Solu – a cloud platform for real-time genomic pathogen surveillance

Moilanen TJ[1], Lehtinen J[1], Visuri K[1], Sihvonen S[1]

1) Solu Healthcare Oy, Helsinki, Finland

Abstract

Summary

- Genomic surveillance is extensively used for tracking public health outbreaks and healthcare-associated pathogens. Despite advances in bioinformatics pipelines, practical infrastructure, expertise and security challenges hinder continuous surveillance.
- Solu addresses this gap with a cloud-based platform (https://platform.solugenomics.com) that integrates genomic data into a real-time, privacy-focused surveillance system.
- In our initial validation, Solu's accuracy for taxonomy assignment, antimicrobial resistance genes, and phylogenetics, was comparable to established pathogen surveillance pipelines.
- By enabling reliable, user-friendly, and privacy-focused genomic surveillance, Solu has the potential to bridge the gap between cutting-edge research and practical, widespread application in healthcare settings.

Availability

The platform is available for free academic use at platform.solugenomics.com.

Introduction

Bacterial and fungal pathogens, along with their antimicrobial resistance, are causing an increasing burden on healthcare and public health (1,2). Advances in microbial genomics have significantly enhanced infection prevention and outbreak surveillance by providing detailed information about pathogen species, antimicrobial resistance, and phylogenetics. (3)

However, as sequencing costs have decreased, data processing has become a significant bottleneck in adopting genomic approaches (4). To address this bottleneck, several pathogen analysis pipelines have emerged recently, including nf-core, TheiaProk, Galaxy, ASA3P, Nullarbor, and Bactopia (5–10). Despite these advancements, the infrastructure for continuous surveillance remains inadequate, with computational resources and trained personnel constituting a major challenge (11).

Most existing pipelines are designed for single-use runs (5-10), and require the user to set up and manage their own infrastructure (5-6, 8-10). One-off analyses are not suitable for continuous surveillance, where new sequencing data is often generated in small, regular

batches (e.g., weekly) (12, 13). To maintain relevance and accuracy, these new results must be seamlessly integrated with previous analyses (12, 13).

Moreover, academia-led projects developed under FAIR (Findable, Accessible, Interoperable, Reusable) principles often lack the necessary privacy focus to meet the stringent requirements of healthcare providers (14). Healthcare providers must adhere to stringent legal requirements, such as the U.S. HIPAA Privacy Rule (15), ruling out many academic tools for clinical genomic surveillance.

Healthcare facilities need access to scalable, user-friendly, and privacy-first infrastructure for ongoing genomic surveillance. In this manuscript, we present a method that meets these needs by integrating genomic characterization and epidemiology into a robust web application.

Methods

We present Solu, a cloud platform for real-time genomic surveillance.

Cloud infrastructure implementation for ongoing surveillance

The cloud infrastructure of Solu is built on three principles: real-time integration, security, and robustness.

Real-time integration

In Solu's infrastructure, each new upload integrates with the cumulative surveillance data, rather than initiating separate pipeline runs.

A highly automated pipeline runner (Figure 1) enables this process, automatically executing new analyses and updating existing ones as inputs change. For instance, when a user uploads a *Salmonella enterica* sample, the platform automatically re-computes the phylogenetics and clustering for all *Salmonella enterica* samples to ensure results remain current.

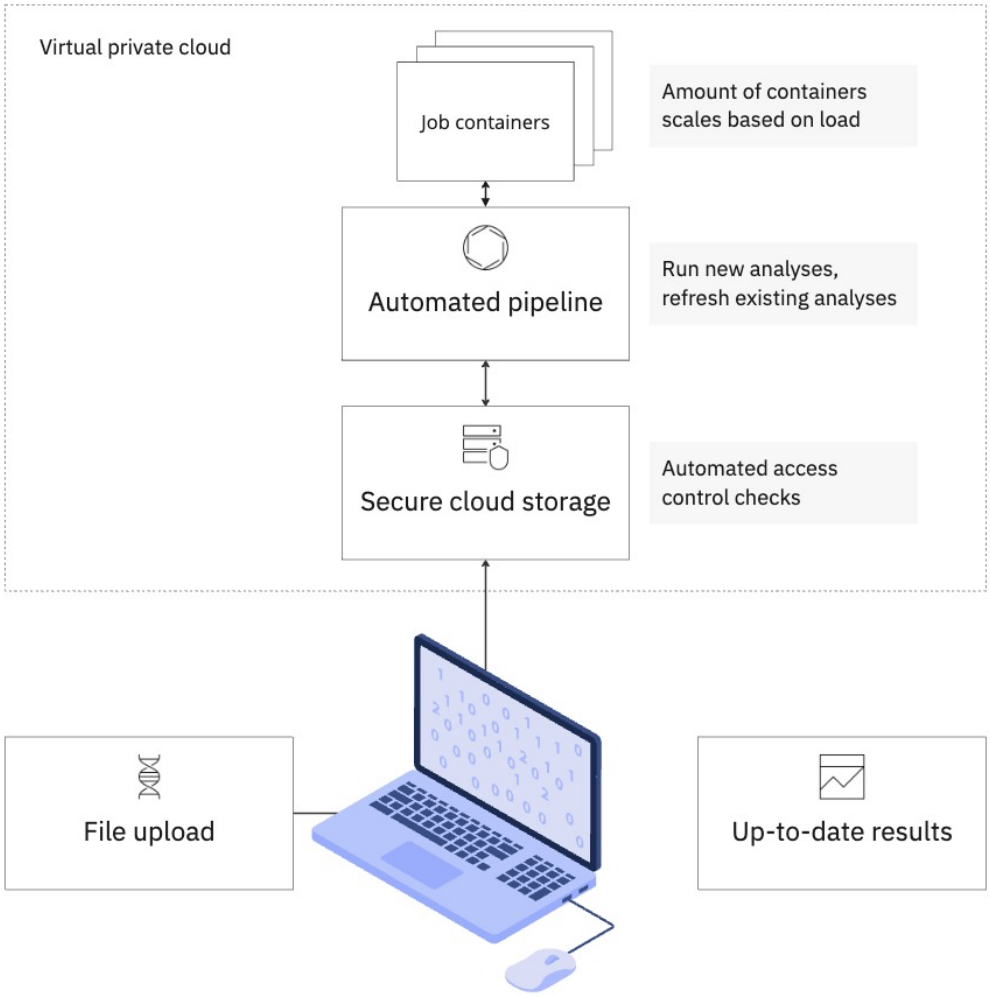

Figure 1. Solu's infrastructure

Security

Data is stored in a secure, single-location cloud storage, each with explicitly set read and write permissions. All computations occur within a private network, monitored by automated access control checks. Solu adheres to the U.S. HIPAA rule (15), implementing strict data security protocols, including appropriate access permissions, authorization, continuous monitoring, encryption, staff training, and other cybersecurity measures. This is a key reason why the platform's code is not open-sourced.

Robustness

The infrastructure is designed to handle variable loads while maintaining fast results even during peak usage. Computation-intensive work is performed by containers separate from the main application, with the number of containers scaling based on demand. If an analysis step fails, it can be re-run without disrupting the entire pipeline. Additionally, the infrastructure

dynamically assigns optimally powerful machines to each job, based on the specific memory and CPU requirements of different pipeline steps.

Bioinformatics pipeline

The platform uses WGS reads or assembled genomes as inputs. Here, we present an overview of the pipeline steps. A full description of tools, versions, and parameters used is available in the supplementary files and at https://www.solugenomics.com/documentation.

Genomic characterization

Raw reads are quality checked using FastQC (16) and quality corrected with fastp (17). The pre-processed reads are assembled using Shovill (18). Assembly quality is assessed with Quast (19), and the genome size is compared to an expected range provided by the NCBI genome API (20). Assembly files are standardized using any2fasta (21).

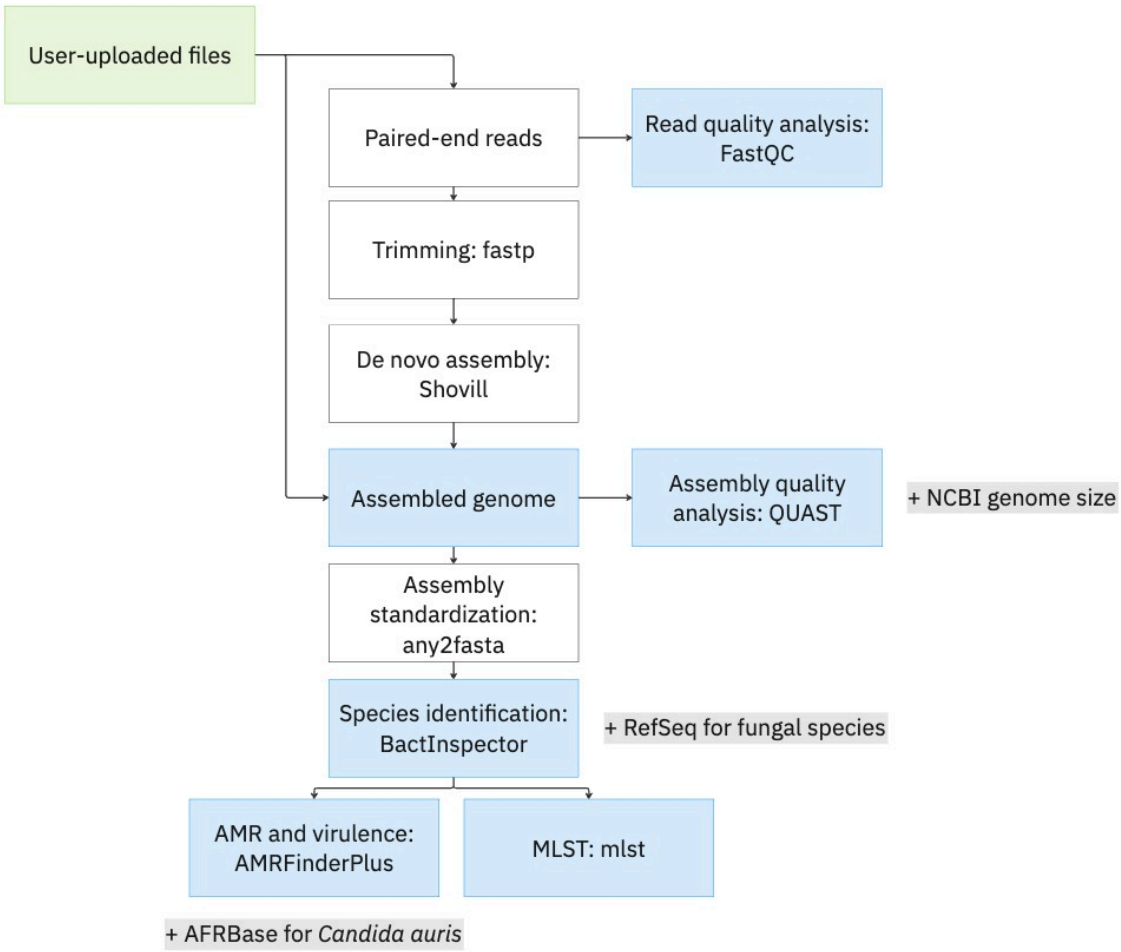

Figure 2. Solu's single-sample pipeline steps

Species identification and MLST are determined with BactInspector (22) and mlst (23). To identify fungal species, we have augmented BactInspector's default database with fungal reference sequences from RefSeq (24).

The antimicrobial resistance (AMR) and virulence genes of bacterial samples are annotated with AMRFinderPlus (25). To detect antifungal resistance, we have augmented AMRFinder's default database with known Candida auris resistance mutations from AFRBase (26).

Phylogenetics

Solu uses both reference-based and reference-free phylogenetic pipelines, depending on the pathogen being analyzed (Figure 3).

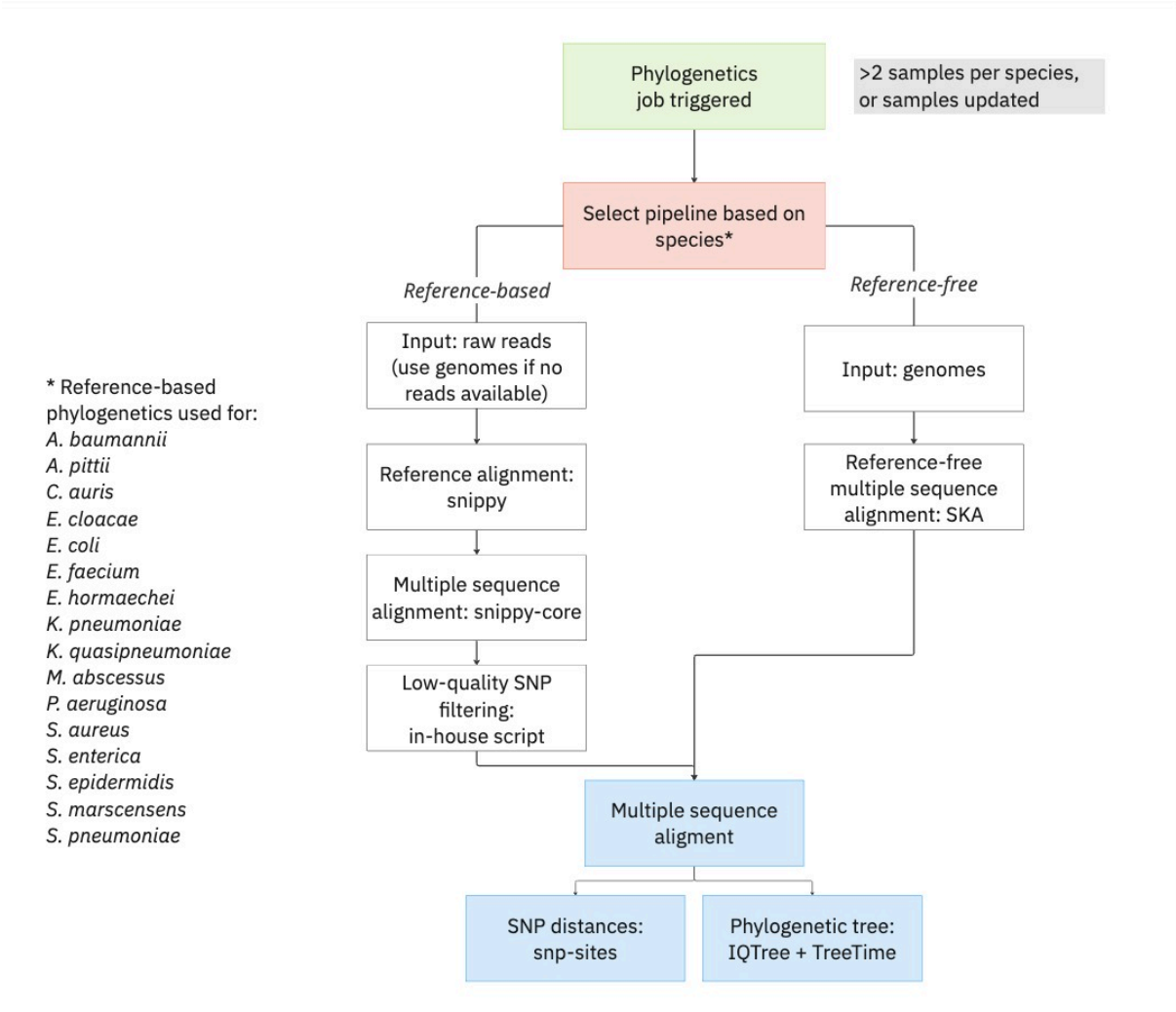

Figure 3. Solu's phylogenetic pipelines

For commonly analyzed species, the reference-based pipeline aligns each genome to a reference genome with Snippy (27) and creates a multiple sequence alignment with snippy-core. Low-quality SNPs are filtered away using an in-house algorithm described in the supplementary material.

For other species, we use the reference-free pipeline which creates the multiple sequence alignment with SKA (28).

SNP distances are counted from the resulting multiple sequence alignment using snp-sites. Samples are clustered using a 20-SNP single-linkage clustering threshold.

A phylogenetic tree for each species is constructed with IQ-tree (29). The resulting tree is midpoint-rooted using TreeTime (30). Both IQ-tree and TreeTime are run using the Augur toolkit (31).

Evaluation

To assess the accuracy of the Solu platform, we evaluated its performance across four microbial datasets: *Staphylococcus aureus*, *Enterococcus faecium*, *Candida auris* and *Salmonella enterica*.

The *S.aureus*, *E.faecium* and *C.auris* datasets were used to validate Solu's species identification, clade assignment, and antimicrobial resistance prediction. The *Salmonella* dataset was derived from an epidemiologically validated outbreak investigation, and was used to assess Solu's phylogenetic reconstruction.

We compared Solu's outputs to established bioinformatics pipelines commonly used in public health genomics: NCBI Pathogen Detection for species identification and AMR, and kSNP3 for phylogenetics. For *Candida auris,* we compared results against the original research publications, due to the limited amount of validated fungal WGS pipelines.

To quantify the concordance between Solu and the benchmark tools, we computed metric scores for species, clade, and AMR predictions. For the phylogenetic trees, we used the TreeDist R package (32) to measure the similarity between the Solu and kSNP3 trees.

All analyses were conducted using paired-end FASTQ reads from the European Nucleotide Archive. A summary of the datasets and comparison methods is presented in Table 1.

Table 1. Overview of the evaluation

Species	**BioProject accession**	**Number of samples**	**Variables compared**	**Pipelines used for comparison**
Staphylococcus aureus	PRJNA722830 (33)	10	Species, AMR	NCBI Pathogen Detection
Enterococcus faecium	PRJNA797179 (34)	10	Species, AMR	NCBI Pathogen Detection
Candida auris	PRJNA328792 (35), PRJNA816104 (36)	17	Species, clade, AMR	Original research publications
Salmonella enterica	PRJNA170556 (37)	23	Species, phylogenetic tree	TheiaProk (kSNP3) (6)

Results

Evaluation results

The Solu platform successfully completed the analysis for all 60 samples, including assembly, genomic characterization, and phylogenetics. Screen-shot of the platform's home screen is shown in Figure 4. This workspace, including all samples and results, is also available at https://platform.solugenomics.com/w/solu-publication-1.

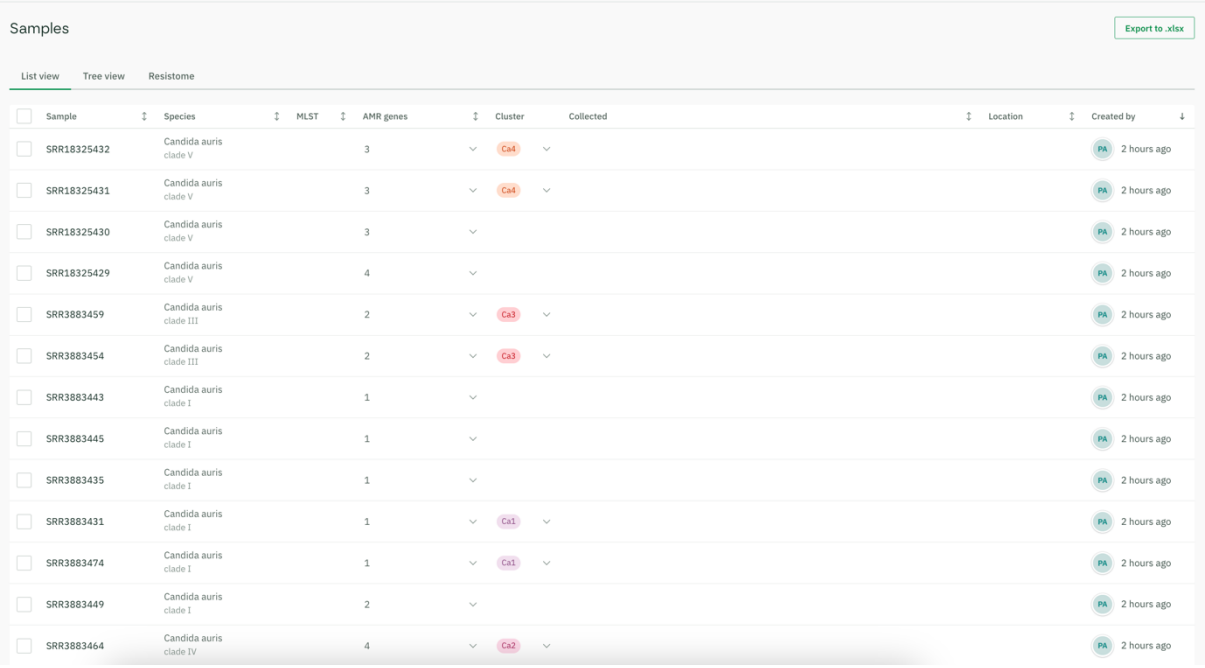

Figure 4. Overview of the platform

Species and clade assignment

Solu accurately identified the species for all 60 samples and correctly assigned the subspecies and serovar of the *Salmonella* Bareilly samples, as well as the clades of the *Candida auris* samples. Full results for each sample are presented in the supplementary file S2. This demonstrates the platform's high accuracy in taxonomic classification.

Antimicrobial resistance

The antimicrobial resistance (AMR) gene detection results from Solu were largely consistent with those obtained using the comparison pipelines (Table 2). NCBI's pipeline detected the abc-f and blaR1 genes in many of the *S.aureus* samples, which were not found by Solu. Additionally, NCBI's pipeline found a few other AMR genes in the *E. faecium* dataset that Solu did not detect. These differences are likely due to NCBI's use of Hidden Markov Models to identify distant functional relatives to genes in the reference gene catalog (38), and are not expected to have significant clinical implications. Full AMR results for each sample are presented in the supplementary file S2.

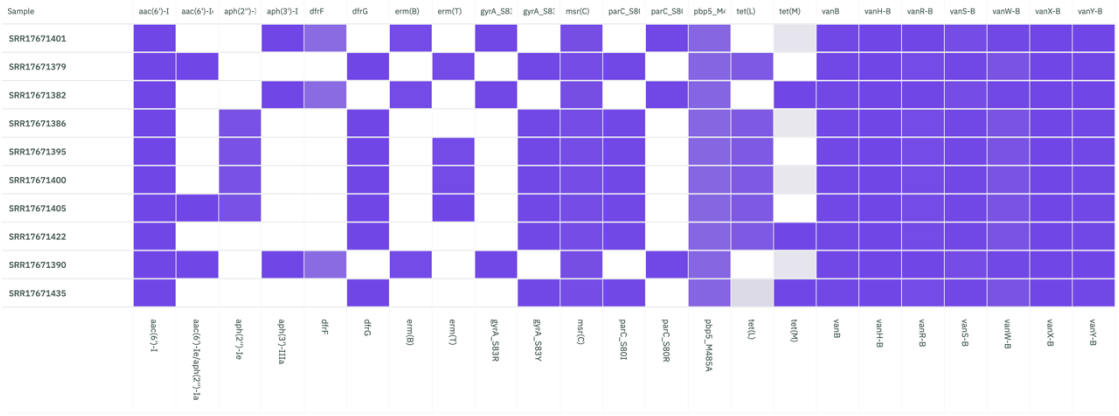

Figure 5. A heatmap of AMR genes for *E. faecium*

For the *Candida auris* samples, Solu's results were also highly concordant with the reference publications. The platform identified the same key resistance markers, such as the ERG11 and Tac1b mutations, as reported in the literature. However, two specific mutations (ERG11_I466L and Tac1b_D559G) described in the Spruijtenburg et al. study were not present in Solu's database and, consequently, were not detected by the platform. Both of these genes have only hypothetical contributions towards antifungal resistance (35). Solu's results included several other resistance-associated mutations that were not detected in the reference articles: V704L and K143R in CDR1 and E343D, K177R, N335S and V125A in ERG11.

Table 2. Summary of AMR detection results

Species	Genes detected by both Solu and the reference	Genes detected only by Solu	Genes detected only by the reference
Staphylococcus aureus	33	1	2
Enterococcus faecium	23	1	4
Candida auris	3	6	2

Phylogenetics

Solu produced a phylogenetic tree for the *Salmonella* Bareilly dataset (Figure 6) with a similar topology to the reference tree, where the outbreak samples are separate from the outgroups.

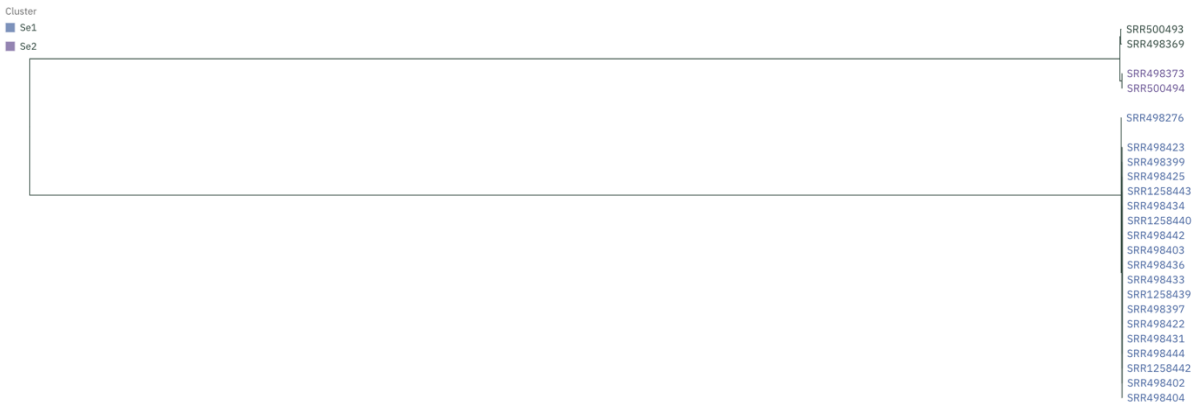

Figure 6. Solu's phylogenetic tree of the *Salmonella* Bareilly dataset

In the computational tree comparison (Table 3), Solu's phylogenetic tree had high similarity to the reference tree, with smaller distances than the kSNP3 tree used as a comparison.

Table 3. Results of the computational tree comparison

Method/Pipeline	**Kendall-Colijn distance (40)**	**Robinson-Foulds distance (41)**
Original publication (used as the reference)	0.00	0
Solu reference-based phylogeny pipeline	6.16	2
kSNP3 workflow	18.95	10

Discussion

Solu integrates the latest advancements in genomic characterization and epidemiology into an easy-to-use web application. Our initial results demonstrate that Solu's pipeline shows promise as an accurate alternative to traditional bioinformatics pipelines.

This was an initial, proof-of-concept evaluation of Solu's performance, with limitations in datasets, methodology and scope. Future work, comparing large datasets to a larger number of established bioinformatics tools, is required to fully assess Solu's capabilities.

By focusing on a robust, privacy-focused infrastructure, Solu facilitates broader adoption of genomic pathogen surveillance, potentially bridging the gap between research and practice.

Availability and requirements

- The Solu platform runs in the browser and is available on all operating systems.
- Project homepage: platform.solugenomics.com
- Dataset: https://platform.solugenomics.com/w/solu-publication-1

References

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022;399(10325):629-55.
2. Vallabhaneni S, Mody RK, Walker T, Chiller T. The Global Burden of Fungal Diseases. Infect Dis Clin North Am. 2016;30(1):1-11.
3. Eyre DW. Infection prevention and control insights from a decade of pathogen whole-genome sequencing. J Hosp Infect. 2022;122:180-6.
4. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect. 2018;24(4):342-9.
5. Ewels PA, Peltzer A, Fillinger S. et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 2020;38: 276–8.
6. Libuit KG, Doughty EL, Otieno JR, Ambrosio F, Kapsak CJ, Smith EA, et al. Accelerating bioinformatics implementation in public health. Microb Genom. 2023;9(7):mgen001051.
7. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research. 2022;50(W1):W345–W351.
8. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, et al. ASA3P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol 2020;16(3):e1007134.
9. Seemann T. Nullarbor. Available at: https://github.com/tseemann/nullarbor.
10. Petit RA 3rd, Read TD. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 2020;5(4):e00190-20.
11. Afolayan AO, Bernal JF, Gayeta JM, et al. Overcoming Data Bottlenecks in Genomic Pathogen Surveillance. Clin Infect Dis. 2021;73(Suppl_4):S267-S274.
12. Roberts LW, Forde BM, Hurst T, et al. Genomic surveillance, characterization and intervention of a polymicrobial multidrug-resistant outbreak in critical care. Microb Genom. 2021;7(3):mgen000530.
13. Forde BM, Bergh H, Cuddihy T, et al. Clinical Implementation of Routine Whole-genome Sequencing for Hospital Infection Control of Multi-drug Resistant Pathogens. Clin Infect Dis. 2023;76(3):e1277-e1284.
14. Stacey D, Wulff K, Chikhalla N, Bernardo T. From FAIR to FAIRS: Data security by design for the global burden of animal diseases. Agron J. 2022;114(5):2693-9.
15. U.S. Department of Health & Human Services. HIPAA Privacy Rule. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Accessed May 29, 2023.
16. Babraham bioinformatics. FastQC. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
17. Shifu Chen. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107.
18. Seemann T. Shovill. Available at: https://github.com/tseemann/shovill.
19. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
20. GenBank. Available at: https://www.ncbi.nlm.nih.gov/genbank/genome-size-check/.
21. Seemann T. any2fasta. Available at: https://github.com/tseemann/any2fasta.
22. Underwood A. BactInspector. Available at: https://gitlab.com/antunderwood/bactinspector.

23. Seemann T. mlst. Available at: https://github.com/tseemann/mlst.
24. RefSeq. Available at: https://www.ncbi.nlm.nih.gov/refseq/.
25. Feldgarden M, Brover V, Gonzalez-Escalona N, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):12728.
26. Aakriti Jain, Neelja Singhal, Manish Kumar. AFRbase: a database of protein mutations responsible for antifungal resistance, Bioinformatics. 2023;39(11):btad677.
27. Seemann T. snippy. Available at: https://github.com/tseemann/s.
28. Harris SR. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. bioRxiv 453142.
29. B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 2020;37:1530-4.
30. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1):vex042.
31. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.
32. Smith, M.R. TreeDist: Distances between Phylogenetic Trees. R package version 2.7.0. Comprehensive R Archive Network.
33. Mardziah, Che & Al-Trad, Esra'a & Puah, Suat Moi, et al. Whole genome sequence analysis of methicillin-resistant Staphylococcus aureus indicates predominance of the EMRSA-15 (ST22-SCCmec IV[2b]) clone in Terengganu, Malaysia. International Journal of Antimicrobial Agents. 2021;58:21004039.
34. Permana B, Harris PNA, Runnegar N, et al. Using Genomics To Investigate an Outbreak of Vancomycin-Resistant Enterococcus faecium ST78 at a Large Tertiary Hospital in Queensland. Microbiol Spectr. 2023;11(3):e0420422.
35. Lockhart SR, Etienne KA, Vallabhaneni S, et al. Simultaneous Emergence of Multidrug-Resistant Candida auris on 3 Continents Confirmed by Whole-Genome Sequencing and Epidemiological Analyses [published correction appears in Clin Infect Dis. 2018 Aug 31;67(6):987]. Clin Infect Dis. 2017;64(2):134-140. doi:10.1093/cid/ciw691.
36. Spruijtenburg B, Badali H, Abastabar M, et al. Confirmation of fifth Candida auris clade by whole genome sequencing. Emerg Microbes Infect. 2022;11(1):2405-2411. doi:10.1080/22221751.2022.2125349.
37. Timme RE, Rand H, Shumway M, et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. PeerJ. 2017;5:e3893. Published 2017 Oct 6. doi:10.7717/peerj.3893.
38. Smith MR. Information theoretic Generalized Robinson-Foulds metrics for comparing phylogenetic trees. Bioinformatics. 2020;36:5007–5013. doi: 10.1093/bioinformatics/btaa614.
39. HMM Catalog [Internet]. Bethesda(MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2024 May 29]. Available from: https://www.ncbi.nlm.nih.gov/pathogens/docs/HMM_catalog/.
40. Kendall M, Colijn C. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. Molecular biology and evolution. 2016;33(10):2735–43.
41. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Mathematical biosciences. 1981;53(1-2):131-147.