Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France – August, 2025
Paper No. ICBES 156
DOI: 10.11159/icbes25.156

Novel Computational Pipeline for Comparative Whole-Genome Methylation Profiling In Bacteria

Michaela Zbudilová[1], Markéta Jakubíčková[1], Helena Vítková[1]
[1]Department of Biomedical Engineering, Brno University of Technology
Technická 12, Brno, Czech Republic
xzbudi00@vutbr.cz; jakubickova@vut.cz
vitkovah@vut.cz

Abstract **-** DNA methylation is an important epigenetic mechanism in bacteria, involved in processes such as gene regulation, virulence, and antimicrobial resistance. Despite advances in long-read sequencing technologies, comparing methylation patterns across multiple bacterial strains remains challenging, largely due to genomic variability and the lack of standardized analytical approaches. In this study, we present a novel pipeline for the analysis of DNA methylation in *Klebsiella pneumoniae* using data from nanopore sequencing. The workflow supports two alignment strategies: mapping to a universal reference genome or to a sample-specific reference constructed *de novo*. This flexible approach enables both broad inter-strain comparisons and detailed strain-specific analyses, facilitating consistent and interpretable investigation of methylation patterns across diverse bacterial genomes.

Keywords**:** nanopore sequencing, DNA methylation, whole genome sequencing, alignment, Klebsiella pneumoniae

1. Introduction

Methylation of deoxyribonucleic acids is the process of adding a methyl group to one of the nucleotides. This mechanism plays a crucial role in epigenetic regulation, which controls gene expression and cell DNA repair mechanisms. Methylation occurs across a wide range of organisms, from bacteria to plants and humans, and is involved in essential biological processes. [1]

Methylation can be detected on each type of nucleotide. In all cases, enzymes DNA methyltransferases add methyl group ($-CH_3$) to one of the nucleotides. Some methylation types are more common, especially 5-methylcytosine (5mC), 6-methyladenine (6mA) and 4-methylcytosine (4mC), where the numbers refer to the carbon atom position on the nitrogenous base where the methyl group is added. [2] Every type of methylation has specific characteristics, such as the types of organisms in which it occurs and the effects it has on them. DNA methylation in bacteria primarily regulates gene expression, while in mammals (with a specific focus on *Homo sapiens*), DNA methylation can influence X chromosome inactivation in order to silence one of the two female sex chromosomes during the early stages of embryonic development. [3]

Methylation detection is a process used to identify potential methylated positions in the genome during profiling techniques. These techniques can focus on a sample as a whole in contrast with typing techniques, which focus on the analysis of one part of the genome across all samples. The most commonly used profiling techniques for methylation detection are liquid chromatography (the first profiling technique used for this type of detection), electrophoresis, and sequencing techniques, which are nowadays primarily used because of their simplicity and detection speed. [4]

Nanopore sequencing from Oxford Nanopore Technologies (ONT) brought the biggest revolution in methylation detection. This sequencing technique is based on passing a sequence from one side of the membrane to another side through a small hole called a nanopore. During this transfer, we measure changes in electric current on the membrane because each nucleotide changes the electric current in a specific way, so after that, we are able, during basecalling, to assign to every current level specific nucleotide. Therefore, the output from nanopore sequencing is an electric signal over time that also contains information about methylation. [5]

This study focuses on *K. pneumoniae*, a bacterium that plays a significant role in human health. This bacterium can cause a wide range of diseases, like inflammatory diseases of the lungs or urinary tract infections, which are very common in real life. The main disadvantage in treating *K. pneumonia* infections is that this bacterium can often resist to antibiotics or rapidly develop resistance. Thus, managing these infections is more complicated than it appears. Therefore, detecting methylation in *K. pneumonia* can help with a detailed analysis of the genome and bacterial physiology, which can reveal

what causes antibiotic resistance and how we can prevent this. [6] In this study, we analyzed 10 samples of closely related *K. pneumoniae* strains. Although these strains share a similar genotype, they exhibit phenotypic differences. Therefore, we focused on methylation profiling, as it may help explain these differences.

Here, we presented a novel computational pipeline for methylation identification, which utilized data from nanopore sequencing and two approaches for their localization in bacteria genomes. The first method involves aligning reads with detected methylation sites to a single universal reference, which is common for all strains and makes it possible to study how methylation affects specific positions in the genome. The second approach aligns the reads of each sample to its own specific reference, enabling a detailed analysis of individual genomes, where we can identify unique genes in the individual samples that have some significant detected methylations.

2. Materials and Methods

The proposed pipeline for methylation analysis involves several steps. Firstly, the obtained ONT sequencing data are basecalled with methylation site detection. Secondly, the basecalled reads are mapped to reference genomes to identify where the methylations occur. Here, we proposed two possibilities - map all sequenced genomes to one reference, i.e. universal, or map each sequenced genome to its *de novo* assembly sequence.

2.1. Data, DNA extraction and sequencing process

Overall, 10 samples of *K. pneumoniae* were processed. These samples were obtained from the University Hospital Brno in the Czech Republic. First, DNA had to be extracted from the samples. The sequencing kit SQK-RBK114-24 (ONT, UK) was used to create a library from extracted DNA. This process involves adding barcodes to the samples, connecting adaptors and purification. The prepared library was then loaded to the flowcell FLO-MIN114 (ONT, UK), and after that, the sequencing was initiated on the ONT platform MinION Mk1C (ONT, UK). The resulting raw data was stored in POD5 files, where, besides signals, the information about possible methylated signals was saved.

The result summary table for all 10 samples can be seen in Table 1.

Table 1: The summary table for 10 processed samples of *K. pneumonia*

Name of sample	Number of reads [-]	Average read length [bp]
KP825	154 544	2670
KP1248	156 482	2543
KP1267	223 532	3145
KP1344	114 993	2942
KP1622	127 118	4038
KP1651	93 425	4705
KP1658	137 357	4035
KP1666	196 428	2778
KP1785	129 344	1912
KP1814	198 502	2896

2.2. Generation of reference genome for methylation mapping

Two mapping strategies were used for methylation pattern analysis. The first was to use a single universal reference for all genomes for read mapping, and the second one was to use the individual *de novo* assembled sequences of each sequenced genome.

In the case of universal reference, only one reference downloaded from the GenBank database on NCBI (National Center for Biotechnology Information) was used. It was the genome sequence of *K. pnemoniae* (NCBI Reference Sequence: NC_012731.1, [12]).

For the approach involving specific references (*de novo* assemblies), the workflow was as follows. The sequencing data were basecalled and demultiplexed using Dorado (v0.7.3, [7]) with *dna_r10.4.1_e8.2_400bps_sup@v5.0.0* model. The super accuracy model was chosen to secure accurate methylation detection of three basic types of methylation – 4mC, 5mA and 6mC. During basecalling, the parameter *min_qscore* was set to 10; thus, only high-quality reads would be further processed. After basecalling and demultiplexing, possible contaminations were filtered. Then, the FASTQ files were aligned to the reference sequence of the human genome (GCF_000001405.40) using Minimap2 (v2.28, [8]). Reads that do not map to the human genome were extracted using Samtools (v1.10, [9]) and used in the next step. The third step, *de novo* assembly, was run with Flye (v2.9.5, [10]). From Flye, assembled sequences in FASTA format were obtained and used for reads mapping. The last step was to annotate the data using DFAST (v1.3.1, [11]), where GenBank files were received, always with one chromosome sequence and one or two plasmid sequences.

2.3. Identification of methylation sites

The reference sequences generated or downloaded in the previous step were used for methylation site identification. Firstly, the sequencing data, which were basecalled with methylation detection in the previous step, were mapped to the reference sequence using Dorado. The second step was to create BedMethyl tables using Modkit (v0.2.6, [13]), which contain the positions of detected methylation sites in the sample. The final data were filtered and processed for further analysis. The detected methylations were selected only when their methylation probability was over 90%.

This process was run for each sample two times according to the mapping strategy – firstly with universal reference and secondly with sample-specific reference. In the end, two BedMethyl tables were obtained for each sample.

To enable proper interpretation of the data, the detected methylated positions (both from the universal reference and from the specific references) needed to be appropriately annotated in order to determine the genomic regions in which the methylation occurred. The resulting GenBank files for the specific references were therefore used for annotation, as well as the GenBank file for the universal reference. Thanks to this annotation, it was possible to observe which genes were affected by methylation, and these results allowed for a meaningful data interpretation.

3. Results and Discussion

3.1 Methylation sites identification with universal reference

Using a universal reference for aligning the samples enabled appropriate interpretation of the data based on genomic positions, as the detected methylation sites did not differ across samples. Using the annotated reference genome, all the genes that were methylated were identified. In this case, the number of methylations was still very high for each sample, so the threshold for the methylation probability parameter was increased from 90% to 95%.

A summary binary table was created from all samples, where it can be seen if the gene was methylated (marked as 1) or not (marked as 0). This matrix was used to calculate how the hierarchical clustering of input samples should look like. The distances between each sample were calculated as Jaccard distances to construct this type of phylogenetic tree. The resulting dendrogram is shown in Figure 1.

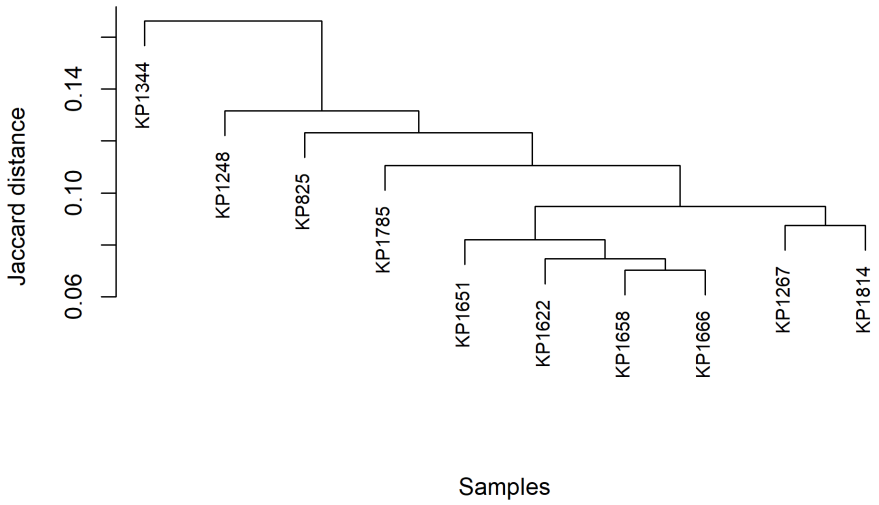

Figure 1: Dendrogram of samples aligned with universal reference

This dendrogram shows that the more methylated genes two samples share, the closer they are to each other (they have a smaller Jaccard distance), and the more similar they are. The most similar samples are KP1658 and KP1666, as well as the pair KP1267 and KP1814. In contrast, sample KP1344 is the most distant from the others, indicating it shares fewer methylated genes with the remaining samples.

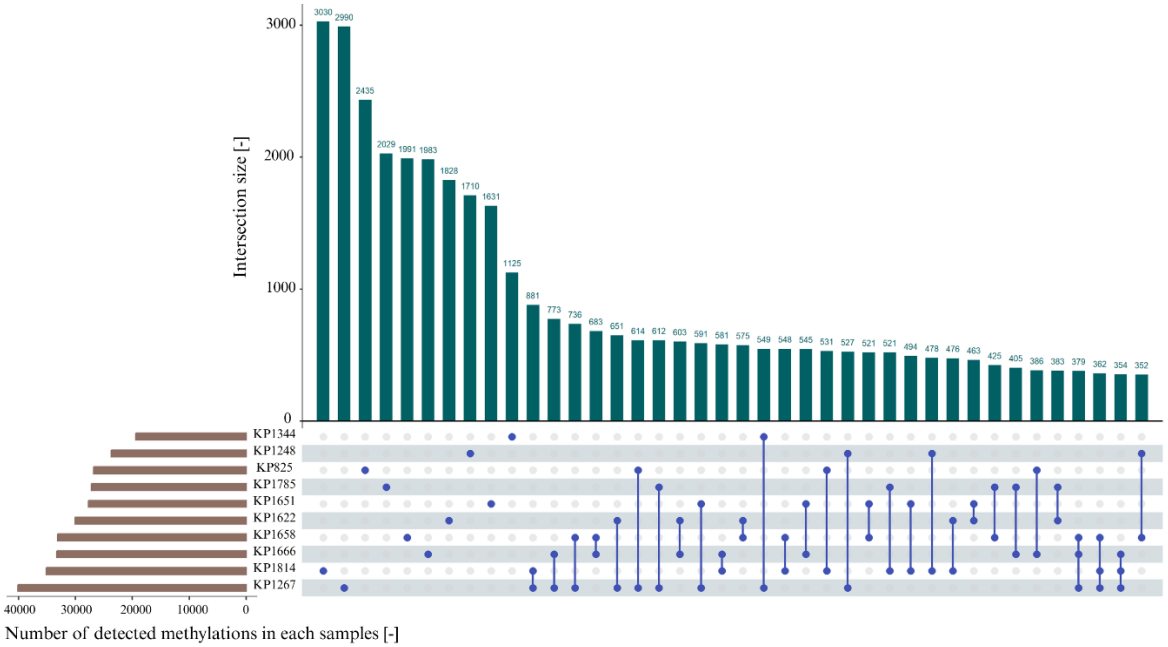

Figure 2: Upset plot of samples aligned with universal reference. On the left side of the plot is the total number of detected methylation sites, the bars on the top of the plot show the intersection size and the dots below every graph show which samples are involved in this specific intersection.

These similarities between samples can also be shown in a special type of Venn diagram used to visualize results from a more complex perspective called an "Upset plot", see Figure 2. The first ten bars represent the number of unique methylation sites for each sample. For example, the first bar for sample KP1814 has 3030 unique methylation positions that other samples do not have. Connecting two or more dots indicates a unique intersection of shared methylated positions for these samples. The eleventh bar has an intersection size of 881 for samples KP1814 and KP1267. That means these two samples are very similar, corresponding to the dendrogram in Figure 1.

The final visualization tool used for samples aligned to the universal reference was DNAPlotter (v18.2.0, [14]). This tool enables the visualization of detected positions across samples. Using DNAPlotter, the positions of methylated genes in each sample were visualized in a single graph (see Figure 3).

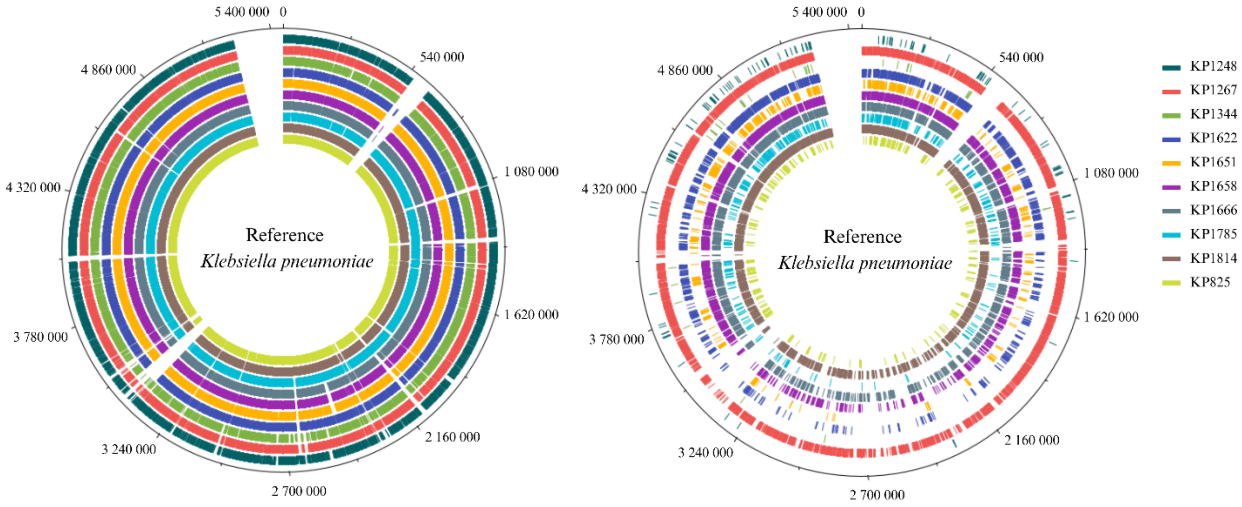

Figure 3: DNAPlots for– on the left for methylation with methylation probability > 95%, on the right with methylation probability > 98% Each colourful circle represents one sample. The black outer circle represents the reference genome sequence. The graph on the left shows the representation of methylated genes, where at least one methylation had methylation probability higher than 95%. On the right, the methylation probability threshold is higher (98%).

All the samples show a similar distribution of methylated genes with methylation probability higher than 95%. However, when the methylation probability threshold was increased to 98%, differences between the samples became more noticeable. The sample KP1267 has a higher number of methylated genes than the others. This may indicate that KP1267 exhibits higher epigenetic activity compared to the other samples.

3.2 Methylation sites identification with specific reference

As mentioned earlier, aligning samples to specific references can give us more detailed observations about our them. The main difference between using a universal and a specific reference is the presence of additional DNA sequences beyond just the chromosome. A detailed analysis of each sample enables us to examine not only chromosome sequences but also plasmid sequences.

While exploring genomic methylated positions in universal alignment results, the type of strand (+/-) was not considered. In contrast, methylated positions must now be analyzed separately for each strand and both chromosomal and plasmid DNA. This approach provides more information about our data.

In Figure 4, the unique methylated genes for each sample can be seen.

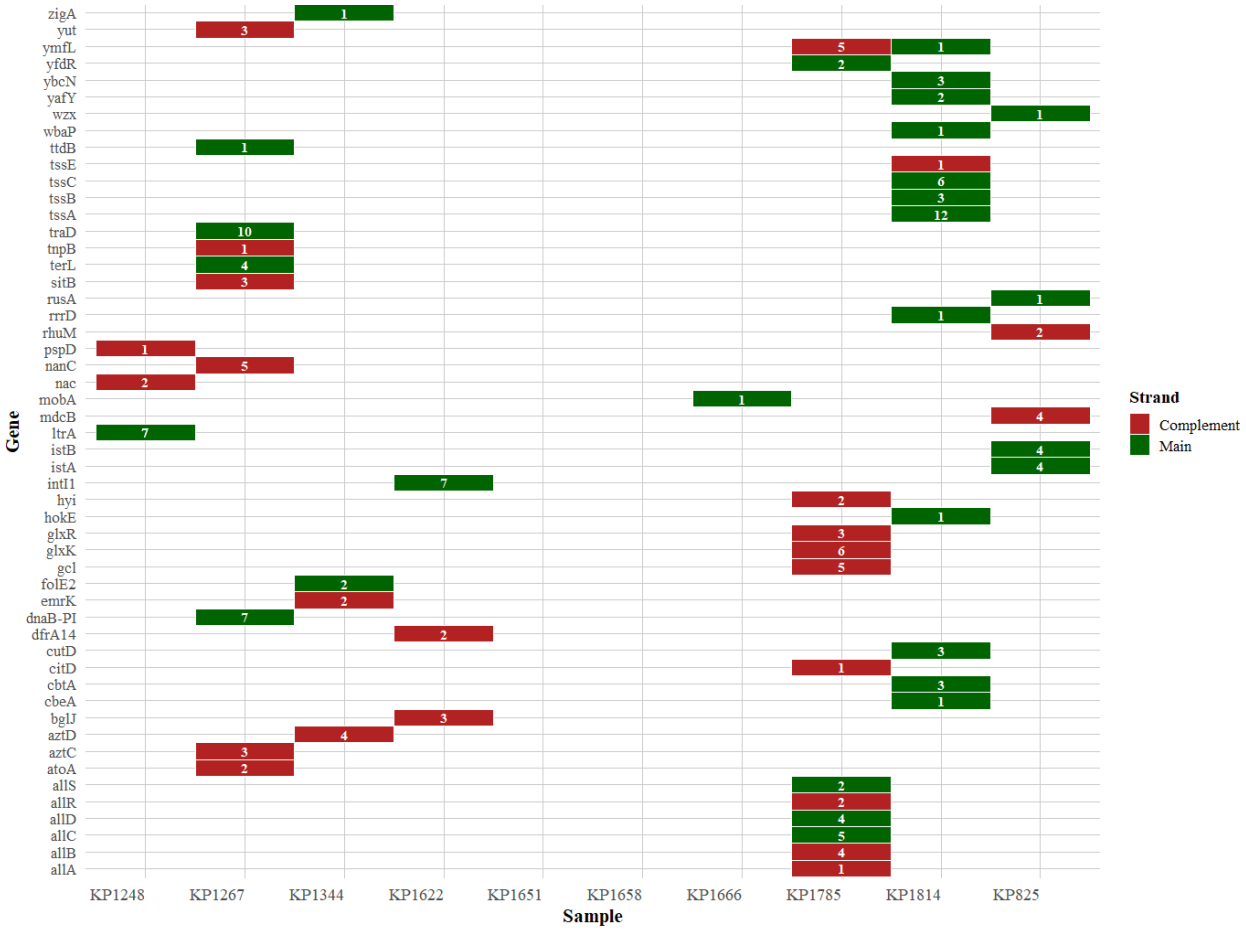

Figure 4: Unique genes for all the samples - green boxes are for genes on main (+) strand, red boxes are for genes on complement (-) strand. The number inside each box corresponds to the number of methylated positions that were detected in that particular gene.

Unique genes were detected in all samples. Eight of ten samples had at least one unique gene, two samples KP1651 and KP1658 do not have any unique genes. Sample KP1785 has the highest number of unique genes, and additionally, many of them have a high number of detected methylation sites. Genes *traD* and *tssA* have the highest number of detected methylated positions and may serve as potential markers with relevant biological significance.

4. Conclusion

This study compares two approaches for nanopore reads with detected methylation sites alignment to detect DNA methylation in *K. pneumoniae*: one using a universal reference genome and another using a specific reference genome for each sample. The main advantage of using universal reference is that all samples are aligned to the same reference, and the positions of methylation sites are consistent across all samples. In contrast, using a specific reference (each sample is assembled with its own specific reference) leads to a more detailed analysis of each sample – for example, the specific genes in samples can be easily lost during alignment to universal reference, with alignment to a specific reference all unique genes in the sample will be preserved.

The results demonstrate that the choice of reference significantly affects methylation analysis and should be chosen based on the research goals. Key differences were revealed in gene-level methylation, with sample KP1785 showing the most unique methylated genes. Genes like *traD* and *tssA* may represent potential epigenetic markers. The analysis showed how

important the appropriate choice of alignment type is. We focused on automating the analysis, which in the future can be used to create open-source tools that can easily process and analyze the input samples according to presets and targeted outputs.

In conclusion, methylation detection provides a valuable extension to genomic analysis, as it not only complements genotypic studies but also aids in predicting phenotypic traits. This additional layer of epigenetic information can greatly enhance our understanding of bacterial function and behaviour under varying environmental conditions, ultimately offering deeper insight into microbial adaptation and diversity. As research in this area progresses, integrating methylation analysis into routine genomic workflows may become essential for comprehensive bacterial profiling in both clinical and environmental contexts.

Acknowledgements

This work was supported by a grant project from the Czech Science Foundation [GA23-05845S].

References

[1] T. Phillip, "The role of methylation in gene expression" in Nature Education, 1(1):116, 2008.

[2] W. Hu, L. Guan, M. Li, and P. Fariselli, "Prediction of DNA Methylation based on Multi-dimensional feature encoding and double convolutional fully connected convolutional neural network", PLOS Computational Biology, vol. 19, no. 8, Aug. 2023, doi: 10.1371/journal.pcbi.1011370.

[3] B. Panning, "X-chromosome inactivation: the molecular basis of silencing", Journal of Biology, vol. 7, no. 8, 2008, doi: 10.1186/jbiol95.

[4] S. Li and T. O. Tollefsbol, "DNA methylation methods: Global DNA methylation and methylomic analyses", Methods, vol. 187, 2021, doi: 10.1016/j.ymeth.2020.10.002.

[5] S. Tamang, "Oxford Nanopore Sequencing: Principle, Protocol, Uses, Diagram", Microbe Notes. Accessed: Apr. 09, 2025. [Online]. Available: https://microbenotes.com/oxford-nanopore-sequencing/

[6] V. Ballén, Y. Gabasa, C. Ratia, R. Ortega, M. Tejero, and S. Soto, "Antibiotic Resistance and Virulence Profiles of Klebsiella pneumoniae Strains Isolated From Different Clinical Sources: a flexible prokaryotic genome annotation pipeline for faster genome publication", Frontiers in Cellular and Infection Microbiology, vol. 11, no. 6, Sep. 2021, doi: 10.3389/fcimb.2021.738223.

[7] Oxford Nanopore Technologies, "Dorado." [Online]. Available: https://dorado-docs.readthedocs.io/en/latest/basecaller/basecall_overview/.

[8] H. Li and I. Birol, "Minimap2: pairwise alignment for nucleotide sequences", Bioinformatics, vol. 34, no. 18, Sep. 2018, doi: 10.1093/bioinformatics/bty191.

[9] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, "Twelve years of SAMtools and BCFtools: pairwise alignment for nucleotide sequences", GigaScience, vol. 10, no. 2, Jan. 2021, doi: 10.1093/gigascience/giab008.

[10] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs: pairwise alignment for nucleotide sequences", Nature Biotechnology, vol. 37, no. 5, 2019, doi: 10.1038/s41587-019-0072-8.

[11] Y. Tanizawa, T. Fujisawa, Y. Nakamura, and J. Hancock, "DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication", *Bioinformatics*, vol. 34, no. 6, Mar. 2018, doi: 10.1093/bioinformatics/btx713.

[12] K.-M. Wu, L.-H. Li, J.-J. Yan, N. Tsao, T.-L. Liao, H.-C. Tsai, C.-P. Fung, H.-J. Chen, Y.-M. Liu, J.-T. Wang, C.-T. Fang, S.-C. Chang, H.-Y. Shu, T.-T. Liu, Y.-T. Chen, Y.-R. Shiau, T.-L. Lauderdale, I.-J. Su, R. Kirby, S.-F. Tsai, "Genome Sequencing and Comparative Analysis of Klebsiella pneumoniae NTUH-K2044, a Strain Causing Liver Abscess and Meningitis: pairwise alignment for nucleotide sequences", *Journal of Bacteriology*, vol. 191, no. 14, Jul. 2009, doi: 10.1128/JB.00315-09.

[13] A. Rand, C. Wright, and M. Stobier, "Modkit." [Online]. Available: https: //github.com/nanoporetech/modkit.

[14] T. Carver, N. Thomson, A. Bleasby, M. Berriman, and J. Parkhill, "DNAPlotter: circular and linear interactive genome visualization", *Bioinformatics*, vol. 25, no. 1, Jan. 2009, doi: 10.1093/bioinformatics/btn578.