This manuscript was automatically generated on September 2, 2021.

Authors

Yersinia pestis Phylodynamics

Rate Variation: Biological Trait or Methodological Artifact?

Previous work has documented substantial rate variation both between and within populations of *Y. pestis* [1,2]. We therefore began by testing if this characteristic was still present in our updated genomic dataset, which is notably larger and more diverse than those used in previous studies.

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 (ie. strict clock) (Figure 1). While there is a statistically significant positive relationship between date and genetic distance (P-value=4.959 x 10^{-13}), an extremely low coefficient of determination (R^2=0.09) indicates there is tremendous variation that is not accounted for.

The rate variation observed in *Y. pestis* (Figure 1) presents a curious case of the time dependency of molecular rates [3]. Rate variation correlates with the sampling time frame, in which populations sampled over several millennia (`0.PRE`), have less variation than those sampled over centuries (`1.PRE`), or only a few decades (`2.MED`). We identify four inter-related mechanisms that drive the observed patterns of rate variation in *Y. pestis*:

1. A slow, long-term substitution rate.
2. A high, short-term mutation rate.
3. Methodological artifacts.
4. Population-specific rate variation.

Mechanisms of Rate Variation

Slow, Long-Term Substitution Rate

The substitution rate of *Y. pestis* has previously been estimated to range from 1 x 10^{-8} to 2 x 10^{-8} subs/site/year, [1,2] or 1 substitution every 10-25 years. Amongst bacterial pathogens, this is one of the slowest rates observed [4] and means that *Y. pestis* lineages often cannot be differentiated until several decades have passed. This question of how much time must pass before sufficient molecular change occurs is referred to as the phylodynamic threshold [5].

In application, we can see this in the finding that isolates from population `1.PRE` dated to the medieval Black Death (1348-1353) are indistinguishable clones, whereas those from subsequent centuries are phylogenetically distinct (Figure 1). This highlights a significant limitation of *Y. pestis* phylogenetics, as comparisons over short time scale (<10 years) have limited resolution and can be easily biased by noisy mutations.

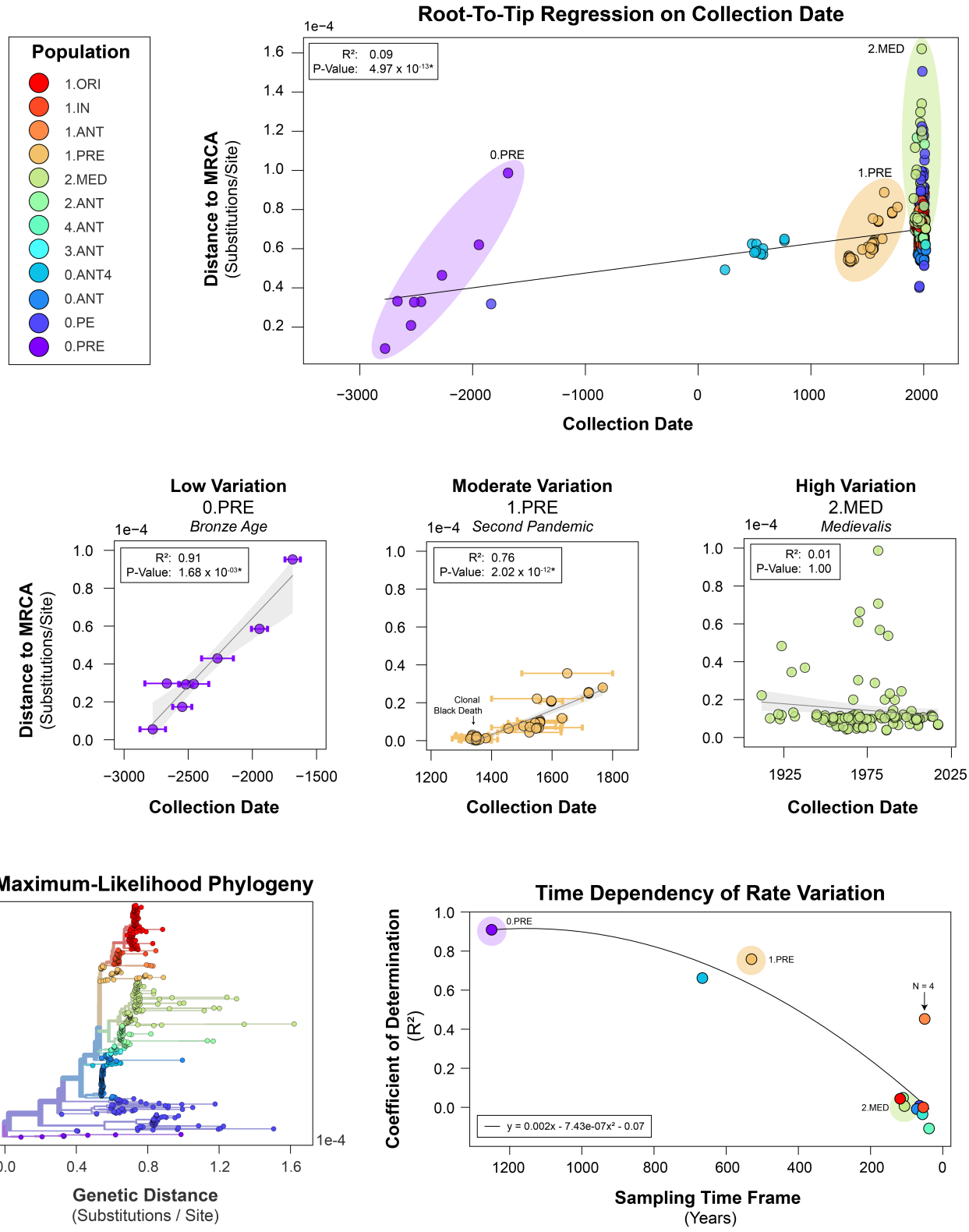

Figure 1: Rate variation in *Yersinia pestis* using root-to-tip regression.

High, Short-Term Mutation Rate

Since it can take decades for a substitution to become fixed in *Y. pestis* populations, rate estimates are highly susceptible to the influence of transient mutations. In whole-genome sequencing, it is common to capture both fixed substitutions in the population and transient mutations found in a single isolate. These singleton mutations manifest as long external branches, and may arise from "true" biological variation, particularly when a population is experiencing exponential growth and is sparsely sampled, or from methodological "artifacts" due to sequencing error.

The global phylogeny of *Y. pestis* is heavily impacted by these transient mutations or long external branches. While these outliers are found ubiquitously throughout the phylogeny, several populations are disproportionately affected including *orientalis* (`1.ORI`), *pestoides* (`0.PE`), *medievalis* (`2.MED`), and *intermedius* (`1.IN`) (Table 1). Given the extensive presence of these apparent outliers, inclusion or exclusion of these samples may have profound impacts on the models used to estimate population sizes, molecular clocks, and migration events.

Table 1: Long branch outliers across *Y. pestis* populations.

Population	Samples	Outliers	% Outliers
1.ORI	116	37	32
1.IN	39	3	8
1.ANT	4	0	0
1.PRE	40	0	0
2.MED	116	11	9
2.ANT	54	3	6
4.ANT	11	0	0
3.ANT	11	0	0
0.ANT4	12	0	0
0.ANT	103	1	1
0.PE	86	20	23
0.PRE	8	0	0

Methodological Artifacts

One way to separate out these conflicting signals is by identifying deviant mutation patterns. Notably, we observe a correlation between branch length and extreme values of the transition to transversion ratio (TsTv) (Figure 4). As low TsTv ratios have previously been associated with false positive variants [6], we hypothesize these outliers represent methodological artifacts.

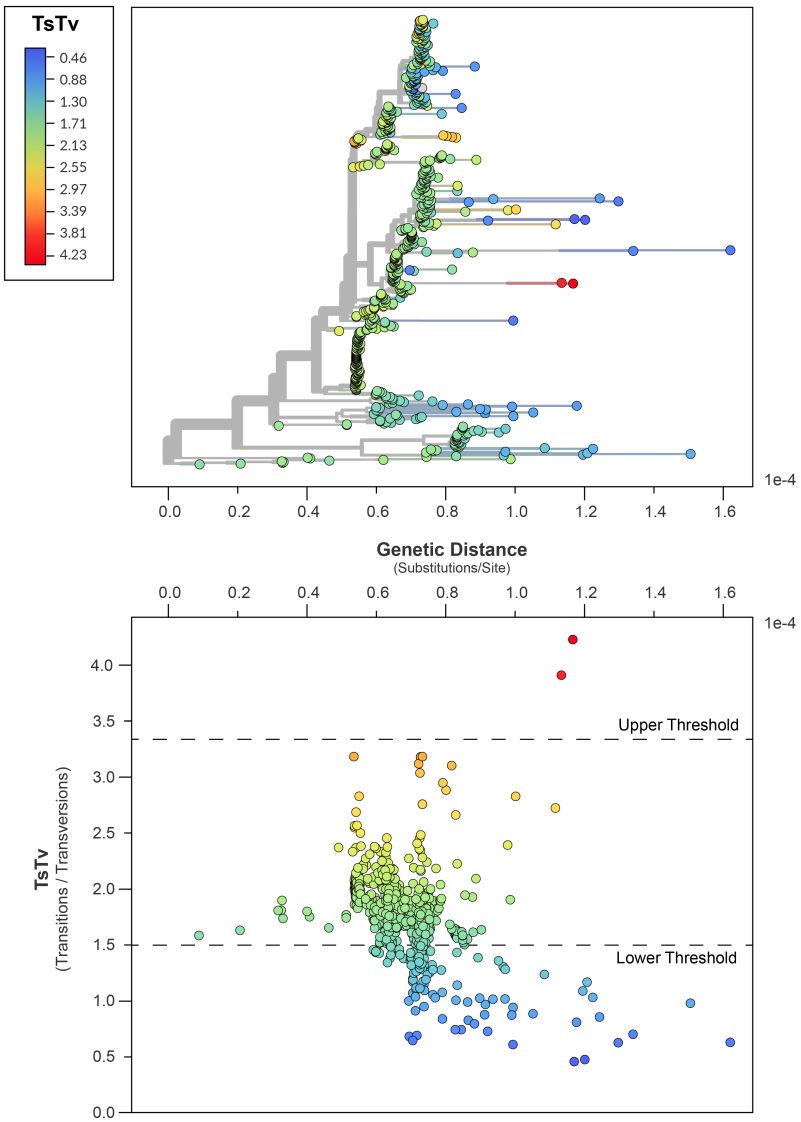

Figure 2: Long external branches in the global *Y. pestis* phylogeny are associated with extreme TsTv ratios.

Population-Specific Rate Variation

To Be Done!

Consequences

1. **A species-wide molecular clock analysis is highly unstable.**

   - The MCMC shows poor mixing and fails to converge at an estimate for key model parameters such as the mean substitution rate.
   - Eliminating tip-date uncertainty and fixing the tree topology do not improve the model.

2. **Analyzing populations in isolation stabilizes the molecular clock analysis.**

   - Temporal signal detected in 9/12 populations.
   - Good mixing and convergence.

3. **Populations with long branch outliers have higher mean substitution rates.**

   - The rates of `0.PE`, `2.MED`, `1.ORI`, and `1.IN` are "artificially" inflated by samples with false-positive variants. tMRCA estimates for these populations will be too young!
   - Populations without false positive samples have overlapping estimates of the mean substitution rate and standard deviation.
   - Mean substitution rate ranges from 3×10^{-8} to 8×10^{-8}.

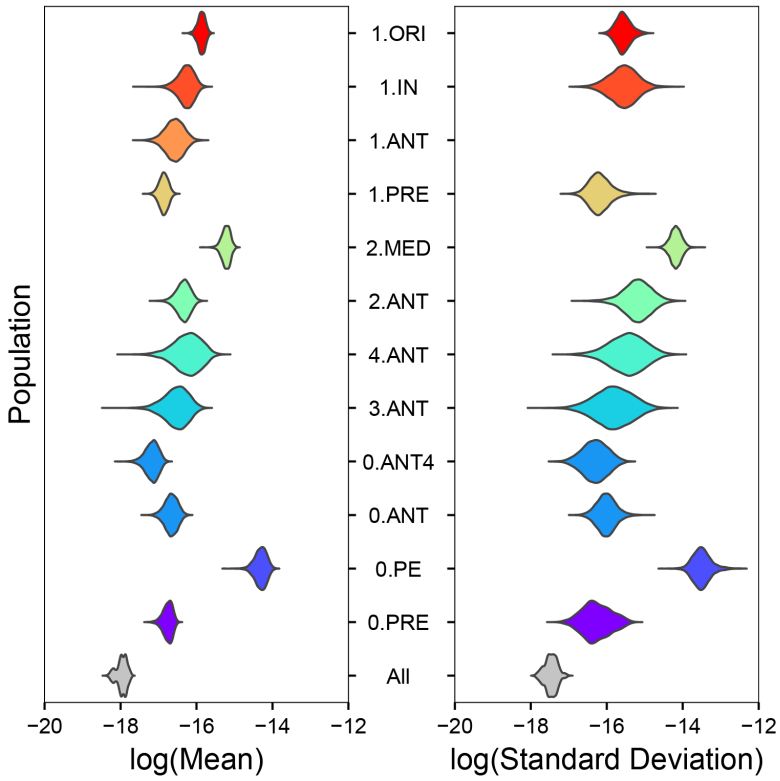

Figure 3: Mean substitution rate and standard deviation.

4. **The mean substitution rate of *Y. pestis* has been considerably under-estimated.**

   - Previously thought to be $1x10^{-8}$. No population is observed to have a rate that slow.
   - tMRCA estimates based on this rate will be too old!

5. **Removing long branch outliers drastically changes phylogeography patterns.**

   - Filtering out long branches recovers a statistically significant pattern of isolation-by-distance (IBD) for almost all populations!

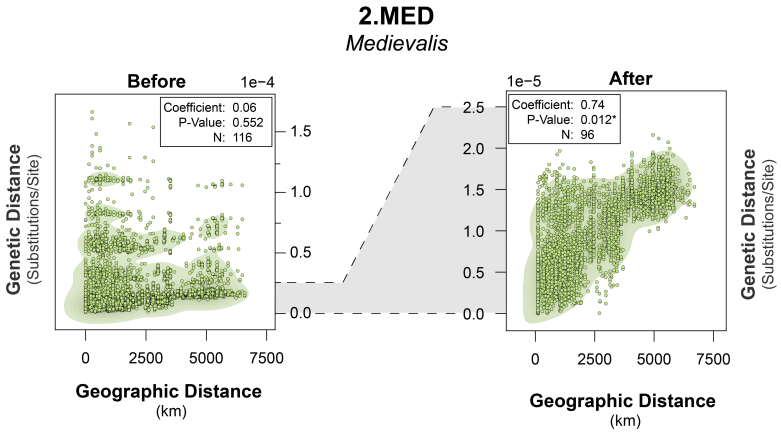

Figure 4: Isolation by distance of `2.MED` before and after long-branch filtering.

References

1. **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

2. **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

3. **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

4. **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

5. **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

6. **Variant discovery in targeted resequencing using whole genome amplified DNA**
   Amit R. Indap, Regina Cole, Christina L. Runge, Gabor T. Marth, Michael Olivier
   BMC Genomics (2013-07-10) https://doi.org/10.1186/1471-2164-14-468
   DOI: 10.1186/1471-2164-14-468