Plague Phylodynamics and Phylogeography

This manuscript (permalink) was automatically generated from ktmeaton/obsidian-public@bd0cba1c on May 18, 2021.

Authors

- **Katherine Eaton**
  0000-0001-6862-7756 · ktmeaton
  McMaster Ancient DNA Center; Department of Anthropology, McMaster University

- **Leo Featherstone**
  0000-0002-8878-1758
  The Peter Doherty Institute For Infection and Immunity , University of Melbourne

- **Sebastian Duchene**
  0000-0002-2863-0907 · sebastianduchene
  The Peter Doherty Institute For Infection and Immunity , University of Melbourne

- **Hendrik Poinar**
  0000-0002-0314-4160
  McMaster Ancient DNA Center; Department of Anthropology, McMaster University

Keywords

- Plague
- Yersinia pestis
- Phylodynamics
- Phylogeography

Introduction

Plague has an impressively long and expansive history as a human pathogen. The earliest evidence of the plague bacterium Yersinia pestis comes from ancient DNA studies dating its emergence to at least the Neolithic [1,2]. Since then, Y. pestis has traveled extensively due to ever-expanding global trade networks [3] and the ability to infect a diverse array of mammalian hosts [4]. Few regions of the ancient and modern world remain untouched by this disease, as plague has an established presence on every continent except Oceania [5].

Accompanying this prolific global presence is unnervingly high mortality. The infamous medieval Black Death is estimated to have killed more than half of Europe's population [**???**]. This virulence can still be observed in the post-antibiotic era, where case fatality rates range from 22-71% [6]. As a result, plague maintains its status as a disease that is of vital importance to current public health initiatives.

The intriguingly high mortality that is repeatedly seen throughout history brings together diverse researchers with interests spanning the modern period, history, and even prehistory. This intersection has brought about novel insight to render what was once invisible, visible. For example, investigating the ecology of ancient rats [7] and reconstructing the genome of Black Death-era Y. pestis [**???**]. However, this breadth of research also reflects the observation that plague has traveled through immensely diverse populations, cultures, and landscapes. Thus it is unsurprising that any consensus on 'universal' disease dynamics or experiences are rare to uncover. For example, within China alone there are 11 natural plague foci, each characterized by distinct environmental factors, bacteriological properties, and host-vector interactions [8]. As a result, significant debate has emerged on topics such as the severity of past pandemics [9], their geographic origins [**???**], and the mechanisms of spread [**???**].

TO BE DONE:

- Introduce the genomic composition of Y. pestis and mechanism of evolution.
- Introduce the topics phylodynamics and phylogeography and what is known so far.
- Introduce the problem(s) and our objective(s).

Materials and Methods

Data Collection

Y. pestis genome sequencing projects were retrieved from the NCBI databases using NCBImeta [10]. 1657 projects were identified and comprised three genomic types: 586 modern assembled, 184 ancient unassembled, and 887 modern unassembled genomes. The 887 modern unassembled genomes were excluded from this project, as the wide variety of laboratory methods and sequencing strategies precluded a standardized workflow. Future work will investigate computationally efficient methods for integrating this data.

Collection location, collection date, and collection host metadata were curated by cross-referencing the original publications. Collection location was transformed to latitude and longitude coordinates using GeoPy and the Nominatim API for OpenStreetMap [11,12,13]. Coordinates were standardized at a sub-country resolution, taking the centroid of the parent province/state. Collection dates were standardized according to their year, and recording uncertainty arising from missing data and radiocarbon estimates. Collection host was the most diverse field with regards to precision, ranging from colloquial nomenclature (*"rat"*) to a genus species taxonomy (*"Meriones libycus"*). For the purposes of this study, collection host was recorded as *Human*, *Non-Human*, or *Not Available*, given the inability to differentiate non-human mammalian hosts.

Genomes were removed if no associated date or location information could be identified in the literature, or if there was documented evidence of laboratory manipulation. After curation, 600 genomes remained, with 539 (90%) being modern in origin and 61 (10%) being ancient.

Two additional datasets were required for downstream analyses. First, *Y. pestis* strain CO92 (GCA_000009065.1) was used as the reference genome for sequence alignment and variant annotation. Second, *Yersinia pseudotuberculosis* strains NCTC10275 (GCA_900637475.1) and IP32953 (GCA_000834295.1) served as an outgroup to root the maximum likelihood phylogeny.

Sequence Quality Criteria

Alignment

Ancient unassembled genomes were downloaded from the SRA databases in FASTQ format using the SRA Toolkit [14]. Pre-processing and alignment to the reference genome was performed using the nf-core/eager pipeline, a reproducible workflow for ancient genome reconstruction [15]. Ancient genomes were removed if the number of sites covered at a minimum depth of 3X was less than 70% of the reference genome.

Modern assembled genomes were aligned to the reference genome using Snippy, a pipeline for core genome alignments [16]. Modern genomes were removed if the number of sites covered at a minimum depth of 10X was less than 70% of the reference genome.

A multiple sequence alignment was constructed using the Snippy Core module of the Snippy pipeline. The output alignment was filtered to only include chromosomal variants and to exclude sites that had more than 5% missing data.

Phylogenetic Reconstruction

Model selection was performed using Modelfinder which identified the K3Pu+F+I model as the optimal choice based on the Bayesian Information Criterion (BIC) [17]. A maximum-likelihood phylogeny was then estimated across 10 independent runs of IQTREE [18]. Branch support was evaluated using 1000 iterations of the ultrafast bootstrap approximation, with a threshold of 95% required for strong support [**???** Improving Ultrafast].

Subsampled Datasets

To improve the performance of Bayesian analysis, three datasets were constructed.

Name	Total Size	Composition	Purpose
Full	N=600	Single alignment of all genomes.	Alignment, maximum-likelihood phylogeny.
Clade	N=600	Separated by clade	Dating and phylogeography
Reduced	N=200	Subsampled	Demonstrate dating instability

Phylodynamics

Phylogeography

Results

Composition

Phylogeny

Divergence-scaled phylogeny of Y. pestis (Figure 1).

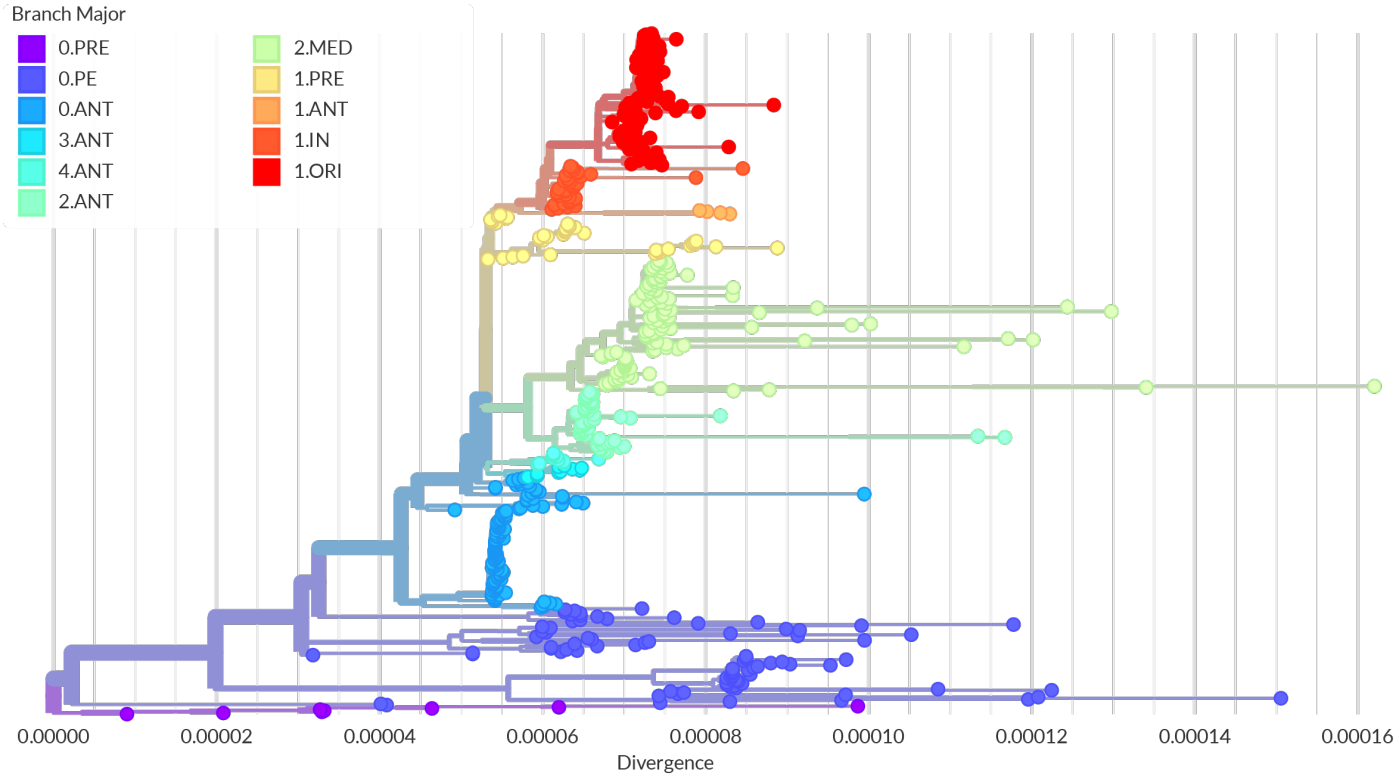

Figure 1: *Yersinia pestis* phylogeny. (Significant SVG editing required)|800

Phylodynamics

Molecular Clock

- *Y. pestis* has extreme rate variation.
- A Root to Tip Regression on collection date confirms this, as the Coefficient of Determination (R^2) is 0.09, revealing a poor fit to a simple linear model (Table 1).
- To some extent, this variation can be explained by examining the clades in isolation (Figure 2).
- Finding an appropriate evolutionary model is key to estimating historic events, like clade emergence (Figure 3).

Table 1: Temporal signal statistics by clade

Branch	Clade	Origin	R^2	p-value
all	all	Ancient, Modern	0.09	3.81E-14
0	0.PRE	Ancient	0.91	1.53E-04*
0	0.PE	Modern	0.01	2.25E-01
0	0.ANT4	Ancient	0.66	7.84E-04*
0	0.ANT	Modern	-0.01	7.35E-01
1	1.ANT	Modern	0.45	2.03E-01
1	1.IN	Modern	0.0	3.24E-01
1	1.ORI	Modern	0.04	1.32E-02*
1	1.PRE	Ancient	0.76	1.68E-13*
2	2.ANT	Modern	0.05	5.96E-02
2	2.MED	Modern	0.01	1.86E-01
3	3.ANT	Modern	-0.04	4.39E-01
4	4.ANT	Modern	-0.11	8.80E-01

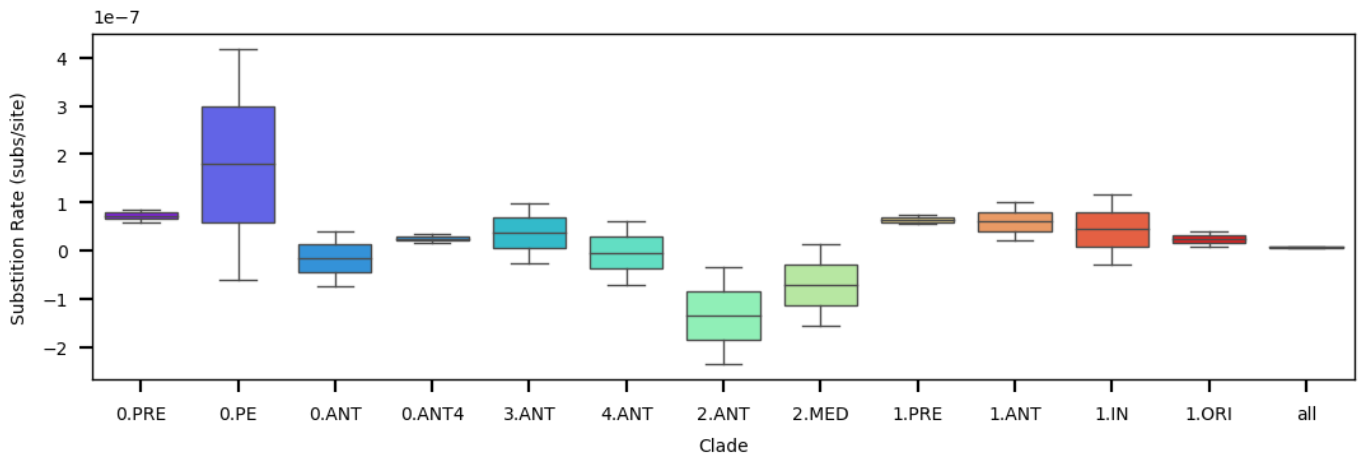

Figure 2: Rate variation by clade.

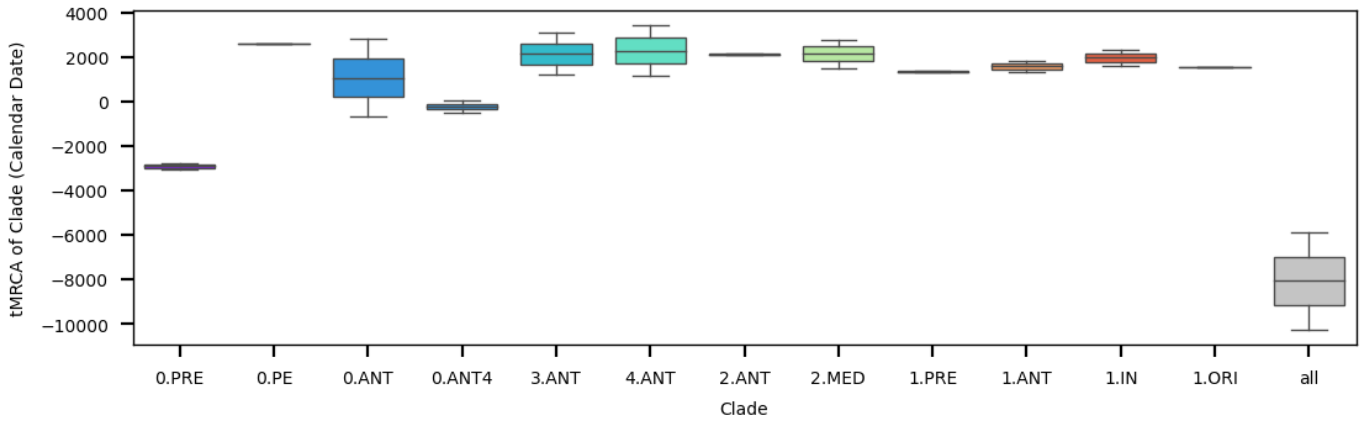

Figure 3: tMRCA by clade.

Relaxing the Clock

- Relaxed clock MCMC runs produce a high Coefficient of Variation indicating a relaxed model is favored over a strict model (Figure 4). However, these runs do not converge, suggesting there is too much rate variation to confidently estimate key parameters such as the mean Substitution Rate or tMRCA.

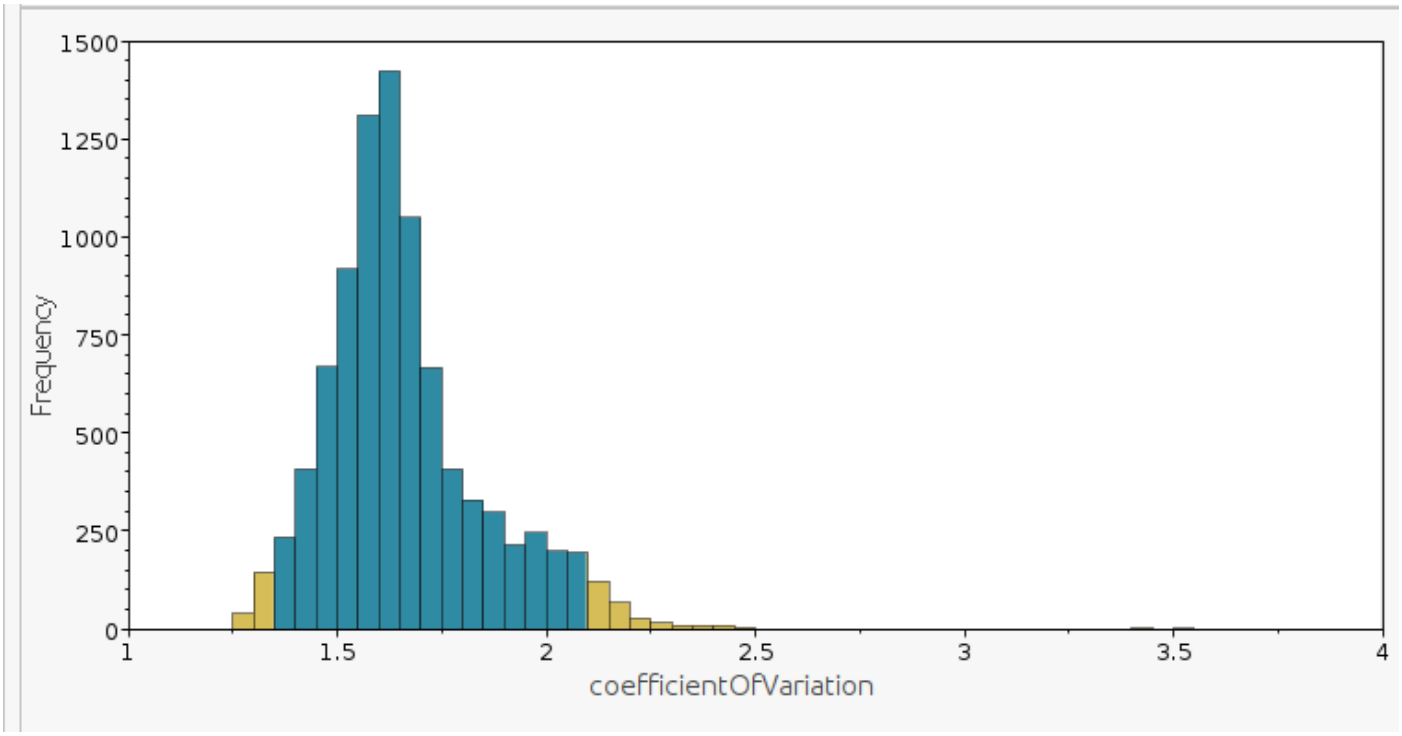

Figure 4: Coefficient of variation.

- A strict clock and relaxed clock have overlapping distributions with similar peaks for the Tree Height (blue: strict, green: relaxed) (Figure 5).

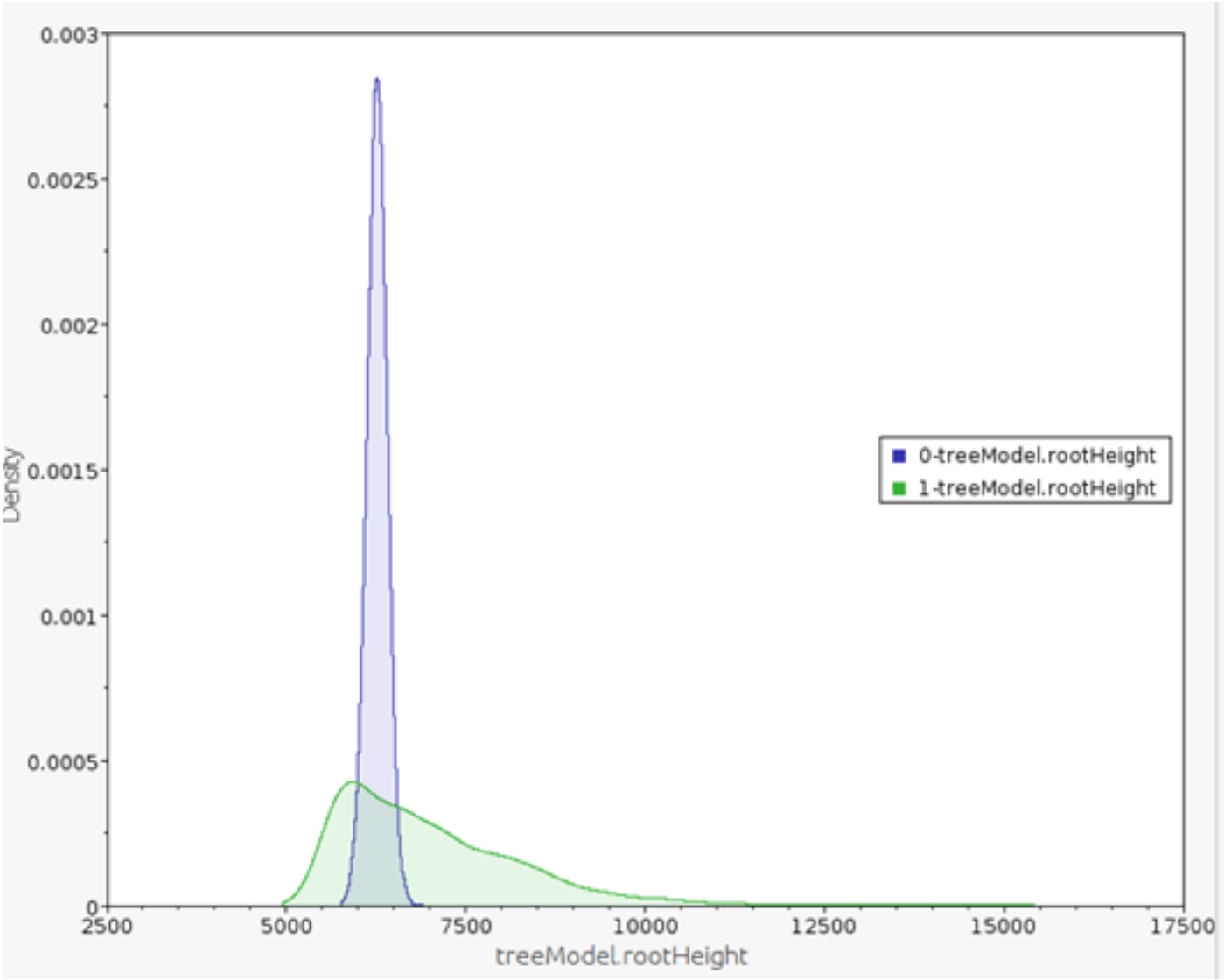

Figure 5: Tree height comparison.

- When estimating a Substitution Rate for all of *Y. pestis*, a [[Clock Model|strict clock]] and relaxed clock produce different estimates (green: strict, orange: relaxed) (Figure 6).

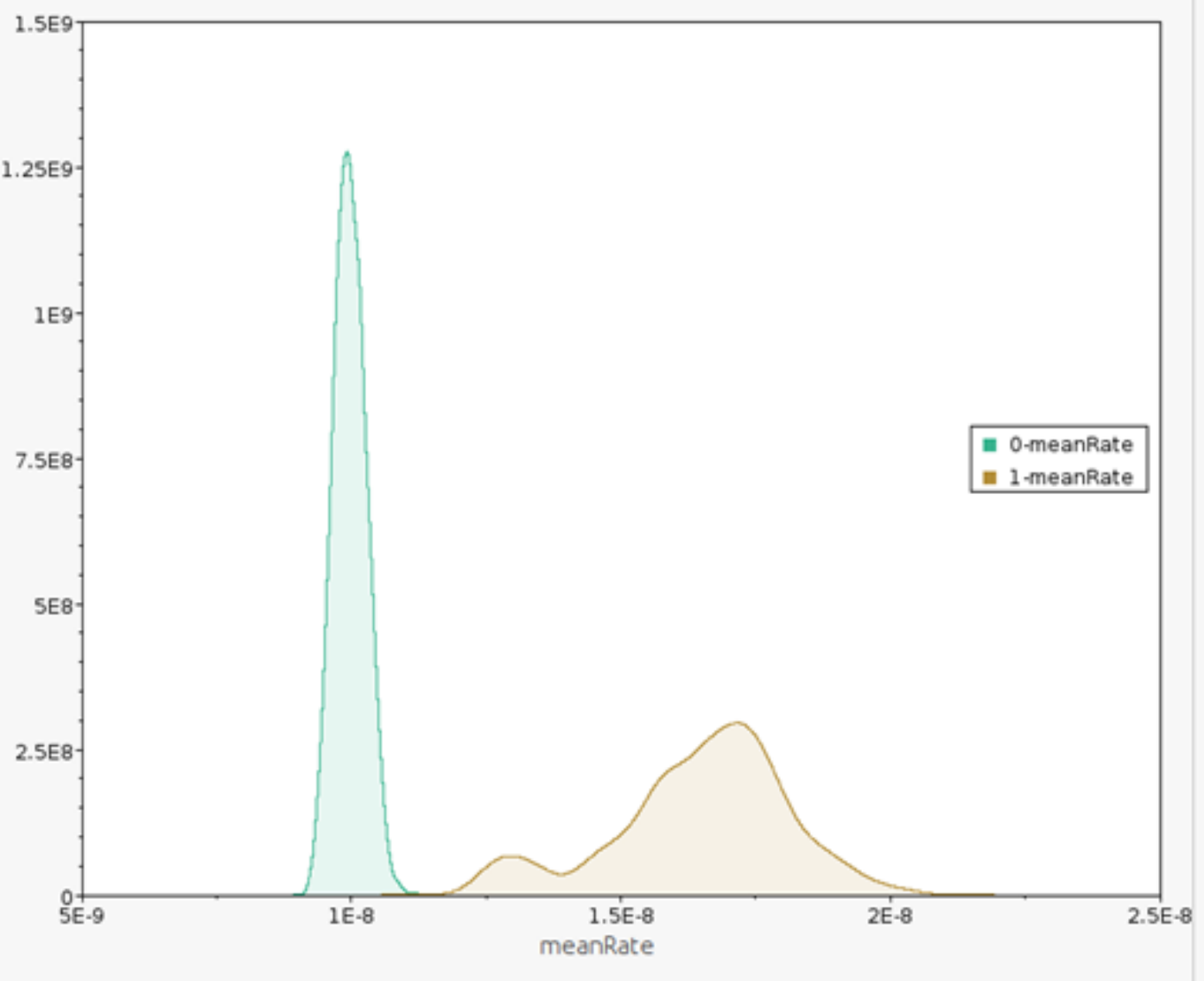

Figure 6: Substitution rate comparison.

- There doesn't appear to be clustering of rates. Branches with high rates are next to those with low rates (Figure 7).

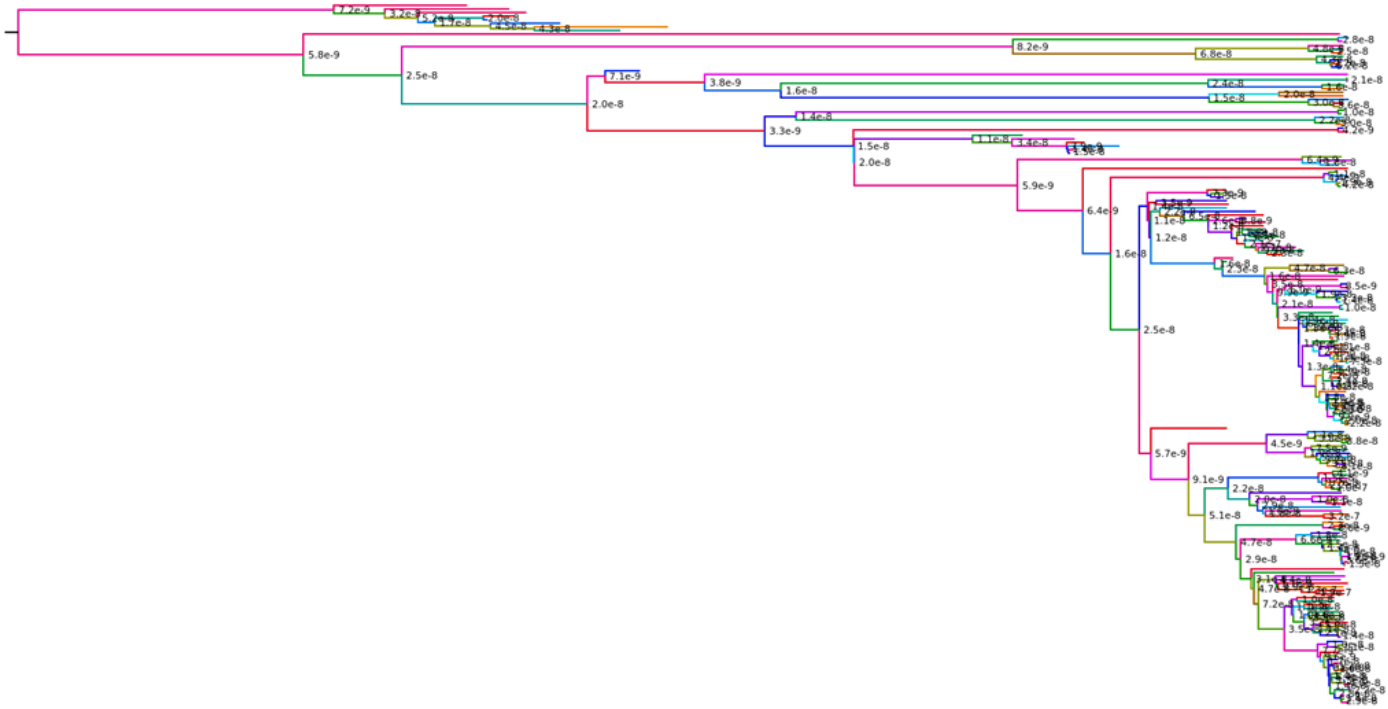

Figure 7: Time tree colored by rate.

Discussion

Conclusion

References

1. **The Stone Age Plague and Its Persistence in Eurasia**
   Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, Michal Feldman, Rimantas Jankauskas, Ivor Janković, … Johannes Krause
   Current Biology (2017-12-04)
   DOI: 10.1016/j.cub.2017.10.025 · PMID: 29174893

2. **Emergence and spread of basal lineages of *Yersinia pestis* during the Neolithic Decline**
   Nicolás Rascovan, Karl-Göran Sjögren, Kristian Kristiansen, Rasmus Nielsen, Eske Willerslev, Christelle Desnues, Simon Rasmussen
   Cell (2019-01-10) https://www.cell.com/cell/abstract/S0092-8674(18)31464-8
   DOI: 10.1016/j.cell.2018.11.005 · PMID: 30528431

3. **Trade routes and plague transmission in pre-industrial Europe**
   Ricci P. H. Yue, Harry F. Lee, Connor Y. H. Wu
   Scientific Reports (2017-10-11) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636801/
   DOI: 10.1038/s41598-017-13481-2 · PMID: 29021541 · PMCID: PMC5636801

4. **Yersinia pestis–etiologic agent of plague**
   R. D. Perry, J. D. Fetherston
   Clinical Microbiology Reviews (1997-01)
   PMID: 8993858 · PMCID: PMC172914

5. **Plague**
   World Health Organization
   (2017-10-31) https://www.who.int/news-room/fact-sheets/detail/plague

6. **Plague around the world in 2019**
   Eric Bertherat
   Weekly Epidemiological Record (2019-06-21) https://apps.who.int/iris/bitstream/handle/10665/325481/WER9425-en-fr.pdf

7. **Rats, Communications, and Plague: Toward an Ecological History**
   Michael McCormick
   The Journal of Interdisciplinary History (2003-07-01) https://doi.org/10.1162/002219503322645439
   DOI: 10.1162/002219503322645439

8. **Comparative and evolutionary genomics of *Yersinia pestis***
   Dongsheng Zhou, Yanping Han, Yajun Song, Peitang Huang, Ruifu Yang
   Microbes and Infection (2004-11-01) http://www.sciencedirect.com/science/article/pii/S1286457904002357
   DOI: 10.1016/j.micinf.2004.08.002

9. **The Justinianic Plague: An inconsequential pandemic?**
   Lee Mordechai, Merle Eisenberg, Timothy P. Newfield, Adam Izdebski, Janet E. Kay, Hendrik Poinar
   Proceedings of the National Academy of Sciences (2019-12-17) http://www.pnas.org/content/116/51/25546
   DOI: 10.1073/pnas.1903797116 · PMID: 31792176

10. **NCBImeta**
Katherine Eaton
NCBImeta (2019) https://github.com/ktmeaton/NCBImeta

11. **GeoPy: A Python client for several popular geocoding web services.**
Kostya Esmukov
(2020-12) https://github.com/geopy/geopy

12. **Nominatim: A tool to search OpenStreetMap data.**
Sarah Hoffman
(2020-12) https://github.com/osm-search/Nominatim

13. **Planet dump retrieved from https://planet.osm.org**
OpenStreetMap Contributors
(2017) https://www.openstreetmap.org

14. **ncbi/sra-tools**
NCBI - National Center for Biotechnology Information/NLM/NIH
(2021-05-18) https://github.com/ncbi/sra-tools

15. **Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager**
James A. Fellows Yates, Thiseas C. Lamnidis, Maxime Borry, Aida Andrades Valtueña, Zandra Fagernäs, Stephen Clayton, Maxime U. Garcia, Judith Neukamm, Alexander Peltzer
PeerJ (2021-03-16) https://peerj.com/articles/10947
DOI: 10.7717/peerj.10947

16. **Snippy: Rapid haploid variant calling and core genome alignment.**
Torsten Seemann
(2020-03-08) https://github.com/tseemann/snippy

17. **ModelFinder: fast model selection for accurate phylogenetic estimates**
Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, Lars S. Jermiin
Nature Methods (2017-06) http://www.nature.com/articles/nmeth.4285
DOI: 10.1038/nmeth.4285

18. **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era**
Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, Robert Lanfear
Molecular Biology and Evolution (2020-05-01) https://academic.oup.com/mbe/article/37/5/1530/5721363
DOI: 10.1093/molbev/msaa015