Genome Analysis of *Pseudomonas alloputida*

Sukhman Grewal*, Surya Appana*, Jeffrey Gao*, Juhi Jadav*
Department of Bioengineering, University of California, Berkeley
BIOE 131 – Fall 2025

I. Introduction

Pseudomonas alloputida is a Gram-negative bacterium within the *P. Putida* group. This bacterium is commonly found in soil and wastewater systems and is defined as a metabolically flexible chemoorganotroph, alongside others in its group. Known for their nutritional versatility, central metabolism, and capacity to degrade aromatic and xenobiotic compounds, *Pseudomonas alloputida* and related strains are recognized as species essential to carbon cycling, bioremediation, and the degradation of industrial pollutants.[2], [3]

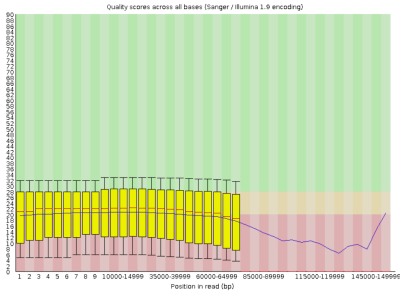
The genomes of the *P. Putida* group contain various collections of catabolic enzymes, stress-response systems, and transporters that support growth in chemically variable and contaminated environments. Genome-based analysis often allows for insight into the organization of metabolic versatility and ecological specialization. This process often reveals clusters or modules of genes that match different ecological specializations. Genome annotation clarifies the interactions of central carbon and nitrogen with biosynthetic systems, aromatic degradation pathways, etc. Phylogenetic comparisons show further emphasis on how metabolic traits evolved across varying, yet closely related lineages, while highlighting the contributions of horizontal gene transfer and the expansion of particular gene families.

II. Methods

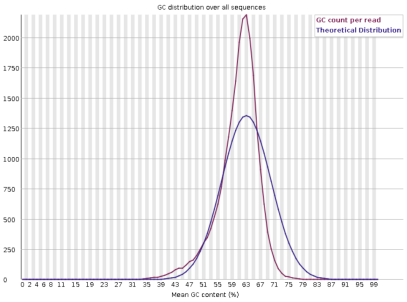
A. Filtering Sequencing Data and Quality Control

Our sequencing data includes single-end long reads (sequenced with Nanopore) and paired-ends short reads (sequenced with Illumina). FASTQC[4] was used for quality control of raw and processed reads.

Several filtration steps were executed on long and paired-ends short reads. To remove short and low-quality reads from the long reads data, Filtlong,[5] a tool for filting long reads by quality, was applied with a minimum read length of 1000 (reads shorter than 1000 base pairs were dropped), keeping the top 90% of reads by quality score. For the paired-ends reads, Cutadapt[6] was executed to remove adapters and Trimmomatic[7] was executed with sliding window sizes of 4, 8, and 12 along with a sliding window minimum average quality of 15.



(a) Per-base sequence quality.



(b) Per-sequence GC content

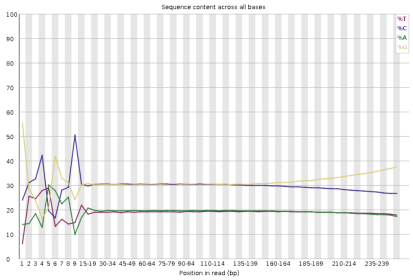
Fig. 1: Per-base sequence quality and per-sequence GC content is poor for raw long reads data.

B. Genome Assembly and Quality Assessment

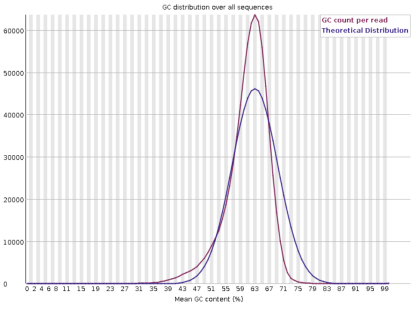
Flye,[10] a de novo assembler for Nanopore long reads, was used to assemble the long reads data. Finally, CheckM[8] was applied to check for genome completeness and possible biological contamination, and QUAST[9] was applied to determine structural assembly quality. Assembly was not executed on the paired-ends short reads data as a result sufficient performance of long reads assembly.

C. Genome Annotation

Genome annotation of the *Pseudomonas alloputida* assembly was performed using the RASTtk (Rapid Annotation using Subsystems Technology toolkit) pipeline within KBase. RASTtk is a workflow that integrates gene prediction, functional assignment, and identification of structural RNAs (rRNAs, tRNAs) and mobile genetic elements. Because RASTtk incorporates both broad protein family databases and highly targeted feature callers for rRNAs, tRNAs, transposases, and repeat-associated elements, it is well suited for



(a) Per-base sequence content.



(b) Per-sequence GC content

Fig. 2: Per-base sequence quality and per-sequence GC content is poor for raw paired-ends short reads data.

generating a comprehensive overview of bacterial genome content.

The genome assembly used as input, consisting of a single contig with an overall length of 6,031,726 bp and a GC content of 61.8%. Following annotation, RASTtk produced an object (RAST_annotation_v2) that includes predicted protein-coding genes (CDS), RNA genes, repetitive elements, and mobile genetic factors. These features are meant to support the subsequent analysis for our team's pathways and evolution sections, where these gene models can be used to infer metabolic capacity and phylogenetic relationships.

D. Metabolic Model Reconstruction and Pathway View

The draft genome-scale metabolic model, generated from the annotated assembly, was examined using the ModelSEED FBAModel viewer in kBase. Pathway content was inspected through the "Pathways" tab, which reports KEGG map IDs, reaction counts, and compound counts for each module in the reconstruction. These data were used to identify major functional categories, including central carbon metabolism, nitrogen and amino-acid utilization, lipid and envelope biosynthesis, cofactor pathways, and xenobiotic degradation systems. No flux balance analysis or gapfilling simulations were performed; all interpretations were based solely on pathway presence within the annotated metabolic model.

E. Comparative Genomics and Phylogeny

To determine the evolutionary placement of our *Pseudomonas alloputida*, we performed phylogenetic analyses at both the species level and individual gene level. First,

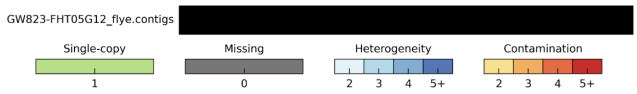


Fig. 3: CheckM indicates 100% completeness and no contamination of Flye long-reads assembly.

a species tree was created using the Insert Genome into SpeciesTree App with 10 neighbors. This was used as the baseline to compare taxonomic placement according to genes.

We analyzed two classes of genes. The 16s and rpoB genes were selected as housekeeping genes for evaluating replication machinery placement: they are highly conserved among bacterial strains due to their essential role in the key processes of transcription and translation. The argC gene (which is involved in arginine biosynthesis) is a housekeeping gene due to the essential role of arginine in *P. alloputida*'s growth. benA (benzoate dioxygenase) and catA (catechol 1,2-dioxygenase) were also chosen for analysis since they are aromatic-degradation genes that are important in relevant pathways. For each gene, homologs were obtained using Find Homologs and were then aligned using MUSCLE for DNA. ML phylogenetic trees were subsequently generated using the Build Phylogenetic Tree app. To evaluate vertical versus horizontal evolutionary relationships, we compared these gene trees to the species tree.

III. Results

A. Read Quality and Assembly Statistics

Quality Control assessment with FASTQC demonstrated poor per-base sequence quality scores and per-sequence GC content for raw long reads (see Figure 1). For raw paired-ends short reads, FASTQC shows poor per-base sequence content (particularly for the first 10-15 positions) and poor per-sequence GC content (see Figure 2). We also observed the presence of adapters in the paired-ends reads data. Filtration of long and paired-ends short reads data did not significantly improve quality scores, per-base sequence content, or GC content.

Running Flye for long reads assembly and QUAST yielded one contig of length 6.03 Mb, no ambiguous bases (gapless assembly), and 5792 genes, which is typical for bacterial genomes of this size. CheckM analysis indicated that the assembly was of very high quality, with ~100% completeness and no detectable contamination (Fig. 3). All expected single-copy marker genes were present exactly once, with no missing or duplicated markers, confirming that the genome represents a complete and uncontaminated single-strain assembly.

Given these high-quality results, we proceeded with using the Flye assembly for annotation and later analyses and did not perform assembly of paired-ends short reads.

B. Annotation Summary and Feature Content

Our annotation identifies 6,304 coding sequences (CDS), as well as groups of RNA including 22 rRNA and 70 tRNA genes. Specifically, we found 7 16S, 7 23S, and 8 5S rRNA genes. These rRNA genes make up the 3 core components of the bacterial ribosomal RNAs.[1] 16S rRNA genes encode the small subunit of the ribosome, meant for aligning mRNAs and tRNAs in translation. 23S rRNA genes encode structural and catalytic components of the large ribosomal subunit and are important for peptide bond formation and overall ribosomal stability. 5S rRNA genes, although much shorter, play more of an architectural role in stabilizing interactions between the 23S rRNA and ribosomal proteins.

In addition to core ribosomal features, the genome displays multiple classes of mobile genetic elements. We found that it was important to identify contents of transposases, recombinases, and mobile element proteins.

Transposases are enzymes that catalyze the movement of insertion sequences (IS), and thus promote genomic rearrangements, horizontal gene transfer, and structural variation. Sixteen transposase genes were identified, with 7 of them belonging to the IS30 family. The IS30 family is associated with replicative transposition mechanisms and is well known to contribute to genome plasticity in environmental bacteria, such as *Pseudomonas,* which we are studying. One gene corresponding to the IS150 insertion sequence family (transposase InsK) was also detected. Even inactive transposases serve as signatures of past mobility events, and help provide insights into the evolutionary dynamic of bacteria.

The annotation also contained 8 recombinase genes and 22 mobile element proteins. Recombinases moderate site-specific DNA rearrangements and play roles in important DNA processes such as repair, integration, and excision. Mobile element proteins, which include factors encoded by phages or transposons, contribute to the overall genome diversification and adaptability within the environment. For a genome with substantial evolutionary fluidity, there must be a combined presence of transposases, recombinases, and mobile element proteins.

A draft metabolic model was generated using the ModelSEED pipeline with OMEGGA to evaluate the functional potential of the genome. The resulting model contained 1,438 enzymatic and transport reactions linked to 2,837 genes, representing roughly 45% of the coding sequences in the genome. These results aim to provide a systems-level view of the genome's functional capacity, and it shows that our species has relatively high metabolic versatility.

C. Pathway Reconstruction and Metabolic Capabilities

Using the ModelSEED reconstruction, the annotated genome was examined with the purpose of documenting the presence of major metabolic modules. Inspection of the KEGG-linked pathways table revealed complete central carbon metabolism, a wide range of nitrogen and amino acid utilization, Gram-negative lipid and envelope biosynthesis, and a large capacity for xenobiotic degradation.

Name	Map ID	Rxn Count	Cpd Count	Source
Glycolysis / Gluconeogenesis	map00010	57	31	KEGG
Citrate cycle (TCA cycle)	map00020	28	23	KEGG
Pentose phosphate pathway	map00030	55	38	KEGG
Pentose and glucuronate interconversions	map00040	72	55	KEGG
Fructose and mannose metabolism	map00051	79	55	KEGG
Galactose metabolism	map00052	60	47	KEGG
Ascorbate and aldarate metabolism	map00053	57	47	KEGG
Fatty acid biosynthesis	map00061	183	50	KEGG
Fatty acid elongation	map00062	53	41	KEGG
Fatty acid degradation	map00071	85	57	KEGG

Showing 1 to 10 of 163 entries

Fig. 4: Subset of the metabolic pathways reconstructed from the RASTtk-derived metabolic model for *Pseudomonas alloputida*. The full ModelSEED Pathways table includes 163 KEGG pathways.

Central carbon and energy metabolism was represented, alongside Glycolysis/Gluconeogenesis (map00010), the TCA cycle (map00020), the Pentose Phosphate Pathway (map00030), and pyruvate metabolism (map00520). These modules, when combined, contain upwards of 200 reactions. They support the organism's chemoorganotrophic nature alongside its ability to generate biomass and reducing power. Both glycolysis and Entner-Doudoroff-PP routes are shown as feeding into the TCA cycle, thus indicating the metabolic versatility expected of *Pseudomonas* strains.

The pathways that were involved in nitrogen and amino-acid metabolism were also well populated. The model includes alanine, aspartate, and glutamate metabolism (53 reactions), valine leucine, and isoleucine biosynthesis and degradation, and cysteine and methionine metabolism. Arignine and proline metabolism (125 reactions, 93 compounds) and nitrogen metabolism (56 reactions, 42 compounds). A nitrogen metabolism pathway (map 000910) further supports the utilization of different nitrogen sources, proving to be consistent with ecological adaptation to various soil and wastewater environments.

Lipid, cell-envelope, and co-factor biosynthesis pathways also matched the expectations of a Gram-negative bacterium. Fatty acid biosynthesis (map00061) and fatty acid degradation (map00071) were strongly represented, alongside glycerophospholipid metabolism (map00564), lipopolysaccharide biosynthesis, and peptidoglycan assembly. Cofactor pathways like riboflavin (map00740), vitamin B6 (map00750), thiamine (map00730), nicotinate and nicotinamide (map00760), and pantothenate/CoA biosynthesis (map00770) indicate that the genome encoded the complete de novo routes for key enzymatic cofactors.

A notable feature of the reconstruction was the range of xenobiotic and aromatic compound degradation pathways. These included benzoate degradation (map00362), aminobenzoate degradation (map00627), xylene (map00622),

toluene (map00623), polycyclic aromatic hydrocarbons (map00624), and pathways for chlorinated compounds like chloroalkane and chloroalkene degradation (map00625), alongside chlorocyclohexane/chlorobenzene degradation (map00361). The appearance of these modules in the reconstruction also matches the scientifically acknowledged ecological role of *P. alloputida* as a adaptable, degrading bacterium with the ability to use aromatic and pollutant-based carbon sources.

Name	Map ID	Rxn Count	Cpd Count	Source
Glycosphingolipid biosynthesis - ganglio series	map00604	38	31	KEGG
Pyruvate metabolism	map00620	71	31	KEGG
Dioxin degradation	map00621	22	58	KEGG
Xylene degradation	map00622	40	38	KEGG
Toluene degradation	map00623	47	43	KEGG
Polycyclic aromatic hydrocarbon degradation	map00624	82	105	KEGG
Chloroalkane and chloroalkene degradation	map00625	41	40	KEGG
Naphthalene degradation	map00626	47	60	KEGG
Aminobenzoate degradation	map00627	93	85	KEGG
Fluorene degradation	map00628	19	36	KEGG

Fig. 5: Subset of xenobiotic and aromatic compound degradation pathways identified in the reconstructed metabolic model of *Pseudomonas alloputida*. KEGG maps (e.g., dioxin, xylene, toluene, polycyclic aromatic hydrocarbon, and aminobenzoate degradation) emphasize the strain's potential for bioremediation of industrial and environmental pollutants.

Combined, these pathway assignments are an indication that the analyzed genome encodes complete central metabolism, fully structured amino acid and nitrogen pathways, the Gram-negative envelope, and a broad range of aromatic and xenobiotic degradation modules, thus supporting *Pseudomonas alloputida's* niche in polluted soil and wastewater environments.

D. Phylogenetic Placement and Comparative Insights

Our species tree correctly placed our isolate within the *Pseudomonas putida* phylogenetic group of *Gammaproteobacteria*.

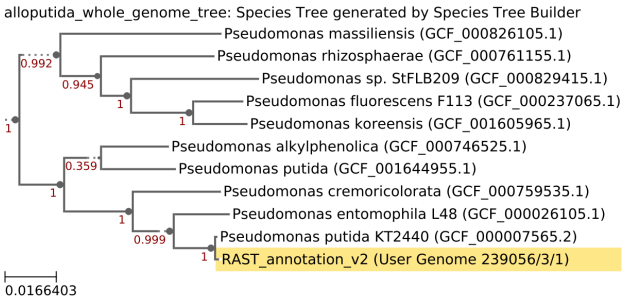


Fig. 6: Phylogenetic placement of the isolate based on the whole genome.

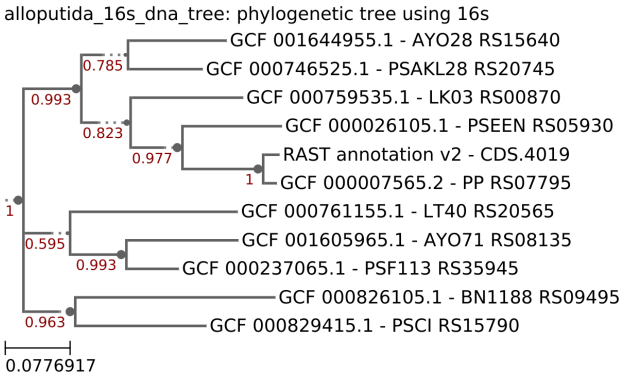


Fig. 7: Maximum-likelihood phylogeny of 16s rRNA sequences from *P. alloputida* and related *Pseudomonas* species.

Trees of our housekeeping genes (16s, rpoB, argC) followed this general species tree structure, with *P. putida* being the closest homolog for all three genes and *P. entomophila* being the second closest homolog for two of three (it was still fairly close for rpoB). The 16s phylogeny is shown in Figure 7). argC exhibited a consistent phylogeny to the expected species tree, indicating vertical transfer of genes - this is likely a gene that is conserved among related *Pseudomonas* species. Since argC is a gene that codes for arginine, an amino acid, and is thus part of a fundamental pathway, it is under strong purifying selection pressure so its sequence will be highly conserved.

On the other hand, phylogenetic analysis of the aromatic degradation genes benA and catA did not neatly follow the species tree. In both cases, multiple paralogs were revealed in the *P. alloputida* genome, with each paralog forming distinct clades. Occasionally, some paralogs showed clustering with homologs of more distantly related *Pseudomonas* species, such as *P. fluorescens* in the benA tree (Figure 8). These slight deviations from the expected phylogeny are likely indicative of horizontal gene transfer. In particular, since aromatic compound degradation can occur in a wide range of environments, these observations likely reflect gene duplication and acquisition events that have expanded the species' ability to metabolize in various conditions.

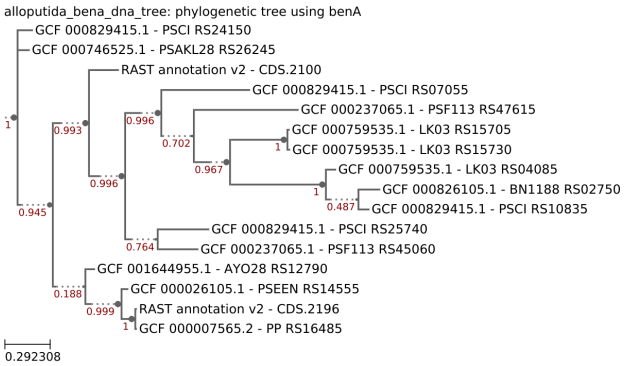


Fig. 8: Phylogenetic analysis of benA paralogs from the assembled *P. alloputida* genome and homologs from other *Pseudomonas* species.

In general, housekeeping genes exhibited congruent phylogenies, whereas adaptive genes had slight discrepancies compared to the species tree. This difference in phylogenies emphasize the different evolutionary pressures placed on essential, core genes versus environmentally-specific genes.

IV. Discussion

Throughout assembly, annotation, metabolic reconstruction, and phylogenetic analysis the *Pseudomonas alloputida* isolate is consistent with the biological expectations of a metabolically adaptable member of the *P. putida* group. The reconstructed genome shows a high level of completeness, a clear representation of integral genes, and annotation that supports interpretation of central metabolism and ecological functions. The annotated metabolic model highlights the range of carbon and energy pathways characteristics of both soil and wastewater-associated strains. The presence of complete modules for glycolysis, the Entner-Doudoroff pathway, the TCA cycle, and the pentose phosphate pathway indicate a large amount of versatility in carbon flux. Overall, as a whole, these pathways support a chemooganotrophic biology and allow the strain to metabolize sugars, organic acids, and aromatic intermediates that are oftentimes found in natural and anthropogenic environemnts. The representation of nitrogen and amino-acid metabolism, alongside arginine and nproline metabolism, and branched-chain amino acid pathways indicates the capacity of the strain to grow on various nitrogen sources, thus reinforcing the known abilities and roles of *P. alloputida* in chemically differing habitats.[2]

Beyond metabolic reconstruction, the assembly and annotation and assembly reinforce the interpretation of the genome. The long-read assembly produced a single contig genome with GC content and size that is consistent with other *P. putida* strains. The RASTtk annotation recovered complete rRNA and tRNA sets alongside multiple transposases and IS elements. These elements are common in environmental *Pseuodomonas* strains and allow a mechanism for the acquisition or diversification of the observed catabolic pathways.

Comparative and phylogenetic analyses also support this reasoning as while the whole genome species tree sorted the isolate alongside the other members of *P. putida*, several pathway associated genes like *benA* and *catA* revealed deviating topologies. This outcome could potentially be attributed to gene-level horizontal transfer as the patterns normally analog with adaption to chemically varying soil and wastewater environments.[3]

Via the integration of assembly quality, annotation content, pathway reconstruction, and gene-level phylogenies, the combined results present the isolate as one that encodes stable central metabolic functions along with adapting and evolving degradation pathways. Although limited by a reliance on draft assembly and predictions founded in annotation, the analyses ultimately support the ecological interpretation of this strain as one that can be defined as a metabolically versatile degrader in polluted environments.

V. Conclusion

Our analysis shows that the *Pseudomonas alloputida* Nanopore long-reads sequencing data we assembled is a metabolically flexible bacterium well suited for life in diverse and pollutant-rich environments. Its complete genome, broad metabolic capabilities, and combination of conserved and adaptive genes suggest an organism that is both ecologically resilient and capable of utilizing a wide range of chemical substrates. These findings reinforce the broader role of members of the *Pseudomonas* genus in environmental cycling and bioremediation. Future work may involve experimentally validating specific degradation pathways or drawing genomic and metabolic comparisons between different genus members.

VI. Contributions Statement

Surya Appana had the “assembly" role, conducting filtration and assembly of raw reads with appropriate tools along with quality control assessments and explained the relevant methods and results. Surya also set up the initial team meeting and methods of communication.

Jeffrey Gao conducted all baseline annotations and analysis of components for the Annotations role. This included running the RASTtk annotation and ModelSEED pipelines, and identifying key genes and elements within the annotation output. Jeffrey contributed all subsections regarding genome annotation within this report, including section C of Methods and section B of results.

Sukhman Grewal contributed all analyses associated with the Pathways role. This includes examining the RASTtk annotation and the ModelSEED genome-scale metabolic reconstruction, identifying major KEGG pathways, and interpreting modules related to the central carbon metabolism, nitrogen and amino-acid utilization, xenobiotic degradation, etc. Sukhman generated pathway fgiures (via screenshot) from the ModelSEED FBAModel viewer and wrote the full pathways subsection for the team report, including descriptions of metabolic capabilities and their ecological relevance.

In specific this refers to Part D of the Methods and Part C of the Results. Sukhman also wrote the group introduction section alongside the discussion section while also formatting and designing the overall Overleaf LaTeX layout in terms of section/subsection structure, references, citations, styling, etc.

Juhi Jadav conducted all analyses for the Evolution role. This includes species tree construction, determining and analyzing key housekeeping genes (16s, rpoB, argC), and investigating the role of adaptive xenobiotic degradation genes (benA, catA) that were uniquely chosen due to their role in interesting pathways identified by the Pathways role. This includes creating FeatureSets, finding homologs, performing Multiple Sequence Alignments using Muscle, and building ML phylogeny trees using FastTree2 in her individual KBase narrative (all final output trees displayed in the group narrative). Juhi wrote Section E of Methods and Section D of Results.

REFERENCES

[1] J. Chodvadiya, G. Shukla, and G. Sharma, "16S-23S-5S rRNA Database: a comprehensive integrated database of archaeal and bacterial rRNA sequences, alignments, intragenomic heterogeneity, and secondary structures," 2025. doi: 10.1101/2025.06.23.661116.

[2] V. de Lorenzo, D. Pérez-Pantoja, and P. I. Nikel, "Pseudomonas putida KT2440: the long journey of a soil-dweller to become a synthetic biology chassis," *Journal of Bacteriology*, vol. 206, no. e00136-24, 2024. doi: 10.1128/jb.00136-24.

[3] G. Purtschert-Montenegro, G. Cárcamo-Oyarce, M. Pinto-Carbó, *et al.*, "Pseudomonas putida mediates bacterial killing, biofilm invasion and biocontrol with a type IVB secretion system," *Nature Microbiology*, vol. 7, pp. 1547–1557, 2022. doi: 10.1038/s41564-022-01209-6.

[4] S. Andrews, "FastQC: A Quality Control Tool for High Throughput Sequence Data," 2010. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[5] R. Wick, "Filtlong: quality filtering tool for long reads," 2018. Available online at: https://github.com/rrwick/Filtlong.

[6] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, 2011. doi: 10.14806/ej.17.1.200.

[7] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014. doi: 10.1093/bioinformatics/btu170.

[8] 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. doi: 10.1101/gr.186072.114.

[9] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, 2013. doi: 10.1093/bioinformatics/btt086.

[10] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, "Assembly of long, error-prone reads using repeat graphs," *Nature Biotechnology*, vol. 37, pp. 540–546, 2019. doi: 10.1038/s41587-019-0072-8.