BIG DATA, SMALL MICROBES

BIG DATA, SMALL MICROBES: GENOMIC ANALYSIS OF THE PLAGUE BACTERIUM *YERSINIA PESTIS*

BY
KATHERINE EATON, B.A. (HONS)

A THESIS SUBMITTED TO
THE DEPARTMENT OF ANTHROPOLOGY
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2022)	McMaster University
(Department of Anthropology)	Hamilton, Ontario, Canada

TITLE:	Big Data, Small Microbes: Genomic analysis of the plague bacterium *Yersinia pestis*
AUTHOR:	Katherine Eaton B.A. (Hons) Anthropology, University of Alberta
SUPERVISOR:	Dr. Hendrik Poinar
NUMBER OF PAGES:	xii, 67

Lay Abstract

The Plague is a disease that has profoundly impacted human history and is responsible for some of the most fatal pandemics ever recorded. It may surprise many to know that this disease is not a bygone of a past era, but in fact is still present in many regions of the world. Although researchers have been studying plague for hundreds of years, there are many aspects of its epidemiology that are enigmatic. In this thesis, I focus on how DNA from the plague bacterium can be used to estimate *where* and *when* this disease appeared in the past. To do so, I reconstruct the evolutionary relationships between modern and ancient strains of plague, using publicly available data and new DNA sequences retrieved from the skeletal remains of plague victims in Denmark. This work offers a new methodological framework for large-scale genetic analysis, provides a critique on what questions DNA evidence *can* and *cannot* answer, and expands our knowledge of the global diversity of plague.

Abstract

Pandemics of plague have reemerged multiple times throughout human history with tremendous mortality and extensive geographic spread. The First Pandemic (6^{th} - 8^{th} century) devastated the Mediterranean world, the Second Pandemic (14^{th} - 19^{th} century) swept across much of Afro-Eurasia, and the Third Pandemic (19^{th} - 20^{th} century) reached every continent except for Antarctica, and continues to persist in various endemic foci around the world. Despite centuries of historical research, the epidemiology of these pandemics remains enigmatic. However, recent technological advancements have yielded a novel form of evidence: ancient DNA of the plague bacterium *Yersinia pestis.* In this thesis, I explore how genomic data can be used to unravel the mysteries of *when* and *where* this disease appeared in the past. In particular, I focus on phylogenetic approaches to studying this 'small microbe' with 'big data' (i.e. 100s - 1000s of genomes). I begin by describing novel software I developed that supports the acquisition and curation of large amounts of DNA sequences in public databases. I then use this tool to create an updated global phylogeny of *Y. pestis*, which includes ~600 genomes with standardized metadata. I devise and validate a new approach for temporal modeling (i.e. molecular clock) that produces robust divergence dates in pandemic lineages of *Y. pestis.* In addition, I critically examine the questions that genomic evidence *can* and *cannot* address in isolation, such as whether the timing and spread of short-term epidemics can be confidently reconstructed. Finally, I apply this theoretical and methodological insight to a case study in which I reconstruct the appearance, persistence, and disappearance of plague in Denmark during the Second Pandemic. The three papers enclosed in this sandwich-thesis contribute to a larger body of work on the anthropology of plague, which seeks to understand how disease exposure and experience change over time and between human populations. Furthermore, this dissertation more broadly impacts both prospective studies of infectious disease, such as environmental surveillance and outbreak monitoring, and retrospective studies, which seek to date the emergence and spread of past pandemics.

‘You have to know the past to understand the present.’
- Carl Sagan

Acknowledgments

I'd like to thank my parents, Michelle and Michael Eaton. When I was little, I thought you knew everything. And now that I'm writing my doctoral dissertation… I realize you do know everything! I hope when I grow up, I turn out to be just like you <3 Thank you for your love, support, and encouragement over all these years.

To Miriam: Thank you for being my partner, my best friend, my everything. I hope that one day I have 1/10 of your intellect, kindness, and patience. (Maybe we won't hold our breath for that last one.)

To Hendrik Poinar: Thank you for your unending support and enthusiasm. Your mentorship and passion for research has been my rock during the hard times. Thank you for taking leaps of faith and trusting me when I proposed ridiculous project ideas. At least a few of those wound up in this thesis! And most importantly, thank you for traveling to Edmonton in 2013 to give a talk at the University of Alberta!

Thank you to members of my doctoral supervisory committee: Brian Golding, Tracy Prowse, and Nükhet Varlık. I am indebted to you for your generous support, careful guidance, and prompt feedback. Our affectionate motto of 'Keep It Simple Stupid' has played on loop in my head as I prepared this dissertation.

To John Silva, Marcia Furtado, and Delia Hutchinson: Thank you for guiding me through the labyrinth that is McMaster's administration. Your smiling faces up on the 5th floor were always so reassuring. I knew that if I ever had a problem, you would be there to investigate and advocate on my behalf.

To the Plague Team: Jennifer Klunk, what would I do without you? I think everything I've ever known and ever will know comes back to you. Thank you for being a dedicated mentor, a brilliant scientist, and the best companion for dancing in the lab. Madeline Tapson, I dearly miss sharing a desk with you. Your warm and friendly spirit was always comforting, and you opened my eyes

to so many new avenues of plague research. Ravneet Sidhu, I learned so much from training you and I loved that you questioned everything. I'm so excited to be collaborating with you on current and future projects. Michael Klowak and Julianna Stangroom, thank you for your HARD work in screening a dizzying number of plague samples and making the lab such a fun and exciting place to be.

To Emil Karpinski: I always looked forward to our bus rides into campus together. You played such an important role in creating a welcoming atmosphere when I first arrived and throughout my whole degree. Also, you are lab notebooks goals (wow).

To Ana Duggan: You have been a role model for me in so many avenues of my academic, personal, and professional life! You were the first woman I met that was also passionate about computational analysis, and have made me feel more comfortable and confident in my own skin.

To Nathalie Mouttham: Thank you for being such a stellar trainer and friend! I have vivid (but positive) memories of long hours in a laboratory basement doing Phenol Chloroform extractions together and playing 20 questions.

Thank you to all past and present members of the McMaster Ancient DNA Centre. In particular: Melanie Kuch, Matthew Emery, Jess Hider, Samantha Price, Marie-Hélène B.-Hardy, and Dirk Hackenberger. You created such a unique sense of community, and left pretty big shoes to fill!

Thank you to all collaborators who have generously shared their time, energy, and resources with me. In particular, thank you to Rebecca Redfern for working with me on the Roman Londoners project!

To my colleagues at Red Lobster: I worked in many kitchens to fund my education, but working at Red Lobster was by far my favorite. Also to the managers, thank you for letting me have so many free Cheddar Bay Biscuits, they were a crucial component of my student diet (no joke).

Finally, I would like to acknowledge all individuals who financially supported me throughout my doctoral research. I thank Hendrik Poinar, the Department of Anthropology, the MacDATA Institute, the Sherman Centre for Digital Scholarship, McMaster University, the Province of Alberta, and the Social Sciences Research and Humanities Research Council.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

aDNA: Ancient DNA
DNA: Deoxyribonucleic acid
MRCA: Most Recent Common Ancestor
NCBI: National Center for Biotechnology Information
SRA: Sequence Read Archive
tMRCA: Time to the most recent common ancestor

Declaration of Academic Achievement

I, Katherine Eaton, declare that this thesis titled, ‘Big Data, Small Microbes: Genomic analysis of the plague bacterium *Yersinia pestis*’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at McMaster University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at McMaster University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Chapter 1

Introduction

In 2011, I learned about a researcher named Dr. Hendrik Poinar. His team had just published a seminal paper, in which they identified the causative agent of the infamous Black Death (Bos et al., 2011). I discovered that this morbid term describes a pandemic that devastated the world in the 14th century, with unprecedented mortality and spread. In less than 10 years (1346-1353) the Black Death swept across Afro-Eurasia, killing 50% of the population (O. J. Benedictow, 2004). Outbreaks of this new and mysterious disease, often referred to as Plague, reoccurred every 10 years on average (Christensen, 2003). This epidemic cycling continued for 500 long years in Europe, but in Western Asia, the disease never truly disappeared (Varlık, 2020). The 10-year window of the Black Death alone has an estimated global mortality of 200 million people, making it the most fatal pandemic in human history (Sampath et al., 2021), and also one of the most mysterious.

The cryptic nature of this medieval disease led to significant debate among contemporaries. The dominant theory of etiology and transmission at the time was *miasma*, in which diseases were spread through noxious air (Ober & Aloush, 1982). However, alternative theories of communicable disease, such as Galen's "seeds of plague", had been circulating for more than a millennium prior (Nutton, 1983). Like these doctors of antiquity, Ibn al-Khatib, a notable Islamic scholar, took issue with the concept of *miasma*. After studying outbreaks of plague in the 14th century, he proposed an alternative hypothesis in which *minute bodies* were transmissible between humans (Syed, 1981). Like most controversial theories, this idea was not readily embraced. Some 400 years later, the British botanist Richard Bradley wrote a radical treatise on *The Plague* (Bradley, 1721) where he similarly proposed that infectious diseases were caused by living, microscopic agents. Again, this theory was rejected. It was not until the bacteriological revolution of 19th century that this "new" perspective would receive widespread acceptance (Santer, 2009). It is quite remarkable that our modern conceptions of epidemiology and bacteriology can be traced back to diverse "founders" throughout history, many of which were grappling with the perplexing nature of plague.

After it was established that *a* living organism caused the Black Death, the intuitive next step was to precisely identify *the* organism. The symptoms described in historical texts seemed to incriminate bubonic plague (O. J. Benedictow, 2004), a bacterial pathogen that passes from *rodents to humans*, and leads to grotesquely swollen lymph nodes (buboes). On the other hand, the rapid spread of the Black Death suggests this was a contagion primarily driven by *human to human* transmission, which more closely fit the profile of an Ebola-like virus (Scott & Duncan, 2001). In the 1990s and 2000s, geneticists began contributing novel evidence to the debate, by retrieving pathogenic DNA from skeletal remains (Drancourt et al., 1998). The plague bacterium, *Yersinia pestis*, played a central role in these molecular investigations, as researchers sought to either establish or refute its presence in medieval victims (Gilbert et al., 2004b). The competitive nature of this discourse fueled significant technological progress, and over the next decade, the study of ancient DNA became a well-established discipline. However, the origins of the Black Death remained unresolved, due to numerous controversies surrounding DNA contamination and scientific rigor (Cooper & Poinar, 2000).

As an undergraduate student of forensic anthropology, I was fascinated by the rapid pace at which the field of ancient DNA was developing. I attribute my developing academic obsession to two early-career experiences. First, was reading the *highly* entertaining back-and-forth commentaries in academic journals (Gilbert et al., 2004a), where plague researchers would occasionally exchange personal insults (Raoult, 2003). It was clear that these researchers cared *deeply* about their work. Despite the occasional toxicity, I found these displays of passion to be engaging and refreshing, compared to the otherwise emotionally-sterile field of scientific publishing.

The second defining experience, was the perplexing and often frustrating task of diagnosing infectious diseases from skeletal remains. I was intrigued by the idea of reconstructing an individual's life story from their skeleton, and using this information to solve the *mysteries of the dead*. However, while some forms of trauma leave diagnostic marks on bone (ex. sharp force), acute infectious diseases rarely manifest in the skeleton (Brown & Inhorn, 2013; Ortner, 2007) and thus are 'invisible' to an anthropologist. Because of this, I found the new field of ancient DNA to be *extremely* appealing, as it offered a novel solution to this problem. Anthropologists could now directly retrieve the *pathogens* that had infected an individual, and contribute new insight regarding disease exposure and experience throughout human history. These experiences suggested to me that studying the ancient DNA of pathogens would be an exciting, dynamic, and productive line of research for a graduate degree. I'm happy to say that 10 years later, I still agree with this statement, and by writing this dissertation I hope to convince you, the reader, as well.

Which brings us back to Dr. Hendrik Poinar and his team's seminal work on the mysterious Black Death. The study, led by first author Kirsten Bos, had found DNA evidence of the plague bacterium *Y. pestis* in several Black Death victims buried in a mass grave in London (Bos et al., 2011). However, they did not just retrieve a few strands of DNA, they captured millions of molecules (10.5 million to be precise) which allowed them to reconstruct the *Y. pestis* genome, comprising four million DNA bases. The amount of molecular evidence was staggering, and offered irrefutable proof that the plague bacterium was present during the time of Black Death. However, with a sample size of N=1, the genetic link between *Y. pestis* and this ancient pandemic was tentative at best.

Armed with the proposal of finding more evidence of *Y. pestis* in the archaeological record, I applied to work for Dr. Hendrik Poinar at the McMaster Ancient DNA Centre. In 2014, I had the delight and privilege of being accepted into the graduate program at McMaster University. Alongside other members of the "McMaster Plague Team", I set about the daunting task of screening the skeletal remains of more than 1000 individuals for molecular evidence of *Y. pestis*. This material was generously provided by archaeological collaborators, who were similarly invested in the idea that ancient DNA techniques could identify infectious diseases in their sites. These archaeological remains reflected nearly a millennium of human history, with sampling ages ranging from the 9th to the 19th century CE. The geographic diversity was also immense, with individuals sampled across Europe, Africa, and Asia.

Of the 1000+ individuals screened, approximately 30% originated in Denmark. Due to this large sample size, we, the "Plague Team", had the greatest success in identifying ancient *Y. pestis* in this region. Over a period of 5 years, we retrieved *Y. pestis* DNA from 13 Danish individuals dated to the medieval and early modern periods. To contextualize these plague isolates, we reconstructed their evolutionary relationships using a large comparative dataset of global *Y. pestis*. In Chapter 4, I present the results of this collaborative study, which marks the first longitudinal analysis of an ancient pathogen in a single region. I explore whether the genetic evidence of *Y. pestis* aligns with the historical narrative of the Black Death, and whether or not subsequent epidemics can be attributed to long-distance reintroductions. However, while this high-throughput study was the first one I embarked on, as the chapter numbering indicates, it would be the last project I completed due to several unforeseen complications.

While the McMaster Plague Team was busy screening for *Y. pestis*, so too were other ancient DNA centres throughout the world. Between 2011 and 2021, more than 100 ancient *Y. pestis* genomes were published, making plague the *most intensively sequenced historical disease*. The sequencing of modern isolates accelerated in tandem, with over 1500 genomes produced from culture collections of 20th and 21st century plague outbreaks (Z. Zhou et al., 2020). Because of this influx of evidence, the research questions changed accordingly. Geneticists were no longer interested in just establishing the *presence* of *Y. pestis* during the short time frame of the Black Death (1346-1353), they wanted to know *how* it behaved and spread throughout the long 500 years of this pandemic. The longitudinal study design of Chapter 4 was therefore well-positioned to address these nuanced epidemiological questions. However, this novel genetic evidence also introduced new complexities.

It quickly became clear that isolates of *Y. pestis* sampled during epidemic periods were highly similar in terms of genetic content, if not indistinguishable clones (Spyrou et al., 2019). This called into question the resolution of genomic evidence, and whether the geographic origins and spread of the Black Death could be accurately inferred using ancient DNA studies. This was further confounded by the finding that the rate of evolutionary change in *Y. pestis* could vary tremendously (Cui et al., 2013) which led to the discovery that previously published temporal models were erroneous (Wagner et al., 2014). It became increasingly uncertain whether genetic evidence could be used to produce informative estimates of the timing of plague's frequent reemergences (Duchêne et al., 2016). As I read these critical studies, I began developing an idea to address the substantial gaps in our evolutionary theory and methodology concerning the plague bacterium *Y. pestis*. This idea culminated in Chapter 3, where I curated and contextualized the largest global data set of plague genomes. I critiqued the existing spatiotemporal models of plague's evolutionary history, and with the assistance of my co-authors, devised a new methodological approach. This method would then be repurposed for Chapter 4, so that I could infer the emergence and disappearance of *Y. pestis* in Denmark with greater accuracy. However, as the chapter numbering once again reflects, there was one final obstacle.

Synthesizing the largest genomic data set was a lofty ambition, especially considering that there were few software tools available to perform such a task. New plague genomes of *Y. pestis* were being published monthly, and at times even weekly, with such volume that manual tracking became impossible. My excel spreadsheet of genetic metadata became riddled with errors and fields with missing data. The era of "Big Data" had arrived, and I was woefully unequipped to effectively manage this deluge of information. In response, I ventured into the tumultuous waters of software development. In Chapter 2, I describe my original software that automates the acquisition and organization of genetic metadata. Academic publishing in the field of software was a unique experience, as I had to both *produce a scholarly manuscript* and demonstrate *expertise as a service-provider*. This database tool has continually proven to be indispensable, and is the backbone upon which the studies in Chapter 3 and Chapter 4 would be rebuilt upon.

At this point, I re-introduce the dissertation as a collection of three hierarchical, but independently published, studies. I first describe an original piece of software in Chapter 2, which automates the retrieval and organization of publicly available sequence data. In Chapter 3, I outline how this tool was used to generate an updated and curated phylogeny of *Y. pestis*, which yielded novel insight regarding the timing and origins of past pandemics. In this chapter, I also conduct a critical examination of the historical questions that genomic evidence can, or cannot, address. In Chapter 4, I use these theories and methods to reconstruct the emergence and continuity of plague in Denmark over a period of 400 years. I conclude in Chapter 5 with a discussion of the contributions of each study, with a particular focus on their significance within the broader field of anthropology.

Chapter 2

NCBImeta: Efficient and comprehensive metadata retrieval from NCBI databases

Published 03 February 2020 in
The Journal of Open Source Software, 5(46), 1990.
https://doi.org/10.21105/joss.01990

Katherine Eaton[1,2]

[1] McMaster Ancient DNA Centre, McMaster University
[2] Department of Anthropology, McMaster University

2.1 Summary

NCBImeta is a command-line application that downloads and organizes biological metadata from the National Centre for Biotechnology Information (NCBI). While the NCBI web portal provides an interface for searching and filtering molecular data, the output offers limited options for record retrieval and comparison on a much larger and broader scale. NCBImeta tackles this problem by creating a reformatted local database of NCBI metadata based on user search queries and customizable fields. The output of NCBImeta, optionally a SQLite database or text file(s), can then be used by computational biologists for applications such as record filtering, project discovery, sample interpretation, and meta-analyses of published work.

2.2 Background

Recent technological advances in DNA sequencing have propelled biological research into the realm of big data. Due to the tremendous output of Next Generation Sequencing (NGS) platforms, numerous fields have transformed to explore this novel high-throughput data. Projects that quickly adapted to incorporate these innovative techniques included monitoring the emergence of antibiotic resistance genes (Zankari et al., 2012), epidemic source tracking in human rights cases (Eppinger et al., 2014), and global surveillance of uncharacterized organisms (Connor et al., 2015). However, the startling rate at which sequence data are being deposited online have presented significant hurdles to the efficient reuse of published data. In response, there is growing recognition within the computational community that effective data mining techniques are a dire necessity (Mackenzie et al., 2016; Nakazato et al., 2013).

An essential step in the data mining process is the efficient retrieval of comprehensive metadata. These metadata fields are diverse in nature, but often include the characteristics of the biological source material, the composition of the raw data, the objectives of the research initiative, and the structure of the post-processed data. Several software applications have been developed to facilitate bulk metadata retrieval from online repositories. Of the available tools, SRAdb [(Zhu et al., 2013)], the Pathogen Metadata Platform (Chang et al., 2016), MetaSRA (Bernstein et al., 2017), and pysradb (Choudhary, 2019) are among the most widely utilised and actively maintained. While these software extensions offer substantial improvements over the NCBI web browser experience, there remain several outstanding issues.

1. Existing tools assume external programming language proficiency (ex. R, Python, SQL), thus reducing tool accessibility.
2. Available software focuses on implementing access to singular NCBI databases in isolation, for example, the raw data repository the Sequence Read Archive (SRA). This does not empower researchers to incorporate

evidence from multiple databases, as it fails to fully leverage the power of interconnected information within the relational database scheme of NCBI.

3. Existing software provides only intermittent database updates, where users are dependent on developers releasing new snapshots to gain access to the latest information. This gives researchers less autonomy over what data they may incorporate as newer records are inaccessible, and may even introduce sampling bias depending on when the snapshots are generated.

In response, `NCBImeta` aims to provide a more user-inclusive experience to metadata retrieval, that emphasizes real-time access and provides generalized frameworks for a wide variety of NCBI's databases.

2.3 NCBImeta

`NBCImeta` is a command-line application that executes user queries and metadata retrieval from the NCBI suite of databases. The software is written in Python 3, using the `BioPython` module (Cock et al., 2009) to connect to, search, and download XML records with NCBI's E-Utilities (Kans, 2013/2019). The `lxml` package is utilised to perform XPath queries to retrieve nodes containing biological metadata of interest. `SQLite` is employed as the database management system for storing fetched records, as implemented with the `sqlite3` python module. Accessory scripts are provided to supply external annotation files, to join tables within the local database so as to re-create the relational database structure, and finally to export the database as tabular text for downstream analyses. `NCBImeta` currently interfaces with the molecular and literature databases (*Entrez Help*, 2006/2016) described in Table 2.3.1.

Table 2.3.1: NCBI databases supported in NCBImeta.

Database	Description
Assembly	Descriptions of the names and structure of genomic assemblies, statistical reports, and sequence data links.
BioSample	Characteristics of the biological source materials used in experiments.
BioProject	Goals and progress of the experimental initiatives, originating from an individual organization or a consortium.
Nucleotide	Sequences collected from a variety of sources, including GenBank, RefSeq, TPA and PDB.
PubMed	Bibliographic information and citations for biomedical literature from MEDLINE, life science journals, and other online publications.
SRA	Composition of raw sequencing data and post-processed alignments generated via high-throughput sequencing platforms.

The typical workflow of NCBImeta follows four major steps as outlined in Figure 2.3.1. Users first configure the program with their desired search terms. NCBImeta is then executed on the command-line to fetch relevant records and organize them into a local database. Next, the user optionally edits the database to, for example, add their own custom metadata. Finally, the resulting database, kept in SQLite format or exported to text, delivers 100+ biologically-relevant metadata fields to researcher's fingertips. This process not only saves significant time compared to manual record retrieval through the NCBI web portal, but additionally unlocks attributes for comparison that were not easily accessible via the web-browser interface.

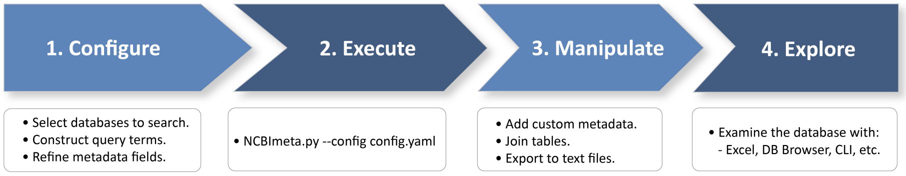

1. Configure
2. Execute
3. Manipulate
4. Explore
• Select databases to search.
• Construct query terms.
• Refine metadata fields.
• NCBImeta.py --config config.yaml
• Add custom metadata.
• Join tables.
• Export to text files.
• Examine the database with:
- Excel, DB Browser, CLI, etc.

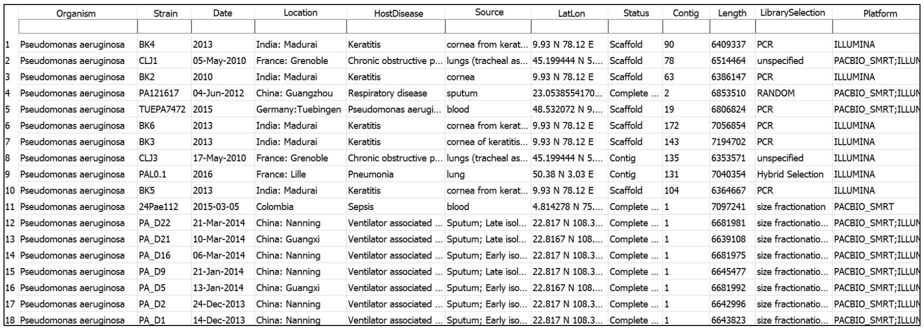
Figure 2.3.1: NCBImeta user workflow.

NCBImeta's implementation offers a novel approach to metadata management and presentation that improves upon the prevously described limitations of existing software in a number of ways. First, NCBImeta is run on the command-line, and the final database can be exported to a text file, thus no knowledge of an external programming language is required to generate or explore the output. Second, a general parsing framework for tables and metadata fields was developed which can be extended to work with diverse database types contained within NCBI's infrastructure. Finally, a query system was implemented for record retrieval that allows users to access records in real-time, as opposed to working with intermittent or out-dated database snapshots.

2.4 Use Case

The following section demonstrates how NCBImeta can be used to obtain current and comprehensive metadata for a pathogenic bacteria, *Pseudomonas aeruginosa*, from various sequencing projects across the globe. *P. aeruginosa* is an opportunistic pathogen associated with the disease cystic fibrosis (CF) and is highly adaptable to diverse ecological niches (Stewart et al., 2014). As such, it is a target of great interest for comparative genomics and there are currently over 15,000 genomic sequence records available which are spread across two or more databases. In cases such as this, it is critical to leverage the tremendous power of these existing datasets while being conscious of the labor typically required to retrieve and contextualize this information. NCBImeta renders the problem of acquiring and sifting through this metadata trivial and facilitates the integration of information from multiple sources.

To identify publicly available *P. aeruginosa* genomes, `NCBImeta` is configured to search through the tables *Assembly* (assembled genomes) and *SRA* (raw data). For additional context, `NCBImeta` is used to retrieve metadata from the *Nucleotide* table for descriptive statistics of the genomic content, from the *BioProject* table to examine the research methodology of the initiative, from *Pubmed* to identify existing publications, and finally from the *Biosample* table to explore characteristics of the biological material. A small subset of the 100+ retrieved columns is shown in Figure 2.4.1, to provide a visual example of the output format and the metadata that is retrieved.



	Organism	Strain	Date	Location	HostDisease	Source	LatLon	Status	Contig	Length	LibrarySelection	Platform
1	Pseudomonas aeruginosa	BK4	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	90	6409337	PCR	ILLUMINA
2	Pseudomonas aeruginosa	CLJ1	05-May-2010	France: Grenoble	Chronic obstructive p...	lungs (tracheal as...	45.199444 N 5....	Scaffold	78	6514464	unspecified	PACBIO_SMRT;ILLUN
3	Pseudomonas aeruginosa	BK2	2010	India: Madurai	Keratitis	cornea	9.93 N 78.12 E	Scaffold	63	6386147	PCR	ILLUMINA
4	Pseudomonas aeruginosa	PA121617	04-Jun-2012	China: Guangzhou	Respiratory disease	sputum	23.0538554170...	Complete ...	2	6853510	RANDOM	PACBIO_SMRT;ILLUN
5	Pseudomonas aeruginosa	TUEPA7472	2015	Germany:Tuebingen	Pseudomonas aerugi...	blood	48.532072 N 9....	Scaffold	19	6806824	PCR	PACBIO_SMRT;ILLUN
6	Pseudomonas aeruginosa	BK6	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	172	7056854	PCR	ILLUMINA
7	Pseudomonas aeruginosa	BK3	2013	India: Madurai	Keratitis	cornea of keratitis...	9.93 N 78.12 E	Scaffold	143	7194702	PCR	ILLUMINA
8	Pseudomonas aeruginosa	CLJ3	17-May-2010	France: Grenoble	Chronic obstructive p...	lungs (tracheal as...	45.199444 N 5....	Contig	135	6353571	unspecified	ILLUMINA
9	Pseudomonas aeruginosa	PAL0.1	2016	France: Lille	Pneumonia	lung	50.38 N 3.03 E	Contig	131	7040354	Hybrid Selection	ILLUMINA
10	Pseudomonas aeruginosa	BK5	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	104	6364667	PCR	ILLUMINA
11	Pseudomonas aeruginosa	24Pae112	2015-03-05	Colombia	Sepsis	blood	4.814278 N 75....	Complete ...	1	7097241	size fractionation	PACBIO_SMRT
12	Pseudomonas aeruginosa	PA_D22	21-Mar-2014	China: Nanning	Ventilator associated ...	Sputum; Late isol...	22.817 N 108.3...	Complete ...	1	6681981	size fractionatio...	PACBIO_SMRT;ILLUN
13	Pseudomonas aeruginosa	PA_D21	10-Mar-2014	China: Guangxi	Ventilator associated ...	Sputum; Late isol...	22.8167 N 108....	Complete ...	1	6639108	size fractionatio...	PACBIO_SMRT;ILLUN
14	Pseudomonas aeruginosa	PA_D16	06-Mar-2014	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6681975	size fractionatio...	PACBIO_SMRT;ILLUN
15	Pseudomonas aeruginosa	PA_D9	21-Jan-2014	China: Nanning	Ventilator associated ...	Sputum; Late isol...	22.817 N 108.3...	Complete ...	1	6645477	size fractionatio...	PACBIO_SMRT;ILLUN
16	Pseudomonas aeruginosa	PA_D5	13-Jan-2014	China: Guangxi	Ventilator associated ...	Sputum; Early iso...	22.8167 N 108....	Complete ...	1	6681992	size fractionatio...	PACBIO_SMRT;ILLUN
17	Pseudomonas aeruginosa	PA_D2	24-Dec-2013	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6642996	size fractionatio...	PACBIO_SMRT;ILLUN
18	Pseudomonas aeruginosa	PA_D1	14-Dec-2013	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6643823	size fractionatio...	PACBIO_SMRT;ILLUN

Figure 2.4.1: A subset of the 100+ metadata columns retrieved for *P. aeruginosa* sequencing projects. Viewed with DB Browser for SQLite (https://sqlitebrowser.org/).

Subsequently, the output of NCBImeta can be used for exploratory data visualization and analysis. The text file export function of NCBImeta ensures downstream compatibility with both user-friendly online tools (ex. Google Sheets Charts) as well as more advanced visualization packages (Wickham, 2016). In Figure 2.4.2, the geospatial distribution of *P. aeruginosa* isolates are plotted alongside key aspects of genomic composition (ex. number of genes).

Finally, NCBImeta can be used to streamline the process of primary data acquisition following careful filtration. FTP links are provided as metadata fields for databases attached to an FTP server (ex. Assembly, SRA) which can be used to download biological data for downstream analysis.

2.5 Future Work

The development of `NCBImeta` has primarily focused on a target audience of researchers whose analytical focus is prokaryotic genomics and the samples of interest are the organisms themselves. Chief among those include individuals pursuing questions concerning epidemiology, phylogeography, and comparative genomics. Future releases of `NCBImeta` will seek to broaden database representation to include gene-centric and transcriptomics research (ex. NCBI's Gene and

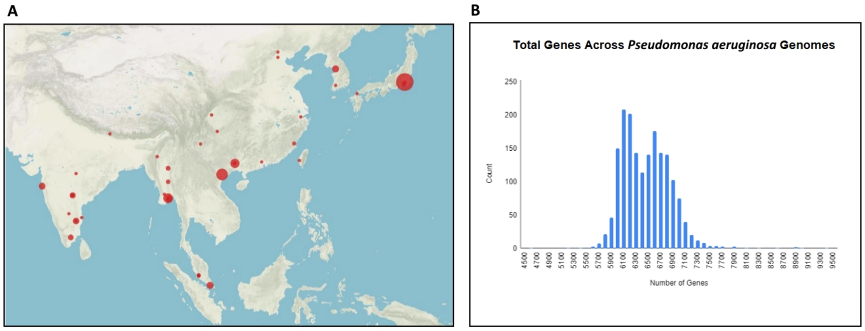


Figure 2.4.2: Metadata visualization of *P. aeruginosa* sequencing projects. A) The geographic distribution of samples in this region highlights a large number originating in Japan. Visualized with Palladio (https://hdlab.stanford.edu/palladio/). B) The number of genes per organism reveals a multi-modal distribution within the species.

GEO databases).

2.6 Availability

NCBImeta is a command-line application written in Python 3 that is supported on Linux and macOS systems. It is distributed for use under the OSD-compliant MIT license (https://opensource.org/licenses/MIT). Source code, documentation, and example files are available on the GitHub repository (https://github.com/ktmeaton/NCBImeta).

2.7 Acknowledgments

I would like to thank Dr. Hendrik Poinar and Dr. Brian Golding for their guidance and support on this project, as well as for insightful conversations regarding biological metadata, the architecture of NCBI, and software deployment. Thank you to Dr. Andrea Zeffiro, Dr. John Fink, Dr. Matthew Davis, and Dr. Amanda Montague for valuable discussions regarding APIs, digital project management, and software publishing. Thank you to all past and present members of the McMaster Ancient DNA Centre and the Golding Lab. This work was supported by the MacDATA Institute (McMaster University, Canada) and the Social Sciences and Humanities Research Council of Canada (#20008499).

Chapter 3

Plagued by a cryptic clock: Insight and issues from the global phylogeny of *Yersinia pestis*

Submitted 17 December 2021 to
Communications Biology.
https://www.researchsquare.com/article/rs-1146895

Katherine Eaton[1,2], Leo Featherstone[3], Sebastian Duchene[3], Ann G. Carmichael[4], Nükhet Varlık[5], G. Brian Golding[6], Edward C. Holmes[7], Hendrik N. Poinar[1,2,8,9,10]

[1]McMaster Ancient DNA Centre, McMaster University, Hamilton, Canada.
[2]Department of Anthropology, McMaster University, Hamilton, Canada.
[3]The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia.
[4]Department of History, Indiana University Bloomington, Bloomington, USA.
[5]Department of History, Rutgers University-Newark, Newark, USA.
[6]Department of Biology, McMaster University, Hamilton, Canada.
[7]Sydney Institute for Infectious Diseases, School of Life & Environmental Sciences and School of Medical Sciences, University of Sydney, Sydney, Australia.
[8]Department of Biochemistry, McMaster University, Hamilton, Canada.
[9]Michael G. DeGroote Institute of Infectious Disease Research, McMaster University, Hamilton, Canada.
[10]Canadian Institute for Advanced Research, Toronto, Canada.

3.1 Abstract

Plague has an enigmatic history as a zoonotic pathogen. This potentially devastating infectious disease will unexpectedly appear in human populations and disappear just as suddenly. As a result, a long-standing line of inquiry has been to estimate when and where plague appeared in the past. However, there have been significant disparities between phylogenetic studies of the causative bacterium, *Yersinia pestis*, regarding the timing and geographic origins of its reemergence. Here, we curate and contextualize an updated phylogeny of *Y. pestis* using 601 genome sequences sampled globally. We perform a detailed Bayesian evaluation of temporal signal in subsets of these data and demonstrate that a *Y. pestis*-wide molecular clock model is unstable. To resolve this, we devised a new approach in which each *Y. pestis* population was assessed independently. This enabled us to recover significant temporal signal in five populations, including the ancient pandemic lineages which we now estimate may have emerged decades, or even centuries, before a pandemic was historically documented from European sources. Despite this, we only obtain robust divergence dates from populations sampled over a period of at least 90 years, indicating that genetic evidence alone is insufficient for accurately reconstructing the timing and spread of short-term plague epidemics. Finally, we identify key historical data sets that can be used in future research, which will complement the strengths and mitigate the weaknesses of genomic data.

3.2 Introduction

Plague has an impressively long and expansive history as a zoonosis of rodents. The earliest "fossil" evidence of the plague bacterium, *Yersinia pestis*, stems from ancient DNA studies which date its first emergence in humans to the Late Neolithic Bronze Age (LNBA) approximately 5000 years ago (Andrades Valtueña et al., 2017). During this time, *Y. pestis* has dispersed globally on multiple occasions due to an ability to infect a variety of mammalian hosts (Perry & Fetherston, 1997) and ever-expanding trade networks (Yue et al., 2017). Few regions of the ancient and modern world remain untouched by this disease, as plague has demonstrated a persistent presence on every continent except Australia and Antarctica (*Plague*, 2017). There are three historically documented pandemics of plague: the First Pandemic (6^{th} to 8^{th} century CE) (Wagner et al., 2014), the Second Pandemic (14^{th} to 19^{th} century CE) (Varlık, 2014), and the Third Pandemic (19^{th} to 20^{th} century CE) (Xu et al., 2014). The advent of each has been marked by a series of outbreaks, such as the medieval Black Death (1346 - 1353 CE), which is estimated to have killed more than half of Europe's population (O. J. Benedictow, 2021).

One long-standing line of inquiry in plague's evolutionary history has been estimating the timing, origins, and spread of these past pandemics. The most intensively researched events have been: (1) the first appearance of *Y. pestis* in human populations (Rasmussen et al., 2015), (2) the onset and progression of the three pandemics (Bos et al., 2011; Cui et al., 2013; Wagner et al., 2014), and (3) the inter-pandemic or "quiescent" periods where *Y. pestis* recedes into wild rodent reservoirs and disappears from the historical record (Green, 2020a; Zeppelini et al., 2018). Our knowledge of these events has deepened considerably in recent years, owing in part to technological advancements in the retrieval and sequencing of ancient DNA alongside new molecular clock dating methods.

Despite intensive interest and methodological advancement, the rate and time scale of evolution in *Y. pestis* remains notoriously difficult to estimate. This is largely attributed to the substantial variation in evolutionary rates that has been documented across the phylogeny (Cui et al., 2013; Spyrou et al., 2019). As a result, considerable debate has emerged over whether *Y. pestis* has no temporal signal (Wagner et al., 2014), thereby preventing meaningful rate estimates, or if some *Y. pestis* populations have such distinct rates that a species-wide signal is obscured (Duchêne et al., 2016). This uncertainty has resulted in radically different rate and date estimates between studies, with node dates shifting by several millennia (Cui et al., 2013; Rasmussen et al., 2015).

The geographic origins and dispersal of past pandemics are similarly contentious, particularly concerning mechanisms of spread and their underlying ecology. This contention concerns competing hypotheses about the relative importance of localized persistence versus long-distance reintroduction (Carmichael, 2015; Guellil et al., 2020; Schmid et al., 2015). Among both sides of this issue, there is an expectation that genomic evidence will play a significant role (Bramanti et al., 2021), if not resolve the debate (Schmid et al., 2015). However, no study to date has statistically evaluated whether *Y. pestis* genomes have sufficient geographic signal to confidently infer ancestral locations and spread.

To address these debates and obstacles, we: (1) curated and contextualized the most recent *Y. pestis* genomic evidence, (2) reviewed and critiqued our current understanding of plague's population structure, (3) devised a new approach for recovering temporal signal in *Y. pestis*, and (4) critically assessed the reliability of phylogeographic analysis. We ground our results and their interpretation using informative historical case studies to demonstrate the methodological and interpretive consequences. We anticipate these results will impact both retrospective and prospective studies of plague, which seek to date the emergence and spread of past pandemics as well as monitor the progression of ongoing outbreaks.

3.3 Results and Discussion

3.3.1 Population Structure

To determine the population structure of *Y. pestis*, we first estimated a maximum likelihood phylogeny using 601 global isolates including 540 modern (89.9%) and

61 ancient (10.1%) strains (Methods). We rooted the tree using two genomes of the outgroup taxa *Yersinia pseudotuberculosis*. 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 3.3.1 A, we contextualize the global phylogeny using three nomenclature systems: the metabolic biovars, major branches, 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 analysis.

3.3.1.1 Biovars

The oldest classification system of *Y. pestis* is the biovar nomenclature that uses metabolic differences to define population structure. Accordingly, *Y. pestis* can be categorized into four classical biovars: *Antiqua* (ANT), *Medievalis* (MED), *Orientalis* (ORI), and *microtus/pestoides* (PE) (Devignat, 1951; D. Zhou et al., 2004). Non-classical biovars have also been introduced, such as the *Intermedium* biovar (IN), which reflects a transitional state from *Antiqua* to *Orientalis* (Li et al., 2009). The biovar system is simple in application, as it largely focuses on two traits: the ability to ferment glycerol and reduce nitrate (D. Zhou et al., 2004). However, this simplicity is offset by the growing recognition of regional inconsistencies in metabolic profiles (Kutyrev et al., 2018). This is further exacerbated by the sequencing of non-viable, "extinct" *Y. pestis* for which metabolic sub-typing is challenging (Bos et al., 2011). Researchers have responded to this uncertainty in a variety of ways, by extrapolating existing biovars (Wagner et al., 2014) and creating new pseudo-biovars (PRE) (Rasmussen et al., 2015). Others have foregone the biovar nomenclature altogether in favor of locally-developed taxonomies (Kutyrev et al., 2018). Despite extensive research, it remains unclear which metabolic traits, if any, can be used to classify *Y. pestis* into distinct populations on a global scale.

3.3.1.2 Major Branches

In contrast to the biovar nomenclature which emphasizes phenotype, the major branch nomenclature focuses on genotype. This system divides the global phylogeny of *Y. pestis* into populations according to their relative position to a multifurcation called the "Big Bang" polytomy (Cui et al., 2013). All lineages that diverged prior to this polytomy are grouped into Branch 0 and those diverging after form Branches 1 through 4. Because this multifurcation plays such a central role in this system, there is great interest in estimating its timing and geographic origins (Green, 2020a, 2020b).

3.3.1.3 Time Period

Ancient *Y. pestis* genomes now represent a substantive portion of the known genetic diversity yet cannot be easily classified via direct metabolic testing. An alternative strategy has been employed that incorporates contextual evidence such as the sampling age, historical time period, and potential pandemic associations. In ancient DNA studies, the genetic diversity of *Y. pestis* is commonly divided into four time periods: the Late Neolithic Bronze Age (Rasmussen et al., 2015), the First Pandemic (Wagner et al., 2014), the Second Pandemic (Spyrou et al., 2019), and the Third Pandemic (Cui et al., 2013) (Figure 3.3.1 B).

The key strengths of the time period nomenclature are two-fold. First, it provides a foundation for interdisciplinary discourse, in which the genetic diversity can be contextualized and explained using relevant historical records. Second, this system effectively categorizes the historical outbreaks of plague recorded in Europe. This can be seen in Figure 3.3.1, where the Bronze Age strains (0.PRE) in Europe are replaced by those of the First Pandemic (0.ANT4), which in turn are replaced by strains of the Second Pandemic (1.PRE). However, this "strength" comes at a cost, as this system is far less effective in describing plague populations outside of Europe and incurs two significant risks.

The first risk is artificially grouping unrelated populations. Contemporaneous strains can have distinct evolutionary histories (Spyrou et al., 2018) even when originating from the same plague foci. The *Pestoides* (0.PE) and *Medievalis* (2.MED) biovars are informative examples, as these populations have co-existed in the Caucasus mountains since at least the 20th century (Figure 3.3.1 C). The second risk is artificially separating related populations. The Second and Third Pandemics were previously seen as mutually exclusive events dated to the 14^{th} to 18^{th} century, and the late 19^{th} to mid-20^{th} century respectively (Gage & Kosoy, 2005). Recent historical scholarship has contested this claim and demonstrated that these dating constraints are a product of a Eurocentric view of plague (Varlık, 2014). The Second Pandemic is now recognized to have extended into at least the 19^{th} century (Bolaños, 2019; Varlık, 2020) and the Third Pandemic is hypothesized to have began as early as the 18^{th} century (Tan et al., 2002; Xu et al., 2014). Phylogenetic analysis reveals genetic continuity between these two events, as the Third Pandemic (1.ORI) is a direct descendant of the Second Pandemic (1.PRE) (Spyrou et al., 2016). What remains unknown is the extent of temporal overlap, and as such, it is unclear how to distinguish these pandemics using genetic evidence.

A final limitation is that several populations are curiously excluded from the pandemic nomenclature altogether. 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 (Eroshenko et al., 2021). In particular, the *Medievalis* population (2.MED) has dispersed throughout Asia (Figure 3.3.1) with the fastest spread velocity of any *Y. pestis* lineage (Xu et al., 2014). Despite its epidemiological significance, Branch 2 populations across Asia continue to be overlooked in the pandemic taxonomy of *Y. pestis*. As ancient DNA sampling strategies expand in geographic scope, and as more non-European historical sources are brought to bear, it will be important to consider how best to refashion the historical period nomenclature to encompass this diversity.

3.3.1.4 Integrative Approach

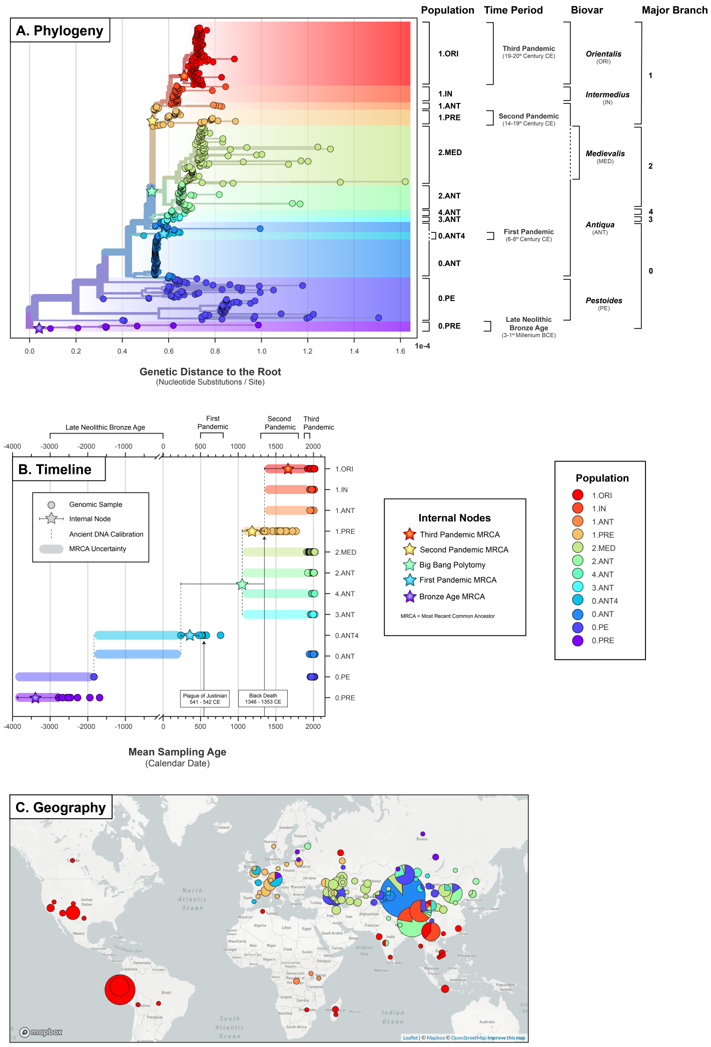
There exists no current classification system which comprehensively represents the global population structure of *Y. pestis*. Instead, integrative approaches have been previously used in large comparative studies of *Y. pestis* (Cui et al., 2013; Morelli et al., 2010). We therefore take the intersection of the three taxonomic systems discussed previously and describe 12 populations for further statistical analysis (Figure 3.3.1, Table S1). In the following sections, we highlight the novel insight and issues that arise when this population structure is explicitly incorporated into molecular clock models and phylogeographic reconstructions.

3.3.2 Estimating Rates of Evolutionary Change

The extent of rate variation present in our updated genomic data set is notably larger than those depicted in previous studies (Pisarenko et al., 2021; Spyrou et al., 2019) . A root-to-tip regression on sampling age reproduces the finding that substitution rates in *Y. pestis* are poorly represented by a simple linear model or "strict clock" (Figure 3.3.2 A). We found a very low coefficient of determination (R^2=0.09) that indicates a large degree of unaccounted variation. This finding suggests that evolutionary change in *Y. pestis* may be more appropriately estimated using a "relaxed clock", where rate variation is explicitly modeled. To test this hypothesis, we performed a Bayesian Evaluation of Temporal Signal (BETS) (Duchene, Lemey, et al., 2020). In brief, this method tested four model configurations including: (1) a strict clock, (2) a relaxed clock, (3) the inclusion of sampling ages, and (4) no sampling ages. Configurations with no sampling ages explicitly test for the presence of temporal signal. A comparison of the model likelihoods, or Bayes factors, was then used to assess the degree of temporal signal.

BETS was inconclusive when attempting to fit a single clock to the updated global diversity of *Y. pestis*. The Markov chain Monte Carlo (MCMC) inference exhibited poor sampling of parameter space (effective sample size, ESS < 200) across all model configurations, even when we reduced sources of variation by removing tip date uncertainty, fixing the tree topology, and removing up to 70% of the genomes. The poor performance of a single clock model is consistent with several other studies, in which low ESS values were observed (Rasmussen et al., 2015) and divergence dates could not be estimated (Wagner et al., 2014). A single clock model is not a viable approach for *Y. pestis*, as there is excessive rate variation across the global phylogeny, which likely explains node-dating disparities between previous studies (Cui et al., 2013; Morelli et al., 2010; Rasmussen et al., 2015).

In contrast to the single clock approach, we observed substantial improvements when each population was assessed independently. All model parameters



Powered by Nextstrain (Hadfield et al *Bioinformatics*)

Figure 3.3.1: The phylogenetic and spatiotemporal diversity of 601 *Y. pestis* genomes. Populations were defined by integrating three nomenclature systems: the major branches, biovars, and time periods. **A**: The maximum likelihood phylogeny of *Y. pestis* with branch lengths scaled by genetic distance from the root in the number of nucleotide substitutions per site. The tree was rooted using two genomes of the outgroup taxa *Y. pseudotuberculosis*, which were pruned before visualization. **B**: The mean sampling age of each genome with internal node dates bounded by ancient DNA calibrations. **C**: The sampling location of each genome with coordinates standardized to the centroid of the associated province/state.

in our Bayesian analysis demonstrated MCMC convergence with ESS values well above 200 and we detected temporal signal in 9 out of 12 *Y. pestis* populations (Table S2). Several of these appeared more clock-like than others, which was observed through the root-to-tip regression and Bayesian rate estimation. For example, we found rate variation to be low in the Bronze Age (R^2=0.92), moderate in the Second Pandemic (R^2=0.76) and high in *Medievalis* (R^2>=0.02). Our results indicate that population specific models are a more effective approach for estimating substitution rates across the global phylogeny.

To demonstrate the application of our molecular clock method and the interpretive consequences, we explored three outcomes as case studies. First, as a control, we examined *Y. pestis* populations that had (i) no temporal signal. These "negative" cases inform us about the minimum sampling time, or phylodynamic threshold, required to obtain robust temporal estimates in *Y. pestis*. Second, we examined populations with (ii) irreproducible estimates between studies, such as the time to most recent common ancestor (tMRCA). We discuss how sampling bias drives this outcome, and how it can be identified and corrected with ancient DNA calibrations where available. Finally, we identify the populations with the most (iii) informative rates and dates. We discuss how these molecular dates change our understanding of pandemic "origins" and complement recent historical scholarship.

3.3.2.1 No Temporal Signal

We found several *Y. pestis* populations with no detectable temporal signal that include the *Intermedium* (1.IN) and *Antiqua* biovars (2.ANT, 3.ANT). Despite being sampled over a period as long as 84 years (2.ANT), these populations have not accumulated sufficient evolutionary change to yield informative divergence dates. This limited diversity is identifiable in the maximum likelihood phylogeny as populations with the highest density of nodes sitting close to their roots (Figure 3.11.6). Out of caution, we also consider the rates and dates associated with the *Antiqua* population 4.ANT to be non-informative, as it has a similar node distribution, with a smaller number of samples (N=12) collected over an even shorter time frame (38 years).

Our results show that for robust temporal estimates to be obtained, *Y. pestis* must be sampled over multiple decades at minimum. This time frame is largely consistent with the finding that *Y. pestis* has one of the slowest substitution rates observed among bacterial pathogens (Duchêne et al., 2016). Here we found that all populations had a median rate of less than 1 substitution per year (Figure 3.3.2 C, Table S4), with the lowest rate in *Antiqua* (0.ANT) at 1 substitution every 14.1 years (1.7 x 10^{-8} subs/site/year) and the highest rate in *Pestoides* (0.PE) at 1 substitution every 1.1 years (2.1 x 10^{-7} subs/site/year). In application, this means that *Y. pestis* lineages often cannot be differentiated until at least several decades have passed, a concept referred to in the literature as the phylodynamic threshold (Lam & Duchene, 2021).

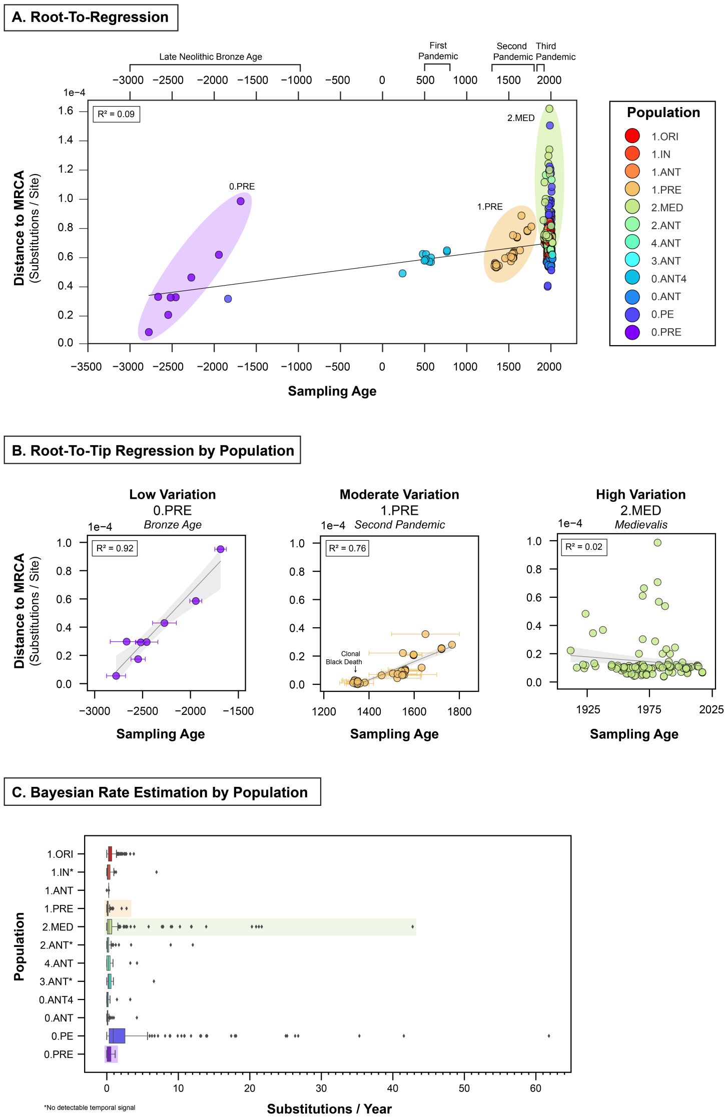


Figure 3.3.2: Substitution rate variation in *Y. pestis*. **A**: A root-to-tip regression on mean sampling age using all genomes from the maximum likelihood phylogeny. **B**: A root-to-tip regression on mean sampling age by population. The distance to the population MRCA was calculated using subtrees extracted from the maximum likelihood phylogeny. **C**: Bayesian substitution rates within and between populations. For each branch in the maximum clade credibility (MCC) trees, we extracted the estimated substitution rate (subs/site/year) and converted this to subs/year based on an alignment of 4,229,098 genomic sites.

The phylodynamic threshold has been rigorously explored in other pathogens, such as SARS-CoV-2 (Duchene, Featherstone, et al., 2020), but not explicitly in *Y. pestis*. The challenges in reconstructing intra-epidemic plague diversity have been noted previously. For example, several isolates from the Second Pandemic dated to the medieval Black Death (1346-1353) are indistinguishable clones (Spyrou et al., 2016), extinguishing any hope of reconstructing its spread from genetic evidence alone. Our median rate estimation for the Second Pandemic (1.PRE) of 1 substitution every 9.5 years (2.5 x 10^{-8} subs/site/year) is congruent with this finding. The clonal nature of the Black Death is not an exceptional event, but rather the norm based on the sampling time frame. Our results highlight a significance limitation and cautionary note for plague research, as genetic evidence alone is not suitable for reconstructing the timing of short-term, episodic epidemics.

3.3.2.2 Irreproducible Estimates

We observed two populations with detectable temporal signal associated with substantial node-dating conflicts: the *Pestoides* (0.PE) and *Antiqua* (0.ANT) biovars: both of which are paraphyletic. Conflicts were identified by comparing their estimated time to the most recent common ancestor (tMRCA) to that of their descendant populations. For example, the First Pandemic (0.ANT4) is a descendant clade of the larger *Antiqua* (0.ANT) population based on the maximum likelihood phylogeny (Figure 3.3.3). We would expect the tMRCA of the ancestral 0.ANT to pre-date the First Pandemic, for which ancient DNA calibrations are available. However, the tMRCA of 0.ANT is far too young (95% HPD: 1357 - 1797 CE), and incorrectly post-dates the tMRCA of 0.ANT4 (95% HPD: 39 - 234 CE) by more than a millennium. This outcome is somewhat paradoxical, as these populations have robust temporal signal and yet a critical examination of their divergence dates reveals they are unreliable.

This conflicting pattern has been previously described and attributed to sampling bias (Featherstone et al., 2021; Ho & Duchêne, 2020), specifically, insufficient sampling of basal branches and the presence of extensive rate variation. The two affected populations, *Antiqua* (0.ANT) and *Pestoides* (0.PE), have a low density of nodes at their roots in the maximum likelihood phylogeny (Figure 3.11.6). This pattern is also observed in another *Antiqua* population (1.ANT) which has a small sample size (N=4) and has previously been linked to rate acceleration events (Cui et al., 2013). The dates associated with these three populations (0.PE, 0.ANT, 1.ANT) should be considered non-informative.

These node-dating issues reveal a clear limitation in our approach of estimating divergence times from population-specific models. Defining *Y. pestis* population by time periods has adverse effects, as ancient plague genomes can serve as crucial calibration points for rate changes that are otherwise unsampled in extant populations. In populations with poorly sampled basal branches (0.PE, 0.ANT, 1.ANT) we expect an optimization approach to be more ideal, in which a few closely related populations are merged or select ancient DNA calibrations are introduced (Spyrou et al., 2018). Otherwise, divergence dates in these populations tend to be overly young, sometimes by more than a 1000 years (Pisarenko et al., 2021), and are difficult to replicate between studies (Table 3.3.1).

The inability to infer divergence dates due to sampling bias also has several historical implications. Perhaps the most significant concerns the emergence of plague in Africa which makes up 90% of all modern plague cases (Munyenyiwa et al., 2019), yet for which there remains not a single ancient sequence. Little progress has been made in sampling extant African plague diversity, with this region represented by only 1.5% (9/601) of all genomes. Furthermore, the oldest genetic evidence of African plague comes from the 1.ANT population, which has only four representative strains. Despite this sparse sampling, researchers have repeatedly attempted to use genomic evidence to date the first appearance of *Y. pestis* in Africa (Cui et al., 2013; Morelli et al., 2010; Pisarenko et al., 2021). The result is a complete lack of congruent dates for this event, as the majority of tMRCA estimates for 1.ANT do not overlap (Table 3.3.1). These divergence dates are of limited value for historical interpretation (Green, 2018; Nyirenda et al., 2018; Pisarenko et al., 2021) and should be treated with great skepticism.

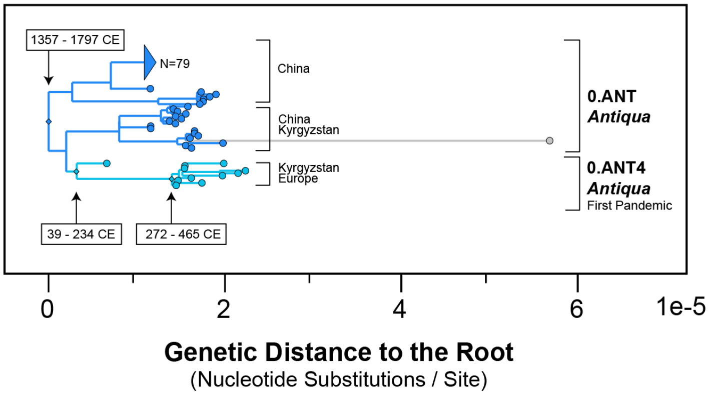


Figure 3.3.3: Ancestor-descendant relationships in the maximum likelihood phylogeny reveal tMRCA conflicts between *Antiqua* (0.ANT) and the First Pandemic (0.ANT4). Node dates (95% HPD) were estimated from the Bayesian analysis, where each population was assessed independently. Grey branches indicate outliers, as defined by the 90% confidence interval of external branch lengths from all populations.

Table 3.3.1: Bayesian estimates of the time to most recent common ancestor (tMRCA) across *Y. pestis* studies. Uncertainty surrounding the tMRCA is represented by the 95% highest posterior density (HPD) interval. A dash indicates the study did not incorporate genomes from the population

Category	Population	Morelli et al. 2010	Cui et al. 2013	Pisarenko et al. 2021	This Study
Informative Dates	1.ORI	-326, 1793	1735, 1863	1744, 1842	1806, 1901
No Temporal Signal	1.IN	-2388, 1606	1500, 1750*	1791, 1897	1651, 1913
Sampling Bias	1.ANT	-4909, 1322	1377, 1650	1483, 1704	1655, 1835
Informative Dates	1.PRE	–	1312, 1353	–	1214, 1315
Informative Dates	2.MED	-583, 1770	1550, 1800*	1413, 1653	1560, 1845
No Temporal Signal	2.ANT	-3994, 1460	1550, 1800*	1373, 1628	1509, 1852
No Temporal Signal	4.ANT	–	1200, 1700*	1611, 1816	1848, 1968
No Temporal Signal	3.ANT	–	1450, 1850*	1531, 1742	1769, 1947
Sampling Bias	0.ANT	-6857, 1199	100, 1100*	1033, 1435	1357, 1797
Informative Dates	0.ANT4	–	–	–	39, 234
Sampling Bias	0.PE	-26641, -598	-4394, 510	-377, 499	1573, 1876
Informative Dates	0.PRE	–	–	–	-3098, -2786

* Visually estimated from the published time-scaled phylogeny.

3.3.2.3 Informative Rates and Dates

Excluding populations with no detectable signal, we identified five populations with potentially informative rates and dates. These include the Bronze Age (0.PRE), *Medievalis* (2.MED), the First Pandemic (0.ANT4), the Second Pandemic (1.PRE), and the Third Pandemic (1.ORI). The Bronze Age marks the first identified appearance of *Y. pestis* in humans, and the three pandemics, along with *Medievalis*, are historically associated with high mortality and rapid spread (Xu et al., 2014). Due to this epidemiological significance, these five populations have been sampled over the longest time frames, ranging from 92 years for the Third Pandemic (1.ORI) to 1250 years for the Bronze Age (1.PRE). This affirms the importance of long-term heterochronous sampling for *Y. pestis*, made possible through the retrieval of ancient DNA (Bos et al., 2011) and recent sequencing of early 20th century culture collections (Eroshenko et al., 2021). By curating and contextualizing this new heterochronous data, we were able to detect temporal signal in extant *Y. pestis* populations without the use of ancient DNA calibrations for the first time.

Our estimates of the tMRCA for the First and Second Pandemics share a common theme, in that the genetic origins potentially pre-date the appearance of plague in traditional (i.e. European) historical narratives. For example, the earliest textual evidence of the Second Pandemic (1.PRE) in Europe comes from the Black Death (1346) (O. J. Benedictow, 2021). However, we estimate the mean tMRCA of this population to be earlier, between 1214 and 1315 CE. Similarly, the first recorded outbreaks of plague during the First Pandemic (0.ANT4) come from the Plague of Justinian (541 CE) (Little, 2007). Instead, we estimate that the strains of *Y. pestis* associated with this pandemic shared a common ancestor between 272 and 465 CE.

One explanation for these disparate timelines is sampling bias, as western European sources dominate both the genetic and historical record. Recent historical scholarship has contested Eurocentric timelines (Hashemi Shahraki et al., 2016; Varlık, 2020) by demonstrating the presence of plague in western Asia far earlier than previously thought. Arabic historical chronicles suggest that the Second Pandemic may have begun as early as the 13th century (Fancy & Green, 2021). Genetic dating appears to support these historical critiques, by expanding the timelines of past pandemics to make space for more diverse historical narratives. An alternative explanation for our earlier dates is tip date uncertainty. The radiocarbon estimates for the majority of ancient *Y. pestis* samples have confidence intervals of ±50 years or more. As we only used the mean sampling age for molecular clock models, it's possible that the true tMRCA intervals are larger and do overlap with historical estimates. How much uncertainty can be included in molecular clock models for *Y. pestis*, while still achieving convergence of parameter estimates, remains to be tested.

In contrast to the ancient pandemics, our temporal estimates of the Third Pandemic were more closely aligned to the historical evidence. We estimated that isolates from the Third Pandemic (1.ORI) shared a common ancestor between 1806 and 1901 CE, which aligns well with the timeline as reconstructed from epidemiological reports. Highly localized plague cases began appearing in southern China and (1772-1880) and later spread globally out of Hong Kong (1894-1901) (Benedict, 1988; Xu et al., 2014; Xu et al., 2019). Our estimate also overlaps with the majority of previous studies, although it is the youngest tMRCA to date (Table 3.3.1). This comparison demonstrates the reproducibility of our estimate, but also reveals how the "origin" story of the Third Pandemic continues to change. The phylogenetic root once estimated to be as old as 326 BCE (Morelli et al., 2010) is now resolved to be much younger (19^{th} century CE). This younger date is particularly intriguing, as a major epidemiological transition occurred in the 19^{th} century with the reemergence of several other notable pathogens Brüssow (2021). Reconstructing the evolutionary history of the Third Plague Pandemic may not only inform us about the epidemiology of plague, but contribute to a broader understanding of the factors that led to reemerging diseases in the modern era (Piret & Boivin, 2021).

Even less is known about the *Medievalis* population whose strains were hypothesized to be responsible for plague outbreaks in the Caspian Sea region which reoccurred throughout the 19^{th} and 20^{th} centuries (Eroshenko et al., 2021). We estimated the tMRCA of *Medievalis* (2.MED) to be between 1560 and 1845 CE, which overlaps with all previously published estimates (Table 3.3.1). While this population was once thought to have emerged as early as 583 BCE, there is now growing consensus that the earliest possible emergence was in the 16^{th} century CE. Interestingly, the Caspian Sea region appears to be a nexus of plague as the only known area where the distributions of both European and Asian *Y. pestis* strains overlap (Figure 3.3.4). This raises the interesting possibility that distinct populations of *Y. pestis* were co-circulating during the Second Pandemic, a hypothesis that ancient DNA from *Medievalis* could help elucidate. In the absence of direct genetic sampling, an alternative approach is to infer the ancestral locations and spread of plague using phylogeographic analysis.

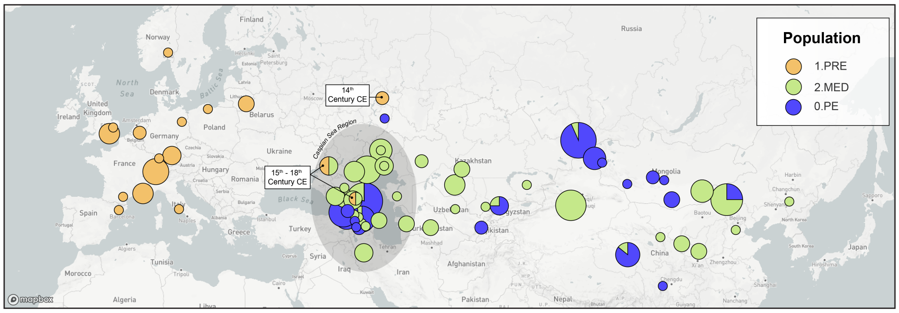


Figure 3.3.4: The geographic distributions of the Second Pandemic (1.PRE), *Medievalis* (2.MED), and *Pestoides* (0.PE) populations. The sampling location of each genome was standardized to the centroid of the associated province and/or state.

3.3.3 Estimating Ancestral Locations and Spread

Phylogeographic inference relies heavily on the degree of geographic signal in the data. To assess this in *Y. pestis*, we tested whether phylogenetic relationships correlate with sampling locations. We identified the closest genetic relative of every genome in our data set, using the maximum likelihood phylogeny. We then recorded whether these genomes were collected from the same location at three levels of resolution: (1) continent, (2) country, and (3) province. As a statistical measure of geographic structure, we reported the percent of genomes that had a closest relative sampled from the same location.

The majority of *Y. pestis* populations (6/12) were localized to a single continent (Figure 3.11.2). Of those distributed across multiple continents, geographic structure ranged from 76% to 99%. At the country level, the degree of geographic structure dropped drastically in some populations (Bronze Age 0.PRE: 38%) while remaining stable in others (3.ANT: 100%). The inverse of this pattern appeared at the province level, where *Antiqua* (3.ANT) dropped to 45% while the Bronze Age (0.PRE) was unchanged. As expected, geographic structure decreases at finer resolutions but the extent to which varies by population.

The factors which appeared to govern these patterns are wide-ranging, but primarily concern mobility. One striking aspect is the difference in host composition between populations driving this signal. We observed a correlation (R^2=0.43) between the degree of geographic structure and the percentage of samples collected from a non-human host (Figure 3.3.5). Populations primarily sampled from rodents and arthropod vectors had far more geographic structure than those found exclusively in humans. This is epidemiologically consistent with the greater mobility of human populations, which disrupt geographic clustering via long-distance spread.

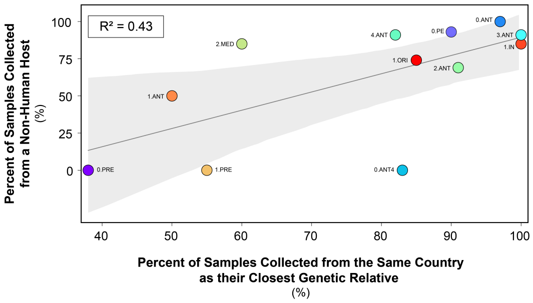


Figure 3.3.5: A linear regression of host-associations on the degree of geographic structure at the country level.

Another factor is the difference in substitution rates relative to migration rates. In populations that are spread faster (locations/year) than they evolve (substitutions/year), geographic structure decreases as identical isolates (clones) are found in different locations. Lineages of the Second Pandemic (1.PRE) exemplify this, as clonal isolates have been sampled across multiple countries (Spyrou et al., 2019), leading to uncertainty regarding the routes of spread. Along similar lines is the disparity in the sizes of locations across the globe used in the analyses. China, for example, is approximately the same size (0.94) as the European continent. Thus, at the country level, plague populations in Asia are sampled over fewer locations and have stronger geographic structure. An informative comparison is *Antiqua* (3.ANT) with 100% structure across China and Mongolia, and the Second Pandemic (1.PRE) with 55% structure across 11 European countries. Plague populations that are distributed over multiple, smaller locations have less geographic structure, leading to greater uncertainty when inferring past migrations.

These observations suggests that phylogeographic inference is best suited to populations that are slow-spreading and/or rapidly-evolving, a significant problem for *Y. pestis*, as it is both a rapidly-spreading and slow-evolving pathogen (Duchêne et al., 2016). To explore how this feature impacts our ability to infer ancestral locations for *Y. pestis* in the past, we independently fit three discrete migration models (Sagulenko et al., 2018) to the maximum likelihood phylogeny using the sampling locations by: (1) continent, (2) country, and (3) province. For each internal node, we extracted the ancestral location with the highest likelihood given the data. To explore whether genomic evidence can provide meaningful geographic estimates, we compared two case studies: the Third Pandemic, which serves as a "control" for our phylogeographic analysis, and the Second Pandemic, where the origins and spread remain contentious due to limited non-European historical evidence.

3.3.3.1 The Third Pandemic (19^{th} - 20^{th} Century CE)

The Third Pandemic of plague was closely monitored by contemporary researchers (Cantlie, 1897). As a result, the geographic origins are well-documented and firmly established. Highly localized plague cases first appeared in Yunnan, China (1772-1800), later spreading throughout the province (1800-1880) (Xu et al., 2014; Xu et al., 2019). Plague then traveled eastwards to the coast (1880-1900), where it dispersed globally out of Hong Kong (1894-1901) (Echenberg, 2002).

We estimated that the Third Pandemic (1.ORI) diverged from an ancestral *Intermedium* (1.IN) population located in Yunnan, China (probability: 1.00) (Figure 3.3.6, Figure 3.11.3). Plague then rapidly diversified (reflected by a polytomy), after which new lineages appeared in North America (probability: 0.99), South America (probability: 1.00), and Africa (probability 1.00). Due to the unresolved branch structure, we could not confidently estimate the routes of this dispersal. The migrations that could be reconstructed all occurred post this radiation, and included endemic cycling in Southeast Asia (China, Myanmar) as well as North America (USA), which led to a re-introduction of plague into South America (Peru).

The strength and specificity of our estimated origin is striking, given that we could not confidently locate the ancestral divergence for any other population (Figure 3.11.3). This may be because the Third Pandemic (1.ORI) is a direct descendant of the *Intermedium* (1.IN) population, which has strong geographic structure at the province level (87%). In addition, isolates from Yunnan fall both basal to, and within, the known diversity of the Third Pandemic (1.ORI). This combination provides strong evidence of the geographic origin, which is congruent with the historical narrative. This level of precision was only possible due to the extensive sampling of non-human hosts. Yunnan is solely represented by rodent and arthropod samples (N=18) and therefore this reservoir would be entirely invisible if only human isolates were used. Like others (Bramanti et al., 2021), we caution that the presence and location of rodent reservoirs should not be inferred from phylogenetic evidence alone. Instead, new modeling approaches have been developed (Kalkauskas et al., 2021) that could leverage multi-disciplinary sources (Xu et al., 2019) to correct for sampling biases in the genomic data.

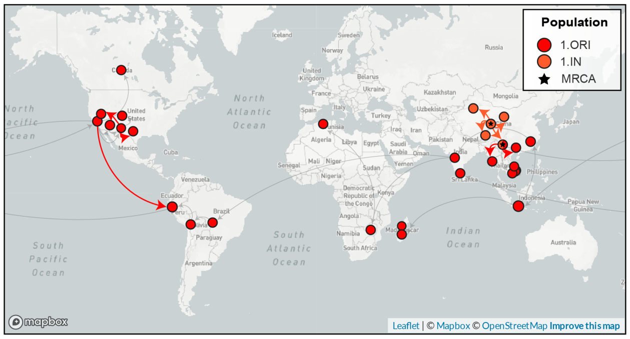


Figure 3.3.6: Geographic origins and spread of the Third Pandemic (1.ORI) and the *Intermedium* (1.IN) population. Ancestral locations were estimated by fitting a discrete migration model to the maximum likelihood phylogeny using sampling locations by province. Arrows reflect the directionality of spread, but not the precise route taken. Grey arrows indicate the migration was poorly supported by the data, with an ancestral likelihood less than 0.95 and/or a branch support bootstrap less than 95%.

3.3.3.2 The Second Pandemic (14th - 19th Century CE)

In comparison to the Third Pandemic, there is far less surviving historical evidence from the Second Pandemic. Historians have identified early accounts of plague appearing in 1346 in the Golden Horde, which encompass Central Asia and Eastern Europe (O. J. Benedictow, 2004). The disease then appears to have spread southward through the Caucasus to reach Western Asia, and westward to the Crimea, from which it dispersed throughout Europe, the Middle East, and North Africa. Plague reoccurred for several centuries in these territories, with successive waves varying in scale from localized epidemics to continent-wide outbreaks (O. J. Benedictow, 2021). In Western Europe, plague receded after 1720 and would not re-emerge again until 1899 (Shadwell, 1899), while in Western Asia the disease never disappeared (Varlık, 2020).

We estimated that the Second Pandemic (1.PRE) diverged from an ancestral population located in China (probability: 0.93) as part of the “Big Bang” polytomy. The ancestral province was poorly resolved, with the most likely location being near Xinjiang (probability): 0.64) which includes the Tien Shan mountains. The location of the Second Pandemic MRCA was also uncertain and estimated to be in Russia (probability: 0.63), specifically in Tatarstan (probability: 0.37) which was part of the Golden Horde. However, these low likelihoods indicate that our estimated origins are poorly supported by the data. With regards to spread, only four migrations could be confidently inferred (likelihood $>$ 0.95) across the full sampling time frame (500 years). The available genetic data therefore provides little definitive evidence as to the spread of plague during the Second Pandemic.

This then begs the question of whether more ancient DNA samples will improve these geographic estimates? As it currently stands, the relationships between all countries could not be resolved during the 14th century, nor among the Baltic states sampled in the 15th century, or between England and Russia in the 17th century. Furthermore, the historical evidence indicates that plague often spread to multiple countries, if not continents, in the span of a decade (Slavin, 2021). This migration rate is far higher than the substitution rate of the Second Pandemic (1.PRE), which accumulates 1 mutation every 9.5 years. The genomic data alone does not have sufficient resolving power to reconstruct the spread of short-term, episodic waves of plague. Instead, this evidence is best used in conjunction with higher-resolution evidence, such as historical case records (Featherstone et al., 2021; Roosen & Curtis, 2018).

3.4 Conclusions

We sought to contribute to five lines of debate concerning the evolutionary history of *Yersinia pestis*. The first, was whether *Y. pestis* has sufficient temporal signal (i.e. molecular clock) to accurately estimate rates and dates. Accordingly, we found that a species-wide clock model was methodologically unstable and did not lead to reproducible estimates. However, we observed significant improvements when each *Y. pestis* population was assessed independently. We therefore recommend this approach for future studies, as the full global diversity of *Y. pestis* can be utilized without down-sampling.

Second, we explored the minimum sampling time frame for *Y. pestis* that yields informative rates and dates. The lowest substitution rate was observed in *Antiqua* (0.ANT) with a median rate of 1 substitution every 14.1 years, meaning

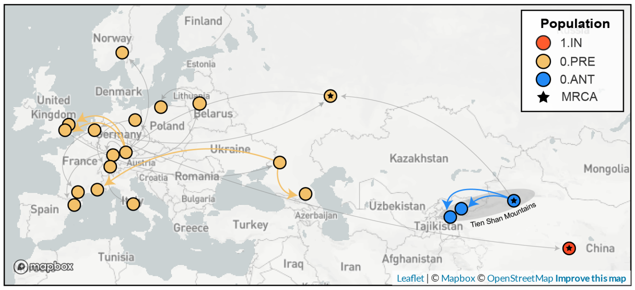


Figure 3.3.7: Geographic origins and spread of the Second Pandemic (1.PRE), the ancestral *Antiqua* (0.ANT) and descendant *Intermedium* (1.IN) populations. Ancestral locations were estimated by fitting a discrete mugration model to the maximum likelihood phylogeny using sampling locations by province. Arrows reflect the directionality of spread, but not the precise route taken. Grey arrows indicate the migration was poorly supported by the data, with an ancestral likelihood less than 0.95 and/or a branch support bootstrap less than 95%..

that some *Y. pestis* lineages cannot be differentiated until several decades have passed. In addition, we found no temporal signal in several populations (1.IN, 2.ANT, 3.ANT) which have been sampled over a period as long as 84 years. Genetic evidence alone may not be suitable for reconstructing the timing of short-term, epidemic events of plague.

In the third instance we tackled node dating disparities between studies. We explored how phylogenetic sampling bias drives this outcome, and how it can be detected and remedied with ancient DNA calibrations. In particular, we focused on the non-overlapping tMRCA estimates of the first appearance of *Y. pestis* in Africa (1.ANT). Until sampling strategies diversify, we caution that the published divergence dates for this population, and several others (*Antiqua* 0.ANT, *Pestoides* 0.PE), are of limited value for historical interpretation.

The fourth point revolved around the timing of past pandemics. We observed a common theme in which the genetic dates (tMRCAs) of pandemic *Y. pestis* potentially pre-date the historical dates by several decades or centuries. For example, we estimated the tMRCA of the Second Pandemic to be between 1214 and 1315 CE which pre-dates the Black Death (1346 - 1353 CE). Similarly, we estimated the tMRCA of the First Pandemic to be between 272 and 465 CE, also pre-dating the Plague of Justinian (531 CE). We discussed this disparity in light of methodological concerns such as radiocarbon dating uncertainty and geographic sampling biases that have historically favored European sources. We anticipate that additional samples from non-European locations will add greater clarity to how the timelines of past pandemics can be expanded to include more diverse historical narratives.

Finally, we asked whether *Y. pestis* has sufficient geographic signal to accurately infer ancestral locations and spread. As expected, geographic structure diminished at finer resolutions but also varied by population. We found that the geographic origins of the Third Pandemic (1.ORI) were unambiguously inferred to be in Yunnan province, China (likelihood=1.00) and attributed this to comprehensive sampling of rodent reservoirs. In contrast, we demonstrated how the origins and spread of the Second Pandemic (1.PRE) cannot be resolved from genetic evidence alone, as this population is exclusively sampled from human remains which have high mobility. In isolation, *Y. pestis* genomic evidence may be unsuitable for inferring point migrations and the directionality of spread. Alternatively, new methods which incorporate non-genetic evidence, such as outbreak case-occurrence records (Featherstone et al., 2021), into phylogeographic analysis presents an exciting avenue for interdisciplinary collaboration, as explicitly integrative models will complement the strengths of genetic and historical evidence, while mitigating their respective weaknesses.

3.5 Methods

3.5.1 Data Collection

Y. pestis genome sequencing projects were retrieved from the National Centre for Biotechnology Information (NCBI) databases using NCBImeta v0.7.0 (Eaton, 2020). 1657 projects were identified and comprised three genomic types. 1473 projects came from isolates sampled during the 20th and 21st centuries, which we label as "modern". Of these, (i) 586 projects were available as assembled genomic contigs (FASTA), and (ii) 887 were only available as unassembled sequences (FASTQ). An additional (iii) 184 projects came from skeletal remains with sampling ages older than the 19th century, which we label as "ancient". 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, date, and host metadata were curated by cross-referencing the original publications. Locations were transformed to latitude and longitude coordinates using GeoPy v2.0.0 and the Nominatim API (https://github.com/osm-search/Nominatim) for OpenStreetMap. Coordinates were standardized at the level of country and province/state, using the centroid of each. Collection dates were standardized according to their year and recording uncertainty arising from missing data and radiocarbon estimates. 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 data sets 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.

3.5.2 Sequence Alignment

Modern assembled genomes were aligned to the reference genome using snippy v 4.6.0 (https://github.com/tseemann/snippy), a pipeline for core genome alignments. Default parameters were used, along with the following minimum thresholds: depth of 10X, base quality of 20, mapping quality of 30, major allele frequency of 0.9. Modern genomes were excluded if the number of sites covered at a minimum depth of 10X was less than 70% of the reference genome. After applying this filter, 540 modern genomes remained.

Ancient unassembled genomes were downloaded from the SRA database in FASTQ format using the SRA Toolkit. Pre-processing and alignment to the reference genome was performed using the nf-core/eager pipeline v2.2.1, a reproducible workflow for ancient genome reconstruction (Yates et al., 2021). Default parameters were used, along with the following minimum filters: read length of 35 bp, an edit distance of 0.01, and a 16 bp seed length. Only merged reads were retained from paired end-sequencing projects. Ancient genomes were removed if the number of sites covered at a minimum depth of 3X was less than 70% of the reference genome. After applying this filter, 61 ancient genomes remained.

A multiple sequence alignment was constructed using the snippy core module of the *snippy* v4.6.0 pipeline. The output alignment was filtered to only include chromosomal sites that were present in at least 95% of samples (i.e. a missing data threshold of 5%). The filtered alignment included 10,249 variant positions exclusive to *Y. pestis*, with 3,844 sites shared by at least two strains.

3.5.3 Maximum Likelihood Phylogenetic Analysis

Model selection was performed on the full data set (N=601) using Modelfinder (Kalyaanamoorthy et al., 2017) which identified the K3Pu+F+I model as the optimal choice based on the Bayesian Information Criterion (BIC). The K3P model, also known as K81, estimates substitution rates using three categories, in this case: (1) A<->C equals G<->T, (2) A<->G equals C <->T, and (3) A<->T equals C<->G). The "u+F"suffix indicates that base frequencies will be empirically evaluated and thus are not assumed to be equal. The "+I" suffix indicates that a proportion of the alignment includes invariable sites (i.e. non-SNPS),

A maximum likelihood phylogeny was estimated for this data across 10 independent runs of IQTREE2 (Minh et al., 2020). Branch support was evaluated using 1000 iterations of the ultrafast bootstrap approximation (Hoang et al., 2018), with a threshold of 95% required for strong support.

3.5.4 Data Partitions

The full multiple sequence alignment was alternatively split into 12 populations, referred to as the population data sets. These populations were defined by the intersection of the following nomenclature systems: the major branches, metabolic biovars, and historical time period (Table S1). One sample was excluded, a *Pestoides* isolate from the Bronze Age (Strain RT5, BioSample Accession SAMEA104488961), as this population would be of size N=1.

In an attempt to improve the performance and convergence of molecular clock analyses, a subsampled data set was also constructed. Populations 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 samples, which is a 68% reduction from the original data set.

3.5.5 Estimating Rates of Evolutionary Change

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 the mean sampling age using the *statsmodels* python package. Given the relative simplicity of a regression model, the full data set of 601 genomes was used.

For the second test of temporal signal, a Bayesian Evaluation of Temporal Signal (BETS) was conducted. This consisted of running four model configurations: either with or without sampling dates, and under a strict or uncorrelated lognormal relaxed clock models (strict and UCLN, respectively). We calculated the log marginal likelihood under each model configuration using stepping-stone sampling as implemented in BEAST v1.10 (Suchard et al., 2018). To this end, we ran 200 path steps, each with a Markov chain Monte Carlo (MCMC) of length 10^6 steps. In addition to the clock model we used a constant-size coalescent tree prior, a GTR+gamma nucleotide substitution model.

Importantly, the models involved priors that were proper for all parameters, which is essential for marginal likelihood calculations (Baele et al., 2013). In particular, the molecular clock rate (i.e. the mean of the UCLN clock model or the global rate of the strict clock) had a continuous time Markov chain reference prior (M. A. R. Ferreira & Suchard, 2008), the population size of the constant-size coalescent an exponential prior distribution with mean 10, and the standard deviation of the UCLN had an exponential prior with mean 0.33. Marginal likelihood estimation with stepping- stone sampling does not require from the posterior distribution. To obtain the posterior distribution we used an MCMC of 10^9 steps, sampling every 10^3 steps. For situations where the effective sample size (ESS) of any parameters was below 200 we increased the chain length by 50% and reduced sampling frequency accordingly.

3.5.6 Estimating Ancestral Locations and Spread

To explore underlying phylogeography, we performed ancestral state reconstruction using the maximum likelihood method implemented in TreeTime (Sagulenko et al., 2018). We independently fit three discrete mugration models to the maximum likelihood phylogeny using the sampling locations by: (1) continent, (2) country, and (3) province. The mapping of countries to continents was defined according to the open-source resource GeoJSONRegions (https://geojson-maps.ash.ms/). For each internal node, we extracted the ancestral location with the highest likelihood given the data.

We also conducted a discrete trait analysis in BEAST (Lemey et al., 2009; Suchard et al., 2018). Country of sample origin was chosen as the discrete trait of interest. A coalescent constant population size tree prior was chosen with an exponential prior placed on the effective population size with mean 100000. We modeled evolutionary rate with an uncorrelated relaxed lognormal clock, with a CTMC scale prior on the mean and exponential prior with mean 1/3 on the standard deviation of the underlying lognormal distribution (Drummond et al., 2006). A GTR+gamma nucleotide substitution model with estimated base frequencies for 1.ORI, 1.PRE, 0.ANT4, and 0.PRE. The same settings were used for 2.MED with the exception of swapping the GTR+gamma model to an HKY+gamma model. MCMC chains were run for 10^7 steps with sampling every 10^3 steps. We used logCombiner to combine between 3-5 replicate runs, with 10% burnin, for each clade to achieve ESS above 200 for each parameter and Maximum Clade Credibility (MCC) trees (Drummond & Rambaut, 2007).

3.5.7 Visualization

Data visualization was performed using the python package seaborn (Waskom, 2021) and Auspice (Hadfield et al., 2018), a component of the Nextstrain visualization suite.

3.6 Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council of Canada (#20008499), CIFAR humans and the microbiome program (HP), and the MacDATA Institute (McMaster University, Canada). This research was enabled in part by support provided by Compute Ontario (https://www.computeontario.ca/) and Compute Canada (https://www.computecanada.ca). We would like to thank Madeline Tapson, Dr. Dan Salkeld, and Dr. Jennifer Klunk for their expertise in contextualizing the ecology and evolutionary history of plague. We also thank Jessica Hider and Marie-Hélène B.-Hardy for discussions on the interpretation of genomic data from zoonotic pathogens. We are indebted to Dr. Ana Duggan and Dr. Emil Karpinski for their insight regarding Bayesian methods for phylogenetic analysis. We thank members of the Sherman Centre for Digital Scholarship, including Dr. Andrea Zeffiro, Dr. John Fink, Dr. Matthew Davis, and Dr. Amanda Montague, for their assistance in developing the genomic database. Finally, we would like to thank all past and present members of the McMaster Ancient DNA Centre and the Golding Lab at McMaster University.

3.7 Author Contributions

K.E, G.B.G, and H.N.P designed the study. K.E, L.F., and S.D performed computational analysis. A.G.C and N.V. provided historical sources and interpretation. E.C.H critiqued and revised the computational methods and discussion. G.B.G provided access to computational resources and data storage. H.N.P and G.B.G supervised the study. K.E wrote the manuscript with contributions from all co-authors.

3.8 Competing Interests Statement

The authors declare no competing interests.

3.9 Data Availability

All genomic sequences used in this study are publicly available and were downloaded from the National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/). Genomic metadata and accession numbers are described in Table S8.

3.10 Code Availability

A visual overview of the computational methods is provided in Figure 3.11.7 and is publicly available as a snakemake pipeline (https://github.com/ktmeaton/plague-phylogeography/).

3.11 Supplementary Information

3.11.1 Tables

- Table S1-S8

3.11.2 Figures

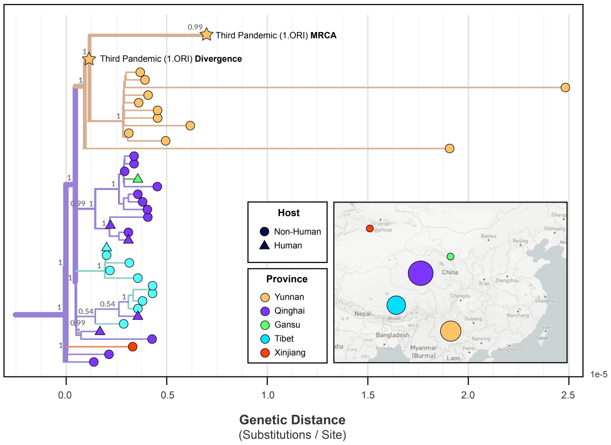


Figure 3.11.1: Geographic structure of the *Intermedium* (1.IN) population which is ancestral to the Third Pandemic (1.ORI). Branches are colored based on a discrete, ancestral state reconstruction using sampling location (province). Branch labels indicate the likelihood of the ancestral location given the data.

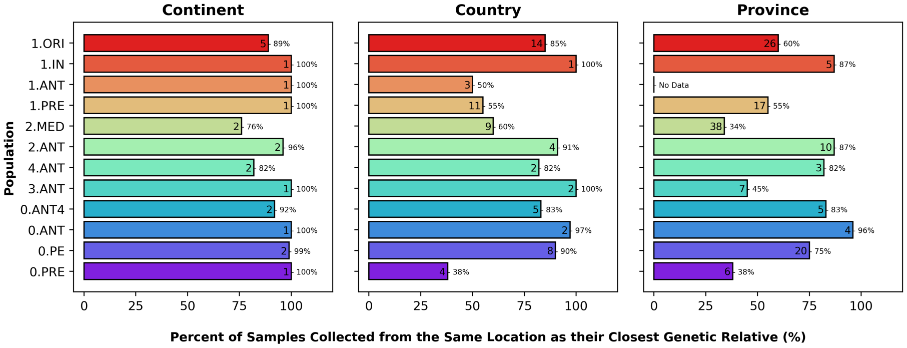


Figure 3.11.2: Geographic structure by population according to the percent of genomes sampled from the same location as their closest genetic relative. Internal bar labels indicate the number of locations sampled.

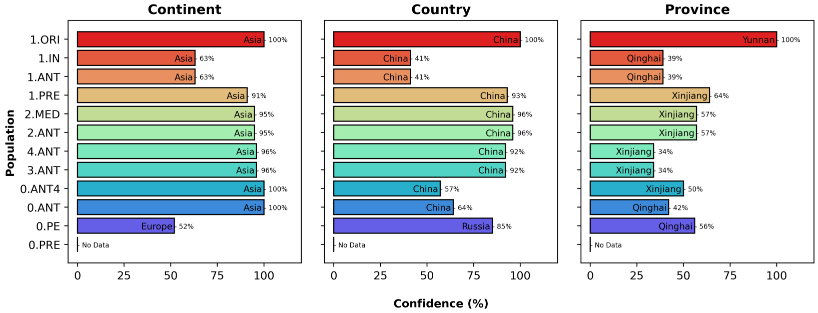


Figure 3.11.3: Geographic locations of the ancestral populations from which each population diverged. Ancestral locations were estimated by fitting discrete mugration models to the maximum likelihood phylogeny using sampling locations. Internal bar labels indicate the location with the highest confidence given the data.

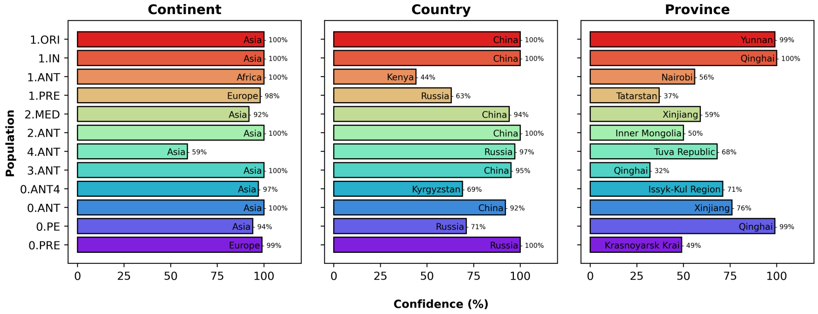


Figure 3.11.4: Geographic locations of the most recent common ancestors (MRCA) by population. Ancestral locations were estimated by fitting discrete mugration models to the maximum likelihood phylogeny using sampling locations. Internal bar labels indicate the location with the highest confidence given the data.

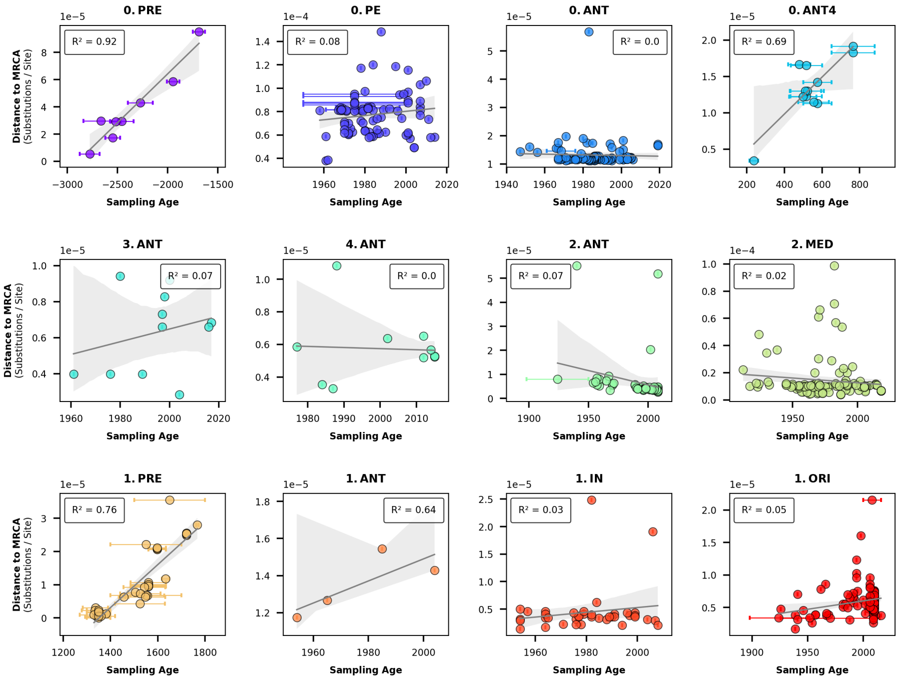


Figure 3.11.5: Population-specific rate variation in *Yersinia pestis* as observed through regressions of root-to-tip distance on sampling age. The distance to the population MRCA was calculated using subtrees extracted from the maximum likelihood phylogeny.

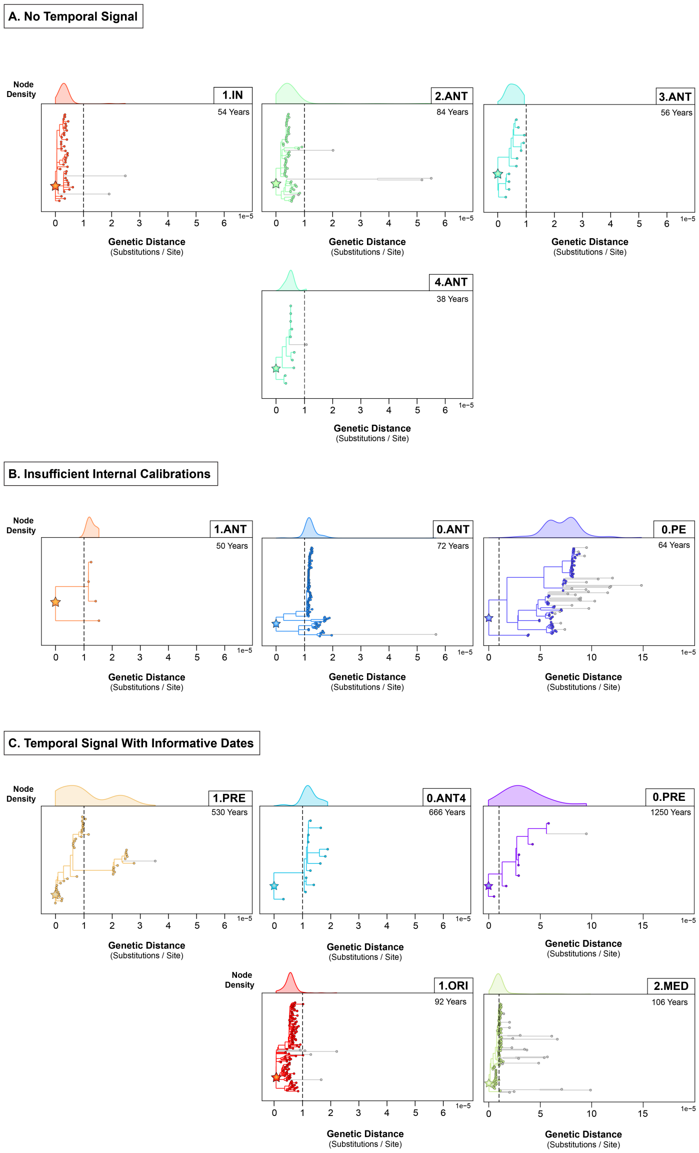


Figure 3.11.6: The subtrees extracted from the maximum likelihood phylogeny for the *Yersinia pestis* populations with (A) no detectable temporal signal, (B) insufficient internal calibrations, and (C) informative rates and dates. Stars indicate the node representing the most recent common ancestor (MRCA). Grey branches indicate outliers, as defined by the 90% confidence interval of external branch lengths from all populations.

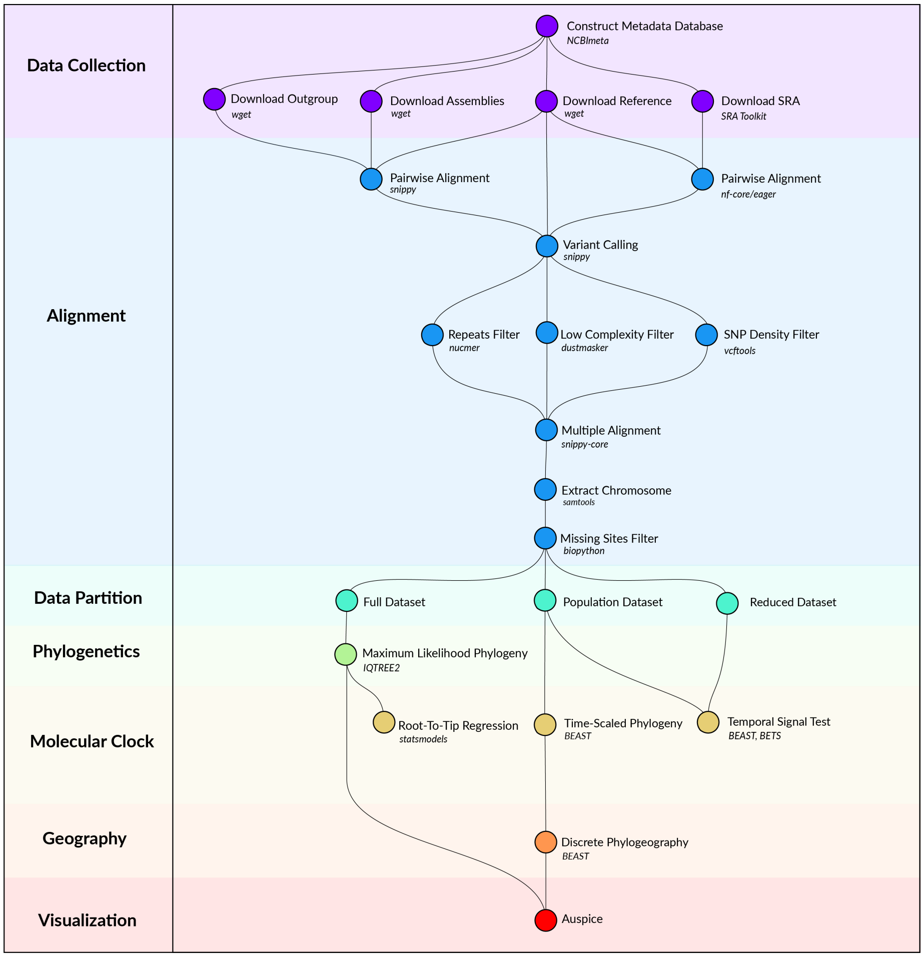


Figure 3.11.7: Computational methods workflow.

Chapter 4

Plague in Denmark (1000-1800): A longitudinal study of *Yersinia pestis*

Prepared 08 December 2021 for submission to
The Proceedings of the National Academy of Sciences

Katherine Eaton*[1,2], Ravneet Sidhu*[1,3], Jennifer Klunk[1,4], Julia Gamble[5], Jesper Boldsen[6], Ann G. Carmichael[7], Nükhet Varlık[8], Sebastian Duchene[9], Leo Featherstone[9], Vaughan Grimes[10], G. Brian Golding[3], Sharon DeWitte[11], Hendrik N. Poinar[1,2,12,13,14]

*Contributed equally.

[1]McMaster Ancient DNA Centre, McMaster University, Hamilton, Canada.
[2]Department of Anthropology, McMaster University, Hamilton, Canada.
[3]Department of Biology, McMaster University, Hamilton, Canada.
[4]Daicel Arbor Biosciences, Ann Arbor, USA.
[5]Department of Anthropology, University of Manitoba, Winnipeg, Canada.
[6]Department of Forensic Medicine, Unit of Anthropology (ADBOU), University of Southern Denmark, Odense, Denmark.
[7]Department of History, Indiana University Bloomington, Bloomington, USA.
[8]Department of History, Rutgers University-Newark, Newark, USA.
[9]The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Australia.
[10]Department of Archaeology, Memorial University of Newfoundland, St. Johns, Canada.
[11]Department of Anthropology, University of South Carolina, Columbia, USA.
[12]Department of Biochemistry, McMaster University, Hamilton, Canada.

[13]Michael G. DeGroote Institute of Infectious Disease Research, McMaster University, Hamilton, Canada.
[14]Canadian Institute for Advanced Research, Toronto, Canada.

4.1 Abstract

The epidemiology of plague in the past is highly controversial, owing to the scarcity and ambiguity of historical evidence. A frequent source of debate is the re-emergence and continuity of plague in Europe during the 14th to 18th centuries CE. Scandinavia is particularly underrepresented in the historical archives, despite having a uniquely long history of plague (5000 years) as revealed through ancient DNA analysis. To better understand the historical epidemiology of plague in this region, we performed in-depth (N=298), longitudinal screening (800 years) for the plague bacterium, *Yersinia pestis*, across 13 archaeological sites in Denmark. We captured the emergence and continuity of *Y. pestis* in this region over a period of 400 years (14th - 17th century CE), for which the plague-positivity rate was 8.3% (3.3% - 14.3% by site). These results deepen the epidemiological link between the plague bacterium, *Y. pestis*, and the unknown *pestilence* that afflicted medieval and early modern Europe. Furthermore, this study paves the way for the next generation of historical disease research, in which hypotheses concerning mortality can be tested using population-scale, genomic evidence from ancient pathogens.

4.2 Introduction

Europe endured a 500-year long pandemic from the 14th to 18th centuries CE (Spyrou et al., 2019). During this period, mysterious outbreaks reoccurred every 10 years with mortality estimates as high as 50% during the infamous Black Death (1346-1353) (Christensen, 2003). Paleogeneticists have increasingly identified the plague bacterium *Yersinia pestis* as the most likely agent, although the epidemiology of this pandemic remains controversial (Ole Jørgen Benedictow, 2016). The major source of debate concerns two aspects: mortality and spread. The ecology of *Y. pestis* is highly complex, and involves both zoonotic "spillover" from rodent populations as well as inter-human transmission (Perry & Fetherston, 1997). As a result, both disease exposure and spread are known to vary between regions and over time (Ole Jørgen Benedictow, 2016). These differences are challenging to reconcile, and have led to significant controversy concerning the location of plague reservoirs in the past (Bramanti et al., 2021).

Recent studies have explored this question by synthesizing genetic evidence (Bramanti et al., 2021) and historical records (Schmid et al., 2015; Yue et al., 2017) across Europe. These sources have significant geographic gaps, such as the complete lack of evidence from Scandinavia in digitized databases (Roosen & Curtis, 2018). This gap has been attributed to the sparseness of historical sources and ambiguity with regards to disease terminology during the medieval period (Christensen, 2003). However, recent ancient DNA research (Rascovan et al., 2019) has revealed that the history of plague in Scandinavia is among the oldest in the world, and established the presence of *Y. pestis* in Sweden 5000 years ago. This raises the possibility of long-term persistence of plague in Scandinavia, with *Y. pestis* re-emerging as a local, endemic disease.

To evaluate the possibility of undocumented plague persistence, we screened for the presence of *Y. pestis* in the Anthropological DataBase Odense University (ADBOU) collection. This extraordinary collection includes preserved and curated skeletal remains from over 16,000 Danish individuals, dated from the Viking Age to the Early Modern period. To ensure a wide variety of locations were represented, we sampled 298 individuals across 13 archaeological sites from the mainland (Jutland), as well as two islands (Funen and Lolland). Based on the skeletal dates, these individuals represent 800 years of population history (1000-1800 CE) which includes both the known pandemic period in Denmark (1350-1657) and the quiescent periods (1000-1350 CE, 1658-1800) for which no outbreaks of plague are historically documented (Ole Jørgen Benedictow, 2016).

4.3 Results and Discussion

We detected *Y. pestis* in 7 archaeological sites using PCR assays and targeted sequencing (Figure 4.3.1 A). Across the 7 sites, 8.3% of individuals (13/157) tested positive for *Y. pestis*, ranging from 3.3% at Ribe Lindegärden to 14.3% at Hågerup. This positivity rate could be considered an underestimate of the 'true' prevalence of *Y. pestis* in Danish populations, due to variable DNA preservation. On the other hand, it may be an overestimate due to the osteological paradox (Wood et al., 1992), in which mortality is selective and the deceased are not representative of the living population. While the exact extrapolation is unclear, our *Y. pestis* positivity rate (3.3 - 14.3%) does align with mortality estimates (5 - 20%) during the later epidemics of the medieval and early modern period (DeWitte & Kowaleski, 2017; Slavin, 2021).

Of the 13 plague-positive individuals, 9 had sufficient sequencing depth (>3X) of the *Y. pestis* chromosome for phylogenetic analysis (Figure 4.3.2 D). To estimate a time-scaled phylogeny and dates for these 9 samples, we fit a relaxed molecular clock to an alignment of *Y. pestis* genomes which included 40 other isolates (Figure 4.3.1 B). We observed that all Danish strains clustered strongly (posterior: 1.0) within the known diversity of medieval and early modern *Y. pestis* in Europe (Figure 4.3.3). We found no evidence to suggest that Neolithic lineages of *Y. pestis* in Scandinavia (5000 YBP) (Rascovan et al., 2019) left descendants in medieval Danish populations. If long-term persistence of *Y. pestis* did occur in this region, it fell outside the geographic and temporal scope of this study.

We found no evidence of *Y. pestis* in Denmark between 1000 and 1300 CE.

The factors influencing the preservation of ancient DNA are wide-ranging and complex, thus the absence of evidence cannot prove evidence of absence. That being said, we sampled a minimum of 85 individuals and a maximum of 165 individuals that pre-date the 14th century (Figure 4.3.2 A). Taking the mean positivity rate observed in this study (8.3%), we would expect to detect *Y. pestis* in 7 to 13 individuals from this time frame if it were present. We therefore interpret our negative results from this period as tentative evidence that *Y. pestis* was a relatively new pathogen in medieval Denmark, that did not become abundant and/or widespread until at least the 14th century.

The earliest evidence of *Y. pestis* in Denmark was found in the town of Ribe. Two individuals were associated with *Y. pestis* from the first half of the 14th century, dated to 1333 (1301-1366) and 1350 (1319-1383). These estimates are highly congruent with the historical record, as the first documented appearance of plague in Denmark was at Ribe in 1349 (Lenz & Hybel, 2016). Furthermore, these strains fell within an unresolved cluster (posterior: 0.15) of samples from Northern and Western Europe (Figure 4.3.3) which has previously been linked to clonal spread of the Black Death (1343-1356) (Spyrou et al., 2019). Our molecular dates support this historical association, albeit only weakly, as the precise epidemic period cannot be resolved due to the large confidence intervals of our estimates (>50 years).

The next period in which we identified *Y. pestis* was in the latter half of the 14th century. A third isolate from Ribe was dated to 1370 (1336-1408) and strongly clustered (posterior: 0.99) with post-Black Death samples from The Netherlands and Russia. These samples have previously been attributed to the *pestis secunda* (1357-1366) (Namouchi et al., 2018), although we find the *pestis tertia* (1364-76) (Slavin, 2021) to be an equally likely candidate. This clade also has broader epidemiological significance, as it is directly ancestral to the Third Pandemic of plague (19th-20th century) (Spyrou et al., 2019). Our results therefore reveal new global connections, as the same lineage that afflicted medieval Danish populations would later re-emerge to cause modern epidemics of plague, including the recent outbreaks in Madagascar (Nguyen et al., 2018).

We observed a gap in the continuity of plague at Ribe, as no *Y. pestis* was detected there between 1408 and 1484. This was surprising, as 86% of individuals (43/50) from this site were archaeologically dated to between 1400 and 1536. Instead, the distribution of *Y. pestis* appeared to shift during this period from the eastern coast of Jutland to the western coast. We recovered 3 distinct, and possibly contemporaneous, isolates of *Y. pestis* from 3 sites near Horsens dated to 1429 (1392-1467), 1433 (1403-1464) and 1457 (1427-1487). These genomes were most closely related to individuals sampled in Germany, Lithuania, Poland, and England (Figure 4.3.3). This geographic association parallels the historical record, in which outbreaks in Denmark coincided with those in the Baltic region (Slavin, 2021). However, recent studies have demonstrated that the directionality and spread of zoonotic diseases cannot be robustly inferred from genomic data alone (Eaton et al., Submitted, 2021; Kalkauskas et al., 2021). Instead, our results establish an epidemiological link between *Y. pestis* and historical case records in Denmark, which could be jointly modeled with greater resolving power (Featherstone et al., 2021) in future work.

In the 16th century, we once again observed *Y. pestis* at Ribe. We dated two *Y. pestis* isolates from this region to 1513 (1484-1546) and 1525 (1494-1560). Furthermore, we also found evidence of *Y. pestis* in the northern site of Faldborg dated to 1594 (1550-1649). As an estimate of plague's disappearance (1649), this is congruent with the historical record which documents the last recorded outbreak of plague in Jutland to last from 1654-1657 (Ole Jørgen Benedictow, 2016). We found no evidence of *Y. pestis* in Denmark after this point, specifically between 1649 and 1800 CE. However, no individuals definitively post-date 1649 CE, although this period could include a maximum of 70 individuals (Figure 4.3.2 A). We would therefore expect to detect *Y. pestis* in 0 to 2 individuals (3.3%) from this time frame if it were present. Our results do not differ from this expectation, and are therefore not informative with regards to the disappearance of *Y. pestis* in Denmark. To address this question, additional samples would be required from the 17th and 18th centuries.

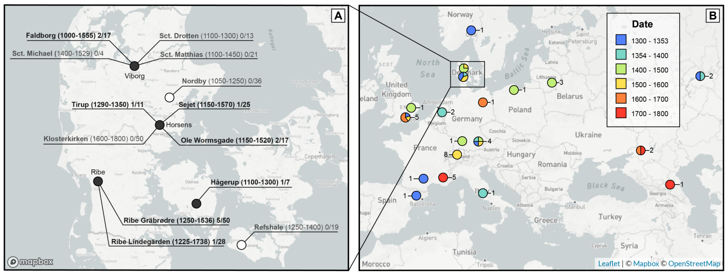
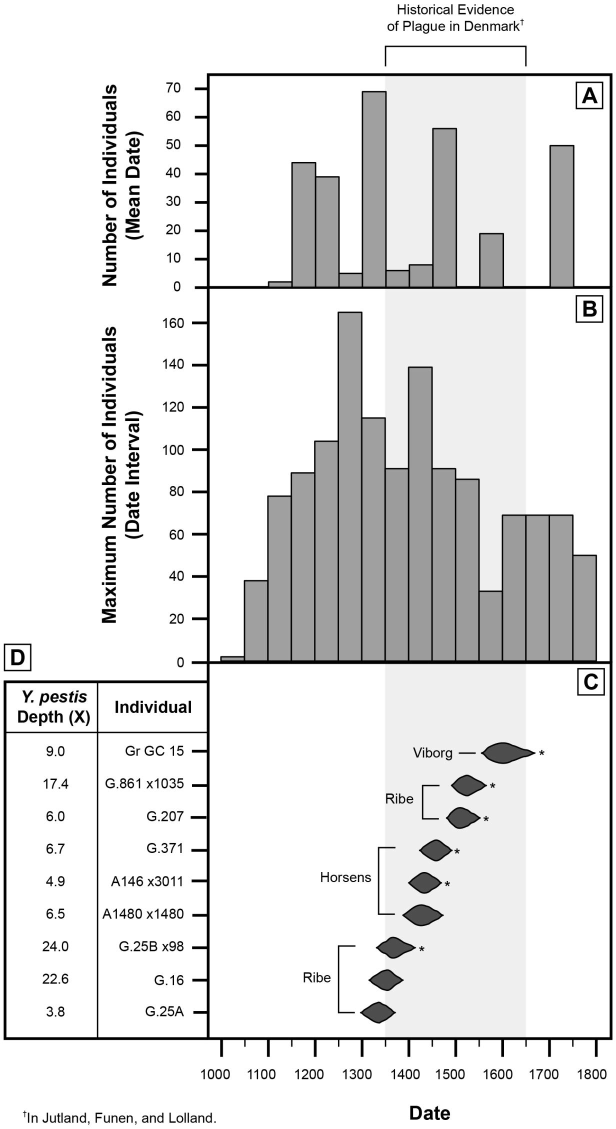


Figure 4.3.1: Geographic distribution of 298 archaeological samples used in this study. **A**. Map of 6 municipalities sampled in Denmark encompassing 13 archaeological sites. Site labels indicate: Archaeological Site (Earliest Date Sampled - Latest Date Sampled) *Y. pestis* positive individuals / total individuals. Plague positive sites are bolded. **B**. Map of 49 *Y. pestis* genomes used for phylogenetic analysis. The sampling locations were standardized to the centroid of the associated province/state. Colors indicate the sampling dates as estimated from the Bayesian molecular clock analysis. Numbered labels indicate the number of genomes sampled from each location.

4.4 Conclusion

This study marks the first population-level analysis of ancient *Y. pestis*, where we performed in-depth (N=298), longitudinal sampling (800 years) within a single country (Denmark). We describe the earliest known appearance of *Y. pestis* in Denmark (14th century), and document the continuity of this pathogen in



†In Jutland, Funen, and Lolland.

Figure 4.3.2: Temporal distribution of archaeological samples used in this study. **A**. Mean skeletal dates for all individuals (N=298). **B**. Skeletal date intervals for all individuals (N=298) using a bin width of 50 years. **C**. Distribution of *Y. pestis* tip-dates for plague-positive individuals (N=9) according to the 95% highest posterior density (HPD) from the Bayesian molecular clock analysis. Asterisks indicate the phylogenetic placement had strong posterior support (>= 0.95). **D**. Mean sequencing depth of the *Y. pestis* chromosome.

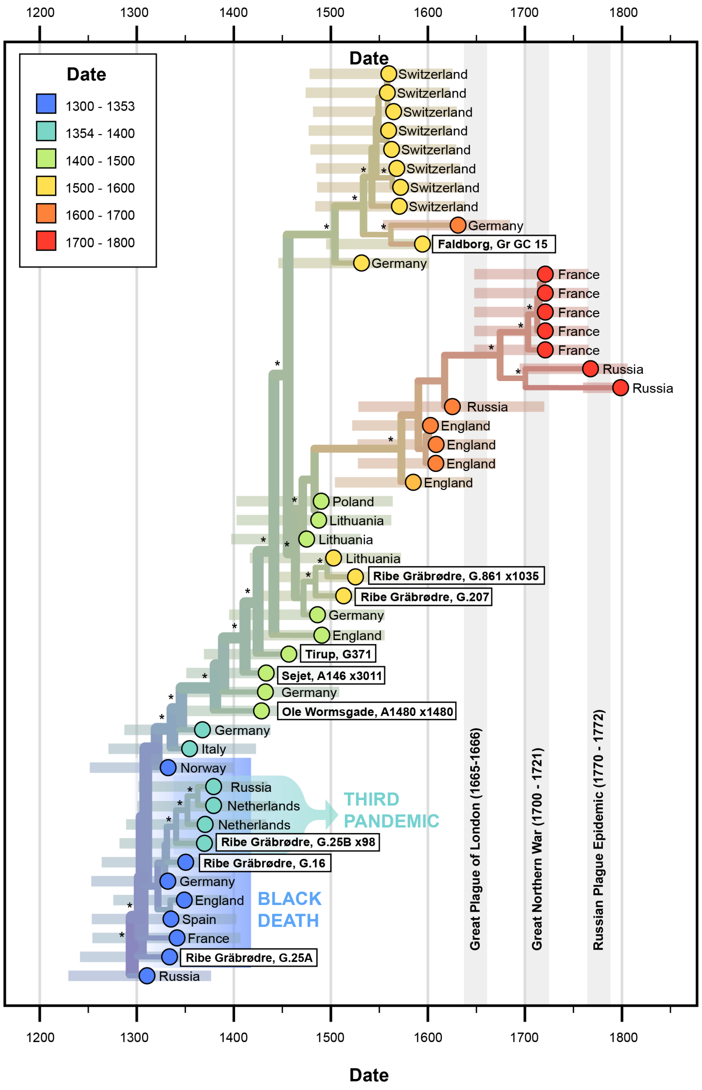


Figure 4.3.3: Maximum-clade credibility (MCC) tree depicting a time-scaled phylogeny of 49 European *Y. pestis* genomes. Asterisks indicate clades with strong posterior support (>=0.95). Colors indicate the mean sampling dates as estimated from the Bayesian molecular clock analysis. Bars indicate tip-dating uncertainty, as represented by the 95% highest posterior density (HPD) interval.

Scandinavia over a period of 400 years (17$^{\text{th}}$ century). Furthermore, we provide the first positivity rates of historical plague from molecular evidence, as we detected *Y. pestis* in 8.3% of Danish individuals. Our phylogenetic analysis was highly congruent with the sparse textual evidence of *pestilence* in Denmark, with regards to the timing of outbreaks and geographic ties to the Baltic region. We also provide novel evidence of plague exposure among Danish populations, such as the site of Tirup, where there is no surviving historical evidence. These results are of importance for both researchers of plague and other infectious diseases, as they (1) illuminate undocumented pathogens in the historical record, (2) reveal new connections between our past and present experience of plague, (3) broaden our understanding of the epidemiology of re-emerging diseases.

4.5 Materials and Methods

We sampled 298 individuals across 13 archaeological sites in Denmark (Figure 4.3.2 A, Table S1). Site occupation dates spanned from the 11$^{\text{th}}$ to the 19$^{\text{th}}$ century CE. We estimated individual date ranges based on burial position, which was categorized according to cultural shifts that occurred in Denmark throughout the medieval and early modern period (Boldsen, 2009). When the original stratigraphic context was preserved, we refined these individual estimates further. For individuals with ambiguous or conflicting archaeological estimates, we performed radiocarbon dating when additional destructive sampling was permitted.

DNA was extracted from teeth and dental pulp according to a specialized protocol for ancient DNA (Dabney et al., 2013). Reagent blanks were introduced as negative controls to monitor DNA contamination in subsequent steps. We screened for plague using a PCR assay that targets the *pla* virulence gene in *Yersinia pestis* (Wagner et al., 2014). Extracts showing amplification in at least 4/6 replicates were converted into paired-end sequencing libraries (Kircher et al., 2012). Targeted capture of the *Y. pestis* genome was performed using previously designed probes (Wagner et al., 2014) and sequenced on an Illumina platform.

Sequenced molecules were aligned to a reference genome using the *nf-core/eager* pipeline (Yates et al., 2021). To phylogenetically place these new samples, we downloaded a comparative dataset of 40 high-coverage *Y. pestis* genomes (>3X) dated between the 14$^{\text{th}}$ and 18$^{\text{th}}$ centuries. We then constructed a multiple alignment with the snippy pipeline (https://github.com/tseemann/snippy), which included 356 variation positions and 4,289,810 constant sites.

To tip-date each genome, we performed a Bayesian Evaluation of Temporal Signal (BETS) (Duchene, Lemey, et al., 2020) with BEAST2 (Bouckaert et al., 2019). We assumed a constant population size and compared the use of a strict clock and an uncorrelated lognormal (UCLN) relaxed clock. Diffuse normal priors were constructed for all tip-dates, using the mean radiocarbon/mortuary date and half the uncertainty as the standard deviation. All Danish samples were assigned equivalent priors with a mean date of 1330 CE and a standard deviation of 115 years. Bayes factors were calculated by comparing the marginal likelihoods of each candidate model, as estimated with a generalized stepping stone (GSS) computation. The model with the highest marginal likelihood was then run for 100,000,000 generations to ensure the effective sample size (ESS) of all relevant parameters was greater than 200.

Data visualization was performed using the python package *seaborn* and *auspice*, a component of the Nextstrain visualization suite (Hadfield et al., 2018).

4.6 Data Availability

Raw sequence reads have been deposited in NCBI BioProject PRJNAXXXXX. Archaeological metadata is provided in the supplementary information (Table S1).

4.7 Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council of Canada (#20008499) and the MacDATA Institute (McMaster University, Canada). This research was enabled in part by support provided by Compute Ontario (https://www.computeontario.ca/) and Compute Canada (https://www.computecanada.ca). We would like to thank Julianna Stangroom, Michael Klowak, Dr. Emil Karpinski, Dr. Matthew Emery, and Dr. Stephanie Marciniak for their assistance in laboratory procedures. We also thank Dr. Ana Duggan for her insight regarding Bayesian methods for phylogenetic analysis. We thank members of the Sherman Centre for Digital Scholarship, including Dr. Andrea Zeffiro, Dr. John Fink, Dr. Matthew Davis, and Dr. Amanda Montague, for their assistance in developing the underlying genomic database. Finally, we would like to thank all past and present members of the McMaster Ancient DNA Centre and the Golding Lab at McMaster University.

4.8 Author Contributions

K.E, R.S, J.K, and H.N.P designed the study. J.G, J.B, and S.D provided access to archaeological sites and materials. V.G performed radiocarbon dating. K.E, R.S, and J.K performed laboratory analysis. A.G.C and N.V. provided historical sources and interpretation. S.D and L.F critiqued and revised the computational methods and discussion. G.B.G provided access to computational resources and data storage. H.N.P and G.B.G supervised the study. K.E wrote the manuscript with contributions from all co-authors.

4.9 Competing Interests Statement

The authors declare no competing interests.

4.10 Supplementary Information

4.10.1 Tables

- Table S1

Chapter 5

Conclusion

5.1 Main Findings and Contributions

In this dissertation, I developed computational methods for genomics research and used them to reconstruct past and present pandemics of plague. In Chapter 2, I resented a novel software called NCBImeta that facilitates the acquisition of sequence data and metadata from the NCBI repository. This specialized tool supports genomics research in the era of big data, where manual processing of abundant sequence records (10,000+) is impossible. As a paper on software development, its contributions and significance to the field of anthropology are understandably unclear. I targeted this article exclusively towards computational biologists because, at the time, few anthropologists had expressed interest in the issue of collecting and curating sequencing data. Reflecting this, NCBImeta has mainly been cited across biological fields including studies of the human microbiome (Agostinetto et al., 2021), plant-associated bacteria in agriculture (Strafella et al., 2021), and emerging infectious diseases in public health (Matthew Gopez & Philip Mabon, *personal communication*, https://github.com/ktmeaton/NCBImeta/pull/9).

In 2021, I took a more active approach in my discipline and used this software to support several bodies of anthropological research. NCBImeta was recently used in an environmental reconstruction of Beringia (Murchie et al., Accepted, 2021), the former land-bridge that facilitated early human migrations into North America from northeast Asia. The study by Murchie et al. furthers our understanding of the peopling of the Americas, and the possible interactions between early human populations and large animals (i.e. megafauna) before the Last Glacial Period (~12,000 years ago). NCBImeta was also recently used to curate sequence data in a case study of the zoonotic disease brucellosis in the 14th century (Hider et al., In Prep). The pioneering work by Hider et al. demonstrates how pathogen DNA preserves differently throughout the body, ranging from being the dominant microorganism in several tissues while being completely absent in others. It raises an important cautionary note for ancient DNA analysis and the anthropology of disease, by empirically demonstrating how sampling strategies can bias our understanding of what diseases were present in past populations.

In Chapter 3, I explored the challenges in estimating *where* and *when* plague appeared in the past, and why these estimates are often not reproducible between studies. I used the software tool from Chapter 2 to collect all publicly available *Y. pestis* genomes, and carefully curated their collection dates, locations, and hosts. My co-authors and I then used this data set for phylodynamic analysis, and devised a new approach for modeling the rates of evolutionary change (i.e. molecular clock). We used these results to explain why divergence dates varied between studies, and outlined a critical framework for identifying which divergence dates should be considered non-informative. In addition, we found that past pandemics of plague may have emerged decades, or even centuries, before they were historically documented in European sources. These early dates are in agreement with recent historical work that examines more diverse (i.e. non-European) sources. Through this finding, we demonstrated how genomic dating plays an important role in expanding the timelines of past pandemics to make space for more diverse narratives.

In contrast to our claims of the 'power' of genomic evidence, a prominent takeway from Chapter 3 was our discussion of the limitations of DNA. In particular, we found that *Y. pestis* genomes in isolation are not suitable for reconstructing evolutionary relationships during short-term epidemics. This is because the evolutionary rate of past pandemic lineages is approximately 1 substitution every 10 years. Isolates collected within this time frame (<10 years) are often identical, which means that the directionality of spread cannot be confidently inferred. To mitigate this weakness, complementary evidence is needed that has a higher temporal resolution. Historical case records are an excellent candidate, where plague cases are recorded annually if not weekly (Roosen & Curtis, 2018). Based on initial comments from readers of the preprint, this conclusion was particularly exciting as it provided guidance on how to avoid over-interpreting ancient DNA evidence, and suggested a new avenue for inter-disciplinary collaboration (Boris Schmidt, *personal communication*).

In Chapter 4, I applied this updated genomic dataset and molecular clock method to a new problem. While in Chapter 3 we were concerned with estimating the first *emergence* of pandemic lineages, in Chapter 4 we reconstructed the *persistence* or *continuity* of ancient pandemics. We designed a unique longitudinal study, where we sampled skeletal remains spanning 800 years (1000 - 1800 CE) dated to before, during, and after the Second Pandemic (14$^{\text{th}}$th - 18$^{\text{th}}$ century). Our sampling strategy focused on Scandinavia, particularly Denmark, as this region is underrepresented in the historical narrative and because the Anthropological DataBase Odense University collection (ADBOU, University of

Southern Denmark) has exquisitely curated over 17,000 skeletal remains dated from the Viking Age (10th century) to the Early Modern Period (18th century). Using ancient DNA techniques, we recovered evidence of *Y. pestis* throughout the 14th to 17th centuries, which perfectly aligns with the historical narrative, limited as it is. Furthermore, our positivity rate for *Y. pestis* (3.3% - 14.3%) overlaps with mortality estimates from several historical outbreaks during the Second Pandemic. Our results strengthen the argument that *Y. pestis* was the causative agent of the Second Pandemic, and suggests that plague was a relatively new disease in medieval Denmark. These findings are being expanded on in two upcoming studies. The first, is an examination of how Danish populations responded to this new disease with regards to changes in the human immune system (Klunk et al., In Prep, 2021). The second, is a reconstruction of how and when virulence in *Y. pestis* became attenuated during the Second Pandemic. Taken together, we anticipate these studies will deepen our understanding of disease exposure and experience in Denmark and across Europe.

5.2 Future Directions

5.2.1 Same 'Plague', New Problems

A reoccurring problem in plague research is how best to integrate multidisciplinary sources, as there is great interest in combining genetic, historical, and environmental records to better understand past pandemics of plague (Dean et al., 2018; Schmid et al., 2015). An approach that is commonly used in ancient DNA studies of *Y. pestis* involves two steps: (1) reconstructing the relationships between epidemics using genetic evidence, and then (2) interpreting those relationships using historical records (Guellil et al., 2020; Namouchi et al., 2018; Spyrou et al., 2019). However, a major limitation of this method is that multidisciplinary sources are *only* integrated during the final interpretation phase. This runs the risk that errors and uncertainty associated with the genetic analysis will propagate, leading to high levels of ambiguity when attempting to provide historical context for this genetic 'noise'.

An alternative method, is to leverage the strengths and mitigate the weaknesses of interdisciplinary sources in a joint phylogenetic analysis. This novel approach treats historical records (ex. location and date of an outbreak) as special 'sequence-free' samples. These records are then combined with DNA evidence to jointly infer a phylogeny, which can then be used to estimate the timing and location of historical events. Recent studies have demonstrated how critical this approach is, as case occurrence records can effectively correct for sampling biases in sparse genomic datasets (Featherstone et al., 2021; Kalkauskas et al., 2021). However, incorporating sequence-free datasets is still a relatively recent method, and to date has only been applied to the study of viruses. Furthermore, it has only been tested on outbreaks occurring over a relatively small geographic area and time range. It remains unknown whether this approach is feasible for bacterial genomics, let alone ancient DNA, where genomes are larger and sampled across greater temporal and geographic scopes. This presents a key line of inquiry for future research, for which the plague bacterium *Y. pestis* would be an excellent case study.

5.2.2 New 'Plague', Same Problems

During the course of this dissertation, my interest in global pandemics turned from an academic curiosity to a lived experience. In 2019, the novel coronavirus SARS-CoV-2 emerged to cause a global pandemic, with over 370 million cases recorded worldwide (2022 January 31). While there are many unique aspects of this pandemic, one that has captured my attention is that it is the first pandemic to be monitored with real-time genomic surveillance (Oude Munnink et al., 2021). Over two million genomic sequences have been deposited in public repositories, which can be used to inform public health responses (Public Health Ontario, 2021). However, this avalanche of data has also caused numerous problems, as researchers are struggling to manage this information and utilize it effectively (Morel et al., 2021). As a result, database tools such as `NCBImeta` presented in Chapter 2, are playing an important role in information management.

One field of ongoing research involves improving the scalability of these tools. For example, `NCBImeta` was developed for a data set of 'only' 15,000 records, and in its current implementation, cannot process the 1+ million SARS-CoV-2 records on NCBI. A second critical avenue is integrating information from multiple repositories, as surveillance data is inconsistently being deposited in national and international databases (CanCOGeN, n.d.; GISAID, n.d.; NCBI, n.d.). Progress towards these two objectives will result in more diverse genomic data being analyzed (geographically and temporally), which may improve of our understanding of transmission and spread between and within countries.

Another parallel between this dissertation and the ongoing pandemic involves spatiotemporal modeling. In Chapter 3, we discovered that in our expanded genomic data set, *Y. pestis*' rate of spread tends to outpace its rate of evolutionary change. This leads to identical *Y. pestis* isolates found across multiple countries, such as the case throughout the Black Death (1346-1353). However, we sporadically observed the opposite trend, in which *Y. pestis* strains collected in a short time frame (<10 years) were *extremely* different. This tremendous diversity in evolutionary rates meant that we were unable to estimate a single molecular clock for *Y. pestis.* These issues, clonal spread and rate variation, were also recently documented in SARS-CoV-2 (R.-C. Ferreira et al., 2021). Ferreira et al. describe this as a paradox in which we *"become increasingly uncertain about the relationships among specific lineages as we collect greater amounts of data"*. This runs counterintuitive to the general expectation in scientific studies that *the more data we collect, the closer we get to the 'truth'.* Overall, this presents a complex theoretical problem that is becoming increasingly prevalent across various disciplines moving into the era of 'big data'.

References

Agostinetto, G., Bozzi, D., Porro, D., Casiraghi, M., Labra, M., & Bruno, A. (2021). *SKIOME Project: A curated collection of skin microbiome datasets enriched with study-related metadata*. 2021.08.17.456635. https://doi.org/10.1101/2021.08.17.456635

Andrades Valtueña, A., Mittnik, A., Key, F. M., Haak, W., Allmäe, R., Belinskij, A., Daubaras, M., Feldman, M., Jankauskas, R., Janković, I., Massy, K., Novak, M., Pfrengle, S., Reinhold, S., Šlaus, M., Spyrou, M. A., Szécsényi-Nagy, A., Tõrv, M., Hansen, S., … Krause, J. (2017). The Stone Age Plague and its persistence in Eurasia. *Current Biology*, *27*(23), 3683–3691.e8. https://doi.org/10.1016/j.cub.2017.10.025

Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, *30*(2), 239–243. https://doi.org/10.1093/molbev/mss243

Benedict, C. (1988). Bubonic plague in nineteenth-century China. *Modern China*, *14*(2), 107–155. https://www.jstor.org/stable/189268

Benedictow, Ole Jørgen. (2016). The Black Death and Later Plague Epidemics in the Scandinavian Countries: Perspectives and Controversies. In *The Black Death and Later Plague Epidemics in the Scandinavian Countries:* De Gruyter Open Poland. https://www.degruyter.com/document/doi/10.1515/9788376560472/html

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

Benedictow, O. J. (2021). *The Complete History of the Black Death*. Boydell Press. https://boydellandbrewer.com/9781783275168/the-complete-history-of-the-black-death/

Bernstein, M. N., Doan, A., & Dewey, C. N. (2017). MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, *33*(18), 2914–2923. https://doi.org/10.1093/bioinformatics/btx334

Bolaños, I. A. (2019). The Ottomans during the global crises of cholera and plague: The View from Iraq and the Gulf. *International Journal of Middle East Studies*, *51*(4), 603–620. https://doi.org/10.1017/S0020743819000667

Boldsen, J. L. (2009). Leprosy in Medieval Denmark — Osteological and epidemiological analyses. *Anthropologischer Anzeiger*, *67*(4), 407–425. https://www.jstor.org/stable/29543069

Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., & Krause, J. (2011). A draft genome of *Yersinia Pestis* from victims of the Black Death. *Nature*, *478*(7370), 506–510. https://doi.org/10.1038/nature10549

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. du, Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bradley, R. (1721). *The Plague at Marseilles: Consider'd with Remarks Upon the Plague in General*. W. Mears. https://books.google.ca/books?id=qQYAmMH1nS4C

Bramanti, B., Wu, Y., Yang, R., Cui, Y., & Stenseth, N. C. (2021). Assessing the origins of the European Plagues following the Black Death: A synthesis of genomic, historical, and ecological information. *Proceedings of the National Academy of Sciences*, *118*(36). https://doi.org/10.1073/pnas.2101940118

Brown, P. J., & Inhorn, M. C. (2013). *The Anthropology of Infectious Disease: International Health Perspectives*. Routledge. https://books.google.com?id=WUj5AQAAQBAJ

Brüssow, H. (2021). What we can learn from the dynamics of the 1889 "Russian flu" pandemic for the future trajectory of COVID-19. *Microbial Biotechnology*, *14*(6). https://doi.org/10.1111/1751-7915.13916

CanCOGeN. (n.d.). *VirusSeq Portal*. Retrieved December 18, 2021, from https://virusseq-dataportal.ca/

Cantlie, J. (1897). The spread of plague. *Transactions. Epidemiological Society of London*, *16*, 15–63. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC554

0063/

Carmichael, A. G. (2015). Plague persistence in Western Europe: A hypothesis. *The Medieval Globe*, *1*(1), 157–191. https://muse.jhu.edu/article/758488

Chang, W. E., Peterson, M. W., Garay, C. D., & Korves, T. (2016). Pathogen metadata platform: Software for accessing and analyzing pathogen strain information. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-1231-2

Choudhary, S. (2019). Pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. *F1000Research*, *8*, 532. https://doi.org/10.12688/f1000research.18676.1

Christensen, P. (2003). “In these perilous times”: Plague and plague policies in early modern Denmark. *Medical History*, *47*(4), 413–450. https://doi.org/10.1017/S0025727300057331

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Connor, T. R., Barker, C. R., Baker, K. S., Weill, F.-X., Talukder, K. A., Smith, A. M., Baker, S., Gouali, M., Pham Thanh, D., Jahan Azmi, I., Dias da Silveira, W., Semmler, T., Wieler, L. H., Jenkins, C., Cravioto, A., Faruque, S. M., Parkhill, J., Wook Kim, D., Keddy, K. H., & Thomson, N. R. (2015). Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella Flexneri*. *eLife*, *4*. https://doi.org/10.7554/eLife.07335

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: Do it right or not at all. *Science (New York, N.Y.)*, *289*(5482), 1139. https://doi.org/10.1126/science.289.5482.1139b

Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L. A., Wang, Z., Guo, Z., Xu, L., Zhang, Y., Zheng, H., Qin, N., Xiao, X., Wu, M., Wang, X., Zhou, D., Qi, Z., Du, Z., … Yang, R. (2013). Historical variations in mutation rate in an epidemic pathogen, *Yersinia Pestis*. *Proceedings of the National Academy of Sciences*, *110*(2), 577–582. https://doi.org/10.1073/pnas.1205750110

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National*

Academy of Sciences, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Dean, K. R., Krauer, F., Walløe, L., Lingjærde, O. C., Bramanti, B., Stenseth, N. Chr., & Schmid, B. V. (2018). Human ectoparasites and the spread of plague in Europe during the Second Pandemic. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(6), 1304–1309. https://doi.org/10.1073/pnas.1715640115

Devignat, R. (1951). Variétés de l'espèce Pasteurella pestis. *Bulletin of the World Health Organization*, *4*(2), 247–263. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2554099/

DeWitte, S. N., & Kowaleski, M. (2017). Black Death Bodies. *Fragments: Interdisciplinary Approaches to the Study of Ancient and Medieval Pasts*, *6*. http://hdl.handle.net/2027/spo.9772151.0006.001

Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O., & Raoult, D. (1998). Detection of 400-year-old Yersinia pestis DNA in human dental pulp: An approach to the diagnosis of ancient septicemia. *Proceedings of the National Academy of Sciences*, *95*(21), 12637–12640. https://doi.org/10.1073/pnas.95.21.12637

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLOS Biology*, *4*(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution*, *6*(2). https://doi.org/10.1093/ve/veaa061

Duchene, S., Lemey, P., Stadler, T., Ho, S. Y. W., Duchene, D. A., Dhanasekaran, V., & Baele, G. (2020). Bayesian evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution*, *37*(11), 3363–3379. https://doi.org/10.1093/molbev/msaa163

Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., Fourment, M., & Holmes, E. C. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, *2*(11). https://doi.org/10.1099/mgen.0.000094

Eaton, K. (2020). NCBImeta: Efficient and comprehensive metadata retrieval from NCBI databases. *Journal of Open Source Software*, *5*(46), 1990. https:

//doi.org/10.21105/joss.01990

Eaton, K., Featherstone, L., Duchene, S., Carmichael, A. G., Varlık, N., Holmes, E. C., Golding, G. B., & Poinar, H. N. (Submitted, 2021). Plagued by a cryptic clock: Insight and issues from the global phylogeny of Yersinia pestis. *Nature Communications*. https://www.researchsquare.com/article/rs-1146895

Echenberg, M. (2002). Pestis redux: The initial years of the third bubonic plague pandemic, 1894-1901. *Journal of World History*, *13*(2), 429–449. https://www.jstor.org/stable/20078978

Entrez Help. (2016). National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK3837/ (Original work published 2006)

Eppinger, M., Pearson, T., Koenig, S. S. K., Pearson, O., Hicks, N., Agrawal, S., Sanjar, F., Galens, K., Daugherty, S., Crabtree, J., Hendriksen, R. S., Price, L. B., Upadhyay, B. P., Shakya, G., Fraser, C. M., Ravel, J., & Keim, P. S. (2014). Genomic epidemiology of the Haitian cholera outbreak: A single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *mBio*, *5*(6). https://doi.org/10.1128/mBio.01721-14

Eroshenko, G. A., Popov, N. V., Al'khova, Z. V., Kukleva, L. M., Balykova, A. N., Chervyakova, N. S., Naryshkina, E. A., & Kutyrev, V. V. (2021). Evolution and circulation of *Yersinia Pestis* in the Northern Caspian and Northern Aral Sea regions in the 20th-21st centuries. *PLOS ONE*, *16*(2), e0244615. https://doi.org/10.1371/journal.pone.0244615

Fancy, N., & Green, M. (2021). Plague and the Fall of Baghdad (1258). *Medical History*, *65*(2), 155–177. https://scholarship.depauw.edu/hist_facpubs/14

Featherstone, L. A., Di Giallonardo, F., Holmes, E. C., Vaughan, T. G., & Duchêne, S. (2021). Infectious disease phylodynamics with occurrence data. *Methods in Ecology and Evolution*, *12*(8), 1498–1507. https://doi.org/10.1111/2041-210X.13620

Ferreira, M. A. R., & Suchard, M. A. (2008). Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics*, *36*(3), 355–368. https://doi.org/10.1002/cjs.5550360302

Ferreira, R.-C., Wong, E., Gugan, G., Wade, K., Liu, M., Baena, L. M., Chato, C., Lu, B., Olabode, A. S., & Poon, A. F. Y. (2021). CoVizu: Rapid analysis and visualization of the global diversity of SARS-CoV-2 genomes. *Virus Evolution*, *7*(2), veab092. https://doi.org/10.1093/ve/veab092

Gage, K. L., & Kosoy, M. Y. (2005). Natural history of plague: Perspectives from more than a century of research. *Annual Review of Entomology*, *50*, 505–528. https://doi.org/10.1146/annurev.ento.50.071803.130337

Gilbert, M. T. P., Cuccui, J., White, W., Lynnerup, N., Titball, R. W., Cooper, A., & Prentice, M. B. (2004a). Response to Drancourt and Raoult. *Microbiology*, *150*(2), 264–265. https://doi.org/10.1099/mic.0.26959-0

Gilbert, M. T. P., Cuccui, J., White, W., Lynnerup, N., Titball, R. W., Cooper, A., & Prentice, M. B. Y. 2004. (2004b). Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology*, *150*(2), 341–354. https://doi.org/10.1099/mic.0.26594-0

GISAID. (n.d.). *GISAID - Initiative*. Retrieved December 18, 2021, from https://www.gisaid.org/

Green, M. H. (2020a). How a microbe becomes a pandemic: A new story of the Black Death. *The Lancet Microbe*, *1*(8), e311–e312. https://doi.org/10.1016/S2666-5247(20)30176-2

Green, M. H. (2020b). The four Black Deaths. *The American Historical Review*, *125*(5), 1601–1631. https://doi.org/10.1093/ahr/rhaa511

Green, M. H. (2018). Putting Africa on the Black Death map: Narratives from genetics and history. *Afriques*, *9*(09). https://doi.org/10.4000/afriques.2125

Guellil, M., Kersten, O., Namouchi, A., Luciani, S., Marota, I., Arcini, C. A., Iregren, E., Lindemann, R. A., Warfvinge, G., Bakanidze, L., Bitadze, L., Rubini, M., Zaio, P., Zaio, M., Neri, D., Stenseth, N. C., & Bramanti, B. (2020). A genomic and historical synthesis of plague in 18th century Eurasia. *Proceedings of the National Academy of Sciences*, *117*(45), 28328–28335. https://doi.org/10.1073/pnas.2009677117

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Hashemi Shahraki, A., Carniel, E., & Mostafavi, E. (2016). Plague in Iran: Its history and current status. *Epidemiology and Health*, *38*. https://doi.org/10.4178/epih.e2016033

Hider, J., Duggan, A. T., Klunk, J., Eaton, K., Long, G. S., Karpinski, E., Golding, G. B., Prowse, T. L., Poinar, H. N., & Fornaciari, G. (In Prep). *Examining pathogen DNA recovery across the remains of a 14th century Italian monk (St. Brancorsini) infected with Brucella melitensis.*

Ho, S. Y. W., & Duchêne, S. (2020). Dating the emergence of human pathogens.

Science, *368*(6497), 1310–1311. https://doi.org/10.1126/science.abc5746

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Kalkauskas, A., Perron, U., Sun, Y., Goldman, N., Baele, G., Guindon, S., & Maio, N. D. (2021). Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLOS Computational Biology*, *17*(1), e1008561. https://doi.org/10.1371/journal.pcbi.1008561

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6, 6), 587–589. https://doi.org/10.1038/nmeth.4285

Kans, J. (2019). Entrez Direct: E-utilities on the UNIX Command Line. In *Entrez Programming Utilities Help*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK179288/ (Original work published 2013)

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), e3–e3. https://doi.org/10.1093/nar/gkr771

Klunk, J., Vilgalys, T., Demeure, C., Cobb, M., Elli, D., Redfern, R., DeWitte, S. N., Gamble, J., Boldsen, J. L., Carmichael, A. G., Varlık, N., Eaton, K., Grenier, J.-C., Golding, G. B., Devault, A., Rouillard, J.-M., Dumaine, A., Missiakas, G. R., Pizarro-Cerdá, J., . . . Barreiro, L. (In Prep, 2021). *Black Death shaped the evolution of immune genes.*

Kutyrev, V. V., Eroshenko, G. A., Motin, V. L., Nosov, N. Y., Krasnov, J. M., Kukleva, L. M., Nikiforov, K. A., Al'khova, Z. V., Oglodin, E. G., & Guseva, N. P. (2018). Phylogeny and classification of *Yersinia Pestis* through the lens of strains from the plague foci of Commonwealth of Independent States. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01106

Lam, A., & Duchene, S. (2021). The impacts of low diversity sequence data on phylodynamic inference during an emerging epidemic. *Viruses*, *13*(1, 1), 79. https://doi.org/10.3390/v13010079

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLOS Computational Biology*, *5*(9), e1000520. https://doi.org/10.1371/journal.pcbi.1000520

Lenz, K., & Hybel, N. (2016). The Black Death. *Scandinavian Journal of*

History, *41*(1), 54–70. https://doi.org/10.1080/03468755.2015.1110533

Li, Y., Cui, Y., Hauck, Y., Platonov, M. E., Dai, E., Song, Y., Guo, Z., Pourcel, C., Dentovskaya, S. V., Anisimov, A. P., Yang, R., & Vergnaud, G. (2009). Genotyping and phylogenetic analysis of *Yersinia Pestis* by MLVA: Insights into the worldwide expansion of Central Asia plague foci. *PLOS ONE*, *4*(6), e6000. https://doi.org/10.1371/journal.pone.0006000

Little, L. K. (2007). *Plague and the End of Antiquity: The Pandemic of 541-750*. Cambridge University Press. https://doi.org/10.1017/CBO9780511812934

Mackenzie, A., McNally, R., Mills, R., & Sharples, S. (2016). Post-archival genomics and the bulk logistics of DNA sequences. *BioSocieties*, *11*(1), 82–105. https://doi.org/10.1057/biosoc.2015.22

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., Serdari, D., Kostaki, E.-G., Mamais, I., Kozlov, A. M., Pavlidis, P., Paraskevis, D., & Stamatakis, A. (2021). Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution*, *38*(5), 1777–1791. https://doi.org/10.1093/molbev/msaa314

Morelli, G., Song, Y., Mazzoni, C. J., Eppinger, M., Roumagnac, P., Wagner, D. M., Feldkamp, M., Kusecek, B., Vogler, A. J., Li, Y., Cui, Y., Thomson, N. R., Jombart, T., Leblois, R., Lichtner, P., Rahalison, L., Petersen, J. M., Balloux, F., Keim, P., . . . Achtman, M. (2010). *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics*, *42*(12), 1140–1143. https://doi.org/10.1038/ng.705

Munyenyiwa, A., Zimba, M., Nhiwatiwa, T., & Barson, M. (2019). Plague in Zimbabwe from 1974 to 2018: A review article. *PLOS Neglected Tropical Diseases*, *13*(11), e0007761. https://doi.org/10.1371/journal.pntd.0007761

Murchie, T., Karpinski, E., Eaton, K., Duggan, A. T., Baleka, S., Zazula, G., MacPhee, R. D. E., Froese, D., & Poinar, H. N. (Accepted, 2021). Pleistocene mitogenomes reconstructed from the environmental DNA of permafrost. *Current Biology*.

Nakazato, T., Ohta, T., & Bono, H. (2013). Experimental design-based functional mining and characterization of high-throughput sequencing data in the Sequence Read Archive. *PLoS ONE*, *8*(10), e77910. https://doi.org/10.1371/journal.pone.0077910

Namouchi, A., Guellil, M., Kersten, O., Hänsch, S., Ottoni, C., Schmid, B. V., Pacciani, E., Quaglia, L., Vermunt, M., Bauer, E. L., Derrick, M., Jensen, A. Ø., Kacki, S., Cohn, S. K., Stenseth, N. C., & Bramanti, B. (2018). Integrative approach using *Yersinia Pestis* genomes to revisit the historical landscape of plague during the Medieval Period. *Proceedings of the National Academy of Sciences*, *115*(50), E11790–E11797. https://doi.org/10.1073/pnas.1812865115

NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved December 18, 2021, from https://www.ncbi.nlm.nih.gov/

Nguyen, V. K., Parra-Rojas, C., & Hernandez-Vargas, E. A. (2018). The 2017 plague outbreak in Madagascar: Data descriptions and epidemic modelling. *Epidemics*, *25*, 20–25. https://doi.org/10.1016/j.epidem.2018.05.001

Nutton, V. (1983). The seeds of disease: An explanation of contagion and infection from the Greeks to the Renaissance. *Medical History*, *27*(1), 1–34. https://doi.org/10.1017/S0025727300042241

Nyirenda, S. S., Hang'ombe, B. M., Simulundu, E., Mulenga, E., Moonga, L., Machang'u, R. S., Misinzo, G., & Kilonzo, B. S. (2018). Molecular epidemiological investigations of plague in Eastern Province of Zambia. *BMC Microbiology*, *18*(1), 2. https://doi.org/10.1186/s12866-017-1146-8

Ober, W. B., & Aloush, N. (1982). The plague at Granada, 1348-1349: Ibn Al-Khatib and ideas of contagion. *Bulletin of the New York Academy of Medicine*, *58*(4), 418–424. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1808550/

Ortner, D. J. (2007). Differential Diagnosis of Skeletal Lesions in Infectious Disease. In *Advances in Human Palaeopathology* (pp. 189–214). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470724187.ch10

Oude Munnink, B. B., Worp, N., Nieuwenhuijse, D. F., Sikkema, R. S., Haagmans, B., Fouchier, R. A. M., & Koopmans, M. (2021). The next phase of SARS-CoV-2 surveillance: Real-time molecular epidemiology. *Nature Medicine*, *27*(9, 9), 1518–1524. https://doi.org/10.1038/s41591-021-01472-w

Perry, R. D., & Fetherston, J. D. (1997). *Yersinia pestis* - etiologic agent of plague. *Clinical Microbiology Reviews*, *10*(1), 35–66.

Piret, J., & Boivin, G. (2021). Pandemics throughout history. *Frontiers in Microbiology*, *11*, 631736. https://doi.org/10.3389/fmicb.2020.631736

Pisarenko, S. V., Evchenko, A. Yu., Kovalev, D. A., Evchenko, Y. M., Bobrysheva, O. V., Shapakov, N. A., Volynkina, A. S., & Kulichenko, A. N. (2021). *Yersinia pestis* strains isolated in natural plague foci of Caucasus and Transcaucasia in the context of the global evolution of species. *Genomics*, *113*(4), 1952–1961. https://doi.org/10.1016/j.ygeno.2021.04.021

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

Public Health Ontario. (2021). *SARS-CoV-2 Whole Genome Sequencing in Ontario, December 14, 2021* (p. 27) [Weekly Epidemiological Summary]. https://www.publichealthontario.ca/-/media/documents/ncov/epi/covid-19-sars-cov2-whole-genome-sequencing-epi-summary.pdf

Raoult, D. (2003). Was the Black Death yersinial plague? *The Lancet Infectious Diseases*, *3*(6), 328. https://doi.org/10.1016/S1473-3099(03)00652-2

Rascovan, N., Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C., & Rasmussen, S. (2019). Emergence and spread of basal lineages of *Yersinia Pestis* during the Neolithic Decline. *Cell*, *176*(1), 295–305.e10. https://doi.org/10.1016/j.cell.2018.11.005

Rasmussen, S., Allentoft, M. E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A. G., Schubert, M., Van Dam, A., Kapel, C. M. O., Nielsen, H. B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M. V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., ... Willerslev, E. (2015). Early divergent strains of *Yersinia Pestis* in Eurasia 5,000 years ago. *Cell*, *163*(3), 571–582. https://doi.org/10.1016/j.cell.2015.10.009

Roosen, J., & Curtis, D. R. (2018). Dangers of noncritical use of historical plague data. *Emerging Infectious Diseases*, *24*(1), 103–110. https://doi.org/10.3201/eid2401.170477

Ryan, E. T. (2011). The cholera pandemic, still with us after half a century: Time to rethink. *PLOS Neglected Tropical Diseases*, *5*(1), e1003. https://doi.org/10.1371/journal.pntd.0001003

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vex042

Sampath, S., Khedr, A., Qamar, S., Tekin, A., Singh, R., Green, R., & Kashyap, R. (2021). Pandemics Throughout the History. *Cureus*, *13*(9), e18136. https://doi.org/10.7759/cureus.18136

Santer, M. (2009). Richard Bradley: A Unified, Living Agent Theory of the Cause of Infectious Diseases of Plants, Animals, and Humans in the First Decades of the 18th Century. *Perspectives in Biology and Medicine*, *52*(4), 566–578. https://doi.org/10.1353/pbm.0.0124

Schmid, B. V., Büntgen, U., Easterday, W. R., Ginzler, C., Walløe, L., Bramanti, B., & Stenseth, N. C. (2015). Climate-driven introduction of the Black Death and successive plague reintroductions into Europe. *Proceedings of the National Academy of Sciences*, *112*(10), 3020–3025. https://doi.org/10.1073/pnas.1412887112

Scott, S., & Duncan, C. J. (2001). *Biology of Plagues: Evidence from Historical Populations*. Cambridge University Press. https://doi.org/10.1017/CBO9780511542527

Shadwell, A. (1899). The plague in Oporto. *The Nineteenth Century: A Monthly Review*, *46*(273), 833–847. https://www.proquest.com/openview/5dd86300bfa45d65/1?pq-origsite=gscholar&cbl=1017

Slavin, P. (2021). Out of the West: Formation of a permanent plague reservoir in south-central Germany (1349–1356) and its implications. *Past & Present*, *252*(1), 3–51. https://doi.org/10.1093/pastj/gtaa028

Spyrou, M. A., Keller, M., Tukhbatova, R. I., Scheib, C. L., Nelson, E. A., Andrades Valtueña, A., Neumann, G. U., Walker, D., Alterauge, A., Carty, N., Cessford, C., Fetz, H., Gourvennec, M., Hartle, R., Henderson, M., von Heyking, K., Inskip, S. A., Kacki, S., Key, F. M., ... Krause, J. (2019). Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia Pestis* genomes. *Nature Communications*, *10*(1, 1), 4470. https://doi.org/10.1038/s41467-019-12154-0

Spyrou, M. A., Tukhbatova, R. I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A. G., Castex, D., Wahl, J., Gazimzyanov, I. R., Nurgaliev, D. K., Herbig, A., Bos, K. I., & Krause, J. (2016). Historical *Y. Pestis* genomes reveal the European Black Death as the source of ancient and modern plague pandemics. *Cell Host & Microbe*, *19*(6), 874–881. https://doi.org/10.1016/j.chom.2016.05.012

Spyrou, M. A., Tukhbatova, R. I., Wang, C.-C., Valtueña, A. A., Lankapalli, A. K., Kondrashin, V. V., Tsybin, V. A., Khokhlov, A., Kühnert, D., Herbig, A., Bos, K. I., & Krause, J. (2018). Analysis of 3800-year-old *Yersinia Pestis* genomes suggests Bronze Age origin for bubonic plague. *Nature Communications*, *9*(1, 1), 2234. https://doi.org/10.1038/s41467-018-04550-9

Stewart, L., Ford, A., Sangal, V., Jeukens, J., Boyle, B., Kukavica-Ibrulj, I., Caim, S., Crossman, L., Hoskisson, P. A., Levesque, R., & Tucker, N. P. (2014). Draft genomes of 12 host-adapted and environmental isolates of *Pseudomonas Aeruginosa* and their positions in the core genome phylogeny. *Pathogens and Disease*, *71*(1), 20–25. https://doi.org/10.1111/2049-632X.12107

Strafella, S., Simpson, D. J., Yaghoubi Khanghahi, M., De Angelis, M., Gänzle, M., Minervini, F., & Crecchio, C. (2021). Comparative Genomics and In Vitro Plant Growth Promotion and Biocontrol Traits of Lactic Acid

Bacteria from the Wheat Rhizosphere. *Microorganisms*, *9*(1, 1), 78. https://doi.org/10.3390/microorganisms9010078

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Syed, I. (1981). Islamic medicine: 1000 years ahead of its times. *Journal of the International Society for the History of Islamic Medicine*, *13*(1), 2–9. https://jima.imana.org/article/view/11925

Tan, J., Liu, Y., Shen, E., Zhu, W., Wang, W., Li, R., & Yang, L. (2002). Towards the atlas of plague and its environment in the People's Republic of China: idea, principle and methodology of design and research results. *Huan Jing Ke Xue*, *23*(3), 1–8.

Varlık, N. (2020). The plague that never left: Restoring the Second Pandemic to Ottoman and Turkish history in the time of COVID-19. *New Perspectives on Turkey*, *63*, 176–189. https://doi.org/10.1017/npt.2020.27

Varlık, N. (2014). New science and old sources: Why the Ottoman experience of plague matters. *The Medieval Globe*, *1*, 193–227. https://scholarworks.wmich.edu/tmg/vol1/iss1/9/

Wagner, D. M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J. W., Enk, J., Birdsell, D. N., Kuch, M., Lumibao, C., Poinar, D., Pearson, T., Fourment, M., Golding, B., Riehm, J. M., Earn, D. J. D., DeWitte, S., Rouillard, J.-M., Grupe, G., … Poinar, H. (2014). *Yersinia pestis* and the Plague of Justinian 541–543 AD: A genomic analysis. *The Lancet Infectious Diseases*, *14*(4), 319–326. https://doi.org/10.1016/S1473-3099(13)70323-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. http://ggplot2.org

Wood, J. W., Milner, G. R., Harpending, H. C., Weiss, K. M., Cohen, M. N., Eisenberg, L. E., Hutchinson, D. L., Jankauskas, R., Cesnys, G., Česnys, G., Katzenberg, M. A., Lukacs, J. R., McGrath, J. W., Roth, E. A., Ubelaker, D. H., & Wilkinson, R. G. (1992). The Osteological Paradox: Problems of Inferring Prehistoric Health from Skeletal Samples. *Current Anthropology*, *33*(4), 343–370. https://www.jstor.org/stable/2743861

Xu, L., Stige, L. Chr., Kausrud, K. L., Ben Ari, T., Wang, S., Fang, X., Schmid, B. V., Liu, Q., Stenseth, N. Chr., & Zhang, Z. (2014). Wet climate and transportation routes accelerate spread of human plague. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1780), 20133159. https://doi.org/10.1098/rspb.2013.3159

Xu, L., Stige, L. C., Leirs, H., Neerinckx, S., Gage, K. L., Yang, R., Liu, Q., Bramanti, B., Dean, K. R., Tang, H., Sun, Z., Stenseth, N. C., & Zhang, Z. (2019). Historical and genomic data reveal the influencing factors on global transmission velocity of plague during the Third Pandemic. *Proceedings of the National Academy of Sciences*, *116*(24), 11833–11838. https://doi.org/10.1073/pnas.1901366116

Yates, J. A. F., Lamnidis, T. C., Borry, M., Valtueña, A. A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., & Peltzer, A. (2021). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, *9*, e10947. https://doi.org/10.7717/peerj.10947

Yue, R. P. H., Lee, H. F., & Wu, C. Y. H. (2017). Trade routes and plague transmission in pre-industrial Europe. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-13481-2

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–2644. https://doi.org/10.1093/jac/dks261

Zeppelini, C. G., DE Almeida, A. M. P., & Cordeiro-Estrela, P. (2018). Ongoing quiescence in the Borborema Plateau Plague focus (Paraiba, Brazil). *Anais Da Academia Brasileira De Ciencias*, *90*(3), 3007–3015. https://doi.org/10.1590/0001-3765201820170977

Zhou, D., Han, Y., Song, Y., Huang, P., & Yang, R. (2004). Comparative and evolutionary genomics of *Yersinia Pestis*. *Microbes and Infection*, *6*(13), 1226–1234. https://doi.org/10.1016/j.micinf.2004.08.002

Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., & Achtman, M. (2020). The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Research*, *30*(1), 138–152. https://doi.org/10.1101/gr.251678.119

Zhu, Y., Stephens, R. M., Meltzer, P. S., & Davis, S. R. (2013). SRAdb: Query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, *14*(1), 19. https://doi.org/10.1186/1471-2105-14-19