Using De Novo Metagenome Assembly for Improved Metagenomic Classification

Josipa Lipovac*, Krešimir Križanović*
* Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, Croatia
josipa.lipovac@fer.hr
kresimir.krizanovic@fer.hr

***Abstract*—Metagenomics is a rapidly growing field that allows for studying complex microbial communities. One of the first steps in the metagenomic analysis is the classification of the organisms present in a sample. This is usually done by comparing sequencing reads to a database of known organisms. With the recent development of long-read sequencing technologies, such as PacBio and Oxford Nanopore Technologies (ONT), it is now possible to generate highly accurate assemblies of genomes from metagenomic samples. This is typically done using a combination of reference-based and de novo assembly approaches. Assembling the genomes from the metagenomic sample, prior to classification, could improve classification results and also aid in identifying new, previously unknown species. However, the evaluation of metagenome assemblies is a challenging task and it is important to assess the quality of the assemblies in order to ensure the accuracy of downstream analyses. In this paper, we provide a detailed overview of metagenomic classification, de novo metagenome assembly process, and evaluation of metagenome assembly, highlighting various tools and techniques currently available for each step. We also present initial results showing that metagenomic classification can benefit from a previously assembled metagenome.**

***Keywords*—metagenomics, metagenomic classification, metagenome assembly, metagenome assembly evaluation**

I. INTRODUCTION

The term microbiota refers to the microbial population present within the human body, including bacteria, viruses, archaea, protozoans, and fungi [1]. Over the course of genetic evolution, humans have formed symbiotic relationships with various microbes. To understand these relationships and the role of microbes in human health and disease, it is necessary to identify these microbes and determine their genome sequences [2].

Metagenomic sequencing is an alternative approach that allows for direct sequencing of a mixture of microbial DNA without the need for isolation, providing a more comprehensive understanding of microbial genomes with diverse characteristics. The field of metagenomics has seen significant growth in recent years due to increased interest in microbial communities and the development of techniques for analyzing their diversity and genetic potential [3].

Partial sequencing of microbiota DNA can provide information on the diversity of a given community, but more comprehensive insights into the genetic potential of the microbiome require the analysis of extended genomic regions or fully reconstructed genomes, which can be obtained through the use of metagenome assemblers [4]. Metagenomics involves sequencing and analysis of the genomic DNA of entire microbial communities in environmental samples, allowing researchers to better understand the makeup of these communities.

PacBio [5] and Oxford Nanopore Technologies (ONT) [6] are DNA sequencing methodologies that generate long reads with average lengths of 10-25 kilobase pairs (kbp) for Pacbio and 10-100 kbp for ONT. Recent advancements in ONT have enabled the production of ultra-long reads exceeding 1 megabase pair (Mbp). In metagenomic sequencing, long reads can be particularly useful as they provide more continuous DNA sequences and facilitate the identification of structural variations within a genome.

De novo metagenome assembly is the process of reconstructing genomic sequences of microorganisms from DNA sequencing reads without the use of reference genomes. This process involves the use of specialized software tools to assemble the reads into longer contiguous sequences, known as contigs. The complexity and computational intensity of this process may be heightened due to the potential for a high number of reads originating from diverse microorganisms, as well as the potential for significant differences between the reads.

One crucial aspect of the metagenomic analysis is metagenomic classification, which typically involves comparing the sequencing reads to databases of known genes and genomes to identify the microorganisms present in the sample. Once the microorganisms present in the sample have been identified, the genes present in the reads can be annotated by assigning functions to the genes based on their known or predicted roles. However, this approach may be problematic when the database is too large or when there is a new, unknown species in the sample. To address these issues, first, the metagenome can be assembled, and then metagenomic classification can be performed using the obtained contigs.

In order for contigs generated through metagenome assembly to be utilized for metagenomic classification, it is essential to accurately evaluate them. Many existing evaluation tools rely on the availability of a reference genome, which can complicate the evaluation process if the sample contains previously unknown species.

II. Metagenomic classification

Metagenomic classification is the process of accurately identifying all species present in a sample and is a challenge because metagenomic sequencing generates genomic data from a mixture of species. During metagenomic classification, reads are assigned to taxonomic groups using various methods. This involves comparing the reads to databases of known reference genomes, which can be used to infer the taxonomic identity of the original organisms. Several classifiers have been developed to help with this process, but choosing the best method requires understanding the characteristics and limitations of each.

The first method for assigning taxonomic labels to unknown reads was using BLAST (basic local alignment and search tool) [7], a high-sensitivity DNA alignment tool. But, due to its high computational demands, it's infeasible for large metagenomic sequencing data.

Metagenomic classifiers can be roughly divided into two groups [8]:

1) **k-mer-based**: Kraken2 [9], Centrifuge [10], CLARK [11], CLARK-S [12], etc.
2) **mapping-based**: MetaMaps [13], MEGAN-LR [14], Minimap2 [15], MetaPhlAn3 [16], etc.

K-mer-based metagenomic classifiers use the presence and frequency of k-mers to classify DNA sequences to species or higher taxonomic groups. They compare the k-mer count of a query sequence to a reference database of labeled sequences and assign the query to the group with the highest similarity score. These classifiers are fast but often have lower accuracy (reporting many false positives) and can be affected by k-mer length and reference database quality. Kraken2 is a k-mer-based metagenomic classifier tool with fast classification and reduced memory usage. Its reference database is built from the NCBI RefSeq database [17] and offers high accuracy, especially at the genus level. However, it may be difficult to run on lower-end computers due to its high memory requirement. MiniKraken [18] is a smaller, less accurate version that requires less memory. Kraken2 also includes features like a spaced seed search scheme and compatibility with the Bracken algorithm [19] for estimating species-level sequence abundance. The Centrifuge uses Burrows-Wheeler Transform (BWT) and Ferragina-Manzini (FM) index for sequence storage and mapping. The transformed string from BWT enables efficient compression and string matching. FM index combines BWT and suffix array for efficient searching. CLARK and CLARK-S classify metagenomic reads by comparing them to a reduced k-mer database and classifying them as the target with the highest shared k-mers. CLARK-S has specific databases for bacteria, viruses, and fungi.

Mapping-based metagenomic classifiers use a sequence alignment algorithm to align the reads generated from metagenomic sequencing to a reference database. The basic idea is to identify the reference sequences that the reads in the query sample map to with the highest degree of similarity and use this information to classify the reads to the species or taxonomic group they belong to. These classifiers have the advantage of being able to classify reads with high accuracy, but they can be computationally intensive and may not be suitable for very large datasets. Minimap2 is a sequence alignment tool that aligns reads to a reference genome. In order for Minimap2 to be used as a metagenomic classifier, its output needs to be reformatted and specially interpreted, as the authors did in a recent study [8] where it achieved fairly successful results. When it encounters a sequence that is not in its database, Minimap2 adds it to the most similar organism, which helps increase its classification accuracy on the genus level. However, it is slower than k-mer-based tools. MetaMaps employs probabilistic scoring to approximate sample composition, while MEGAN-LR, built on MEGAN6, interprets long nucleotide sequences using a translation alignment method. MEGAN-LR assigns reads to taxa using the LCA (Lowest Common Ancestor) algorithm, along with other factors such as lcaCoveragePercent, minSupportPercent, and minPercentReadCover. In contrast, MetaPhlAn3 calculates the relative abundance of taxa by mapping reads to a database of unique, clade-specific marker genes and using coverage scores.

Metagenomic classifiers rely on a pre-computed database. A large and rapidly increasing size of such databases can make metagenomic analysis computationally demanding [20]. The most widely used reference databases are RefSeq complete genomes for microbial species, and the BLAST database [21]. Other commonly used databases include SILVA [22], which contains 16S ribosomal RNA sequences, and Genbank [23], which contains a larger number of genomes with lower quality control standards. Many metagenomics tools allow users to create their own reference database based on a specific set of sequences. While this process can be computationally demanding, especially for large databases, it gives users greater control over the analysis, particularly when studying rare, newly discovered, or highly diverse species. Using a uniform reference database is important when comparing results from different classifiers to prevent potential confounding effects from differences in default databases.

Benchmark papers show that most classification tools that obtain a low false positive rate tend to have a lower recall. Due to the inherent variability in the specific needs of metagenomic analyses, it is challenging to establish a universally superior classification tool. Longer reads tend to improve accuracy, but using only the longest reads can reduce accuracy due to varying read length distributions. Furthermore, the presence of novel species within a metagenomic sample has been found to have a significant impact on classification accuracy, yet this aspect has not been extensively investigated in previous benchmark studies [8] [24]. Currently, most metagenomic tools classify unknown species into a genus or report them as unclassified. A better solution would be having tools

that not only detect unknown species but also provide information on their closest related species. This can be achieved through metagenome assembly and contig classification. A large contig classified under a genus could indicate a previously unknown species in the sample.

III. Metagenome assembly

De novo genome assembly is the process of reconstructing a full genomic sequence from DNA reads without the use of a reference genome. It is more challenging than reference-based genome assembly, which uses a known genome as a starting point. De novo genome assembly is necessary when studying organisms for which no reference genome is available, and it is accomplished using computational algorithms that align the reads and reconstruct the full genomic sequence. Metagenome assembly is the process of reconstructing sequences from a metagenomic sample to obtain a representation of the genomes present in the sample. The full genetic potential of a microbial community can be determined by analyzing extended genomic regions or complete genomes, obtainable through modern sequencing technologies. However, genome assembly of metagenomic data is challenging due to factors such as large data volume, uneven representation of community members, and the presence of multiple strains and closely related microorganisms.

Tools for metagenome assembly can be divided into three groups:

1) **short-reads assemblers**: Omega [25], MetaVelvet [26], IDBA-UD [27], MEGAHIT [28], RayMeta [29], etc.
2) **long-reads assemblers**: Canu [30], metaFlye [31], HifiAsm-meta [32], etc.
3) **hybrid assemblers** metaSPAdes [33], DBG2OLC [34], OPERA-MS [35], etc.

Short-read metagenome assemblers typically have more difficulty reconstructing genomic sequences of the individual microorganisms present in a metagenomic sample compared to long-read metagenome assemblers. Short reads are typically 100-150 bp long and are very accurate (error rate is around 0.5%). On the other hand, long reads are several thousand bp long, so even though they are more error-prone, they provide more information about the genomic sequence, making it easier for the assembler to distinguish between the different genomes and correctly reconstruct the individual sequences and handle repetitive regions of the genome. Assembling a metagenome can be difficult, especially with short reads, but even an incomplete and fragmented assembly can help improve metagenomic classification, particularly in identifying new species. Omega uses hash tables to store the prefix and suffix sequences of each read and then uses these sequences to construct a bi-directed graph by linking the reads with their overlapping sequences. MetaVelvet creates a de Bruijn graph using Velvet and then divides it into subgraphs using coverage peaks of k-mers to separate different microbial genomes. IDBA-UD tries to trim the graph iteratively and merge bubbles using increasing k-mer sizes. MEGAHIT combines the process of selecting k-mer sizes with a succinct de Bruijn graph and has strong computational efficiency, while RayMeta generates the local depth distribution for each seed path in a de Bruijn graph.

Hybrid metagenome assembling merges the benefits of long reads (contiguous sequences) and short reads (error correction) to achieve more accurate and complete assemblies. A popular metagenome assembler is metaSPAdes, which combines various assembler strengths including the OLC (overlap-layout-consensus) approach, using long reads for assembly and short reads for error correction. DBG2OLC uses OLC and OPERA-MS combines long reads with low coverage and short reads to create a scaffold graph and group contigs into species-specific clusters using a Bayesian clustering algorithm that leverages read depth and long-read connections.

Long-read metagenome assemblers generate accurate and complete assemblies by producing contiguous sequences that can resolve complex regions of the genome. Canu is a long-read metagenome assembler that generates assemblies using the OLC approach. It error-corrects high-noise, high-coverage long reads, then uses "unitigging" to construct contigs from the graph representation of the genome/metagenome. MetaFlye generates genome assemblies from high-coverage long reads. It starts with constructing a de Bruijn graph from input reads, then identifies paths most likely to correspond to genomes. Error correction and polishing are done by aligning reads to the genome assemblies and correcting errors to improve the final genome quality. HifiAsm-meta uses the de Bruijn graph to identify probable paths of individual genomes from low-error PacBio-HiFi reads. Contigs are generated from these paths, which can then be used to reconstruct MAGs (Metagenome-Assembled Genomes) with the binning algorithm.

Most of the current assemblers do not represent complete microbial genomes with a single sequence so different binning algorithms are used. Metagenome binning algorithms group sequences into clusters. Many algorithms use features like TNFs (taxonomic novelty features), k-mer frequencies, and read depth to differentiate sequences and assign them to bins. MetaBAT2 [36] is a metagenome binning tool that uses TNFs and read depths to compute sequence similarities and partitions the graph of similarities into subgraphs using a modified label propagation algorithm (LPA).

IV. Evaluation of metagenome assembly

The evaluation of the quality of contigs assembled through metagenomic sequencing is a challenging problem. High-coverage regions of the contig, which are defined as regions that are overlapped by a large number of reads, are often associated with high-quality contigs. However, the development of a method for quantification

of these observations into a comprehensive quality score for the entire contig is a topic of ongoing research.

The majority of methods employed for the assessment of the quality of metagenomic assemblies depend on the availability of a set of annotated reference genomes for comparative analysis. One such commonly used reference-based method is metaQUAST [37], which aligns the contigs of the assembly to the reference genomes and subsequently computes various statistical measures such as: number of contigs, length of the largest contig, total length, N50, NG50, etc. MetaQUAST is a customized metagenomics version of the QUAST tool [38], which is one of the most widely used tools for genome assembly evaluation. Additionally, metaQUAST provides the capability to detect and visualize misassemblies and extract unrepresented genomic regions. However, it can be challenging to select a reference for novel MAGs from distantly related organisms, and the presence of similar organisms can lead to incorrect contig assignments. For the listed reason, the evaluation of the metagenomic assembly would be much more accurate and realistic if methods without reference requirements were used. An example of such a method is CheckM [39], which be utilized to evaluate the quality of metagenome assemblies by determining their completeness and contamination. The measures used by CheckM are based on the occurrence of specific genetic loci, and thus do not evaluate genome assembly at the level of individual contigs.

Repetitive genomic regions within the same genome or conserved sequences shared among different organisms can lead to assembly errors, including both inter- and intra-genome misassemblies. This is particularly likely to occur when multiple strains that are closely related are present in the same environment [40]. There are several existing reference-free methods for evaluating contig misassembly. ALE [41] provides nucleotide-level likelihood scores for assembled contigs, but not contig-level quality scores. SuRankCo [42] uses machine learning to provide contig quality scores based on length and coverage. VALET [43] detects misassemblies by combining multiple metrics extracted from the alignment of reads to contigs. DeepMAsED [44] is a deep learning-based tool for evaluating metagenomic assemblies by predicting the error rate using a CNN that analyzes k-mer content. The tool is trained on error-free reference assemblies and then used to predict the error rate of new assemblies by comparing k-mer content to the reference assemblies. metaMIC [45] is a tool for identifying and correcting misassemblies in metagenomic assemblies by localizing breakpoints. It uses features from both reads and assemblies, such as read coverage and k-mer consistency, to detect both intra- and inter-genome misassemblies. It can be adapted to work with assemblies from various assembler tools.

V. Test experiment

A. Experimental setup

To demonstrate the impact of metagenome assembly on metagenomic classification, we conducted two experiments, one based on species level and the other on strain level classification. First, we generated three metagenomic samples by simulating reads from known references using the Badread tool [46]. Two simulated datasets were used for experimental setup 1 and the third one was used for experimental setup 2. The simulation included ONT and PacBio long reads, incorporating the respective built-in qscore and error models (nanopore2020 and pacbio2016). To investigate differences between ONT and PacBio sequencing technologies, experimental setup 1 included simulated reads for both of them. Because the results showed no significant difference between the technologies, experimental setup 2 included only ONT simulated reads.

The presence of similar or low abundant organisms in the metagenomic sample may result in poor classification or assembly, however, we are using simulated reads with sufficient coverage. In contrast to actual metagenomic samples, our samples are characterized by a reduced number of organisms. The selection of a reduced number of organisms was done deliberately to facilitate the evaluation of the entire method. In the first sample, organisms were chosen based on their shared taxonomic levels, specifically order. The second metagenomic sample was simulated using organisms that are representative of those present in Zymo D6331 Gut Microbiome Standard metagenomic sample, where their relative abundance is well-established. This methodology ensures the preservation of authentic inter-organism associations present in the original sample. Simulated samples also contain errors, random reads as well as chimeric reads and thereby additionally reflect real samples. In addition, the second dataset contains five strains of E.coli, which also reflects the real situation, because metagenomic samples generally contain multiple strains of the same species, which makes strain level classification more challenging.

Once we simulated the reads, we performed metagenomic classification using the Kraken2 as a tool that provides the best balance between speed and accuracy. After the initial classification of raw reads, we assembled metagenomic samples using the metaFlye as a state-of-the-art metagenome assembly tool and additionally performed classification on the obtained contigs. Our main hypothesis was that assembling the metagenomic sample could enhance strain level classification and reduce the number of false positives.

To compare the classification outcomes, we took into account the abundance of individual organisms. Specifically, we estimated the approximate abundance of assembled contigs by considering their length and coverage. To calculate the approximate number of reads utilized for assembling a single contig, we applied the following

formula:

$$num\_of\_reads = ceil(\frac{contig\_len * contig\_cov}{mean\_read\_len}) \quad (1)$$

where $contig\_len$ and $contig\_cov$ are the length and coverage of the contig for which we estimate the number of reads, and $mean\_read\_len$ is the mean length of all reads in the dataset.

B. Results and discussion

Table I presents the number of false positive classifications on species and strain level for reads and contigs classification. From the results, it can be seen that with contigs classification the number of false positive species and strain identifications is significantly lower.

To reduce the number of incorrect classifications due to sequencing errors and other causes, a strain or species is considered identified if at least 50 reads are assigned to it. The used threshold (50 reads) was determined based on the study [8]. However, further investigation is necessary to establish the optimal threshold. Setting the threshold too high may result in true positive (TP) classifications reduction.

Table II shows calculated abundances for experimental setup 1. It can be seen that the classification at the species level is quite successful in both cases, but the abundance values of the contigs classification are still closer to the actual abundance values for all organisms. It can also be seen that there is no significant difference in the results for ONT and PacBio datasets. What additionally contributed to the success of the classification is the fact that all organisms in the sample are equally represented. However, even in this less complex scenario, it can be seen that the classification of reads reports a higher number of false positives and that the estimation of abundance is more accurate with the classification of contigs.

The calculated abundance values for experimental setup 2 are shown in Table III. The Table III does not show the results for 4 organisms: *Bifidobacterium adolescentis* (taxid: 367928), *Clostridioides difficile* (taxid: 1121308), *Escherichia coli* (taxid: 2605619), and *Roseburia hominis* (taxid: 585394). strain level classification for these organisms resulted in strains not being detected at all or being detected with abundance orders of magnitude lower than expected. Additionally, for 2 organisms (*Prevotella corporis* and *Veillonella rogosae*) the exact strain level was not known because the strain TAXID of those species is not present in the Kraken2 database. Abundance values for those two organisms were calculated at the species level.

In Table III, abundance values closer to the true abundances are presented in bold. The results show that there is a significant improvement in abundance values for contigs classification in 9 out of 14 cases. Additionally, in the case of non-bolded organisms with TAXIDs 83334 and 941322, the abundance values are quite different from the true values, but they are still better in the case of contigs classification.

TABLE I: **Number of FP classifications**

		reads classification	contigs classification
species level	experimental setup 1 ONT	14	2
	experimental setup 1 PacBio	15	1
	experimental setup 2 ONT	7	1
strain level	experimental setup 2 ONT	61	31

The results of experimental setup 2 show that, in general, strain level classification is significantly worse compared to species-level classification. Four of the strains were not detected by either method (reads and contigs), and in several cases reported abundance differs significantly from the true value. However, regardless of the fact that strain level classification in general requires further improvement, it is evident that the classification after assembling the metagenome gives better results in the context of the abundance of expected strains.

VI. Conclusion

The process of metagenomic analysis includes several steps, including metagenomic classification, de novo metagenome assembly, and the evaluation of metagenome assemblies. A variety of tools and techniques are available for each step, including long-read sequencing technologies such as PacBio and ONT, which have increased the accuracy of assemblies. It is challenging to identify a single tool that is superior to others, as most tools are well-suited to specific data or metagenomic analysis tasks.

One crucial aspect for future advancements in the field is the simulation of realistic metagenomic samples, which can greatly enhance the successful implementation of new tools for metagenomic classification, assembly, and assembly evaluation. The use of realistic datasets can provide a better understanding of the complexity of real-world samples and improve the performance of the new tools.

While single-genome assembly techniques have seen significant success in reconstructing a majority of genomes, metagenomic assemblers continue to face challenges in achieving high-quality reconstruction of the organisms present in a sample. It is necessary to develop appropriate methods for properly segregating reads according to the organisms present in the sample. Additionally, in order to evaluate the performance of metagenomic assemblers, quality assessment tools are needed, which particularly have difficulties when similar organisms are present in the sample, as it makes it challenging for these tools to distinguish between them.

In the context of this research, two simple experiments were conducted to test if there are indications of improvement in metagenomic classification with previous metagenome assembly. The datasets that we used in these experiments, although they contain errors as well as ran-

TABLE II: **Results for experimental setup 1.**

	ONT			PacBio		
species name - TAXID	true abundance	reads abundance	contigs abundance	true abundance	reads abundance	contigs abundance
Staphylococcus aureus - 1280	7,9%	8,7%	**7,9%**	8,0%	8,8%	**8,0%**
Staphylococcus epidermidis - 1282	7,0%	10,1%	**6,9%**	7,1%	10,6%	**7,0%**
Streptococcus mutans - 1309	5,7%	6,9%	**5,9%**	5,8%	7,6%	**5,7%**
Streptococcus agalactiae - 1311	5,3%	3,7%	**5,4%**	5,3%	3,6%	**5,3%**
Enterococcus faecalis - 1351	8,2%	10,4%	**8,4%**	8,2%	10,7%	**8,2%**
Bacillus cereus - 1396	15,0%	9,1%	**14,7%**	14,9%	9,3%	**14,9%**
Lactobacillus gasseri - 1596	6,1%	4,1%	**6,0%**	6,1%	3,4%	**6,1%**
Pseudomonas aeruginosa - 287	19,4%	20,1%	**19,2%**	19,3%	18,6%	**19,4%**
Acinetobacter baumannii - 470	10,5%	12,8%	**10,8%**	10,5%	12,8%	**10,6%**
Escherichia coli - 562	14,9%	14,0%	**14,7%**	14,8%	14,5%	**14,9%**

TABLE III: **Results for experimental setup 2.**

species name - strain TAXID	true abundance	reads abundance	contigs abundance
Akkermansia muciniphila - 349741	4,9%	0,2%	**6,1%**
Bacteroides fragilis - 295405	10,0%	**8,8%**	12,1%
Clostridium perfringens - 451752	6,5%	1,8%	**7,8%**
Enterococcus faecalis - 936153	5,9%	29,0%	**7,2%**
Escherichia coli strain 1 - 83334	10,7%	7,2%	13,9%
Escherichia coli strain 2 - 2778656	9,4%	3,4%	**7,8%**
Escherichia coli strain 3 - 941322	9,7%	0,9%	3,0%
Escherichia coli strain 5 - 83333	8,8%	1,5%	1,0%
Faecalibacterium prausnitzii - 657322	6,3%	8,7%	**7,4%**
Fusobacterium nucleatum - 469607	4,4%	6,8%	**5,2%**
Lactobacillus fermentum - 1381124	4,4%	**5,1%**	5,5%
Prevotella corporis - 28128	5,8%	8,8%	**7,1%**
Salmonella enterica - 59201	8,8%	11,0%	**10,6%**
Veillonella rogosae - 423477	4,4%	6,9%	**5,5%**

dom and chimeric reads, do not fully reflect the real situation due to the composition of the sample. However, even in such a less complex case, the results of the experiments show that the classification after the assembly of the metagenome significantly improves the estimation of the actual number of organisms in the sample, both at the species level and at the strain level. In addition, the experiments show that assembling the metagenome results in a smaller number of false positive classifications.

Further research of this approach is necessary, which would include simulation of more realistic datasets (larger number of species, more uneven true abundances, more strains of the same species, organisms that are not present in the database...) and testing other tools for classification and metagenome assembly. Also, it is necessary to investigate a method of evaluating the metagenome assembly to be used for further classification, including finding a correct way to detect contigs that do not contribute to the correct classification due to poor or incorrect assembly.

VII. ACKNOWLEDGMENTS

This research was funded by the Croatian Science Foundation under the grant IP-2018-01-5886 (SIGMA) and by the European Union through the European Regional Development Fund, under the grant KK.01.1.1.01.0009 (DATACROSS).

REFERENCES

[1] S. M. Bakhtiar, J. G. LeBlanc, E. Salvucci, A. Ali, R. Martin, P. Langella, J.-M. Chatel, A. Miyoshi, L. G. Bermúdez-Humarán, and V. Azevedo, "Implications of the human microbiome in inflammatory bowel diseases," *FEMS Microbiology Letters*, vol. 342, no. 1, pp. 10–17, 05 2013. [Online]. Available: https://doi.org/10.1111/1574-6968.12111

[2] T. Korem, D. Zeevi, J. Suez, A. Weinberger, T. Avnit-Sagi, M. Pompan-Lotan, E. Matot, G. Jona, A. Harmelin, N. Cohen *et al.*, "Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples," *Science*, vol. 349, no. 6252, pp. 1101–1106, 2015.

[3] A. Latorre-Pérez, P. Villalba-Bermell, J. Pascual, and C. Vilanova, "Assembly methods for nanopore-based metagenomic sequencing: a comparative study," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[4] A. L. Lapidus and A. I. Korobeynikov, "Metagenomic data assembly–the way of decoding unknown microorganisms," *Frontiers in Microbiology*, vol. 12, p. 613791, 2021.

[5] A. Rhoads and K. F. Au, "Pacbio sequencing and its applications," *Genomics, proteomics & bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015.

[6] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, "The oxford nanopore minion: delivery of nanopore sequencing to the genomics community," *Genome biology*, vol. 17, no. 1, pp. 1–11, 2016.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[8] J. Marić, K. Križanović, S. Riondet, N. Nagarajan, and M. Šikić, "Benchmarking metagenomic classification tools for long-read sequencing data," *BioRxiv*, pp. 2020–11, 2021.

[9] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with kraken 2," *Genome biology*, vol. 20, no. 1, pp. 1–13, 2019.

[10] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, "Centrifuge: rapid and sensitive classification of metagenomic sequences," *Genome research*, vol. 26, no. 12, pp. 1721–1729, 2016.

[11] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, "Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers," *BMC genomics*, vol. 16, no. 1, pp. 1–13, 2015.

[12] R. Ounit and S. Lonardi, "Higher classification sensitivity of short metagenomic reads with clark-s," *Bioinformatics*, vol. 32, no. 24, pp. 3823–3825, 2016.

[13] A. T. Dilthey, C. Jain, S. Koren, and A. M. Phillippy, "Strain-level metagenomic assignment and compositional estimation for long reads with metamaps," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[14] D. H. Huson, B. Albrecht, C. Bağcı, I. Bessarab, A. Gorska, D. Jolic, and R. B. Williams, "Megan-lr: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs," *Biology direct*, vol. 13, no. 1, pp. 1–17, 2018.

[15] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018.

[16] F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. M. Thomas *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3," *Elife*, vol. 10, p. e65088, 2021.

[17] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "Ncbi reference sequences (refseq): current status, new features and genome annotation policy," *Nucleic acids research*, vol. 40, no. D1, pp. D130–D135, 2012.

[18] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, pp. 1–12, 2014.

[19] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, "Bracken: estimating species abundance in metagenomics data," *PeerJ Computer Science*, vol. 3, p. e104, 2017.

[20] H. Y. Simon, K. J. Siddle, D. J. Park, and P. C. Sabeti, "Benchmarking metagenomics tools for taxonomic classification," *Cell*, vol. 178, no. 4, pp. 779–794, 2019.

[21] S. McGinnis and T. L. Madden, "Blast: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic acids research*, vol. 32, no. suppl_2, pp. W20–W25, 2004.

[22] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[23] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "Genbank," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D34–D38, 2005.

[24] D. M. Portik, C. T. Brown, and N. T. Pierce-Ward, "Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–39, 2022.

[25] B. Haider, T.-H. Ahn, B. Bushnell, J. Chai, A. Copeland, and C. Pan, "Omega: an overlap-graph de novo assembler for metagenomics," *Bioinformatics*, vol. 30, no. 19, pp. 2717–2722, 2014.

[26] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, "Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads," in *Proceedings of the 2nd ACM conference on bioinformatics, computational biology and biomedicine*, 2011, pp. 116–124.

[27] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin, "Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, 2012.

[28] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph," *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, 2015.

[29] S. Boisvert, F. Raymond, É. Godzaridis, F. Laviolette, and J. Corbeil, "Ray meta: scalable de novo metagenome assembly and profiling," *Genome biology*, vol. 13, no. 12, pp. 1–13, 2012.

[30] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation," *Genome research*, vol. 27, no. 5, pp. 722–736, 2017.

[31] M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. Smith *et al.*, "metaflye: scalable long-read metagenome assembly using repeat graphs," *Nature Methods*, vol. 17, no. 11, pp. 1103–1110, 2020.

[32] X. Feng, H. Cheng, D. Portik, and H. Li, "Metagenome assembly of high-fidelity long reads with hifiasm-meta," *Nature Methods*, pp. 1–4, 2022.

[33] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaspades: a new versatile metagenomic assembler," *Genome research*, vol. 27, no. 5, pp. 824–834, 2017.

[34] C. Ye, C. M. Hill, S. Wu, J. Ruan, and Z. S. Ma, "Dbg2olc: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[35] D. Bertrand, J. Shaw, M. Kalathiyappan, A. H. Q. Ng, M. S. Kumar, C. Li, M. Dvornicic, J. P. Soldo, J. Y. Koh, C. Tong *et al.*, "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes," *Nature biotechnology*, vol. 37, no. 8, pp. 937–944, 2019.

[36] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, p. e7359, 2019.

[37] A. Mikheenko, V. Saveliev, and A. Gurevich, "Metaquast: evaluation of metagenome assemblies," *Bioinformatics*, vol. 32, no. 7, pp. 1088–1090, 2016.

[38] A. Mikheenko, A. Prjibelski, V. Saveliev, D. Antipov, and A. Gurevich, "Versatile genome assembly evaluation with quast-lg," *Bioinformatics*, vol. 34, no. 13, pp. i142–i150, 2018.

[39] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome research*, vol. 25, no. 7, pp. 1043–1055, 2015.

[40] C. Kingsford, M. C. Schatz, and M. Pop, "Assembly complexity of prokaryotic genomes using short reads," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–11, 2010.

[41] S. C. Clark, R. Egan, P. I. Frazier, and Z. Wang, "Ale: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies," *Bioinformatics*, vol. 29, no. 4, pp. 435–443, 2013.

[42] M. Kuhring, P. W. Dabrowski, V. C. Piro, A. Nitsche, and B. Y. Renard, "Surankco: supervised ranking of contigs in de novo assemblies," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–7, 2015.

[43] N. D. Olson, T. J. Treangen, C. M. Hill, V. Cepeda-Espinoza, J. Ghurye, S. Koren, and M. Pop, "Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes," *Briefings in bioinformatics*, vol. 20, no. 4, pp. 1140–1150, 2019.

[44] O. Mineeva, M. Rojas-Carulla, R. E. Ley, B. Schölkopf, and N. D. Youngblut, "Deepmased: evaluating the quality of metagenomic assemblies," *Bioinformatics*, vol. 36, no. 10, pp. 3011–3017, 2020.

[45] S. Lai, S. Pan, C. Sun, L. P. Coelho, W.-H. Chen, and X.-M. Zhao, "metamic: reference-free misassembly identification and correction of de novo metagenomic assemblies," *Genome Biology*, vol. 23, no. 1, pp. 1–21, 2022.

[46] R. R. Wick, "Badread: simulation of error-prone long reads," *Journal of Open Source Software*, vol. 4, no. 36, p. 1316, 2019. [Online]. Available: https://doi.org/10.21105/joss.01316