frontiers | Frontiers in Microbiology

TECHNOLOGY AND CODE
published: 16 June 2022
doi: 10.3389/fmicb.2022.797997

Using Evolutionary Analyses to Refine Whole-Genome Sequence Match Criteria

Arthur W. Pightling, Hugh Rand and James Pettengill*

Biostatistics and Bioinformatics Staff, Office of Analytics and Outreach, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, United States

Whole-genome sequence databases continue to grow. Collection times between samples are also growing, providing both a challenge for comparing recently collected sequence data to historical samples and an opportunity for evolutionary analyses that can be used to refine match criteria. We measured evolutionary rates for 22 *Salmonella enterica* serotypes. Based upon these measurements, we propose using an evolutionary rate of 1.97 single-nucleotide polymorphisms (SNPs) per year when determining whether genome sequences match.

Keywords: whole-genome sequence, evolutionary analyses, *Salmonella enterica*, outbreak investigation, evolutionary rate, resident strain, match, closely related genetically

OPEN ACCESS

Edited by:
David Rodriguez-Lazaro, University of Burgos, Spain

Reviewed by:
Ken-ichi Lee, National Institute of Infectious Diseases (NIID), Japan
Sandeep Tamber, Health Canada, Canada

****Correspondence:***
Arthur W. Pightling
arthur.pightling@fda.hhs.gov

Specialty section:
This article was submitted to Food Microbiology, a section of the journal Frontiers in Microbiology

Received: *19 October 2021*
Accepted: *16 May 2022*
Published: *16 June 2022*

Citation:
Pightling AW, Rand H and Pettengill J (2022) Using Evolutionary Analyses to Refine Whole-Genome Sequence Match Criteria. Front. Microbiol. 13:797997. doi: 10.3389/fmicb.2022.797997

BACKGROUND

Whole-genome sequence (WGS) databases exist for many important bacterial pathogens[1] and the analysis of such data is routine during surveillance and outbreak investigations (Allard et al., 2016). WGS data is commonly used to identify matching isolates. (i.e., isolates that arose from a recent source of contamination; Pightling et al., 2018). Determination of isolate matches usually relies on estimates of genomic distances and tree topologies; generally, these determinations do not incorporate collection times of isolates (Pightling et al., 2018; Wang et al., 2018). Omitting this temporal information is satisfactory when the time spans between collection dates are small, but evolutionary changes could lead to incorrectly inferring mismatches when the time spans are large. That is, isolates collected only a year apart are expected to have fewer genomic differences than isolates that were collected 10 years apart; applying the same genomic distance match criteria to each pair may not be appropriate. Thus, it is necessary to develop match criteria that incorporate the time spans separating collection times. Here, we present the estimated evolutionary rates of bacteria representing 22 *Salmonella enterica* serotypes that were collected from US food manufacturers and show how those rates can be incorporated into distance analyses that are used to assess matches between *S. enterica* genomes.

[1]ncbi.nlm.nih.gov/pathogens/organisms/

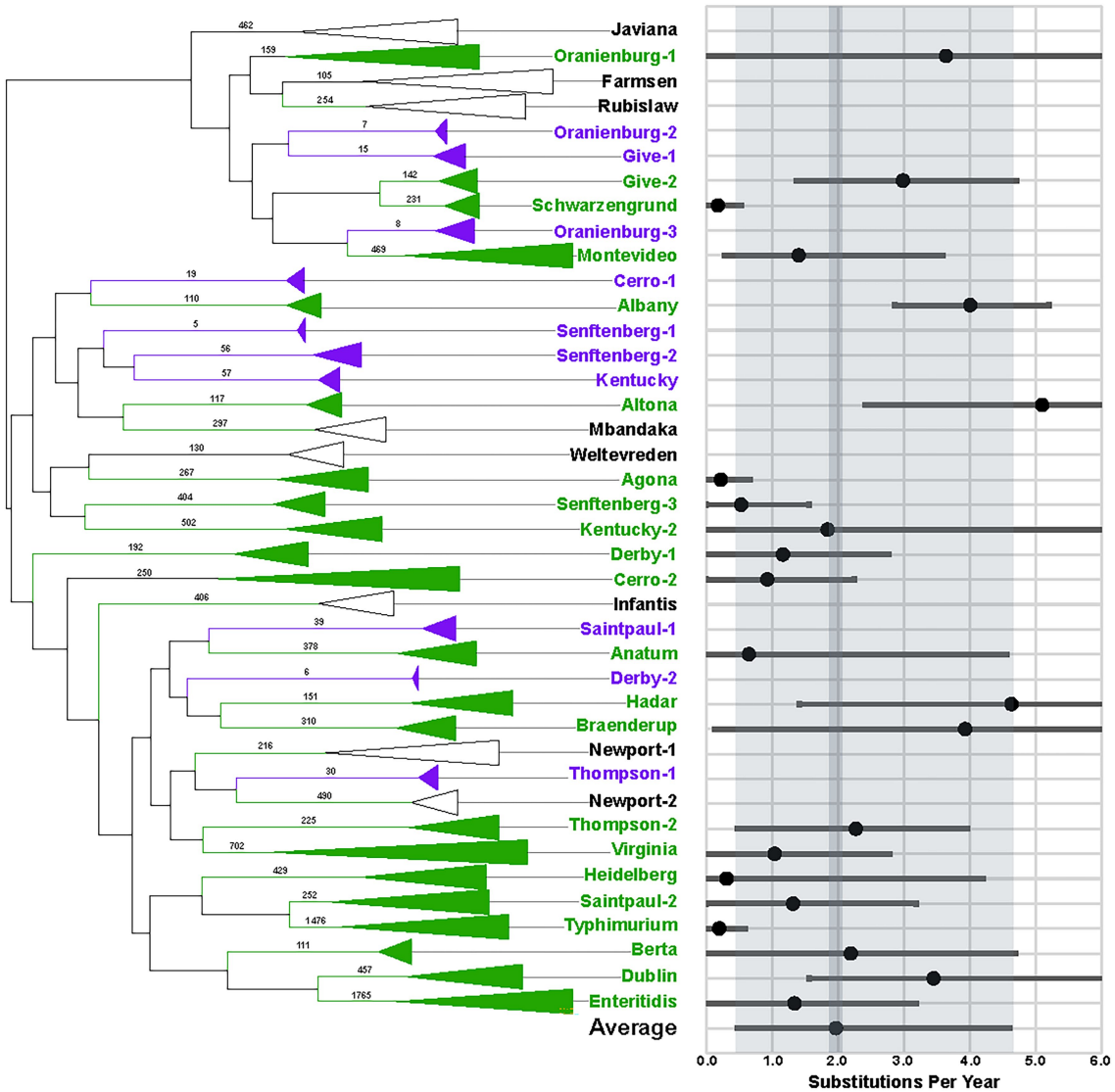

FIGURE 1 | Phylogeny and evolutionary rate measurements of 29 *Salmonella enterica* serotypes. Values on branches of the phylogenetic tree indicate the numbers of taxa that comprise each clade. Green labels indicate that ≥100 taxa were available for evolutionary analyses, while purple labels show clades with <100 taxa that were not analyzed. Clades that have ≥100 taxa but did not exhibit clock-like behavior are labeled with black. Results of BEAST analyses are indicated in the graph with black dots. Bars indicate 95% HPD values. Gray shading shows the averages of the serotypes.

MATERIALS AND METHODS

Dataset

We identified 15,580 *S. enterica* genome sequences that: (1) originated in the United States from 2014 to 2019, (2) were generated on Illumina platforms, and (3) were submitted to the National Center for Biotechnology Information's Pathogen Detection (NCBI's) portal[2] by the U.S. Food and Drug Administration's Center for Food Safety and Applied Nutrition (**Supplementary Table 1**).

[2]ncbi.nlm.nih.gov/pathogens

Phylogenetic Analysis

We assembled 15,580 *S. enterica* genome sequences with SPAdes v3.13.0 (Bankevich et al., 2012) and used SeqSero v1.0.1 (Zhang et al., 2015) to predict their serotypes. We selected 11,701 taxa for further analysis that represent the 29 serotypes with at least 100 isolates (**Supplementary Table 2**). We defined open reading frames with PROKKA v1.12 (Seemann, 2014) and used BLAST v2.7.1+ (Altschul et al., 1990) to find 1,152 loci that comprise an extended multi-locus sequence typing (MLST) scheme (Pettengill et al., 2016). Sequence data were aligned with MAFFT v7.305b (mafft -adjustdirection infile > outfile; Katoh and Standley, 2013). Parsimony informative single-nucleotide polymorphisms (SNPs) were identified and concatenated into a single alignment

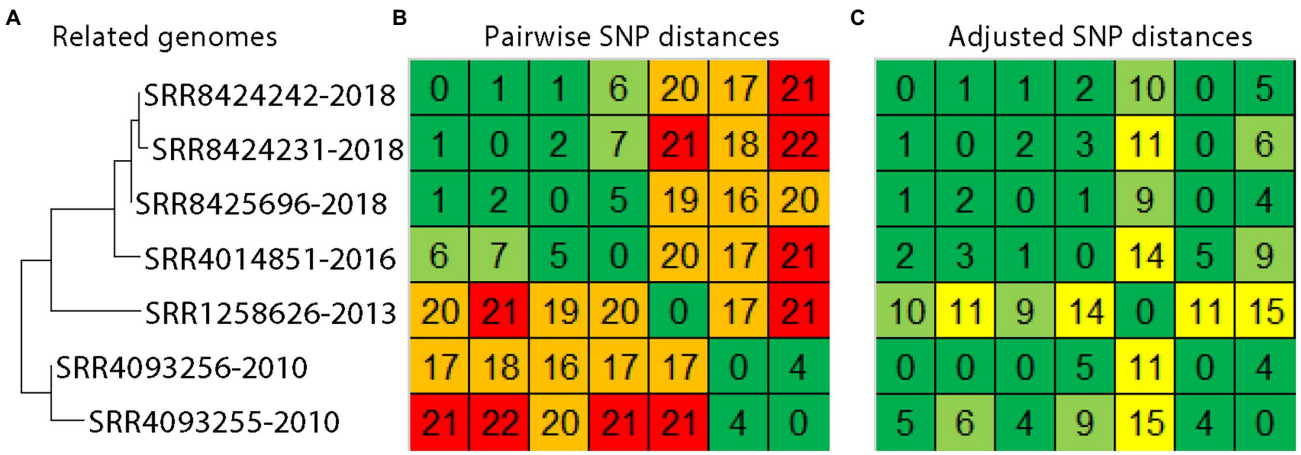

FIGURE 2 | Phylogenetic and pairwise SNP analysis of *Salmonella enterica* collected from the same facility over time. A phylogeny **(A)** and pairwise SNP distance matrix **(B)** was generated with the CFSAN SNP pipeline. Distances were adjusted manually to reflect the time spans separating collections at the rate of 1.97 SNPs/year **(C)**.

with FASconCAT-G v1.04 (FASconCAT-G_v1.04.pl. -o -j -s; Kuck and Longo, 2014). The supermatrix was phylogenetically analyzed with FastTree v2.1.11 SSE3 (FastTreeMP -fastest -nt -gtr < FcC_supermatrix.fas > tree; Price et al., 2009, 2010). The resulting tree was edited with FigTree v1.4.4.[3]

Evolutionary Rate Measurements

We estimated evolutionary rates for 22 lineages that are comprised of at least 100 taxa. We generated lineage-specific phylogenetic analyses using the CFSAN SNP Pipeline (Davis et al., 2015). We then investigated lineages within those trees with TempEst v1.5.3 (Rambaut et al., 2016) to determine which exhibit clock-like behavior and to remove long-branches. Those lineages with clock-like behavior were analyzed further. We generated alignments of assemblies for each clock-like lineage with SKA v1.0, using default settings, and identified regions of recombination with Gubbins v1.4.5 (-first-tree-builder fasttree -tree-builder fasttree -first-model JC -model GTRCAT; Croucher et al., 2015). Regions of recombination were masked and the resulting alignments were analyzed with BEAST v2.6.2, using default settings (Bouckaert et al., 2019). We used the General time reversible (GTR) nucleotide substitution (Tavaré, 1986) model with both the Strict and Relaxed (Drummond et al., 2006) Log Normal clock models and the Coalescent Constant (Drummond et al., 2005) demographic model for 10^8 generations, sampling every 5,000 generations. BEAST outputs were visualized with Tracer v1.7.1 (Rambaut et al., 2018). BEAST runs with effective sampling sizes of at least 200 were analyzed further. For four lineages, both the Strict and Relaxed models had effective sampling sizes of at least 200. In these cases, nested sampling was used to select the best-fitting models (Bouckaert et al., 2019). The numbers of SNPs per year were calculated by multiplying the rates estimated with BEAST and the numbers of unmasked sites in the alignments. The slowest rates are reported for those serotypes in which multiple lineages were analyzed.

[3]http://tree.bio.ed.ac.uk/software/figtree

RESULTS AND DISCUSSION

We identified the most common *S. enterica* serotypes that were isolated from food and environmental samples in the United States and submitted to the NCBI by the US Food and Drug Administration from 2014 to 2019. We then phylogenetically analyzed the WGS data (**Figure 1**). Interestingly, we found that 24.2% (7/29) of the serotypes examined are polyphyletic (Cerro, Derby, Give, Oranienberg, Saintpaul, Senftenberg, and Thompson), which has been documented elsewhere (Timme et al., 2013) and further supports that serotypes are not always reliable for estimating genomic similarity (Wattiau et al., 2011).

We estimated evolutionary rates for 22 lineages (**Figure 1**, green labels and **Supplementary Table 3**). The average evolutionary rate measured is 1.97 single-nucleotide polymorphisms (SNPs) per year (**Figure 1**, dark gray shading), with an average highest posterior density (HPD) interval of 0.48-4.61 SNPs/year (**Figure 1**, light gray shading). The slowest evolutionary rate is 0.18 SNPs/year for *S.* Schwarzengrund (HPD 0-0.53), while the fastest is 5.10 SNPs/year for *S.* Altona (HPD 2.41-7.72). However, since most of the rates measured fall within the average HPD interval, disparities between lineages are less likely to represent true evolutionary differences than to reflect the variability inherent in rate estimates. Thus, we propose that the average of 1.97 SNPs/year be used for determining whether genomes match, while being mindful that evolutionary rates for lineages are likely to vary, depending upon the conditions that they are exposed to.

These results can be used when establishing matches between *S. enterica* genome sequences. For instance, evolutionary rates can be applied to genome sequence data that were collected at different times. As a test case, we used the evolutionary rates observed here to adjust SNP distances for a real-life genome sequence data-set (**Figure 2A**). *Salmonella enterica* were isolated from samples that were taken from the environment of a food processing facility and products that originated from that facility. The samples were collected over

a time span of 8 years. *Salmonella enterica* that were collected in 2010 are estimated to be 16-22 SNPs distant from *S. enterica* collected from the same firm in 2018 (**Figure 2B**, orange and red squares). These SNP distances may obscure the true relationships between genomes, being beyond the range that may often be considered when assessing matches. However, adjusting SNP distances by the average evolutionary rate of 1.97 SNPs/year (i.e., initial SNP distance-[1.97 SNPs/year*(collection date difference in years)] = adjusted SNP distance, with a minimum of 0 SNPs) yields a range of 0-6 SNPs (**Figure 2C**, yellow and green squares), which more accurately reflects their shared origin.

CONCLUSION

As genome sequence databases continue to grow, evolutionary analyses are increasingly important for assessing matches between isolates that are separated by ever greater gaps in time. By applying an evolutionary rate of 1.97 SNPs per year, the time spans separating sample collections can be accounted for. The rate proposed here provides general guidance but should not be used in a strict manner, since conditions in individual cases or lineages may vary. This approach will help to elucidate relationships between bacteria, even as changes accumulate, and to reduce bias that may be introduced when comparing WGS data.

DATA AVAILABILITY STATEMENT

The dataset supporting the conclusions of this article are available in the National Center for Biotechnology Information repository. BEAST XML files are available at figshare (https://doi.org/10.6084/m9.figshare.19617234).

AUTHOR CONTRIBUTIONS

AP, HR, and JP conceived the study and wrote the manuscript. AP performed the analyses. HR and JP provided materials and funding. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the US Food and Drug Administration, Center for Food Safety and Applied Nutrition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.797997/full#supplementary-material

REFERENCES

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* 54, 1975–1983. doi: 10.1128/JCM.00081-16

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., et al. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43:e15. doi: 10.1093/nar/gku1196

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., et al. (2015). CFSAN SNP pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *Peer J. Comput. Sci.* 1:e20. doi: 10.7717/peerj-cs.20

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88. doi: 10.1371/journal.pbio.0040088

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192. doi: 10.1093/molbev/msi103

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kuck, P., and Longo, G. C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.* 11:81. doi: 10.1186/s12983-014-0081-x

Pettengill, J. B., Pightling, A. W., Baugher, J. D., Rand, H., and Strain, E. (2016). Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of *Salmonella* samples. *PLoS One* 11:e0166162. doi: 10.1371/journal.pone.0166162

Pightling, A. W., Pettengill, J. B., Luo, Y., Baugher, J. D., Rand, H., and Strain, E. (2018). Interpreting whole-genome sequence analyses of foodborne Bacteria for regulatory applications and outbreak investigations. *Front. Microbiol.* 9:1482. doi: 10.3389/fmicb.2018.01482

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian Phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly path-O-gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* 17, 57–86.

Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., et al. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol. Evol.* 5, 2109–2123. doi: 10.1093/gbe/evt159

Wang, Y. U., Pettengill, J. B., Pightling, A., Timme, R., Allard, M., Strain, E., et al. (2018). Genetic diversity of *Salmonella* and *Listeria* isolates from

food facilities. *J. Food Prot.* 81, 2082–2089. doi: 10.4315/0362-028X.JFP-18-093

Wattiau, P., Boland, C., and Bertrand, S. (2011). Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl. Environ. Microbiol.* 77, 7877–7885. doi: 10.1128/AEM.05527-11

Zhang, S., Yin, Y., Jones, M. B., Zhang, Z., Deatherage Kaiser, B. L., Dinsmore, B. A., et al. (2015). *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J. Clin. Microbiol.* 53, 1685–1692. doi: 10.1128/JCM.00323-15

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.