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 (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 B.A. (Hons) Anthropology, University of Alberta
SUPERVISOR:	Dr. Hendrik Poinar
NUMBER OF PAGES:	ix, 12

Lay Abstract

"A lay abstract of not more 150 words must be included explaining the key goals and contributions of the thesis in lay terms that is accessible to the general public."

Abstract

Abstract here (no more than 300 words).

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

Acknowledgments

Acknowledgments go here.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

aDNA: Ancient DNA
DNA: Deoxyribonucleic acid
NCBI: National Center for Biotechnology Information
SRA: Sequence Read Archive

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 *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 Black Death.

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 seem 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 fits 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 snide and personal insults (Raoult, 2003). It was clear that these researchers cared *deeply* about their work. Despite the 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 trying to diagnose 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 confirmed 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). But 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 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 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

Katherine Eaton[1,2]

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

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

Submitted 06 December 2021 to
Nature Communications.
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.

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*[1,2], Leo Featherstone[3], Sebastian Duchene[3], Ann G. Carmichael[4], Nükhet Varlık[5], G. Brian Golding[6], Hendrik N. Poinar[1,2,8,9,10]

*Contributed equally.

[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.
[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.

5 Conclusion

As a paper on software development, its contributions and significance to the field of anthropology are understandably unclear. I admittedly targeted this article exclusively towards computational biologists because, at the time, few anthropologists had expressed interested in the issue of collecting and curating sequence data from online repositories. However, since its publication, my software has been used to support several bodies of anthropological research.

The database software NCBImeta was recently used to support an environmental reconstruction of Beringia (Murchie et al., In Prep), 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).

This tool 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 2019, my relationship with infectious diseases transformed from an intellectual curiosity to a lived experience. The emergence of the novel coronavirus (SARS-CoV-2) triggered a global pandemic, operating on a scale that had not been seen for a 100 years. For years, I had written grants to fund my plague research u... *you have to know the past to understand the present* (Sagan, 1980).

References

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

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.

Brown, P. J., & Inhorn, M. C. (2013). *The Anthropology of Infectious Disease: International Health Perspectives*. Routledge.

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

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

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

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

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.*

Murchie, T., Karpinski, E., Eaton, K., Duggan, A. T., Baleka, S., Zazula, G., MacPhee, R. D. E., Froese, D., & Poinar, H. N. (In Prep). *No bones about it: Pleistocene mitogenomes reconstructed from the environmental DNA of permafrost.*

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.

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

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

Sagan, C. E. (1980). One Voice in the Cosmic Fugue. In *Cosmos: A Personal Voyage* (No. 2). Arlington, VA: Public Broadcasting Service.

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

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), 4470. https://doi.org/10.1038/s41467-019-12154-0

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.

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

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