Plagued by a cryptic clock

**Can we date the global phylogeny of *Yersinia pestis*?**

This manuscript ([permalink](https://rawcdn.githack.com/ktmeaton/obsidian-public/18455b83/academic/Plague%20Phylodynamics%20and%20Phylogeography%20Paper.html)) was automatically generated from [ktmeaton/obsidian-public@18455b83](https://github.com/ktmeaton/obsidian-public/tree/18455b83) on September 15, 2021.

## Authors

* **Katherine Eaton** ORCID icon [0000-0001-6862-7756](https://orcid.org/0000-0001-6862-7756)  
   McMaster Ancient DNA Center; Department of Anthropology, McMaster University
* **Leo Featherstone** ORCID icon [0000-0002-8878-1758](https://orcid.org/0000-0002-8878-1758)  
   The Peter Doherty Institute For Infection and Immunity , University of Melbourne
* **Sebastian Duchene** ORCID icon [0000-0002-2863-0907](https://orcid.org/0000-0002-2863-0907)  
   The Peter Doherty Institute For Infection and Immunity , University of Melbourne
* **Ann Carmichael**  
   Indiana University Bloomington
* **Nükhet Varlık** ORCID icon [0000-0001-6870-5945](https://orcid.org/0000-0001-6870-5945)  
   University of South Carolina
* **Brian Golding** ORCID icon [0000-0002-7575-0282](https://orcid.org/0000-0002-7575-0282)  
   Department of Biology, McMaster University
* **Eddie Holmes** ORCID icon [0000-0001-9596-3552](https://orcid.org/0000-0001-9596-3552)  
   University of Sydney
* **Hendrik Poinar** ORCID icon [0000-0002-0314-4160](https://orcid.org/0000-0002-0314-4160)  
   McMaster Ancient DNA Center; Department of Anthropology, McMaster University

## Abstract

**Background**: A popular objective in plague research is dating historical events using a molecular clock. But despite intensive interest and methodological advancement, *Yersinia pestis* remains notoriously difficult to model. As a result, debate has emerged concerning the applicability and accuracy of molecular clock approaches in plague phylogenetics. In response, we test a variety of molecular clock models on an updated and curated *Y. pestis* phylogeny.

**Results**: All iterations of a species-wide clock were analytically unstable. Partitioning the data by population improved modeling performance, but often resulted in divergence dates that were too young. This effect was most noticeable in populations with no ancient DNA calibrations, even when the sampling time frame was as long as 100 years.

**Conclusions**: The majority of *Y. pestis* populations have detectable temporal signal. However, accurate node-dating presents an optimization problem, in which calibrating samples are maximized and rate variation is limited.

## 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](#ref-Vo6ReJPm),[2](#ref-AQa9Tn4j)]. 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](#ref-1093vihdz),[4](#ref-uM6Rh5Fu)]. 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](#ref-h4WAqbKy)].

Accompanying this prolific global presence is unnervingly high mortality. The infamous medieval Black Death (1346-1353) is estimated to have killed more than half of Europe’s population [[6](#ref-13ET92iS0)]. This virulence can still be observed in the post-antibiotic era, where case fatality rates range from 22-71% [[7](#ref-Nd9WUU1j)]. As a result, plague maintains its status as a disease that is of vital importance for public health research.

A long-standing line of inquiry in plague research has been estimating evolutionary rates in order to date internal nodes. Key areas of the phylogeny that have been intensively researched are the first emergence of *Y. pestis* in human populations [[8](#ref-ZvyYZ3sx)], the “Big Bang” polytomy [[9](#ref-qLAL8Don)], and the onset of past pandemics [[10](#ref-vk7LFPKc),[11](#ref-ONhNS9aO),[12](#ref-ACt53Sow)]. Recent technological advancements, such as ancient DNA sequencing and new molecular clock methods, have enabled researchers to reach further back in time with increasingly complex models. But despite this intensive interest and methodological advancement, *Y. pestis* remains notoriously difficult to model using a molecular clock approach.

This difficulty can largely be attributed to the substantial rate variation that has been documented across the phylogeny of *Y. pestis* [[11](#ref-ONhNS9aO),[13](#ref-1CPIgshmC)]. As a result, considerable debate has emerged over whether *Y. pestis* has absolutely no temporal signal [[12](#ref-ACt53Sow)], or if populations have such distinct rates that a species-wide signal is obscured [[13](#ref-1CPIgshmC),[14](#ref-1cTipQcd)]. This uncertainty has produced radically different temporal models between studies, with node dates shifted by as much as several millennia [[8](#ref-ZvyYZ3sx),[11](#ref-ONhNS9aO)]. Thus a comprehensive understanding of plague’s molecular clock, or lack thereof, is necessary before we can begin to untangle when and where this disease appeared in the past.

A critical step in reconstructing the evolutionary history of an organism is exploring the degree of population or genetic structure [[15](#ref-lYwCAR8r)]. This structure develops as a populations subdivide and diversify in isolation, producing a pattern of lower diversity within groups and higher diversity between them. This knowledge can then be used to add nuance to phylogenetic analyses and interpretations, by explicitly modeling these unique evolutionary histories. However, there is little consensus concerning the genetic structure of *Y. pestis* on a global scale, and it was recently proposed [[16](#ref-S7uUqUQG)] that our understanding of population structure should be revised according to the latest genomic data.

## Results and Discussion

### Population Structure

We began our explanation of population structure by estimating a maximum-likelihood phylogeny, using 601 global isolates including 540 modern (89.9%) and 61 ancient (10.1%) strains. In addition, two genomes of the outgroup taxa *Yersinia pseudotuberculosis* were included to root the tree. The alignment consisted of 10,249 variant positions exclusive to *Y. pestis*, with 3,844 sites shared by at least two strains. Following phylogenetic estimation, we pruned the outgroup taxa from the tree to more closely examine the genetic diversity of *Y. pestis*.

In Figure [1](#fig:divtree_map_timeline), we contextualize the global phylogeny using three widely-used nomenclature systems: the major branches, metabolic biovars, and historical time periods. In the following section, we compare and critique each system, identify any incongruent divisions and uncertainty, and explore an integrative approach for molecular clock analyses.

#### Biovar

The oldest system to date is the biovar nomenclature, which uses metabolic differences to define population structure. *Y. pestis* can be categorized into four classical biovars: *antiqua* (ANT), *medievalis* (MED), *orientalis* (ORI), and *microtus*/*pestoides* (PE) [[17](#ref-15WOrftK3),[18](#ref-cIuFxbEY)]. Non-classical biovars have also been introduced, such as the *intermedium* biovar (IN), which reflects a transitional state from *antiqua* to *orientalis* [[19](#ref-66GmpEaG)]. The biovar system is simple in application, as it largely focuses on two traits: the ability to ferment glycerol and reduce nitrate [[18](#ref-cIuFxbEY)]. However, this simplicity is offset by the growing recognition of regional inconsistencies in metabolic profiles [[16](#ref-S7uUqUQG)], which weakens its broader applicability. This is further exacerbated by the sequencing of non-viable, ‘extinct’ *Y. pestis*, for which metabolic sub-typing is impossible [[10](#ref-vk7LFPKc)]. Researchers have responded to this uncertainty in a variety of ways, by creating pseudo-biovars (PRE) [[8](#ref-ZvyYZ3sx)] or extrapolating existing ones [[12](#ref-ACt53Sow)]. Other still have foregone the *biovar* nomenclature altogether in favor of locally-developed taxonomies [[16](#ref-S7uUqUQG)]. Despite extensive research, it remains unclear which metabolic traits, if any, can be used to classify *Y. pestis* into distinct populations at a global scale.

#### Major Branch

In contrast to the biovar nomenclature which emphasizes phenotype, the major branch nomenclature focuses on the evolutionary relationships between strains. This system divides the global phylogeny of *Y. pestis* into populations according to their relative position to the “Big Bang” polytomy [[11](#ref-ONhNS9aO)]. All lineages that diverged prior to this polytomy are grouped into Branch 0 and those diverging after form Branches 1-4. Because this multifurcation plays such a central role in this system, there is great interest in estimating its timing and geographic origins [[9](#ref-qLAL8Don)]. However, the epidemiological significance of the “Big Bang” polytomy remains unclear, as no definitive phenotype has been identified that correlates with the observed branching pattern [**???**]. While the major branch system excels at reconstructing the evolutionary relationships between candidate populations, it struggles to connect these relationships to other biological changes.

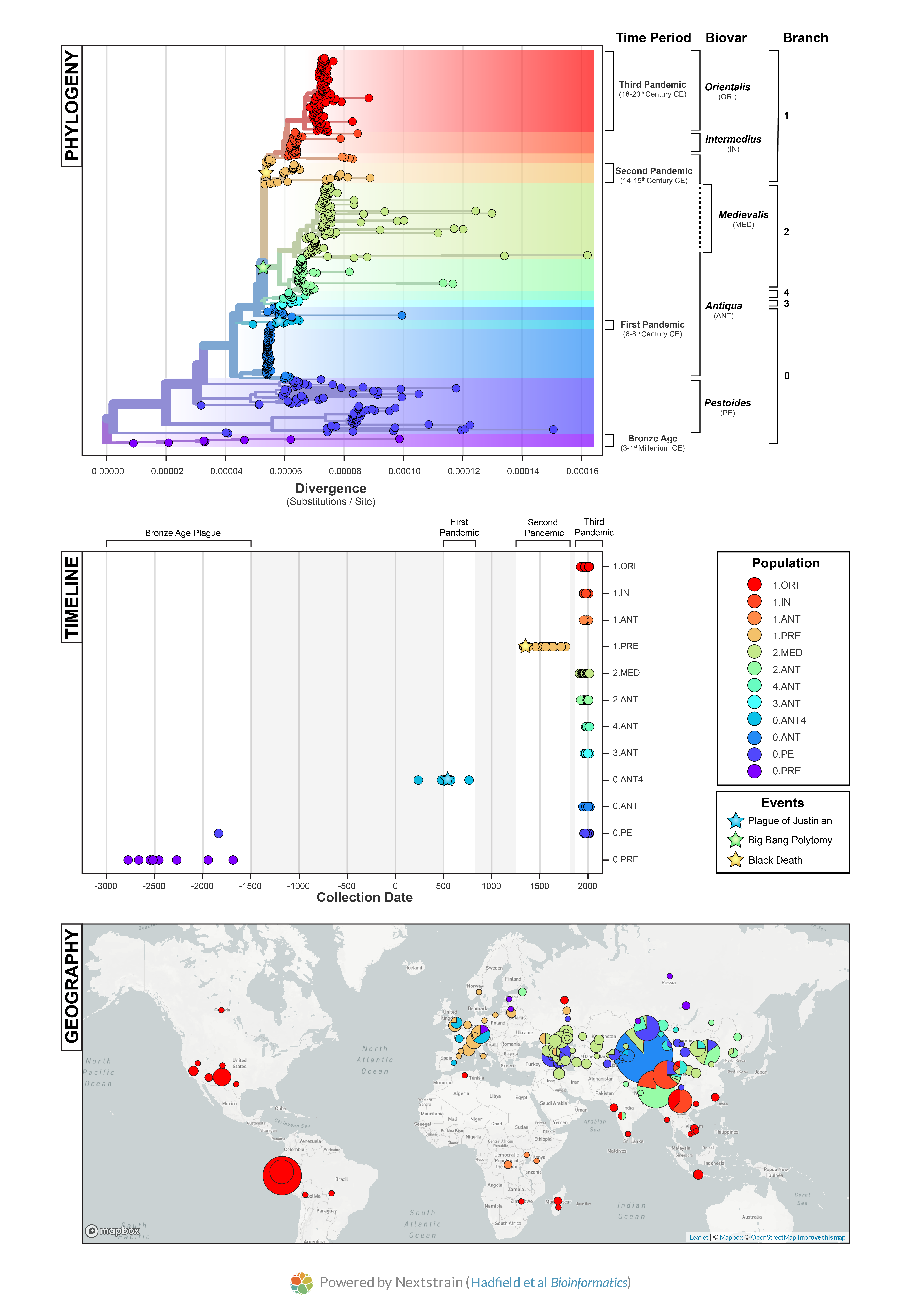


Figure 1: The phylogenetic, temporal, and geographic distribution of *Yersinia pestis* genomes. Top: The maximum-likelihood phylogeny. Middle: The timeline of collection dates. Bottom: The global geographic distribution.

#### Time Period

As previously mentioned, the sequencing of ancient *Y. pestis* poses a problem for classification, as direct metabolic testing is impossible for these non-viable samples. Ancient DNA researchers thus use an alternative strategy, by incorporating contextual evidence such as the collection date or associated time period. The known genetic diversity of *Y. pestis* has been most commonly divided into four time periods: the Bronze Age (3rd to 1st millennium BCE) [[8](#ref-ZvyYZ3sx)] , the First Pandemic, (6thto 8th century CE) [[12](#ref-ACt53Sow)], the Second Pandemic (14th to 18th century CE) [[13](#ref-1CPIgshmC)], and the Third Pandemic (19th to 20th century CE) [[11](#ref-ONhNS9aO)].

The key strength of this nomenclature is that it provides a foundation for interdisciplinary discourse. However this system runs several risks, like grouping unrelated populations, as contemporaneous strains have been observed to have distinct evolutionary histories [[20](#ref-mkJFtJz5)]. Furthermore, there is growing awareness of the temporal overlap of the Second and the Third Pandemics. Previously, the temporal extents of these events were mutually exclusive, dating from the 14th to 18th century, and the late 19th to mid-20th century respectively [[21](#ref-MoWKnwLZ)]. Recent historical scholarship has contested this claim, and demonstrated that these constraints are a product of a Eurocentric view of plague [[22](#ref-1FNi5AFJf)]. The Second Pandemic is now recognized to have extended into at least the 19th Century [[23](#ref-AbODfcWU),[24](#ref-137GGjfp5)] and the Third Pandemic is hypothesized to have began as early as the 18th century in southern China [[25](#ref-RWkOSyzX)]. Unfortunately, this period of overlap remains genomically unsampled, thus it is unclear where exactly to draw a genetic distinction, if it even exists, between these pandemic events.

Another limitation of this system is that several populations are curiously excluded from the time period nomenclature which emphasizes historically documented pandemics. For example, Branch 2 populations emerged at the same time as, but separate from, the Second Pandemic and have been associated with high mortality epidemics [[26](#ref-3aFEkVra)]. In particular, the *medievalis* population (2.MED) has spread throughout Asia (Figure [1](#fig:divtree_map_timeline)) and was observed to have the fastest spread velocity of any *Y. pestis* lineage [[25](#ref-RWkOSyzX)]. Given this epidemiological significance, it is surprising that Branch 2 populations have been largely overlooked in the pandemic taxonomy of *Y. pestis*. As ancient DNA sampling strategies expand in geographic scope, it will be important to consider how to adapt and expand the historical period nomenclature to encompass this new diversity.

#### Uncertainty

In light of this uncertainty and inconsistencies, no classification system comprehensively represents the global population structure of *Y. pestis*. Instead, integrative approaches have been previously used in large comparative studies of *Y. pestis* [[11](#ref-ONhNS9aO),[27](#ref-Y4CIJOZW)]. We therefore take the intersection of the three taxonomic systems discussed previously, to define 12 populations for statistical analysis (Figure [1](#fig:divtree_map_timeline) Legend). In the following sections, we highlight the novel insight and issues that arise when this population structure is incorporated into molecular clock models.

### Molecular Clock

We began our temporal analysis by exploring the extent of rate variation present in our updated genomic dataset, which is notably larger and more diverse than those used in previous studies [[13](#ref-1CPIgshmC)]. Given this expanded diversity, it is unsurprising that a root-to-tip regression on collection date reproduces the finding that substitution rates in *Y. pestis* are poorly represented by a simple linear model or “strict clock” (Figure [2](#fig:rtt)). While there is a statistically significant relationship between collection date and genetic distance to the root (P-value=4.959 x 10-13), an extremely low coefficient of determination (R2=0.09) indicates there is tremendous variation that is not accounted for.

Thus far, the hypotheses proposed to explain this variation have primarily focused on ecological processes, such as the cycling between endemic and epidemic phases [[11](#ref-ONhNS9aO)] and geographic expansions over large distances [[13](#ref-1CPIgshmC)]. However, we argue that methodological factors must first be taken into account, before investigating more complex ecological factors such as host and landscape.

#### Time Dependency

One striking methodological factor is the time dependency of molecular rates. In Figure [2](#fig:rtt), we show how rate variation in *Yersinia pestis* correlates with the sampling time frame, as populations sampled over several millennia (Bronze Age) appear more “clock-like” than those sampled over multiple centuries (Second Pandemic) or decades (*medievalis*). This correlation is a well-known and widely-documented phenomenon in many organisms [[28](#ref-orBi9QMY)] and occurs due to two conflicting signals: a slower, long-term substitution rate combined with a higher, short-term mutation rate.

Separating out these signals can be extremely challenging and failure to do so has significant consequences when estimating and interpreting a molecular clock. Of particular concern for epidemiological investigations is the risk of artificially inflating the substitution rate due to transient mutations, which can lead to younger node dates. With regards to plague genomics, this may result in incorrect molecular dates linked to key historical events, such as the emergence of pandemic populations. Because of this risk, we first evaluate the presence of spurious mutations in our dataset before attempting to estimate a molecular clock model.

*Y. pestis* is particularly susceptible to the time-dependency of molecular rates, as it has one of the slowest substitution rates observed among bacterial pathogens [[14](#ref-1cTipQcd)]. The substitution rate of *Y. pestis* has previously been estimated to range from 1 x 10-8 to 2 x 10-8 substitutions/site/year [[11](#ref-ONhNS9aO),[13](#ref-1CPIgshmC)], or 1 substitution every 10 to 25 years. In application, this means that *Y. pestis* lineages often cannot be differentiated until multiple decades have passed, a concept generally referred to as the phylodynamic threshold [[29](#ref-hcvZW0sG)].

A historical example of this can be seen during the Second Pandemic, where isolates dated to the medieval Black Death (1348-1353) are nearly indistinguishable clones (Figure [2](#fig:rtt)). A modern example is the *medievalis* population, where the youngest samples (2010s) have diverged little compared to those from a century prior (1910s). This highlights a significant limitation and cautionary note for *Y. pestis* phylogenetics, as comparisons over short time scales (10 to 100 years) may have limited resolving power. Furthermore, the little phylogenetic signal that accumulates in the population may be easily obscured or biased by spurious mutations in a single sample.

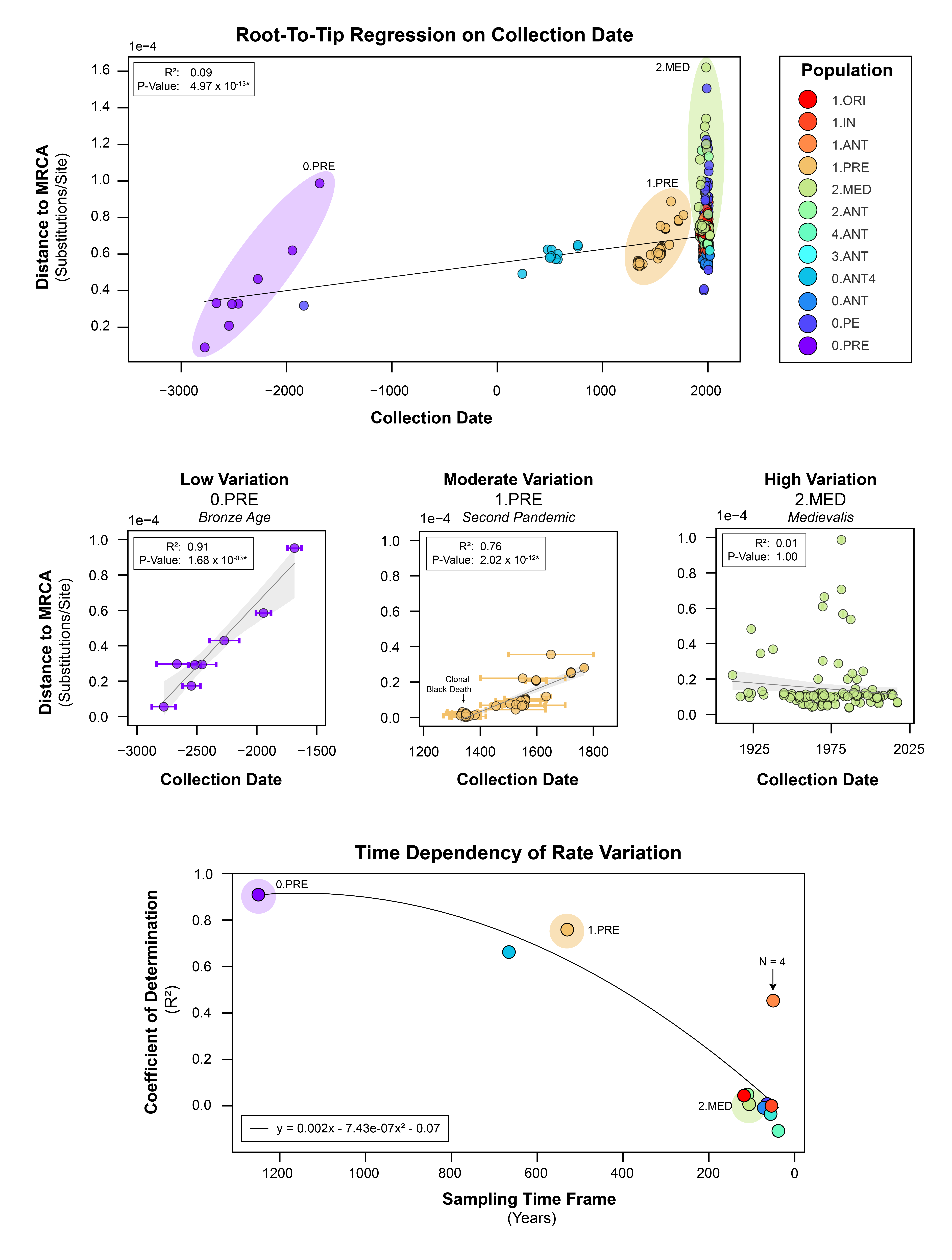


Figure 2: Rate variation in *Yersinia pestis* as observed through a regression of root-to-tip distances on collection date. Top: A species-wide model using all genomes from the maximum-likelihood phylogeny. Middle: Population-specific models based on extracted subtrees from the phylogeny. Bottom: The time-dependency of population-specific rate variation on the sampling time frame.

The *medievalis* population is an extreme example of mutational bias, as several samples collected between 1970 and 1980 are exceptionally divergent (Figure [2](#fig:rtt)). This short-term diversity is largely due to mutations observed in only a single isolate, which manifest as long terminal branches in the maximum-likelihood phylogeny (Figure [1](#fig:divtree_map_timeline),). These mutations may be the result of true biological variation occurring in a wild isolate or false-positive variants due to sequencing error. Given the extensive presence of these potential outliers (SI Figure [6](#fig:long_branches)), inclusion or exclusion of these samples may have profound impacts on clock model parameters.

#### Detecting Temporal Signal

As the root-to-tip regression revealed poor support for a strict clock (Figure [2](#fig:rtt)), our next step was to model the observed rate variation using a relaxed clock. To investigate the degree of temporal structure in *Y. pestis*, we performed a Bayesian evaluation of temporal signal (BETS) test [[30](#ref-zikRADit)]. In brief, this method compares the likelihoods of two different phylogenetic models, one where the true collection dates are used and the other where all collection dates are assumed to be contemporaneous. A comparison of the model likelihoods, or Bayes factors, is then used to assess the degree of temporal signal in the dataset. We performed this test using both a strict clock and an uncorrelated lognormal (UCLN) relaxed clock, and compared their performance in species-wide and population-specific models.

#### Species Clock

The BETS test was inconclusive when attempting to estimate a single clock for all *Y. pestis* populations combined. The Bayesian analysis exhibited poor sampling of the clock parameters, for both a strict and relaxed clock, even when attempting to reduce sources of variation such as decreasing the number of genomes, using fixed tip dates, and fixing the tree topology. This poor performance is consistent with previous analyses [[12](#ref-ACt53Sow)] where robust estimates of model parameters could not be estimated, thus leading to the conclusion that *Y. pestis* lacks temporal signal. We therefore conclude that a single, species-wide clock is not statistically viable to model the global diversity of *Y. pestis*. Furthermore, this approach produced the lowest mean rate (1.6 x 10-8 subs/site/year) of any model tested in this study (Figure [3](#fig:mean_rate_tree_shape)), with the greatest uncertainty surrounding the time to the most recent common ancestor (Figure [4](#fig:tmrca)). We hypothesize that this model misspecification may help explain *Y. pestis* node-dating disparities between studies [[8](#ref-ZvyYZ3sx),[11](#ref-ONhNS9aO)].

#### Population Clocks

In contrast to the species-wide model, separating the genomic dataset by population dramatically stabilized the Bayesian analysis. Temporal signal was detected in 9 out of 12 populations (SI Table [1](#tbl:bets)) and all clock model parameters were well-sampled with effective sample sizes (ESS) greater than 200. In the following sections, we explore whether this approach is methodologically robust and which *Y. pestis* populations are associated with high confidence node-dates.

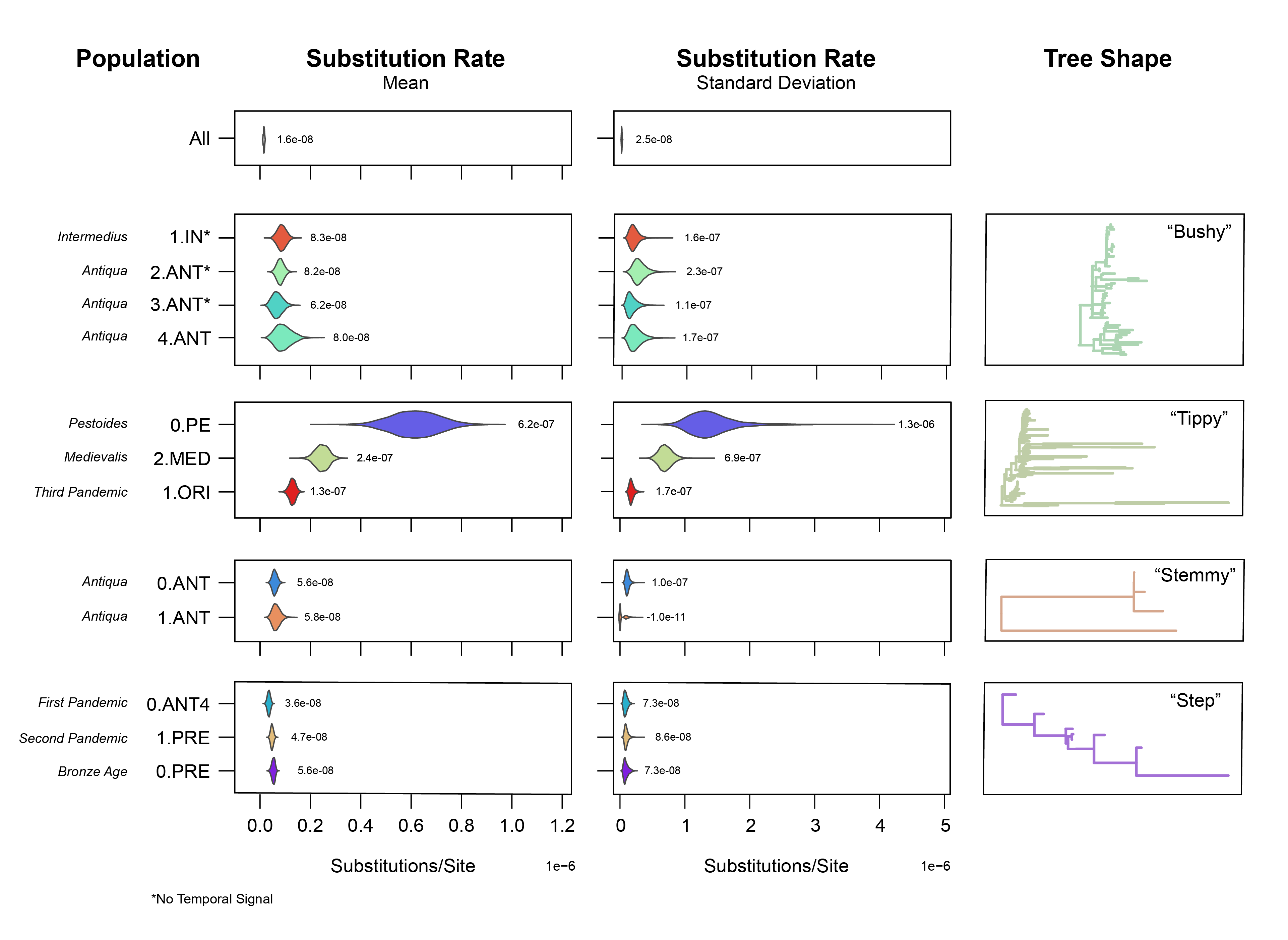


Figure 3: The mean and standard deviation of substitution rates estimated using a relaxed clock method. Each distribution is annotated with the peak value.

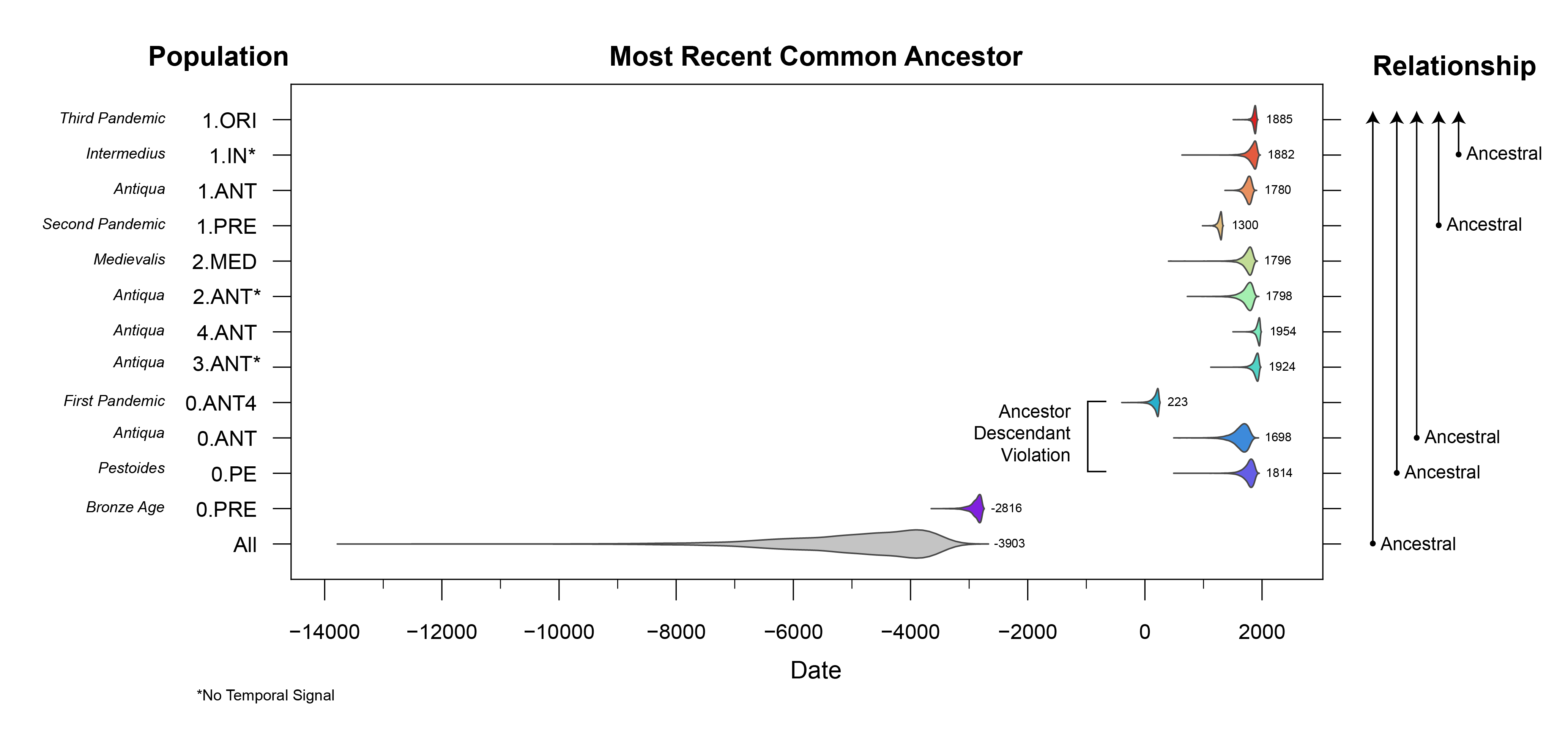


Figure 4: The time to most recent common ancestor estimated using a relaxed clock method. Each distribution is annotated with the peak value.

##### No Detectable Signal

Three populations had no detectable temporal signal: *intermedius* (1.IN) and the *antiqua* populations from Branch 2 (2.ANT) and Branch 3 (3.ANT). According to our results, insufficient substitutions have accumulated in these populations to be considered measurably-evolving, and thus their associated rates and dates are unreliable. Overall, a lack of temporal signal tended to be associated with relatively short and “bushy” trees in the maximum-likelihood phylogeny (Figure [3](#fig:mean_rate_tree_shape)).

##### Inflated Rates and Young Dates

For populations with detectable temporal signal, several stood out as potential outliers. Specifically, the *pestoides* (0.PE),*medievalis* (2.MED), and the Third Pandemic (1.ORI) had mean substitution rates more than the double that of the other populations. Overall we observed that these high substitution rates were associated with long and “tippy” trees. Given the large number of long external branches present in the maximum-likelihood phylogeny (SI Figure [6](#fig:long_branches)), we hypothesize that the faster rates associated with these populations are biased by an increased prevalence of transient mutations. If true, then the node-dates of these populations will be overly young.

Tentative support for this hypothesis can be found by searching for violations of ancestor-descendant dates. The most extreme example is the *pestoides* (0.PE) group, where the root is incorrectly dated to 1814 when it should instead pre-date the First Pandemic (0.ANT4) (Figure [4](#fig:tmrca)). In light of this inconsistency, we caution that node dates associated with these populations should be treated as highly suspect, and are likely underestimates of the true divergence dates (ie. too young).

The remaining populations with temporal signal had much slower, and overlapping, mean substitution rates that ranged from 3.6 to 5.6 x 10-8 subs/site/year. Nevertheless, we still observed several node-dating incongruencies, as the root date of the *antiqua* population 0.ANT was estimated to be 1698 which again, incorrectly post-dates the First Pandemic (0.ANT4). We attribute this incongruency to a “stemmy” topology, where rate variation is poorly modeled in the long, internal branches near the root. The connection of “stemmy” trees with overly young dates has been previously observed and linked to use of the UCLN relaxed clock model [[31](#ref-1F6c3FFf8)]. Correcting the incorrect dates was possible, but required significant model changes for the affected populations. We thus propose that studies seeking to robustly date nodes associated with “stemmy” plague populations, such as 0.ANT and 1.ANT, will require significant model changes, such as substituting the uncorrelated relaxed clock for the random local clock [[31](#ref-1F6c3FFf8)].

The most robust clock estimates were associated with the pre-modern populations from the Bronze Age (0.PRE), the First Pandemic (0.ANT4), and the Second Pandemic (1.PRE). These populations are associated with “step” trees, in which historical samples help to break up long internal branches. These calibrating samples are absolutely critical for estimating divergence times [[32](#ref-1ESS7L1hK)], especially for organisms such as *Y. pestis* which is known to experience sudden and extreme rate changes [[11](#ref-ONhNS9aO),[13](#ref-1CPIgshmC)]. It is therefore not surprising that these populations, which have been rigorously sampled across multiple centuries, have robust and measurable temporal signal.

What is surprising, is that…

## Discussion

* A single, species-model is not viable.
* Sample curation is needed! Tippy populations.
* Calibrations are critical.
* Node dates are generally too young.
* However, as these outlier populations have high epidemiological significance we suggest an important avenue for future research will be to systematically evaluate the genomes associated with long branches and the impacts of their removal.

The use of population-specific clocks recovered temporal signal in 9 out of 12 *Y. pestis* populations. However, this methodological improvement came at the cost of

* Optimizing problem, of trying to retain as many calibration points as possible, while keeping the rate variation to a reasonable level.
* We need to limit the amount of variation in the clock model, but also retain as many calibrating samples as possible.
* In the species model, the main problem was underestimating the rate and overestimating divergence dates (ie. too old). This was due to substantial rate variation beyond what could be modeled using an UCLN relaxed clock.
* In contrast, population-specific models faced the opposite problem. Substitution rates were frequently overestimated producing …
* there is such a high degree of rate variation in *Y. pestis*,
* With population models, the main problem was overestimating the rate and underestimating divergence dates (ie. too young).

## Methods

## Methods

A visual overview of the computational methods is provided in Figure [5](#fig:workflow) and is available as a snakemake pipeline (https://github.com/ktmeaton/plague-phylogeography/).

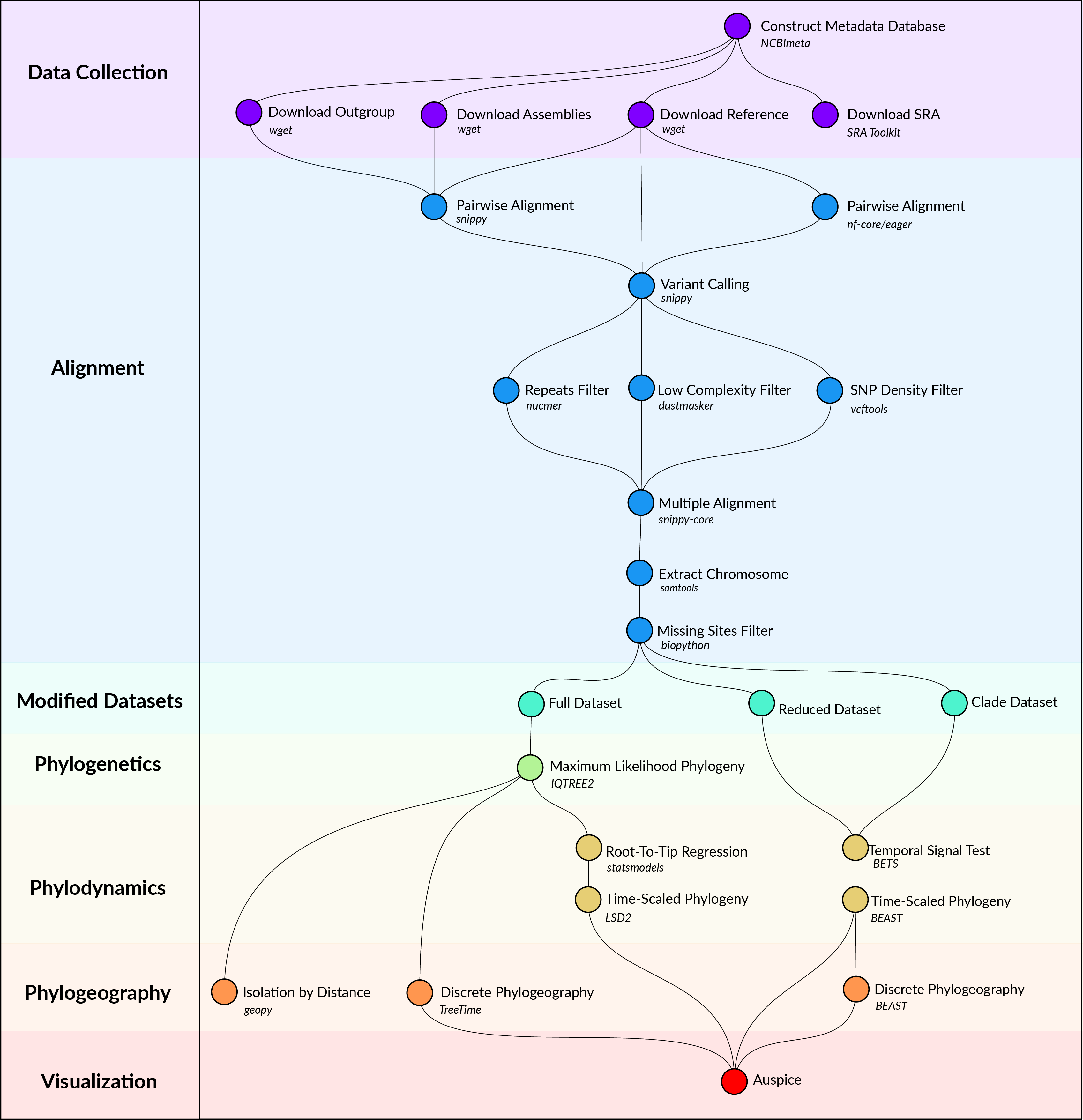


Figure 5: Computational methods workflow.

### Data Collection

*Y. pestis* genome sequencing projects were retrieved from the NCBI databases using NCBImeta [[33](#ref-131fQLiJt)]. 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.

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 [[34](#ref-hxWSqZgm)] and the Nominatim API [[35](#ref-M15a78x5)] for OpenStreetMap [[36](#ref-5rCiNisz)]. 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 at the taxonomic level of Order.

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 [[37](#ref-1DR126iIZ)]. 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 [[38](#ref-puYDXtJ9)]. Pre-processing and alignment to the reference genome was performed using the nf-core/eager pipeline, a reproducible workflow for ancient genome reconstruction [[39](#ref-17yD9OrGW)]. Ancient genomes were removed if the number of sites covered at a minimum depth of 3X was less than 70% of the reference genome.

A multiple sequence alignment was constructed using the Snippy Core module of the Snippy pipeline [[37](#ref-1DR126iIZ)]. 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 is referred to as the *clade* dataset. Clade labeling was derived from the five-branch population structure accompanied by a biovar abbreviation [[11](#ref-ONhNS9aO)]. 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.

To improve the performance and convergence of Bayesian analysis, a subsampled dataset was constructed and is 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 191 representative samples.

### Maximum-Likelihood Phylogeny

Model selection was performed using Modelfinder which identified the K3Pu+F+I model as the optimal choice based on the Bayesian Information Criterion (BIC) [[40](#ref-QZIPWLUx)]. A maximum-likelihood phylogeny was then estimated across 10 independent runs of IQTREE [[41](#ref-mkkgRhHT)]. Branch support was evaluated using 1000 iterations of the ultrafast bootstrap approximation, with a threshold of 95% required for strong support [[42](#ref-12SvE6y3A)].

### Bayesian Evaluation of Temporal Signal

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 using the python package statsmodels [[43](#ref-11e70fLkA)]. Given the relative simplicity of a regression model, the *full* dataset of 601 genomes was used. For the second test of temporal signal, a Bayesian Evaluation of Temporal Signal (BETS) was conducted. As the complexity of Bayesian modeling is computationally intensive, the *reduced* dataset (N=191) was used.

## 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](https://doi.org/10.1016/j.cub.2017.10.025) · PMID: [29174893](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1016/j.cell.2018.11.005) · PMID: [30528431](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1038/s41598-017-13481-2) · PMID: [29021541](https://www.ncbi.nlm.nih.gov/pubmed/29021541) · PMCID: [PMC5636801](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5636801)

4. **Yersinia pestis–etiologic agent of plague**   
R. D. Perry, J. D. Fetherston  
*Clinical Microbiology Reviews* (1997-01)   
PMID: [8993858](https://www.ncbi.nlm.nih.gov/pubmed/8993858) · PMCID: [PMC172914](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://worldcat.org/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. **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](https://doi.org/10.1016/j.cell.2015.10.009) · PMID: [26496604](https://www.ncbi.nlm.nih.gov/pubmed/26496604) · PMCID: [PMC4644222](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644222)

9. **How a microbe becomes a pandemic: a new story of the Black Death**   
Monica H. Green  
*The Lancet Microbe* (2020-12-01) <https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(20)30176-2/abstract>   
DOI: [10.1016/s2666-5247(20)30176-2](https://doi.org/10.1016/S2666-5247(20)30176-2)

10. **A draft genome of *Yersinia pestis* from victims of the Black Death**   
Kirsten I. Bos, Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, Sharon N. DeWitte, Matthias Meyer, Sarah Schmedes, … Johannes Krause  
*Nature* (2011-10) <http://www.nature.com/articles/nature10549>   
DOI: [10.1038/nature10549](https://doi.org/10.1038/nature10549)

11. **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](https://doi.org/10.1073/pnas.1205750110)

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](https://doi.org/10.1016/S1473-3099(13)70323-2)

13. **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](https://doi.org/10.1038/s41467-019-12154-0)

14. **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](https://doi.org/10.1099/mgen.0.000094) · PMID: [28348834](https://www.ncbi.nlm.nih.gov/pubmed/28348834) · PMCID: [PMC5320706](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320706)

15. **Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter?**   
Eleanor M. L. Scerri, Mark G. Thomas, Andrea Manica, Philipp Gunz, Jay T. Stock, Chris Stringer, Matt Grove, Huw S. Groucutt, Axel Timmermann, G. Philip Rightmire, … Lounès Chikhi  
*Trends in Ecology & Evolution* (2018-08) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6092560/>   
DOI: [10.1016/j.tree.2018.05.005](https://doi.org/10.1016/j.tree.2018.05.005) · PMID: [30007846](https://www.ncbi.nlm.nih.gov/pubmed/30007846) · PMCID: [PMC6092560](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6092560)

16. **Phylogeny and classification of *Yersinia pestis* through the lens of strains From the plague foci of Commonwealth of Independent States**   
Vladimir V. Kutyrev, Galina A. Eroshenko, Vladimir L. Motin, Nikita Y. Nosov, Jaroslav M. Krasnov, Lyubov M. Kukleva, Konstantin A. Nikiforov, Zhanna V. Al’khova, Eugene G. Oglodin, Natalia P. Guseva  
*Frontiers in Microbiology* (2018-05-25) <https://www.frontiersin.org/article/10.3389/fmicb.2018.01106/full>   
DOI: [10.3389/fmicb.2018.01106](https://doi.org/10.3389/fmicb.2018.01106)

17. **Variétés de l’espèce Pasteurella pestis**   
R. Devignat  
*Bulletin of the World Health Organization* (1951) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2554099/>   
PMID: [14859080](https://www.ncbi.nlm.nih.gov/pubmed/14859080) · PMCID: [PMC2554099](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2554099)

18. **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](https://doi.org/10.1016/j.micinf.2004.08.002)

19. **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](https://doi.org/10.1371/journal.pone.0006000)

20. **Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague**   
Maria A. Spyrou, Rezeda I. Tukhbatova, Chuan-Chao Wang, Aida Andrades Valtueña, Aditya K. Lankapalli, Vitaly V. Kondrashin, Victor A. Tsybin, Aleksandr Khokhlov, Denise Kühnert, Alexander Herbig, … Johannes Krause  
*Nature Communications* (2018-06-08) <https://www.nature.com/articles/s41467-018-04550-9>   
DOI: [10.1038/s41467-018-04550-9](https://doi.org/10.1038/s41467-018-04550-9)

21. **Natural history of plague: perspectives from more than a century of research**   
Kenneth L. Gage, Michael Y. Kosoy  
*Annual Review of Entomology* (2005)   
DOI: [10.1146/annurev.ento.50.071803.130337](https://doi.org/10.1146/annurev.ento.50.071803.130337) · PMID: [15471529](https://www.ncbi.nlm.nih.gov/pubmed/15471529)

22. **New Science and Old Sources: Why the Ottoman Experience of Plague Matters**   
Nükhet Varlık  
*The Medieval Globe* (2014) <https://scholarworks.wmich.edu/tmg/vol1/iss1/9/>

23. **The Ottomans During the Global Crises of Cholera and Plague: The View from Iraq and the Gulf.**   
Isacar A. Bolaños  
*International Journal of Middle East Studies* (2019-11) <http://www.cambridge.org/core/journals/international-journal-of-middle-east-studies/article/ottomans-during-the-global-crises-of-cholera-and-plague-the-view-from-iraq-and-the-gulf/82A6A0E7CF8BD2138FA230AE4189057E>   
DOI: [10.1017/s0020743819000667](https://doi.org/10.1017/S0020743819000667)

24. **The plague that never left: restoring the Second Pandemic to Ottoman and Turkish history in the time of COVID-19**   
Nükhet Varlık  
*New Perspectives on Turkey* (2020-11) <https://www.cambridge.org/core/journals/new-perspectives-on-turkey/article/plague-that-never-left-restoring-the-second-pandemic-to-ottoman-and-turkish-history-in-the-time-of-covid19/AE81F48AD6FEB71C42AAA2115A0307A9>   
DOI: [10.1017/npt.2020.27](https://doi.org/10.1017/npt.2020.27)

25. **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](https://doi.org/10.1098/rspb.2013.3159)

26. **Evolution and circulation of Yersinia pestis in the Northern Caspian and Northern Aral Sea regions in the 20th-21st centuries**   
Galina A. Eroshenko, Nikolay V. Popov, Zhanna V. Al’khova, Lyubov M. Kukleva, Alina N. Balykova, Nadezhda S. Chervyakova, Ekaterina A. Naryshkina, Vladimir V. Kutyrev  
*PLOS ONE* (2021-02-11) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244615>   
DOI: [10.1371/journal.pone.0244615](https://doi.org/10.1371/journal.pone.0244615)

27. ***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](https://doi.org/10.1038/ng.705) · PMID: [21037571](https://www.ncbi.nlm.nih.gov/pubmed/21037571) · PMCID: [PMC2999892](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2999892)

28. **Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times**   
Simon Y. W. Ho, Matthew J. Phillips, Alan Cooper, Alexei J. Drummond  
*Molecular Biology and Evolution* (2005-07-01) <https://doi.org/10.1093/molbev/msi145>   
DOI: [10.1093/molbev/msi145](https://doi.org/10.1093/molbev/msi145)

29. **Temporal signal and the phylodynamic threshold of SARS-CoV-2**   
Sebastian Duchene, Leo Featherstone, Melina Haritopoulou-Sinanidou, Andrew Rambaut, Philippe Lemey, Guy Baele  
*Virus Evolution* (2020-07-01) <https://doi.org/10.1093/ve/veaa061>   
DOI: [10.1093/ve/veaa061](https://doi.org/10.1093/ve/veaa061)

30. **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](https://doi.org/10.1093/molbev/msaa163)

31. **Clock model makes a large difference to age estimates of long-stemmed clades with no internal calibration: a test using Australian grasstrees**   
Michael D. Crisp, Nate B. Hardy, Lyn G. Cook  
*BMC Evolutionary Biology* (2014-12-19) <https://doi.org/10.1186/s12862-014-0263-3>   
DOI: [10.1186/s12862-014-0263-3](https://doi.org/10.1186/s12862-014-0263-3)

32. **Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times**   
Simon Y. W. Ho, Matthew J. Phillips  
*Systematic Biology* (2009-06-01) <https://doi.org/10.1093/sysbio/syp035>   
DOI: [10.1093/sysbio/syp035](https://doi.org/10.1093/sysbio/syp035)

33. **NCBImeta**   
Katherine Eaton  
*NCBImeta* (2019) <https://github.com/ktmeaton/NCBImeta>

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

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

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

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

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

39. **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](https://doi.org/10.7717/peerj.10947)

40. **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](https://doi.org/10.1038/nmeth.4285)

41. **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](https://doi.org/10.1093/molbev/msaa015)

42. **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](https://doi.org/10.1093/molbev/msx281)

43. **Statsmodels: Econometric and Statistical Modeling with Python**   
Skipper Seabold, Josef Perktold  
(2010) <https://conference.scipy.org/proceedings/scipy2010/seabold.html>   
DOI: [10.25080/majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011)

## Supplementary Information

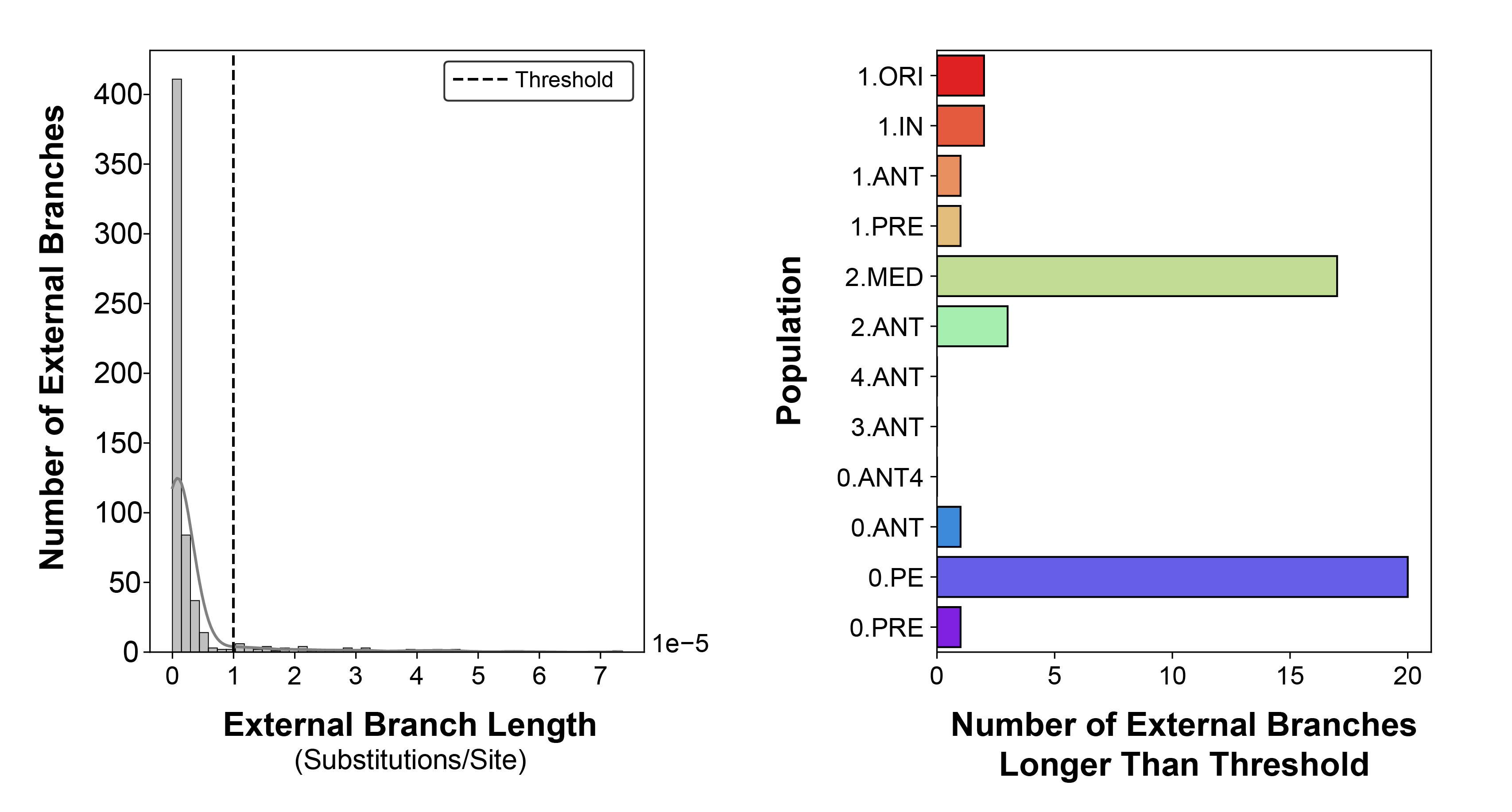


Figure 6: The distribution of external branch lengths across the maximum-likelihood phylogeny. The threshold to be considered a long external branch is set at 1e-5 substitutions/site.

Table 1: Model selection and log marginal likelihoods obtained from a Bayesian evaluation of temporal signal (BETS) test. \*0.PRE had temporal signal according to a strict clock, although the relaxed clock with no dates model had the highest likelihood.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Population | N | Time Span (Years) | Best Model | Bayes Factor | Strict Clock No Dates | Strict Clock Dates | Relaxed Clock No Dates | Relaxed Clock Dates |
| 1.ORI | 117 | 118 | Relaxed Clock | 36 | -5899691 | -5899661 | -5899601 | -5899566 |
| 1.IN | 39 | 54 | – | -10 | -5891399 | -5891403 | -5891344 | -5891355 |
| 1.ANT | 4 | 50 | Relaxed Clock | 13 | -5882596 | -5882587 | -5882595 | -5882582 |
| 1.PRE | 40 | 530 | Relaxed Clock | 44 | -5888140 | -5888130 | -5888082 | -5888038 |
| 2.MED | 116 | 106 | Relaxed Clock | 4 | -5920837 | -5920733 | -5919662 | -5919658 |
| 2.ANT | 54 | 110 | – | -13 | -5892876 | -5892895 | -5892791 | -5892805 |
| 4.ANT | 11 | 38 | Relaxed Clock | 4 | -5886031 | -5886034 | -5886026 | -5886022 |
| 3.ANT | 11 | 56 | – | -11 | -5887497 | -5887506 | -5887495 | -5887506 |
| 0.ANT4 | 12 | 666 | Relaxed Clock | 6 | -5889526 | -5889520 | -5889502 | -5889496 |
| 0.ANT | 103 | 72 | Relaxed Clock | 13298 | -5896014 | -5896016 | -5895880 | -5882582 |
| 0.PE | 85 | 64 | Relaxed Clock | 12 | -5945603 | -5945574 | -5944627 | -5944614 |
| 0.PRE | 8 | 1250 | Relaxed Clock | 83\* | -5892926 | -5892843 | -5892739 | -5892741 |