLAST (Basic Local Alignment Search Tool), a popular tool for taxonomic profiling, stands out among these techniques. By using sequence similarity to find the closest matches between the query sequences and database entries, BLAST—which was first developed for sequence alignment—has been modified for taxonomy assignment. With the aid of BLAST, researchers can determine the taxonomic affiliation of the query sequences by looking at the best hits and their related taxonomy. Its extensive use in metagenomic research has produced insightful data on the diversity and abundance of various taxa, shedding light on the taxonomic makeup of microbial communities [9].

Utilizing the distinctive compositional patterns in DNA or protein sequences, composition-based techniques like k-mer frequencies, Kraken, and MetaPhlAn can identify the taxonomic origins of an organism. These approaches use machine learning or statistical models to infer taxonomic labels based on the presence and abundance of particular sequence characteristics. To assign taxonomic labels, Kraken, for instance, compares query sequences to a precomputed database of reference sequences using a k-mer-based methodology. It creates a taxonomic classification system using the k-mer frequencies, enabling quick and precise taxonomic classifications. The metagenomic taxonomic profiling programme MetaPhlAn, on the other hand, was created exclusively for marker gene analysis. To recognise and count the number of microbial species present in metagenomic samples, it makes use of a set of clade-specific marker sequences [10].

Taxonomic profiling techniques provide useful information, but they can have some drawbacks. The precision and scope of taxonomic assignments may be constrained by fragmented or lacking reference databases. Additionally, due to taxonomy's inherent variety and hierarchical structure, there may be ambiguities when taxa are assigned to various taxonomic levels. Nevertheless, several typical methods, such as BLAST, Kraken, and MetaPhlAn, each with its own advantages and disadvantages, have shown their effectiveness in taxonomic profiling. These methods take into account multiple machine learning techniques, statistical models, and search strategies to meet the diverse requirements of metagenomic datasets.

A. BLAST (Basic Local Alignment Search Tool)

BLAST is a prominent sequence similarity-based algorithm used for taxonomic profiling in metagenomic studies. Originally developed for sequence alignment, BLAST has been adapted to compare query sequences against a reference database of known taxonomic sequences. By leveraging sequence similarity, BLAST identifies the closest matches between the query sequences and the entries in the reference database. Through analyzing the best hits and their associated taxonomy, BLAST allows researchers to infer the taxonomic affiliation of the query sequences. It offers different search implementations tailored to specific sequence types, such as BLASTN for nucleotide sequences and BLASTP for protein sequences. Each implementation employs specific algorithms and techniques suited for their respective sequence types, enabling accurate and efficient comparisons. For example, BLASTN utilizes the megablast algorithm for highly similar nucleotide sequences, while discontinuous megablast is suitable for more divergent sequences. By employing the BLAST algorithm and its various search implementations, taxonomic or functional classification of query sequences based on their similarity to sequences within the reference database can be achieved (see Figure 6) [11].

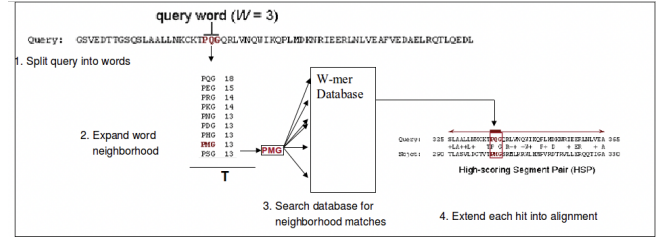


Fig. 6. Schematic Representation of the BLAST Taxonomic Profiling Algorithm

B. Kraken

Kraken is a composition-based method that utilizes k-mer frequencies to perform taxonomic classification of metagenomic sequences. It compares query sequences to a precomputed database of reference sequences and analyzes the presence and abundance of specific k-mers to assign taxonomic labels. Kraken's approach allows for rapid and accurate taxonomic assignments, making it a valuable tool in metagenomic analysis (see Figure 7).

During the classification process, Kraken uses a taxonomy tree, also known as a classification tree, which represents the hierarchical relationships among different taxonomic groups. By traversing the taxonomy tree, Kraken determines the classification path for each query sequence based on the identified k-mers. This path represents the taxonomic lineage or hierarchy to which the sequence belongs, from the root of the tree (e.g., the highest taxonomic level) down to the specific taxonomic group or species.

The accuracy of Kraken's taxonomic assignments heavily relies on the completeness and quality of the reference

database used. A comprehensive and well-curated reference database enhances the accuracy of classification by providing a broader representation of different taxonomic groups. Conversely, an incomplete or biased reference database may lead to misclassifications or missed assignments. Therefore, ensuring the quality and breadth of the reference database is essential for obtaining reliable taxonomic classifications using Kraken.

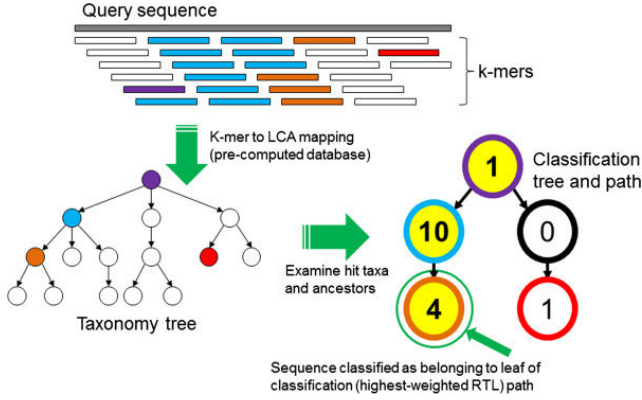


Fig. 7. Illustration of the Kraken Taxonomic Profiling Algorithm

C. MetaPhlAn

MetaPhlAn is a highly specialized method specifically designed for metagenomic taxonomic profiling, with a particular emphasis on marker gene analysis. This innovative approach relies on a carefully curated collection of clade-specific marker sequences, which are essential for accurately identifying and quantifying microbial species within metagenomic samples. By leveraging the power of these marker genes, MetaPhlAn enables researchers to gain deep insights into the taxonomic composition and relative abundance of various taxa present in microbial communities.

The key principle underlying MetaPhlAn's functionality lies in aligning query sequences against the comprehensive marker database it utilizes. Through this alignment process, MetaPhlAn effectively determines the taxonomic origin of the query sequences, providing valuable information about the microbial species present in the sample. This information, in turn, allows researchers to assess the relative abundance of different taxa, gaining a better understanding of the overall taxonomic structure of the microbial community under investigation.

The capacity of MetaPhlAn to precisely analyze and quantify changes in community structure between various samples is one of its significant advantages. Variations in the composition and number of microbial species can be found by comparing the taxonomic profiles of several samples. This capacity is especially helpful in long-term studies or research projects looking at how environmental factors or interventions affect the microbiome. The marker-based approach used by MetaPhlAn provides a trustworthy and effective way to record these differences, facilitating the investigation of intricate microbial dynamics [12].

By offering important insights into the taxonomic makeup of microbial communities, MetaPhlAn also advances our understanding of the microbiome. MetaPhlAn can discover and identify many microbial species by utilizing its marker database, illuminating the richness and diversity of microbial life. Understanding the potential functions and ecological functions of various microorganisms within their distinct communities depends on having this knowledge.

D. Summary of Taxonomic Profiling Algorithms

In taxonomic profiling, BLAST compares query sequences to a reference database based on sequence similarity, offering flexible search methods. Kraken uses a taxonomy tree for categorization using k-mer frequencies to swiftly and precisely assign taxonomic labels. When analyzing changes in community structure, MetaPhlAn focuses on marker gene analysis and uses clade-specific markers to identify and quantify microbial species. Every algorithm has its benefits and is designed to satisfy a need in metagenomic analysis (see Table II).

TABLE II
COMPARISON OF TAXONOMIC PROFILING ALGORITHMS

Algorithm	Approach	Key Features
BLAST	Sequence similarity-based	<ul style="list-style-type: none"> Compares query sequences to a reference database Provides search implementations for different sequence types Comprehensive database for metagenomic research
Kraken	Composition-based	<ul style="list-style-type: none"> Uses k-mer for taxonomic classification Rapid taxonomic assignments Relies on a taxonomy tree for classification
MetaPhlAn	Marker gene analysis	<ul style="list-style-type: none"> Uses clade-specific marker sequences Enables identification and quantification of microbial species Captures changes in community structure between samples

IV. CASE STUDIES AND APPLICATIONS

Numerous research fields have used taxonomic profiling and metagenomics assembly to gain important understanding of microbial communities. Here, we give case examples that demonstrate how these methods have been applied in certain fields, like environmental microbiology and the human gut microbiome, highlighting the knowledge that has been gathered and how it has affected our understanding of microbial communities.

A. Environmental Microbiology

Metagenomics assembly and taxonomic profiling have transformed our understanding of microbial diversity and functional capacity in intricate ecosystems, which has profound implications for environmental microbiology. For instance, metagenomics assembly was used in a marine investigation to rebuild the genomes of uncultivated bacteria. The entire genomes of marine microorganisms were successfully assembled using a combination of overlap-layout-consensus and de Bruijn graph approaches [13]. The taxonomic composition and functional potential of these marine microbial communities were characterized using taxonomic profiling tools like BLAST and composition-based approaches like Kraken. The research illuminated previously unknown microorganisms' ecological functions by identifying functional genes involved in nitrogen cycle and revealing unique microbial lineages.

B. Human Gut Microbiome

The human gut microbiota is extremely important for both health and disease. Understanding the complexity of the gut microbiome and how it affects human physiology has been made possible with the use of metagenomics assembly and taxonomic profiling. Metagenomics assembly methods like MetaSPAdes and MEGAHIT were used to reconstruct microbial genomes from fecal samples of individuals from various cultures in a thorough analysis of the gut microbiome [14]. The gut microbiome's microbial species were identified and categorized using taxonomic profiling techniques like BLAST and Kraken. Researchers identified links between particular gut microorganisms and illnesses like obesity, inflammatory bowel disease, and metabolic disorders by examining the taxonomic mix and functional potential of these microbial communities. These findings have paved the way for targeted interventions and personalized treatments based on the manipulation of the gut microbiome.

C. Overall Impact

Our understanding of microbial communities has been completely transformed by the use of metagenomics assembly and taxonomic profiling in a variety of scientific fields. These methods have provided researchers with unprecedented insights into the composition, potential applications, and ecological functions of microbes in complex ecosystems. Researchers can reconstruct the genomes of uncultivated bacteria through metagenomics assembly, allowing for the study of their genetic makeup and metabolic capabilities. These methods, when combined with taxonomic profiling, have increased our understanding of microbial diversity, discovered new lineages, and clarified the functional contributions of microbes to various ecosystems. These case studies' and applications' learned insights have broad ramifications for many disciplines, including biotechnology, environmental science, and human health. The creation of new treatments, bioremediation techniques, and environmental monitoring methods has been made possible by the ability to characterize microbial communities at a high

resolution level. Moreover, by enabling customized interventions and therapies based on a person's microbiota makeup, better understanding of the relationship between the human microbiome and health has the potential to revolutionize personalized medicine.

ACKNOWLEDGMENT

We would like to express our gratitude to Allah for granting us the knowledge, opportunities, and blessings that have enabled us to undertake this research. We acknowledge His guidance and inspiration throughout our journey.

We would also like to thank our fellow researchers and colleagues for their collaboration and contributions to this project. Their insights and discussions have shaped our understanding and findings.

We appreciate the academic community for their extensive research and literature, which has provided a strong foundation for our study.

Our sincere thanks go to the faculty and staff of our institution, as well as to our families and friends for their unwavering encouragement.

REFERENCES

- [1] Reich JG, Drabsch H, Diumler A. On the statistical assessment of similarities in DNA sequences. *Nucleic Acids Res.* 1984;12(13):5529–5543.
- [2] 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.
- [3] Jeck WR, Reinhardt JA, Baltrus DA. et al. Extending assembly of short DNA sequences to handle error. *Bioinformatics.* 2007;23(21):2942–2944.
- [4] Venter JC, Adams MD, Myers EW. et al. The sequence of the human genome. *Science.* 2001;291(5507):1304–1351.
- [5] Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–829.
- [6] 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.
- [7] Bankevich A, Nurk S, Antipov D. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol.* 2012;19(5):455–477.
- [8] Compeau, P., Pevzner, P. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29, 987–991 (2011). <https://doi.org/10.1038/nbt.2023>.
- [9] Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics.* 2019 Jun;20(6):356–70.
- [10] Comin M, Di Camillo B, Pizzi C, Vandin F. Comparison of microbiome samples: methods and computational challenges. *Briefings in bioinformatics.* 2021 Jan;22(1):88–95.
- [11] Oulas A, Pavlodi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos L. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and biology insights.* 2015 Jan;9:BB1-S12462.
- [12] Gacesa R, Kurilshikov A, Vich Vila A, Sinha T, Klaassen MA, Bolte LA, Andreu-Sánchez S, Chen L, Collij V, Hu S, Dekens JA. Environmental factors shaping the gut microbiome in a Dutch population. *Nature.* 2022 Apr 28;604(7907):732–9.
- [13] Bishara A, Moss EL, Kolmogorov M, Parada A, Weng Z, Sidow A, Dekas AE, Batzoglu S, Bhatt AS. Culture-free generation of microbial genomes from human and marine microbiomes. *BioRxiv.* 2018 Feb 11:263939.
- [14] Karlsson F, Tremaroli V, Nielsen J, Bäckhed F. Assessing the human gut microbiota in metabolic diseases. *Diabetes.* 2013 Oct 1;62(10):3341–9.