Integrative Bioscience Nexus

Biologica Aid

Article

BactoFlow: An Automated and Reproducible Pipeline for High-Resolution Bacterial Genomics

Md. Ataur Rahman Sikdar, Md. Abidur Rahman, Umma Habiba Hafsa, Hossain Md. Faruquee, Sudhangshu Kumar Biswas*

Bacteriophage Biology and Genomics Lab., Department of Biotechnology and Genetic Engineering, Faculty of Biological Sciences, Islamic University, Kushtia-7003, Bangladesh

* Correspondence: skbiswas@btge.iu.ac.bd

Abstract

Next-generation sequencing has transformed bacterial genomics, but fragmented bioinformatics tools restrict accessibility. BactoFlow is a modular, automated, end-to-end pipeline, integrating quality control, mapping, assembly, annotation, resistance and virulence screening, antiviral defense detection and publication-ready visualization. Conda-managed environments and checkpoint-based execution for reproducible, scalable and fault-tolerant genomic analyses. We used BactoFlow to characterize a clinical *Pseudomonas aeruginosa* isolate (SRR33893847) and assigned it by MLST to sequence type ST308, a globally disseminated, high-risk clone with a well-documented association with multidrug resistance. Quality trimming with Trimmomatic retained 92.4% of 1,661,330 raw reads. Subsequent alignment via BWA-MEM enabled FreeBayes to identify 56,587 high-confidence variants, reflecting substantial genomic divergence from the reference strain. De novo assembly with SPAdes produced a 6.92 Mb draft genome distributed across 73 contigs, with an N50 of 412,000 bp and a GC content of 66.11%. Prokka annotation identified 6,380 coding sequences along with 68 tRNA genes and 3 rRNA features. Multi-database functional screening filled out a detailed resistance and virulence profile: CARD detected 51 antimicrobial resistance genes, VFDB flagged 290 virulence-associated factors, and PADLOC identified 12 antiviral defense systems, among them *Gabija, RosmerTA,* and *zorya_type_I*, alongside 3 integrated prophage regions. Overall, these findings validate a genomically well-equipped and clinically significant isolate. BactoFlow offers a scalable, accessible solution for researchers seeking high-resolution bacterial genomics without the overhead of assembling and managing separate tools.

Keywords: Whole-genome sequencing, bacterial genomics, bioinformatic workflow, pipeline automation

Academic Editor: Dr. Hamed Hamuoda

Received: 2 May 2026
Revised: 25 June 2026
Accepted: 27 June 2026
Published: 05 July 2026

1. Introduction

Whole-genome sequencing (WGS) has transformed bacterial genomic studies by becoming the gold standard for epidemic investigation, antimicrobial resistance (AMR) surveillance, and pathogen characterization [1], [2]. The rapidly declining cost of sequencing has resulted in an exponential growth of genomic data. However, this progress has simultaneously exposed a critical analytical bottleneck. Turning raw sequencing output into interpretable biological data requires the coordinated application of several specialized bioinformatics tools, each with distinct dependencies, file formats, and parameters. Navigating this fragmented computational landscape demands considerable expertise and introduces substantial

reproducibility concerns, wherein inconsistent software configurations can yield divergent results from identical datasets [3], [4].

Existing workflow managers, including Nextflow and Snakemake, have improved standardization and scalability. However, they impose steep learning curves and commonly address only discrete segments of the analytical process, either upstream quality control and assembly, or downstream functional profiling, which require manual integration of outputs through error-prone custom scripting [5]. Consequently, no single accessible solution currently consolidates the complete analytical continuum, from raw data retrieval and quality control through *de novo* assembly, multilocus sequence typing (MLST), and simultaneous multi-database functional screening, into a unified and reproducible framework accessible to non-specialist users.

To address this gap, we present BactoFlow, an automated end-to-end bacterial genomics pipeline engineered on three core principles: automation, reproducibility, and biological interpretability. BactoFlow integrates the complete analytical workflow from raw data retrieval to publication-ready visualizations through a checkpoint-based architecture with Conda-managed environments, ensuring resilience and cross-system reproducibility. The pipeline incorporates resistome profiling via the Comprehensive Antibiotic Resistance Database (CARD), virulome characterization via the Virulence Factor Database (VFDB), antiviral defense system detection via PADLOC, and prophage region prediction via PhiSpy. A dedicated Species identification utility further ensures taxonomic metadata accuracy across heterogeneous sequencing inputs [6], [7], [8], [9].

We validate BactoFlow through comprehensive characterization of a clinical *Pseudomonas aeruginosa* sequence type ST308 isolate, a globally disseminated high-risk clone associated with multidrug resistance, nosocomial outbreaks, and enhanced virulence, demonstrating its capacity to generate high-resolution genomic profiles that simultaneously resolve 51 AMR genes, 290 virulence factors, 12 antiviral defense systems, and 56,587 genomic variants from raw sequencing reads [10]. BactoFlow thus provides a scalable, accessible, and reproducible solution for high-throughput bacterial genomics in the era of precision microbiology.

2. Materials and Methods

2.1. Data acquisition and quality control

The paired-end sequencing reads of *Pseudomonas aeruginosa* (SRR33893847) were downloaded from the NCBI Sequence Read Archive via the SRA Toolkit (v3.0.3) [11]. Both compressed (.fastq.gz) and uncompressed FASTQ formats were accepted as valid inputs. FastQC (v0.11.9) was used to analyse adaptor contamination, GC content, and per-base sequence quality; MultiQC (v1.14) was used to compile the results into a single summary report [12], [13], [14]. Raw reads were then subjected to adapter removal and quality trimming using Trimmomatic (v0.39), where a sliding window of 4:15 was applied alongside a minimum read-length threshold of 36 bp to exclude low-quality fragments from further processing [15].

2.2. Reference-based mapping and variant calling

The quality-filtered reads were mapped against the reference genome of *Pseudomonas aeruginosa* strain 2507 using BWA-MEM (v0.7.17). The resulting alignments were sorted and indexed using SAMtools (v1.17) to prepare them for variant analysis [16]. Variant calling was then performed using FreeBayes (v1.3.6), and a quality-score threshold was applied to the raw variant calls to retain only those with sufficient confidence for downstream interpretation [17].

2.3. De novo genome assembly

De novo assembly of the trimmed reads was performed using SPAdes (v3.15.5) under isolate mode, which results contigs and scaffolds by iterating over a range of k-mer lengths. The resulting assemblies were assessed for contiguity using QUAST (v5.2.0), with N50, L50, and total assembly length adopted as the principal evaluation criteria [18], [19].

2.4. Functional annotation and species identification

The assembled scaffolds were submitted to Prokka (v1.14.6) for structural and functional gene prediction, enabling the identification of coding sequences, rRNA, and tRNA features across the draft genome [20]. Multilocus sequence typing (MLST) was carried out using the MLST command-line tool (v2.23) which performs allele-based typing against curated scheme database (pubMLST) to assign a sequence type to the isolate.

2.5. Multi-database genomic screening

A comprehensive genomic screening was carried out to identify Antibiotic Resistance and Virulence genes within the assembled genome. ABRicate (v1.0.1) was used to query the assembled sequences against three established databases: the Comprehensive Antibiotic Resistance Database (CARD), ResFinder, and the Virulence Factor Database (VFDB) [21],

[22], [23]. Beyond resistance and virulence profiling, the genome was further interrogated for prokaryotic defence systems using PADLOC (v1.1.0), and putative prophage regions were delineated with PhiSpy (v5.0.10) [24], [25]. All results were visualized using Matplotlib and Seaborn to produce publication-grade figures.

2.6. Pipeline orchestration and reproducibility

The entire analytical workflow was implemented as a modular Bash-based pipeline capable of processing bacterial whole-genome sequencing data from raw reads through to final outputs in a single execution (**Figure 1**). To safeguard against interruptions, a checkpoint mechanism records each completed step to results/, checkpoints/, allowing the pipeline to resume from the last successful stage rather than restarting from scratch. Computational reproducibility was ensured through Conda environment management, with all tools installed at fixed, version-pinned configurations.

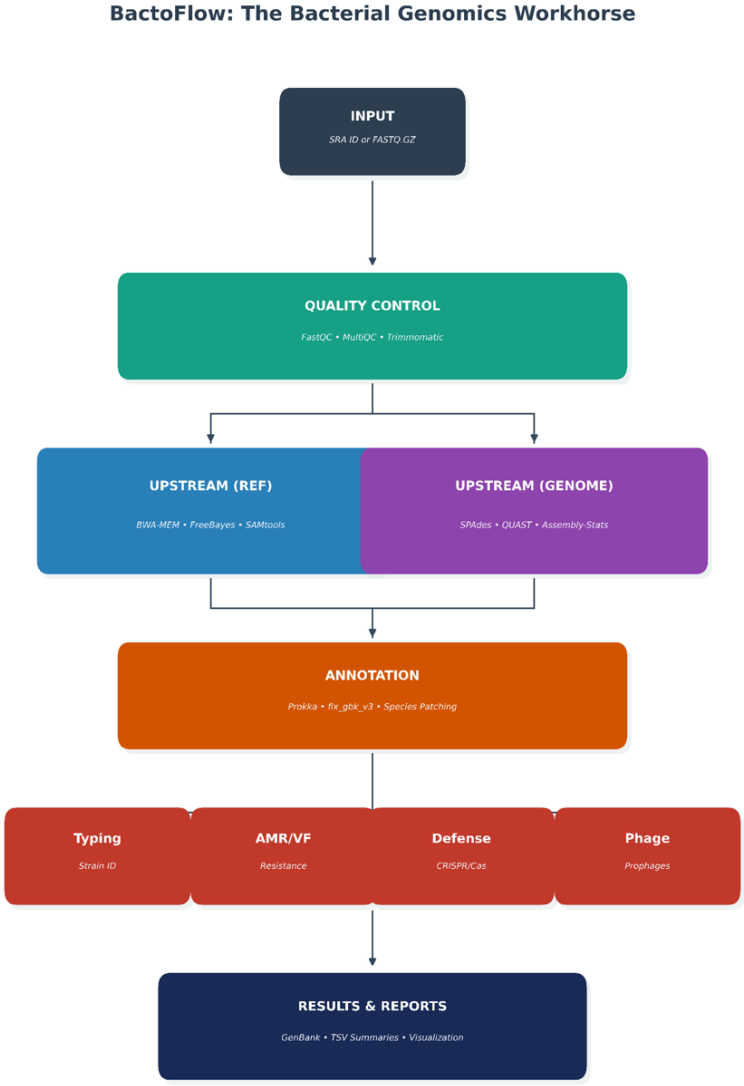

Figure 1. Schematic overview of the BactoFlow pipeline architecture, illustrating the end-to-end workflow from raw data acquisition to comprehensive genomic characterization and visualization.

3. Results

3.1. Case study: genomic characterization of Pseudomonas aeruginosa isolate using BactoFlow

The pipeline BactoFlow has been utilized to analyze a *Pseudomonas aeruginosa* isolate (SRR33893847) for transforming raw sequencing data into a comprehensive genomic profile through a series of automated modular steps (**Table 1**).

3.2. Read quality and pre-processing

Preliminary quality assessment of the raw sequencing data (1,661,330 sequences) indicated high-fidelity reads with average base quality scores above Q30. After adapter trimming and quality-based filtering, 92.4% of the reads were retained

for downstream analysis. The persistence of a significant proportion of high-quality reads ensured sufficient coverage for accurate variant calling and genome assembly.

3.3. Genomic divergence from the reference strain

Mapping the trimmed reads to the reference genome of *Pseudomonas aeruginosa strain 2507* using BWA-MEM resulted in excellent coverage. Subsequent variant calling using FreeBayes identified a total of 56,587 high-confidence variants. The draft genome comprised a significant density of single nucleotide polymorphisms (SNPs) and small insertions/deletions (Indels), highlighting the substantial genomic divergence against the reference strain (**Figure 2**).

Table 1. Summary of genomic characterization metrics for the *Pseudomonas aeruginosa* isolate SRR33893847 generated by the BactoFlow pipeline.

Category	Parameter	Value
Isolate	Species	*Pseudomonas aeruginosa*
	Accession	SRR33893847
	Sequence Type	ST308 (High-Risk Clone)
Quality Control	Raw Sequences	1,661,330 reads
	Reads Retained Post-Trimming	92.4%
	Mean Base Quality	> Q30
Assembly	Total Assembly Size	6.92 Mbp
	Number of Contigs	73
	N50	412,000 bp
	L50	7
	GC Content	66.11%
Annotation	Protein-Coding Sequences (CDS)	6,380
	tRNA Genes	68
	rRNA Features	3
Variants	High-Confidence Variants	56,587
Resistome	AMR Genes	51
Virulome	Virulence Factors	290
Defense Systems	Antiviral Defense Systems	12
Prophage	Integrated Prophage Regions	3

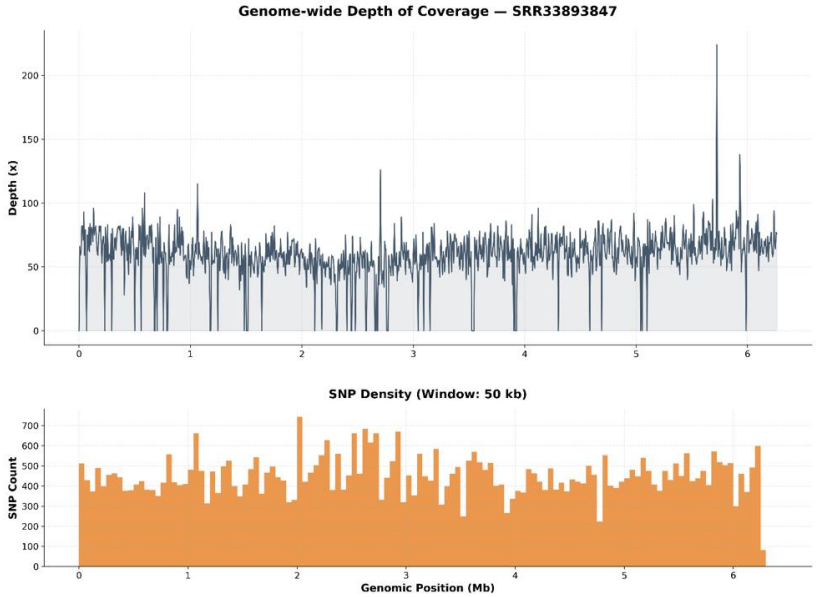

Figure 2. Genome-wide sequencing depth and variant distribution.

3.4. Assembly contiguity and genome characteristics

The genome was assembled *de novo* using the SPAdes assembler in isolate mode. The final assembly spanned 6.92 Mbp across 73 large contigs exhibiting high contiguity with an N50 value of 412,000 bp and an L50 of 7. The average GC

content was determined to be 66.11%. These metrics confirm a high-quality genomic reconstruction suitable for detailed functional annotation (**Table 1**).

3.5. Genomic feature inventory and sequence typing

Annotation of the assembly using Prokka identified 6,380 protein-coding sequences (CDS), 68 tRNA genes, and 3 rRNA features. Multi-locus sequence typing (MLST) definitively assigned the isolate to sequence type ST308, a globally recognized high-risk clone associated with multidrug resistance and enhanced virulence.

3.6. Downstream screening

Comprehensive screening of the functional genomic landscape revealed an extensively equipped genome. Screening against the Comprehensive Antibiotic Resistance Database (CARD) identified 51 antimicrobial resistance (AMR) genes (**Figure 3**); 290 virulence-associated factors were detected via the Virulence Factor Database (VFDB); the PADLOC framework identified 12 unique antiviral defense systems including *Gabija, RosmerTA,* and *zorya_type_I* suggesting a robust immune repertoire against mobile genetic elements (**Figure 4**); and PhiSpy analysis predicted 3 integrated prophage regions distributed across the assembly scaffolds, contributing to overall genomic plasticity (**Figure 5**).

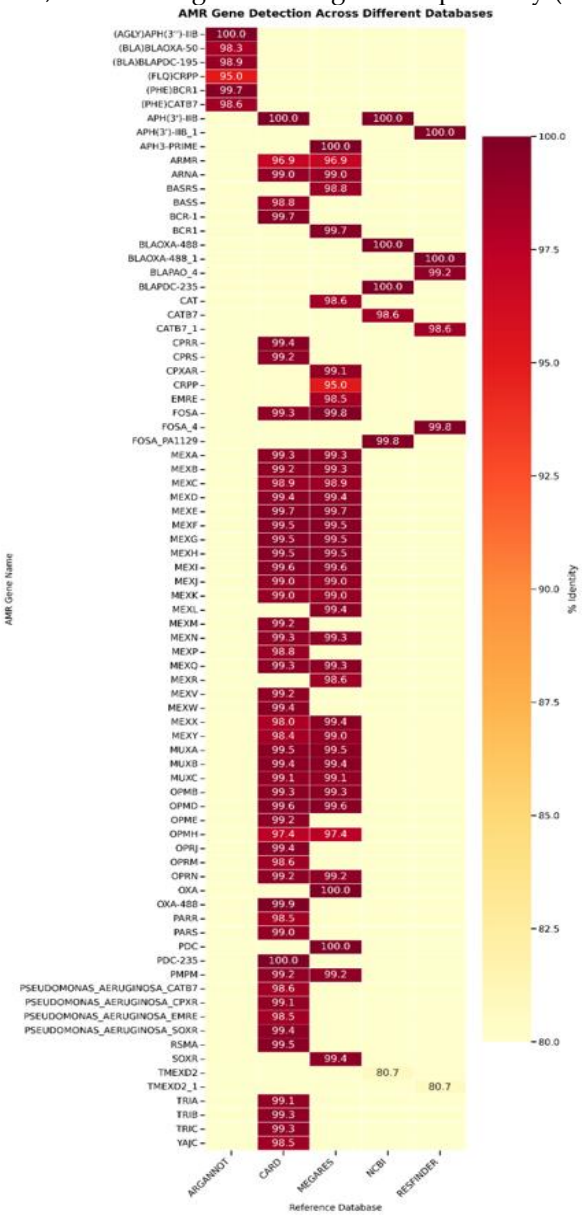

Figure 3. Heatmap displaying the percentage identity of antimicrobial resistance (AMR) genes identified through the integration of five curated databases: ARGANNOT, CARD, MEGARES, NCBI, and RESFINDER.

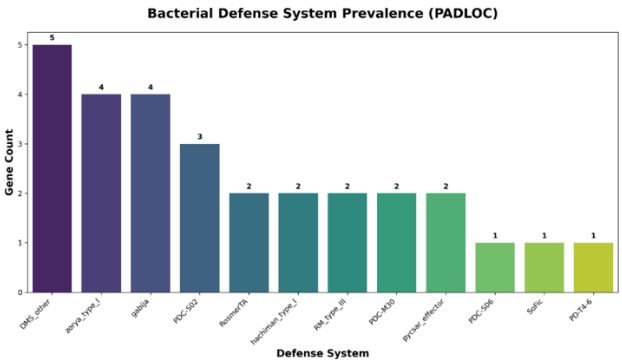

Figure 4. Bar chart represents the diversity and gene composition of 12 unique defense systems identified by the PADLOC framework.

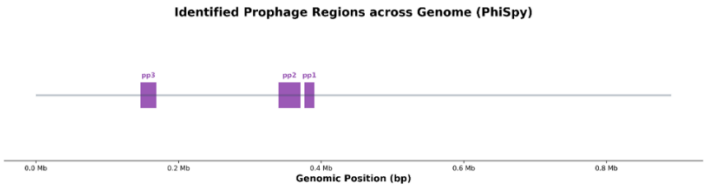

Figure 5. Genomic map illustrating the localization and span of the three integrated prophage regions (pp1–pp3) identified across the assembly scaffolds by the PhiSpy module.

4. Discussion

The validation of BactoFlow with a clinical isolate of *Pseudomonas aeruginosa* (ST308) demonstrated consistent efficacy across all the modules, confirming its ability to convert raw sequencing data into well-structured and interpretable genomic outputs within a unified analytical framework. The metrics for quality control and assembly during the pre-processing and reconstruction stages were consistent with the known genomic features of *P. aeruginosa*, demonstrating the pipeline's capacity to preserve sequencing accuracy throughout these stages [26], [27].

Reference-based variant calling identified the genomic divergence characteristic of the *Pseudomonas aeruginosa* strain against the reference, while sequence typing definitively assigned the isolate to a globally recognized high-risk sequence type (ST308), reinforcing the precision of BactoFlow's mapping and typing modules in clinically relevant scenarios [28], [29]. Functional annotation produced a biologically coherent inventory of genomic features, supporting the reliability of the assembly-to-annotation within the pipeline.

Functional screening against CARD, VFDB, PADLOC, and PhiSpy yielded outputs consistent with the anticipated repertoire of high-risk *P. aeruginosa* lineages. Collectively verifying the database integration modules instead of representing independent characterization. The concordance between predicted functional profiles and established lineage-specific expectations provides meaningful evidence of the pipeline's analytical accuracy.

Overall, these findings establish BactoFlow as a reproducible, end-to-end genomic analysis platform accessible beyond specialist bioinformatics environments. While short-read sequencing imposes inherent constraints, particularly in resolving repetitive regions and localizing mobile genetic elements, these limitations are readily addressable through future hybrid long-read integration. As antimicrobial resistance continues to rise globally, this automated genomic pipeline offers a practical and scalable solution for surveillance and outbreak response.

5. Conclusion

BactoFlow integrates quality control, assembly, annotation, and multi-database resistance, virulence, and defense-system screening into a single automated, reproducible pipeline. Applied to a clinical *Pseudomonas aeruginosa* isolate, the pipeline accurately identified it as high-risk sequence type ST308 and revealed an extensive resistance, virulence, and defense profile, 51 AMR genes, 290 virulence factors, and 12 defense systems, alongside 56,587 genomic variants, underscoring the isolate's clinical significance. Its checkpoint-based, Conda-managed design ensures reliable, reproducible results and can be easily upgraded, including with future long-read data. Overall, BactoFlow provides a scalable, accessible tool for high-resolution bacterial genomics.

Author contributions: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Visualization, Writing-original draft: MARS; Conceptualization, Methodology, Writing - original draft: MAR; Conceptualization, Methodology, Writing - original draft: UHH; Writing - review & editing: HMF; Conceptualization, Methodology, Supervision, Writing - review & editing: SKB.

Data availability: The BactoFlow pipeline, including all source code, configuration files, Conda environment specifications, and documentation, is openly available at https://github.com/mdarsikdar/BactoFlow.git under an open-source license.

Conflicts of interest: There are no potential conflicts of interest.

Funding: Authors received no funding from an external source.

References

[1] C. U. Köser *et al.*, "Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*," *N Engl J Med*, vol. 369, no. 3, pp. 290–292, Jul. 2013, doi: 10.1056/NEJMc1215305.
[2] J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, and E. Trees, "Next-generation sequencing technologies and their application to the study and control of bacterial infections," *Clinical Microbiology and Infection*, vol. 24, no. 4, pp. 335–341, Apr. 2018, doi: 10.1016/j.cmi.2017.10.013.
[3] The Bioconda Team *et al.*, "Bioconda: sustainable and comprehensive software distribution for the life sciences," *Nat Methods*, vol. 15, no. 7, pp. 475–476, Jul. 2018, doi: 10.1038/s41592-018-0046-7.
[4] J. Leipzig, "A review of bioinformatic pipeline frameworks," *Brief Bioinform*, p. bbw020, Mar. 2016, doi: 10.1093/bib/bbw020.
[5] P. A. Ewels *et al.*, "The nf-core framework for community-curated bioinformatics pipelines," *Nat Biotechnol*, vol. 38, no. 3, pp. 276–278, Mar. 2020, doi: 10.1038/s41587-020-0439-x.
[6] B. P. Alcock *et al.*, "CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database," *Nucleic Acids Research*, vol. 51, no. D1, pp. D690–D699, Jan. 2023, doi: 10.1093/nar/gkac920.
[7] B. Liu, D. Zheng, S. Zhou, L. Chen, and J. Yang, "VFDB 2022: a general classification scheme for bacterial virulence factors," *Nucleic Acids Research*, vol. 50, no. D1, pp. D912–D917, Jan. 2022, doi: 10.1093/nar/gkab1107.
[8] L. J. Payne *et al.*, "PADLOC: a web server for the identification of antiviral defence systems in microbial genomes," *Nucleic Acids Research*, vol. 50, no. W1, pp. W541–W550, Jul. 2022, doi: 10.1093/nar/gkac400.
[9] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.
[10] D. M. Mrochen *et al.*, "Wild rodents and shrews are natural hosts of Staphylococcus aureus," *International Journal of Medical Microbiology*, vol. 308, no. 6, pp. 590–597, Aug. 2018, doi: 10.1016/j.ijmm.2017.09.014.
[11] R. Leinonen, H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration, "The Sequence Read Archive," *Nucleic Acids Research*, vol. 39, no. Database, pp. D19–D21, Jan. 2011, doi: 10.1093/nar/gkq1019.
[12] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
[13] E. W. Sayers *et al.*, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 50, no. D1, pp. D20–D26, Jan. 2022, doi: 10.1093/nar/gkab1112.
[14] R. M. Leggett, R. H. Ramirez-Gonzalez, B. J. Clavijo, D. Waite, and R. P. Davey, "Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics," *Front. Genet.*, vol. 4, 2013, doi: 10.3389/fgene.2013.00288.
[15] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
[16] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[17] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," Jul. 20, 2012, *arXiv*: arXiv:1207.3907. doi: 10.48550/arXiv.1207.3907.
[18] A. Bankevich *et al.*, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.
[19] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, doi: 10.1093/bioinformatics/btt086.
[20] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.
[21] A. G. McArthur *et al.*, "The Comprehensive Antibiotic Resistance Database," *Antimicrob Agents Chemother*, vol. 57, no. 7, pp. 3348–3357, Jul. 2013, doi: 10.1128/AAC.00419-13.
[22] E. Zankari *et al.*, "Identification of acquired antimicrobial resistance genes," *Journal of Antimicrobial Chemotherapy*, vol. 67, no. 11, pp. 2640–2644, Nov. 2012, doi: 10.1093/jac/dks261.

[23] L. Chen, "VFDB: a reference database for bacterial virulence factors," *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D325–D328, Dec. 2004, doi: 10.1093/nar/gki008.
[24] L. J. Payne *et al.*, "PADLOC: a web server for the identification of antiviral defence systems in microbial genomes," *Nucleic Acids Research*, vol. 50, no. W1, pp. W541–W550, Jul. 2022, doi: 10.1093/nar/gkac400.
[25] S. Akhter, R. K. Aziz, and R. A. Edwards, "PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies," *Nucleic Acids Research*, vol. 40, no. 16, pp. e126–e126, Sep. 2012, doi: 10.1093/nar/gks406.
[26] C. K. Stover *et al.*, "Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen," *Nature*, vol. 406, no. 6799, pp. 959–964, Aug. 2000, doi: 10.1038/35023079.
[27] G. L. Winsor, E. J. Griffiths, R. Lo, B. K. Dhillon, J. A. Shay, and F. S. L. Brinkman, "Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database," *Nucleic Acids Res*, vol. 44, no. D1, pp. D646–D653, Jan. 2016, doi: 10.1093/nar/gkv1227.
[28] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," *Nat Rev Genet*, vol. 13, no. 1, pp. 36–46, Jan. 2012, doi: 10.1038/nrg3117.
[29] R. Bruyndonckx *et al.*, "Change-points in antibiotic consumption in the community, European Union/European Economic Area, 1997–2017," *Journal of Antimicrobial Chemotherapy*, vol. 76, no. Supplement_2, pp. ii68–ii78, Aug. 2021, doi: 10.1093/jac/dkab179.