




# Mashpit: sketching out genomic epidemiology

Tongzhou Xu <sup>1</sup>, Henk C. den Bakker <sup>1</sup>, Xiangyu Deng <sup>1</sup>, and Lee S. Katz <sup>1,2</sup>

<sup>1</sup> Center for Food Safety, University of Georgia, Griffin, GA, United States of America <sup>2</sup> Enteric Diseases Laboratory Branch (EDLB), Centers for Disease Control and Prevention, Atlanta, GA, United States of America

DOI: [10.21105/joss.07306](https://doi.org/10.21105/joss.07306)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Charlotte Soneson](#) 

## Reviewers:

- [@hkaspersen](#)
- [@mberacochea](#)

Submitted: 09 September 2024

Published: 17 December 2024

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

We are in the era of genomic epidemiology. The surveillance of many transmissible diseases is increasingly being conducted through whole genome sequencing of pathogenic agents. One notable example is *Salmonella*, a major foodborne pathogen routinely sequenced by surveillance programs such as PulseNet ([Swaminathan et al., 2001](#)). Large volumes of *Salmonella* genomes from these programs are deposited in database systems including NCBI ([Nadon et al., 2017](#)). These publicly available genomes can be analyzed in a variety of ways such as serotyping ([Zhang et al., 2019](#)), multilocus sequence typing (MLST) ([Zhou et al., 2020](#)), and single nucleotide polymorphism (SNP) typing ([Katz et al., 2017](#)). These analyses provide important laboratory evidence for outbreak surveillance and investigation.

As of August 2024, there are more than 600 thousand *Salmonella* genomes and more than half a million other pathogen genomes at NCBI Pathogen Detection (<https://www.ncbi.nlm.nih.gov/pathogens>). These numbers are expected to increase dramatically and therefore faster analysis methods are needed.

There have been some major advances to scale up bioinformatic analyses to large volumes of pathogenic genomes. One approach is to provide centralized resources that integrate data and analytical tools. For example, NCBI Pathogen Detection combines information from three databases: SRA, GenBank, and BioSample. About once a day, it compares all genomes of a given taxon, separates all genomes into individual clusters using MLST, and then creates a phylogeny for each cluster using SNP analysis. This method is quite comprehensive, but it relies on each sample being public, and it cannot be executed locally.

Another approach is to provide new tools for decentralized and customized manipulation of genomics resources. We observed that an algorithm for genomics called MinHash is well positioned for this purpose. A commonly used software for MinHash is called Mash ([Ondov et al., 2016](#)). Querying with Mash can be about 4 orders of magnitude faster than other common methods like Basic Local Alignment Search Tool (BLAST) and can have a smaller disk footprint ([Camacho et al., 2009](#); [Topaz et al., 2018](#)). Therefore it can be run on more common scientific workstations.

We present Mashpit, a new rapid genomic epidemiology platform to query against these large groups of genomes on a local computer.

## Statement of need

Querying a sample against these magnitudes of genomes is becoming less sustainable, especially for smaller laboratories. Currently, GISAID and NCBI are staying ahead of the curve by producing a global tree of each organism every day ([Shu & McCauley, 2017](#)). This requires herculean efforts, cutting-edge algorithms, and powerful computers. However, smaller laboratories usually

have a scientific workstation or similar equipment, much different than a cluster computing system.

We note that for some organisms like *Salmonella*, queries can be of a sensitive nature. For example, harboring isolates in food production environments that are related to outbreak isolates is often perceived as a potential liability by food establishments, therefore thwarting the efforts to use and share the genomes of these organisms.

To address any needs for speed and sensitivity, we created Mashpit. Mashpit queries genomes locally using Mash, thereby achieving speedy results while keeping any sensitive queries offline.

## Mashpit design

Mashpit consists of three major parts: A MinHash database, its associated metadata, and the MinHash querying.

The database is created with an interface to Mash, called Sourmash (Brown & Irber, 2016). Each genome is imported by sketching it and adding it to a Sourmash signature database. Each genome can also have an entry in the associated metadata. These data include date of isolation, geography, host age range, and other information that could be useful in an epidemiological investigation. Mashpit can build a species database from NCBI Pathogen Detection, termed a Mashpit taxon database, or a custom database from a list of biosample accessions. For each SNP cluster of one species on NCBI Pathogen Detection, the set of all genomes is defined as:

$$G = \{g_1, g_2, \dots, g_n\}$$

where  $n$  is the number of genomes in the cluster. The centroid genome  $g_c$  is calculated as:

$$g_c = \underset{g_i \in G}{\operatorname{argmin}} \sum_{j=1}^n d(g_i, g_j)$$

where  $d(g_i, g_j)$  is the distance between two genomes. The centroid genome represents the most central genome in each SNP cluster, reducing redundancy while retaining representative information for queries. By default, Mashpit will download the latest SNP cluster for specified species and uses a kmer size of 31 and kmer number of 1000 for sketching the genomes.

With the database and its metadata complete, a user could perform a query. The query is an assembly fasta file, which is then sketched and compared against the signature database. The query then returns a tab delimited spreadsheet, sorted by Mash distance, and a phylogenetic tree based on the Mash distance. All associated metadata are included in the spreadsheet.

Mashpit also provides a webserver interface for users to query the database. The webserver is built using Flask and can be run locally or deployed on a server. The webserver provides a user-friendly interface for users to upload their query genomes and view the results.

## Performance

To evaluate the performance of Mashpit, we tested it on a server that runs Ubuntu 20.04.2 with an Intel Xeon CPU E5-2697 v4 2.30GHz and 256GB RAM. We used NCBI Pathogen Detection SNP clusters that were versioned before January 2024. We then randomly selected 1000 newly added genomes for each species added to NCBI Pathogen Detection after January 2024. We measured the elapsed time for querying four major foodborne pathogens: *Salmonella*, *Listeria*, *E. coli*, and *Campylobacter* (Figure 1). We also compared the query results with the true SNP cluster of the query genomes. We calculated the proportion of true SNP clusters appearing among the top hits at various thresholds (Figure 2). The 'threshold' indicates whether the correct SNP cluster is among the top 'threshold number' of query hits. For instance, a threshold of 25 indicates that the correct cluster is among the top 25 hits. Our

findings indicate that *Salmonella* achieved a 70% success rate for true clusters appearing within the top 25 hits, compared to approximately 90% for *Campylobacter*. This variability reflects differences in how species are represented in the database and the limitations of MinHash-based methods for resolving closely related clusters.

For *Salmonella*, which is the most frequently sequenced organism in NCBI Pathogen Detection, many closely related SNP clusters exist due to its extensive representation. Mash, being a MinHash-based method, operates at a resolution that is not always sufficient to distinguish fine-scale differences between these closely related clusters. As a result, users analyzing *Salmonella* should interpret Mashpit results as preliminary and consider following up with higher-resolution methods for definitive SNP cluster assignments.

## Discussion

Mashpit provides a fast and lightweight platform for genomic epidemiology. Its MinHash-based approach enables rapid querying of large datasets on standard scientific workstations, addressing key challenges for laboratories with limited computational resources or privacy concerns.

However, we note that the Mash distance does not correlate well with established distances such as Average Nucleotide Identity (ANI) for closely related genomes (Jain et al., 2018). Therefore it has resolution limits when differentiating closely related clusters, particularly for species like *Salmonella* that are highly represented in databases such as NCBI Pathogen Detection.

Therefore we recommend that this platform is used as a first-pass to filter unrelated samples before using a more established protocol such as MLST. In conclusion, we believe that Mashpit is an essential genomic epidemiology tool.

## Figures

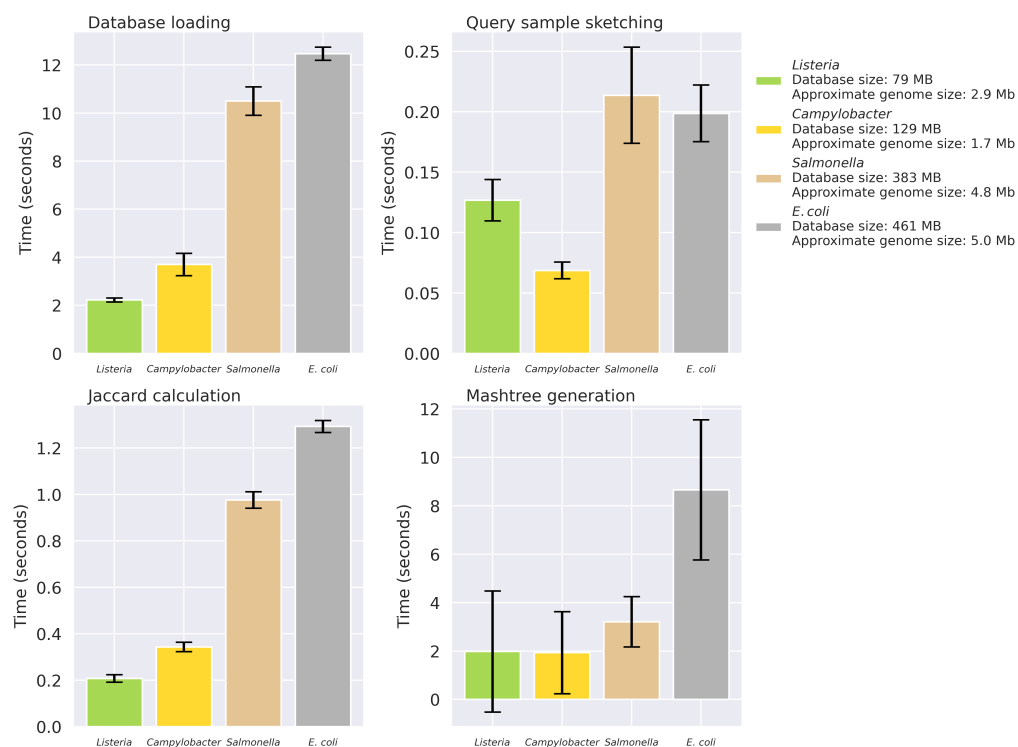
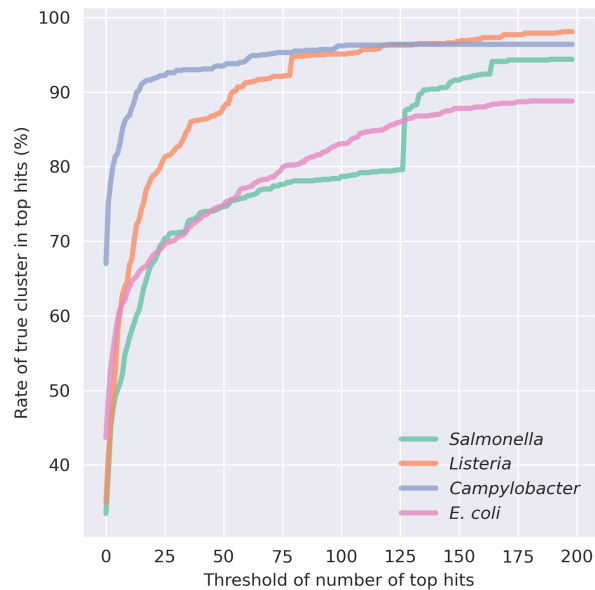


Figure 1: Average query time for four Mashpit taxon databases.



**Figure 2:** Probability of the true SNP cluster being included among the highest-ranking hits at varying thresholds.

## Acknowledgements

Financial support for the development of Mashpit was provided by the Center for Food Safety at the University of Georgia, United States of America. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

## References

- Brown, C. T., & Irber, L. (2016). Sourmash: A library for MinHash sketching of DNA. *Journal of Open Source Software*, 1(5), 27. <https://doi.org/10.21105/joss.00027>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 1–9. <https://doi.org/10.1186/1471-2105-10-421>
- Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1), 5114. <https://doi.org/10.1038/s41467-018-07641-9>
- Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., Van Domselaar, G., Deng, X., & Carleton, H. A. (2017). A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. *Frontiers in Microbiology*, 8, 375. <https://doi.org/10.3389/fmicb.2017.00375>
- Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., Gilpin, B., Smith, A. M., Kam, K. M., Perez, E., & others. (2017). PulseNet international: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Eurosurveillance*, 22(23), 30544. <https://doi.org/10.2807/1560-7917.es.2017.22.23.30544>
- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., &

- Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1), 1–14. <https://doi.org/10.1186/s13059-016-0997-x>
- Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*, 22(13), 30494. <https://doi.org/10.2807/1560-7917.es.2017.22.13.30494>
- Swaminathan, B., Barrett, T. J., Hunter, S. B., Tauxe, R. V., & CDC PulseNet Task Force. (2001). PulseNet: The molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases*, 7(3), 382. <https://doi.org/10.3201/eid0703.017303>
- Topaz, N., Boxrud, D., Retchless, A. C., Nichols, M., Chang, H.-Y., Hu, F., & Wang, X. (2018). BMScan: Using whole genome similarity to rapidly and accurately identify bacterial meningitis causing species. *BMC Infectious Diseases*, 18, 1–9. <https://doi.org/10.1186/s12879-018-3324-1>
- Zhang, S., den Bakker, H. C., Li, S., Chen, J., Dinsmore, B. A., Lane, C., Lauer, A., Fields, P. I., & Deng, X. (2019). SeqSero2: Rapid and improved Salmonella serotype determination using whole-genome sequencing data. *Applied and Environmental Microbiology*, 85(23), e01746–19. <https://doi.org/10.1128/aem.01746-19>
- Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., Achtman, M., Brown, D., Chattaway, M., Dallman, T., Delahay, R., Kornschober, C., & others. (2020). The Enterobase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Research*, 30(1), 138–152. <https://doi.org/10.1101/gr.251678.119>