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

© Copyright by Katherine Eaton,
All Rights Reserved

Doctor of Philosophy (2021)
(Department of Anthropology)

McMaster University
Hamilton, Ontario, Canada

| | |
|---|---|
| TITLE: | Big Data, Small Microbes: Genomic analysis of the plague bacterium *Yersinia pestis* |
| AUTHOR: | Katherine Eaton<br>B.A. (Hons) Anthropology, University of Alberta |
| SUPERVISOR: | Dr. Hendrik Poinar |
| NUMBER OF PAGES: | xi, 26 |

# 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 ($14^{th}$ - 19th$^{th}$ 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' (ie. 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 (ie. 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.

# 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 14$^{\text{th}}$ 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 (Benedictow, 2004). Outbreaks of this new and mysterious disease, often referred to as *The 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 contagion at the time was *miasma*, in which diseases were spread through noxious air (Ober & Aloush, 1982). However, Ibn al-Khatib, a notable Islamic scholar, took issue with this theory. After studying outbreaks of *The Plague* in the 14$^{\text{th}}$ 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 19$^{\text{th}}$ 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, who all happened to be grappling with the perplexing nature of *The 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 (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 retrieve the *precise pathogen* 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 entire *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 $9^{th}$ to the $19^{th}$ 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 $20^{th}$ and $21^{st}$ century plague outbreaks (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.

# 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
Licensed under a Creative Commons Attribution 4.0 International License.

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 1.

Table 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 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.

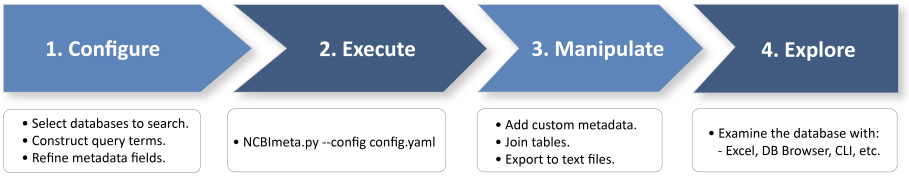


Figure 1: NCBImeta user workflow.

`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, 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: 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 3, 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 GEO databases).

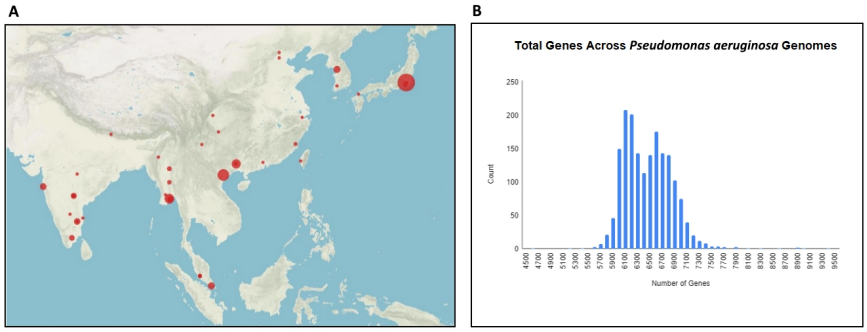


Figure 3: Metadata visualization of *P. aeruginosa* sequencing projects. A) The geographic distribution of samples in this region highlights a large number originating in Japan. Visualized with Palladio (https://hdlab.stanford.edu/palladio/). B) The number of genes per organism reveals a multi-modal distribution within the species.

## 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).

## 2.8 References

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

Submitted 06 December 2021 to
*Nature Communications*.
Transferred 17 December 2021 to
*Communications Biology*.
https://www.researchsquare.com/article/rs-1146895
Licensed under a Creative Commons Attribution 4.0 International License

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.

# 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*
Licensed under a Creative Commons Attribution 4.0 International License

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.

# 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 presented 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 interested 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 (ie. 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 $14^{th}$ 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 (ie. 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 (ie. 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 ($10^{\text{th}}$ century) to the Early Modern Period ($18^{\text{th}}$ century). Using ancient DNA techniques, we recovered evidence of *Y. pestis* throughout the $14^{\text{th}}$ to $17^{\text{th}}$ 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 270 million cases recorded worldwide. 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 (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

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

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

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

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

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

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

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

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

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

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

*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

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, 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

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/

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

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.*

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

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)

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.*

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

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

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/

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

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

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

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

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

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

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

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

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

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

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

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