Plague Phylodynamics and Phylogeography

This manuscript (permalink) was automatically generated from ktmeaton/obsidian-public@6b2b8ae1 on May 29, 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

Abstract

- *Y. pestis* exhibits greater temporal signal than previously thought.
- Clades associated with pandemics exhibit strict clock-like behavior.
- Sampling bias significantly impacts phylogeography reconstructions.

Introduction

Plague has an impressively long and expansive history as a human disease. 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 wide variety 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 other notable 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 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 monomorphic 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.

Two epidemiological transitions that have been extensively researched are the time to Most Recent Common Ancestor (tMRCA) and the advent of historically documented plague pandemics. Substantial progress on these topics has been made as *Y. pestis* is the most intensively sequenced ancient pathogen, and over 100 ancient genomes are available to serve as fossil calibrations. Furthermore, there are now over 1000 publicly-available modern *Y. pestis* genomes [11], offering the potential to model rate variations and dating uncertainty with greater nuance. However, it is unclear whether this additional data will prove useful as it is contentious whether *Y. pestis* demonstrates sufficient temporal signal to robustly estimate a clock model [12]. Different *Y. pestis* datasets have been shown to produce dramatically different patterns of temporal signal from weak support to a complete absence of temporal structure 13. It was thus posited that variations in temporal signal "*may be a property of individual data sets rather than a true species effect*.

While powerful in potential, this recent avalanche of data comes with new challenges inherent to Big Data, with curation of the geospatial metadata proving to be a substantial obstacle.

TO BE DONE:

- Introduce the debate on whether *Y. pestis* has temporal signal and why incongruent findings have emerged.
- Introduce the issue of sampling bias in phylogeography.

Objectives

1. Curate and critique publicly available Y. pestis genomes. Discuss how sampling bias drives our current understanding of plague.
2. To propose a nuanced phylodynamics model.
3. To critique interpretations drawn from phylogeographic approaches?

Materials and Methods

Workflow Overview

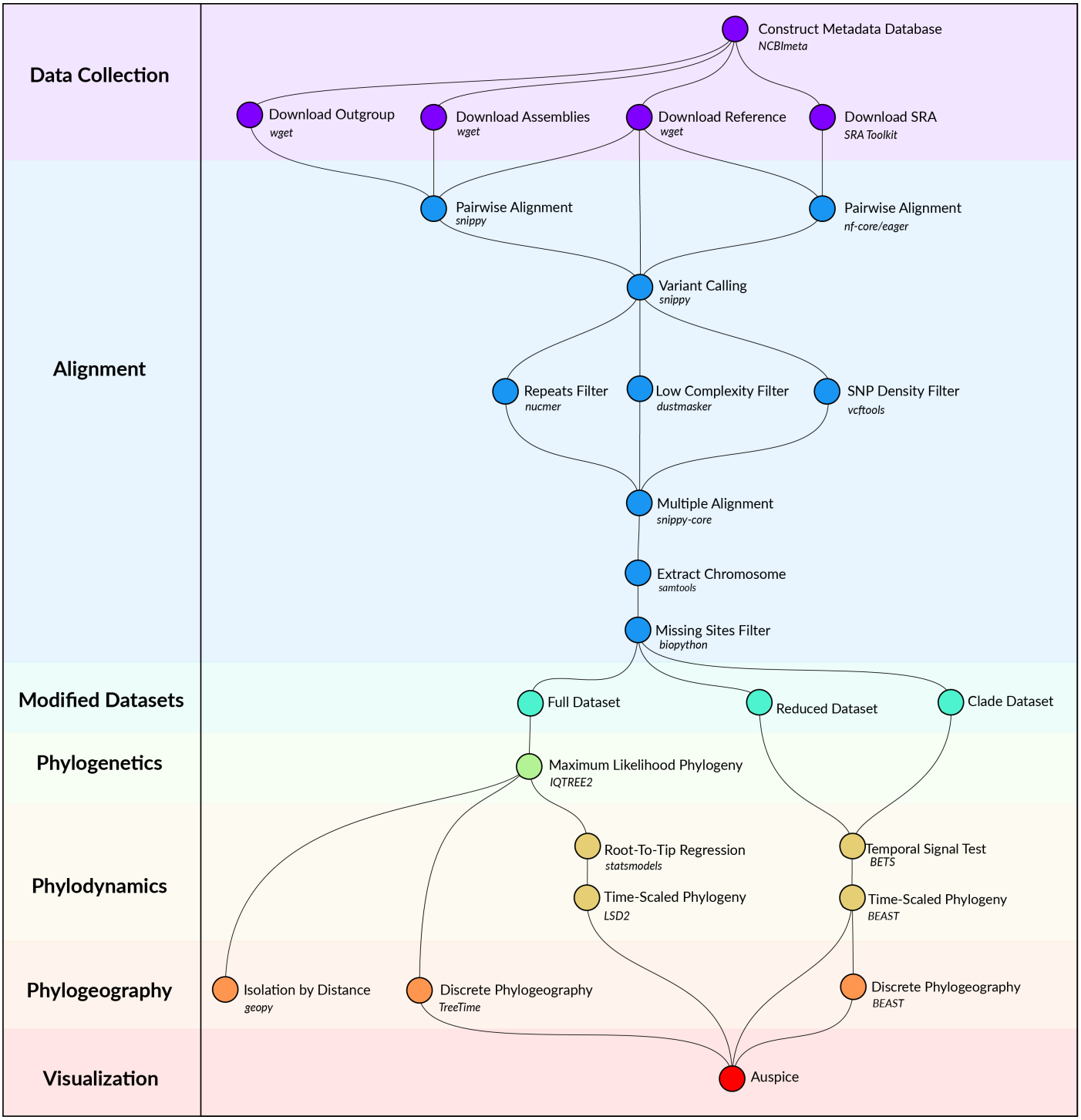

Figure 1: Computational methods workflow.

Data Collection

Y. pestis genome sequencing projects were retrieved from the NCBI databases using NCBImeta [14]. 1657 projects were identified and comprised three genomic types:

- 586 modern assembled
- 184 ancient unassembled
- 887 modern unassembled

The 887 modern unassembled genomes were excluded from this project, as the wide variety of laboratory methods and sequencing strategies precluded a standardized workflow. In contrast, the 184 ancient unassembled genomes were retained given the relatively standardized, albeit specialized, laboratory procedures required to process ancient tissues. Future work will investigate computationally efficient methods for integrating more diverse datasets.

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 [15,16,17]. 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 annotation. Second, *Yersinia pseudotuberculosis* strains NCTC10275 (GCA_900637475.1) and IP32953 (GCA_000834295.1) served as an outgroup to root the maximum likelihood phylogeny.

Alignment

Modern assembled genomes were aligned to the reference genome using Snippy, a pipeline for core genome alignments [18]. Modern genomes were removed if the number of sites covered at a minimum depth of 10X was less than 70% of the reference genome.

Ancient unassembled genomes were downloaded from the SRA database in FASTQ format using the SRA Toolkit [19]. Pre-processing and alignment to the reference genome was performed using the nf-core/eager pipeline, a reproducible workflow for ancient genome reconstruction [20]. Ancient genomes were removed if the number of sites covered at a minimum depth of 3X was less than 70% of the reference genome. It is a typical approach to relax coverage thresholds for ancient genomes relative to their modern counterparts (CITE). The threshold chosen here is commonly used, and aims to strike a balance between variant confidence and sample representation (CITE).

A multiple sequence alignment was constructed using the Snippy Core module of the Snippy pipeline [18]. The output alignment was filtered to only include chromosomal variants and to exclude sites that had more than 5% missing data.

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 partitioning of data by clade is a relatively new approach in Yersinia research and has been implemented used to... study epidemics of interest [21,22].

Clade and subclade labeling was derived from the five-branch population structure accompanied by a biovar abbreviation ([23]. Only one modification was made, in that the subclade associated with the Plague of Justinian (0.ANT4) was considered to be a distinct clade 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 and will be referred to as the *Reduced* dataset. 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 uniquely derived mutations on the estimated substitution rate. An interval of 25 years was identified as striking an optimal balance, resulting in 200 representative samples.

Phylogenetics

Model selection was performed using Modelfinder which identified the K3Pu+F+I model as the optimal choice based on the Bayesian Information Criterion (BIC) [24]. A maximum-likelihood phylogeny was then estimated across 10 independent runs of IQTREE [25]. Branch support was evaluated using 1000 iterations of the ultrafast bootstrap approximation, with a threshold of 95% required for strong support [26].

Phylodynamics

To explore the degree of temporal signal present in the data, two categories of tests were performed . 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 bayesian timetree was estimated using ... as implemented in BEAST.

A maximum-likelihood timetree was estimated using a least-squares approach as implemented in LSD2 [27]. Rate variation was modeled using a lognormal relaxed clock with 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.

> Note: I'm still pondering the best choice of parameters for the LSD2 relaxed clock.

Phylogeography

Geographic location was modeled as a discrete state with transitions following a GTR mugration model as implemented in TreeTime [28].

Results

Population Structure

A maximum-likelihood phylogeny was estimated from 603 genomes (600 *Y. pestis* isolates, 1 *Y. pestis* reference, and 2 *Y. pseudotuberculosis* outgroup taxa). A total of 26,282 Single Nucleotide Polymorphisms (SNPs) were identified with 17,333 sites present in at least two genomes and 16,370 sites present in only one. Following removal of the outgroup taxa, 10,249 sites remained with 3,844 sites shared by at least two genomes and 6,405 sites in only one.

The global population structure of *Y. pestis* is organized according to a vast array of historical, ecological, biochemical, and molecular characteristics (Figure 2). Arguably the most important event, in terms of phylogenetic structure, is the Big Bang Polytomy from which arose four monophyletic clades: `Branches 1-4`. All lineages that diverged prior to this multifurcation are grouped into `Branch 0`.

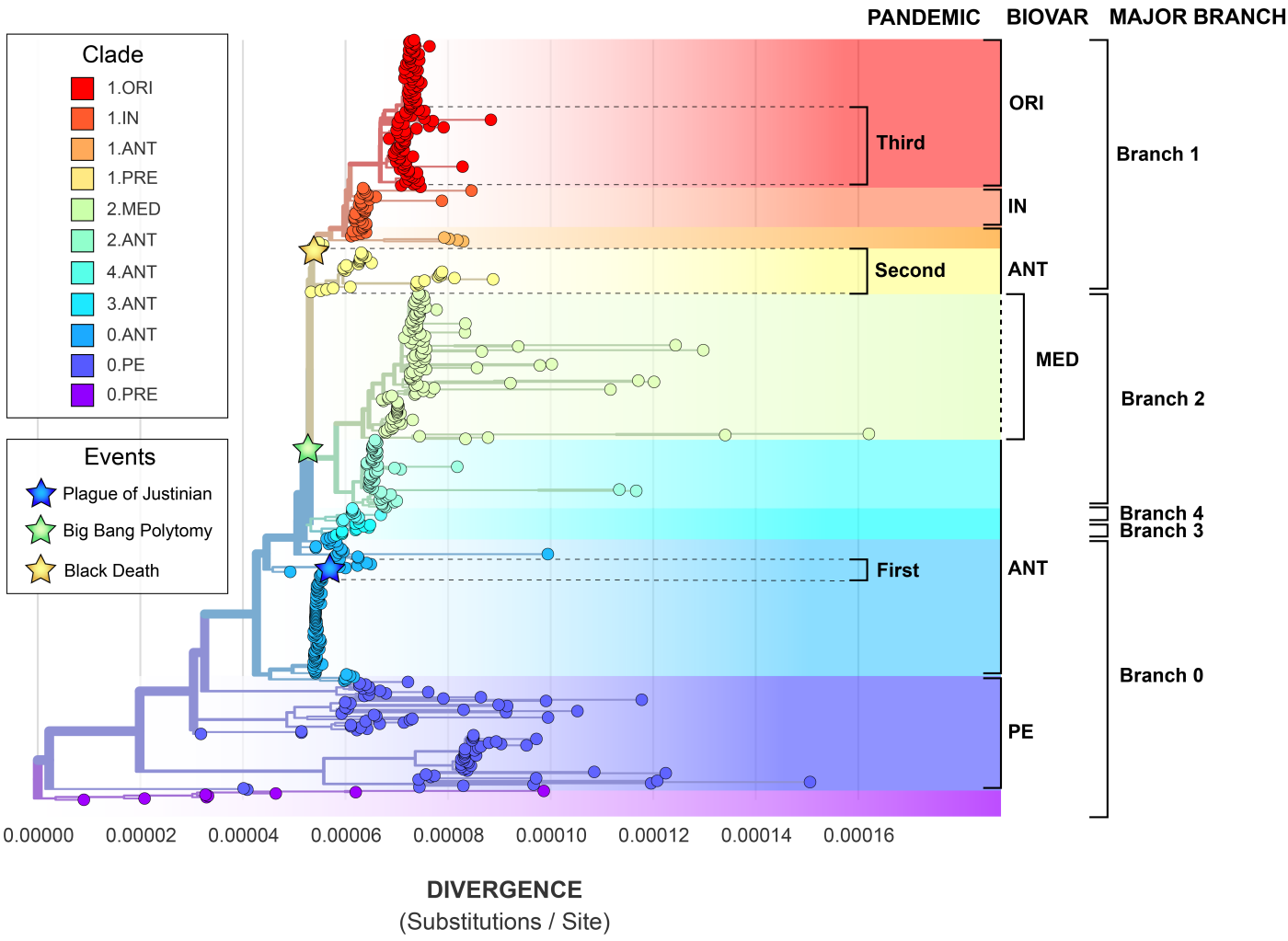

Figure 2: The maximum-likeilhood tree, constructed from 10,249 SNPs, depicts the global population structure of *Y. pestis*.

Each major branch is further subdivided into biovars according to metabolic properties [29,30]. The oldest isolates of plague date to the Late Neolithic Bronze Age following a divergence that pre-dates all known modern lineages [31]. In the absence of metabolic evidence, this clade is designated `0.PRE`. The *microtus* biovar, alias *pestoides* (`PE`), is also a basal clade found in `Branch 0` and while it is typically avirulent in humans, sporadic cases can occur (CITE).

The other inhabitant of `Branch 0` is biovar *antiqua* (`ANT`) which is the ancestral state prior to the Big Bang Polytomy and continues to be isolated from all major branches. `Branch 1` encapsulates a transition from *antiqua* (`ANT`) through the *intermedium* biovar (`IN`) and into *orientalis* (`ORI`). `Branch 2` includes the transition from *antiqua* (`ANT`) to *medievalis* (`MED`), which was once hypothesized to be associated with the Medieval Black Death, but is now known to be a distinct emergence. `Branch 3` and `Branch 4` are exclusively composed of *antiqua* (`ANT`) strains.

Pandemics of Plague

Additionally, several lineages of plague have been associated with historically documented plague pandemics. The First Pandemic (6th - 8th century CE) began with the Plague of Justinian and proceeded to devastate the Byzantine Empire of the Mediterranean world (CITE). A unique emergence of *Y. pestis* within the *antiqua* biovar of Branch 0 (`0.ANT4`) is thought to derive from this pandemic given spatiotemporal overlap of the skeletal remains from which this lineage was retrieved [12,32].

Similarly, variants of the the *antiqua* biovar of Branch 1 are thought to have given rise to the Second Pandemic. This well-documented pandemic began with the infamous Black Death and swept across most of Eurasia from the 14th to 19th centuries (CITE). The divergence of ancient *Y. pestis* dated to this time period pre-dates all other `Branch 1` lineages, with several samples placed directly at the base of `Branch 1`. To mark this unique phylogenetic positioning, this clade is designated `1.PRE` rather than a subclade of `1.ANT`.

The third documented pandemic of plague, alias the *Modern Pandemic*, spread globally from the end of the 18th Century and until the mid-20th Century. There is little dispute that a new lineage of plague emerging from Branch 1 as biovar *orientalis* (`1.ORI`) was the causative agent of this pandemic. While the World Health Organization (WHO) declared the third pandemic over in 1950 (CITE)), this lineage continues to re-emerge to cause recent epidemics such as the 2010 plague in Peru and the Madagascar Outbreaks of 2017.

The Three Pandemic Problem

While the pandemic clade nomenclature provides an excellent foundation for historical discussion, there are several problems with this system. First is the growing awareness of the spatiotemporal overlap of the Second and the Third Pandemic. Previously, the temporal extents of these events were mutually exclusive, dating from the 14th-18th century, and the 19th-20th century respectively. Recent historical scholarship has contested this claim, and demonstrated that these constraints are a product of a Eurocentric view of plague (CITE Nukhet). The Second Pandemic is now known to have extended into the 19th Century in parts of the Ottoman Empire, with the latest epidemics dating to 1819 (CITE). Similarly, the Third Pandemic is now hypothesized to have began as early as 1772 in southern China [33]. It remains unclear where to draw the distinction, if it even exists, between the Second and Third Pandemic.

Another limitation of the pandemic nomenclature is the complete disconnection of `Branch 2` to any pandemic-related events. This is surprising given that several criteria of a pandemic pathogen are fulfilled by `Branch 2` lineages, namely extensive spread and virulence. `Branch 2` genomes of *Y. pestis* have been collected from all throughout Eurasia, stretching from the Caucasus, to India, and to eastern China (Figure 3). Furthermore, clade 2.MED was demonstrated to have the highest spread velocity of any *Y. pestis* clade [34]. And finally, while Branch 2 isolates are not historically linked to the Third Pandemic Proper, clade 2.MED has been implicated in numerous modern plague outbreaks.

As historical plague scholarship extends beyonds the bounds of Western Europe, the traditional narrative of three pandemic plague becomes unstable.

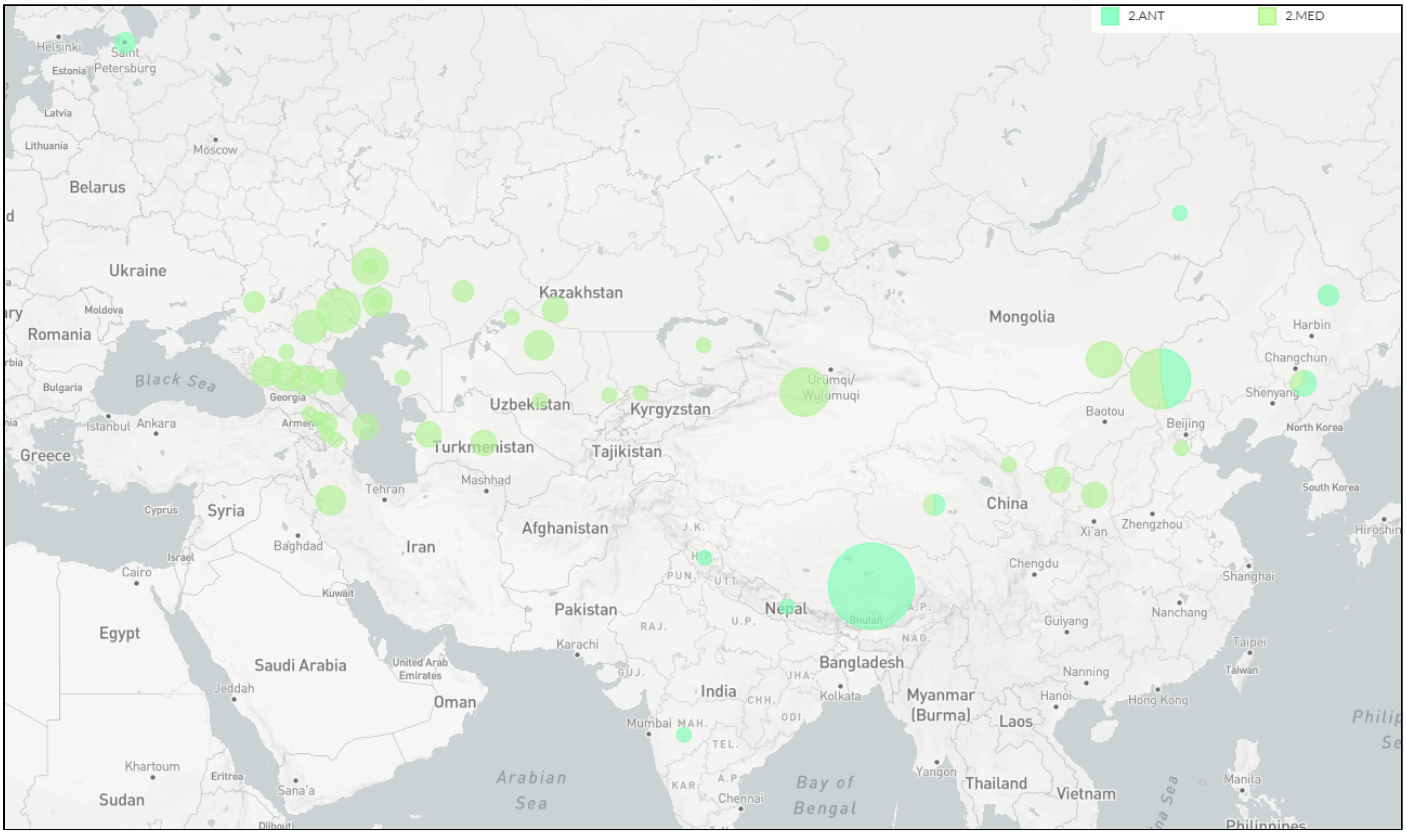

Figure 3: The geographic distribution of *Y. pestis* `Branch 2`.

Hosts

Clades of *Y. pestis* are additionally defined by virulence in particular hosts. For example, the Pestoides clade is frequently avirulent in humans, although sporadic cases of human plague do still occur (CITE). In general, all branches of *Y. pestis* are capable of causing plague in humans and the species barrier between wild rodents and human popuations is crossed frequently (FIGURE).

While clades of ancient *Y. pestis* are exclusively associated with humans, this is more likely due to the sampling strategies of ancient DNA studies which have prioritized human skeletal remains over zooarchaeological remains. Given that no other clades across the *Y. pestis* phylogeny show a specificty for human hosts, ... isolate aDNA *Y. pestis* from rats.

> Plague can cause disease in humans at any time, and from anywhere. There are virtually no lineages that are “safe” for humans. Thus plague as a disease is treated with exceptional caution.

> No lineage of modern plague has been observed to exclusively infect humans (?) and thus ... attention to multi-host ecology.

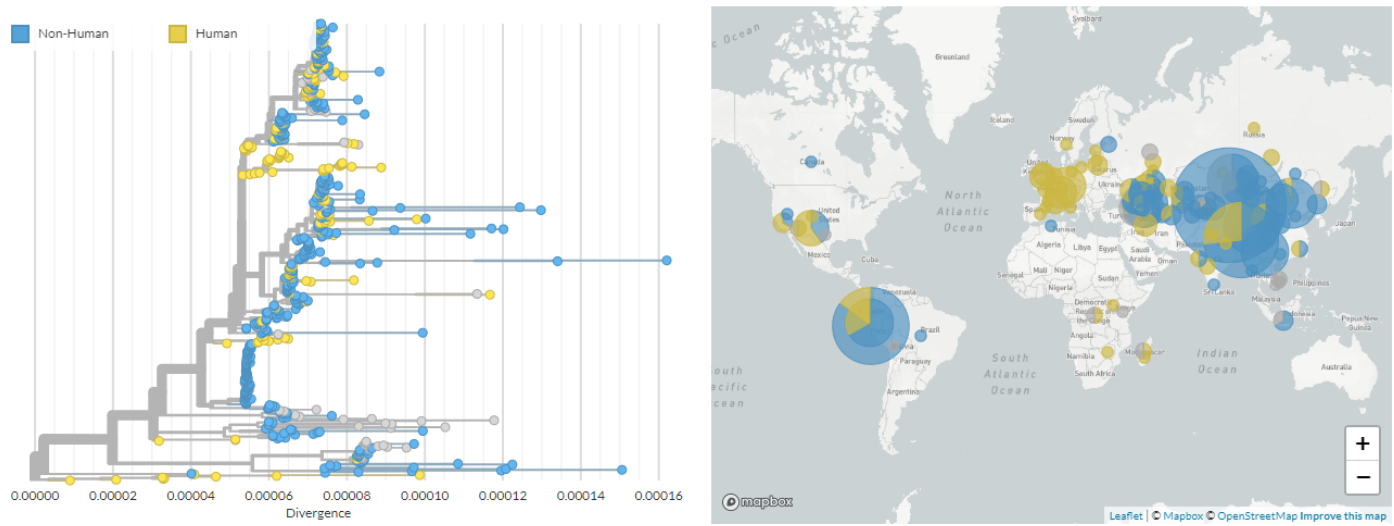

Figure 4: Distribution of human vs. non-human samples.

Phylodynamics

> How has the known diversity of plague changed over time?

Temporal Distribution

Historical strains of plague isolated prior to the 20th century cluster into three time periods: the Late Neolithic Bronze Age (LNBA), Roman Period/Late Antiquity(?), and the Medieval Period (5). For example, all strains of *Y. pestis* isolated in the Roman period group together in Branch 0 and form a distinct subclade within 0.ANT. Similarly, all Medieval strains of *Y. pestis* cluster together in Branch 1 as clade 1.PRE. The LNBA is a notable exception as mixing between clades 0.PE and 0.PRE is observed.

However, collection date and collection location are highly confounded as ancient DNA sampling strategies have predominantly targeted Western Europe. Due to this sampling bias, it is challenging to evaluate whether multiple strains of plague have co-occured in human populations in the past.

An additional observation is that the temporal sampling strategy of genomic data reflects greater interest in *Y. pestis* as a historical pathogen, rather than a public health threat to modern humans. One example of this is that the Medieval Plague in Western Europe (Clade 1.PRE) has more representative samples than all of the African continent (Clade 1.ANT). Sequencing initiatives are greatly needed that shift the balance away from Eurocentrism and encompass a greater diversity of affected populations.

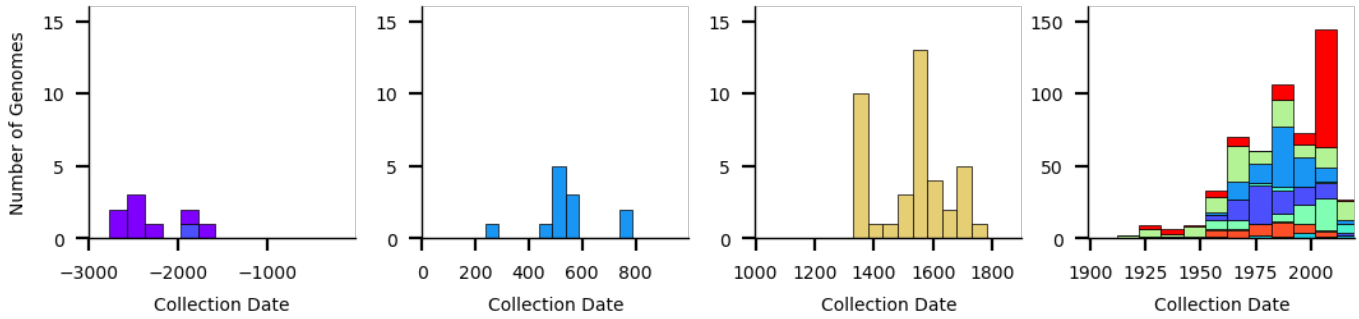

Figure 5: Temporal distribution of *Y. pestis* genomes.

Temporal Signal

> **TDLR;**
>
> - *Y. pestis* has more temporal signal than previously thought.
> - A root-to-tip regression is a poor measure of temporal signal.

Y. pestis has been unambiguously shown to exhibit substantial rate variation both between and within clades [22,23]. It is therefore unsurprising that extreme rate variation is also observed in this study. The Coefficient of Determination (R^2) of the linear regression for the full dataset is extremely low at 0.09 (Table 1). This suggests that a simple linear model, such as the strict clock model, is overall a poor fit for the data.

> What were the BETS results for the reduced dataset?

Taken together, these findings seem to support the observation that *Y. pestis* as a whole does not demonstrate temporal signal.

However, visual inspection of the Root to Tip Regression hints that rate variation may be partially explained by clade-specific clocks (Figure 6). When examining each clade in isolation, evidence of strict clock-like behavior is recovered in both the Root to Tip Regression and the BETS Bayes factors (Table 1, Figure 7). Intriguingly, clades that have been associated with historically documented pandemics are well-modeled by a strict clock including the First Pandemic clade `0.ANT4`, the Second Pandemic clade `1.PRE` and the Third Pandemic clade `1.OR1`. The final clade which demonstrates strict clock behavior is the Late Neolithic Bronze Age group `0.PRE`. As this time period pre-dates historical records, there is little opportunity to estimate mortality rates. However, this similarity may lend support to the hypothesis that an early pandemic of plague was occurring at that time 2. It may be a useful avenue of research to investigate to what extent rate variation is a useful predictor for 'pandemic potential'.

While a Root to Tip Regression can be useful tool to explore temporal signal, it has several known limitations. Namely the underlying assumption of strict clock-behavior and the non-independence of data points [35]. A BETS analysis counters both of these limitations, and is overall more sensitive given that multiple clock models can be tested. The superior performance of the BETS test can be seen in (Table 1). The Root to Tip Regression detects temporal signal in 5/12 clades while BETS detects signal in 7/12 clades. Furthermore, in all cases the [[Clock Model|relaxed clock]] proves to have stronger support, even when the regression Coefficient of Determination is high. The conclusion is that a root-to-tip regression is a relatively poor measure of temporal signal in *Y. pestis*.

> Wait for an update on 0.PRE to confirm this.

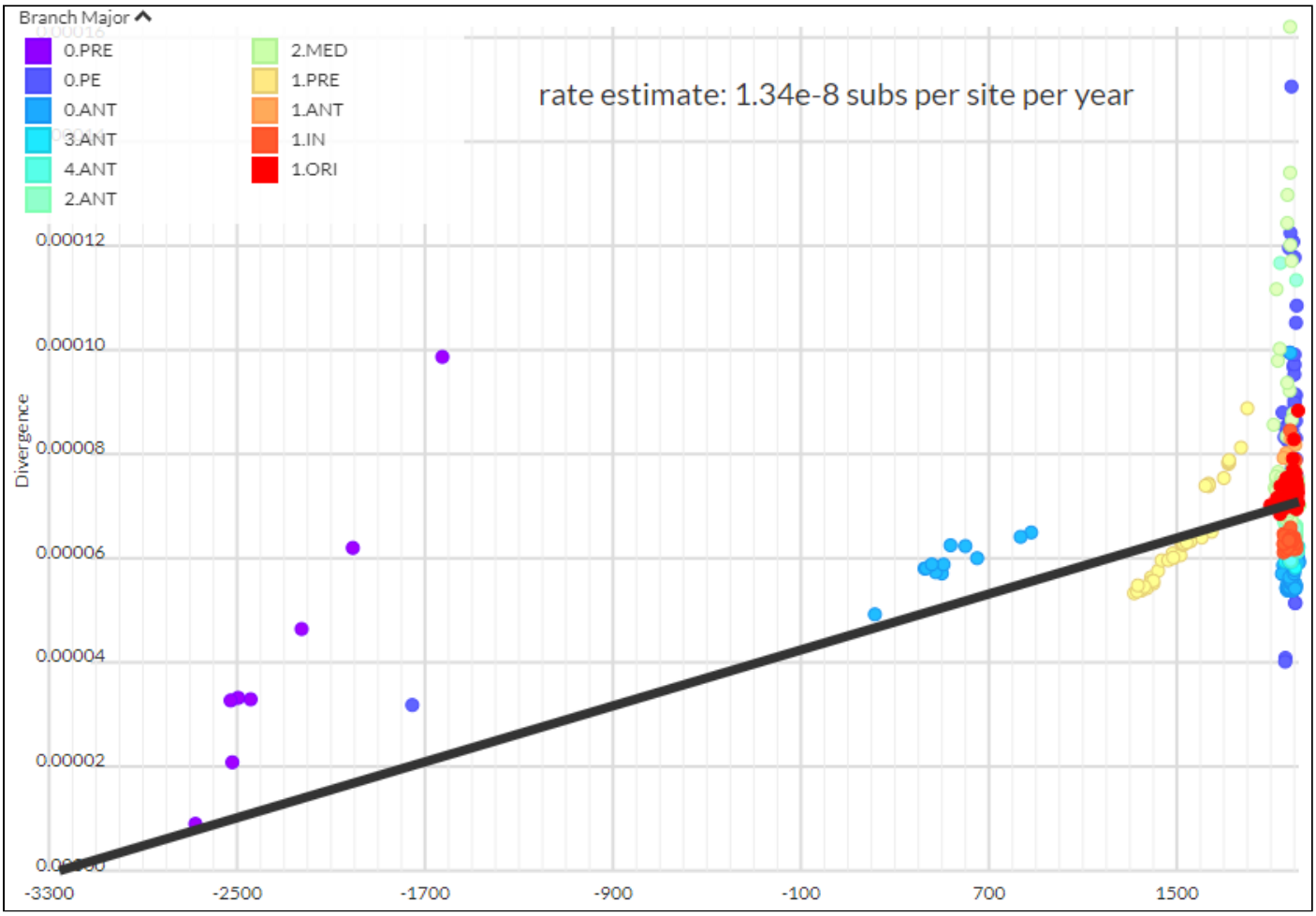

Figure 6: Root to Tip Regression of Y. pestis on sampling date, colored by clade. (PLACEHOLDER)

> So what about Branch 2, since I was so adamant that it could be an undocumented historical pandemic?

Table 1: Temporal signal statistics by clade based on a root-to-tip linear regression. A * indicates a significant p-value or bayes factor.

Branch	Clade	Origin	RTT R^2	RTT p-value	Strict Clock BF	Relaxed Clock BF
All	All	Ancient, Modern	0.09	3.81E-14*	–	–
1	1.ORI	Modern	0.04	1.32E-02*	29.6*	35.7*
1	1.IN	Modern	0.0	3.24E-01	-3.9	-10.2
1	1.ANT	Modern	0.45	2.03E-01	8.9*	12.6*
1	1.PRE	Ancient	0.76	1.68E-13*	10.1*	44.1*
2	2.MED	Modern	0.01	1.86E-01	–	–
2	2.ANT	Modern	0.05	5.96E-02	-20.8	-13.7
4	4.ANT	Modern	-0.11	8.80E-01	-2.9	3.7*
3	3.ANT	Modern	-0.04	4.39E-01	-9.6	-11.4
0	0.ANT	Modern	-0.01	7.35E-01	-2.3	-6.5
0	0.ANT4	Ancient	0.66	7.84E-04*	5.3*	5.9*
0	0.PE	Modern	0.01	2.25E-01	-82.1	12.4*
0	0.PRE	Ancient	0.91	1.53E-04*	83.0*	-2.9

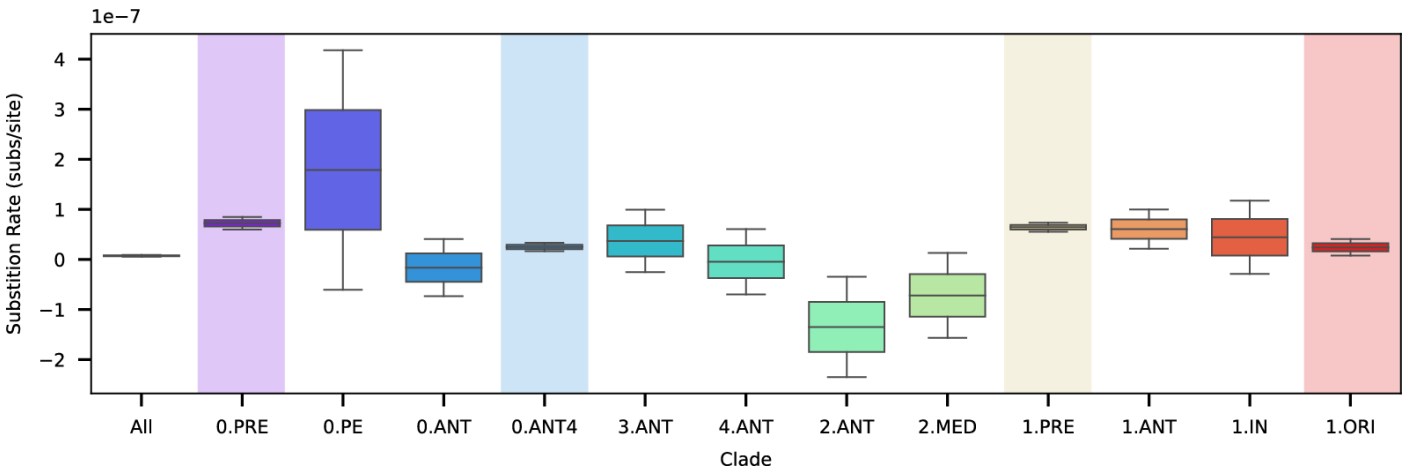

Figure 7: Mean substitition rate uncertainty by clade based on a non-parametric bootstrap of the root-to-tip linear regression. Highlighted clades show statistical support for a strict clock.

Clock Model

- Relaxed clock MCMC runs produce a high Coefficient of Variation indicating a relaxed model is favored over a strict model (Figure 8). 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.

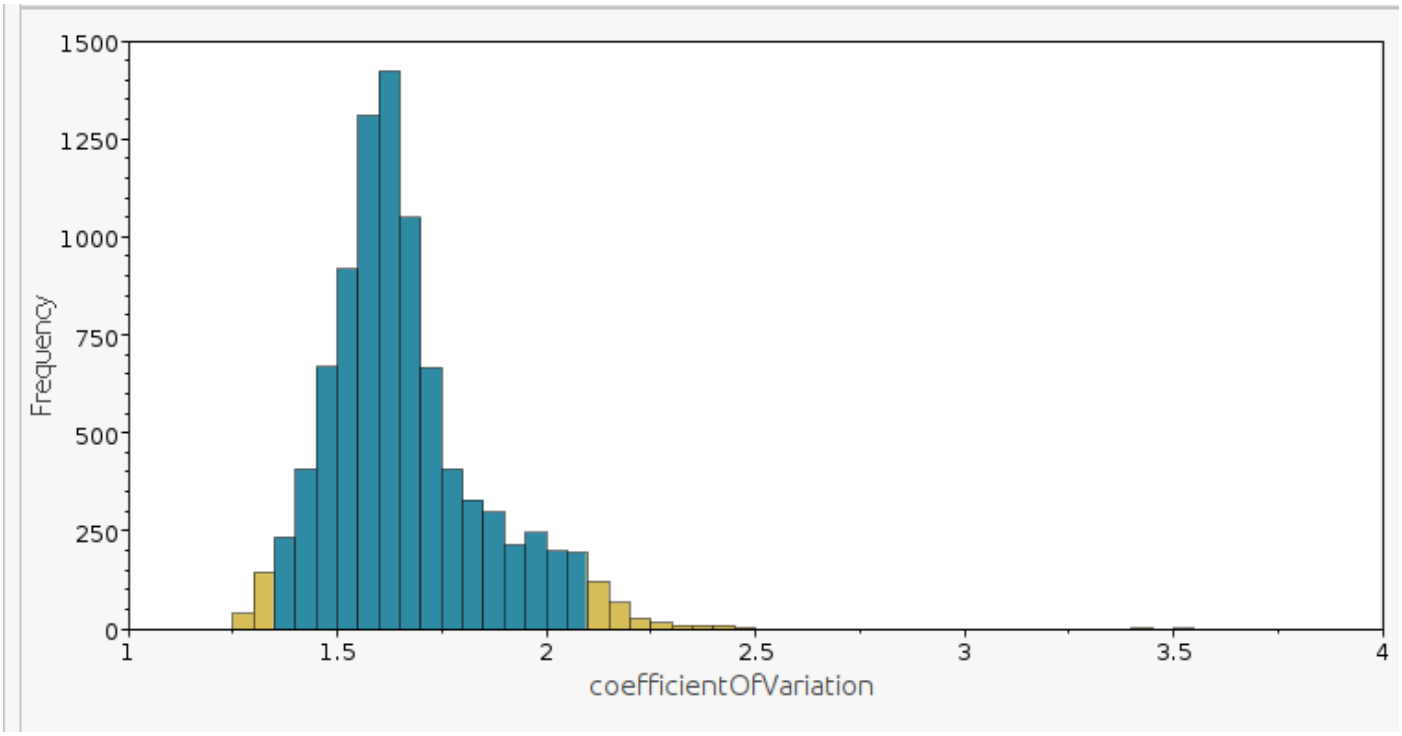

Figure 8: Coefficient of variation.

- A strict clock and relaxed clock have overlapping distributions with similar peaks for the Tree Height (blue: strict, green: relaxed) (Figure 9).

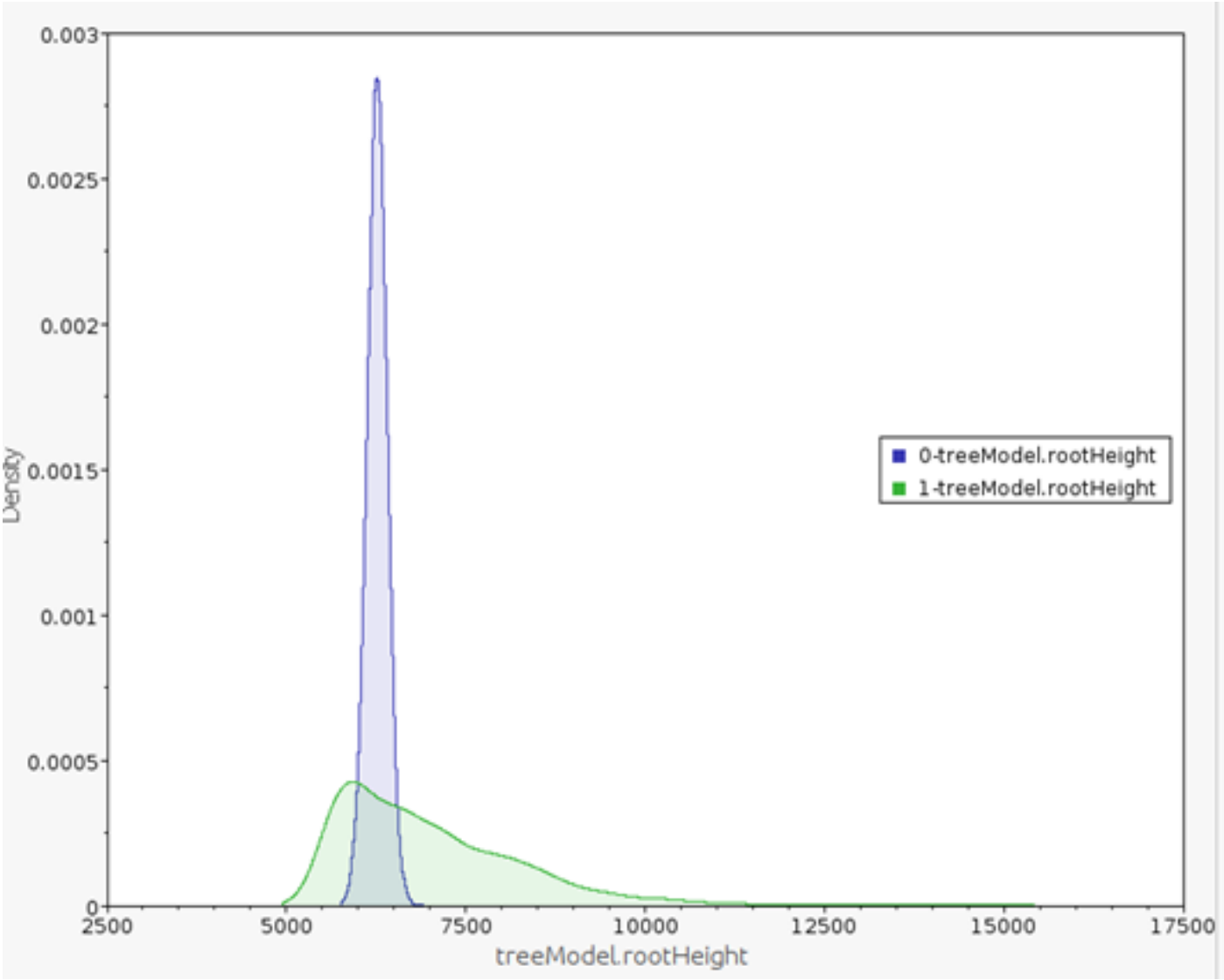

Figure 9: Tree height comparison.

- 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 10).

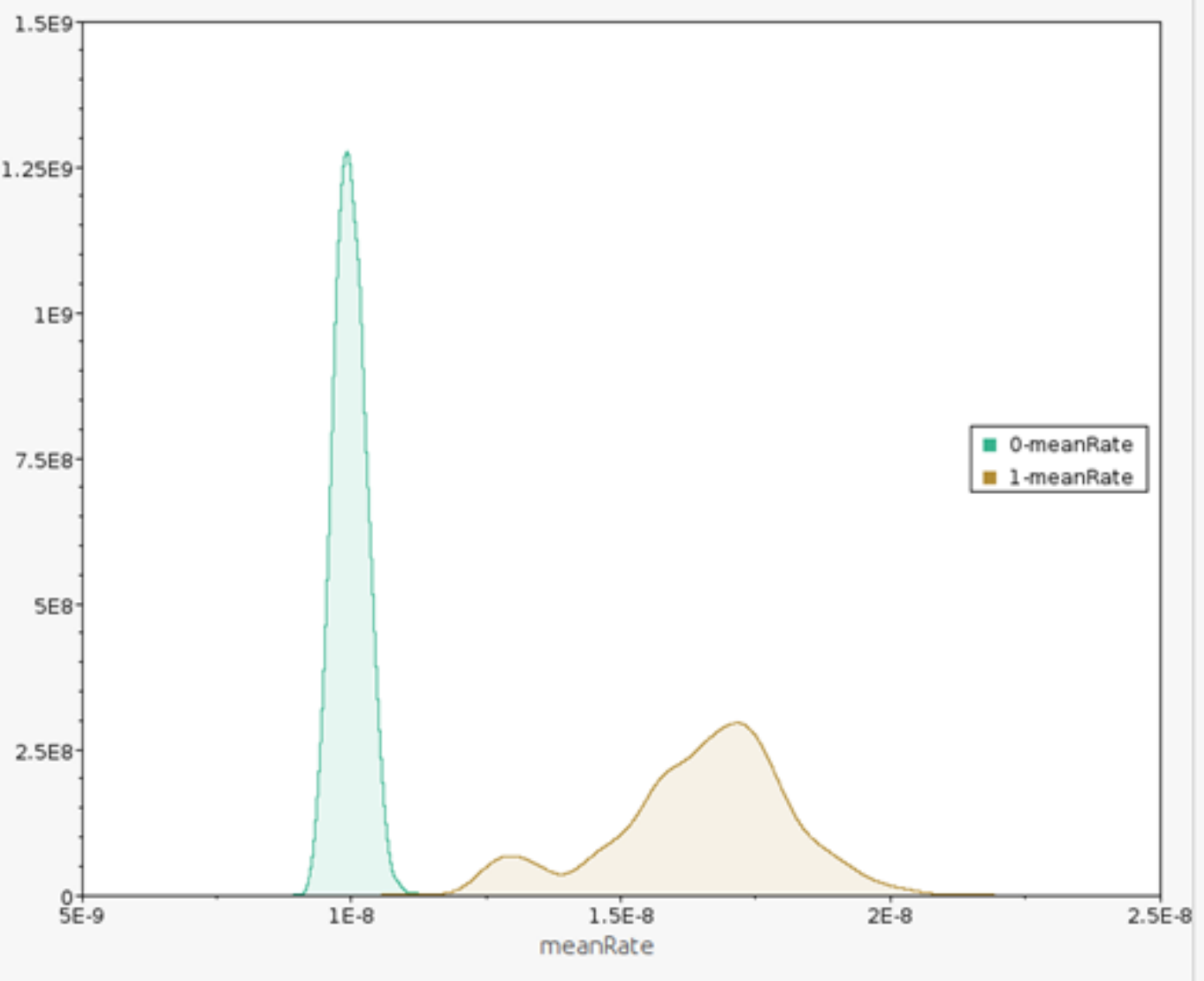

Figure 10: Substitution rate comparison.

- There doesn't appear to be clustering of rates. Branches with high rates are next to those with low rates (Figure 11).

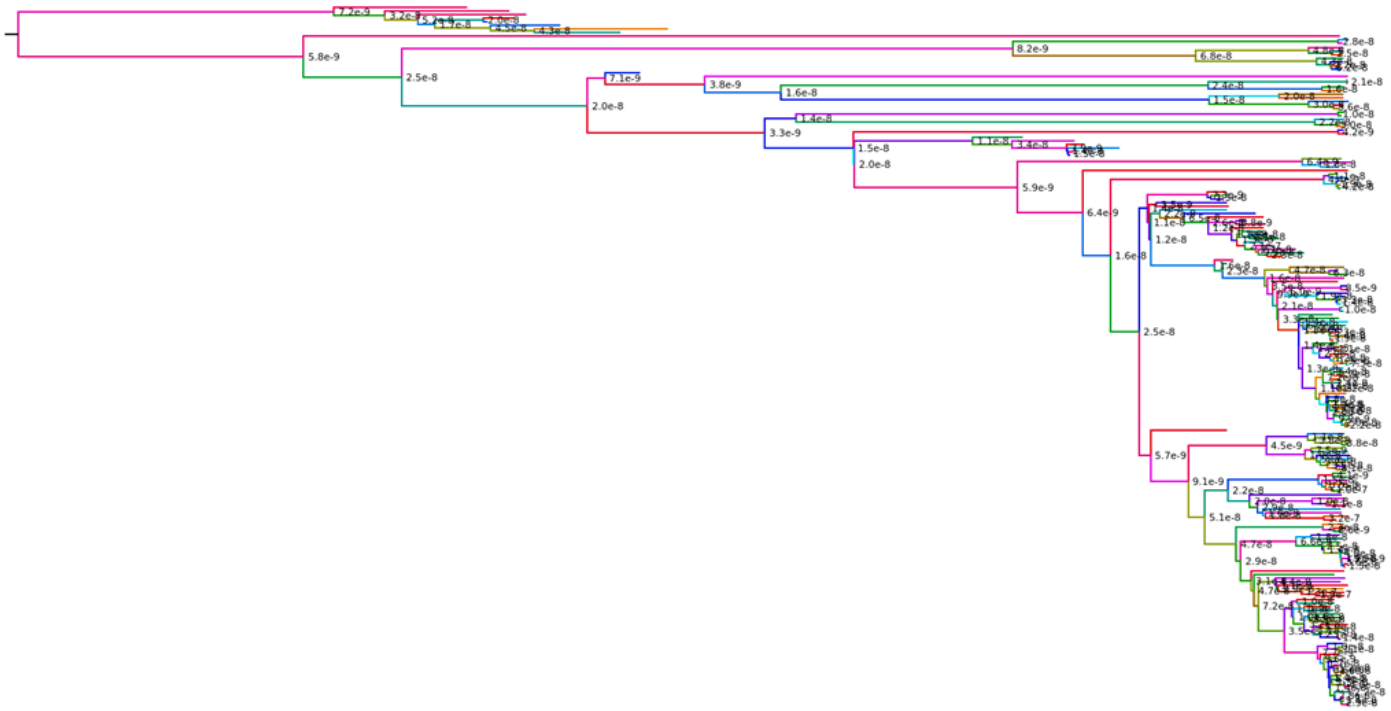

Figure 11: Time tree colored by rate.

Mean Substitution Rate

> **TLDR;**
>
> - The mean substitution rate of *Y. pestis* is of little interpretive value given substantial rate variation and the time-dependency on sampling date.

Y. pestis has one of the slowest substitution rates observed in a bacterial pathogen (Table 2). Given the tremendous variation observed in modern plague ecology, it is surprising that the evolutionary rate does not reflect this need to rapidly adapt to changing environments. However, this slow rate makes perfect sense when viewed in the context of rate *time-dependency*, wherein the observed substitution rate decreases as the sampling time frame increases. Furthermore, given that the full dataset shows no temporal signal, likely due to the lineage-specific variation showed in Figure 7, the mean substitution rate is of little interpretive value.

The time-dependency does not hold in the clade datasets. Clades sampled in a narrow time frame (ex. `1.ORI`) can have a slower rate than clades with wider sampling times (ex. `0.PRE`).

Table 2: Substitution rates of bacterial pathogens.

Organism	Disease	Substitution Rate (subs/site year^{-1})	Sampling Time (years)	Study
Yersinia pestis	Plague	1.42 x 10^{-8}	4687	This Study
Mycobacterium leprae	Leprosy	1.56 x 10^{-8}	1993	[13]
Mycobacterium tuberulcosis	Tuberculosis	5.39 x 10^{-8}	895	[13]
Neisseria meningitis	Meningitis	6.05 x 10^{-8}	59	[13]
Salmonella enterica	Typhoid	7.60 x 10^{-8}	84	[13]

Organism	Disease	Substitution Rate (subs/site year^{-1})	Sampling Time (years)	Study
Pseudomonas aeruginosa	Pneumonia	3.36 x 10^{-7}	35	[13]

> The Big Bang Polytomy is surrounded by clades with varying degrees of temporal signal. The branches surrounded the Big Bang Polytomy [23] show some of the most dramatic rate acceleration to accomodate.

Phylogeography

Geographic Distribution

The geographic distribution of *Y. pestis* reflects a complex dispersal history (Figure 12 .B). Many regions have been colonized by diverse strains of *Y. pestis*. This diversity can be contemporaneous, such as endemic foci in the Caucausus and Western China (Clade 0.PE). Alternatively, this diversity may occur over multiple centuries through distinct re-introductions and extinctions, as seen in the historical epidemics of Europe (Clades 0.ANT and 1.PRE). In these examples, a relatively large amount of genetic diversity appears in a small geographic range. In contrast, regions such as the Americas have been colonized by a single strain of *Y. pestis* (Clade 1.ORI) which shows a relatively small amount of genetic diversity over a tremendously large geographic range.

An important consideration is that the geographic sampling strategy of *Y. pestis* genomes (Figure 12 .A) does not reflect the known distribution of modern plague [34], let alone historical pandemics. Nor does it adequately characterize the most heavily affected regions of the world, namely Madagascar and the Democratic Republic of the Congo [5]. The over-sampling of East Asia has been previously described by 36 and considerably drives the hypothesis that *Y. pestis* originated in China [23,37]. This once established hypothesis is now in contention, as the most basal strains of *Y. pestis* (Clades 0.PRE and 0.PE) have been isolated from all across Eurasia.

> The geographic distribution of *Y. pestis* genomes reflects the *sampling history* more so than the *natural history* of the disease.

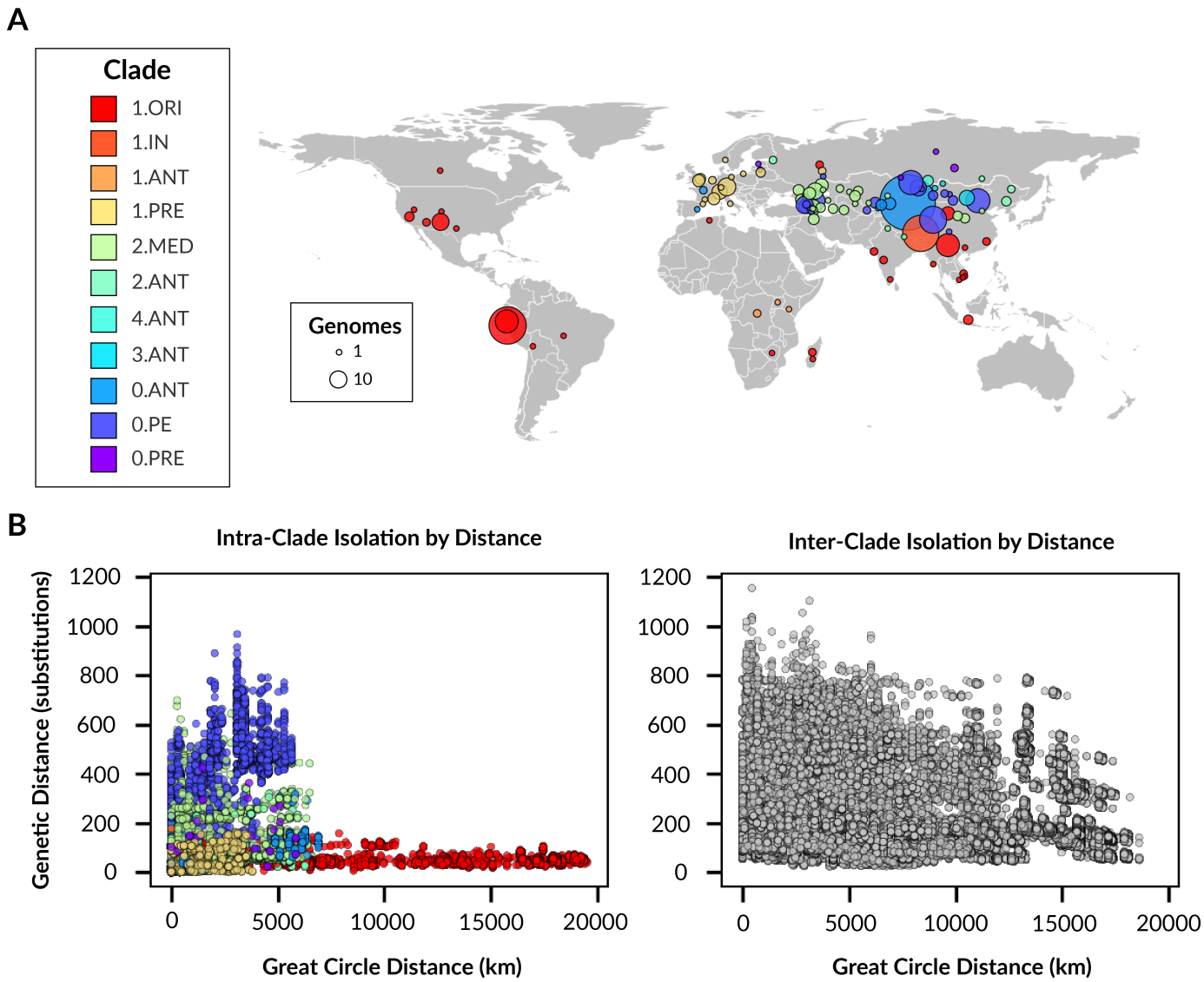

Figure 12: Spatiotemporal distribution of *Y. pestis* genomes. **A**: Geographic distribution, **B**: Isolation by distance as a function of geographic distance and genetic distance.

Ancestral Reconstruction

The confidence with which ancestral location could be estimated is described in Table 3 and visualized in Figure 13. Across the entire tree, 77% of internal nodes could be estimated with high confidence (>=0.95).

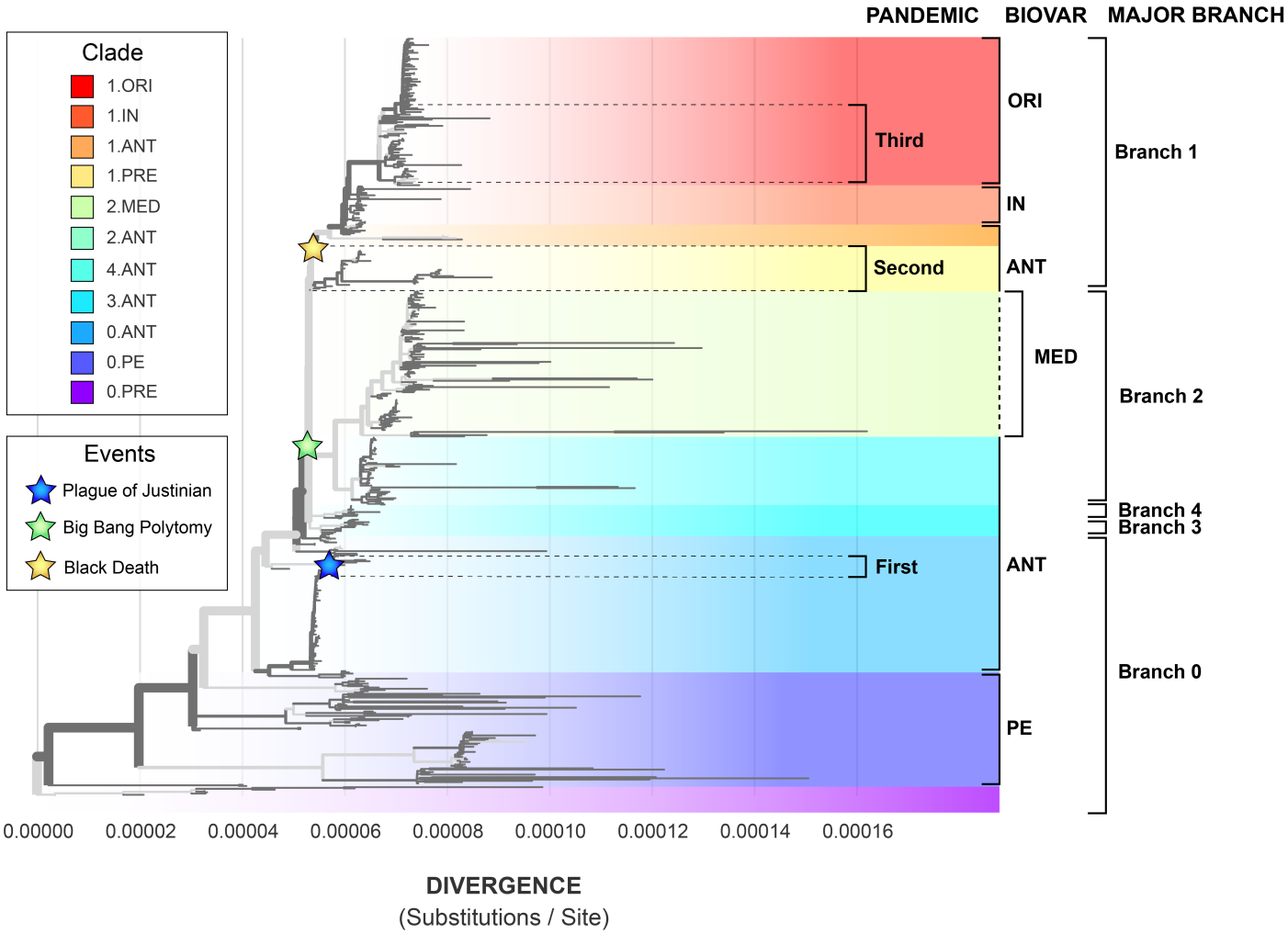

Figure 13: Discrete state phylogeography confidence. High confidence branches (>=0.95) are colored black, low confidence branches are colored light grey.

Table 3: Discrete state phylogeography confidence.

Clade	Total Nodes	Tips	Internal Nodes	High Confidence Nodes	Percent High Confidence
All	1201	601	600	461	76.83
1.ORI	233	117	116	93	80.17
1.IN	78	39	39	37	94.87
1.ANT	7	4	3	0	0.0
1.PRE	80	40	40	30	75.0
2.MED	231	116	115	62	53.91
2.ANT	107	54	53	47	88.68
4.ANT	21	11	10	6	60.0
3.ANT	21	11	10	6	60.0
0.ANT	207	103	104	98	94.23
0.ANT4	23	12	11	8	72.73
0.PE	169	84	85	70	82.35
0.PRE	15	8	7	3	42.86

The clade associated with the highest confidence (95% of nodes) is the *intermedium* biovar (`INT`) that falls just basal to the Third Pandemic clade `1.ORI` (Figure 14). The root of this clade is estimated to originate in Qinghai Province China, followed by independent radiations to Xinjian, Gansu, Tibet, and Yunnan. The lineage associated with Yunnan province then gives rise to `1.ORI`. Overall this geographic ancestry is consistent with the known history of the Third Pandemic.

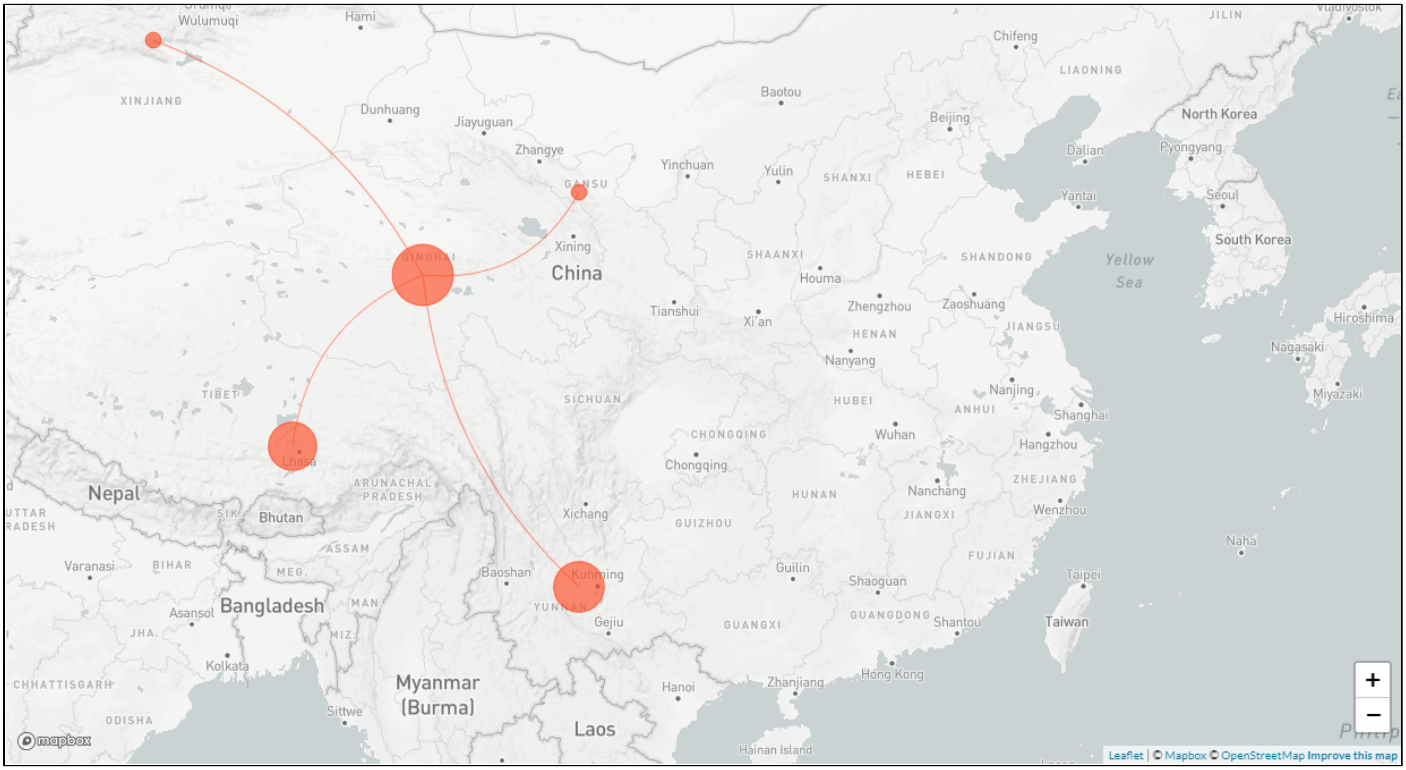

Figure 14: Geographic distribution of `1.IN` and inferred migration history.

The clade with the next highest confidence is the basal *antiqua* biovar `[[0.ANT]`. However, this clade is dominated by 50 years (1956-2006) of environment surveillance in the Junggar Basin of Xinjiang.

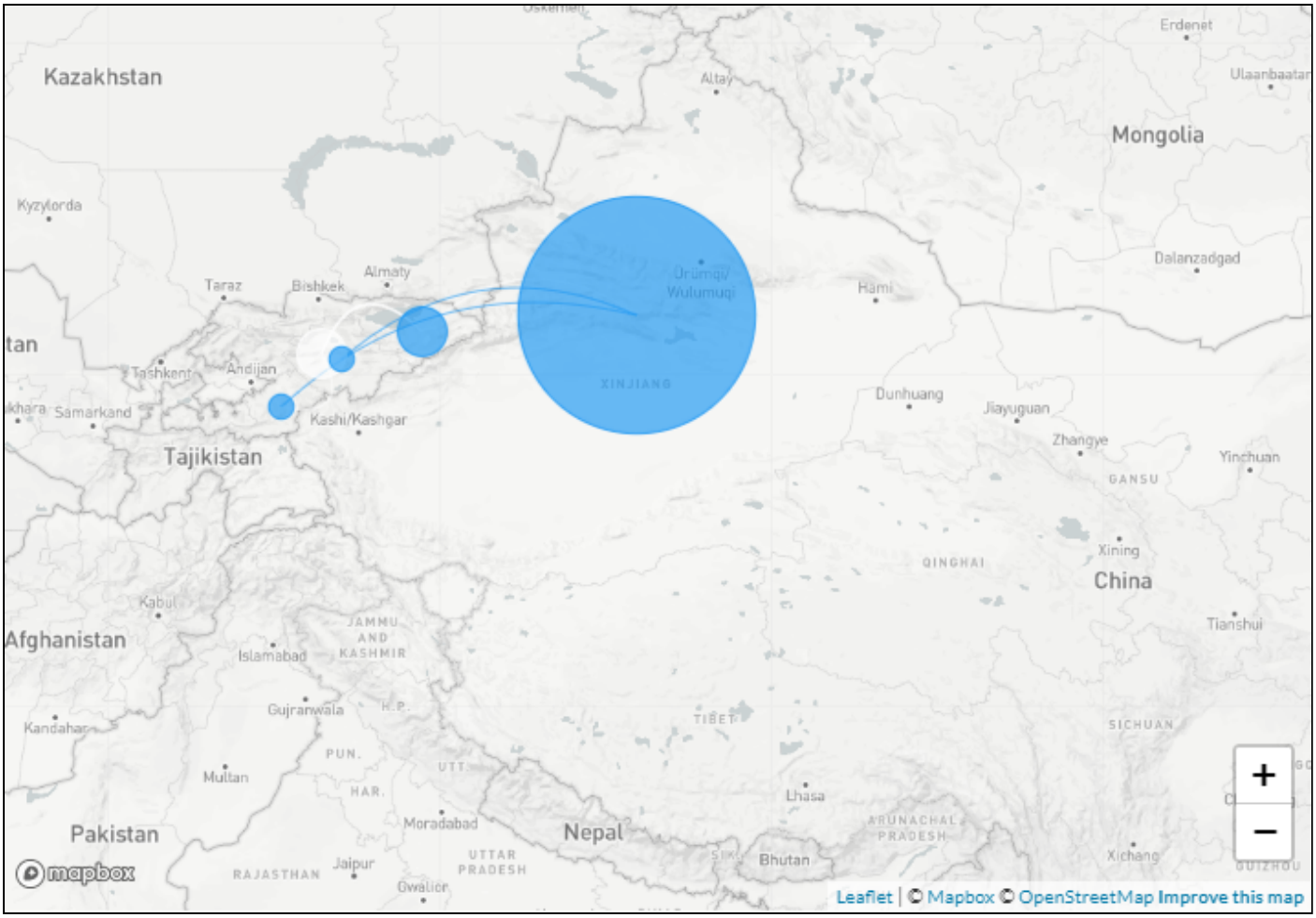

Figure 15: Geographic distribution of `0.ANT` and inferred migration history.

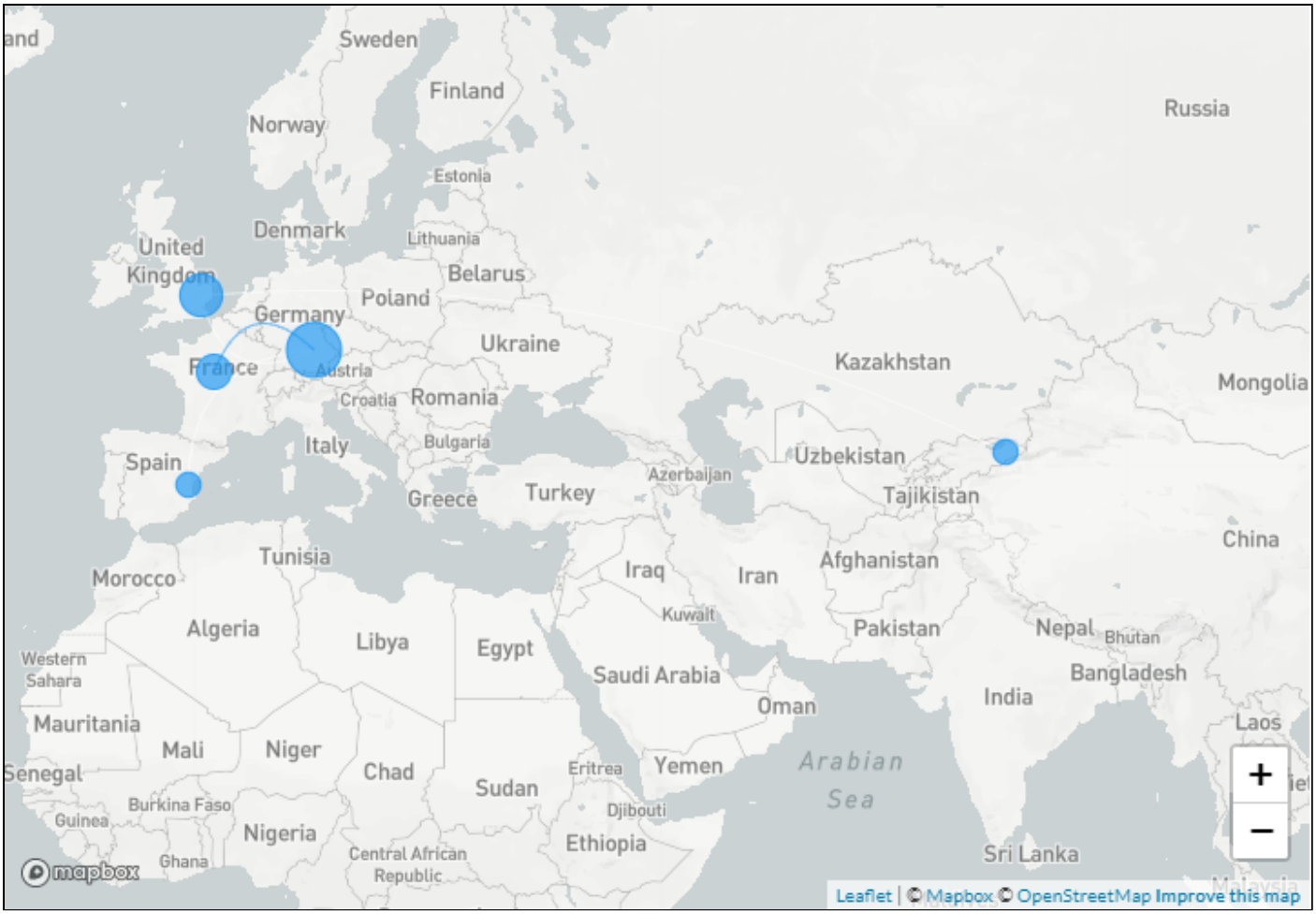

Figure 16: Geographic distribution of `0.ANT4` and inferred migration history.

Conclusion

Appendix

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. **The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity**
Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Mark Achtman
Genome Research (2020-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6961584/
DOI: 10.1101/gr.251678.119 · PMID: 31809257 · PMCID: PMC6961584

12. ***Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis**
David M Wagner, Jennifer Klunk, Michaela Harbeck, Alison Devault, Nicholas Waglechner, Jason W Sahl, Jacob Enk, Dawn N Birdsell, Melanie Kuch, Candice Lumibao, … Hendrik Poinar
The Lancet Infectious Diseases (2014-04) https://linkinghub.elsevier.com/retrieve/pii/S1473309913703232
DOI: 10.1016/s1473-3099(13)70323-2

13. **Genome-scale rates of evolutionary change in bacteria**
Sebastian Duchêne, Kathryn E. Holt, François-Xavier Weill, Simon Le Hello, Jane Hawkey, David J. Edwards, Mathieu Fourment, Edward C. Holmes
Microbial Genomics (2016-11-30) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320706/
DOI: 10.1099/mgen.0.000094 · PMID: 28348834 · PMCID: PMC5320706

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

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

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

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

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

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

20. **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

21. **Genomic Insights into a Sustained National Outbreak of Yersinia pseudotuberculosis**
Deborah A. Williamson, Sarah L. Baines, Glen P. Carter, Anders Gonçalves da Silva, Xiaoyun Ren, Jill Sherwood, Muriel Dufour, Mark B. Schultz, Nigel P. French, Torsten Seemann, … Benjamin P. Howden
Genome Biology and Evolution (2016-12-01) https://doi.org/10.1093/gbe/evw285
DOI: 10.1093/gbe/evw285

22. **Phylogeography of the second plague pandemic revealed through analysis of historical Yersinia pestis genomes**
Maria A. Spyrou, Marcel Keller, Rezeda I. Tukhbatova, Christiana L. Scheib, Elizabeth A. Nelson, Aida Andrades Valtueña, Gunnar U. Neumann, Don Walker, Amelie Alterauge, Niamh Carty, … Johannes Krause
Nature Communications (2019-10-02) https://www.nature.com/articles/s41467-019-12154-0
DOI: 10.1038/s41467-019-12154-0

23. **Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis***
Y. Cui, C. Yu, Y. Yan, D. Li, Y. Li, T. Jombart, L. A. Weinert, Z. Wang, Z. Guo, L. Xu, … R. Yang
Proceedings of the National Academy of Sciences (2013-01-08) http://www.pnas.org/cgi/doi/10.1073/pnas.1205750110
DOI: 10.1073/pnas.1205750110

24. **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

25. **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

26. **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

27. **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

28. **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

29. **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

30. **Genotyping and Phylogenetic Analysis of Yersinia pestis by MLVA: Insights into the Worldwide Expansion of Central Asia Plague Foci**
Yanjun Li, Yujun Cui, Yolande Hauck, Mikhail E. Platonov, Erhei Dai, Yajun Song, Zhaobiao Guo, Christine Pourcel, Svetlana V. Dentovskaya, Andrey P. Anisimov, … Gilles Vergnaud
PLOS ONE (2009-06-22) https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006000
DOI: 10.1371/journal.pone.0006000

31. **Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago**
Simon Rasmussen, Morten Erik Allentoft, Kasper Nielsen, Ludovic Orlando, Martin Sikora, Karl-Göran Sjögren, Anders Gorm Pedersen, Mikkel Schubert, Alex Van Dam, Christian Moliin Outzen Kapel, … Eske Willerslev
Cell (2015-10-22) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644222/
DOI: 10.1016/j.cell.2015.10.009 · PMID: 26496604 · PMCID: PMC4644222

32. **Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750)**
Marcel Keller, Maria A. Spyrou, Christiana L. Scheib, Gunnar U. Neumann, Andreas Kröpelin, Brigitte Haas-Gebhard, Bernd Päffgen, Jochen Haberstroh, Albert Ribera i Lacomba, Claude Raynaud, … Johannes Krause
Proceedings of the National Academy of Sciences (2019-06-18) https://www.pnas.org/content/116/25/12363
DOI: 10.1073/pnas.1820447116 · PMID: 31164419

33. **Wet climate and transportation routes accelerate spread of human plague**
Lei Xu, Leif Chr. Stige, Kyrre Linné Kausrud, Tamara Ben Ari, Shuchun Wang, Xiye Fang, Boris V. Schmid, Qiyong Liu, Nils Chr. Stenseth, Zhibin Zhang
Proceedings of the Royal Society B: Biological Sciences (2014-04-07) https://royalsocietypublishing.org/doi/10.1098/rspb.2013.3159
DOI: 10.1098/rspb.2013.3159

34. **Historical and genomic data reveal the influencing factors on global transmission velocity of plague during the Third Pandemic**
Lei Xu, Leif C. Stige, Herwig Leirs, Simon Neerinckx, Kenneth L. Gage, Ruifu Yang, Qiyong Liu, Barbara Bramanti, Katharine R. Dean, Hui Tang, … Zhibin Zhang
Proceedings of the National Academy of Sciences (2019-06-11) https://www.pnas.org/content/116/24/11833
DOI: 10.1073/pnas.1901366116 · PMID: 31138696

35. **Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations**
Sebastian Duchene, Philippe Lemey, Tanja Stadler, Simon YW Ho, David A Duchene, Vijaykrishna Dhanasekaran, Guy Baele
Molecular Biology and Evolution (2020-11-01) https://doi.org/10.1093/molbev/msaa163
DOI: 10.1093/molbev/msaa163

36. **Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics**
Maria A. Spyrou, Rezeda I. Tukhbatova, Michal Feldman, Joanna Drath, Sacha Kacki, Julia Beltrán de Heredia, Susanne Arnold, Airat G. Sitdikov, Dominique Castex, Joachim Wahl, … Johannes Krause
Cell Host & Microbe (2016-06) http://linkinghub.elsevier.com/retrieve/pii/S1931312816302086
DOI: 10.1016/j.chom.2016.05.012

37. ***Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity**
Giovanna Morelli, Yajun Song, Camila J. Mazzoni, Mark Eppinger, Philippe Roumagnac, David M. Wagner, Mirjam Feldkamp, Barica Kusecek, Amy J. Vogler, Yanjun Li, … Mark Achtman
Nature Genetics (2010-12)
DOI: 10.1038/ng.705 · PMID: 21037571 · PMCID: PMC2999892