arXiv:submit/6610895 [q-bio.GN] 1 Jan 1970

MicroTrace: A Lightweight R Tool for SNP-Based Pathogen Clustering in Outbreak Detection

Kaitao Lai[1]

2025-07-02

[1] University of Sydney

1 Summary

MicroTrace is an open-source R tool that performs SNP-based hierarchical clustering to detect potential transmission clusters from pathogen whole-genome sequencing (WGS) data. Designed for epidemiologists, microbiologists, and genomic surveillance teams, it processes SNP distance matrices and outputs dendrograms and cluster tables with optional metadata integration. MicroTrace enables reproducible outbreak detection workflows with minimal setup.

2 Statement of Need

This tool was motivated by my own experience supporting hospital outbreak investigations where rapid, reproducible, and interpretable clustering was required—but often slowed by the complexity or inflexibility of available pipelines. During exploratory analyses of SNP distance matrices, I observed that closely related isolates often exhibited SNP distances in the lowest decile of all pairwise comparisons. As such, MicroTrace uses the **10th percentile of SNP distances** as a conservative default threshold for outbreak definition—prioritizing sensitivity in early cluster detection. This approach reflects empirical patterns in genomic epidemiology (Payne et al. 2021) while remaining adjustable for local context.

Although whole-genome sequencing (WGS) has revolutionized pathogen surveillance (K"oser et al. 2012), outbreak detection workflows often remain fragmented and require specialized tools or pipelines. Tools like Snippy (Seemann 2015) provide variant calling, and visualization platforms such as GrapeTree offer phylogenetic context, but a lightweight, scriptable R solution for direct clustering from SNP matrices is lacking.

MicroTrace addresses this need by providing: - Automated hierarchical clustering with configurable SNP thresholds - Distance distribution-based threshold recommendation - Metadata integration for spatiotemporal context - Scripted reproducibility and a markdown-based HTML report

This simplicity supports real-time response in clinical microbiology and public health settings.

3 MicroTrace Design and Features

MicroTrace consists of modular R functions and follows a clear processing workflow:

- `distance_loader()`: Loads a pairwise SNP distance matrix from CSV
- `auto_threshold_suggestion()`: Suggests a clustering threshold (e.g., 10th percentile SNP distance) and generates histogram and density plots
- `run_microtrace_clustering()`: Performs UPGMA clustering, cuts the tree at the threshold, joins optional metadata, and generates output files

Additional features include: - Optional metadata integration (collection date, ward, patient ID) - Publication-ready PNG output of dendrogram and SNP distance plots - Markdown-based HTML report template - Intra-cluster SNP summary statistics (mean, SD, range) - Unit tests for all core functions

4 Simulated Dataset Example

A simulated dataset (10 samples) demonstrates two genetically distinct clusters. The SNP distances among samples from the same hospital ward (Ward A) are 0–3, while those from Ward B are 7–10. When clustered with a 5-SNP threshold, MicroTrace correctly identifies two distinct groups.

Metadata are merged into the output table, allowing alignment of genetic clusters with sample origin. The example illustrates utility for localized outbreak tracking in hospitals or communities.

5 Reproducibility and Testing

MicroTrace includes a `tests/` directory with automated tests using `testthat`, covering: - Distance matrix loading - Threshold estimation - Cluster output generation - Output dimensions and object class checks

This supports reproducibility and continuous integration.

6 Visualization and Reporting

The `MicroTrace_Report.Rmd` template provides a user-friendly, customizable HTML report. It summarizes the clustering results, visualizes the SNP distance histogram, density plot, dendrogram, and presents metadata-aware statistics. This facilitates communication with infection control or public health teams.

7 Software Repository

The source code for MicroTrace is freely available on GitHub at:
https://github.com/biosciences/MicroTrace

8 Acknowledgements

The idea for MicroTrace originated during my involvement in antimicrobial resistance surveillance projects, where I observed a recurring gap between high-throughput WGS data and pragmatic, interpretable clustering tools usable by clinical microbiologists. This led me to develop a self-contained R solution focused on outbreak clustering logic, metadata overlay, and report generation. I thank the research community of University of Sydney for fostering applied genomics thinking and encouraging a culture of lightweight, real-time analytics in infectious disease genomics.

The author acknowledges colleagues from University of Sydney for insights on real-time SNP clustering in public health genomics.

References

K"oser, C. U., M. J. Ellington, E. J. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. G. Holden, et al. 2012. “Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology.” *PLoS Pathogens* 8 (8): e1002824. https://doi.org/10.1371/journal.ppat.1002824.

Payne, M., D. J. Ingle, M. Valcanis, T. Seemann, et al. 2021. “Enhancing Genomics-Based Outbreak Detection of Endemic Salmonella Enterica Serovar Typhimurium Using Dynamic Thresholds.” *Microbial Genomics* 7 (6): 000310. https://doi.org/10.1099/mgen.0.000310.

Seemann, Torsten. 2015. “Snippy: Rapid Haploid Variant Calling and Core Genome Alignment.” https://github.com/tseemann/snippy.