Plague Phylodynamics and Phylogeography

This manuscript (permalink) was automatically generated from ktmeaton/obsidian-public@5b5b06dd on May 19, 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 and the ability to infect a diverse array of mammalian hosts [3,4]. 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 [6]. This virulence can still be observed in the post-antibiotic era, where case fatality rates range from 22-71% [7]. As a result, plague maintains its status as a disease that is of vital importance to current public health initiatives.

This high priority disease status is unsurprising given that *Y. pestis* is a member of the Enterobacteriaceae family. This family includes enteric pathogens such as *Escherichia coli* and *Salmonella typhi* that are commonly transmitted by contaminated food and water. In comparison, the plague bacterium is unique among this family due to a striking difference in host habitat and transmission. *Y. pestis* commonly resides in the blood of its mammalian hosts and can be transmitted to new hosts through an infectious fleabite [8]. In addition to these tissues, the [[Yersinia pestis|plague bacterium]] is also capable of colonizing parts of the mammalian immune system including the lymphatic and reticuloendothelial systems. The large diversity of media in which *Y. pestis* has adapted to colonize is particularly surprising given that it only recently (within the last 20,000 years) diverged as a clone of its parent species Yersinia pseudotuberculosis [9].

Despite a close genetic similarity between *Y. pestis* and *Y. pseudotuberculosis*, in which they share 97% gene identity, they differ widely in their transmission and pathogenecity [10]. Whereas *Y. pseudotuberculosis* causes gastrointestinal disease and is transmitted by the food-borne route, *Y. pestis* is primarily transmitted between mammalian hosts by fleas and causes septicemia, pneumonia, and lymphadenitis. Because of this apparent contradiction of genetic homogeneity and diverse phenotypes, an extensive body of research has formed to address how, when, and where, these epidemiological shifts occurred.

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).

Objectives

1. Synthesize and curate publicly available Y. pestis genomes.
2. To propose a nuanced phylodynamics model.
3. To critique interpretations drawn from phylogeographic approaches?

Materials and Methods

Data Collection

Y. pestis genome sequencing projects were retrieved from the NCBI databases using NCBImeta [11]. 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 [12,13,14]. 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.

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 database in FASTQ format using the SRA Toolkit [15]. Pre-processing and alignment to the reference genome was performed using the nf-core/eager pipeline, a reproducible workflow for ancient genome reconstruction [16]. 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 [17]. 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) [18]. A maximum-likelihood phylogeny was then estimated across 10 independent runs of IQTREE [19]. Branch support was evaluated using 1000 iterations of the ultrafast bootstrap approximation, with a threshold of 95% required for strong support [20].

Modified Datasets

To investigate the influence of between-clade variation in substitution rates, the multiple alignment was separated into the major clades of *Y. pestis*, which will be referred to as the *Clade* dataset. The subclade associated with the Plague of Justinian (0.ANT4) was considered to be a distinct clade separate from its parent (0.ANT) due to its geographic, temporal, and ecological uniqueness. In total, 12 clades were considered and are described in Table 1.

To improve the performance and convergence of Bayesian analysis, a subsampled dataset was constructed. Clades that contained multiple samples drawn from the same geographic location and the same time period were reduced to one representative sample. The sample with the shortest terminal branch length was prioritized, to diminish the influence of derived mutations on the estimated substitution rate. An interval of 25 years was identified as striking an optimal balance, resulting in 200 representative samples.

Phylodynamics

To investigate the degree of temporal signal present in the data, two tests were formed . The first was a root-to-tip (RTT) regression on collection date. This linear model is a simple approach to explore whether the data follows a strict clock model. Uncertainty in the model parameters, namely the mean substitution rate and tMRCA, were estimated using 1000 iterations of the non-parametric bootstrap on the residuals.

While RTT is a practical approach, it has two main limitations: 1) No rate variation is accounted for, and 2) The data are not independent observations due to shared internal branch lengths. Therefore to complement this approach, a bayesian evaluation of temporal signal (BETS) was performed.

A maximum-likelihood timetree was estimated using a least-squares approach as implemented in LSD2 [21]. Rate variation was modeled using a lognormal relaxed clock using the default parameters for the mean (1.0) and the standard deviation (0.2). The outgroup *Y. pseudotuberculosis* was used to root the tree and then subsequently removed.

A bayesian timetree was estimated using ... as implemented in BEAST.

Phylogeography

Geographic location was modeled as a discrete state with transitions following a GTR mugration model as implemented in TreeTime [22].

Results

Curated Public Dataset

After curation, 600 genomes remained, with 539 (90%) being modern in origin and 61 (10%) being ancient. The geographic distribution of samples is shown in Figure 1 .

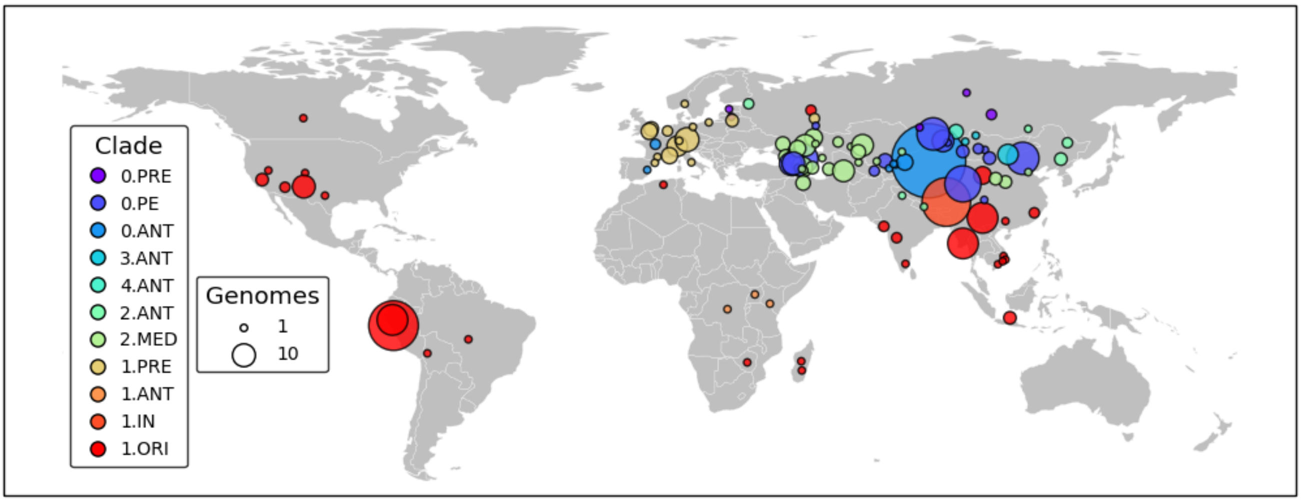

Figure 1: Geographic distribution of *Yersinia pestis* genomes.

Phylogeny

Divergence-scaled phylogeny of Y. pestis (Figure 2).

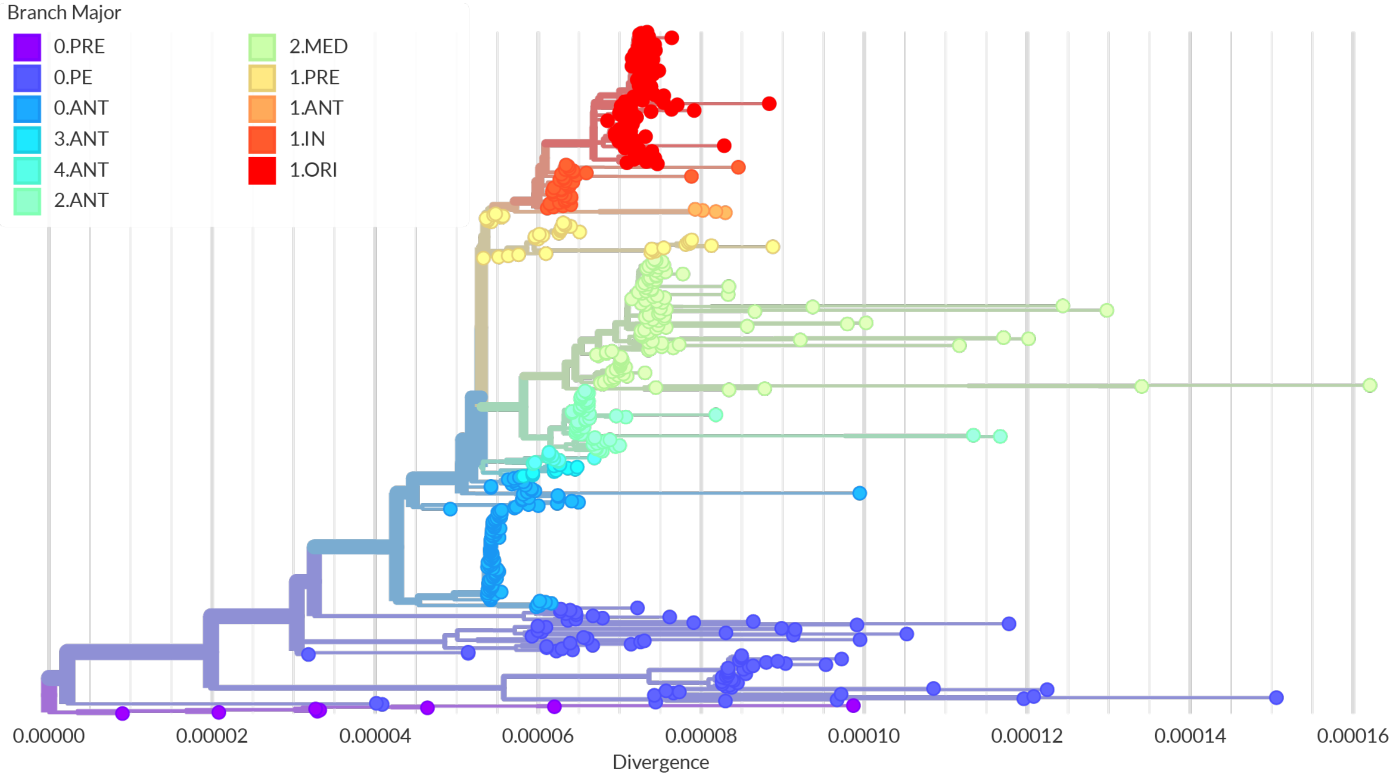

Figure 2: *Yersinia pestis* phylogeny. (Significant SVG editing required)|800

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 3).
- Finding an appropriate evolutionary model is key to estimating historic events, like clade emergence (Figure 4).

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

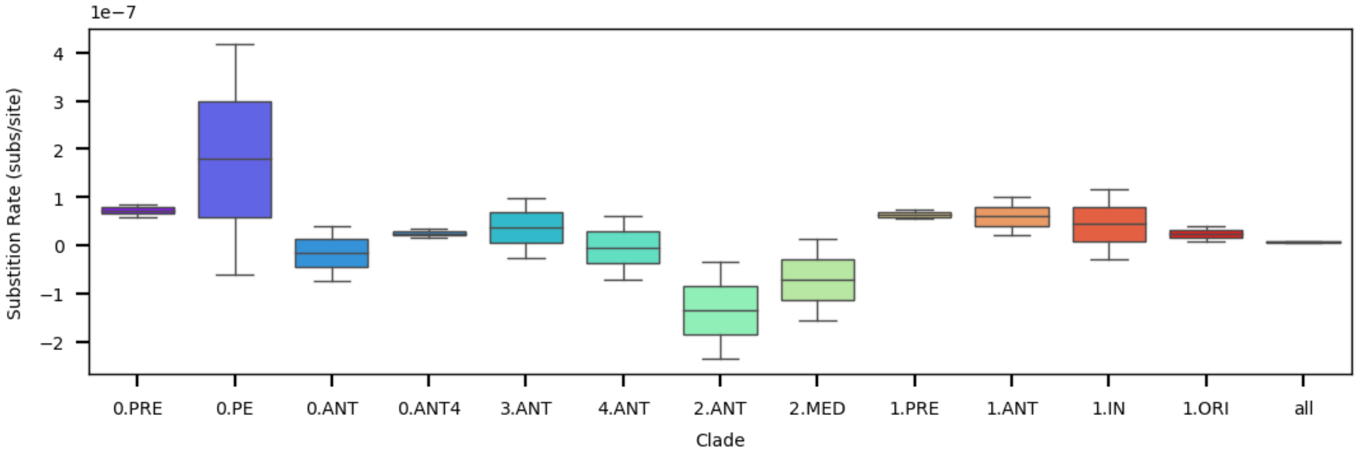

Figure 3: Rate variation by clade.

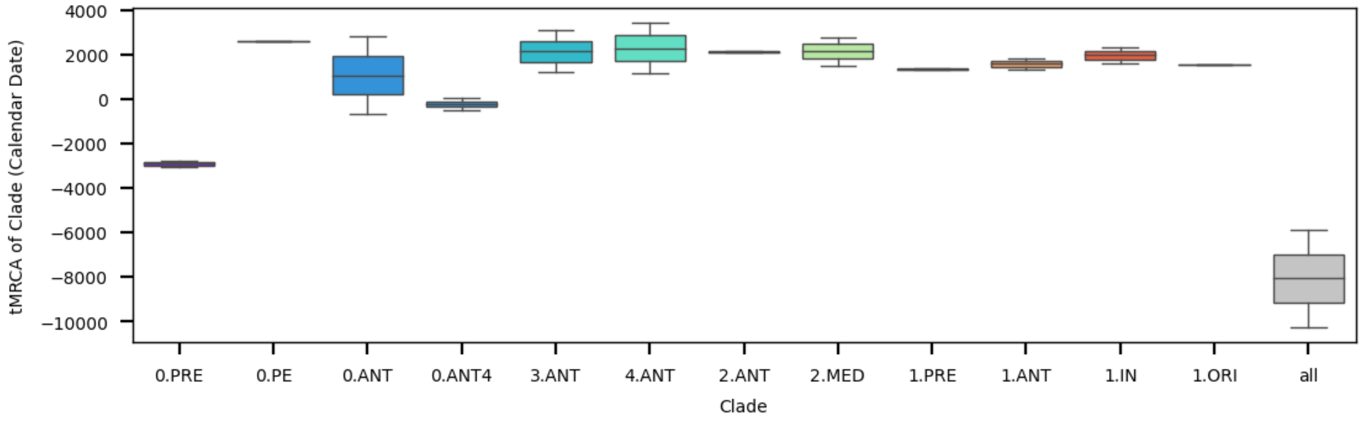

Figure 4: tMRCA by clade.

Relaxing the Clock

- Relaxed clock MCMC runs produce a high Coefficient of Variation indicating a relaxed model is favored over a strict model (Figure 5). 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.

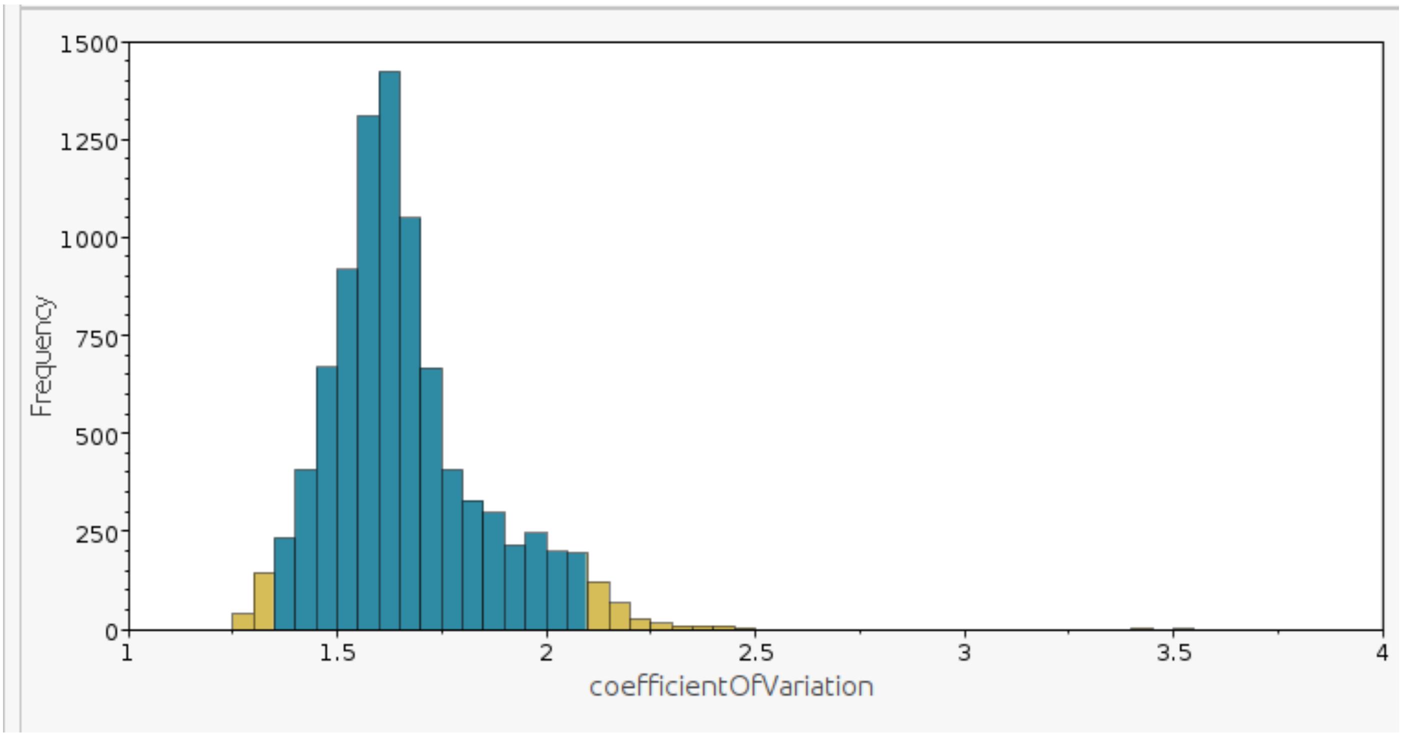

Figure 5: Coefficient of variation.

- A strict clock and relaxed clock have overlapping distributions with similar peaks for the Tree Height (blue: strict, green: relaxed) (Figure 6).

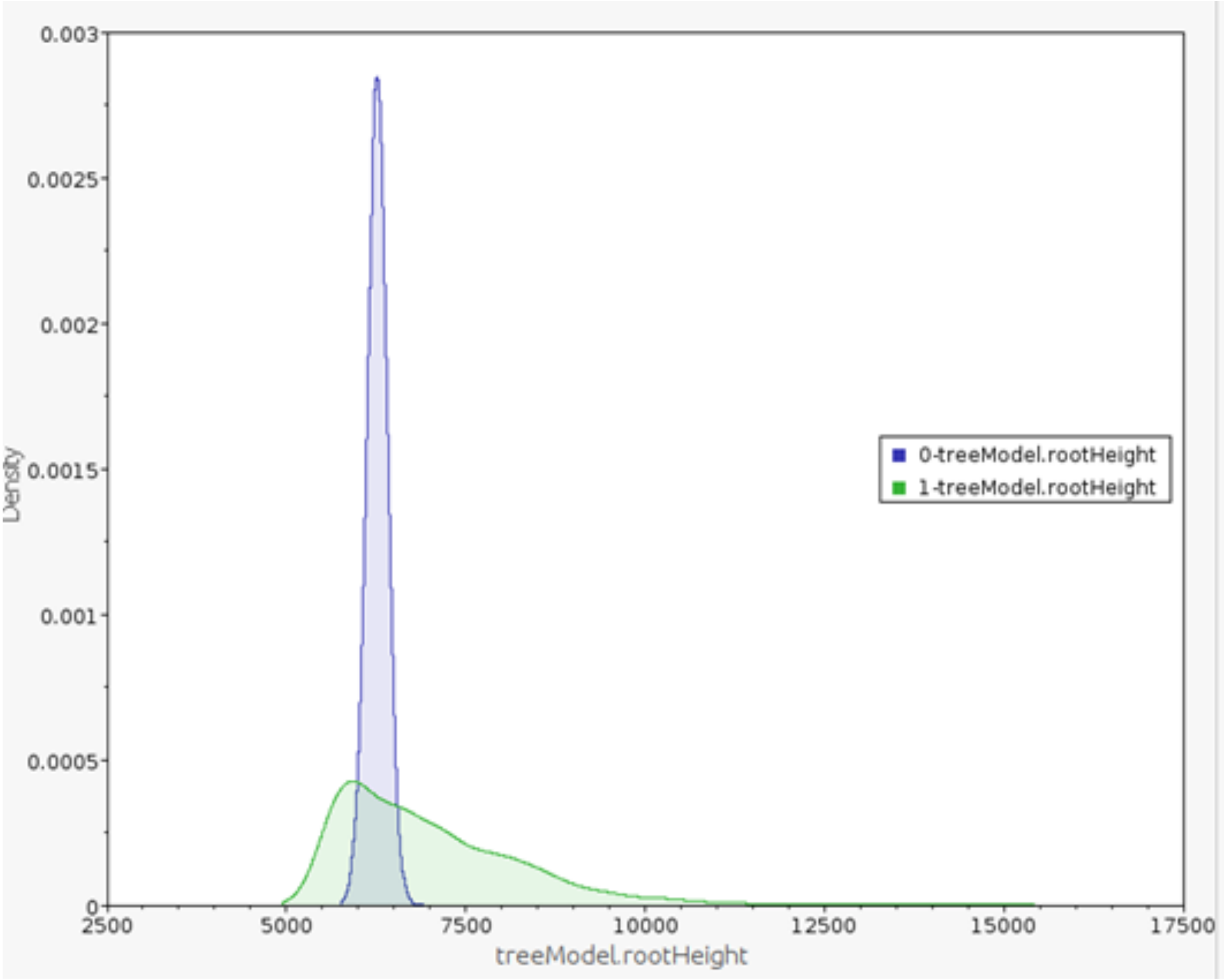

Figure 6: Tree height comparison.

- 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 7).

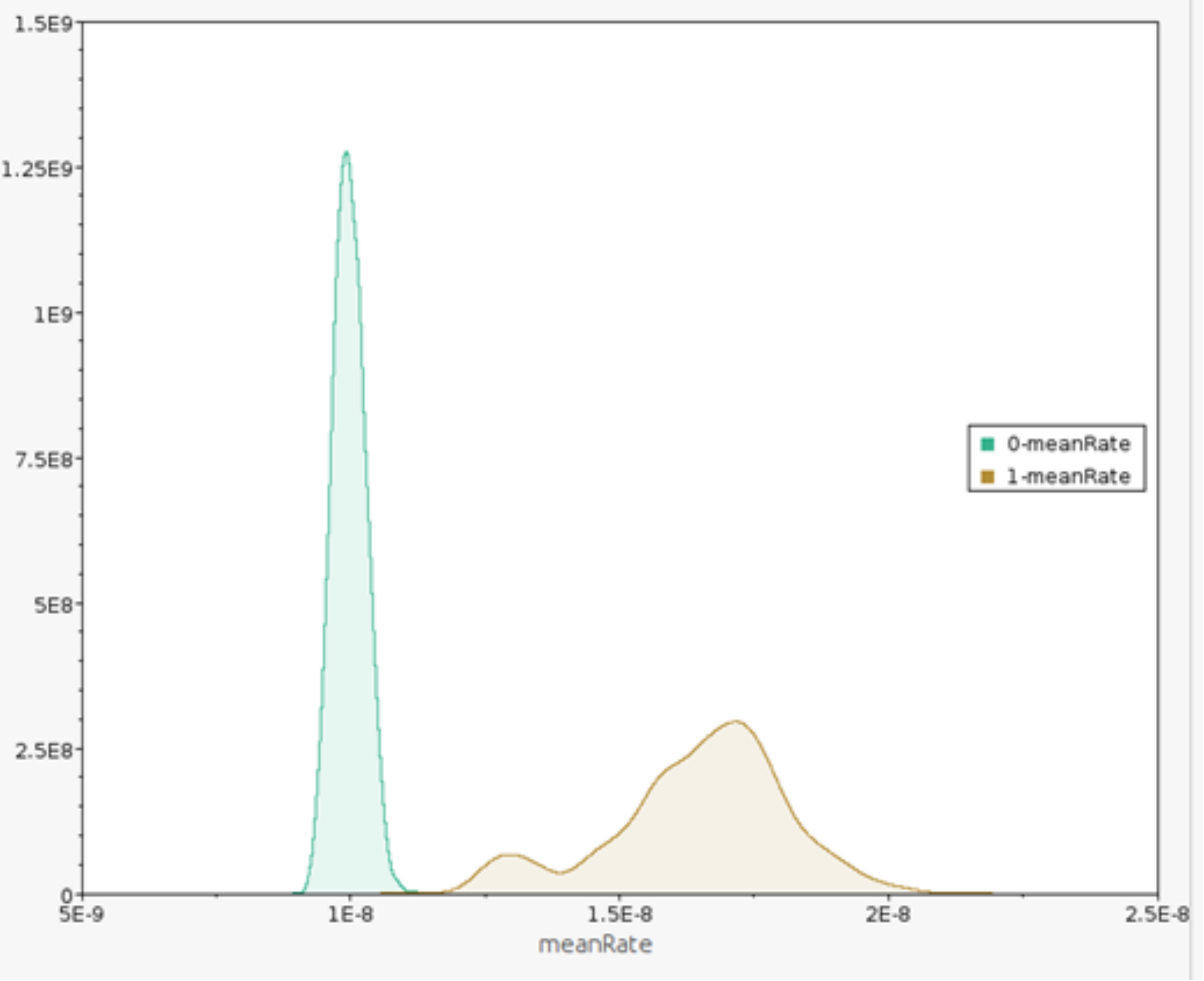

Figure 7: Substitution rate comparison.

- There doesn’t appear to be clustering of rates. Branches with high rates are next to those with low rates (Figure 8).

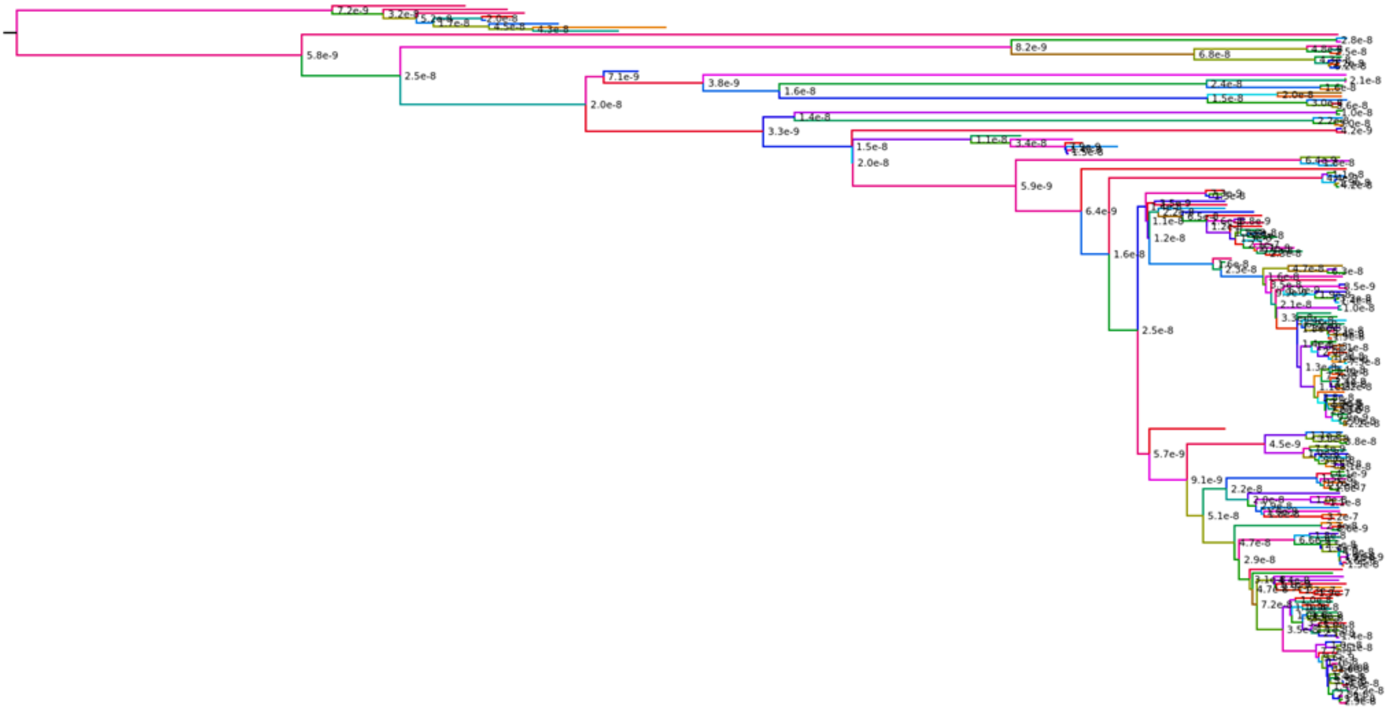

Figure 8: Time tree colored by rate.

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. **The Black Death, 1346-1353: The Complete History**
   O. J. Benedictow
   Boydell Press (2004)
   ISBN: 0-85115-943-5

7. **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

8. **Recent trends in plague ecology**
   K Gage, M Kosoy
   (2006) http://reviverestore.org/wp-content/uploads/2015/02/Gage-and-Kosoy_USGS-Blk-footed-ferret-symp_2006-copy.pdf

9. **Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis**
   M. Achtman, K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, E. Carniel
   Proceedings of the National Academy of Sciences of the United States of America (1999-11-23)
   DOI: 10.1073/pnas.96.24.14043 · PMID: 10570195 · PMCID: PMC24187

10. **Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis**
P. S. G. Chain, E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, … E. Garcia
Proceedings of the National Academy of Sciences (2004-09-21) http://www.pnas.org/cgi/doi/10.1073/pnas.0404012101
DOI: 10.1073/pnas.0404012101

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

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

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

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

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

16. **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

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

18. **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

19. **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

20. **UFBoot2: Improving the Ultrafast Bootstrap Approximation**
Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, Le Sy Vinh
Molecular Biology and Evolution (2018-02-01) https://academic.oup.com/mbe/article/35/2/518/4565479
DOI: 10.1093/molbev/msx281

21. **Fast Dating Using Least-Squares Criteria and Algorithms**
Thu-Hien To, Matthieu Jung, Samantha Lycett, Olivier Gascuel
Systematic Biology (2016-01) https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syv068
DOI: 10.1093/sysbio/syv068

22. **TreeTime: Maximum-likelihood phylodynamic analysis**
Pavel Sagulenko, Vadim Puller, Richard A Neher
Virus Evolution (2018-01-08) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5758920/
DOI: 10.1093/ve/vex042 · PMID: 29340210 · PMCID: PMC5758920