Genomic Insights into Pre- and Post-Contact Human Pathogens in the New World

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Åshild Joanne Vågene
aus Stavanger, Norwegen

Tübingen
2018

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.12.2018
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Johannes Krause
2. Berichterstatter:	Prof. Nicholas J. Conard, PhD

Acknowledgements

I am grateful to have received valuable advice, support and inspiration from many people throughout my doctoral studies.

First and foremost I would like to extend my thanks and gratitude to Prof. Johannes Krause for being a dependable and encouraging supervisor throughout my doctoral studies. To Dr. Kirsten Bos for teaching me the ropes in the lab and for fruitful project mentorship over the years. To Dr. Alexander Herbig for providing invaluable guidance in bioinformatics analyses and problem solving.

To Prof. Nick Conard for creating an enjoyable working environment at the INA and for kindly agreeing to supervise my dissertation.

My sincere thanks to Prof. Anne Stone, Prof. Jane Buikstra, Dr. Tanvi Honap, Prof. Noreen Tuross and Dr. Michael Campana for stimulating discussions and inspiring collaboration.

To Prof. Christina Warinner for her fountain of knowledge and enthusiastic collaboration on the Teposcolula-Yucundaa project.

A special thanks to Antje Wissgott and Guido Brandt and all the technicians at the MPI-SHH for their support in the lab and for performing the in-solution captures.

Further thanks to Alissa and Philip for translating my dissertation summary.

To my fellow 'pathogen-people', thank you for your input and advice throughout many group meetings.

To all the past and present members of the Paleogenetics group at INA, University of Tübingen and DAG MPI-SHH for creating such lively and fun work environments. To my gals: Alissa, Kerttu and Joanna for their continued friendship, I look forward to many future get-togethers! Thanks to Betsy and Susanna for being awesome office-mates and for insightful TB discussions. To Maria, Marcel, Cosimo, Alex I., Michal, Alex P. and Verena who have worked alongside me since my time in Tübingen.

To Rhian and Natalie for a decade of friendship and for proofreading my dissertation.

To my parents: Cynthia and Sveinulf Vågene, for their love and unconditional support throughout my studies and for fostering my scientific curiosity from a young age.

Table of Contents

Abbreviations

aDNA	ancient DNA
BCE	Before Common Era
bp	base pairs
C	Cytosine
CE	Common Era
DNA	deoxyribonucleic acid
HTPS	High-throughput sequencing
MALT	MEGAN ALignment Tool
MTBC	*Mycobacterium tuberculosis* Complex
PCR	Polymerase Chain Reaction
qPCR	quantitative PCR
SNP	Single Nucleotide Polymorphism
T	Thymine
TB	tuberculosis
UDG	Uracil DNA glycosylase
UV	ultraviolet
YBP	years before present

Summary

Ancient pathogen genomes recovered from archaeological human remains provide primary evidence for infectious diseases that circulated amongst human populations in the past, as well as valuable insights regarding past pathogen macroecology, host-adaptation, spread and emergence. In this dissertation, I leverage recent molecular and computational advances to detect and characterize the genomes of pathogenic bacteria that affected indigenous human populations in the New World pre- and post- 16th century European contact.

In the first paper I investigate strain members of the *Mycobacterium tuberculosis* complex (MTBC), the causative agent of tuberculosis, in geographically dispersed human populations from inland Colombia and coastal Peru. I reconstructed and analyzed three pre-contact MTBC genomes belonging to the sub-clade *Mycobacterium pinnipedii,* associated with infection in seals and sea lions today. *M. pinnipedii* strains are hypothesized to have spread to coastal Peruvians via seal-to-human transmission (Bos et al., 2014). However, this ecological model does not explain the spread of *M. pinnipedii* to inland Colombian human populations. Different ecological models accounting for its inland spread are discussed.

The second paper presents *Salmonella enterica* ssp. *enterica* Paratyphi C, a cause of enteric fever, as a strong candidate pathogen for the post-contact 1545-1550 CE "*cocoliztli*" outbreak at Teposcolula-Yucundaa in Southern Mexico. The pathogenic agent of this outbreak is unknown from archaeological and ethnohistorical evidence. In this study I used a broad-scale computational screening approach, the MEGAN ALignment Tool (MALT), and was able to detect *S. enterica* DNA against a complex environmental DNA background in ten individuals buried in the epidemic cemetery. This was done without prior knowledge of the target organism, thus demonstrating the efficiency of such an approach in identifying ancient pathogens in archaeologically preserved tissues. Genome-wide analyses of the ten *S.* Paratyphi C genomes are presented.

In a third paper I applied MALT to screen for ancient pathogen DNA in the remains of a sub-adult individual from a deviant burial, dated to the 16th or 17th century, in Moneen Cave, Ireland. Negative results are reported for the presence of ancient pathogen DNA. Instead this paper focuses on the reconstruction and analysis of this individual's mitochondrial genome.

Zusammenfassung

Genome von Pathogenen, die aus menschlichen Überresten archäologischer Kontexte geborgen wurden, liefern einen direkten Nachweis für Infektionskrankheiten, die in menschlichen Population der Vergangenheit kursierten, und bieten wertvolle Einblicke in die Makroökologie (prä-)historischer Krankheitserreger, Anpassung der Wirte, und ihre Verbreitung und Erscheinung. In der vorliegenden Arbeit nutze ich neueste molekularbiologische und computergestützte Methoden, um pathogene Bakterien, die die Populationen der ‚Neuen Welt' vor und nach dem Kontakt mit Europäern im 16. Jh befielen, zu erfassen und genomisch zu beschreiben.

Im ersten Artikel untersuche ich Angehörige des *Mycobacterium tuberculosis* Komplexes (MTBC), dem Erreger von Tuberkulose, in geographisch verstreuten menschlichen Populationen des kolumbianischen Inlands und der peruanischen Küste. Ich konnte drei präkolumbianische MBTC-Genome der Untergruppe *Mycobacterium pinnipedii*, die heutzutage mit Infektionen in Seehunden und Seelöwen assoziiert werden, rekonstruieren und analysieren. Es wird vermutet, dass sich *M. pinnipedii*-Erregerstämme durch eine Übertragung von Seehunden auf die menschlichen Populationen der peruanischen Küste verbreitet haben (Bos, et al. 2014), jedoch erklärt diese ökologische Modell nicht die Verbreitung von *M. pinnipedii* auf die menschliche Populationen im Inland Kolumbiens. Daher werden abweichende ökologische Modelle betrachtet, die mit der Verbreitung ins südamerikanische Inland vereinbar sind.

Der zweite Artikel präsentiert *Salmonella enterica* ssp. *enterica* Paratyphi, einen Erreger des Typhus, als einen naheliegenden Kandidaten für den „*Cocoliztli*" - Ausbruch in Teposcolula-Yucundaa in Süd-Mexiko um 1545 - 1550 n.Chr., nach dem Eintreffen der Europäer. Der Erreger dieser Epidemie konnte bisher aus keinen archäologischen oder ethnohistorischen Befunden ermittelt werden. In dieser Studie wandte ich ein breit angelegtes Screening-Verfahren an, das MEGAN ALignment Tool (MALT), und konnte so *S. enterica*-DNA vor dem komplexen Hintergrund mikrobieller-DNA nachweisen. Dies wurde ohne *a prioi*-Wissen über einen anvisierten Erreger erreicht, welches die Wirksamkeit dieser Anwendung darin zeigt, alte Pathogene in archäologisch erhaltenem biologischen Gewebe zu identifizieren. Des Weiteren werden in der Studie Analysen der genomweiten Daten der zehn S. Paratyphi C Genome vorgestellt.

In einem dritten Artikel wandte ich MALT an, um DNA pathogener Organismen in den Überresten sub-adulter Männer aus einer Sonderbestattung, der 16. oder 17. Jahrhunderts, in der Moneen-Höhle in Irland nachzuweisen. Ich konnte keine antike Erreger DNA feststellen. Daher fokussiert sich die Untersuchung auf die Rekonstruktion und Analyse der mitochondrialen Genome der Individuen.

List of publications

The following three papers are included and discussed in this dissertation:

I. **Å. J. Vågene***, T. P. Honap*, K. M. Harkins, M. Rosenberg, F. Cárdenas-Arroyo, L. P. Leguizamón, J. Arnett, J. E. Buikstra, A. Herbig, A. C. Stone, K. I. Bos, J. Krause (2018). Geographically dispersed zoonotic tuberculosis in pre-contact New World human populations. *Manuscript*.

II. **Å. J. Vågene***, A. Herbig*, M. G. Campana, N. M. Robles García, C. Warinner, S. Sabin, M. A. Spyrou, A. Andrades Valtueña, D. Huson, N. Tuross, K. I. Bos, J. Krause (2018). Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution* 2:520-528.

III. **Å. J. Vågene**, J. Krause and K. I. Bos. (2016). Metagenomic analysis and mitochondrial genome reconstruction of the post-medieval individual from Moneen Cave. In: Dowd, M. (Ed.), *Archaeological Excavations in Moneen Cave, The Burren, Co. Clare*. Oxford, England: Archaeopress Publishing Ltd. (pp. 49-52).

*equal contributors

Reprints were made available with permission of the respective publishers.

Additionally, I am a co-author on the following articles published during my graduate studies:

K. I. Bos, G. Jäger, V. J. Schuenemann, **Å. J. Vågene**, M. A. Spyrou, A. Herbig, K. Nieselt, J. Krause. (2015). Parallel detection of ancient pathogens via array-based DNA capture. *Philosophical Transactions Royal Society London B Biological Sciences* 370:20130375.

A. E. Mann, S. Sabin, K. Ziesemer, **Å. J. Vågene**, H. Schroeder, A. T. Ozga, K. Sankaranarayanan, C. A. Hofman, J. A. Fellows Yates, D. C. Salazar-Garcia, B. Frohlich, M. Aldenderfer, M. Hoogland, C. Read, G. R. Milner, A. C. Stone, C. M. Lewis Jr., J. Krause, C. Hofman, K. I. Bos, C. Warinner. (2018). Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Scientific Reports* 8:9822.

I am also a shared first co-author on the following manuscript written during my graduate studies:

T. P. Honap*, **Å. J. Vågene***, M. S. Rosenberg, A. Herbig, A. T. Ozga, K. M. Harkins, C. Warinner, C. M. Lewis Jr., J. E. Buikstra, K. I. Bos, J. Krause, A. C. Stone. (2018). Mycobacterium tuberculosis Lineage 4 genomes from post-contact era North America. *Manuscript in preparation.*

*equal contributors

Own Contributions

I. I carried out DNA extraction for two of the samples and generated DNA libraries for three of the samples, included in this study, while at the University of Tübingen. I prepared all samples for in-solution capture and sequencing at the Max Planck Institute for the Science of Human History (MPI-SHH), together with Dr. Tanvi Honap from Arizona State University. I performed all data analyses of the shotgun and capture data, generated all figures (except Figure 1 in the main manuscript) and tables. I drafted the main manuscript and wrote the supplement. I co-edited the main manuscript together with Dr. Kirsten Bos, group leader at MPI-SHH in Jena. My work was carried out under the supervision of Prof. Johannes Krause.

II. I conducted labwork for all samples in this study together with Dr. Michael Campana (Harvard University), Maria Spyrou (MPI-SHH) and Susanna Sabin (MPI-SHH). I conducted the whole-genome array capture with support from Antje Wissgott, technician at the MPI-SHH. I performed all data analyses of the *S. enterica* sequencing data with support from Dr. Alexander Herbig, group leader at the MPI-SHH. I drafted the main manuscript and co-edited it together with Dr. Kirsten Bos (MPI-SHH). I wrote supplementary methods sections 2-14, and co-wrote supplementary methods section 1 with Prof. Noreen Tuross at Harvard University, and Prof. Christina Warinner, group leader at the MPI-SHH in Jena. I generated Figure 2 of the main manuscript together with Prof. Christina Warinner. I generated Figure 3 of the main manuscript, supplementary Figures 3-9 and all tables. My work was carried out under the supervision of Prof. Johannes Krause.

III. I carried out all the labwork, conducted all data analyses and wrote the manuscript under the supervision of Dr. Kirsten Bos and Prof. Johannes Krause.

1.Introduction

1.1 General introduction

Infectious diseases resulting from the transmission, spread and adaptation of a multitude of pathogenic microbes to the human host have significantly impacted human demography and biology, shaping cultural, political and economic landscapes (Dobson & Carper, 1996; J. F. Lindahl & Grace, 2015; McNeill, 1998). Analyses of microbial and human genetic material document this shared co-evolutionary history (Bliven & Maurelli, 2016; Karlsson, Kwiatkowski, & Sabeti, 2014), where pathogens represent one of the strongest selective forces to have acted on human populations (Fumagalli et al., 2011). However, relatively little is known about the deep-time evolutionary histories of pathogens and commensal microbes that emerged prior to the 20th century.

A pathogen's survival depends on its ability to persist, reproduce and adapt to new hosts and environments (Bliven & Maurelli, 2016). Emerging pathogens are those entering a new host and environment for the first time, while those classified as re-emerging have maintained a long-term relationship with a host and have re-emerged due to changes in conditions that are favorable to the spread and growth of that microbe's population (Jones et al., 2008). Human population growth, mobility, trade, antibiotic resistance, association with animals, and man-made environmental changes are all drivers of pathogen emergence (Cohen, 2000; Jones et al., 2008; Wolfe, Dunavan, & Diamond, 2007). Recent studies indicate that pathogens carried by wildlife and transmitted to humans through zoonotic events (animal-associated pathogens that can be transmitted to, and infect, humans) represent the most significant threat to global human health today (Jones et al., 2008). Approximately 61% of all recognized human pathogens are zoonotic in origin (Woolhouse, 2002). Many pathogens are estimated, or hypothesized, to have emerged in tandem with increased human population growth, and the adoption of agriculture and sedentism during the Neolithic Revolution, beginning ~10,000 YBP. The increase in human contact with animals in a setting of higher population densities is thought to have facilitated increased transmission, adaptation and exchange of pathogenic microbes between animals and humans (Harper & Armelagos, 2013). Today, human mobility, migration and trade have reached a pinnacle, and as a consequence, pathogens that emerge in one country are transported

between countries and continents at a rapid pace (Cohen, 2000; Institute of Medicine (US) Forum on Microbial Threats, 2006).

The introduction of Old World pathogens by European colonizers to New World indigenous populations in the 16th century caused devastatingly high rates of mortality via epidemic outbreaks that continued well into the 19th century (Dobyns, 1993). This represents one of the most catastrophic disease events to have occurred in recorded human history. Some historical epidemic events are well understood, such as the Black Death (1346-1353 CE) (Benedictow, 2004; Bos et al., 2011; Spyrou et al., 2016). However, much remains unknown about the 16th century New World epidemics due to a lack of, or conflicting, historical and archaeological evidence (Cook & Lovell, 2001; Dobyns, 1993). The disease exchange between the Old World and the New World was not equal, wherein European colonizers did not transport many infectious diseases back to the Old World (Crosby, 2003). This has led to debate over what the disease burden of New World human populations was and which pathogens were circulating amongst them prior to European contact (Drake & Oxenham, 2013; Larsen, 1994).

Today, modern medical efforts such as public health programs, antibiotics, vaccines and improved palliative care have done much to lessen the burden of infectious disease (Cohen, 2000; Dye, 2014). However, despite such advances, the worldwide number of human deaths caused by bacterial, viral and parasitic infections has decreased slowly, reducing only from 16 million to 15 million between 1990 and 2010 (Dye, 2014). Additionally, an unprecedented number of pathogens have emerged in recent decades with over 30 new pathogens identified over the course of the past 30 years (Mukherjee, 2017). Technological and societal changes continue to play a role in the emergence, and re-emergence, of human pathogens (Cohen, 2000; Wolfe et al., 2007).

The emergence and re-emergence of pathogens in a new or old host species is facilitated when pathogens are able to accumulate beneficial mutations in response to host immunity defenses, allowing them to be successful in the new host environment (Siddle & Quintana-Murci, 2014). The constant competition and counter-adaptation between host and pathogen has been dubbed the evolutionary ‘arms-race’ (Siddle & Quintana-Murci, 2014). Over time pathogenic strains may have outcompeted their

predecessors leading to strain replacement, loss of temporal structure and loss of past genetic diversity. Analyses based solely on extant strains may therefore produce factitious divergence time estimates and false impressions of the pathogen's virulence and diversity in the past (Achtman, 2016; Achtman, Zhou, & Didelot, 2015; Firth et al., 2010; Pimenoff, Houldcroft, Rifkin, & Underdown, 2018). Thus, in order to design effective preventative strategies and treatments for pathogens that cause clinically important infectious diseases today, it is vital to understand their evolutionary histories prior to the modern clinical era.

To achieve this, ancient pathogen genomes from different time periods and geographic regions are required for comparison with modern genomic datasets in order to constructively trace their evolutionary histories, emergence, diversity, macroecology and geographic range through time. It is in this area of research that the field of ancient DNA, and more specifically 'ancient pathogen genomics', is uniquely able to address such outstanding evolutionary questions for pathogens that have had, and continue to have, major effects on the course of human history. In this dissertation ancient DNA is used to investigate pathogens that circulated amongst pre- and post-contact New World human populations.

1.2 Ancient pathogen genomics: state of the art and technical challenges

Ancient DNA (aDNA) is the collective term used to refer to DNA extracted from the remains of organisms that have been deceased for a sufficient length of time for their DNA to undergo heavy degradation. As a research area within the field of ancient DNA, the study of ancient pathogens and commensal microorganisms is gaining increased attention from scientists, largely due to molecular and computational advances that have propelled this field forward in the last decade. In 2011 the first ancient pathogen genomes from the bacterium *Yersinia pestis* – the causative agent of the plague – were published, demonstrating that the reconstruction of complete bacterial genomes from archaeologically preserved host tissue is feasible (Bos et al., 2011). To date, a number of ancient pathogen genomes from a variety of species have been recovered and reconstructed from the bones, teeth and dental calculus of humans that perished up to several thousand years ago (Bos et al., 2014; Bos et al., 2011; Devault et al., 2017; Harkins & Stone, 2015; G. L. Kay et al., 2014; Schuenemann et al., 2013; Warinner et al., 2014; Zhou et al., 2018).

These advances were made feasible by the development of high-throughput sequencing (HTPS), allowing researchers to generate whole genomes without the need to target specific genomic regions via the polymerase chain reaction (PCR), the only method previously available (Krause, 2010). PCR amplifies (copies many times) regions of interest by using synthetic oligonucleotide primers complementary to conserved regions in the genome that flank regions of interest (Mullis et al., 1986). Prior to HTPS, researchers investigating ancient pathogens using PCR were mostly focused on determining the presence/absence of ancient pathogen DNA using one or a few genetic markers, which provided low (if any) resolution regarding the phylogenetic placement or evolutionary history of the pathogens (Donoghue et al., 2005; Harkins et al., 2015; Montiel et al., 2012; Muller, Roberts, & Brown, 2014). There are numerous additional drawbacks associated with PCR and its application to the study of ancient DNA, such as the exceedingly high cost-per-base sequenced, and the difficulty associated with verifying PCR results as true ancient DNA (R. E. Green et al., 2009; Willerslev & Cooper, 2005). This shed doubt on some of the earliest ancient pathogen findings (M. T. Gilbert et al., 2004; M. T. P. Gilbert et al., 2004; Shapiro, Rambaut, & Gilbert, 2006; Wilbur et al., 2009).

One major challenge in the study of ancient pathogen DNA is the identification of skeletal remains of individuals in the archaeological or historical record that died while infected, and that have detectable levels of pathogen DNA preserved. Only a few infectious diseases cause distinctive lesions to form in bone, these include tuberculosis, syphilis, brucellosis and leprosy, and do so only in cases of prolonged infection (Ortner, 2003). The absence of distinctive lesions is due to the lack of skeletal involvement in progression of the disease, the individual dying in the early phases of the disease before lesions are able to form, or a short period of infectivity (Ortner, 2003; Wood et al., 1992). Individuals that exhibit bone or soft tissue lesions characteristic of a particular disease have been targeted by ancient DNA researchers for the reason that they are easily visible in the archaeological record. Targeted sampling of skeletal remains, based on such lesions, has led to the successful reconstruction of the causative agents of leprosy (Mendum et al., 2014; Schuenemann, Avanzi, et al., 2018; Schuenemann et al., 2013), tuberculosis (Bos et al., 2014; Gemma L. Kay et al., 2015), syphilis (Schuenemann, Kumar Lankapalli, et al., 2018) and brucellosis (G. L. Kay et al., 2014).

Historical documents linking a particular infectious agent to an epidemic event have also proven useful, as was the case for several studies on *Yersinia pestis*, the causative agent of many well-documented plague outbreaks (Bos et al., 2016; Bos et al., 2011; Feldman et al., 2016; Spyrou et al., 2016). However, historical documents can be ambiguous in their descriptions of infectious diseases as they do not comply with modern clinical descriptions, are uninformed by the germ theory of disease, and may suffer from cultural biases and inaccuracies in translation.

The development of broad-scale screening approaches for pathogenic and commensal microorganisms, both at the molecular and computational level, have significantly advanced the area of 'ancient pathogen genomics' (Bos et al., 2015; Devault et al., 2014; Devault et al., 2017). The ability to search for the DNA traces of pathogenic microbes without prior indication from osteological or historical evidence has made it possible to screen for, and target, the genomes of all known pathogens that have been genetically characterized today. Thus, researchers are now able to screen for pathogen DNA in the remains of individuals for which no osteological or written evidence implicating a particular pathogen exists. Broad-range laboratory based molecular techniques have concentrated on pathogen detection via fluorescence-hybridization-based microarray technology (Devault et al., 2014) and identification via DNA enrichment of specific, seemingly unique, regions of pathogen genomes (Bos et al., 2015). Conversely, computational methods have focused on the analysis of sequence data generated from non-enriched libraries that were subsequently screened against human microbiome datasets (Devault et al., 2017) or datasets comprising all currently known bacterial and viral species for which complete genomes exist (paper II).

After an organism has died, its genetic material is no longer subject to cellular repair mechanisms, and it therefore begins to degrade due to a variety of factors such as oxidative, hydrolytic and enzymatic damage (Dabney, Meyer, & Paabo, 2013; T. Lindahl, 1993; Sawyer, Krause, Guschanski, Savolainen, & Paabo, 2012). These forms of degradation accumulate over time and the rates at which they accumulate are influenced by the temperature and humidity of the burial environment (Kistler, Ware, Smith, Collins, & Allaby, 2017). Oxidative and hydrolytic damage cause chemical changes in individual bases of the DNA, and hydrolytic damage can cleave the sugar-

phosphate backbone causing single-strand breaks. Intracellular nucleases are no longer controlled after cell death, therefore exposing the DNA to enzymatically driven fragmentation. Overall, degradation manifests itself as high rates of DNA fragmentation, depurination and cytosine deamination on the terminal ends of DNA fragments (A. Briggs, Stenzel, & Johnson, 2007; Dabney, Meyer, et al., 2013). The latter is commonly used as a measure to determine if the DNA is ancient (A. Briggs et al., 2007; Jonsson, Ginolhac, Schubert, Johnson, & Orlando, 2013; Sawyer et al., 2012).

To date, RNA viruses that make up a number of important human pathogens have not been the subjects of ancient pathogen genetic studies. The methods developed for the study of ancient RNA are comparably underdeveloped to that of aDNA (Fordyce, Kampmann, van Doorn, & Gilbert, 2013; Guy, 2014). This is largely driven by the fact that RNA is more prone to post-mortem degradation than DNA, because the chemical structure of RNA is less stable and makes it more prone to hydrolytic damage (Fordyce et al., 2013). Additionally, it is highly susceptible to degradation by RNAses, and is therefore less likely to survive the initial stages of decomposition of organic tissues (Fordyce et al., 2013; Guy, 2014).

A further complicating factor in the study of ancient pathogen and commensal microbes is the environmental contamination of archaeological skeletal remains derived from the burial environment, especially by soil-dwelling microbes, some of which share close genetic affinity with microbes of interest (Warinner et al., 2017). Ancient microbial DNA is usually preserved at very low quantities in skeletal remains (bodily tissues such as dental calculus and paleofeces are an exception to this (E. J. Green & Speller, 2017)) and the introduction of a large amount of environmental contamination stands the risk of ‘drowning-out’ the ancient pathogen signal to undetectable levels or causing false-positive results. Incorporation of appropriate screening and authentication efforts can help avoid false positive results, as outlined by Warinner et al. (2017) and Key, Posth, Krause, Herbig, and Bos (2017). Computational screening methods that take the known diversity of environmental microbes into account, such as the MALT pipeline presented in paper II, reduce to some degree the bias in the computational detection of ancient pathogen DNA against a DNA background derived from genetically similar environmental microbes. In addition, several authentication methods have been developed to distinguish true aDNA from

modern contamination, with some methods taking advantage of the, somewhat predictable, degradation patterns often found in ancient DNA (Jonsson et al., 2013; Key et al., 2017; Prufer et al., 2010; Warinner et al., 2017).

Once a pathogen of interest has been identified a genome can be generated through direct sequencing of a DNA library or through the use of whole-genome targeted capture enrichment, i.e. fishing out the DNA of interest using synthetic oligonucleotide probes. Genome-wide studies have significantly advanced our understanding of the evolutionary histories of pathogens that circulated amongst humans in the past, allowing researchers to analyse ancient genomes in many of the same ways they would analyse modern genomes. Specific revelations learned from the analysis of ancient pathogen genomes pertain to their acquisition of virulence factors, divergence time estimates, vector dynamics, strain diversity, geographic range and host adaptation (Achtman, 2016; Andam, Worby, Chang, & Campana, 2016; Harkins & Stone, 2015). The lack of external calibration points from archaeological sources that document the presence of pathogens throughout human history makes ancient pathogen genomes invaluable for molecular dating analyses (Achtman, 2016; Leonardi et al., 2017). These are vital for understanding the age at which pathogens emerged, and the rates at which they have accumulated genetic changes.

1.3 Infectious disease in the pre- and post-contact New World

The first humans to colonize the New World travelled via the Bering Land Bridge that connected Siberia and Alaska between 15,000-11,000 YBP, when this landmass was last exposed (Dyke, 2004; Jakobsson et al., 2017; Pedersen et al., 2016). The earliest evidence for human occupation in the New World appears in the archaeological record 14,500 YBP (Gilbert et al., 2008; Skoglund & Mathieson, 2018). The number of founding individuals is estimated to have been extremely low, with one recent estimate consisting of ~250 individuals (Fagundes et al., 2018). Genetic evidence supports this severe human population bottleneck and indicates that the geographic distribution of human genetic structure in the Americas was established soon after its initial colonization and received limited gene flow from genetically different regions before the arrival of Europeans in the 16th century (Llamas et al., 2016; O'Rourke, Hayes, & Carlyle, 2000; Reich et al., 2012). Given the timing of the initial settlement of the New World, the founding individuals had not been exposed to the Old World

infectious diseases that emerged, or re-emerged, during the Neolithic Revolution. Instead, it is hypothesized that indigenous New World populations evolved immunity genes that were adapted to the local pathogen environment encountered in the New World (Merbs Charles, 1992; Ramenofsky, 2003), a trend observed world-wide for humans who have settled in new environments (Fumagalli et al., 2011).

The majority of what is known about pathogens that circulated amongst indigenous New World populations before European contact has been acquired from the study of skeletal and mummified human remains (Drake & Oxenham, 2013; Ramenofsky, 2003). Based on osteological assessments and microscopic observations of parasites, researchers assert that tuberculosis (TB), treponematosis (syphilis, yaws, bejel and pinta) and a multitude of parasitic infections, such as leishmaniasis and chagas disease, were prevalent during the pre-contact era (Drake & Oxenham, 2013; Larsen, 1994; Roberts & Buikstra, 2003). Many studies have also reported periosteal changes consistent with soft tissue infection, systemic bacterial infection and trauma in New World human skeletal remains, but these are not indicative of a particular pathogen (Larsen, 1994). Ancient DNA studies support the archaeological observations of pre-contact tuberculosis via the PCR detection of *Mycobacterium tuberculosis* Complex (MTBC) DNA (Salo, Aufderheide, Buikstra, & Holcomb, 1994) and the reconstruction of complete MTBC genomes (Bos et al., 2014) (see section 1.4). However, much remains unknown about the burden of infectious disease among indigenous New World populations in the pre-contact era, and it has been questioned whether the population size and density was high enough to sustain epidemic scale outbreaks (Drake & Oxenham, 2013). The population density in Mesoamerica is estimated to have been high in areas associated with certain well-established civilizations. Therefore, it has been suggested that the demise of several Mesoamerican cultural groups (Teotihuacan, Maya, Zapotec, Mixtec) between 750-950 CE may have been related to epidemic outbreaks of disease (Acuna-Soto, Stahle, Therrell, Gomez Chavez, & Cleaveland, 2005). Overall, studies of pre-contact New World human remains have refuted the earlier misconception that the New World was ‘disease-free’ prior to the arrival of European colonizers (Larsen, 1994).

It is widely agreed upon that the introduction of Old World pathogens to the New World by European colonizers, beginning in 1492 CE, caused the disease burden

of the indigenous people to increase markedly (Acuna-Soto, Stahle, Therrell, Griffin, & Cleaveland, 2004; Cook, 1998; Cook & Lovell, 2001; Crosby, 1976; Dobyns, 1993; Joralemon, 1982; Ubelaker, 1976). The devastating consequences were manifested as several major and numerous minor epidemic outbreaks, which continued well into the 19th century, and are reported to have caused up to 90-95% population loss in some cases (Dobyns, 1993). The causative pathogens responsible for some of these outbreaks have been attributed to Old World infectious diseases such as smallpox, measles, influenza and mumps (Cook & Lovell, 2001; Crosby, 1976; Dobyns, 1993; Fields, 2008). However, the causative agents of many post-contact New World epidemics remain unknown.

Mexico was particularly heavily affected, enduring three major epidemics in the 16th century alone. The first occurring in 1521 CE was likely caused by smallpox, followed by two epidemics beginning in 1545 CE and 1576 CE, both locally known as '*cocoliztli*', meaning 'pestilence' in Nahuatl (the widespread Aztec language) (Cook, 1998; Cook & Lovell, 2001). The pathogenic agents responsible for the 1545 CE and 1576 CE epidemics remain unclear based on historical and archaeological evidence alone (Cook & Lovell, 2001; Fields, 2008; Warinner, Robles García, Spores, & Tuross, 2012). The 1545 CE epidemic has been ascribed as one the most devastating epidemiological events to have affected Mesoamerica in the post-contact era, however, it is not known how many people perished. One estimate suggests between 60-90% of the population of Mexico and Guatemala was affected (Acuna-Soto, Stahle, Cleaveland, & Therrell, 2002; Cook & Lovell, 2001). The pathogenic cause of the 1545 CE epidemic is investigated using ancient DNA in paper II of this dissertation.

The exceedingly high levels of population loss documented for New World populations in reaction to Old World diseases led some to hypothesize that New World populations were more genetically susceptible to the pathogens due to low genetic diversity in immunity related genes and lacked previous exposure (Black, 1992; Crosby, 1976). However, as Lindo et al. (2016) note, the assumption of homogeneity among specific immune genes is based on observations of surviving communities of Native American populations. Therefore, these assumptions fail to account for genetic diversity of immune genes that may have existed prior to contact (Lindo et al., 2016). What can be asserted is that the population of the New World experienced strong

selective force from pathogens and that the human population structure of the New World was intensively shaped by pathogen driven selection during the last 500 years.

1.4 Pre-contact tuberculosis in the New World

Tuberculosis (TB) is a bacterial disease caused by members of the *Mycobacterium tuberculosis* complex (MTBC). The MTBC consists of seven human adapted lineages belonging to *Mycobacterium sensu strictu* and *Mycobacterium africaanum*, along with several animal-adapted strains and the ancestral "smooth tubercle bacilli" *Mycobacterium cannettii* (Gagneux, 2018). Human-adapted Lineage 4 strains predominantly associated with Europe remain the main pathogenic cause of human TB across the New World today, after being introduced during 16th c. European colonization (Gagneux, 2018; O'Neill et al., 2018; Pepperell et al., 2011). Skeletal pathologies document the presence of TB in indigenous New World populations many centuries before European contact (Roberts & Buikstra, 2003; Stone, Wilbur, Buikstra, & Roberts, 2009), an observation that was difficult to reconcile with the dominance of modern Lineage 4 strains today. The earliest pre-contact TB cases appear in South American human remains in 290 CE in Chile, and possibly as early as 160 BCE in Peru (Allison, Gerszten, Munizaga, Santoro, & Mendoza, 1981; Roberts & Buikstra, 2003). Cases of skeletal lesions consistent with tuberculosis infection become more prevalent in South America from 700 CE onwards, with the first cases appearing in North America in 900 CE (Roberts & Buikstra, 2003). In South America, the highest density of cases exist in Peru and Chile, prompting researchers to speculate if this was the region where human MTBC initially developed in the pre-contact New World (Roberts & Buikstra, 2003). Currently, skeletal TB lesions have been identified in human remains at more than 99 pre-contact archaeological sites across South and North America (Drake & Oxenham, 2013; Roberts & Buikstra, 2003).

Pulmonary TB infection occurs when MTBC bacteria are transmitted via droplet infection to the lungs; this is the most common form of TB. Extrapulmonary TB occurs when the bacteria infecting the lungs spread to other organs in the body, or when MTBC bacteria are introduced via the consumption of, or contact with, infected animal tissues leading to gastrointestinal disease (Lee, 2015). If an active infection persists it may spread via the blood or lymphatic system to other organs such as bones, joints, lymph nodes, gastrointestinal organs, pleura and meninges (Lee, 2015; Monack,

Mueller, & Falkow, 2004). When a skeletal MTBC infection is maintained over a prolonged period of time, lesions form in the bone as part of an inflammatory response. However, only certain vertebral lesions are considered pathognomonic of prolonged MTBC infection, even though infection may occur in any skeletal element (Ortner, 2003; Roberts & Buikstra, 2003). Such skeletal lesions can be identified in the archaeological record and have been used by researchers to target human remains for molecular screening for the presence of preserved MTBC DNA (Arriaza, Salo, Aufderheide, & Holcomb, 1995; Bos et al., 2014; Braun, Collins Cook, & Pfeiffer, 1998; Harkins et al., 2015; Konomi, Lebwohl, Mowbray, Tattersall, & Zhang, 2002; Salo et al., 1994).

PCR studies provided the first molecular confirmation of pre-contact tuberculosis in the New World (Arriaza et al., 1995; Braun et al., 1998; Konomi et al., 2002; Salo et al., 1994). In a recent study by Bos et al. (2014) three complete MTBC genomes were isolated from the skeletal remains of three individuals from three distinct sites located along the Osmore River Drainage on the southern coast of Peru. These genomes belong to *Mycobacterium pinnipedii*, an MTBC lineage that is most commonly associated with infection in pinnipeds (seals and sea lions) today (Boardman et al., 2014; de Amorim et al., 2014; Jurczynski et al., 2011; A. Kiers, Klarenbeek, Mendelts, Van Soolingen, & Koeter, 2008; Loeffler et al., 2014). Molecular dating analyses incorporating these ancient genomes yielded an estimated date of MTBC emergence of approximately 6,000 YBP, thus seemingly precluding the transfer of MTBC from the Old World during initial human settlement of the New World, between 15,000 to 11,000 YBP (Bos et al., 2014) (see Discussion section 5.5). Instead, these *M. pinnipedii* infections in Peruvian humans are hypothesized to have arisen from ancient zoonotic events mediated by infected pinniped tissues obtained via hunting (Bos et al., 2014). The pinniped-human transmission model of the bacterium is, however, incompatible with the observation of TB lesions at inland sites across the New World. In paper I, DNA extracted from inland human remains from Colombia and coastal highland Peru are analyzed for the purpose of better understanding the past diversity of ancient MTBC strains across South America.

2. Goals and Objectives

The unifying objective of this dissertation was to apply cutting-edge molecular and computational techniques to detect, recover, reconstruct and analyze genome-wide DNA data from ancient pathogens extracted from human skeletal remains preserved in the archaeological record.

This objective is realized through a series of three studies that, combined, tackle the three primary goals of this dissertation: 1) to genetically determine the causative pathogenic agents of archaeologically and historically recorded infectious disease events in the New World, by analyzing pre- and post-contact era human remains; 2) to improve our understanding of the evolutionary histories, past genetic diversity and macroecology of ancient pathogens; 3) to apply a novel computational tool (MEGAN ALignment Tool) to screen for and detect ancient pathogen DNA in metagenomic DNA samples without prior knowledge of the target organism(s).

The studies presented in this dissertation (papers I, II and III) provide results that supplement archaeological, biological and historical findings. Many aspects of infectious disease in the pre- and post- contact Americas remain unknown. This dissertation aims to shed light on two such archaeologically and historically significant questions in papers I and II.

In paper I, I investigated pre-contact tuberculosis in the New World. This paper is a direct continuation of the findings presented in Bos et al. (2014). The aim of paper I is to shed light on the following questions: what was the diversity of MTBC strains that circulated in pre-contact South America? Did the previously discovered Peruvian MTBC strains (Bos et al., 2014) affect human populations elsewhere in South America? Were the Peruvian strains present at inland sites, and if so, did transmission occur with the involvement of animal hosts or via human-to-human transmission?

The aim of Paper II is to shed light on the causative pathogenic agent(s) responsible for the 1545 CE '*cocoliztli*' epidemic at Teposcolula-Yucundaa in southern Mexico. The skeletal collection that I had access to for this study did not indicate a causative agent based on skeletal morphology, and nor did the historical evidence point to a clear pathogenic agent. Therefore, the second aim of paper II was to tackle the question faced by those who study infectious disease in the past: how to screen skeletal remains for molecular traces of pathogenic agents where skeletal and historical evidence is absent or unclear regarding the causative agent. A broad-scale screening

approach using a novel computational tool (MALT) was applied with the aim of addressing this issue.

In paper III, MALT was also applied with the aim of screening for ancient pathogens. In this case, I applied it to the remains of an individual buried in Moneen Cave in Ireland. A pathogen associated with a stigmatized infectious disease may account for the remains of this individual being sequestered from the normal burial ground.

3. Methods

Precautions for avoiding contamination

Many of the molecular and computational methods applied in papers I, II and III were specifically adapted and developed for ancient DNA work. Due to the degraded condition and low abundance of ancient DNA, strict precautions are required during sample preparation, DNA extraction and the building of DNA libraries to avoid further contamination by human and/or environmental DNA, as well as cross-contamination between samples. The wet-lab procedures were carried out in pre-PCR cleanroom facilities dedicated to ancient DNA work at the University of Tübingen, Arizona State University and The Max Planck Institute for the Science of Human History (MPI-SHH).

Before entering the cleanroom all researchers took the required precautions to protect the samples from human contamination by dressing in full body suits, face masks, hairnets, protective footwear and two to three pairs of gloves – the outermost glove was changed frequently to reduce cross-contamination between the different processing steps. Surfaces in the cleanroom facilities were routinely UV irradiated and cleaned with a diluted solution of bleach to minimize DNA cross-contamination. When appropriate, chemical reagents were UV irradiated to remove contaminant DNA, and re-usable items, such as sampling tools, were routinely sterilized using diluted bleach and UV irradiation between each use.

Sample processing and extraction

The skeletal elements chosen for pathogen screening varied depending on the goal of the study. In paper I ribs and vertebrae carrying skeletal lesions consistent with prolonged MTBC infection were chosen. In papers II and III tooth pulp-chambers were sampled since they had proven to be a good source of ancient pathogen DNA in previous studies (Andrades Valtuena et al., 2017; Bos et al., 2011; Feldman et al., 2016; Spyrou et al., 2016). The pulp-chambers of teeth are highly vascularized and somewhat protected from the external environment by the hard enamel surface of the tooth crown, two factors that will maximize one's chance of finding molecular traces of blood-borne pathogens.

Dirt was removed from surfaces of the bones and teeth prior to sampling, and bone samples processed by collaborators at the University of Arizona were then wiped with a diluted bleach solution and UV irradiated in a further effort to remove environmental DNA contamination. I did not expose the samples that I processed at the University of Tübingen to bleach solution for fear of degrading the target DNA. Teeth were cross-sectioned at the cementoenamel junction, separating the tooth crown from the root, and dentine was sampled from the inside of the crown using a sterile dental drill. Sampling of the vertebrae and ribs was also carried out using a dental drill. The vast majority of samples were extracted using a silica-based protocol specifically developed for the retrieval of extra short DNA fragments, a characteristic feature of ancient DNA (Dabney, Knapp, et al., 2013). In paper I, one sample (AD82) was previously extracted using an older protocol by Rohland and Hofreiter (2007).

DNA Library preparation

DNA libraries compatible with Illumina sequencing platforms were built by ligating adapters to both ends of the extracted DNA fragments (Meyer & Kircher, 2010) and by subsequently attaching unique 8bp index pairs to the ends of the adapters through PCR extension (Kircher, Sawyer, & Meyer, 2012). The universal priming sites included in both the adapter and index sequences allow all DNA fragments to be amplified simultaneously during PCR. Uracil sites occur in ancient DNA as a result of deaminated cytocines (C), therefore polymerases that are able to incorporate uracil sites during amplification by replacing them with a thymine (T) were used (Heyn et al., 2010). Pfu Turbo Cx Hotstart Polymerase (Agilent Technologies) and AmpliTaq Gold (Applied Biosystems) were used in paper I, and only the former was used in papers II and III. These polymerases allow the characteristic damage pattern of DNA to remain intact and quantifiable after sequencing (Heyn et al., 2010). Sample extracts determined to be positive for pathogen DNA were made into new libraries treated with uracil DNA glycosylase (UDG) and endonuclease VIII to repair and remove the deaminated bases from the ancient DNA (A. W. Briggs et al., 2010). This was done to facilitate the generation of high-quality data for genome reconstruction through downstream capture experiments.

All PCR, quantitative PCR (qPCR) and post-indexing experiments were carried out in modern lab facilities. Two polymerases with minimal fragment length and GC

bias were used for subsequent amplifications: AccuPrime Pfx Polymerase (Invitrogen) and Herculase II Fusion (Agilent Technologies) (Dabney & Meyer, 2012). While working in the facilities at the University of Tübingen, it was discovered that AccuPrime Pfx Polymerase (Invitrogen) is prone to creating primer-dimers during amplification; it was replaced with Herculase II Fusion (Agilent Technologies) in the modern lab facilities in the University of Tübingen and MPI-SHH.

Pathogen Screening

Molecular and computational screening approaches were used according to the goals of the various studies. In paper I, a two-tiered molecular screening approach was used to target ancient MTBC DNA. The first approach was a qPCR assay targeting three different mycobacterial gene regions less than 100bp. This assay was applied to DNA extracts and measured the amplification of the target region in real time. The positive extracts were then converted into libraries and screened by targeting five complete mycobacterial genes according to Bos et al. (2014) using the in-solution capture protocol presented by Maricic, Whitten, and Paabo (2010).

A computational approach for screening shotgun-sequenced data was used in papers II and III. The shotgun strategy entails the direct sequencing of libraries without prior molecular enrichment. The MEGAN ALignment Tool (MALT) (paper II) was used to screen shotgun data for traces of ancient pathogen DNA. MALT is a fast and sensitive tool that carries out the alignment and analysis of metagenomic sequencing data against a specified database and assigns reads to their taxonomic node of best fit. It was applied for pathogen screening in papers II and III using a database consisting of all complete bacterial genomes available through NCBI RefSeq (December 2016). A second database consisting of the full NCBI RefSeq database was also used in paper II for the purpose of screening for DNA viruses. MALT was also used to assess the metagenomic composition of the sequenced libraries in papers I and II.

Whole-genome enrichment

After screening, all sample libraries positive for ancient pathogen DNA were converted into UDG-treated libraries (A. W. Briggs et al., 2010) intended for whole-genome capture. UDG-treated libraries were captured for the purpose of generating high-quality data for targeted bacterial pathogen genome reconstruction. Non-UDG

libraries were also captured in order to use the on-target data to generate deamination plots of the captured product (Jonsson et al., 2013). Two different capture methods were used in paper II: array capture (Burbano et al., 2010; Hodges et al., 2009) and in-solution capture (Fu et al., 2013). Array capture was first applied for the capture of three strong positive *S. enterica* samples (paper II). With the discovery of seven additional weak positive samples, in-solution capture, which had been newly implemented in the MPI-SHH modern lab, was applied to all ten positive samples (paper II). Comparison of samples captured using both methods (and the same probe design) revealed that in-solution capture generated more on-target data. In-solution capture had also worked successfully in generating genome-wide data from the seven weak positive samples (see Discussion section 5.2). With this in mind, in-solution capture was chosen over array capture for the enrichment of MTBC DNA in paper I, where all positive samples contained low amounts of target DNA (see Fig. 1 in Discussion section 5.2).

Synthetic oligonucleotide probes required for both MTBC and *S. enterica* whole-genome capture assays were designed based on the target organism(s). The *S. enterica* assay consisted of 67 chromosome/assembly and 45 plasmid reference sequences (paper II). In contrast, the MTBC assay in paper I was based on a single genome: the reconstructed hypothetical ancestor for MTBC (Comas et al., 2010). *S. enterica* is genetically diverse bacterial species (Alikhan, Zhou, Sergeant, & Achtman, 2018), which is why a large number of genomes were included in the probe design. Conversely MTBC strains have highly conserved genome sequences retaining >99% similarity (Brosch et al., 2000). Additionally, many genetically similar soil-dwelling mycobacterial species may be present in the DNA extracts and libraries due to contamination from the burial environment. Therefore, in order to increase capture specificity only the MTBC ancestor reference was used. This reference genome is basal to the MTBC and does not skew the capture towards a specific extant strain type.

Data analyses

The GUI-based EAGER pipeline incorporates a series of computational tools used to process sequencing data (Peltzer et al., 2016). EAGER was used in all three papers for the pre-processing of sequencing data (Peltzer et al., 2016; Schubert, Lindgreen, & Orlando, 2016), mapping (Li & Durbin, 2009), duplicate removal (Peltzer et al., 2016), SNP calling (McKenna et al., 2010) and generating deamination plots

(Jonsson et al., 2013). For study specific analyses, refer to the papers and corresponding supplementary materials presented in the Appendices.

4. Results

4.1 Pre-contact tuberculosis genomes from South American human populations

I. **Å. J. Vågene***, T. P. Honap*, K. M. Harkins, M. Rosenberg, F. Cárdenas-Arroyo, L. P. Leguizamón, J. Arnett, J. E. Buikstra, A. Herbig, A. C. Stone, K. I. Bos, J. Krause (2018). Geographically dispersed zoonotic tuberculosis in pre-contact New World human populations. *Manuscript*.

* equal contributors

Synopsis

Paper I presents three novel ancient *Mycobacterium tuberculosis* complex (MTBC) genomes isolated using genome-wide in-solution capture applied to DNA extracted from the skeletal remains of three individuals from geographically distinct sites in South America. These individuals pre-date European arrival in the New World and come from two sites located in inland Colombia, and one from high-land coastal Peru. These MTBC genomes phylogenetically cluster with *Mycobacterium pinnipedii* strains, which are most commonly found circulating in pinnipeds (seals and sea lions) today (Cousins et al., 2003).

Three previously published pre-contact *M. pinnipedii* genomes isolated from human skeletal remains from Peru (Bos et al., 2014) also come from the same geographic region as the Peruvian genome presented in paper I. This finding is interpreted as a result of an ancient zoonotic event that facilitated the spread of *M. pinnipedii* from pinnipeds to these coastal Peruvian humans via the consumption or manipulation of infected seal tissues (Bos et al., 2014). Together all four Peruvian genomes support this hypothesis based on their geographic location and the history of pinniped exploitation associated with this coastal region. However, skeletal evidence from sites across North and South America demonstrates the presence of pre-contact human tuberculosis infections at inland sites, which are geographically incompatible with a model of direct pinniped-to-human transmission. The Colombian *M. pinnipedii* genomes come from inland sites located too far from the coast to support direct human-to-pinniped transmission (paper I). Different scenarios for the spread and transmission

of *M. pinnipedii* to individuals at inland Colombian sites are discussed, including the potential of human adaptation and/or animal mediated dissemination.

Metagenomic analyses of the shotgun and genome-wide capture data show an abundance of DNA from genetically similar soil mycobacteria that infiltrated and contaminated the samples. The presence of this mycobacterial background affected the efficiency of both the whole-genome enrichment of ancient MTBC DNA as well as downstream analyses.

Together, the data presented in paper I demonstrate the ability of ancient *M. pinnipedii* strains to cause human infection in the past and point to a more complex transmission route than simple pinniped to human transfer in the case of the inland Colombian individuals. Additionally, these data provide further molecular support to the archaeological evidence for dispersed tuberculosis infections circulating amongst pre-contact humans in the New World.

4.2 Post-contact epidemic disease in Mexico

II. **Å. J. Vågene***, A. Herbig*, M. G. Campana, N. M. Robles García, C. Warinner, S. Sabin, M. A. Spyrou, A. Andrades Valtueña, D. Huson, N. Tuross, K. I. Bos, J. Krause (2018). Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution* 2:520-528.

* equal contributors

Synopsis

Paper II investigates the pathogenic cause of an epidemic outbreak that occurred in 1545-1550 CE at Teposcolula-Yucundaa in southern Mexico, by making use of cutting-edge computational techniques to screen for ancient pathogen DNA. The epidemic cemetery at Teposcolula-Yucundaa is unique in that it is the only known cemetery that is historically linked to 1545 CE epidemic. The 1545 CE outbreak, referred to as '*cocoliztli*' by the indigenous Mexican population, a generic term meaning 'pestilence' in Nahuatl (the Aztec language), was one of the most devastating to occur in the New World after European arrival (Cook, 1998; Cook & Lovell, 2001). The 1545 CE *cocoliztli* affected large parts of Mexico and Guatemala and is estimated to have killed between 60-90% of the indigenous population (Cook & Lovell, 2001). It

was one of many infectious disease outbreaks that occurred after the arrival of the Europeans in the 16th century. The pathogenic cause of this epidemic has been debated for over a century and is not known based on historical or archaeological evidence alone.

DNA extracts from teeth sampled from 24 individuals buried in the epidemic cemetery and 5 individuals buried in the pre-contact cemetery at Teposcolula-Yucundaa were screened for genetic traces of ancient pathogens using a novel computational tool called the MEGAN ALignment Tool (MALT), officially presented in paper II. MALT, is a metagenomic analysis tool that rapidly aligns reads from DNA sequencing data to the genetic elements included in the users' database of choice, assigning reads to the taxonomic node of best fit. The MALT pipeline used is specifically tailored to the screening and detection of ancient microbial DNA against a complex background of environmental contaminant DNA. A comparative runtime analysis of MALT and the gold standard Basic Local Alignment Search Tool (BLAST) revealed a greater than 200-fold improvement in computation time over BLASTn.

MALT analysis of the sample shotgun data revealed the presence of putative ancient pathogen DNA belonging to *Salmonella enterica* subsp. *enterica* in ten individuals from the epidemic cemetery. Three samples (Tepos_10, Tepos_14 and Tepos_35) showed a particularly strong genetic signal that indicated that the specific strain-type might be *Salmonella enterica* subsp. *enterica* Paratyphi C, a bacterial cause of enteric (paratyphoid) fever. This finding was followed up with whole-genome capture to enrich for any ancient *S. enterica* strains that might be present in the samples. Successful genome captures yielded genome-wide *S.* Paratyphi C data from all ten individuals. Samples from five of the individuals yielded sufficient genetic data to allow complete *S. Paratyphi* C genome reconstruction and phylogenetic placement. The data from the remaining five samples were also determined to belong to *S.* Paratyphi C through SNP-based analyses.

The results presented in paper II provide the first primary evidence for the potential pathogenic cause of the 1545-1550 CE epidemic at Teposcolula-Yucundaa, where *S.* Paratyphi C is proposed as a strong candidate pathogen. The efficiency of MALT in detecting ancient pathogen DNA is successfully demonstrated. This study

further illustrates that bioinformatics tools such as MALT can overcome the hurdle of detecting human pathogens in archaeological human remains, even in cases where the candidate organism is not known based on prior evidence.

4.3 Pathogen screening and mitochondrial genome reconstruction of an individual from Moneen Cave, Ireland

III. **Å. J. Vågene**, J. Krause and K. I. Bos. (2016). Metagenomic analysis and mitochondrial genome reconstruction of the post-medieval individual from Moneen Cave. In: Dowd, M. (Ed.), *Archaeological Excavations in Moneen Cave, The Burren, Co, Clare*. Oxford, England: Archaeopress Publishing Ltd. (pp. 49-52).

Synopsis

The human remains belonging to a 14 to 16-year-old juvenile, dated to the 16th or 17th century, were excavated from Moneen Cave, Co. Clare, Ireland. The practice of interning a single individual in a cave breaks with Christian burial practices in Ireland at the time (Garattini, 2007). The reason for this individual's internment in Moneen Cave is unknown and the focus of paper III was to screen for the presence of ancient pathogen DNA to determine if this individual suffered from an infectious disease at the time of death. The stigma associated with certain diseases might provide an explanation for why this individual's remains were sequestered to an abnormal burial location.

MALT was used to screen for the presence of ancient pathogen DNA in shotgun data generated from DNA extracted from a tooth. A molecular signal of the oral microbiome was detected, however molecular traces of DNA-based bacterial or viral pathogens were absent from the data. Although it was not possible to detect any ancient pathogen DNA, it was possible to use the shotgun data to determine the gender of the individual as male using a computational tool (Skoglund, Stora, Gotherstrom, & Jakobsson, 2013), confirming previous genetic tests for this individual based on PCR methods (Taylor, 2016). It was also possible to reconstruct the complete mitochondrial genome of this individual to an average coverage of 14.5-fold. The mitochondrial genome belongs to haplogroup J2b1b1, which is associated with modern populations from Europe and the Near East (Pierron et al., 2011). The mitochondrial genome was

also investigated for the presence of mutations that cause Leber's hereditary optic neuropathy (LHON), a genetic disease that causes individuals to become prone to blindness and deafness (Achilli et al., 2012), but none were found.

Although it was not possible to detect any pathogens that could be responsible for the death of this individual, it cannot be concluded that this individual did not suffer from an infectious disease at the time of death. Ancient pathogen DNA is not guaranteed to preserve over time and it may be difficult, or impossible, to detect when the environmental background is high.

5. Discussion

Here I discuss the impact of the findings presented in this dissertation as they relate to the historical and archaeological context of the New World: what they tell us, what they do not, and which questions are raised that should be addressed by future research. I also discuss their contribution to our understanding of the transmission, spread and macroecology of *Mycobacterium pinnipedii* and *Salmonella enterica ssp. enterica* Paratyphi C in the past and what it means for our understanding of these pathogens in the present. To begin, I discuss the methods applied and challenges faced in the study of 'ancient pathogen genomics'.

5.1 Environmental contamination (papers I, II)

Papers I and II demonstrate that the detection and retrieval of ancient microbial DNA is encumbered by the presence of contaminating environmental DNA, particularly when environmentally derived bacterial species that are genetically similar to the target organism are abundant.

In paper I, I found that high backgrounds of environmental mycobacterial DNA from soil dwelling species were a complicating factor in the enrichment and analysis of ancient MTBC DNA. Non-MTBC mycobacterial DNA was co-enriched alongside the targeted ancient MTBC DNA during capture due to the lenient hybridization specificity. This is a well-known phenomenon, and has been exploited in studies seeking to enrich genetic material from organisms that do not have reference genomes, by using probes based on genetically similar organisms (Burbano et al., 2010; Immel et al., 2015). The enrichment of non-MTBC mycobacterial species during hybridization capture was detected via the abundance of unexpected heterozygous (multiallelic) positions in the three novel ancient *M. pinnipedii* genomes. The contaminating reads could not be filtered out through the use of stringent mapping criteria and their persistence in the dataset manifested as unexpectedly long genome branch lengths in phylogenetic analyses. They occurred because a subset of the SNPs introduced by the contaminating reads infringed on the threshold used to define homozygous SNP calls. These branch lengths are comparatively much longer than those of the contemporaneous *M. pinnipedii* genomes published by Bos et al. (2014). Following the assumption that all genomes in the *M. pinnipedii* clade evolved at a constant rate, the

branch lengths of the novel genomes that I present in paper I are in discord with this assumption. For this reason, molecular dating analyses of the MTBC was not carried out in paper I. The abnormally long branch lengths would affect the dating results by causing the age of the *M. pinnipedii* clade, and the full MTBC phylogeny, to artificially appear older, because the number of mutations to the most recent common ancestor would be higher.

Similar issues were observed for one of the *S*. Paratyphi C genomes in paper II, Tepos_10, wherein this genome harbored an abnormally high number of heterozygous positions as a result of contaminating reads compared to the other two high-coverage genomes. The level of infringement these contaminating SNPs had on our homozygous SNP calls was again reflected in the abnormally long branch length of Tepos_10.

Despite being geographically distant, both the Colombian and Peruvian *M. pinnipedii* genomes presented in paper I are affected by environmental mycobacterial contamination. This is due to the fact that the human remains were either buried directly into the soil (Colombian) or the burial tomb was disturbed post-internment exposing the remains to the natural elements (Peruvian). Based on this, it appears that soil-dwelling mycobacteria are ubiquitous across South America. However, it is worth noting that the data for the previously published Peruvian *M. pinnipedii* genomes contain much less contaminating mycobacterial DNA (Bos et al., 2014). These genomes were isolated from naturally mummified individuals buried in tombs that remained intact until excavation. Similarly, MTBC genomes isolated from individuals buried in lead coffins in Vác, Hungary, showed that the extracted DNA data was largely uncontaminated by environmental mycobacteria, allowing researchers to identify co-infections of multiple MTBC strains belonging to the human-adapted *Mycobacterium tuberculosis sensu stricto* Lineage 4 (Gemma L. Kay et al., 2015). The presence of multiple co-infecting strains of MTBC may also result in the occurrence of multiallelic positions when the sequence data is mapped, however this would not be expected to occur to the extent observed for the *M. pinnipedii* genomes or *S*. Paratyphi C genomes in papers I and II. If multiple strains of the same bacterium were present in the samples, this could result in conflicting positioning of the strains in the phylogenetic trees, manifested through low bootstrap values. The bootstrap values for all ancient genomes presented in papers I and II are high, indicating that the phylogenetic positioning is

correct and that the long branch lengths are the result of unique mutations that accumulated at the terminal branches of the tree, and do not occur (or very rarely occur) at phylogenetically significant positions. Thus, the contaminating DNA merely results in artificially long terminal branch-lengths for the affected genomes. This indicates that the burial environment plays a huge role in predicting a study's success in isolating 'clean' uncontaminated ancient MTBC data from the archaeological record.

Among the *S.* Paratyphi C samples from Teposcolula-Yucundaa in southern Mexico (paper II), only one (Tepos_10) out of five for which complete genome data could be analyzed was seemingly affected. This result perhaps indicates that environmental contamination of genetically similar bacterial species may pose less of an issue to the study of ancient *Salmonella* DNA. However, further studies from different geographic regions will be needed to make this assessment.

5.2 Preservation, authentication and capture enrichment (papers I, II, III)

After an organism has died its DNA progressively decays over time, becoming increasingly fragmented and accumulates chemical damage. These physical changes are universal to all ancient DNA extracted from archaeologically or historically preserved tissues, and can be used to assess DNA preservation and to determine if the DNA in question is truly ancient (Dabney, Meyer, et al., 2013; Key et al., 2017; Prufer et al., 2010; Sawyer et al., 2012). Due to its degraded state, ancient DNA is often preserved in low abundance and it can be difficult to retrieve adequate amounts required for in-depth genetic analyses. Additionally, the presence of environmentally derived DNA may reduce the relative abundance of ancient DNA by 'drowning out' the ancient molecules. In papers I, II and III, I use measures of abundance, fragmentation and chemical damage patterns to assess microbial and human DNA sequencing data.

The relative abundance of target DNA in a sample, in relation to all other DNA, is referred to as 'endogenous DNA', a term commonly applied to DNA originating from a host's tissue, but in the context of paper I and II it is applied to the target pathogen DNA detected in human samples. In papers I and II, I found that the target pathogen DNA detected in the non-enriched sequence data was low across all samples that yielded genome-wide data after capture. In paper I, I found that the percentage of endogenous un-enriched MTBC DNA ranged from 0.02 to 0.03% for the four samples,

while the percentage of endogenous *S.* Paratyphi C DNA in paper II ranged from 0.047 to 0.077% for the three strong positives and 0.001 to 0.008% for the seven weak positives (see Fig. 1).

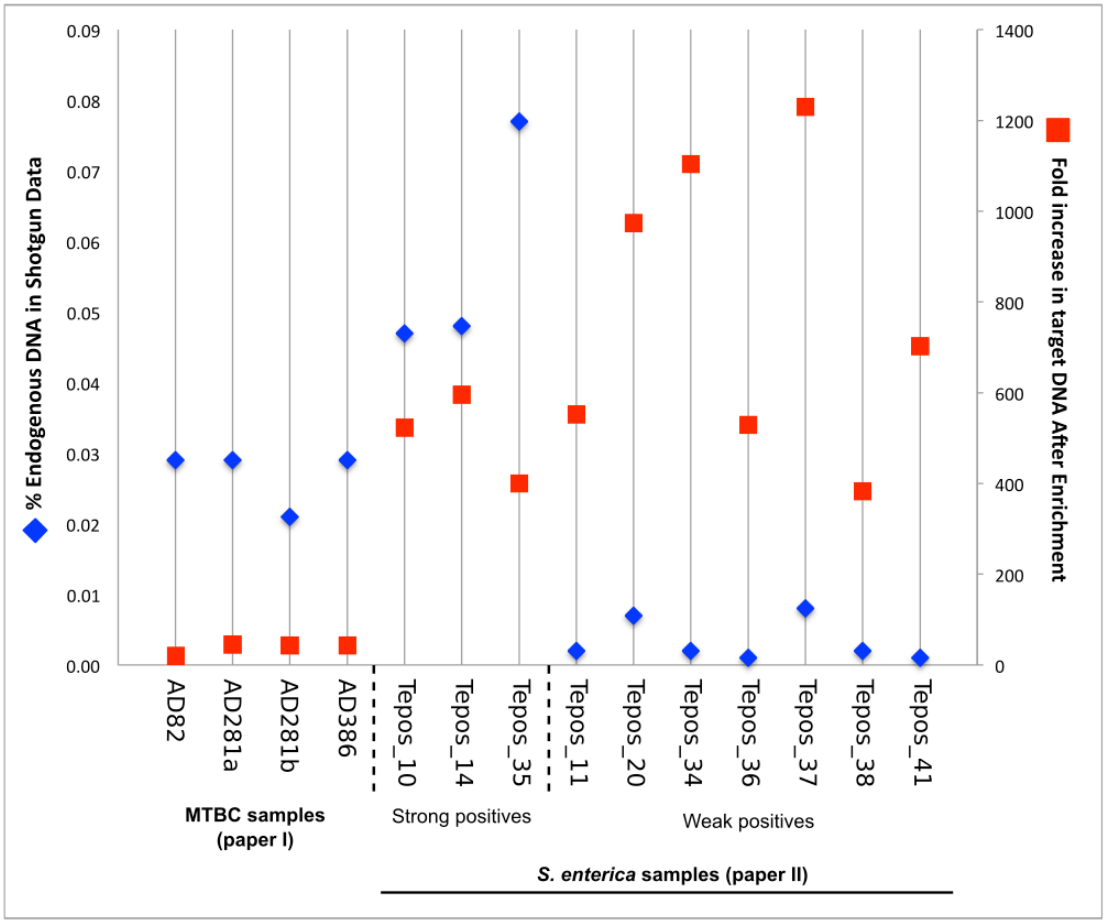


Figure 1 | Percentage of endogenous target DNA and fold-enrichment after pathogen capture for sample libraries presented in papers I and II. The endogenous DNA and fold enrichment for the MTBC samples was calculated based on UDG-treated shotgun and capture data (paper I); while for the *S. enterica* samples it was calculated based on non-UDG treated shotgun and capture data (paper II). The plot shows the percentage of endogenous pathogen DNA (blue diamond) and fold enrichment of target pathogen DNA after whole-genome in-solution capture (red square) for samples that yielded genome-wide data analyzed in papers I and II. MTBC samples AD281a and AD281b come from different skeletal elements from the same individual. The data from both samples was combined after whole-genome capture for genome reconstruction. The samples positive for *S. enterica* DNA are grouped into strong positive and weak positive samples.

In paper II the strong-positive *S. enterica* samples detected during the MALT pathogen screening drove the decision to include the seven weak-positive samples in the whole-genome capture, based on the notion that the few reads observed in the weak-positive samples might also be of ancient origin. In Fig. 1 a comparison of the percentage of endogenous DNA in non-enriched data and the fold-increase of the target DNA after in-solution capture is shown. MTBC samples enriched poorly during capture in comparison to the *S*. Paratyphi C samples. The poor enrichment efficiencies in paper I are accounted for by the abundance of DNA from genetically related environmental mycobacterial species co-enriched during capture (see section 5.1). In paper II non-specific enrichment interfered with the downstream analyses for the Tepos_10 *S*. Paratyphi C genome, however the fold-enrichment for Tepos_10 was comparatively much better at 523-fold than for any of the MTBC samples, which all had less than 46-fold enrichment. The retrieval of genome-wide data for the seven weak-positive *S*. Paratyphi C samples, containing as little as 0.001% endogenous *S. enterica* DNA (paper II) before capture, demonstrates the power of the in-solution capture technique (Fu et al., 2013).

Two of the seven *S*. Paratyphi C weak-positive samples yielded sufficient genomic coverage to allow phylogenetic positioning, while the genome-wide data for the remaining five was determined to be derived from genetically similar *S*. Paratyphi C strains based on SNP analyses. In-depth genetic analyses were not possible for any of the weak positive samples. However, this data has great value in that it substantially increased the number of positive individuals from the Grand Plaza epidemic cemetery at the Teposcolula-Yucundaa site from three to ten, out of a total of 24 samples. Thus, the argument for *S*. Paratyphi C being a candidate pathogen for the cause of the 1545 CE *cocoliztli* epidemic at the site of Teposcolula-Yucundaa is strengthened.

In comparison to epidemic mass graves previously investigated for ancient *Yersinia pestis* DNA, such as “Marktplatz” Ellwangen, Germany (1 positive out of 67 individuals tested; 1.5% positive) and Saints Màrtirs Just i Pastor, Barcelona, Spain (1 positive out of 18 individuals; 5.5% positive) (Spyrou et al., 2016), the detection of ancient *S*. Paratyphi C DNA from 41% of tested individuals from the Grand Plaza epidemic cemetery at Teposcolula-Yucundaa is extremely high. Reasons for these differences likely relate to the environment of the burial location, age of the samples

and microbial characteristics on the part of the pathogens, as well as the mode of infectivity of the pathogen and its effect on the skeletal elements sampled.

Expected fragment lengths for ancient DNA were observed in both papers I and II, where the un-enriched UDG treated paired-end MTBC data had a median fragment length of 38-50bp, that increased to 48-68bp after hybridization capture. Likewise, for the *S*. Paratyphi C DNA in paper II, the non-UDG paired-end un-enriched data for the three strong-positive samples (Tepos_10, tepos_14 and Tepos_35) had a median fragment length of 43-53bp, that increased to 50-63bp after capture. Target DNA fragment lengths are known to increase after enrichment due to the preferential capture of longer fragments, a known bias of hybridization capture techniques (Spyrou et al., 2018). A downside to this phenomenon is that modern contaminating DNA tends to have longer fragment lengths and will therefore be favored over shorter ancient DNA fragments during capture.

In papers I and II, I compared the damage patterns between the endogenous ancient pathogen and corresponding ancient human DNA in order to gauge if there were differences in the rates of chemical degradation. In paper II, the damage pattern for the non-enriched *S*. Paratyphi C DNA (~17-23%) was determined to be similar to that of the corresponding human DNA (~14-28%). However, in paper I, the damage pattern for the enriched MTBC DNA (~7-8.7%) was lower in comparison to the human DNA (~14-14.75%), despite my attempt to computationally exclude mis-mapping contaminant mycobacterial reads. Although the persistence of mis-mapping reads could influence the low MTBC DNA damage patterns, it is known that the lipid rich mycobacterial cell wall affords some protection to the DNA against degrading environmental factors, as has previously been observed for ancient *Mycobacterium leprae* DNA (Schuenemann et al., 2013).

The human DNA analyzed in papers I and II was generated during shotgun sequencing or as an off-target co-enriched by-product of the whole-genome pathogen captures. Negative results were reported for ancient pathogen DNA for the individual from Moneen Cave, Ireland (paper III). However, the un-enriched sequence data contained 8.2% endogenous human DNA, which was enough to allow the reconstruction of this individual's mitochondrial genome. Comparatively, the human

endogenous DNA reported for the 28 individuals investigated from the two cemetery sites at Teposcolula-Yucundaa was much lower, ranging from 0.01 to 2.1%, with one sample having 26.3%. Warm temperatures, temperature fluctuations, increased age of the samples and humidity of the burial environment have all been linked to the increased degradation of ancient DNA (Kistler et al., 2017; Smith et al., 2001; Willerslev & Cooper, 2005). In terms of sample age, the Moneen Cave and Teposcolula-Yucundaa individuals were roughly contemporaneous. Therefore, the warmer climate of Teposcolula-Yucundaa, Mexico may have contributed to the overall poor preservation of human DNA at this site.

5.3 Pathogen Screening (papers I, II and III)

Both targeted molecular and broad-scale computational approaches were used to screen for the presence of ancient pathogen DNA in papers I, II and III.

In paper I, a two-tiered system of targeted molecular techniques, consisting of a qPCR assay followed by gene capture, was used to determine which samples were positive for MTBC DNA. Like all targeted screening methods, this approach was contingent on prior knowledge of which pathogenic organism to target. In paper I this knowledge was derived from skeletal markers of pathology consistent with prolonged MTBC infection, therefore only samples from individuals carrying such lesions were screened. The DNA extracts positive for MTBC via qPCR were re-screened using gene capture for the purpose of excluding false positive samples from the whole-genome capture, a costly and time intensive experiment. Comparatively, the qPCR assay targeted three mycobacterial regions less than 100 bp in length, while the gene capture assay targeted five mycobacterial genes totaling several thousand bp in length. Additionally, sequencing of the gene capture product facilitated a thorough evaluation of the DNA itself, allowing standard measures of ancient DNA authenticity to be applied (see section 5.2). Conversely, the sequence of the qPCR-amplified product remains unknown and there is no way to authenticate the amplified fragments as ancient MTBC DNA versus contaminant DNA (Harkins et al., 2015; Muller, Roberts, & Brown, 2016), making this method more prone to false-positive results. Despite these limitations, qPCR has been applied to several ancient DNA studies for the purpose of pathogen screening, as it is a relatively cheap and fast way of screening a large number of extracts (Feldman et al., 2016; Harkins et al., 2015; Schuenemann et al., 2011;

Spyrou et al., 2016). This was the underlying reason for not directly applying gene capture to all samples, as was the approach adopted by Bos et al. (2014).

Another major drawback to the use of targeted methods, whether molecular or computational, is the inability to simultaneously screen for other pathogenic or commensal microbes of interest. This limitation exists because targeted methods are driven by hypotheses generated on the basis of skeletal pathology, historical evidence and/or archaeological context. Such hypothesis-driven research fixates on one pathogen from the start and does not leave room for exploration of other pathogens or co-infections caused by other pathogens. The MEGAN ALignment Tool (MALT) (published in paper II) represents a broad-scale screening approach, developed for the purpose of screening for ancient microbial DNA against complex metagenomic backgrounds. MALT affords a high degree of flexibility to researchers in that it can be used with any customized database. Broad-scale approaches such as this facilitate a new type of 'opportunistic' research within the field of ancient pathogen genomics, where hypotheses about organismal, biological and historical research questions are formed based on the screening results. Broad-scale screening approaches such as MALT represent a game changing technological advance for the field of ancient pathogen genomics, providing opportunities for researchers to access information about microbial evolution and the disease histories of past human populations that was previously unattainable, as illustrated in paper II.

The power and sensitivity of MALT is demonstrated across all three papers. However, like all other broad-scale screening techniques it is subject to biases inherent to database selection or probe design. It is therefore recommended that all screening results be supported by enough genetic data to verify the findings as described by Prufer et al. (2010), Key et al. (2017) and Warinner et al. (2017).

Even when computational methods, such as MALT, use databases consisting of all known genomes, therein lies a limitation in that only organisms that have been genetically characterized can be screened for. Another bias exists wherein molecular protocols used to prepare DNA libraries eliminate molecular traces of RNA pathogens that may have been present. Different molecular protocols are required for the preparation of RNA libraries, and the protocols for ancient RNA are comparatively

underdeveloped to those for DNA (Fordyce et al., 2013; Guy, 2014). Therefore, the presence of RNA viruses was not investigated in papers II and III. Today RNA viruses represent a large number of clinically relevant human pathogens, and future research investigating the potential for recovering ancient virus RNA would be of great relevance to our understanding of past human infectious diseases.

5.4 Archaeological and historical significance of findings (papers I and II)

The ancient pathogen genomes that I present in papers I and II increase our understanding of the pathogenic burden that afflicted New World indigenous human populations pre- and post- 16th century European contact. These genomes are direct evidence for the human-pathogen relationship through time, and provide information that is otherwise inaccessible to researchers studying human infectious disease through historical and osteological evidence alone.

Paper I is a direct continuation of the study published by Bos et al. (2014), which presents MTBC genomes recovered from the skeletal remains of three ancient Peruvian humans dated to 1028-1280 CE, all coming from the same narrow geographic region in southern Peru. The three new MTBC genomes presented in paper I, which I helped generate and conducted analysis for, provide continued molecular support for the longstanding archaeological observation of pre-contact human TB infections in the New World (Allison et al., 1981; Roberts & Buikstra, 2003). The concept of pre-contact human TB was previously difficult to reconcile with the observation that the New World is dominated by Lineage 4 strains common to Europe today (Roberts & Buikstra, 2003), which were introduced to the New World during European contact (O'Neill et al., 2018; Pepperell et al., 2011). The findings that I present in paper I are also geographically significant, because they extend the molecular support for *M. pinnipedii* beyond the Osmore River drainage in Peru to include two inland Colombian sites (paper I) (see Fig. 2).

As previously described, the pathogenic cause of the 1545-1550 CE '*cocoliztli*' epidemic, which affected an estimated 60-90% of Mexico and Guatemala's indigenous population (Cook & Lovell, 2001), has been debated for over a century and remains unknown based solely on historical and archaeological evidence alone (paper II). In paper II, I analyzed DNA extracted from human remains from the only epidemic

cemetery, to date, that has been historically linked to the 1545 CE '*cocoliztli'* epidemic. The uniqueness of the Teposcolula-Yucundaa cemetery reflects a disparity of the substantial number of indigenous individuals estimated to have died during the 1545 CE epidemic and the apparent lack of human remains discovered in the archaeological record. MALT (paper II), allowed me to screen these human remains in a way that made full use of the sequencing data generated from them by screening for all complete bacterial and DNA virus genomes. The public availability of this data (as well as that generated in paper I upon publication), means that it can continue to be investigated by researchers making use of the continually growing genome databases.

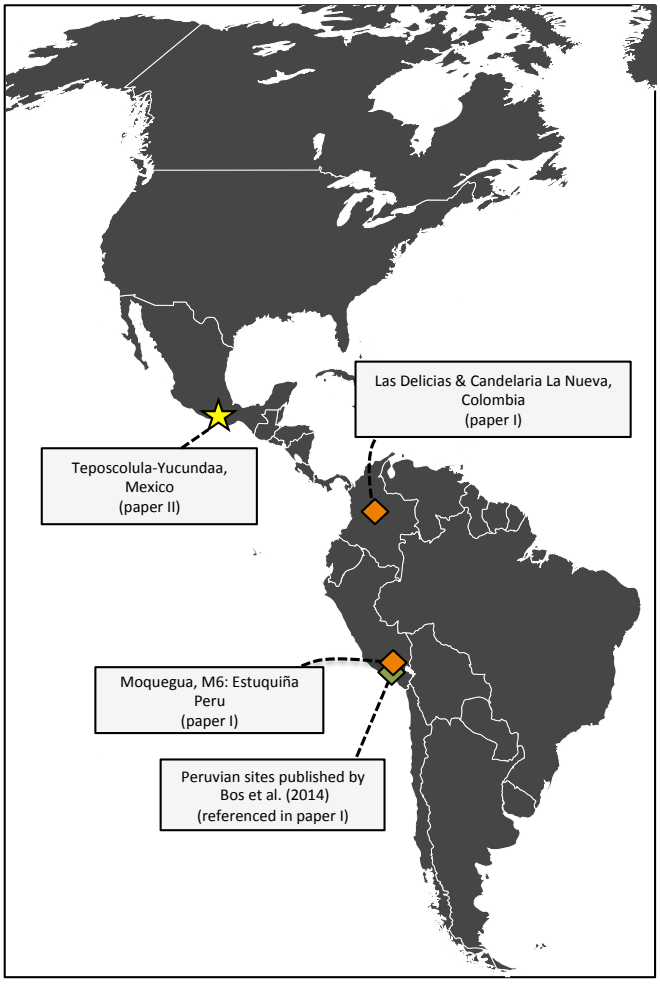


Figure 2 | Map of New World sample sites from papers I and II.

The findings of paper II represent the first direct evidence for the involvement of any pathogen in the 1545 CE epidemic, however this finding is only applicable to the site of Teposcolula-Yucundaa (Fig. 2). In order to investigate the geographic range of *S.* Paratyphi C pathogen during the 1545 CE epidemic it would first require the discovery and excavation of further burial sites associated with this epidemic.

The significance of paper II not only lies in the identification of a candidate pathogen for the epidemic at Teposcolula-Yucundaa, but also in the way that it was detected. The use of MALT circumvented the need to specify a target organism based on prior information. This study thus demonstrates the possibilities of what can be accomplished now that it is possible to traverse this barrier, thus, significantly increasing the number of past pathogens that can be accessed and studied using ancient DNA.

5.5 Transmission, spread and macroecology (papers I and II)

M. pinnipedii and *S*. Paratyphi C are both intracellular facultative bacterial pathogens that have the potential to cause persistent infections in the human host, if the initial infection is not cleared by the immune system (Monack et al., 2004). This is an important characteristic to consider when contextualizing the transmission and spread of pathogens, as persistent infections allow pathogens to survive within restricted host populations.

Members of the MTBC are able to cause two types of persistent infections: 1) latent pulmonary (lung) infections where bacteria survive long-term inside macrophages within granulomas, with the potential to become reactivated and cause active disease later on; 2) disseminated infections to extrapulmonary organs (e.g. bones and joints, lymph nodes, meninges and pleura) that are never completely cleared by the immune system, leading individuals to become chronic carriers of active infection for many years, even throughout their lifetime (Lee, 2015; Monack et al., 2004). Currently one quarter of the world's human population is estimated to carry a latent pulmonary MTBC infection (Houben & Dodd, 2016).

In 1 to 6% of cases of human infection with *S.* Typhi and *S*. Paratyphi A, *Salmonella enterica* strain-types that like *S*. Paratyphi C cause enteric fever, individuals

contract a symptomless persistent infection. The infection can persist in the gall bladder, gastrointestinal tract and bone marrow, during which time bacteria are actively shed through fecal matter for months or years after the initial infection (Gunn et al., 2014; Monack et al., 2004). Persistent infections of *S.* Paratyphi C have yet to be systematically studied due to the rarity of this strain today, but it is thought that it also has the capability to cause prolonged symptomless infections (Gunn et al., 2014).

In paper II it is suggested that indigenous populations in the New World were first introduced to this pathogen after the arrival of 16th century European colonizers. Chronic carriers of S. Paratyphi C voyaging from Europe may have been able to survive the cross-Atlantic transit with the infection intact. This hypothesis is supported by an *S.* Paratyphi C genome from 12th century Norway (Zhou et al., 2018), demonstrating that the pathogen was indeed present in Europe prior to the 1545-1550 CE '*cocoliztli'* epidemic. Further research on the diversity of European and New World *S.* Paratyphi C strains will be needed in order to confidently rule out its presence in the New World prior to European contact. Based on the findings from paper II, it can only be asserted that *S.* Paratyphi C was present at Teposcolula-Yucundaa during the 1545 CE epidemic. Currently, we have no knowledge of whether or not it was infecting individuals at other sites. However, if *S.* Paratyphi C was widespread, latent human infections may have been an essential factor in this pathogen's spread and transmission via the human host.

The four Peruvian *M. pinnipedii* genomes, two of which I aided in generating (paper I), are interpreted to have stemmed from an ancient zoonotic event linked to the manipulation and consumption of infected pinniped tissues. The two *M. pinnipedii* genomes recovered from inland Colombian human remains were likely transmitted via human-to-human and/or via terrestrial animal(s)-to-human contact (paper I). Latent MTBC infections are known to occur in humans and to some extent in animals (Patel, Jhamb, & Singh, 2011), and may have been an important part of transporting and establishing the infection at the inland Colombian sites (paper I). Colombia is geographically distant from Peru and the finding of *M. pinnipedii* in these two locations suggests that *M. pinnipedii* may have been widespread across South America in the past. Today, *M. pinnipedii* is known to affect non-pinniped animal hosts as well as humans, both in wild and captive settings (Cousins et al., 2003; Jurczynski et al., 2011; Albert Kiers, Klarenbeek, Mendelts, Van Soolingen, & Koëter, 2008). However, no

wild animal populations other than pinnipeds are currently known to maintain within-species transmission of *M. pinnipedii*. A pathogen's ability to infect a wide host-range also increases the geographic range at which it can travel. Currently the data is too sparse to indicate transmission routes that *M. pinnipedii* might have taken, abetted by the fact that all ancient strains are currently derived from humans (Bos et al., 2014; paper I). Furthermore, it remains to be seen which lineages were causing human MTBC infection in pre-contact North and Central America. Future ancient DNA research will be fundamental to our understanding pre-contact MTBC in these New World regions.

The above interpretation of how *M. pinnipedii* was introduced to and spread across the New World is contingent on the estimated date of MTBC emergence. Molecular dating methods incorporating ancient MTBC genomes corroborate a MTBC emergence date of approximately 6,000 YBP (Bos et al., 2014; Gemma L. Kay et al., 2015). Currently, this seems to be the best supported date (O'Neill et al., 2018), although other studies arrive at older dates of MTBC emergence (Comas et al., 2013; Refregier et al., 2016). A 6,000 YBP date of emergence precludes the spread of MTBC strains to the New World via Pleistocene human migrations across the Bering Strait more than 11,000 YBP. Bos et al. (2014) dates the split of the *M. pinnipedii* and *M. microti* lineages to between 1,439-2,510 YBP (511 CE - 560 BCE). The earliest appearance of skeletal lesions consistent with prolonged tuberculosis infections (occurring in multiple individuals at one site) in the New World appears in 281 CE (Allison et al., 1981). Thus, the molecular and archaeological dates do not contradict each other, nor the hypothesis that pinnipeds introduced *M. pinnipedii* to the New World. It is within this framework that the results of paper I were discussed. If future inclusion of a wider diversity of ancient MTBC strains changes the date of MTBC emergence significantly, it may call for a re-interpretation of how *M. pinnipedii* was introduced to, and spread across, the New World.

5.6 Contrasting past and present prevalence rates and disease manifestations (papers I and II)

Tuberculosis caused by *M. pinnipedii*, and enteric fever caused by *S.* Paratyphi C, are rarely reported in cases of human infections today (Cousins et al., 2003; Albert Kiers et al., 2008; Liu et al., 2009; Wain, Hendriksen, Mikoleit, Keddy, & Ochiai, 2015). However, other strain members belonging to the MTBC and *S. enterica* spp.

enterica cause a high number of human infections, which respectively account for an estimated 1.7 million (World Health Organization, 2017) and 200,000 deaths annually (Buckle, Walker, & Black, 2012; Crump, Luby, & Mintz, 2004). Because *M. pinnipedii* and *S*. Paratyphi C are rarely implicated in modern human infections, it was somewhat unexpected to recover these pathogen strain types from archaeological human remains. Hence, the results I present in this dissertation provide new, and added, insight into the prevalence and capability of these pathogens to cause human infection in the past (papers I and II).

On the rare occasions that *M. pinnipedii* is associated with modern human infection it has occurred as a result of prolonged contact with infected animals held in captivity. In these interactions, humans are the spillover host and are not known to transfer the pathogen to other humans. To date, human-to-human transfer of an animal associated MTBC strain has only been observed for *M. bovis* (Gonzalo-Asensio et al., 2014; Rivero et al., 2001). Therefore, the discovery of *M. pinnipedii* in archaeological human remains from Peru and Colombia was unexpected (Bos et al., 2014). The interpretation of the *M. pinnipedii* transmission chain in the pre-contact New World is uncertain regarding whether human-to-human transmission occurred or whether other animal hosts were involved in the transmission of this pathogen to both the inland Colombian sites, as well as to the coastal highland Peruvian site (Moquegua, Estuquiña: M6) (paper I). Although, Moquegua, Estuquiña: M6 is located ~67 km from the coast, it sits at 1000m above sea level (paper I), making direct human-to-pinniped contact less likely than for the other coastal Peruvian sites situated in the low-lands of the Osmore River valley (Bos et al., 2014). Transmission chains of animal-associated MTBC are highly difficult to elucidate based on modern data (Allen, 2017; Nugent, 2011) and will be impossible to define in the past, although certain animal hosts may be identified through the analysis of ancient animal remains. Additionally, little is known about the underlying molecular mechanisms that determine host range (Gagneux, 2018).

Regardless of the transmission chain, these data illustrate that *M. pinnipedii* was geographically widespread in the past and may have been a significant strain-type affecting indigenous human South American populations. At Moquegua, Estuquiña: M6 9% of the total population, and 15% of adult remains (for which sex could be determined) exhibit skeletal lesions consistent with prolonged extrapulmonary MTBC

infection (Buikstra & Williams, 1991). Such prevalence rates are consistent with an epidemic-scale outbreak of chronic tuberculosis. Thus, the *M. pinnipedii* genome (AD82) generated from this population, may demonstrate the past potential of this pathogen to contribute to large-scale human outbreaks in the past (assuming that multiple individuals were affected by this same strain-type), even though such behavior is not observed for the pathogen today (paper I).

Four human-specific *S. enterica* ssp. *enterica* strains cause enteric fever: *S.* Typhi and *S*. Paratyphi A, B and C. Of these, *S.* Paratyphi C is rarely identified as the cause of human infection today (Liu et al., 2009; Wain et al., 2015). By detecting *S.* Paratyphi C in 10 out of the 24 individuals, I provide strong indication for its involvement as at least one of the pathogenic agents, of the 1545-1550 *cocoliztli* epidemic at Teposcolula-Yucundaa (paper II). These results illustrate the capacity of *S.* Paratyphi C to contribute to an epidemic-scale outbreak ~450 years ago. The presence of *S.* Paratyphi C at Teposcolula-Yucundaa during the 16th century (paper II) and 12th century Trondheim, Norway (Zhou et al., 2018) may indicate that this pathogen was more prevalent in the past, and that modern prevalence rates may not be comparable.

6. Concluding Remarks

The addition of ancient pathogen genomes to modern datasets yields unparalleled insight into pathogen evolution over time, whereby time-stamped ancient genomes can be used to address research questions relating to the 'where, how and when' of geographic distribution, host adaptation and emergence. Papers I and II presented in this dissertation extend the current knowledge base relating to pre-contact MTBC in South America, provide the first molecular evidence for the unknown cause of a post-contact epidemic in Mexico, and illustrate the efficiency and application of a novel computational tool (MALT) for the screening and detection of ancient pathogens from complex metagenomic sequence data (paper II). These findings form a foundation for future studies concerning the human infectious disease burden of indigenous New World populations to build upon; as well as providing valuable information for the continued exploration of the evolutionary histories of ancient MTBC and *S.* Paratyphi C. Sampling of geographically and temporally diverse ancient and modern genomes in future studies will be essential to elucidating the evolutionary histories of these pathogens. In regards to the intrinsic challenges faced by all ancient pathogen studies, computational tools such as MALT (paper II) will be instrumental in driving the field of 'ancient pathogen genomics' forward by facilitating the detection and study of the multitude of pathogens whose presence in the archaeological record cannot be gleaned from osteological and historical evidence alone.

References

Achilli, A., Iommarini, L., Olivieri, A., Pala, M., Hooshiar Kashani, B., Reynier, P., . . . Carelli, V. (2012). Rare primary mitochondrial DNA mutations and probable synergistic variants in Leber's hereditary optic neuropathy. *PLoS One, 7*(8), e42242. doi:10.1371/journal.pone.0042242

Achtman, M. (2016). How old are bacterial pathogens? *Proc Biol Sci, 283*(1836). doi:10.1098/rspb.2016.0990

Achtman, M., Zhou, Z., & Didelot, X. (2015). Formal Comment to Pettengill: The Time to Most Recent Common Ancestor Does Not (Usually) Approximate the Date of Divergence. *PLoS One, 10*(8), e0134435. doi:10.1371/journal.pone.0134435

Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K., & Therrell, M. D. (2002). Megadrought and megadeath in 16th century Mexico. *Emerging Infectious Diseases, 8*(4), 360-362.

Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Gomez Chavez, S., & Cleaveland, M. K. (2005). Drought, epidemic disease, and the fall of classic period cultures in Mesoamerica (AD 750-950). Hemorrhagic fevers as a cause of massive population loss. *Med Hypotheses, 65*(2), 405-409. doi:10.1016/j.mehy.2005.02.025

Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Griffin, R. D., & Cleaveland, M. K. (2004). When half of the population died: the epidemic of hemorrhagic fevers of 1576 in Mexico. *FEMS Microbiol Lett, 240*(1), 1-5. doi:10.1016/j.femsle.2004.09.011

Alikhan, N. F., Zhou, Z., Sergeant, M. J., & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. *PLoS Genet, 14*(4), e1007261. doi:10.1371/journal.pgen.1007261

Allen, A. R. (2017). One bacillus to rule them all? - Investigating broad range host adaptation in Mycobacterium bovis. *Infect Genet Evol, 53*, 68-76. doi:10.1016/j.meegid.2017.04.018

Allison, M. J., Gerszten, E., Munizaga, J., Santoro, C., & Mendoza, D. (1981). Tuberculosis in Pre-Columbian Andean Populations. In J. E. Buikstra (Ed.), *Prehistoric Tuberculosis in the Americas* (pp. 49-61). Evanston, Illinois: Northwestern University Archaeological Program.

Andam, C. P., Worby, C. J., Chang, Q., & Campana, M. G. (2016). Microbial Genomics of Ancient Plagues and Outbreaks. *Trends Microbiol, 24*(12), 978-990. doi:10.1016/j.tim.2016.08.004

Andrades Valtuena, A., Mittnik, A., Key, F. M., Haak, W., Allmae, R., Belinskij, A., . . . Krause, J. (2017). The Stone Age Plague and Its Persistence in Eurasia. *Curr Biol, 27*(23), 3683-3691 e3688. doi:10.1016/j.cub.2017.10.025

Arriaza, B. T., Salo, W., Aufderheide, A. C., & Holcomb, T. A. (1995). Pre-Columbian tuberculosis in northern Chile: molecular and skeletal evidence. *Am J Phys Anthropol, 98*(1), 37-45. doi:10.1002/ajpa.1330980104

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

Black, F. L. (1992). Why did they die? *Science, 258*(5089), 1739-1740.

Bliven, K. A., & Maurelli, A. T. (2016). Evolution of Bacterial Pathogens Within the Human Host. *Microbiol Spectr, 4*(1). doi:10.1128/microbiolspec.VMBF-0017-2015

Boardman, W. S., Shephard, L., Bastian, I., Globan, M., Fyfe, J. A., Cousins, D. V., . . . Woolford, L. (2014). Mycobacterium pinnipedii tuberculosis in a free-ranging Australian fur seal (Arctocephalus pusillus doriferus) in South Australia. *J Zoo Wildl Med, 45*(4), 970-972. doi:10.1638/2014-0054.1

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., . . . Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature, 514*(7523), 494-497. doi:10.1038/nature13591

Bos, K. I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S. A., . . . Poinar, H. N. (2016). Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *Elife, 5*, e12994. doi:10.7554/eLife.12994

Bos, K. I., Jager, G., Schuenemann, V. J., Vagene, A. J., Spyrou, M. A., Herbig, A., . . . Krause, J. (2015). Parallel detection of ancient pathogens via array-based DNA capture. *Philos Trans R Soc Lond B Biol Sci, 370*(1660), 20130375. doi:10.1098/rstb.2013.0375

Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., . . . Krause, J. (2011). A draft genome of Yersinia pestis from victims of the Black Death. *Nature, 478*(7370), 506-510. doi:10.1038/nature10549

Braun, M., Collins Cook, D., & Pfeiffer, S. (1998). DNA fromMycobacterium tuberculosisComplex Identified in North American, Pre-Columbian Human Skeletal Remains. *Journal of Archaeological Science, 25*(3), 271-277. doi:https://doi.org/10.1006/jasc.1997.0240

Briggs, A., Stenzel, U., & Johnson, P. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 14616-14621.

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Paabo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res, 38*(6), e87. doi:10.1093/nar/gkp1163

Brosch, R., Gordon, S. V., Pym, A., Eiglmeier, K., Garnier, T., & Cole, S. T. (2000). Comparative genomics of the mycobacteria. *Int J Med Microbiol, 290*(2), 143-152. doi:10.1016/S1438-4221(00)80083-1

Buckle, G. C., Walker, C. L., & Black, R. E. (2012). Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. *J Glob Health, 2*(1), 010401. doi:10.7189/jogh.02.010401

Buikstra, J. E., & Williams, S. R. (1991). Tuberculosis in the Americas: Current perspectives. In D. Ortner & A. C. Aufderheide (Eds.), *Human palaeopathology: Current syntheses and future options* (pp. 161-172). Washington, D.C.: Smithsonian Institution Press.

Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., . . . Pääbo, S. (2010). Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science, 328*(5979), 723-725. doi:10.1126/science.1188046

Cohen, M. L. (2000). Changing patterns of infectious disease. *Nature, 406*(6797), 762-767. doi:10.1038/35021206

Comas, I., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., . . . Gagneux, S. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet, 42*(6), 498-503. doi:10.1038/ng.590

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., . . . Gagneux, S. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nature Genetics, 45*(10), 1176-U1311. doi:10.1038/ng.2744

Cook, N. D. (1998). *Born to Die: Disease and New World Conquest, 1492-1650*: Cambridge University Press.

Cook, N. D., & Lovell, W. G. (2001). *Secret Judgments of God: Old World Disease in Colonial Spanish America*: University of Oklahoma Press.

Cousins, D. V., Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., . . . Bernardelli, A. (2003). Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *Int J Syst Evol Microbiol, 53*(Pt 5), 1305-1314. doi:10.1099/ijs.0.02401-0

Crosby, A. W. (1976). Virgin soil epidemics as a factor in the aboriginal depopulation in America. *William Mary Q, 33*, 289-299.

Crosby, A. W. (2003). *The Columbian exchange: biological and cultural consequences of 1492* (Vol. 2). United States of America: Greenwood Publishing Group.

Crump, J. A., Luby, S. P., & Mintz, E. D. (2004). The global burden of typhoid fever. *Bull World Health Organ, 82*(5), 346-353.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., . . . Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A, 110*(39), 15758-15763. doi:10.1073/pnas.1314445110

Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques, 52*(2), 87-94. doi:10.2144/000113809

Dabney, J., Meyer, M., & Paabo, S. (2013). Ancient DNA damage. *Cold Spring Harb Perspect Biol, 5*(7). doi:10.1101/cshperspect.a012567

de Amorim, D. B., Casagrande, R. A., Alievi, M. M., Wouters, F., De Oliveira, L. G., Driemeier, D., . . . Ferreira-Neto, J. S. (2014). Mycobacterium pinnipedii in a stranded South American sea lion (Otaria byronia) in Brazil. *J Wildl Dis, 50*(2), 419-422. doi:10.7589/2013-05-124

Devault, A. M., McLoughlin, K., Jaing, C., Gardner, S., Porter, T. M., Enk, J. M., . . . Poinar, H. N. (2014). Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. *Sci Rep, 4*, 4245. doi:10.1038/srep04245

Devault, A. M., Mortimer, T. D., Kitchen, A., Kiesewetter, H., Enk, J. M., Golding, G. B., . . . Pepperell, C. S. (2017). A molecular portrait of maternal sepsis from Byzantine Troy. *Elife, 6*. doi:10.7554/eLife.20983

Dobson, A. P., & Carper, E. R. (1996). Infectious Diseases and Human Population HistoryThroughout history the establishment of disease has been a side effect of the growth of civilization. *BioScience, 46*(2), 115-126. doi:10.2307/1312814

Dobyns, H. F. (1993). Disease Transfer at Contact. *Annual Review of Anthropology, 22*, 273-291.

Donoghue, H. D., Marcsik, A., Matheson, C., Vernon, K., Nuorala, E., Molto, J. E., . . . Spigelman, M. (2005). Co-infection of Mycobacterium tuberculosis and Mycobacterium leprae in human archaeological samples: a possible explanation for the historical decline of leprosy. *Proceedings of the Royal*

Society B: Biological Sciences, 272(1561), 389-394. doi:10.1098/rspb.2004.2966
Drake, A., & Oxenham, M. (2013). Disease, climate and the peopling of the Americas. *Historical Biology, 25*(5-6), 565-597. doi:10.1080/08912963.2012.725728
Dye, C. (2014). After 2015: infectious diseases in a new era of health and development. *Philos Trans R Soc Lond B Biol Sci, 369*(1645), 20130426. doi:10.1098/rstb.2013.0426
Dyke, A. S. (2004). An outline of North American deglaciation with emphasis on central and northern Canada. In J. Ehlers & P. L. Gibbard (Eds.), *Developments in Quaternary Sciences* (Vol. 2, pp. 373-424): Elsevier.
Fagundes, N. J. R., Tagliani-Ribeiro, A., Rubicz, R., Tarskaia, L., Crawford, M. H., Salzano, F. M., & Bonatto, S. L. (2018). How strong was the bottleneck associated to the peopling of the Americas? New insights from multilocus sequence data. *Genet Mol Biol, 41*(1 suppl 1), 206-214. doi:10.1590/1678-4685-GMB-2017-0087
Feldman, M., Harbeck, M., Keller, M., Spyrou, M. A., Rott, A., Trautmann, B., . . . Krause, J. (2016). A High-Coverage Yersinia pestis Genome from a Sixth-Century Justinianic Plague Victim. *Mol Biol Evol, 33*(11), 2911-2923. doi:10.1093/molbev/msw170
Fields, S. L. (2008). *Pestilence and Headcolds: Encountering Illness in Colonial Mexico*: Columbia University Press.
Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C., & Rambaut, A. (2010). Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol, 27*(9), 2038-2051. doi:10.1093/molbev/msq088
Fordyce, S. L., Kampmann, M. L., van Doorn, N. L., & Gilbert, M. T. (2013). Long-term RNA persistence in postmortem contexts. *Investig Genet, 4*(1), 7. doi:10.1186/2041-2223-4-7
Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., & Paabo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A, 110*(6), 2223-2227. doi:10.1073/pnas.1221359110
Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet, 7*(11), e1002355. doi:10.1371/journal.pgen.1002355
Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol*. doi:10.1038/nrmicro.2018.8
Garattini, C. (2007). Creating memories: material culture and infantile death in contemporary Ireland. *Mortality, 12*(2), 193-206. doi:10.1080/13576270701255172
Gilbert, M. T., Cuccui, J., White, W., Lynnerup, N., Titball, R. W., Cooper, A., & Prentice, M. B. (2004). Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology, 150*(Pt 2), 341-354. doi:10.1099/mic.0.26594-0
Gilbert, M. T., Jenkins, D. L., Gotherstrom, A., Naveran, N., Sanchez, J. J., Hofreiter, M., . . . Willerslev, E. (2008). DNA from pre-Clovis human coprolites in Oregon, North America. *Science, 320*(5877), 786-789. doi:10.1126/science.1154116

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

Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., . . . Guilhot, C. (2014). Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A, 111*(31), 11491-11496. doi:10.1073/pnas.1406693111

Green, E. J., & Speller, C. F. (2017). Novel Substrates as Sources of Ancient DNA: Prospects and Hurdles. *Genes (Basel), 8*(7). doi:10.3390/genes8070180

Green, R. E., Briggs, A. W., Krause, J., Prüfer, K., Burbano, H. a., Siebauer, M., . . . Pääbo, S. (2009). The Neandertal genome and ancient DNA authenticity. *The EMBO journal, 28*, 2494-2502. doi:10.1038/emboj.2009.222

Gunn, J. S., Marshall, J. M., Baker, S., Dongol, S., Charles, R. C., & Ryan, E. T. (2014). Salmonella chronic carriage: epidemiology, diagnosis, and gallbladder persistence. *Trends Microbiol, 22*(11), 648-655. doi:10.1016/j.tim.2014.06.007

Guy, P. L. (2014). Prospects for analyzing ancient RNA in preserved materials. *Wiley Interdiscip Rev RNA, 5*(1), 87-94. doi:10.1002/wrna.1199

Harkins, K. M., Buikstra, J. E., Campbell, T., Bos, K. I., Johnson, E. D., Krause, J., & Stone, A. C. (2015). Screening ancient tuberculosis with qPCR: challenges and opportunities. *Philos Trans R Soc Lond B Biol Sci, 370*(1660), 20130622. doi:10.1098/rstb.2013.0622

Harkins, K. M., & Stone, A. C. (2015). Ancient pathogen genomics: insights into timing and adaptation. *J Hum Evol, 79*, 137-149. doi:10.1016/j.jhevol.2014.11.002

Harper, K. N., & Armelagos, G. J. (2013). Genomics, the origins of agriculture, and our changing microbe-scape: time to revisit some old tales and tell some new ones. *Am J Phys Anthropol, 152 Suppl 57*, 135-152. doi:10.1002/ajpa.22396

Heyn, P., Stenzel, U., Briggs, A. W., Kircher, M., Hofreiter, M., & Meyer, M. (2010). Road blocks on paleogenomes--polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Res, 38*(16), e161. doi:10.1093/nar/gkq572

Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., . . . Hannon, G. J. (2009). Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc, 4*(6), 960-974. doi:10.1038/nprot.2009.68

Houben, R. M., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med, 13*(10), e1002152. doi:10.1371/journal.pmed.1002152

Immel, A., Drucker, D. G., Bonazzi, M., Jahnke, T. K., Munzel, S. C., Schuenemann, V. J., . . . Krause, J. (2015). Mitochondrial Genomes of Giant Deers Suggest their Late Survival in Central Europe. *Sci Rep, 5*, 10853. doi:10.1038/srep10853

Institute of Medicine (US) Forum on Microbial Threats. (2006). *The Impact of Globalization on Infectious Disease Emergence and Control: Exploring the Consequences and Opportunities: Workshop Summary*. Washington (DC).

Jakobsson, M., Pearce, C., Cronin, T. M., Backman, J., Anderson, L. G., Barrientos, N., . . . O'Regan, M. (2017). Post-glacial flooding of the Bering Land Bridge dated to 11cal ka BP based on new geophysical and sediment records. *Clim. Past, 13*(8). doi:10.5194/cp-13-991-2017

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature, 451*(7181), 990-993. doi:10.1038/nature06536

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics, 29*(13), 1682-1684. doi:10.1093/bioinformatics/btt193

Joralemon, D. (1982). New World Depopulation and the Case of Disease. *Journal of Anthropological Research, 38*(1), 108-127.

Jurczynski, K., Lyashchenko, K. P., Gomis, D., Moser, I., Greenwald, R., & Moisson, P. (2011). Pinniped tuberculosis in Malayan tapirs (Tapirus indicus) and its transmission to other terrestrial mammals. *J Zoo Wildl Med, 42*(2), 222-227. doi:10.1638/2009-0207.1

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nat Rev Genet, 15*(6), 379-393. doi:10.1038/nrg3734

Kay, G. L., Sergeant, M. J., Giuffra, V., Bandiera, P., Milanese, M., Bramanti, B., . . . Pallen, M. J. (2014). Recovery of a medieval Brucella melitensis genome using shotgun metagenomics. *MBio, 5*(4), e01337-01314. doi:10.1128/mBio.01337-14

Kay, G. L., Sergeant, M. J., Zhou, Z., Chan, J. Z.-M., Millard, A., Quick, J., . . . Pallen, M. J. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature communications, 6*, 6717. doi:10.1038/ncomms7717

Key, F. M., Posth, C., Krause, J., Herbig, A., & Bos, K. I. (2017). Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet, 33*(8), 508-520. doi:10.1016/j.tig.2017.05.005

Kiers, A., Klarenbeek, A., Mendelts, B., Van Soolingen, D., & Koeter, G. (2008). Transmission of Mycobacterium pinnipedii to humans in a zoo with marine mammals. *Int J Tuberc Lung Dis, 12*(12), 1469-1473.

Kiers, A., Klarenbeek, A., Mendelts, B., Van Soolingen, D., & Koëter, G. (2008). Transmission of Mycobacterium pinnipedii to humans in a zoo with marine mammals. *International Journal of Tuberculosis and Lung Disease, 12*, 1469-1473.

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

Kistler, L., Ware, R., Smith, O., Collins, M., & Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res, 45*(11), 6310-6320. doi:10.1093/nar/gkx361

Konomi, N., Lebwohl, E., Mowbray, K., Tattersall, I., & Zhang, D. (2002). Detection of mycobacterial DNA in Andean mummies. *J Clin Microbiol, 40*(12), 4738-4740.

Krause, J. (2010). From Genes to Genomes: What is New in Ancient DNA? *Mitteilungen der Gesellschaft für Urgeschichte, 19*, 11-34.

Larsen, C. S. (1994). In the wake of Columbus: Native population biology in the postcontact Americas. *American Journal of Physical Anthropology, 37*(S19), 109-154. doi:10.1002/ajpa.1330370606

Lee, J. Y. (2015). Diagnosis and treatment of extrapulmonary tuberculosis. *Tuberc Respir Dis (Seoul), 78*(2), 47-55. doi:10.4046/trd.2015.78.2.47

Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H., Alquraishi, S. A., . . . Orlando, L. (2017). Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst Biol, 66*(1), e1-e29. doi:10.1093/sysbio/syw059

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Lindahl, J. F., & Grace, D. (2015). The consequences of human actions on risks for infectious diseases: a review. *Infect Ecol Epidemiol, 5*, 30048. doi:10.3402/iee.v5.30048

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature, 362*(6422), 709-715. doi:10.1038/362709a0

Lindo, J., Huerta-Sanchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., . . . Malhi, R. S. (2016). A time transect of exomes from a Native American population before and after European contact. *Nat Commun, 7*, 13175. doi:10.1038/ncomms13175

Liu, W. Q., Feng, Y., Wang, Y., Zou, Q. H., Chen, F., Guo, J. T., . . . Liu, S. L. (2009). Salmonella paratyphi C: genetic divergence from Salmonella choleraesuis and pathogenic convergence with Salmonella typhi. *PLoS One, 4*(2), e4510. doi:10.1371/journal.pone.0004510

Llamas, B., Fehren-Schmitz, L., Valverde, G., Soubrier, J., Mallick, S., Rohland, N., . . . Haak, W. (2016). Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv, 2*(4), e1501385. doi:10.1126/sciadv.1501385

Loeffler, S. H., de Lisle, G. W., Neill, M. A., Collins, D. M., Price-Carter, M., Paterson, B., & Crews, K. B. (2014). The seal tuberculosis agent, Mycobacterium pinnipedii, infects domestic cattle in New Zealand: epidemiologic factors and DNA strain typing. *J Wildl Dis, 50*(2), 180-187. doi:10.7589/2013-09-237

Maricic, T., Whitten, M., & Paabo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One, 5*(11), e14004. doi:10.1371/journal.pone.0014004

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research, 20*, 1297-1303. doi:10.1101/gr.107524.110

McNeill, W. H. (1998). *Plagues and peoples*: Anchor Books.

Mendum, T. A., Schuenemann, V. J., Roffey, S., Taylor, G. M., Wu, H., Singh, P., . . . Stewart, G. R. (2014). Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics, 15*, 270. doi:10.1186/1471-2164-15-270

Merbs Charles, F. (1992). A new world of infectious disease. *American Journal of Physical Anthropology, 35*(S15), 3-42. doi:10.1002/ajpa.1330350603

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc, 2010*(6), pdb prot5448. doi:10.1101/pdb.prot5448

Monack, D. M., Mueller, A., & Falkow, S. (2004). Persistent bacterial infections: the interface of the pathogen and the host immune system. *Nat Rev Micro, 2*(9), 747-765.

Montiel, R., Solorzano, E., Diaz, N., Alvarez-Sandoval, B. A., Gonzalez-Ruiz, M., Canadas, M. P., . . . Malgosa, A. (2012). Neonate human remains: a window of opportunity to the molecular study of ancient syphilis. *PLoS One, 7*(5), e36371. doi:10.1371/journal.pone.0036371

Mukherjee, S. (2017). Emerging Infectious Diseases: Epidemiological Perspective. *Indian J Dermatol, 62*(5), 459-467. doi:10.4103/ijd.IJD_379_17

Muller, R., Roberts, C. A., & Brown, T. A. (2014). Genotyping of ancient Mycobacterium tuberculosis strains reveals historic genetic diversity. *Proceedings of the Royal Society B-Biological Sciences, 281*(1781). doi:ARTN 20133236

10.1098/rspb.2013.3236

Muller, R., Roberts, C. A., & Brown, T. A. (2016). Complications in the study of ancient tuberculosis: Presence of environmental bacteria in human archaeological remains. *Journal of Archaeological Science, 68*, 5-11. doi:10.1016/j.jas.2016.03.002

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol, 51 Pt 1*, 263-273.

Nugent, G. (2011). Maintenance, spillover and spillback transmission of bovine tuberculosis in multi-host wildlife complexes: a New Zealand case study. *Vet Microbiol, 151*(1-2), 34-42. doi:10.1016/j.vetmic.2011.02.023

O'Neill, M. B., Shockey, A. C., Zarley, A., Aylward, W., Eldholm, V., Kitchen, A., & Pepperell, C. S. (2018). Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *bioRxiv*.

O'Rourke, D. H., Hayes, M. G., & Carlyle, S. W. (2000). Spatial and temporal stability of mtDNA haplogroup frequencies in native North America. *Hum Biol, 72*(1), 15-34.

Ortner, D. J. (2003). *Identification of pathological conditions in human skeletal remains* (2 ed.): Academic Press.

Patel, K., Jhamb, S. S., & Singh, P. P. (2011). Models of latent tuberculosis: their salient features, limitations, and development. *J Lab Physicians, 3*(2), 75-79. doi:10.4103/0974-2727.86837

Pedersen, M. W., Ruter, A., Schweger, C., Friebe, H., Staff, R. A., Kjeldsen, K. K., . . . Willerslev, E. (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature, 537*(7618), 45-49. doi:10.1038/nature19085

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. *Genome biology, 17*, 60. doi:10.1186/s13059-016-0918-z

Pepperell, C. S., Granka, J. M., Alexander, D. C., Behr, M. A., Chui, L., Gordon, J., . . . Feldman, M. W. (2011). Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 6526-6531. doi:10.1073/pnas.1016708108

Pierron, D., Chang, I., Arachiche, A., Heiske, M., Thomas, O., Borlin, M., . . . Letellier, T. (2011). Mutation rate switch inside Eurasian mitochondrial haplogroups: impact of selection and consequences for dating settlement in Europe. *PLoS One, 6*(6), e21543. doi:10.1371/journal.pone.0021543

Pimenoff, V. N., Houldcroft, C. J., Rifkin, R. F., & Underdown, S. (2018). The Role of aDNA in Understanding the Coevolutionary Patterns of Human Sexually Transmitted Infections. *Genes (Basel), 9*(7). doi:10.3390/genes9070317

Prufer, K., Stenzel, U., Hofreiter, M., Paabo, S., Kelso, J., & Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biol, 11*(5), R47. doi:10.1186/gb-2010-11-5-r47

Ramenofsky, A. (2003). Native American disease history: past, present and future directions. *World Archaeology, 35*(2), 241-257. doi:10.1080/0043824032000111407

Refregier, G., Abadia, E., Matsumoto, T., Ano, H., Takashima, T., Tsuyuguchi, I., . . . Sola, C. (2016). Turkish and Japanese Mycobacterium tuberculosis sublineages share a remote common ancestor. *Infect Genet Evol, 45*, 461-473. doi:10.1016/j.meegid.2016.10.009

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., . . . Ruiz-Linares, A. (2012). Reconstructing Native American population history. *Nature, 488*(7411), 370-374. doi:10.1038/nature11258

Rivero, A., Marquez, M., Santos, J., Pinedo, A., Sanchez, M. A., Esteve, A., . . . Martin, C. (2001). High rate of tuberculosis reinfection during a nosocomial outbreak of multidrug-resistant tuberculosis caused by Mycobacterium bovis strain B. *Clin Infect Dis, 32*(1), 159-161. doi:10.1086/317547

Roberts, C. A., & Buikstra, J. E. (2003). *The Bioarchaeology of Tuberculosis: A Global View on a Reemerging Disease*: University Press of Florida.

Rohland, N., & Hofreiter, M. (2007). Ancient DNA extraction from bones and teeth. *Nat Protoc, 2*(7), 1756-1762. doi:10.1038/nprot.2007.247

Salo, W. L., Aufderheide, A. C., Buikstra, J., & Holcomb, T. A. (1994). Identification of Mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy. *Proc Natl Acad Sci U S A, 91*(6), 2091-2094.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Paabo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One, 7*(3), e34131. doi:10.1371/journal.pone.0034131

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes, 9*, 88. doi:10.1186/s13104-016-1900-2

Schuenemann, V. J., Avanzi, C., Krause-Kyora, B., Seitz, A., Herbig, A., Inskip, S., . . . Krause, J. (2018). Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLoS Pathog, 14*(5), e1006997. doi:10.1371/journal.ppat.1006997

Schuenemann, V. J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., . . . Poinar, H. N. (2011). Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. *Proc Natl Acad Sci U S A, 108*(38), E746-752. doi:10.1073/pnas.1105107108

Schuenemann, V. J., Kumar Lankapalli, A., Barquera, R., Nelson, E. A., Iraiz Hernandez, D., Acuna Alonzo, V., . . . Krause, J. (2018). Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Negl Trop Dis, 12*(6), e0006447. doi:10.1371/journal.pntd.0006447

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jager, G., Bos, K. I., . . . Krause, J. (2013). Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science, 341*(6142), 179-183. doi:10.1126/science.1238286

Shapiro, B., Rambaut, A., & Gilbert, M. T. P. (2006). No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). *International Journal of Infectious Diseases, 10*(4), 334-335. doi:10.1016/j.ijid.2006.02.006

Siddle, K. J., & Quintana-Murci, L. (2014). The Red Queen's long race: human adaptation to pathogen pressure. *Curr Opin Genet Dev, 29*, 31-38. doi:10.1016/j.gde.2014.07.004

Skoglund, P., & Mathieson, I. (2018). Ancient Human Genomics: The First Decade. *Annu Rev Genomics Hum Genet*. doi:10.1146/annurev-genom-083117-021749

Skoglund, P., Stora, J., Gotherstrom, A., & Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science, 40*(12), 4477-4482. doi:10.1016/j.jas.2013.07.004

Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B., & Collins, M. J. (2001). Neanderthal DNA. Not just old but old and cold? *Nature, 410*(6830), 771-772. doi:10.1038/35071177

Spyrou, M. A., Tukhbatova, R. I., Feldman, M., Drath, J., Kacki, S., Beltran de Heredia, J., . . . 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. doi:10.1016/j.chom.2016.05.012

Spyrou, M. A., Tukhbatova, R. I., Wang, C. C., Valtuena, A. A., Lankapalli, A. K., Kondrashin, V. V., . . . Krause, J. (2018). Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. *Nat Commun, 9*(1), 2234. doi:10.1038/s41467-018-04550-9

Stone, A. C., Wilbur, A. K., Buikstra, J. E., & Roberts, C. A. (2009). Tuberculosis and leprosy in perspective. *Am J Phys Anthropol, 140 Suppl 49*, 66-94. doi:10.1002/ajpa.21185

Taylor, M. (2016). DNA analysis of the human skeletal remains. In M. Dowd (Ed.), *Archaeological Excavations in Moneen Cave, The Burren, Co. Clare* (pp. 47-48). Oxford, England: Archaeopress Publishing Ltd.

Ubelaker, D. H. (1976). Prehistoric New World population size: Historical review and current appraisal of North American estimates. *American Journal of Physical Anthropology, 45*(3), 661-665. doi:10.1002/ajpa.1330450332

Wain, J., Hendriksen, R. S., Mikoleit, M. L., Keddy, K. H., & Ochiai, R. L. (2015). Typhoid fever. *Lancet, 385*(9973), 1136-1145. doi:10.1016/S0140-6736(13)62708-7

Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiss, C. L., Burbano, H. A., . . . Krause, J. (2017). A Robust Framework for Microbial Archaeology. *Annu Rev Genomics Hum Genet*. doi:10.1146/annurev-genom-091416-035526

Warinner, C., Robles García, N., Spores, R., & Tuross, N. (2012). Disease, Demography, and Diet in Early Colonial New Spain: Investigation of a Sixteenth-Century Mixtec Cemetery at Teposcolula Yucundaa. *Latin American Antiquity, 23*(4), 467-489.

Warinner, C., Rodrigues, J. F., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., . . . Cappellini, E. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nat Genet, 46*(4), 336-344. doi:10.1038/ng.2906

Wilbur, A. K., Bouwman, A. S., Stone, A. C., Roberts, C. A., Pfister, L.-A., Buikstra, J. E., & Brown, T. A. (2009). Deficiencies and challenges in the study of ancient tuberculosis DNA. *Journal of Archaeological Science, 36*(9), 1990-1997.

Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proc Biol Sci, 272*(1558), 3-16. doi:10.1098/rspb.2004.2813

Wolfe, N. D., Dunavan, C. P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature, 447*(7142), 279-283. doi:10.1038/nature05775

Wood, J. W., Milner, G. R., Harpending, H. C., Weiss, K. M., Cohen, M. N., Eisenberg, L. E., . . . Wilkinson, R. G. (1992). The Osteological Paradox: Problems of Inferring Prehistoric Health from Skeletal Samples [and Comments and Reply]. *Current Anthropology, 33*(4), 343-370.

Woolhouse, M. E. (2002). Population biology of emerging and re-emerging pathogens. *Trends Microbiol, 10*(10 Suppl), S3-7.

World Health Organization. (2017). Tuberculosis. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/

Zhou, Z., Lundstrom, I., Tran-Dien, A., Duchene, S., Alikhan, N. F., Sergeant, M. J., . . . Achtman, M. (2018). Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr Biol, 28*(15), 2420-2428 e2410. doi:10.1016/j.cub.2018.05.058

List of Figures

Appendices:
Papers I, II and III

Paper I

Å. J. Vågene, T. P. Honap, K. M. Harkins, M. Rosenberg, F. Cárdenas-Arroyo, L. P. Leguizamón, J. Arnett, J. E. Buikstra, A. Herbig, A. C. Stone, K. I. Bos, J. Krause (2018).
Geographically dispersed zoonotic tuberculosis in pre-contact New World human populations.
Manuscript.

Geographically dispersed zoonotic tuberculosis in pre-contact New World human populations

Authors: Åshild J. Vågene[1,2,†], Tanvi Honap[3,†,§], Kelly M. Harkins[4,‡], Michael S. Rosenberg[3], Felipe Cárdenas-Arroyo[5], Laura Paloma Leguizamón[5], Judith Arnett[4,6], Jane E. Buikstra[4] Alexander Herbig[1,2], Anne C. Stone [4,7,8,*], Kirsten I. Bos[1,2,*] and Johannes Krause[1,2,*]

Affiliations:
[1]Max Planck Institute for the Science of Human History, Jena, Germany
[2]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany
[3]School of Life Sciences, Arizona State University, Tempe, Arizona, USA
[4]School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, USA
[5]Colombian Institute of Anthropology and History (ICANH), Bogotá, Colombia
[6]University of the Andes, School of Medicine, Colombia
[7]Center for Evolution and Medicine, Arizona State University, Tempe, Arizona, USA
[8]Institute of Human Origins, Arizona State University, Tempe, Arizona, USA

[†]These authors contributed equally to this work
*Corresponding authors
[§]Current address for T.H.: Department of Anthropology, University of Oklahoma, Norman, Oklahoma, USA
[‡] Current address for K.H.: Claret Bioscience LLC., Santa Cruz, California, USA

ABSTRACT

Zoonotic transmissions of animal-adapted pathogens with broad host ranges pose an increasing threat to human and animal health. Recently, strains belonging to the *Mycobacterium tuberculosis* complex (MTBC) most commonly found today in pinnipeds (*M. pinnipedii*) were recovered from coastal Peruvian human remains pre-dating European arrival in the Americas. This was interpreted to result from an ancient zoonotic event through frequent contact with infected living seals or their tissues. Skeletal evidence indicates the presence of tuberculosis in North and South America before contact, though in geographical locations that are incompatible with zoonotic acquisition via direct contact with infected pinnipeds. Here we investigate the relationships between pre-contact MTBC strains through the analysis of three new MTBC genomes from contemporaneous human remains in coastal Peru and inland Colombia. All three individuals were infected with *M. pinnipedii* strains, and while zoonotic pinniped transmission remains possible for the Peruvian case, it does not easily account for its presence in inland Colombia. We explore different scenarios for disease

transmission, including the potential of human adaptation and/or animal mediated dissemination. Together these data demonstrate the ability of ancient *M. pinnipedii* strains to cause human infection in the past and point to a more complex transmission route than simple pinniped to human transfer of tuberculosis for the inland pre-contact era Colombian individuals.

INTRODUCTION

Tuberculosis (TB) is the leading causes of human death due to a single infectious agent (World Health Organization, 2017a). It is estimated that a quarter of the world's population carries a latent infection (Houben & Dodd, 2016), with 1.7 million deaths reported in 2016 (World Health Organization, 2017b). The majority of human TB infections are caused by *Mycobacterium tuberculosis sensu stricto* and *M. africanum* strains that comprise seven human-adapted lineages (Daniela Brites & Gagneux, 2017). These, together with several animal-associated strains and the ancestral "smooth tubercle bacilli" *M. canettii*, form the *Mycobacterium tuberculosis* complex (MTBC).

Today, the number of reported human TB cases caused by animal-associated strains is low. In 2010, for example, the WHO estimated only 121,268 TB cases of a zoonotic origin (World Health Organization, 2015), however this is likely an underrepresentation since clinical strain assignment is rarely performed (Olea-Popelka et al., 2017). The majority of human zoonotic cases are attributed to *M. bovis*, which predominantly causes TB infection in cattle, and to a lesser extent *M. caprae*, which is associated with domestic sheep and goats (Pesciaroli et al., 2014). The notion that MTBC is comprised of host-specific strains (Brosch et al., 2002) is continually challenged by examples that document inter-species transfer between wild, captive and domesticated animals, and humans (Daniela Brites et al., 2018; Coscolla et al., 2013; Malone & Gordon, 2017; Nugent, 2011; Olea-Popelka et al., 2017; Pesciaroli et al., 2014). Our limited knowledge about the zoonotic capacities of these strains makes them an underappreciated threat to both human and animal health in many parts of the world (Allen, 2017; Nugent, 2011).

Genomic data from ancient bacterial and viral organisms have revealed important details regarding disease ecology in the past (Achtman, 2016; Andam, Worby, Chang, & Campana, 2016; Harkins & Stone, 2015). In 2014, Bos *et al.* reported the complete genome sequences for three MTBC strains recovered from archaeological skeletal remains of individuals from the southern coast of Peru, which pre-date European contact in the New World. Skeletal pathology has indicated the presence of tuberculosis in indigenous New World populations long before European contact (Roberts & Buikstra, 2003; Stone, Wilbur, Buikstra, & Roberts, 2009), an observation that was difficult to reconcile with the dominance of modern European *M. tuberculosis* lineages in the New World today (Coscolla & Gagneux, 2014). These ancient Peruvian strains are distinct from human-adapted MTBC and are genetically most closely related to modern *M. pinnipedii*. The *M. pinnipedii* lineage is associated with pinnipeds (seals and

sea lions) and rarely causes human infection today (Cousins et al., 2003; Forshaw & Phelps, 1991; Kiers, Klarenbeek, Mendelts, Van Soolingen, & Koeter, 2008). Ancient zoonotic events related to the manipulation and consumption of infectious seal tissues likely accounts for their transmission to pre-contact human populations who occupied this coastal region (Bos et al., 2014).

The archaeological observations of human skeletal and soft tissue remains displaying pathological changes consistent with TB infection have been recorded in many pre-contact archaeological sites across South and North America (Roberts & Buikstra, 2003; Stone et al., 2009). The earliest cases come from Peru and northern Chile and are dated to ~700 C.E., with possible cases occurring as early as 290 C.E. (Allison, Gerszten, Munizaga, Santoro, & Mendoza, 1981; Roberts & Buikstra, 2003; Salo, Aufderheide, Buikstra, & Holcomb, 1994). Peru and northern Chile also have the highest density of archaeological TB cases in South America. By contrast, affected human remains from northern regions of South America are fewer in number. Cases of pre-contact TB in North America begin to appear in the archaeological record after 900 C.E., and the majority of affected individuals are located at inland sites in the midcontinent and southwestern U.S.A. (Roberts & Buikstra, 2003). The presence of putative TB pathologies at inland sites across the Americas – in places where direct contact with pinniped and coastal human populations would have been unlikely – invites a more thorough sampling of ancient human remains to better understand past genetic diversity and the temporal and geographic distribution of MTBC strains circulating in the pre-contact New World.

Here we present three additional ancient South American MTBC genomes recovered from pre-contact human skeletal remains. All three strains belong to the *M. pinnipedii* lineage and thus reveal its widespread geographical presence. Importantly, two of the individuals from whom genomes were recovered derive from inland Colombian sites where human contact with infectious pinniped tissues is unlikely to have occurred. This finding reveals a more complex transmission pathway that moves beyond a simple pinniped-to-human zoonotic event. Our results highlight the zoonotic potential of *M. pinnipedii* in the past and its capacity for causing human infection.

RESULTS

Screening

A total of ten individuals showing signs of skeletal TB were screened for this study. Four DNA extracts (AD82, AD281b, AD382, and AD386) were considered positive for MTBC DNA based on qPCR assay results. Positive qPCR results for AD82 and AD281a have already been published elsewhere (Harkins et al., 2015). Second-tier screening via gene capture, targeting five mycobacterial genes: *rpoB*, *gyrA*, *gyrB, katG* and *mtp40*, yielded positive results for AD281b and AD386 wherein all five genes were sufficiently covered. AD382 did not yield any coverage of the *mtp40* gene and was

therefore not included in the downstream whole-genome in-solution capture. The MTBC gene capture results for AD82 were also previously published (Bos et al., 2014). AD281c was not tested via qPCR or gene capture, but was included in the whole genome in-solution capture based on the positive results for AD281a and AD281b, as all three samples are from the same individual. The qPCR and gene capture screening results are summarized in Supplementary Table 1.

Ancient Colombian and Peruvian genomes

Based on MTBC genome capture data, we reconstructed complete genomes for three ancient MTBC strains, each from a separate individual. Two strains, AD281 and AD386, were reconstructed from individuals from the respective Colombian sites of Las Delicias and Candelaria La Nueva, located in the modern city of Bogotá, which is situated more than 600 km inland and ~2,640 m above sea level in the Eastern Cordillera of the Andes (Supplementary section 1). The third strain, AD82, was recovered from an individual from the Moquegua, M6: Estuquiña site, located at the origin of the Osmore River in the Osmore drainage, ~57 km inland from the southern coast of Peru, situated 2000m above sea level in the Andes (Rice, Conrad, & Buikstra, 1990) (Supplementary section 1). We captured three libraries generated from different skeletal elements of the AD281 individual (AD281a, AD281b and AD281c), and constructed the genome based on capture data from AD281a and AD281b (see below). Radiocarbon dates indicate that AD281 (1265-1380 C.E.) pre-dates European contact, while AD386 (1450-1640 C.E.) overlaps with European presence in the Bogotá region, beginning 1536-1537 C.E. (Francis, 2007; Kurella, 1998) (see Supplementary Table 2). The site from which individual AD82 was excavated has been archaeologically dated to the pre-contact period (1250-1470 C.E.) (Sharratt, 2017).

After in-solution capture, a comparison of shotgun and capture data for the UDG treated libraries yielded enrichment efficiencies of 20-, 45-, 43- and 43-fold for AD82, AD281a, AD281b and AD386, respectively. These samples had endogenous DNA content ranging from 0.95% to 2.08% after capture (Supplementary Table 9). We used one of the samples (58U) from which an ancient MTBC genome had been reconstructed earlier (Bos, 2014) as a positive control. In this study, the 58U library showed 42.98% endogenous DNA after capture – indicating that the low percent endogenous DNA observed in the other libraries was not caused by faulty experimental conditions. Metagenomic analyses, carried out using the Megan ALignment Tool (MALT) (version 0.3.8), of the UDG-treated shotgun data from the sample libraries revealed that 0.004-0.012% of reads were assigned to members of the MTBC and 9.7-14.3% could be assigned to other prokaryotic or eukaryotic organisms, including non-MTBC mycobacteria (Supplementary Tables 3, 4; Supplementary Figure 1). The remaining reads from our sequencing data could not be assigned to an entry in our database, which consisted of all complete genomes in the NCBI Nucleotide (nt) database (7th Dec. 2016), when using a sensitive 85% minimum identity threshold. These contaminant reads persist after capture, where the number of MALT assigned MTBC reads represented between 0.3-1.2% or reads and non-MTBC reads rose to 19.2-25.1%,

demonstrating that our MTBC capture assay also enriches DNA from closely related mycobacterial and non-mycobacterial species that have infiltrated our samples (Supplementary Fig. 1; Supplementary Tables 3, 4; Materials and Methods). The sequenced capture product for AD281c did not meet our threshold of 0.4% MALT-assigned endogenous MTBC reads after capture and was therefore not sequenced deeper for genome reconstruction (Supplementary Table 4). Additionally, we compared the ratios of MTBC and non-MTBC mycobacterial reads present in three of our samples (AD82, AD281b, AD386) to those of the other Peruvian samples (Bos et al., 2014) – the data for which are considered to have a low background of contaminant mycobacterial reads. This procedure assessed the level of contaminant mycobacterial reads present in our samples and allowed us to anticipate whether such closely genetically related bacteria could interfere with downstream analyses. MALT results for the non-UDG (nU) treated shotgun data, generated using both 85% and 95% "minimum percent identity", were used to calculate ratios of MTBC and non-MTBC mycobacterial reads (refer to Supplementary Tables 5, 6, 7, 8). The previously published Peruvian samples (54nU, 58nU, and 64nU) contain between 0.029 to 0.459 non-MTBC mycobacterial reads for every MTBC assigned read at 85% identity; and is similarly between 0.014 to 0.078 when using 95% identity (Supplementary Table 8). The ratio of non-MTBC mycobacterial DNA in our three new genomes is comparatively higher with 139, 9.4 and 14.2 non-MTBC mycobacterial reads for every MTBC assigned read at 85% identity and 29, 1.92 and 2.6 reads at 95% identity for AD82, AD281b and AD386 respectively (Supplementary Table 8). This demonstrates that our three sample libraries contain a substantial amount of non-MTBC mycobacterial DNA in comparison to the samples published by Bos et al. (2014) and that such reads persist after the application of stringent (95% ID) filtering criteria.

The AD281a and AD281b genomes, retrieved from a vertebra and a rib, respectively, from the same individual, were determined to be identical, and the data were therefore combined (Methods and Materials; Supplementary Figure 2). From this point onwards, we refer to this composite genome as AD281.

The sequencing data generated from the captured UDG-treated libraries yielded genomes reconstructed at an average depth of coverage of 14.9-, 15.3- and 10.8-fold with 80%, 81% and 83% of the ancestral MTBC reference covered at least 5-fold for AD82, AD281 and AD386 respectively (Supplementary Table 9). All negative controls were negative for MTBC DNA after in-solution capture (Supplementary Table 9).

Capture data from the non-UDG treated libraries were used to determine deamination patterns for the enriched ancient MTBC DNA and co-enriched ancient human DNA. Ancient DNA damage occurs at the ends of degraded DNA fragments in the form of cytosine deamination that accumulates over time (Sawyer, Krause, Guschanski, Savolainen, & Paabo, 2012). The deamination patterns for the MTBC DNA in the AD82, AD281, and AD386 non-UDG genome data were initially respectively estimated to be 4.16 %, 8.58%, and 8.11% on the first base of the 5-prime end of the

reads (Supplementary Table 9). We deemed these damage patterns to be lower than expected, especially in the case of AD82. In an attempt to ameliorate these low damage rates, we used a selective stringent mapping approach to discard reads that contained a high number of mismatches not contained in the four bases at the ends of the reads – where deaminated bases tend to occur at highest frequency. Clipped reads that mapped using stringent criteria were then filtered out of the non-clipped non-UDG dataset and used to generate damage patterns (see Materials and Methods). Reappraisal of the damage patterns yielded 7.69% (AD82), 8.72% (AD281) and 7.32% (AD386) on the first base of the 5-prime end of the reads, thus confirming the presence of ancient DNA (Sawyer et al., 2012) (Supplementary Table 10; Materials and Methods). The deamination pattern observed for the human DNA was higher (~14% for all samples) than that for MTBC (Supplementary Table 11). The lower rates of DNA damage observed for the MTBC reads could be due to the thicker mycobacterial cell wall, whereby mycolic acid in the cell wall decreases permeability and potentially protects the DNA over time from hydrolytic and enzymatic degradation, as has previously been suggested for ancient *M. leprae* DNA (Schuenemann et al., 2013).

Diversity amongst ancient *M. pinnipedii* genomes

A dataset comprising our three ancient genomes, the three previously published Peruvian *M. pinnipedii* genomes and 266 modern MTBC genomes was used for phylogenetic analyses (Supplementary Table 12). Our ancient genomes were phylogenetically clustered with the previously published ancient Peruvian and modern *M. pinnipedii* genomes and retained the same phylogenetic positioning regardless of the tree building method (Fig. 2; Supplementary Fig. 3, 4). The two Colombian genomes (AD281 and AD386) are closely related and form a clade basal to the ancient Peruvian and modern *M. pinnipedii* diversity. The new Peruvian genome (AD82) clusters with those previously published from the region, but diverges earlier.

Despite all six ancient *M. pinnipedii* genomes being almost contemporaneous, longer branch lengths were observed in our new ancient genomes that indicate a higher number of derived positions than those observed for the other ancient Peruvian genomes (54U, 58U and 64U) (Bos et al., 2014). This is unexpected under the assumption of a constant evolutionary rate and could be a sign of rate heterogeneity within the diversity of ancient *M. pinnipedii* strains, if they represent true branch lengths. An investigation of heterozygous variant calls showed that AD82, AD281 and AD386 have much higher numbers of heterozygous sites (Supplementary Fig. 5). We believe this is best explained by the presence of genetically similar non-target DNA stemming from environmental taxa that infiltrated the bones and were co-enriched during the capture. These reads remain in our mapping despite our adherence to stringent mapping parameters. The heterozygous positions infringe upon our threshold for calling homozygous variant calls (90% of reads in agreement or more), making Bayesian-based molecular dating impractical. Regardless, all ancient genomes show some degree of branch shortening.

SNP analysis of protein coding genes

1,137 variant positions occur in coding regions in at least one of the six ancient *M. pinnipedii* genomes. Of these 686 are non-synonymous: 656 create non-synonymous amino acid changes, 30 SNPs create pseudogenes through the loss of two start-codons (START_LOST) and gain of 28 stop-codons (STOP_GAINED). 451 represent synonymous changes, and an additional 22 occur in non-coding RNA related genes (Supplementary Table 13). We chose to limit our investigation to non-synonymous SNPs that are shared by two or more ancient *M. pinnipedii* genomes. We did not investigate variant calls that were unique to any of our single genomes due to the high amount of contaminating reads in our data that may contribute to false positive SNP calls.

We used an updated version of MultiVCFanalyzer (v. 0.87) (Bos et al., 2014) (https://github.com/alexherbig/MultiVCFAnalyzer) to collate and filter the SNP calls according to our thresholds. Our results showed that a SNP previously called in Rv3768 (position 4,214,338) was not unique to the three ancient Peruvian genomes published by Bos et al. (2014), but was in fact shared by all genomes in the *M. pinnipedii* clade. The results also showed that an additional nsSNP occurring in gene *accD3* was unique to the three Peruvian genomes. *accD3* encodes acetyl-CoA carboxylase carboxyl transferase subunit beta, which plays a key role in the biosynthesis of mycolic acid (Gande et al., 2004).

24 genes within the *M. pinnipedii* clade contain multiple non-synonymous SNPs occurring in branches shared by at least two strains. Several are noteworthy with regard to the ancient genomes. The ancient Colombian genomes share a non-synonymous SNP in the *malQ* gene, which is also a pseudogene in the six modern *M. pinnipedii* genomes. *malQ* is also pseudogenized in six Lineage 3 genomes, five of which form a subclade. We also note that one Lineage 1 strain (L1_V23210) and two Lineage 6 strains (L6_N0089 and L6_N1058) contain non-synonymous SNPs in the same base position (2,015,930) in codon 516, and that a further sixteen strains distributed across the phylogeny have non-synonymous SNPs in *malQ*. *malQ* encodes 4-alpha-glucanotransferase which degrades maltose and maltodextrins, resulting in their conversion to glucose for use as a carbon source by bacteria (Boos & Shuman, 1998; Sato, Okamoto-Shibayama, & Azuma, 2015). *E. coli* was shown to be unable to grow on maltose when this gene was knocked out (Pugsley & Dubreuil, 1988). The two Colombian strains share two additional non-synonymous SNPs, one in both *recA* and *recN*; three of the modern *M. pinnipedii* genomes (Pinnipedii_7739, Pinnipedii_7011 and Pinnipedii_G01222) also share a non-synonymous SNP in *recN*. The proteins produced by *recA* (Rv2737c) and *recN* (Rv1696) are integral to the repair of double-stranded DNA breaks and homologous recombination. RecN is one of the first responders to a site where a double-stranded break has occurred. The presence of RecN is crucial when DNA breakage occurs at multiple sites, where a possible function of the protein is to act as a cohesin during homologous recombinational repair (Odsbu & Skarstad, 2014). A role of RecA is to recruit RecN to the sites where double strand

breaks have occurred (Keyamura, Sakaguchi, Kubota, Niki, & Hishida, 2013). In *Bacillus subtilis,* the presence of RecA is thought to be necessary for inducing the expression level of RecN (Cardenas, Gandara, & Alonso, 2014). The Colombian genomes also share two non-synonymous SNPs in gene Rv1065, which encodes for a hypothetical protein. One of these SNPs creates a stop codon (STOP_GAINED) causing Rv1065 to become pseudogenized. Furthermore, two non-synonymous SNPs occur in the *mmpL12* gene within the *M. pinnipedii* clade; one is shared by the two Colombian genomes, and the other is shared by the six modern *M. pinnipedii* genomes. *mmpL12* is thought to encode for a probable transmembrane transport protein, but no further function of this protein has been characterized (http://www.uniprot.org/uniprot/P9WJT7).

Cellular systems of ion import and export play a crucial role in bacterial osmoregulation, intracellular homeostasis and membrane potential. Bos et al. (2014) remarked on the presence of metal ion transporters within the *M. pinnipedii* clade as well as along the branch shared by *M. pinnipedii* and *M. microti*. Additionally, they reported positive selection acting on codon 62 in the *ctpA* gene that encodes the CtpA copper efflux protein in the three Peruvian genomes (54U, 58U and 64U) (Bos et al., 2014). Therefore, we investigated the presence of SNPs in ion transporter genes within the *M. pinnipedii* clade. We identify three non-synonymous SNPs occurring in three different genes that are shared by all *M. pinnipedii* genomes in our phylogeny. They occur in *kdpA* (Rv1029) – involved in potassium import (Huang, Pedersen, & Stokes, 2017), *ctpE* (Rv0908) – involved in calcium uptake (Gupta, Shrivastava, & Sharma, 2017) and Rv3679 linked to the transport of an undetermined anion (https://mycobrowser.epfl.ch/genes/Rv3679). In addition to *ctpA*, there are two further non-synonymous SNPs shared by two Peruvian genomes, 54U and 58U, occurring in genes *arsC* (Rv2643) and *mgtE* (Rv0362), which are involved in metal ion transportation. *arsC* is involved in the export and detoxification of arsenic compounds across the cell's membrane. In *M. tuberculosis*, ArsC is a fusion protein, also referred to as ACR3-ArsC, that functions both as an arsenite transporting domain and an arsenate reductase domain (Wu, Song, & Beitz, 2010), and may potentially confer arsenic resistance (Cole et al., 1998). MgtE is a highly selective channel that regulates the intracellular concentration of magnesium (Mg2+), a crucial ion involved in many cellular processes (Hattori et al., 2009). An additional non-synonymous SNP in *mgtE* in the *M. pinnipedii* clade is shared by three modern strains (Pinnipedii_7739, Pinnipedii_7011 and Pinnipedii_G01222). Additionally, a non-synonymous SNP in *cysA1*, involved in the transport of sulfate and thiosulfate is shared by the six modern *M. pinnipedii* strains (https://mycobrowser.epfl.ch/genes/Rv2397c).

DISCUSSION

Our results provide further support for the findings that South American pre-contact tuberculosis documented by bioarchaelogical evidence was caused by *M. pinnipedii*, indicating that in antiquity *M. pinnipedii* had a geographic range beyond the Peruvian

coast (Bos et al., 2014). Our genome (AD82) from Moquegua, M6: Estuquiña is the fourth to be recovered from individuals inhabiting the Osmore River Valley, a narrow geographic region on the southern coast of Peru. The peoples inhabiting this specific region as well as the wider region of coastal Peru and northern Chile in the first millennium C.E. are known to have exploited seal tissues for sustenance and tool manufacture (Benson, 2012; Donnan, 1978; Reinhard & Urban, 2003; Swenson, 2006). Consumption of, or close contact with, infectious seal tissues likely provided multiple opportunities for zoonotic infection. Skeletal lesions consistent with prolonged TB infection are prevalent in the Moquegua, M6: Estuquiña skeletal assemblage, where they occur in ~9% percent of the general population and in 19.2% of adult males and 9.8% in adult females (Buikstra & Williams, 1991) (Supplementary section 1), thus suggesting epidemic proportions of chronic TB infection. The higher percentage of males with skeletal TB lesions could indicate that men were more frequently exposed to MTBC bacteria or that they were more likely to sustain a prolonged infection allowing skeletal lesions to form.

Our detection and recovery of *M. pinnipedii* genomes from Colombian individuals that lived more than 600 km inland point to an ecological model for this pathogen that went beyond simple seal-to-human zoonotic transmission events, as have been proposed for the coastal Peruvian individuals (Bos et al., 2014). Furthermore, archaeological evidence suggestive of pinniped tissue exploitation is lacking from these inland regions and the Colombian coastline is outside the range of modern pinniped species from which *M. pinnipedii* has so far been recovered (Bastida et al., 1999; Boardman et al., 2014; de Amorim et al., 2014), although their range may have extended farther north in the past (Fig. 1). Therefore, alternative hypotheses to that of pinniped-to-human transmission need to be considered to account for *M. pinnipedii*'s geographically broad dispersal in South America at this time.

The possibility of *M. pinnipedii* having adapted to and spread within humans after zoonotic transfer could account for the patterns noted in our data. The Colombian individuals were Muisca, who were known to form an essential part of far-reaching trade networks that went beyond the Andean region, facilitating contact between coastal and inland populations (Kurella, 1998). Trade would have provided an opportunity for *M. pinnipedii* to be brought inland via human movements. Since *M. pinnipedii* is a member of an animal-adapted clade that diverged from a predominantly human pathogen, a scenario of human-to-human transmission of the pathogen would constitute an example of a human-adapted pathogen becoming re-adapted to the human host, a phenomenon already observed in nosocomial outbreaks of *M. bovis* in Spain (Gonzalo-Asensio et al., 2014; Rivero et al., 2001). Today, *M. pinnipedii* is known to cause human infection occasionally, but only under circumstances of prolonged exposure to infected animals (Cousins et al., 2003; Kiers et al., 2008).

The western coast of South America has been proposed as the geographic region where human TB first developed in the New World (Roberts & Buikstra, 2003). This is

supported by archaeological data that demonstrate Peru and northern Chile to have the highest density and earliest cases of pre-contact human TB-like skeletal lesions (Roberts & Buikstra, 2003; Stone et al., 2009). Subsequent adaptation to, or maintenance by, humans and/or other terrestrial animal hosts could account for the later emergence of TB-associated skeletal lesions in North American populations that appear, mostly at inland sites, as early as 900 AD (Roberts & Buikstra, 2003). To test this hypothesis it will be necessary to analyze human skeletal remains from pre-contact North American sites for the presence of *M. pinnipedii.*

The ability of *M. pinnipedii* to affect a range of host species, a trait increasingly observed for many MTBC strains (Jurczynski et al., 2011; Kiers et al., 2008; Nugent, 2011; Zachariah et al., 2017), coupled with the high diversity of terrestrial mammal species in South America invites a conceptual exploration of different transmission pathways that may have facilitated its geographic spread. Transmission of *M. pinnipedii* from seals to cattle grazing on the New Zealand coastline has been observed (Loeffler et al., 2014), as well as transmission, independent of pinniped involvement, between captive zoo animals (Jurczynski et al., 2011). Additionally, pathogenicity experiments have shown guinea pigs, one of the few domesticated animals in South America, to be susceptible to *M. pinnipedii* (Cousins et al., 2003), as well as human-adapted MTBC strains (Clark, Hall, & Williams, 2014). Studies have demonstrated the complexity of *M. bovis* transmission pathways involving a variable range of host species that depend on a region's local ecology (Allen, 2017; Nugent, 2011). It is feasible that animal-to-human transmission could have occurred independent of, or in concert with, human-to-human transmission, allowing it to spread across South America via a terrestrial route. Dispersal across the geographical expanse considered here would require the involvement of at least one, if not multiple, terrestrial host species able to maintain the infection. Frequent contacts would be required between humans and any non-human host to facilitate transmission, thus making domesticated animals or widespread rodent species potential candidates. Future study of archaeological faunal material may elucidate the contributions of animal host(s) in past transmission networks.

The current phylogeny reveals the human-derived *M. pinnipedii* strains to occupy the most basal positions of this clade. Taken at face value, our data could be interpreted as a human transfer of the pathogen to pinnipeds rather than the reverse. Although the transfer of human-associated MTBC strains to other species is a known phenomenon (Ameni et al., 2011; Zachariah et al., 2017), we regard this phylogenetic structure to reflect sampling biases since all ancient MTBC genomes currently derive from humans.

Our Colombian and Peruvian genomes contribute two new branches to the *M. pinnipedii* clade, both of which are basal to the divergence between the previously published ancient Peruvian and modern *M. pinnipedii* genomes. The two Colombian genomes form the most basal branch, followed by Peruvian genome AD82. Overall, our genomes extend the known diversity of ancient *M. pinnipedii* strains that circulated in pre-contact South America (Figure 2). Additionally, all six ancient genomes were

assigned date ranges that partially overlap within the ~600-year time interval between 1028-1640 C.E (Bos et al., 2014) (see Results; Supplementary Table 2). The near contemporaneous presence of paraphyletic strains within the *M. pinnipedii* clade at geographically distant locations indicates that multiple introductions of *M. pinnipedii* may have occurred in South America. Near contemporaneous strain diversity is also observed on a narrower geographic scale in the Osmore River Valley, where the basal positioning of the AD82 genome to all other ancient Peruvian genomes illustrates the presence of a variety of lineages. This strain diversity might reflect multiple individuals having become infected by different *M. pinnipedii* strains circulating in ancient pinniped populations through the repeated exploitation of infected seal tissues. It is important to note that we do not currently know if the three previously published Peruvian genomes (54U, 58U and 64U) evolved from a common ancestor that was introduced to humans on a single occasion – becoming human adapted and spreading via human-to-human transmission – or whether this reflects a subset of the diversity evolving in pinniped populations that was transferred to humans. If the strain diversity in this region originated from the same pinniped-to-human introduction event of a single strain, one would expect the AD82 genome to fall into a monophyletic clade with the other Peruvian strains given their proximity in time and space, which it does not. With regard to the phylogeographic distribution observed for the Colombian strains, we believe that multiple introductions of *M. pinnipedii* from pinniped populations to human and/or terrestrial animal populations is the most parsimonious explanation for their spread to these inland locations.

Two studies to date have incorporated ancient genome data into Bayesian molecular dating analyses of the MTBC phylogeny. Bos et al. (2014) calculated the root of the MTBC to have emerged less than 6000 YBP, with an estimated substitution rate of $4.6x10^{-8}$ per nucleotide per year. Kay, et al. (2015) observed a similar substitution rate ($5.00x10^{-8}$ per nucleotide per year) when they dated the emergence of human adapted lineage 4. The similarity between these substitution rates provides support for the young date of MTBC emergence put forward by Bos et al. (2014). This emergence date of 6000 YBP precludes the possibility of humans having brought MTBC to the New World during the initial waves of settlement approximately 15,000 YBP (Skoglund & Reich, 2016). Future Bayesian-based molecular dating attempts that make use of ancient MTBC genomes from different lineages will be instrumental in corroborating this young emergence date and shaping the way we contextualize our findings.

Our metagenomic evaluation of the sequencing data generated from our sample libraries shows that non-MTBC mycobacterial DNA was enriched alongside the true ancient MTBC DNA during hybridization capture (Supplementary Fig. 1; Supplementary Table 4). The enrichment of non-MTBC mycobacterial DNA during capture points to the well-known genetic similarity between environmental mycobacteria and MTBC strains (Malhotra, Vedithi, & Blundell, 2017). Therefore, not all non-MTBC mycobacterial reads are filtered out of the data when applying stringent mapping parameters in BWA (Li & Durbin, 2009) (see Methods). Our investigations

of heterozygous SNP positions in the alignment of the ancient genomes confirm the presence of non-specific reads (Supplementary Fig. 5) and likely account for the unexpectedly long phylogenetic branch-lengths observed for our ancient genomes (Fig 2; Supplementary Fig. 2, 3). Our analyses show that environmentally derived mycobacterial DNA can be a confounding factor in the capture and analysis of MTBC DNA. The genomes in this study were retrieved from skeletal remains that had been exposed to direct contact with soil and moisture, likely accounting for the high percentage of contamination by environmental mycobacterial DNA (Supplementary section 1). Conversely, the Peruvian *M. pinnipedii* genomes generated by Bos et al. (2014) were recovered from individuals interred in arid stone-lined tombs, reducing exposure to soil and moisture, likely accounting for the low percentage of non-MTBC mycobacterial DNA observed in these samples. Therefore, we urge researchers seeking to analyze MTBC DNA from metagenomic samples to assess their data for the presence of non-MTBC mycobacterial DNA to avoid inadvertent misinterpretations of results and misidentification of true MTBC DNA.

The functional implications of the SNPs identified by our study are unknown but could be the result of selective pressures. For example, the occurrence of non-synonymous SNPs in *recA* and *recN*, two genes involved in the repair of double-strand breaks, might indicate that selective pressure has acted on this repair pathway in the ancient Colombian strains and certain modern *M. pinnipedii* strains. However, the functional implications of these SNPs are unknown. RecA is thought to be essential for the survival of intracellular pathogens, such as MTBC (Sander et al., 2001). Previous research has shown that *recA* knockout mutants cause increased chromosome-fragmentation, higher susceptibility to DNA-damaging agents and apoptosis-like cell death (Erental, Kalderon, Saada, Smith, & Engelberg-Kulka, 2014; Kouzminova, Rotman, Macomber, Zhang, & Kuzminov, 2004; Sander et al., 2001). Surprisingly, a knockout in *recA* in the *M. bovis* BCG strain, the basis for the TB vaccine, did not cause it to become attenuated, thus indicating that a functional *recA* is not required to cause infection (Sander et al., 2001). The non-synonymous SNP in *malQ* shared by the Colombian genomes and the gene's pseudogenization in all six modern genomes, could be indicative of an adaptive change related to a reduced availability of maltose, thus rendering MalQ, which is essential to maltose metabolism (Weiss, Skerra, & Schiefner, 2015), obsolete. This may account for the convergent evolution in the pseudogenization of *malQ* observed for the modern *M. pinnipedii* strains and six human adapted Lineage 3 strains. We also note an accumulation of SNPs within genes involved in import/export of different ions in the *M. pinnipedii* clade, especially along the branch shared by two ancient Peruvian genomes (58U and 54U) where 2 out of 7 non-synonymous SNPs occur in genes involved in trans-membrane ion transport. As previously noted (Bos et al., 2014), this could be related to adaptation to ion availability in the host.

These new genomes extend our knowledge of the phylogeographic and genetic diversity amongst ancient *M. pinnipedii* strains previously known to be circulating in pre-contact South American populations (Bos et al., 2014). By doubling the number of

ancient MTBC genomes retrieved from the Americas, we highlight *M. pinnipedii*'s capacity for human infection in unexpected inland locations where contact with marine mammals would have been limited or non-existent. Screening archeological fauna for MTBC DNA may be of great relevance to elucidate the evolution and ecology of animal associated MTBC strains in the Americas and elsewhere. Future research into the genomics of animal MTBC infections in general may contribute to a better understanding of its past history and transmission networks in both historical and modern contexts.

MATERIALS & METHODS

Sampling & DNA extraction

DNA was extracted from bone powder sampled from the ribs and vertebrae of nine individuals excavated from six different pre-/peri-contact archaeological sites across Colombia (Supplementary Table 1). All individuals displayed lesions compatible with long-term infection by a member of the MTBC. One bone from each individual was sampled, except for individual AD281, where three bones were subsampled: one vertebra (AD281a) and two ribs (AD281b, AD281c).

All samples (except AD281a and AD281c) were processed in the cleanroom facilities at Arizona State University (ASU), U.S.A. Debris and dirt was removed from the surface of the bones using a sterilized Dremel tool, the bones were then subsampled. Following this the bone sub-samples were wiped with 10% bleach solution followed by distilled water, and UV irradiated for 1 minute on each side, and subsequently powdered using the 8000M Mixer/Mill (SPEX). DNA extraction was carried out using a protocol tailored for ancient DNA (Dabney et al., 2013), using between 50-100 mg of bone powder for each sample. Extracts were eluted in 100 µL EBT buffer pre-heated to 65°C. An extraction blank (negative control) was introduced in each batch of extractions to check for possible contamination introduced during the extraction process. All DNA extracts and extraction blanks were quantified using the Qubit dsDNA High Sensitivity assay (Life Technologies).

Samples AD281a and AD281c were sampled in the cleanroom facilities at the University of Tübingen, Germany. A dental drill was used to drill bone powder, ~50 mg of bone powder from each sample was extracted (Dabney et al., 2013), during the lysis step samples were rotated for at least 16 hours. One negative control and one positive control (bone powder from an ancient cave bear) were included in the extraction batch. DNA extracts were eluted in 100 µL of TET (10 mM Tris-Cl, pH 8.0; 1 mM EDTA, pH 8.0; 0.05% Tween-20).

Screening for MTBC DNA

DNA extracts were screened for the presence of MTBC DNA using quantitative PCR (qPCR) assays and an in-solution hybridization capture.

qPCR assays: Undiluted extracts and extraction blanks were tested for MTBC DNA using three TaqMan qPCR assays. A 1:10 dilution of each extract was also used to test for presence of inhibitory substances in the ancient DNA extracts. The first qPCR assay (rpoB2 assay) targets a region of the *rpoB* gene, which is a single-copy gene found in all bacteria and codes for RNA polymerase subunit B. This assay uses a TaqMan probe that binds to an MTBC-specific sequence in the gene (Harkins et al., 2015); however, due to lack of sequence data for numerous mycobacterial species, this assay might test positive for closely-related mycobacterial species as well. The other two assays target regions of the multi-copy insertion elements IS6110 and IS1081 that are found in the MTBC (Collins & Stephens, 1991; Eisenach, Cave, Bates, & Crawford, 1990; Klaus et al., 2010; McHugh, Newport, & Gillespie, 1997). Genomic DNA from *M. tuberculosis* H37Rv was used to create DNA standards for the qPCR assays. Ten-fold serial dilutions ranging from one to 1,000,000 copy numbers of the genome per µL were used to plot a standard curve for quantification purposes. Non-template controls (PCR-grade water) were also included on each qPCR plate. DNA extracts, extraction blanks, and non-template control were run in triplicate whereas DNA standards were run in duplicate. qPCR reactions were run in a 20 µL total volume: 10 µL of TaqMan 2X Universal MasterMix, 0.2 µL of 10mg/mL RSA, and 2 µL of sample (DNA, standard, or non-template control). Primers and probe were added at optimized concentrations as given in Housman et al. (2015). The qPCR assays were carried out on an Applied Biosystems 7900HT thermocycler with the following conditions: 50°C for 2 minutes, 95°C for 10 minutes, and 50 cycles of amplification at 95°C for 15 seconds and 60°C for 1 minute. The results were visualized using SDS 2.3. Both amplification and multicomponent plots were used to classify the replicates of the extracts as positive or negative. An extract was considered to be positive for a qPCR assay if two or more replicates out of three were positive.

qPCR assay results for extracts from samples AD281(a) and AD82 have been previously published elsewhere (Harkins et al., 2015).

In-solution hybridization gene capture: DNA extracts considered to be positive for one or more qPCR assays were converted into double-indexed libraries (Kircher, Sawyer, & Meyer, 2012; Meyer & Kircher, 2010) using 10-20 µL of extract, following the procedure given in Bos et al. (2014). Libraries were indexed using AmpliTaq Gold (Life Technologies) for 20 cycles and quantified using the DNA1000 assay on the Bioanalyzer 2100 (Agilent) and the KAPA Library Quantification kit (Kapa Biosystems). A library blank (negative control) was included in each batch of samples that underwent library preparation. All libraries, including library blanks, were target-enriched using an in-solution capture protocol at ASU. The libraries were target-enriched for five genes (Bos et al., 2014; Maricic, Whitten, & Paabo, 2010) - the *rpoB*,

gyrA, *gyrB* and *katG* genes commonly found in all mycobacterial species, and the *mtp40* gene believed to be unique to MTBC strain types, although it is not present in all (Weil, Plikaytis, Butler, Woodley, & Shinnick, 1996). Enriched libraries were amplified to a concentration of 10^{13} copies per reaction using AccuPrime Pfx DNA polymerase (Life Technologies) and quantified using the Bioanalyzer 2100 (Agilent) and the KAPA Library Quantification kit (Kapa Biosystems). The libraries were pooled at equimolar concentrations and sequenced on an Illumina MiSeq using V2 chemistry (2×150 bp) at the DNASU Sequencing Center, Tempe, USA.

The sequenced reads were trimmed and merged using SeqPrep (https://github.com/jstjohn/SeqPrep) using default parameters, except the minimum overlap for merging was modified to 11. Merged reads were mapped to the hypothetical MTBC ancestor reference genome (Comas et al., 2010) using the Burrows-Wheeler Aligner (bwa v0.7.5) (Li & Durbin, 2009). Stringent mapping parameters (-l 1000, -n 0.1) were used in an attempt to avoid reads from closely related soil-dwelling mycobacteria from mapping. Contaminating DNA from soil swelling bacteria is known to be present in archaeological tissues interred in the soil (Vågene et al., 2018). SAMtools v0.1.19 (Li et al., 2009) was used to filter the mapped reads at a minimum mapping quality threshold of Q30, remove duplicate reads and reads that map equally well to more than one position in the genome. The resulting BAM files were visually analyzed using Geneious R7 (Biomatters) to determine the percentage of the targeted genes covered at least one-fold. Samples where more than 50% of all five targeted genes were covered at least one-fold were chosen for MTBC whole-genome enrichment.

Sample AD82

Sample AD82 was previously screened for the presence of MTBC DNA via gene capture and qPCR (Bos et al., 2014; Harkins et al., 2015), but did not meet the previously set requirements for being included in whole genome capture (Bos et al., 2014). In this study we used 60 µL of extract to make a UDG treated Illumina library (Kircher et al., 2012; Meyer & Kircher, 2010) that we included in our whole-genome in-solution capture.

Probe design

Single-stranded probes for in-solution capture were designed using a computationally extrapolated ancestral genome of the *Mycobacterium tuberculosis* Complex (MTBC) (Comas et al., 2010), which was generated by reverting phylogenetically informative positions in the H37Rv reference genome (NC_000962.1) to their ancestral state, here on referred to as MTBC_anc reference. The probes were 52bp in length with 5bp tiling, yielding a set of unique 852,164 probes after the removal of duplicate and low complexity probes. The probe set was raised to 980,000 by random filling of probes. A linker sequence (5-prime CACTGCGG 3-prime) was attached to each probe sequence, resulting in 60-base probes, which were printed on a custom-designed 1 million-feature array (Agilent). The printed probes were cleaved off the array, biotinylated, and prepared for capture according to Fu et al. (2013).

MTBC whole genome in-solution capture and data evaluation

Libraries treated with uracil DNA glycosylase (UDG; ancient DNA damage removed) were generated for the five sample libraries deemed positive for ancient MTBC DNA: AD82, AD281a, AD281b, AD281c, AD386. Whole-genome in-solution capture (Fu et al., 2013) was performed for UDG treated and non-UDG treated libraries, and each library was captured separately in a single well on a 96-well plate. Negative controls were pooled equimolar and captured together in a single well. A previously published sample from which an MTBC genome had been captured (58U) (Bos et al., 2014) was included in our capture as a positive control. Prior to capture all sample libraries were amplified using Herculase to a minimum concentration of 200 ng/µL. The protocol for in-solution capture was carried out according to Fu et al. (2013) with the exception that Cot-1 DNA was excluded from the capture master-mix, therefore 7.5 µL of DNA template was added instead of 5 µL, as listed in the original protocol.

Initial paired-end sequencing (2x75bp) of the capture product was carried out on part of a HiSeq 4000 lane (Supplementary Table 9). De-indexing of the sequenced capture data was carried out using bcl2fastq (Illumina; http://support.illumina.com/downloads/bcl2fastq-conversion-software-v217.html).
The EAGER pipeline (Peltzer et al., 2016) (v.1.92.55) was used to pre-process, map and estimate damage patterns. Specifically, AdapterRemoval v. 2.2.0 was used to clip adapters and merge paired-end reads with an overlap of at least 10bp, and only merged reads were kept for downstream analyses. Captured sequence data were mapped to the MTBC_anc reference using BWA (v. 0.7.12) (Li & Durbin, 2009), where lenient parameters were used for non-UDG treated libraries (-l 16, -n 0.01, -q 37), as mismatches in the form of deamination are expected to occur in this data, and stringent parameters were used for UDG treated libraries (-l 32, -n 0.1, -q 37) where the chemical lesions caused by deamination had been removed. All negative controls were mapped to the MTBC_anc with lenient parameters, regardless of library treatment. Duplicate removal was executed using DeDup v. 0.12.2 and the assessment of DNA damage patterns with mapDamage 2.0 (Jonsson, Ginolhac, Schubert, Johnson, & Orlando, 2013).

Metagenomic analysis of capture and shotgun data

Shotgun data for the UDG treated libraries and non-UDG treated libraries (excluding AD281a and AD281c) were generated to evaluate the capture efficiency and the metagenomic profiles of our samples. Libraries were paired-end (2x75bp) shotgun sequenced on part of a HiSeq 4000 lane or single-end (1x75bp) sequenced as part of a Next-Seq 500 run (Supplementary Table 7, 9). Pre-processing and mapping of the data was carried out as described above, using stringent mapping parameters for the UDG treated shotgun data. Capture efficiencies were calculated based on the number of quality-filtered reads before duplicate removal.

The pre-processed shotgun and capture data were analyzed using the Megan ALignment Tool (MALT) (version 0.3.8) (Vågene et al., 2018). The shotgun (non-

UDG) and capture (UDG) data previously generated for the three published Peruvian MTBC genomes (Bos et al., 2014) were also processed with MALT. These data were analyzed using a database comprised of all complete genomes in the NCBI Nucleotide (nt) database downloaded from ftp://ftp-trace.ncbi.nih.gov/blast/db/FASTA/ on the 7th December 2016 that was created using *malt build* (v. 0.3.8). The purpose of this was to assess the amount of non-MTBC DNA in our shotgun and capture data, particularly with regard to the amount of non-MTBC mycobacterial DNA. Two MALT runs were performed. The first using 85 as the "minimum percent identity" parameter (--minPercentIdentity), this is a more sensitive alignment criteria, mimicking the stringency of DNA-binding during hybridization capture. The second using 95 as the "minimum percent identity", allowing fewer mismatches in the reads aligned to the database. BlastN mode and SemiGlobal alignment were applied. All other parameters were set to default with the exception of the minimum support parameter (--minSupport) that was set to 1 and a top percent value (--topPercent) of 1. MEGAN6 v.6.12.3 (Huson et al., 2016) was used to view the MALT results. Taxon tables of the MALT results for the shotgun and capture data is shown in Supplementary Tables 4, 5, 6.

The MALT results for the UDG treated shotgun and capture data were used to generate pie-charts illustrating the ratio of MTBC versus non-MTBC mycobacterial data present in our samples before and after capture (Supplementary Fig. 1; Supplementary Table 3, 4). The MALT results for the non-UDG treated shotgun data for our samples and the dataset previously generated for the published Peruvian samples (Bos et al., 2014) were used to compare ratios investigating the number of MTBC versus non-MTBC mycobacterial reads between the two datasets (Supplementary Table 5, 6, 8).

Deep sequencing, genome reconstruction, and phylogenetic analysis

Based on the results from the initial sequencing of the capture products, we sequenced the captured UDG treated libraries deeper for samples AD82, AD281a, AD281b and AD386 to facilitate full genome reconstruction. Two libraries, AD281a and AD281b, originate from a vertebrae and a rib respectively, from the same individual. An additional captured library generated from another rib (AD281c) from this same individual was not sequenced further, because it was determined to contain higher levels of contaminating environmental DNA and lower amounts of endogenous ancient MTBC DNA (see Supplementary Figure 1, Supplementary Tables 3, 4). Deeper sequencing was carried out as part of several single-end HiSeq 4000 (1x75bp) sequencing runs.

The additional sequencing data was adapter-clipped and quality filtered, excluding reads shorter than 30 bp. This data was combined with the merged and quality filtered paired-end data. The merged data were then treated as single-end data. MarkDuplicates, which consider only the 5-prime end of reads, was used for duplicate removal, because the true 3-prime end is often not observed in single-end sequences. The combined data for each library were in all other respects subjected to the same mapping and variant

calling pipeline as described above, where stringent parameters were used during BWA mapping to the MTBC_anc reference. Variant calling was also executed with the EAGER pipeline during the processing of the deeper sequenced data using the GATK UnifiedGenotyper (DePristo et al., 2011) where the 'EMIT_ALL_SITES' function was activated, providing a call for all variant or non-variant bases in the *vcf* file.

The dataset of ancient and modern MTBC genomes compiled by Bos et al. (2014) was used in this study for phylogenetic comparison, with the exception of the ancient Hungarian genome that is a composite of two strains (Chan et al., 2013). The same *vcf* files previously generated for the 259 modern genomes were used in this study (Bos et al., 2014). The capture data for the three ancient Peruvian genomes were reprocessed using our pipeline in the same manner as our ancient data. Two recently published modern animal-associated MTBC genomes, *M. microti* strain 12 and one *M. mungi*, were included in our dataset (Alexander, Larsen, Robbe-Austerman, Stuber, & Camp, 2016; Zhu et al., 2016). The assembled genome for '*M. microti* strain 12' (CP010333.1) was downloaded and converted into artificial read data (100bp reads with 1bp tiling), and was subsequently mapped to the MTBC_anc reference as single-end data. The paired-end *M. mungi* BM22813 (sample accession no. SRS1434640) data was concatenated and also mapped as single-end data.

A SNP alignment consisting of homozygous positions for our four ancient genomes (AD82, AD281a, AD281b, AD386), 261 ancient and modern genomes from Bos *et al.* (2014), *M. microti* strain 12 and *M. mungi* BM22813 was generated using MultiVCFanalyzer (v. 0.87) (Bos et al., 2014) (https://github.com/alexherbig/MultiVCFAnalyzer). Homozygous positions were typed at positions where 90% or more of the reads covering it were in agreement and where a minimum of 3 reads were covering the position. The resulting SNP alignment file was used to construct an initial Maximum Parsimony tree with 500 replicates using MEGA6, all missing and ambiguous data were excluded.

The genomes generated from AD281a and AD281b appeared to be identical in the initial phylogeny. This was confirmed by visual inspection of the 492 variant positions shared between the AD281a and AD281b genomes where a base call could be made for both genomes. The sequence data from the two libraries were combined for all remaining analyses, and are collectively referred to as AD281.

A new SNP alignment was generated, now including the combined AD281 library data yielding a higher average coverage for this sample. SNP calling with MultiVCFAnalyzer was repeated and variant calls were made when a minimum of 5 reads covered a position. This SNP alignment file was used to build phylogenies using the Neighbour joining and Maximum Parsimony methods in MEGA6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013), and Maximum Likelihood method using RAxML v.8 (Stamatakis, 2014) (Fig. 2; Supplementary Figures 3, 4). All positions with missing data were excluded from the dataset.

The allelic frequencies of positions that appear as 'heterozygous' were investigated due to the longer branch lengths observed for our Colombian and Peruvian genomes compared to the previously published contemporaneous Peruvian genomes (Bos 2014). Heterozygous variants were called at positions with a minimum of 5-fold coverage and where all positions with a SNP allele frequency between 10-90% were typed as heterozygous. All ancient *M. pinnipedii* genomes were investigated and compared. The distribution of SNP allele frequencies for all six ancient samples is shown in Supplementary Figure 5.

Deamination rates analysis

The deamination rates observed for the non-UDG treated MTBC capture sequences were considered to be low, especially for AD82. In an attempt to exclude cross mapping reads from closely related soil dwelling mycobacteria from artificially lowering the true deamination rates we clipped 4bp off the 3-prime and the 5-prime ends of all reads in the non-UDG data to remove the bases most likely affected by deamination for samples AD82, AD281 (combined AD281a and AD281b) and AD386. The clipped reads were mapped to the MTBC_anc reference with BWA using stringent parameters (-l 32, -n 0.1, -q 37). The IDs for the stringently mapped reads were used to extract the corresponding non-clipped reads from the non-UDG treated sequence data, mapping with BWA to the MTBC_anc reference with lenient parameters and mapDamage 2.0 were subsequently repeated.

The non-UDG MTBC whole genome capture data was also mapped to the human genome reference (hg19) using BWA with lenient parameters (-l 16, -n 0.01, -q 37). Deamination patterns were determined using mapDamage 2.0 (Jonsson et al., 2013) for the human reads that were co-enriched during the MTBC capture.

SNP analysis of protein coding genes

SNP analysis was carried out for a subset of genomes, selected from the full dataset that was used in phylogenetic analyses. This subset included all six ancient and six modern *M. pinnipedii* genomes, two *M. microti* genomes and three human adapted *M. tuberculosis* Lineage 6 (L6) genomes, which functioned as outgroups during SNP filtering (Supplementary Table 13). A SNP table of homozygous variant positions called at a minimum of 5-fold coverage where 90% or more reads supported a base call was generated using MultiVCFanalyzer. The variant calls were annotated using *SnpEff* (v. 3.1) (Cingolani et al., 2012) with standard parameters, except that the region for upstream or downstream gene regions where variants were called was limited to 100 bp (-ud 100). A custom annotation database for non-protein-coding and protein-coding gene in the H37Rv genome was used (Bos et al., 2014), because this was the genome for which the design of the hypothetical ancestral MTBC genome based on (Comas et al., 2010). SNP positions occurring on shared branches within the *M. pinnipedii* clade were investigated.

Table 1. Genome mapping statistics for UDG treated capture libraries

Sample ID	# processed reads before mapping	# Unique quality filtered mapped reads	Endogenous DNA (%) - quality filtered reads	Mean Fold Coverage	% of Genome Covered at least 5-fold	Median fragment length (bp)
AD82	449,538,686	1,002,140	0.71	14.95	80.31	75
AD281	170,865,535	1,003,454	1.30	15.34	80.81	76
AD386	328,622,471	736,677	1.66	10.80	83.07	71


1
Northern
pinniped range
Southern
pinniped range
2
1. SOUTH WEST BOGOTÁ, COLOMBIA
OLARTE
LAS DELICIAS
MADELENA
Río Tunjuelo
GALICIA
FATIMA
CIUDAD BOLIVAR
TUN JUELITO
CANDELARIA LA NUEVA
0
500m
2. PERU
MOQUEGUA, M6: ESTUQUIÑA
EL YARAL
Río Osmore
EL ALGODONAL
ILO
CHIRIBAYA ALTA
0
10km
Archaeological site, this study
Archaeological site, Bos et al. 2014
Modern city

Figure 1. Maps indicating the modern pinniped range and locations of archaeological sites that have yielded ancient *M. pinnipedii* genomes. The topmost panel shows a map of South and Central America illustrating the southern and northern pinniped range. The ranges of *Arctocephalus australis* (Cárdenas-Alayza, Oliveira, & Crespo, 2016), *Arctocephalus galapagoensis* (F. Trillmich, 2015), *Arctocephalus philippii* (Aurioles-Gamboa, 2015), *Mirounga leonina* (Hofmeyr, 2015), *Otaria byronia* (Cárdenas-Alayza, Crespo, & Oliveira, 2016) and *Zalophus wollebaeki* (F. Trillmich, 2015) are overlaid for the southern hemisphere and *Zalophus californianus* (Aurioles-Gamboa & Hernández-Camacho, 2015) is shown for the northern hemisphere. All range data is from the IUCN red list of threatened species (http://www.iucnredlist.org/); the middle panel shows a zoomed in view of the locations for the Colombian sites; the bottom panel shows zoomed in view of the Osmore River Valley in Peru. The locations of the sites that yielded *M. pinnipedii* genomes in this study and the published study by Bos et al. (2014) are shown. Graphic produced by Michelle O'Reilly.

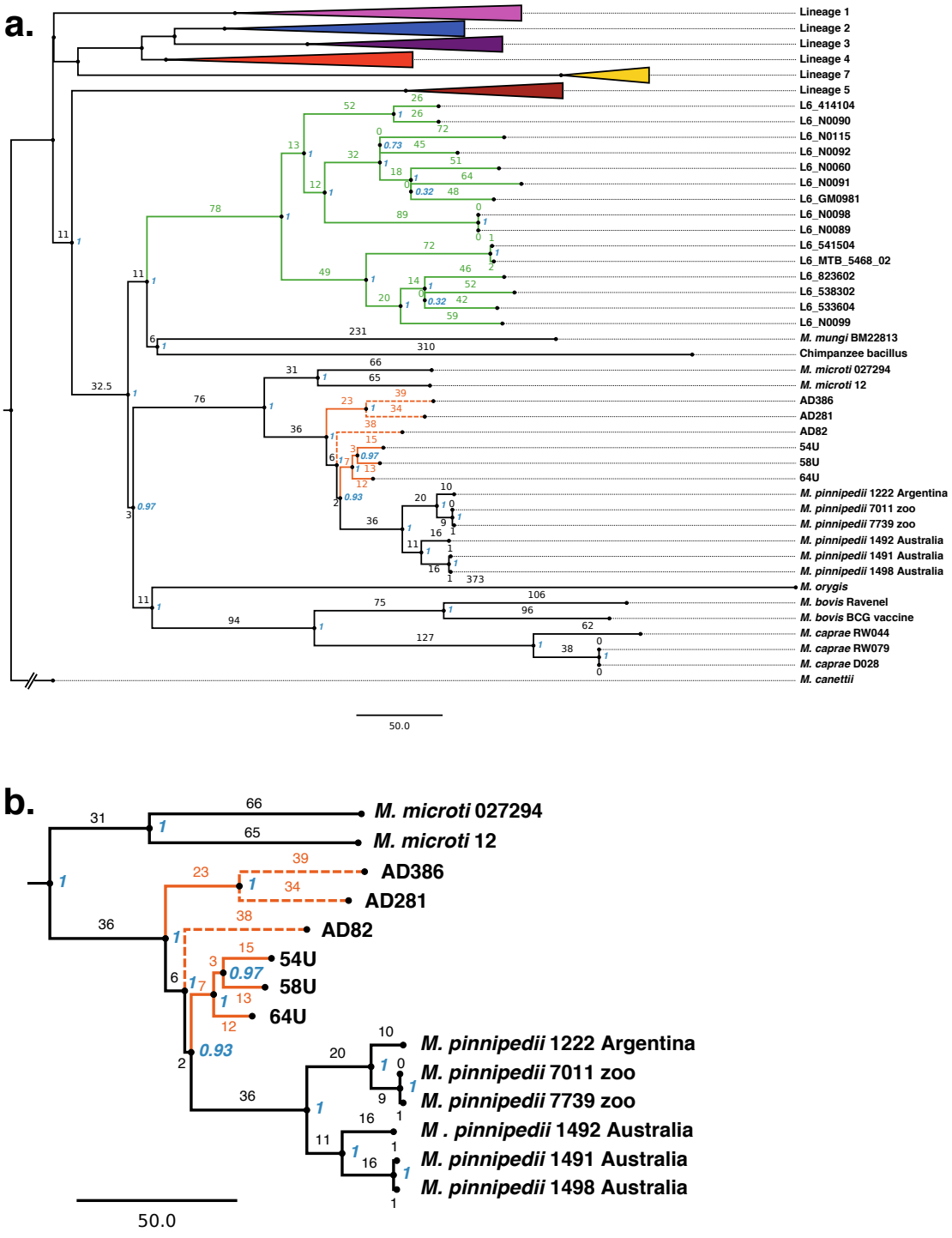


Figure 2. Maximum Parsimony MTBC phylogeny. The tree was constructed using the full dataset of 266 genomes, including the six ancient genomes that are highlighted in orange. The tree was constructed by excluding all missing and ambiguous data, using 500 bootstrap replicates and is based on 14,263 positions out of a possible 44,273. Bootstrap support (blue) and branch lengths are marked. Our three ancient genomes fall together with other ancient Peruvian genomes within the *M. pinnipedii* clade. a) the full MTBC phylogenetic tree where human-adapted lineages 1-5 and 7 have been collapsed, b) a zoomed in view of the *M. microti* and *M. pinnipedii* clades. Genomes AD82, AD281 and AD386 display longer branch lengths compared to the contemporaneous

ancient Peruvian genomes (AD54, AD58, AD64), and are marked by dashed branches. These longer branches are due to contaminant reads from closely related organisms causing false-positive variant calls. The contaminating reads were co-enriched during the capture and persist in the genome alignments despite the application of stringent mapping criteria.

References

Achtman, M. (2016). How old are bacterial pathogens? *Proc Biol Sci, 283*(1836). doi:10.1098/rspb.2016.0990

Alexander, K. A., Larsen, M. H., Robbe-Austerman, S., Stuber, T. P., & Camp, P. M. (2016). Draft Genome Sequence of the Mycobacterium tuberculosis Complex Pathogen M. mungi, Identified in a Banded Mongoose (Mungos mungo) in Northern Botswana. *Genome Announc, 4*(4). doi:10.1128/genomeA.00471-16

Allen, A. R. (2017). One bacillus to rule them all? - Investigating broad range host adaptation in Mycobacterium bovis. *Infect Genet Evol, 53*, 68-76. doi:10.1016/j.meegid.2017.04.018

Allison, M. J., Gerszten, E., Munizaga, J., Santoro, C., & Mendoza, D. (1981). Tuberculosis in Pre‐Columbian Andean Populations. In J. E. Buikstra (Ed.), *Prehistoric Tuberculosis in the Americas* (pp. 49-61). Evanston, Illinois: Northwestern University Archaeological Program.

Ameni, G., Vordermeier, M., Firdessa, R., Aseffa, A., Hewinson, G., Gordon, S. V., & Berg, S. (2011). Mycobacterium tuberculosis infection in grazing cattle in central Ethiopia. *Vet J, 188*(3), 359-361. doi:10.1016/j.tvjl.2010.05.005

Andam, C. P., Worby, C. J., Chang, Q., & Campana, M. G. (2016). Microbial Genomics of Ancient Plagues and Outbreaks. *Trends Microbiol, 24*(12), 978-990. doi:10.1016/j.tim.2016.08.004

Aurioles-Gamboa, D. (2015). *Arctocephalus philippii*. The IUCN Red List of Threatened Species 2015: e.T2059A61953525.

Aurioles-Gamboa, D., & Hernández-Camacho, J. (2015). *Zalophus californianus*. The IUCN Red List of Threatened Species 2015: e.T41666A45230310. . Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2015-4.RLTS.T41666A45230310.en

Bastida, R., Loureiro, J., Quse, V., Bernardelli, A., Rodriguez, D., & Costa, E. (1999). Tuberculosis in a wild subantarctic fur seal from Argentina. *J Wildl Dis, 35*(4), 796-798. doi:10.7589/0090-3558-35.4.796

Benson, E. P. (2012). *The Worlds of the Moche on the North Coast of Peru*. Austin, Texas: University of Texas Press.

Boardman, W. S., Shephard, L., Bastian, I., Globan, M., Fyfe, J. A., Cousins, D. V., . . . Woolford, L. (2014). Mycobacterium pinnipedii tuberculosis in a free-ranging Australian fur seal (Arctocephalus pusillus doriferus) in South Australia. *J Zoo Wildl Med, 45*(4), 970-972. doi:10.1638/2014-0054.1

Boos, W., & Shuman, H. (1998). Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation. *Microbiol Mol Biol Rev, 62*(1), 204-229.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., . . . Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature, 514*(7523), 494-497. doi:10.1038/nature13591

Brites, D., & Gagneux, S. (2017). The Nature and Evolution of Genomic Diversity in the Mycobacterium tuberculosis Complex. In S. Gagneux (Ed.), *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*. Springer International Publishing AG: Springer International Publishing.

Brites, D., Loiseau, C., Menardo, F., Borrell, S., Boniotti, M. B., Warren, R., . . . Gagneux, S. (2018). A New Phylogenetic Framework for the Animal-adapted Mycobacterium tuberculosis Complex. *bioRxiv*.

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., . . . Cole, S. T. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci U S A, 99*(6), 3684-3689. doi:10.1073/pnas.052548299

Buikstra, J. E., & Williams, S. R. (1991). Tuberculosis in the Americas: Current perspectives. In D. Ortner & A. C. Aufderheide (Eds.), *Human palaeopathology: Current syntheses and future options* (pp. 161-172). Washington, D.C.: Smithsonian Institution Press.

Cardenas, P. P., Gandara, C., & Alonso, J. C. (2014). DNA double strand break end-processing and RecA induce RecN expression levels in Bacillus subtilis. *DNA Repair (Amst), 14*, 1-8. doi:10.1016/j.dnarep.2013.12.001

Cárdenas-Alayza, S., Crespo, E., & Oliveira, L. (2016). *Otaria byronia*. The IUCN Red List of Threatened Species 2016: e.T41665A61948292. Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T41665A61948292.en

Cárdenas-Alayza, S., Oliveira, L., & Crespo, E. (2016). *Arctocephalus australis*. The IUCN Red List of Threatened Species 2016: e.T2055A45223529. . Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T2055A45223529.en

Chan, J. Z.-M., Sergeant, M. J., Lee, O. Y.-C., Minnikin, D. E., Besra, G. S., Pap, I., . . . Pallen, M. J. (2013). Metagenomic analysis of tuberculosis in a mummy. *The New England journal of medicine, 369*, 289-290. doi:10.1056/NEJMc1302295

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*(2), 80-92. doi:10.4161/fly.19695

Clark, S., Hall, Y., & Williams, A. (2014). Animal models of tuberculosis: Guinea pigs. *Cold Spring Harb Perspect Med, 5*(5), a018572. doi:10.1101/cshperspect.a018572

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., . . . Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature, 393*(6685), 537-544. doi:10.1038/31159

Collins, D. M., & Stephens, D. M. (1991). Identification of an insertion sequence, IS1081, in Mycobacterium bovis. *FEMS Microbiol Lett, 67*(1), 11-15.

Comas, I., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., . . . Gagneux, S. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet, 42*(6), 498-503. doi:10.1038/ng.590

Coscolla, M., & Gagneux, S. (2014). Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin Immunol, 26*(6), 431-444. doi:10.1016/j.smim.2014.09.012

Coscolla, M., Lewin, A., Metzger, S., Maetz-Rennsing, K., Calvignac-Spencer, S., Nitsche, A., . . . Leendertz, F. H. (2013). Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. *Emerg Infect Dis, 19*(6), 969-976. doi:10.3201/eid1906.121012

Cousins, D. V., Bastida, R., Cataldi, A., Quse, V., Redrobe, S., Dow, S., . . . Bernardelli, A. (2003). Tuberculosis in seals caused by a novel member of the Mycobacterium tuberculosis complex: Mycobacterium pinnipedii sp. nov. *Int J Syst Evol Microbiol, 53*(Pt 5), 1305-1314. doi:10.1099/ijs.0.02401-0

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., . . . Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle

Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad Sci U S A, 110*(39), 15758-15763. doi:10.1073/pnas.1314445110

de Amorim, D. B., Casagrande, R. A., Alievi, M. M., Wouters, F., De Oliveira, L. G., Driemeier, D., . . . Ferreira-Neto, J. S. (2014). Mycobacterium pinnipedii in a stranded South American sea lion (Otaria byronia) in Brazil. *J Wildl Dis, 50*(2), 419-422. doi:10.7589/2013-05-124

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi:10.1038/ng.806

Donnan, C. B. (1978). *Moche Art of Peru: Pre-Columbian Symbolic Communication*. California: Museum of Cultural History, University of California.

Eisenach, K. D., Cave, M. D., Bates, J. H., & Crawford, J. T. (1990). Polymerase chain reaction amplification of a repetitive DNA sequence specific for Mycobacterium tuberculosis. *J Infect Dis, 161*(5), 977-981.

Erental, A., Kalderon, Z., Saada, A., Smith, Y., & Engelberg-Kulka, H. (2014). Apoptosis-like death, an extreme SOS response in Escherichia coli. *MBio, 5*(4), e01426-01414. doi:10.1128/mBio.01426-14

Forshaw, D., & Phelps, G. R. (1991). Tuberculosis in a captive colony of pinnipeds. *J Wildl Dis, 27*(2), 288-295. doi:10.7589/0090-3558-27.2.288

Francis, J. M. (2007). *Invading Colombia*

Spanish Accounts of the Gonzalo Jiménez de Quesada Expedition of Conquest: Pennsylvania State University Press.

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H. A., Kelso, J., & Paabo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A, 110*(6), 2223-2227. doi:10.1073/pnas.1221359110

Gande, R., Gibson, K. J., Brown, A. K., Krumbach, K., Dover, L. G., Sahm, H., . . . Eggeling, L. (2004). Acyl-CoA carboxylases (accD2 and accD3), together with a unique polyketide synthase (Cg-pks), are key to mycolic acid biosynthesis in Corynebacterianeae such as Corynebacterium glutamicum and Mycobacterium tuberculosis. *J Biol Chem, 279*(43), 44847-44857. doi:10.1074/jbc.M408648200

Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., . . . Guilhot, C. (2014). Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A, 111*(31), 11491-11496. doi:10.1073/pnas.1406693111

Gupta, H. K., Shrivastava, S., & Sharma, R. (2017). A Novel Calcium Uptake Transporter of Uncharacterized P-Type ATPase Family Supplies Calcium for Cell Surface Integrity in Mycobacterium smegmatis. *MBio, 8*(5). doi:10.1128/mBio.01388-17

Harkins, K. M., Buikstra, J. E., Campbell, T., Bos, K. I., Johnson, E. D., Krause, J., & Stone, A. C. (2015). Screening ancient tuberculosis with qPCR: challenges and opportunities. *Philos Trans R Soc Lond B Biol Sci, 370*(1660), 20130622. doi:10.1098/rstb.2013.0622

Harkins, K. M., & Stone, A. C. (2015). Ancient pathogen genomics: insights into timing and adaptation. *J Hum Evol, 79*, 137-149. doi:10.1016/j.jhevol.2014.11.002

Hattori, M., Iwase, N., Furuya, N., Tanaka, Y., Tsukazaki, T., Ishitani, R., . . . Nureki, O. (2009). Mg(2+)-dependent gating of bacterial MgtE channel underlies Mg(2+) homeostasis. *EMBO J, 28*(22), 3602-3612. doi:10.1038/emboj.2009.288

Hofmeyr, G. J. G. (2015). *Mirounga leonina*. The IUCN Red List of Threatened Species 2015: e.T13583A45227247. Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2015-4.RLTS.T13583A45227247.en

Houben, R. M., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med, 13*(10), e1002152. doi:10.1371/journal.pmed.1002152

Housman, G., Malukiewicz, J., Boere, V., Grativol, A. D., Pereira, L. C., Silva Ide, O., . . . Stone, A. C. (2015). Validation of qPCR Methods for the Detection of Mycobacterium in New World Animal Reservoirs. *PLoS Negl Trop Dis, 9*(11), e0004198. doi:10.1371/journal.pntd.0004198

Huang, C. S., Pedersen, B. P., & Stokes, D. L. (2017). Crystal structure of the potassium-importing KdpFABC membrane complex. *Nature, 546*(7660), 681-685. doi:10.1038/nature22970

Huson, D. H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., . . . Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol, 12*(6), e1004957. doi:10.1371/journal.pcbi.1004957

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics, 29*(13), 1682-1684. doi:10.1093/bioinformatics/btt193

Jurczynski, K., Lyashchenko, K. P., Gomis, D., Moser, I., Greenwald, R., & Moisson, P. (2011). Pinniped tuberculosis in Malayan tapirs (Tapirus indicus) and its transmission to other terrestrial mammals. *J Zoo Wildl Med, 42*(2), 222-227. doi:10.1638/2009-0207.1

Keyamura, K., Sakaguchi, C., Kubota, Y., Niki, H., & Hishida, T. (2013). RecA protein recruits structural maintenance of chromosomes (SMC)-like RecN protein to DNA double-strand breaks. *J Biol Chem, 288*(41), 29229-29237. doi:10.1074/jbc.M113.485474

Kiers, A., Klarenbeek, A., Mendelts, B., Van Soolingen, D., & Koeter, G. (2008). Transmission of Mycobacterium pinnipedii to humans in a zoo with marine mammals. *Int J Tuberc Lung Dis, 12*(12), 1469-1473.

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

Klaus, H. D., Wilbur, A. K., Temple, D. H., Buikstra, J. E., Stone, A. C., Fernandez, M., . . . Tam, M. E. (2010). Tuberculosis on the north coast of Peru: skeletal and molecular paleopathology of late pre-Hispanic and postcontact mycobacterial disease. *Journal of Archaeological Science, 37*(10), 2587-2597. doi:10.1016/j.jas.2010.05.019

Kouzminova, E. A., Rotman, E., Macomber, L., Zhang, J., & Kuzminov, A. (2004). RecA-dependent mutants in Escherichia coli reveal strategies to avoid chromosomal fragmentation. *Proc Natl Acad Sci U S A, 101*(46), 16262-16267. doi:10.1073/pnas.0405943101

Kurella, D. (1998). The Muisca, Chiefdoms in Transition. In E. M. Redmond (Ed.), *Chiefdoms and Chieftaincy in the Americas* (pp. 189-216): University Press of Florida.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Loeffler, S. H., de Lisle, G. W., Neill, M. A., Collins, D. M., Price-Carter, M., Paterson, B., & Crews, K. B. (2014). The seal tuberculosis agent, Mycobacterium pinnipedii, infects domestic cattle in New Zealand: epidemiologic factors and DNA strain typing. *J Wildl Dis, 50*(2), 180-187. doi:10.7589/2013-09-237

Malhotra, S., Vedithi, S. C., & Blundell, T. L. (2017). Decoding the similarities and differences among mycobacterial species. *PLoS Negl Trop Dis, 11*(8), e0005883. doi:10.1371/journal.pntd.0005883

Malone, K. M., & Gordon, S. V. (2017). Mycobacterium tuberculosis Complex Members Adapted to Wild and Domestic Animals. In S. Gagneux (Ed.), *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control* (pp. 135-154). Cham: Springer International Publishing.

Maricic, T., Whitten, M., & Paabo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One, 5*(11), e14004. doi:10.1371/journal.pone.0014004

McHugh, T. D., Newport, L. E., & Gillespie, S. H. (1997). IS6110 homologs are present in multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. *J Clin Microbiol, 35*(7), 1769-1771.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc, 2010*(6), pdb prot5448. doi:10.1101/pdb.prot5448

Nugent, G. (2011). Maintenance, spillover and spillback transmission of bovine tuberculosis in multi-host wildlife complexes: a New Zealand case study. *Vet Microbiol, 151*(1-2), 34-42. doi:10.1016/j.vetmic.2011.02.023

Odsbu, I., & Skarstad, K. (2014). DNA compaction in the early part of the SOS response is dependent on RecN and RecA. *Microbiology, 160*(Pt 5), 872-882. doi:10.1099/mic.0.075051-0

Olea-Popelka, F., Muwonge, A., Perera, A., Dean, A. S., Mumford, E., Erlacher-Vindel, E., . . . Fujiwara, P. I. (2017). Zoonotic tuberculosis in human beings caused by Mycobacterium bovis-a call for action. *Lancet Infect Dis, 17*(1), e21-e25. doi:10.1016/S1473-3099(16)30139-6

Peltzer, A., Jager, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biol, 17*(1), 60. doi:10.1186/s13059-016-0918-z

Pesciaroli, M., Alvarez, J., Boniotti, M. B., Cagiola, M., Di Marco, V., Marianelli, C., . . . Pasquali, P. (2014). Tuberculosis in domestic animal species. *Res Vet Sci, 97 Suppl*, S78-85. doi:10.1016/j.rvsc.2014.05.015

Pugsley, A. P., & Dubreuil, C. (1988). Molecular characterization of malQ, the structural gene for the Escherichia coli enzyme amylomaltase. *Mol Microbiol, 2*(4), 473-479.

Reinhard, K., & Urban, O. (2003). Diagnosing ancient diphyllobothriasis from Chinchorro mummies. *Memorias Do Instituto Oswaldo Cruz, 98*, 191-193. doi:Doi 10.1590/S0074-02762003000900028

Rice, D. S., Conrad, G. W., & Buikstra, J. (1990). Investigaciones en Estuquiña: Descripciones Preliminares, 1985-1986. In L. K. Watanabe, M. E. Moseley, & F. Cabieses (Eds.), *Trabajos Anqueologicos en Moquegua, Peru* (Vol. 3, pp. 39-

93). Peru: Programa Contisuyu del Museuo Peruana de Sciences de la Salud, Southern Peru Copper Corporation.
Rivero, A., Marquez, M., Santos, J., Pinedo, A., Sanchez, M. A., Esteve, A., . . . Martin, C. (2001). High rate of tuberculosis reinfection during a nosocomial outbreak of multidrug-resistant tuberculosis caused by Mycobacterium bovis strain B. *Clin Infect Dis, 32*(1), 159-161. doi:10.1086/317547
Roberts, C. A., & Buikstra, J. E. (2003). *The Bioarchaeology of Tuberculosis: A Global View on a Reemerging Disease*: University Press of Florida.
Salo, W. L., Aufderheide, A. C., Buikstra, J., & Holcomb, T. A. (1994). Identification of Mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy. *Proc Natl Acad Sci U S A, 91*(6), 2091-2094.
Sander, P., Papavinasasundaram, K. G., Dick, T., Stavropoulos, E., Ellrott, K., Springer, B., . . . Bottger, E. C. (2001). Mycobacterium bovis BCG recA deletion mutant shows increased susceptibility to DNA-damaging agents but wild-type survival in a mouse infection model. *Infect Immun, 69*(6), 3562-3568. doi:10.1128/IAI.69.6.3562-3568.2001
Sato, Y., Okamoto-Shibayama, K., & Azuma, T. (2015). Glucose-PTS Involvement in Maltose Metabolism by Streptococcus mutans. *Bull Tokyo Dent Coll, 56*(2), 93-103. doi:10.2209/tdcpublication.56.93
Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., & Paabo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One, 7*(3), e34131. doi:10.1371/journal.pone.0034131
Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jager, G., Bos, K. I., . . . Krause, J. (2013). Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science, 341*(6142), 179-183. doi:10.1126/science.1238286
Sharratt, N. (2017). Steering Clear of the Dead: Avoiding Ancestors in the Moquegua Valley, Peru. *American Anthropologist, 119*(4), 645-661.
Skoglund, P., & Reich, D. (2016). A genomic view of the peopling of the Americas. *Curr Opin Genet Dev, 41*, 27-35. doi:10.1016/j.gde.2016.06.016
Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033
Stone, A. C., Wilbur, A. K., Buikstra, J. E., & Roberts, C. A. (2009). Tuberculosis and leprosy in perspective. *Am J Phys Anthropol, 140 Suppl 49*, 66-94. doi:10.1002/ajpa.21185
Swenson, E. R. (2006). Competitive Feasting, Religious Pluralism and Decentralized Power in the Late Moche Period. In W. H. Isbell & H. Silverman (Eds.), *Andean Archaeology III: North and South* (pp. 112-142). Boston, MA: Springer US.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol, 30*(12), 2725-2729. doi:10.1093/molbev/mst197
Trillmich, F. (2015). *Arctocephalus galapagoensis*. The IUCN Red List of Threatened Species 2015: e.T2057A45223722. Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2015-2.RLTS.T2057A45223722.en
Trillmich, F. (2015). *Zalophus wollebaeki*. The IUCN Red List of Threatened Species 2015: e.T41668A45230540. Retrieved from http://dx.doi.org/10.2305/IUCN.UK.2015-2.RLTS.T41668A45230540.en
Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., . . . Krause, J. (2018). Salmonella enterica genomes from victims of a major

sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution*. doi:10.1038/s41559-017-0446-6

Weil, A., Plikaytis, B. B., Butler, W. R., Woodley, C. L., & Shinnick, T. M. (1996). The mtp40 gene is not present in all strains of Mycobacterium tuberculosis. *J Clin Microbiol, 34*(9), 2309-2311.

Weiss, S. C., Skerra, A., & Schiefner, A. (2015). Structural Basis for the Interconversion of Maltodextrins by MalQ, the Amylomaltase of Escherichia coli. *J Biol Chem, 290*(35), 21352-21364. doi:10.1074/jbc.M115.667337

World Health Organization. (2015). WHO estimates of the global burden of foodborne diseases. Retrieved from http://www.who.int/foodsafety/areas_work/foodborne-diseases/ferg/en/

World Health Organization. (2017a). *Global tuberculosis report 2017*. Retrieved from Geneva, Switzerland:

World Health Organization. (2017b). Tuberculosis. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/

Wu, B., Song, J., & Beitz, E. (2010). Novel channel enzyme fusion proteins confer arsenate resistance. *J Biol Chem, 285*(51), 40081-40087. doi:10.1074/jbc.M110.184457

Zachariah, A., Pandiyan, J., Madhavilatha, G. K., Mundayoor, S., Chandramohan, B., Sajesh, P. K., . . . Mikota, S. K. (2017). Mycobacterium tuberculosis in Wild Asian Elephants, Southern India. *Emerg Infect Dis, 23*(3), 504-506. doi:10.3201/eid2303.161741

Zhu, L., Zhong, J., Jia, X., Liu, G., Kang, Y., Dong, M., . . . Chen, F. (2016). Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res, 44*(2), 730-743. doi:10.1093/nar/gkv1498

Acknowledgements: This work was supported by the Max Planck Society (J.K.), the European Research Council (ERC) starting grant APGREID (to J.K.), Social Sciences and Humanities Research Council of Canada postdoctoral fellowship grant 756-2011-501 (to K.I.B.), National Science Foundation grants BCS-1063939 (to A.C.S and J.E.B) and BCS-1515163 (to A.C.S, J.E.B and M.S.R). The Wenner Gren Foundation for Anthropological Research (to A.C.S). We are grateful to Guido Brandt for assistance with laboratory work, and to Michelle O'Reilly for graphical support.

Author Contributions: A.C.S., J.E.B, J.K. and K.I.B conceived the investigation. A.H., K.I.B., T.H., Å.J.V., A.C.S., and J.K. designed experiments. J.E.B., F.C.A., L.P.L and J.A. identified samples for analyses and provided archaeological information. Å.J.V., T.H., K.I.B. and K.M.H. performed laboratory work. Å.J.V., T.H., A.H., M.S.R. and K.I.B. performed analyses. Å.J.V. and K.I.B. wrote the manuscript with contributions from all co-authors.

Competing financial interests: The authors declare no competing financial interests.

Geographically dispersed zoonotic tuberculosis in pre-contact New World human populations

SUPPLEMENTARY INFORMATION

Authors: Åshild J. Vågene[1,2,†], Tanvi Honap[3,†,§], Kelly M. Harkins[4,‡], Michael S. Rosenberg[3], Felipe Cárdenas-Arroyo[5], Laura Paloma Leguizamón[5], Judith Arnett[4,6], Jane E. Buikstra[4] Alexander Herbig[1,2], Anne C. Stone [4,7,8,*], Kirsten I. Bos[1,2,*] and Johannes Krause[1,2,*]

Affiliations:
[1]Max Planck Institute for the Science of Human History, Jena, Germany
[2]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany
[3]School of Life Sciences, Arizona State University, Tempe, Arizona, USA
[4]School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, USA
[5]Colombian Institute of Anthropology and History (ICANH), Bogotá, Colombia
[6]University of the Andes, School of Medicine, Colombia
[7]Center for Evolution and Medicine, Arizona State University, Tempe, Arizona, USA
[8]Institute of Human Origins, Arizona State University, Tempe, Arizona, USA

[†]These authors contributed equally to this work
*Corresponding authors
[§]Current address for T.H.: Department of Anthropology, University of Oklahoma, Norman, Oklahoma, USA
[‡] Current address for K.H.: Claret Bioscience LLC., Santa Cruz, California, USA

CONTENTS:

Refer to CD for Supplementary Tables 3, 5, 6, 9, 12 and 13

Supplementary Section 1: Archaeological Provenance and Context

The four samples that yielded data for the reconstruction of three MTBC genomes represent one individual (AD82) buried at the site of Moquegua, M6: Estuquiña located in the upper Moquegua Valley in Peru, and two individuals buried at separate sites in the modern city of Bogotá in Colombia, named Las Delicias (AD281) and Candelaria La Nueva (AD386).

Moquegua, M6: Estuquiña, Peru

The site of Moquegua, M6: Estuquiña is situated on a plateau above the Osmore River Drainage, ca. 2000 m above sea level (Rice, Conrad, & Buikstra, 1990), and is located in the upper Moquegua Valley which is separated from the coast by a mountain barrier (Sharratt, 2017). This site was occupied during the Terminal Middle Horizon (1000-1250 C.E.) and Late Intermediate Period (1250-1470 C.E.) (Sharratt, 2017; Williams, 1990). Excavations at Moquegua, M6: Estuquiña were carried out in 1985. They were led by the general project director Donald S. Rice (Southern Illinois University) and Jane E. Buikstra (Arizona State University) who directed the mortuary excavations. Moquegua, M6: Estuquiña was inhabited by the Altiplano people, their subsistence practices were mainly agrarian (Rice et al., 1990; Sharratt, 2017; Williams, 1990). They were a cultural group distinct form any other inhabiting the Osmore River Valley.

245 tombs were excavated at Moquegua, M6: Estuquiña, 185 of which had already been opened and disturbed prior to excavation, likely by looters (Williams, 1990). Individual AD82 was excavated from Tomb 629, which had previously been disturbed and its burial context left exposed to the surrounding environment. This exposure could account for the high amount of mycobacterial background from soil dwelling mycobacteria present in our sample library form this individual.

Jane E. Buikstra, Sloan Williams and Niki R. Clark carried out mortuary excavations and osteological analyses. Analyses revealed skeletal lesions consistent with prolonged tuberculosis (TB) infection to be prevalent at Moquegua, M6: Estuquiña. Of the investigated individuals 37 had skeletal TB lesions, constituting a minimum of 8.9% of the total population. Amongst adult individuals for which sex could be determined 9.8% of females and 19.2% of males displayed skeletal TB lesions affecting vertebrae and/or ribs (Buikstra & Williams, 1991). Thus, indicating that men may have been at higher risk of contracting TB, or that they were more likely to sustain a prolonged infection allowing skeletal lesions to form.

Excavations of the residential areas, carried out by Donald S. Rice and Geoffrey W. Conrad revealed that faunal remains were included among the grave goods at the site. Skeletal remains from guinea pigs and llama feet were common finds, as well as fish and bird remains (Williams, 1990). No pinniped remains were reportedly found.

Colombian sites

The archaeological sites of Las Delicias and Candelaria La Nueva are named after the *barrios* (neighborhoods) of Bogotá, Colombia's capital city, in which they are located. Bogotá is situated on the 'Sabana de Bogotá', a plateau in the Eastern Cordillera in the

Andes. Bogotá is approximately 2,640 m above sea level and more than 600 km inland from the nearest coastline.

The sites of Las Delicias and Candelaria La Nueva, which are separated by only 3.3 km, were excavated as part of archaeological rescue efforts carried out by researchers at the *Instituto Colombiano de Antropología e Historia* (ICANH). Both sites are located in the territory that belonged to the Muisca confederation, a highly organized confederation of tribes, which occupied the area between 950-1550 C.E. until European colonization changed the societal structure (Cifuentes Toro, 1987). The tribes of the Muisca confederation are known to have relied on long-distance trade networks in order to trade salt, emeralds, cotton and clothing for crops cultivated in warmer and lower elevated terrain, such as cocoa, even reaching coastal regions (Cárdenas Arroyo, 1994; Garcia, 2012; Kurella, 1998). The subsistence practices of the Muisca were primarily agricultural, with high production of corn and beans, supplemented with hunting and freshwater animals caught in lakes and rivers (Garcia, 2012; Kurella, 1998).

Based on human osteological surveys conducted by Felipe Cárdenas-Arroyo (ICANH), Jane E. Buikstra (ASU), Judith Arnett (University of the Andes) and Laura Paloma Leguizamón (ICANH), only the individuals included in this study displayed skeletal lesions consistent with prolonged TB infection at Candelaraia La Nueva (individual AD386) and Las Delicias (individuals AD281, AD387). However, another six individuals from other Muisca territory sites included in this study displayed skeletal lesions characteristic of TB infection – indicating that it was not uncommon in the Muisca confederation, although it has not been recorded at such high prevalence rates as those recorded for Moquegua, M6: Estuquiña. Samples from these six individuals were screened using qPCR and gene-capture methods, but did not yield positive results for MTBC DNA preservation (Supplementary Table 1).

Las Delicias

Braida Enciso led the rescue excavations at Las Delicias. The archaeological site was located on an alluvial terrace of the Tunjuelito River. Five strata were found; one of them (depth: 30 centimeters) contained evidence of human occupation (pers. comm. Laura Paloma Leguizamón). This evidence consisted of tombs, remains of posts from dwellings, evidence of storage in the form of pits, musical instruments crafted from animal bone, charcoal, seeds, pottery, lithics, and animal and human bone remains (Cardenas Arroyo, 1993; Enciso Ramos, 1991). Archaeological evidence points to this site having been used for domestic, agricultural, and funerary activities. The skeletal remains of at least 18 human individuals were recovered from various tombs at the site (Instituto Colombiano de Antropologia e Historia (ICANH)), along with the remains of deer, birds, fish and guinea pigs (Cardenas Arroyo, 1993; Instituto Colombiano de Antropologia e Historia (ICANH)). Paleodemographic analysis revealed that 66.6% of individuals were infants and that adults are highly underrepresented compared to the neighboring site of Candelaria La Nueva where the demography of the cemetery is more equally distributed in terms of age and sex (Cardenas Arroyo, 1993).

Radiocarbon dates from the site have been generated from two different types of material. Two dating efforts were commissioned by ICANH from charcoal material. One charcoal sample was associated with a dwelling feature from a domestic context yielding a date of 790 ± 82 C.E. (1230 ± 70 B.P; Beta-39874). The second charcoal

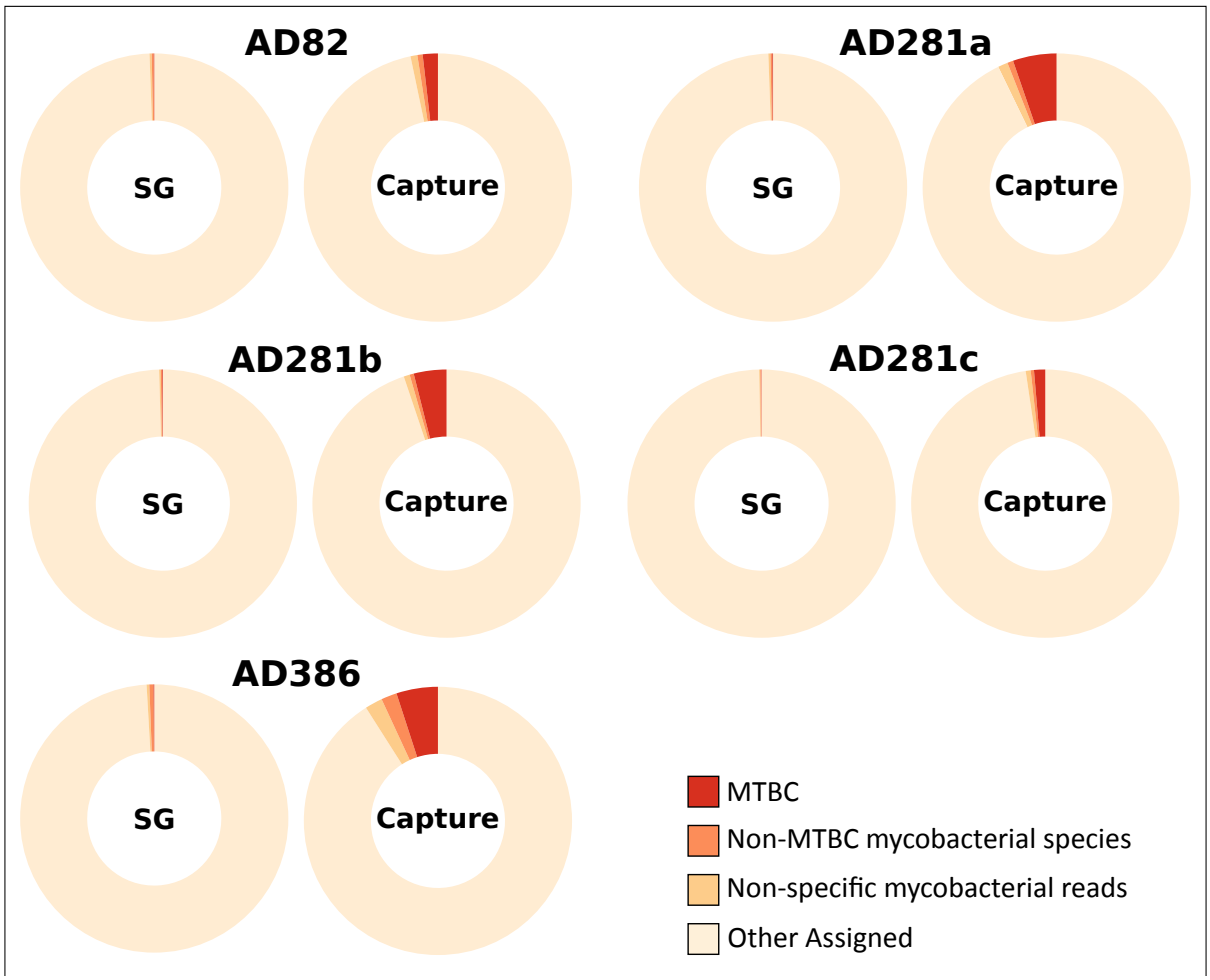
sample was from a pit associated with a grave and yielded a date of 980 ± 66 C.E. (1060 ± 70 B.P.; Beta-39874). Another date (this study) was generated from a sample from a rib from individual 281 (LD-X-011), which yielded a calibrated date of 1265-1380 C.E. (Supplementary Table 2). Based on these radiocarbon dates Las Delicias seems to have been occupied during both the early (950-1250 C.E.) and late (1250- 1550 C.E.) Muisca period (Delgado, 2016).

Candelaria La Nueva

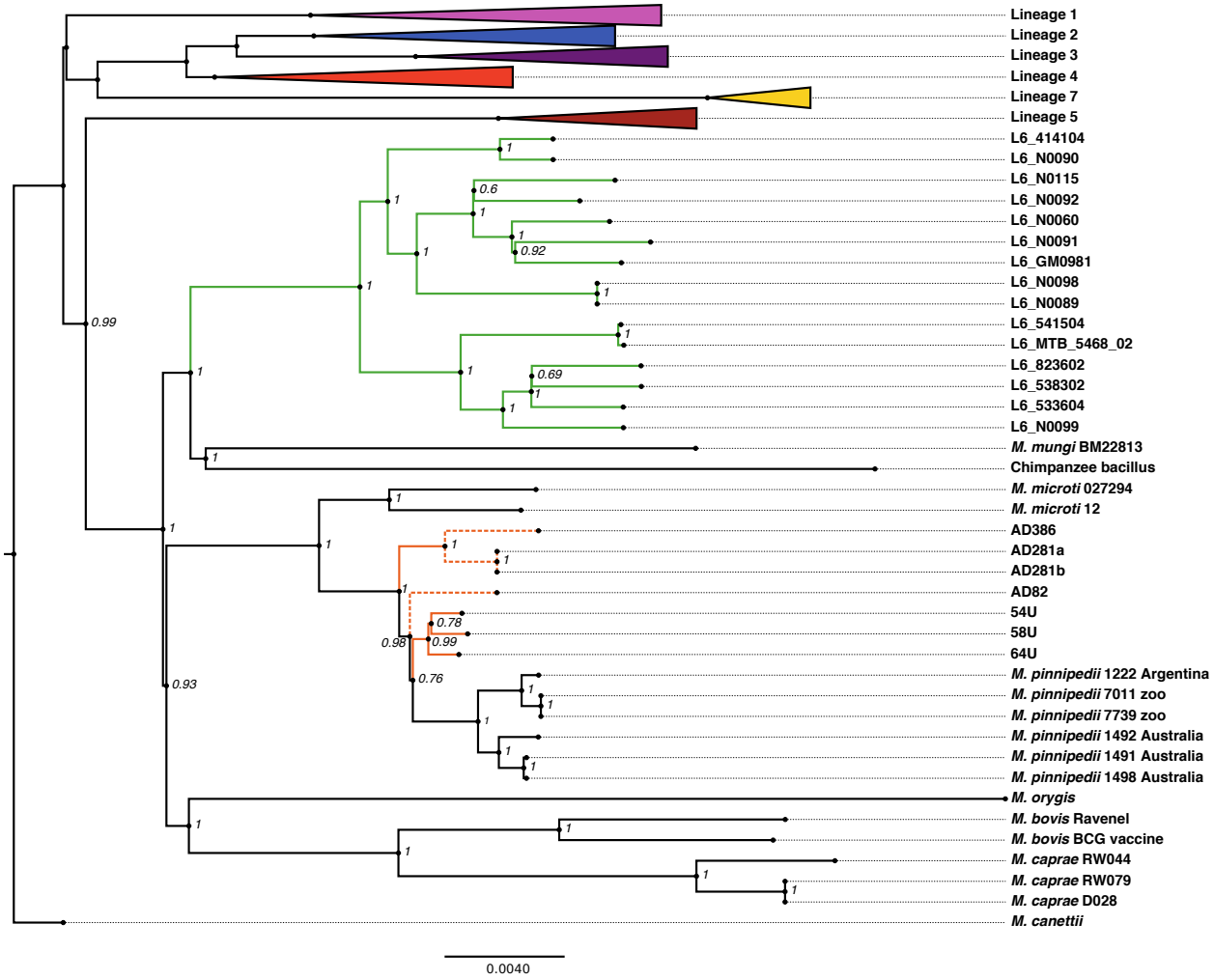
Arturo Cifuentes and Leonardo Moreno led the rescue excavations at Candelaria La Nueva, located on a colluvial terrace near the Tunjuelito River (Cifuentes Toro, 1987). Four strata were found, one of them (depth: 30-40 centimeters) revealed evidence of human occupation. This evidence consisted of tombs, traces of posts from dwellings, storage traces and pits, charcoal, seeds, pottery, lithics, and animal and human bone remains (pers. comm. Laura Paloma Leguizamón). Archaeological evidence shows that this site was used for both domestic and funerary activities. 48 disturbed or destroyed human graves were identified at the site, and the skeletal remains of 37 individuals were collected (Instituto Colombiano de Antropologia e Historia (ICANH)). Paleodemographic analyses revealed that the maximum age for the Candelaria La Nueva population was 45 years, with a high mortality rate among females aged 20-30 years (Zajec, 1989). The faunal remains recovered from the site belong to deer, armadillo, guinea pig, mollusks, and an unidentified feline species (Instituto Colombiano de Antropologia e Historia (ICANH)).

Radiocarbon dating has been carried out for three individuals excavated from Candelaria La Nueva. Two dates were published by Therrien and Enciso (1992), using samples from two separate individuals, yielding the following dates: 1250 ± 110 C.E. (700 ± 110 BP; GX-18839-G) and 1175 ± 110 (775 ±110 BP; GX-18840-G). Radiocarbon dating for Individual 386 (87-X-005) (this study) yielded a calibrated date of 1450-1640 C.E. (Supplementary Table 2). These dates indicate that the dated individuals lived during the Muisca period (950-1550 C.E.) and, potentially in the case of Individual 386, overlapped with the arrival of Spanish colonizers in 1536 (Kurella, 1998).

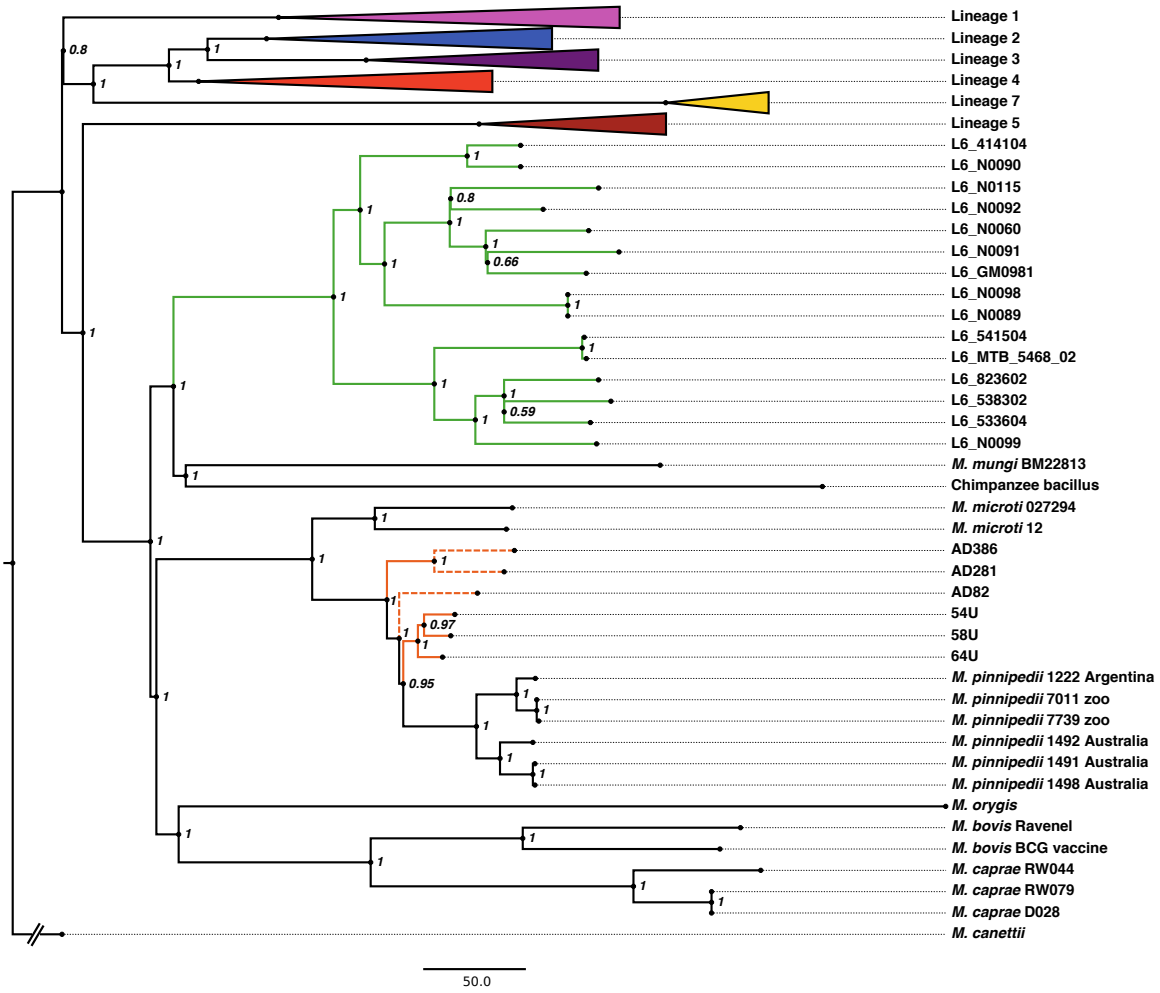
Previous ancient DNA work has been carried out on samples from 14 of the individuals excavated from Candelaria La Nueva, and this research sought to determine the mitochondrial haplogroup of the individuals via Restriction Fragment Length Polymorphism (RFLP) analysis (Jara, 2010). All ancient individuals were found to carry haplogroup A, a common haplogroup in the Americas, including individual AD386 (87-X-005) from whom an *M. pinnipedii* genome was reconstructed in this study.



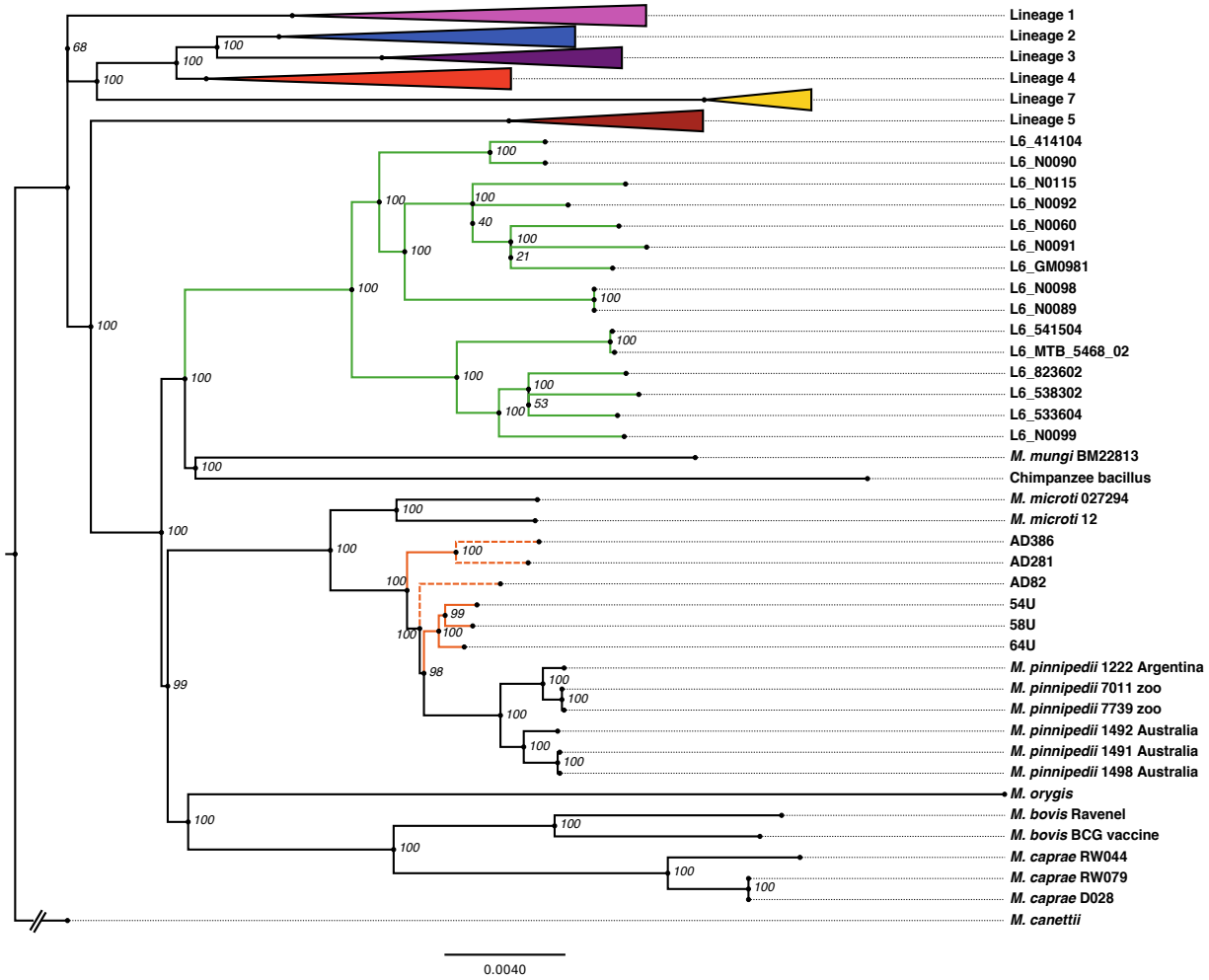
Supplementary Figure 1. MALT analysis of mycobacterial read proportions in sample libraries before and after capture. Shotgun and capture data for the UDG treated libraries, from which whole genomes were generated, were analyzed with MALT using a database constructed from the full nucleotide database available through NCBI RefSeq (December 2016). Pie charts displaying proportions of mycobacterial reads assigned by MALT amongst samples positive for MTBC DNA before and after in-solution capture. The proportions shown are based on the reads that could be assigned by MALT to a taxon included in the database when using an 85% identity filter. SG= un-enriched shotgun data, Capture = data after whole-genome enrichment.



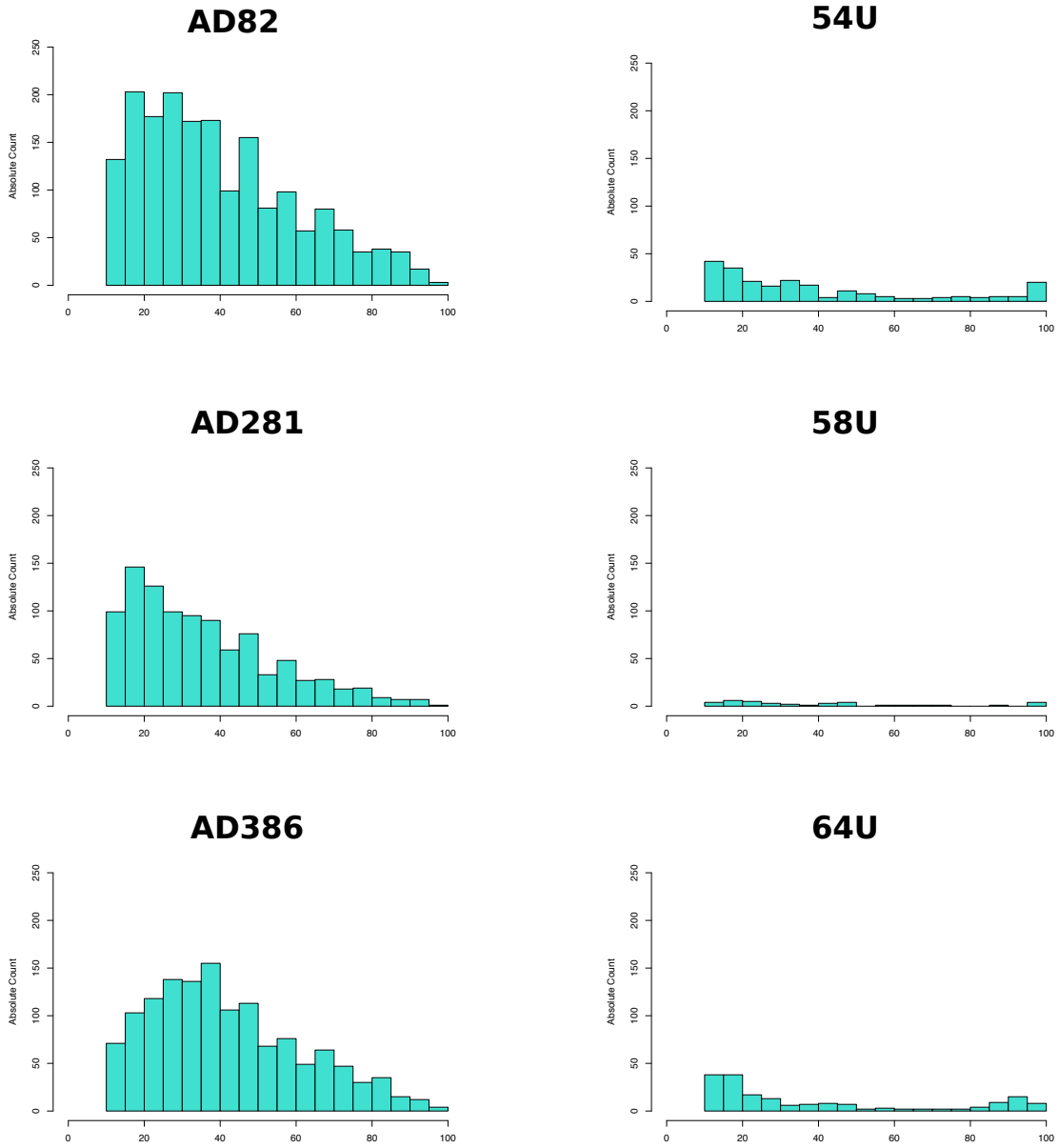
Supplementary Figure 2. Neighbour joining MTBC phylogeny with AD281a and AD281b separated. The phylogenetic tree was constructed using the full dataset of 267 genomes including the seven ancient genomes where the genomic data from AD281a and AD281b are separated. Ancient genomes are highlighted in orange. Human-adapted lineages 1-5 and 7 have been collapsed. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 10,765 positions out of a possible 44,304. Our three ancient genomes fall together with other ancient Peruvian genomes within the *M. pinnipedii* clade; due to their artificially longer branch lengths these branches are marked with a dashed line.



Supplementary Figure 3. Neighbour joining MTBC phylogeny. The phylogenetic tree was constructed using the full dataset of 266 genomes including the six ancient genomes, which are highlighted in orange. Human-adapted lineages 1-5 and 7 have been collapsed. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates, and is based on 14,263 positions out of a possible 44,273. Our three ancient genomes fall together with other ancient Peruvian genomes within the *M. pinnipedii* clade; due to their artificially longer branch lengths these branches are marked with a dashed line.



Supplementary Figure 4. Maximum Likelihood MTBC phylogeny. The phylogenetic tree was constructed using the full dataset of 266 genomes including the six ancient genomes, which are highlighted in orange. Human-adapted lineages 1-5 and 7 have been collapsed. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates, and is based on 14,263 positions out of a possible 44,273. Our three ancient genomes fall together with other ancient Peruvian genomes within the *M. pinnipedii* clade; due to their artificially longer branch lengths these branches are marked with a dashed line.



Supplementary Figure 5. Histograms of SNP allele frequency distributions for six ancient *M. pinnipedii* genomes. The x-axis shows the SNP allele frequencies as a percentage. All variants where the SNP allele frequency is higher than 10% and lower than 100% are shown.

Supplementary Table 1. Overview of samples and corresponding qPCR and gene capture screening results

Sample ID	Archaeological Site	Date Radiocarbon (rc) OR archaeological (arch.)	Archaeological ID	Skeletal Element	MTBC qPCR screening				MTBC gene capture screening % of gene covered >= 1X					
					rpoB2	IS6110	IS1081	Source	rpoB	katG	mtp40	gyrA	gyrB	Source
AD382	Soacha, Colombia	--	T-1	Vertebral body	Pos	Pos	Pos	This study	34.6	5.8	0	3.5	10.6	This study
AD383	Soacha, Colombia	--	T-35	Vertebral body	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD384	Soacha, Colombia	--	T-42	Vertebral body	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD385	Marin, Colombia	--	36	Vertebral body	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD386	La Candelaria Nueva, Colombia	cal 1450-1640 C.E. (rc)	87-X-005	Vertebral body	Pos	Pos	Pos	This study	37.7	20.9	39.1	25.2	31.1	This study
AD387	Las Delicias, Colombia	--	LD-IX-005	Vertebral body	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD281a	Las Delicias, Colombia	cal 1256-1385 C.E. (rc)	LD-X-011	Vertebral body	Pos	Pos	Pos	Harkins *et al.* 2015	--	--	--	--	--	--
AD281b/AD388				rib	Pos	Pos	Pos	This study	75.3	72.2	61.8	73.4	86.8	This study
AD281c				rib	n/a	n/a	n/a	--	--	--	--	--	--	--
AD389	Tecquendama, Colombia	--	Ent 1A	Vertebral body	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD390	Bugala Grande, Colombia	--	Tumba 1	Rib	Neg	Neg	Neg	This study	--	--	--	--	--	--
AD82	Moquegua, M6 : Estuquiña, Peru	1250-1470 C.E. (arch.)	M6-4165	Vertebral body	Pos	Pos	Pos	Harkins *et al.* 2015	49.67	34.33	28.03	69.46		Bos *et al.* 2014

Supplementary Table 2. Radiocarbon dates generated as part of this study

Sample ID	Site	Arch. ID	Skeletal Element	Laboratory number	14C Age (years B.P.)	Calibarted Age (C.E.)	Confidence interval
AD281b	Las Delicias	LD-X-011	Rib	Beta-434387	460±30	1265-1380	95
AD386	Candelaria La Nueva	87-X-005	Vertebral body	Beta-434386	150±30	1450-1640	95

Supplementary Table 4. Values used to evaluate the level of MTBC reads in UDG-treated shotgun and capture data relative to non-target DNA
The UDG capture data for all samples is based on total number of reads generated across multiple sequencing runs

			Assigned by MALT 85% identity								
Sample ID	**Data type**	**# Total pre-processed reads**	**# Total reads assigned by MALT**	**Mycobacterium reads (SUM)**	**Mycobacterium reads (Assigned)**	**MTBC reads (SUM)**	**Reads assigned to Non-MTBC mycobacterial species**	**Total non-MTBC reads assigned by MALT**	**% MTBC MALT-assigned reads out of total reads sequenced**	**% MTBC reads out of total reads assigned in MALT**	**% non-MTBC MALT-assigned reads out of total reads sequenced**
AD82	Shotgun	1055419	130388	736	338	82	316	130072	0.007	0.063	12.324
	Capture	449538686	113712632	3735942	939760	2074941	721241	112991391	0.461	1.825	25.135
AD281a	Shotgun	830092	80371	450	204	68	178	80193	0.008	0.085	9.661
	Capture	56932159	12113723	870018	144376	640145	85497	12028226	1.124	5.284	21.127
AD281b	Shotgun	3905086	378733	1682	745	309	628	378105	0.007	0.081	9.682
	Capture	113933376	22010684	1136613	159818	867452	109343	21901341	0.761	3.941	19.223
AD281c	Shotgun	652852	75899	219	103	23	93	75806	0.003	0.030	11.611
	Capture	3059617	729054	17051	4357	9807	2887	726167	0.320	1.345	23.734
AD386	Shotgun	6414097	919244	8396	2995	795	4606	914638	0.012	0.086	14.260
	Capture	328622471	81219888	7321404	1715869	4064962	1540573	79679315	1.237	5.005	24.246

Supplementary Table 7. Mapping statistics for the non-UDG treated shotgun data mapped to the MTBC_anc

Sample ID (nU=non-UDG)	Data source	Paired-End (PE) or Single-End (SE) data	# raw paired end reads	# reads after adapter clipping and merging (PE) prior mapping	# of merged reads	% merged reads	# mapped reads prior to duplicate removal	# quality filtered mapped reads prior to duplicate removal	# of duplicates removed	Mapped reads after duplicate removal	Endogenous DNA (%)	Endogenous DNA (%) quality filtered reads
54nU	Bos et al. 2014	PE	141470	70210	69941	99,617	460	377	2	375	0.658	0.539
58nU	Bos et al. 2014	PE	224448	112368	111020	98,8	745	579	7	572	0.671	0.522
64nU	Bos et al. 2014	PE	353846	155631	116954	75,148	983	927	3	924	0.841	0.793
AD82-nU	Bos et al. 2014	PE	156706	77437	76884	99,286	108	72	0	72	0.14	0.094
AD281b-nU	This study	SE	4140257	4077863	n.a.	n.a.	2554	1573	185	1388	0.063	0.039
AD386-nU	This study	SE	4267108	4225997	n.a.	n.a.	2275	1493	197	1296	0.054	0.035

Table continued…

Sample ID (nU=non-UDG)	Duplication factor	Mean coverage	Standard deviation in coverage	Coverage >= 1-fold (%)	Coverage >= 2-fold (%)	Coverage >= 3-fold (%)	% damage of first 5-prime base	% damage of second 5-prime base	Average fragment length (bp)	Median fragment length (bp)	GC content (%)
54nU	1.005	0.0038	0.149	0.23	0.03	0.02	0.1778	0.0813	44.25	42	59.29
58nU	1.012	0.0061	0.3074	0.32	0.02	0.02	0.1157	0.1123	47.25	44	61.88
64nU	1.003	0.0077	0.1041	0.72	0.02	0.01	0.1362	0.1039	36.62	35	64.6
AD82-nU	1	0.0009	0.0449	0.06	0.01	0.01	0.0556	0.0952	53.26	49.5	61.16
AD281b-nU	1.133	0.0159	0.4947	0.62	0.13	0.08	0.0719	0.0423	50.4	47	59.82
AD386-nU	1.152	0.0157	0.4763	0.7	0.11	0.07	0.1087	0.0661	53.51	50	59.52

n.a.= not applicable

Supplementary Table 8. Ratios of MTBC and non-MTBC mycobacterial reads in all six ancient samples
Comparative calculations were made based on the results of two different MALT runs, one executed with 85% and the other with 95% "minimum percent identity"

Sample ID (nU=non-UDG)	# reads after pre-processing	Total Assigned reads 85ID	85ID Mycobacterium SUM	85ID Mycobacterium Assigned	85ID Non-MTBC mycobacterial reads	85ID MTBC SUM	Ratio: Non-MTBC Mycobacterial reads/MTBC SUM - 85ID	Ratio: Mycobacterium Assigned/MTBC SUM - 85ID	% MTBC SUM reads out of total	% MTBC SUM reads out of assigned in MALT
AD82-nU	77437	14672	140	30	139	1	139.000	30.000	0.001	0.007
AD281b-nU	4077863	416089	3339	863	3019	320	9.434	2.697	0.008	0.077
AD386-nU	4225997	602157	7185	1681	6714	471	14.255	3.569	0.011	0.078
54nU	70210	10271	299	32	94	205	0.458	0.156	0.292	1.996
58nU	112368	20630	318	4	9	309	0.029	0.013	0.275	1.498
64nU	155631	88835	866	20	27	839	0.032	0.024	0.539	0.944

Table continued…

Sample ID (nU=non-UDG)	# reads after pre-processing	Total Assigned reads 95ID	95ID Mycobacterium SUM	95ID Mycobacterium Assigned	95ID Non-MTBC mycobacterial reads	95ID MTBC SUM	Ratio: Non-MTBC Mycobacterial reads/MTBC SUM - 95ID	Ratio: Mycobacterium Assigned/MTBC SUM - 95ID	% MTBC SUM reads out of total	% MTBC SUM reads out of assigned in MALT
AD82-nU	77437	2860	30	2	29	1	29.000	2.000	0.001	0.035
AD281b-nU	4077863	101610	654	144	430	224	1.920	0.643	0.005	0.220
AD386-nU	4225997	235946	1080	210	780	300	2.600	0.700	0.007	0.127
54nU	70210	3836	207	8	15	192	0.078	0.042	0.273	5.005
58nU	112368	8513	300	4	4	296	0.013	0.013	0.263	3.477
64nU	155631	77091	845	15	17	828	0.020	0.018	0.532	1.074

SUM=summarized, reads assigned to that node and nodes below

Supplementary Table 10. Mapping statistics for the captured non-UDG treated sensitive/stringent filtered data mapped to MTBC_anc

Sample ID (nU=non-UDG)	# of Raw Reads prior Clip & Merge (C&M)	# reads after C&M prior mapping	# mapped reads prior RMDup	# of Duplicates removed	Mapped Reads after RMDup	Endogenous DNA (%)	Cluster Factor	Mean Coverage	std. dev. Coverage
AD281-nU	56396	56396	56396	11177	45219	100	1.247	0.7438	2.2462
AD386-nU	54138	54138	54138	7772	46366	100	1.168	0.828	2.2937
AD82-nU	23565	23565	23565	6469	17096	100	1.378	0.2696	1.9848

Table continued…

Sample ID (nU=non-UDG)	Coverage >= 1X in %	Coverage >= 2X in %	Coverage >= 3X in %	Coverage >= 4X in %	Coverage >= 5X in %	DMG 1st Base 3'	DMG 2nd Base 3'	DMG 1st Base 5'	DMG 2nd Base 5'	average fragment length	median fragment length	GC content in %
AD281-nU	42.64	16.4	6.06	2.24	0.93	0.0868	0.0576	0.0872	0.0545	72.56	68	61.13
AD386-nU	45.78	19.19	7.73	3,0	1.25	0.0735	0.0575	0.0732	0.0598	78.78	76	61.13
AD82-nU	17.68	3.04	0.77	0.38	0.27	0.0744	0.0563	0.0769	0.059	69.56	65	61.23

Supplementary Table 11. Mapping statistics of non-UDG treated capture data to the human genome (hg19)

Sample ID (nU=non-UDG)	Library treatment	Total # raw reads	# of reads after pre-processing	# of Merged Reads	% Merged Reads	# mapped reads prior to duplicate removal	# quality filtered mapped reads prior to duplicate removal	# of duplicates removed	Mapped reads after duplicate removal	Endogenous DNA (%)	Endogenous DNA (%) quality filtered reads
AD82-nU	non-UDG	6720190	3407914	3236339	94,965	1343	485	28	457	0.041	0.015
AD281-nU	non-UDG	6412450	6412450	6412450	95,235	69912	42101	2069	40032	1.090	0.657
AD386-nU	non-UDG	12954162	6654250	6187681	92,988	195483	139077	8587	130490	3.159	2.248

Table continued…

Sample ID (nU=non-UDG)	Duplication factor	Mean coverage	Standard deviation in coverage	Coverage >= 1-fold (%)	# of reads on mitochondrium	Average Coverage on mitochondrium	MT/NUC Ratio	% damage of first 5-prime base	% damage of second 5-prime base	Average fragment length (bp)	Median fragment length (bp)	GC content (%)
AD82-nU	1.061	0.0000	0.003	0	4	0.017	1873.62	0.1475	0.079	62.78	61	54.29
AD281-nU	1.052	0.0009	0.031	0.09	118	0.531	577.73	0.1405	0.084	71.38	68	48.53
AD386-nU	1.066	0.0028	0.059	0.27	1268	6.173	2253.88	0.1452	0.085	65.77	62	45.12

References

Buikstra, J. E., & Williams, S. R. (1991). Tuberculosis in the Americas: Current perspectives. In D. Ortner & A. C. Aufderheide (Eds.), *Human palaeopathology: Current syntheses and future options* (pp. 161-172). Washington, D.C.: Smithsonian Institution Press.

Cardenas Arroyo, F. (1993). Paleodieta y Paleodemografia en Poblaciones Arqueologicas Muiscas (Sitios Las Delicias y Candelaria). *Revista Colombiana de Antropologia, XXX*.

Cárdenas Arroyo, F. (1994). Reconstrucción química de la paleodieta en restos arqueológicos humanos del territorio

muisca. *Revista Eres. Serie de arqueología,, 5*(1), 71-82.

Cifuentes Toro, J. A. (1987). *Proyecto de rescate arqueológico de la avenida Villavicencio: barrio Candelaria la Nueva, Bogotá*. Retrieved from Bogotá : ICAN:

Delgado, M. (2016). Stable isotope evidence for dietary and cultural change over the Holocene at the Sabana de Bogotá region, Northern South America. *Archaeological and Anthropological Sciences, 10*(4), 817-832.

Enciso Ramos, B. E. (1991). Arqueología de rescate, en el barrio las Delicias, Bogotá. *Revista Colombiana de Antropología, 28*, 155-160.

Garcia, J. L. (2012). *The Foods And Crops Of The Muisca: A Dietary Reconstruction Of The Intermediate Chiefdoms Of Bogota (bacata) And Tunja (hunza), Colombia.* (Master of Arts (M.A.) Masters Thesis), University of Central Florida. Retrieved from http://purl.fcla.edu/fcla/etd/CFE0004199

Instituto Colombiano de Antropologia e Historia (ICANH). Consultar la colección ósea del ICANH (en el Museo Nacional de Colombia). Retrieved from http://www.icanh.gov.co/servicios_ciudadano/tramites_servicios/servicios/laboratorio_arqueologia/4550&download=Y

Instituto Colombiano de Antropologia e Historia (ICANH). Registro Materiales Arqueológicos. Retrieved from http://www.icanh.gov.co/nuestra_entidad/grupos_investigacion/arqueologia/parques_asociados/5151

Jara, N. P. (2010). Aplication of authenticity criteria in mitochondrial studies on archaic bone remains from a prehispanic Muisca population. *Colombia Médica, 41*(4).

Kurella, D. (1998). The Muisca, Chiefdoms in Transition. In E. M. Redmond (Ed.), *Chiefdoms and Chieftaincy in the Americas* (pp. 189-216): University Press of Florida.

Rice, D. S., Conrad, G. W., & Buikstra, J. (1990). Investigaciones en Estuquiña: Descripciones Preliminares, 1985-1986. In L. K. Watanabe, M. E. Moseley, & F. Cabieses (Eds.), *Trabajos Anqueologicos en Moquegua, Peru* (Vol. 3, pp. 39-93). Peru: Programa Contisuyu del Museuo Peruana de Sciences de la Salud, Southern Peru Copper Corporation.

Sharratt, N. (2017). Steering Clear of the Dead: Avoiding Ancestors in the Moquegua Valley, Peru. *American Anthropologist, 119*(4), 645-661.

Therrien, M., & Enciso, B. (1992). Avances de investigación: una re-investigacion arqueológica en la sabana de Bogotá. *Boletín Museo del Oro, 31*, 130-131.

Williams, S. R. (1990). *The skeletal biology o f Estuquina: A Late Interm ediate Period site in southern Peru.* Northwestern University, Evanston, Illinois.

Zajec, D. N. (1989). *Paleodemografía de la población Muisca del sitio Candelaria la Nueva: informe final de trabajo de campo*. Retrieved from Bogotá : Universidad de los Andes:

Paper II

Å. J. Vågene, A. Herbig, M. G. Campana, N. M. Robles García, C. Warinner, S. Sabin, M. A. Spyrou, A. Andrades Valtueña, D. Huson, N. Tuross, K. I. Bos, J. Krause (2018).
Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico.
Nature Ecology & Evolution 2:520-528.

Salmonella enterica genomes from victims of a major 16th century epidemic in Mexico

Authors: Åshild J. Vågene[1,2,#], Alexander Herbig[1,2,#*], Michael G. Campana[3, 4†], Nelly M. Robles García[5], Christina Warinner[1], Susanna Sabin[1], Maria A. Spyrou[1,2], Aida Andrades Valtueña[1], Daniel Huson[6], Noreen Tuross[3,*], Kirsten I. Bos[1,2,*] and Johannes Krause[1,2*]

Affiliations:

[1]Max Planck Institute for the Science of Human History, Jena, Germany.

[2]Institute for Archaeological Sciences, University of Tübingen, Tübingen, Germany.

[3]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA.

[4]Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland.

[5]INAH, National Institute of Anthropology and History, Mexico, Teposcolula-Yucundaa Archaeological Project.

[6]Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Tübingen, Germany.

[#]These authors contributed equally to this work

*Correspondence to: A.H.: herbig@shh.mpg.de; N.T.: tuross@fas.harvard.edu; K.I.B.: bos@shh.mpg.de; J.K.: krause@shh.mpg.de

†Current address: M.G.C.: Smithsonian Conservation Biology Institute, Center for Conservation Genomics, 3001 Connecticut Avenue NW, Washington, DC 20008, USA.

Summary paragraph: Indigenous populations of the Americas experienced high mortality rates during the early contact period as a result of infectious diseases, many of which were introduced by Europeans. Most of the pathogenic agents that caused these outbreaks remain unknown. Through the introduction of a new metagenomic analysis tool called MALT applied here to search for traces of ancient pathogen DNA, we were able to identify *Salmonella enterica* in individuals buried in an early contact era epidemic cemetery at Teposcolula-Yucundaa, Oaxaca in southern Mexico. This cemetery is linked, based on historical and archaeological evidence, to the 1545-1550 CE epidemic that affected large parts of Mexico. Locally this epidemic was known as "*cocoliztli*", the pathogenic cause of which has been debated for more than a century. Here we present genome-wide data from ten individuals for *Salmonella enterica* subsp. *enterica* serovar Paratyphi C, a bacterial cause of enteric fever. We propose that *S.* Paratyphi C be considered a strong candidate for the epidemic population decline during the 1545 *cocoliztli* outbreak at Teposcolula-Yucundaa.

Introduction

Infectious diseases introduced to the New World following European contact led to successive outbreaks in many regions of the Americas that continued well into the 19th century. These often caused high mortality and thus contributed a central, and often underappreciated, influence on the demographic collapse of many indigenous populations[1-4]. Population declines linked to regionally specific epidemics are estimated to have reached as high as 95%[3], and their genetic impact based on recent population-based studies of ancient and modern human exome and mitochondrial data attest to their scale[5,6]. One hypothesis posits that the increased susceptibility of New World populations to Old World diseases facilitated European conquest, whereby rapidly disseminating diseases severely weakened indigenous populations[2], in some cases even in advance of European presence in the region[2,7]. Well-characterized infections such as smallpox, measles, mumps, and influenza are known causes of later contact era outbreaks; however, the diseases responsible for many early contact period New World epidemics remain unknown and have been the subject of scientific debate for over a century[2-4,7,8].

Morphological changes in skeletal remains[9] and ethnohistorical accounts[10] are often explored as sources for understanding population health in the past, although both provide only limited resolution and have generated speculative and at times conflicting hypotheses about the diseases introduced to New World populations[2,3,7,11,12]. Most infectious diseases do not leave characteristic markers on the skeleton due to either their short periods of infectivity, the death of the victim in the acute phase before skeletal changes formed, or a lack of osteological involvement[9]. While historical descriptions of infectious disease symptoms can be detailed, they are subject to cultural biases, suffer from translational inaccuracies, lack a foundation in germ theory, and describe historical forms of a condition that may differ from modern manifestations[8,11]. Additionally, differential diagnosis based on symptoms alone can be unreliable even in modern contexts since many infectious diseases have similar clinical presentations.

Genome-wide studies of ancient pathogens have proven instrumental in both identifying and characterizing past human infectious diseases. These studies have largely been restricted to skeletal collections where individuals display physical changes consistent with particular infections[13-15], an historical context that links a specific pathogen to a known epidemiological event[16], or an organism that was identified via targeted molecular screening without prior indication of its presence[17]. Recent attempts to circumvent these limitations have concentrated on broad-spectrum molecular approaches focused on pathogen detection via fluorescence-hybridization-based microarray technology[18], identification via DNA-enrichment of certain microbial regions[19] or computational screening of non-enriched sequence data against human microbiome data sets[20]. These approaches offer improvements, but remain biased in the bacterial taxa used for species-level assignments. As archaeological material is expected to harbour an abundance of bacteria that stem from the depositional context,

omission of environmental taxa in species assignments can lead to false-positive identifications. Additional techniques for authenticating ancient DNA have been developed[21,22], including the identification of characteristic damage patterns caused by the deamination of cytosines[23], methods that evaluate evenness of coverage of aligned reads across a reference genome, or length distributions that consider the degree of fragmentation, where ancient molecules are expected to be shorter than those from modern contaminants[24].

A typical NGS dataset from an ancient sample comprises millions of DNA sequencing reads, making taxonomic assignment and screening based on sequence alignments computationally challenging. The gold standard tool for alignment-based analyses is the Basic Local Alignment Tool (BLAST)[25], due to its sensitivity and statistical model. However, the computational time and power BLAST requires to analyse a typical metagenomic dataset is often prohibitive.

Here we introduce the MEGAN alignment tool (MALT), a program for the fast alignment and analysis of metagenomic DNA sequencing data. MALT contains the same taxonomic binning algorithm, i.e. the naïve LCA (Lowest Common Ancestor) algorithm (for reviews see[26,27]), implemented in the interactive metagenomics analysis software MEGAN[28]. Like BLAST, MALT computes 'local' alignments between highly conserved segments of reads and references. MALT can also calculate 'semi-global' alignments where reads are aligned end-to-end. In comparison to protein alignments or local DNA alignments, semi-global DNA alignments are more suitable for assessing various quality and authenticity criteria that are commonly applied in the field of paleogenetics.

We applied our MALT screening pipeline (Supplementary Figures 1, 2) using a database of all complete bacterial genomes available in NCBI RefSeq to non-enriched DNA sequence data from the pulp chamber of teeth collected from indigenous individuals excavated at the site of Teposcolula-Yucundaa, located in the highland Mixteca Alta region of Oaxaca, Mexico[29,30]. The site contains both pre-contact and contact era burials, including the earliest identified contact era epidemic burial ground in Mexico[30,31] (Fig. 1; Supplementary Methods 1). This is the only known cemetery historically linked to the *cocoliztli* epidemic of 1545-1550 CE[30], described as one of the principal epidemiological events responsible for the cataclysmic population decline of 16th century Mesoamerica[7,32]. This outbreak affected large areas of central Mexico and Guatemala, spreading perhaps as far south as Peru[7,30]. Via the MALT screening approach, we were able to identify ancient *Salmonella enterica* DNA in the sequence data generated from this archaeological material, to the exclusion of DNA stemming from the complex environmental background. While the pathogenic cause of the *cocoliztli* epidemic is ambiguous based on ethnohistorical evidence[7,8,30], we report the first molecular evidence of microbial infection with genome-wide data from *S. enterica* subsp. *enterica* serovar Paratyphi C (enteric fever) isolated from ten epidemic-associated contact era burials.

Results

The individuals included in this investigation were excavated from the contact era epidemic cemetery located in the Grand Plaza (administrative square) (n=24) and the pre-contact churchyard cemetery (n=5) at Teposcolula-Yucundaa between 2004 and 2010 (ref. 30) (Fig. 1; Supplementary Table 1; Supplementary Methods 1). Previous work demonstrated ancient DNA preservation at the site through the identification of New World mitochondrial haplogroups in 48 individuals, 28 of which overlap with this study[30]. Additionally, oxygen isotope analysis identified them as local inhabitants[30]. Thirteen individuals included in this study were previously radiocarbon dated[31] yielding dates that support archaeological evidence that the Grand Plaza (n=10) and churchyard (n=3) contain contact and pre-contact era burials, respectively (Supplementary Table 1). The Grand Plaza is estimated to contain >800 individuals, most interred in graves containing multiple persons. The excavated individuals contribute to a demographic profile consistent with an epidemic event[29,30].

Tooth samples were processed according to protocols designed for ancient DNA work (Supplementary Methods 2). An aggregate soil sample from the two burial grounds was analysed in parallel to gain an overview of environmental bacteria that may have infiltrated our samples. Pre-processed sequencing data of approximately one million paired-end reads per tooth were analysed with MALT using a curated reference database of 6247 complete bacterial genomes, comprising all those available in NCBI RefSeq (December 2016). Our approach limits ascertainment biases and false positive assignments that could result from databases deficient in environmental taxa (Supplementary Methods 3). A runtime analysis revealed a 200-fold improvement in computation time for MALT in comparison to BLASTn (see Methods). Results were visualized in MEGAN[28] and taxonomic assignments were evaluated with attention to known pathogenic species. Reads taxonomically assigned by MALT ranged from 4,842 to 44,315 for the samples. Assigned reads belonging to bacterial constituents of human oral and soil microbiota are present in varying proportions amongst the samples (Fig. 2; Supplementary Table 2). Of note, three teeth (Tepos_10, Tepos_14 and Tepos_35) had between 365 to 659 reads assigned to *Salmonella enterica*, a known cause of enteric fever in humans today. Of the *S. enterica* strains present in the database, *S.* Paratyphi C had the highest number of assigned reads (Supplementary Methods 3). Mapping these three metagenomic datasets to the *S.* Paratyphi C RKS4594 genome (NC_012125.1) revealed the characteristic pattern of damage expected of ancient DNA (Supplementary Fig. 3; Supplementary Methods 3; Supplementary Table 4). Subsequently, the sequencing data for all samples was mapped to the human genome (hg19), revealing a similar level of damage in the human reads for Tepos_10, Tepos_14 and Tepos_35, thus providing further support for the ancient origin of the *S. enterica* reads (Supplementary Methods 4; Supplementary Table 5). An additional seven individuals from the Grand Plaza cemetery and one negative control harboured low numbers of assigned *S. enterica* reads ranging from 4 to 51 (Supplementary Table 2). These were considered as potential weak-positive samples. One negative control was found to

contain 15 reads assigned to *S. enterica*, and a further four contained one or two reads, as did nine sample libraries, seven of which were not included in downstream analyses. The soil library and remaining sample libraries were void of *S. enterica* reads (Fig. 2; Supplementary Methods 3; Supplementary Table 2). An additional MALT screen for traces of viral DNA revealed one notable taxonomic hit to *Salmonella* phage Vi II-E1, a phage associated with *Salmonella* serovars that produce the Vi capsule antigen[33], which includes *S.* Paratyphi C (Supplementary Methods 3; Supplementary Table 3).

To further authenticate and elucidate our findings we performed whole-genome targeted array and in-solution hybridization capture[34,35], using probes designed to encompass modern *S. enterica* genome diversity (Supplementary Methods 5, 6, 7; Supplementary Table 6). All five pre-contact samples, the soil sample, one post-contact sample putatively negative for *S. enterica* based on our MALT screening, all negative controls, and both UDG-treated (DNA damage removed) and non-UDG treated libraries from the ten putatively positive samples (Tepos_10, Tepos_11, Tepos_20, Tepos_14, Tepos_34, Tepos_35, Tepos_36, Tepos_37, Tepos_38, Tepos_41) were included in the capture (Supplementary Methods 6, 7).

Mapping and genotyping of the captured Illumina sequenced reads was performed using the *S.* Paratyphi C reference genome (NC_012125.1) (Supplementary Methods 6, 7, 8; Supplementary Table 7). Capture of *S. enterica* DNA was successful for the ten positive samples yielding a minimum of 33,327 unique reads per UDG treated library. The remaining bone samples, soil sample, and negative controls were determined to be negative for ancient *S. enterica* DNA with the exception of one negative control (EB2-091013) that had likely become cross-contaminated during processing (see Supplementary Methods 8; Supplementary Table 7). Five complete genomes were constructed for Tepos_10, Tepos_14, Tepos_20, Tepos_35 and Tepos_37, covering 95%, 97%, 67%, 98% and 74% of the reference at a minimum of 3-fold coverage and yielding an average genomic coverage of 33-, 36-, 4.6-, 96- and 5.5-fold, respectively (Table 1). Artificial reads generated *in silico* for 23 complete genomes included in the probe design were also mapped to the *S.* Paratyphi C RKS4594 reference (Supplementary Methods 8; Supplementary Table 6) and phylogenetic comparison revealed that the five ancient genomes clustered with *S.* Paratyphi C (Fig. 3; Supplementary Figures 4, 6; Supplementary Methods 8). The phylogenetic positioning was retained when the whole dataset was mapped to and genotyped against the *S.* Typhi CT18 reference genome (NC_003198.1) (Supplementary Figure 5; Supplementary Table 8), the most common bacterial cause of enteric fever in humans today. This result excludes the possibility of a reference bias. Despite all five ancient genomes being contemporaneous, the Tepos_10 genome was observed to contain many more derived positions. An investigation of heterozygous variant calls showed that Tepos_10 has a much higher number of heterozygous sites. We believe this is best explained by the presence of genetically similar non-target DNA that co-enriched in the capture for this sample alone. Based on the pattern of allele frequencies, this genome was excluded from downstream analyses (Supplementary Methods 9; Supplementary Figure 7).

Tepos_20 and Tepos_37 were also excluded due to their genomic coverage of less than 6-fold, which allowed more reliable SNP calling at a minimum of 5-fold coverage for Tepos_14 and Tepos_35. Subsequent phylogenetic tree construction with 1000 bootstrap replicates revealed 100% support and branch shortening for the Tepos_14 and Tepos_35 genomes in all phylogenies, supporting their ancient origin (Fig. 3; Supplementary Figure 8).

SNP analysis for the ancient genomes together with the reference dataset yielded a total of 203,256 variant positions amongst all 25 genomes. Our analyses identified 681 positions present in one or both of the ancient genomes, where 133 are unique to the ancient lineage (Supplementary Methods 10; Supplementary Table 9). Of these, 130 unique SNPs are shared between Tepos_14 and Tepos_35, supporting their close relationship and shared ancestry. The *ydiD* gene involved in the breakdown of fatty acids[36] and the *tsr* gene related to the chemotaxic response system[37] were found to contain multiple non-synonymous SNPs (nsSNPs) unique to the ancient genomes (Supplementary Methods 10). Seven homoplastic and four tri-allelic variant positions were detected in the ancient genomes (Supplementary Methods 10; Supplementary Tables 10a, 10b).

A region of the *pil* operon consisting of five genes, *pilS*, *pilU*, *pilT*, *pilV* and *rci,* was found in our ancient genomes and was absent in the *S.* Paratyphi C RKS4594 genome[38] (Supplementary Methods 12; Supplementary Table 12). This region is located in Salmonella Pathogenicity Island 7 (SPI-7), and encodes a type IVB pili[39,40]. The version of *pilV* in our ancient genomes is thought to facilitate bacterial self-aggregation, a phenomenon that potentially aids in invasion of host tissues[39,40] (for details see Supplementary Methods 12). A further presence/absence analysis was performed to evaluate additional virulence factors. These results are summarized in Supplementary Fig. 9 and Supplementary Methods 13 (see also Supplementary Table 13).

The *S.* Paratyphi C RKS4594 strain harbours a virulence plasmid, pSPCV, which was included in our capture design. It is present at 10- to 224-fold average coverage for the five genomes (Supplementary Methods 14; Supplementary Tables 14, 15).

Non-UDG capture reads mapped to the *S.* Paratyphi C genome (NC_012125.1) for Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41, i.e. those that did not yield full genomes, had damage patterns characteristic of ancient DNA (Supplementary Figure 3). To further verify these reads as true ancient *S.* Paratyphi C reads we investigated 45 SNPs unique to Tepos_14 and Tepos_35 that are included in our phylogenetic analysis (Supplementary Methods 11; Supplementary Table 11). Of the 45 positions, between 6 and 29 were identified at minimum 1-fold in these lower coverage genomes. All of these were in agreement with the unique SNPs present in the high-coverage ancient genomes, thus confirming their shared ancestry.

Discussion

Interpretations of ethnohistorical documents have suggested some form of typhus or enteric (typhoid/paratyphoid) fever (from the Spanish "tabardillo", "tabardete", and "*tifus mortal"*), viral haemorrhagic fever, measles, or pneumonic plague as potential causes of the *cocoliztli* epidemic of 1545 CE (for ref. see Supplementary discussion 1). These diseases present symptoms similar to those that were recorded in the *cocoliztli* outbreak such as red spots on the skin, bleeding from various body orifices, and vomiting (Supplementary discussion 1; Supplementary Figure 10). Given the non-specific nature of these symptoms, additional sources of data are needed to clarify which disease(s) was/were circulating. Previous investigation of sequencing data generated from the Teposcolula-Yucundaa material did not identify DNA traces of ancient pathogens; however, *S. enterica* was not considered as a candidate[41]. Here we have isolated genome-wide data of ancient *S.* Paratyphi C from ten Mixtec individuals buried in the *Grand Plaza* epidemic cemetery at Teposcolula-Yucundaa, indicating that enteric fever was circulating in the indigenous population during the *cocoliztli* epidemic of 1545-50 CE. As demonstrated here, MALT offers a sensitive approach for screening non-enriched sequence data in search for unknown candidate bacterial pathogens involved in past disease outbreaks, even to the exclusion of a dominant environmental microbial background. Most importantly, it offers the advantage of extensive genome-level screening without the need to specify a target organism, thus avoiding ascertainment biases common to other screening approaches. Fast metagenomic profiling tools that are based on k-mer matching such as KRAKEN[42] or specific diagnostic marker regions such as MetaPhlan2[43] have limitations in ancient DNA applications. Complete alignments are needed to authenticate candidate taxonomic assignments, and a small number of marker regions might not provide sufficient resolution for identification, as target DNA is often present in low amounts. Our focus on only bacterial and DNA viral taxa limits our resolution in identifying other infectious agents that may have been present in the population during the Teposcolula-Yucundaa epidemic.

Although our discussions here have focused on a single pathogenic organism, the potential of its having acted synergistically with other pathogen(s) circulating during the epidemic must be considered. The concept of syndemics and the complex biosocial factors that influence infectious disease transmission and severity are well-documented in both modern and historical contexts[44,45]. We are currently limited to the detection of bacterial pathogens and DNA viruses included in the NCBI genomic database, though the resolution offered by MALT analyses will increase as this database grows. We have not investigated the presence of RNA viruses, since methods for RNA retrieval from archaeological tissues are underdeveloped and not supported by our current protocols[46].

We confidently exclude an environmental organism as the source for our ancient genomes on the basis that 1) *S.* Paratyphi C is restricted to humans, 2) it is not known to freely inhabit soil (our soil sample was negative for *Salmonella* during screening and

after capture), 3) the deamination patterns observed for the ancient human and *S.* Paratyphi C reads are characteristic of authentic ancient DNA, and 4) the ancient *S.* Paratyphi C genomes display expected branch-shortening in all constructed phylogenies. Moreover, we recovered all ancient genomic data from the pulp-chambers of teeth collected *in situ*, increasing the likelihood of our having identified a bacterium that was present in the victim's blood at the time of death. *S. enterica* introduction via post-burial disturbance is unlikely because the graves in the *Grand Plaza* were dug directly into the thickly paved floor at the site and historical records indicate that Teposcolula-Yucundaa was abandoned shortly after the epidemic ended in 1552 CE[29,30].

S. Paratyphi C is one of >2600 identified *S. enterica* serovars distinguished by their antigenic formula[47]. Only four serovars (*S.* Typhi and *S.* Paratyphi A, B, C), all of which cause enteric fever, are restricted to the human host[47]. Today *S.* Typhi and *S.* Paratyphi A cause the majority of reported cases[48]. *S.* Paratyphi C is rarely reported[38,48]. Infected individuals shed bacteria long after the termination of symptoms[47], and in the case of *S.* Typhi infection, 1-6% of individuals become asymptomatic carriers[49]. Following the hypothesis that this disease was introduced via European contact, it is conceivable that asymptomatic European carriers who withstood the cross-Atlantic voyage could have introduced *S.* Paratyphi C to Mesoamerican populations in the 16th century. First hand descriptions of the 1545 *cocoliztli* epidemic suggest that both European and Mixtec individuals were susceptible to the disease[7,50], with one estimate of a 60-90% population decline in New Spain during this period[7].

The additional SPI-7 genes detected through indel analysis are reported to vary in presence/absence amongst modern *S.* Paratyphi C strains[38,40], and are suspected to cause increased virulence when the inverted repeats in *pilV* allow the Rci recombinase to shuffle between its two protein states (Supplementary Methods 12). This may support an increased capacity for our ancient strains to cause an epidemic outbreak. However, the overall mechanisms through which *S.* Paratyphi causes enteric fever remain unclear. The nsSNPs in the *ydiD* and *tsr* genes may signify adaptive processes, and comparison with a greater number of *S.* Paratyphi C genomes may clarify this[51].

Today, *S.* Paratyphi C is rare in Europe and the Americas, with more cases identified across Africa and Asia[52,53]. Based on multilocus sequence typing (MLST) data from modern *S.* Paratyphi C strains, no clear phylogeographic pattern has been observed[52]. However, the presence of a 1200 CE *S.* Paratyphi C genome in Norway indicates its presence in Europe in the pre-contact era[51], which would be necessary for it to be considered an Old World disease. However, based on the small number of pre-contact individuals that we have screened, we cannot exclude the presence of *S.* Paratyphi C at Teposcolula-Yucundaa prior to European arrival. A local origin for the *cocoliztli* disease has been proposed elsewhere[54]. Historical accounts offer little perspective on its origin since neither the indigenous population nor the European colonizers had a pre-existing name for the disease[7,8,30]. Spanish colonial documents refer to it as

pujamiento de sangre ('full bloodiness'), while the indigenous Aztec population of Central Mexico called it *cocoliztli*, a generic term meaning 'pestilence' in Nahuatl[7,8] (see Supplementary discussion 1).

Little is known about the past severity and worldwide incidence of enteric fever, first determined to be distinct from typhus in the mid-nineteenth century[55]. Enteric fevers are regarded as major health threats across the world[48], causing an estimated ~27 million illnesses in 2000, the majority of which were attributed to *S.* Typhi[56]. Due to the rarity of *S.* Paratyphi C diagnoses, mortality rates are not established for this particular serovar. Today, outbreaks predominantly occur in developing countries. *S.* Typhi/Paratyphi are commonly transmitted through the faecal-oral route via ingestion of contaminated food or water[57]. Changes imposed under Spanish rule such as forced relocations under the policy of *congregación*, altered living arrangements, and new subsistence farming practices[29,30] compounded by drought conditions[32] could have disrupted existing hygiene measures, facilitating *S.* Paratyphi C transmission.

Our study represents a first step towards a molecular understanding of disease exchange in contact era Mexico. The 1545 *cocoliztli* epidemic is regarded as one of the most devastating epidemics in New World history[7,32]. Our findings contribute to the debate concerning the causative agent of this epidemic at Teposcolula-Yucundaa, where we propose that *S.* Paratyphi C be considered. We introduced MALT, a novel fast alignment and taxonomic assignment method. Its application to the identification of ancient *Salmonella enterica* DNA within a complex background of environmental microbial contaminants speaks to the suitability of this approach, and its resolution will improve as the number of available reference genomes increases. This method may be eminently useful for studies wishing to identify pathogenic agents involved in ancient and modern disease, particularly for cases where candidate organisms are not known *a priori.*

Methods

The MALT algorithm

MALT is based on the seed-and-extend paradigm and consists of two programs, *malt-build* and *malt-run*.

First, *malt-build* is used to construct an index for the given database of reference sequences. To do so, malt-build determines all occurrences of spaced seeds[58-60] in the reference sequences and places them into a hash table[61].

Following this, *malt-run* is used to align a set of query sequences against the reference database. To this end, the program generates a list of spaced seeds for each query and then looks them up in the reference hash table, which is kept in main memory. Using the x-drop extension heuristic[25], a high-scoring ungapped alignment anchored at the seed is computed and is used to decide whether or not a full alignment should be constructed. Local or semi-global alignments are computed using a banded implementation[62] of the Smith-Waterman[63] or Needleman-Wunsch[64] algorithms, respectively. The program then computes the bit-score and expected value (E-value) of the alignment and decides whether to keep or discard the alignment depending on user-specified thresholds for bit-score, E-value or percent identity. The application of *malt-run* is illustrated in Supplementary Figure 1.

The MALT screening pipeline

In order to use MALT in ancient DNA contexts to screen for bacterial DNA and to assess the taxonomic composition of ancient bacterial communities we applied the following workflow (Supplementary Figure 2). First we used *malt-build* to construct a MALT index on all complete bacterial genomes in GenBank[65]. This was done only once, and is rebuilt only when the target database requires updating. We align reads to the reference database using *malt-run* in semi-global mode. MALT generates output in RMA format and in SAM format. The former can be used for interactive analysis of taxonomic composition in MEGAN[28] and the latter for alignment-based assessment of damage patterns and other authenticity criteria.

Sample provenience

The site of Teposcolula-Yucundaa is situated on a mountain ridge in the Mixteca Alta region of Oaxaca, Mexico. Prior archaeological excavation at this site revealed a large epidemic cemetery located in the Grand Plaza – the town's administrative centre, and an additional cemetery in the churchyard (Fig. 1; Supplementary Methods 1). 24 teeth were collected from individuals buried in the Grand Plaza cemetery and five from individuals buried in the churchyard cemetery (Supplementary Methods 2; Supplementary Table 1). Soil samples were also collected from both cemetery sites (Supplementary Methods 2).

DNA extraction and library preparation

DNA extracts and double-stranded indexed libraries compatible with Illumina sequencing were generated using methods tailor-made for ancient DNA[66-68]. This work

was carried out in dedicated ancient DNA cleanroom facilities at the University of Tübingen and Harvard University (Supplementary Methods 2).

Screening with MALT

Amplified libraries were shotgun sequenced. The reads were adapter clipped and merged before being analysed with MALT and the results visualized in MEGAN6[28] (Supplementary Methods 2, 3). Two MALT runs were executed. The first using all complete bacterial genomes available through NCBI RefSeq (December, 2016), and the second using the full NCBI Nucleotide (nt) database (ftp://ftp-trace.ncbi.nih.gov/blast/db/FASTA/) as reference to screen for viral DNA (Fig. 2; Supplementary Methods 3; Supplementary Tables 2, 3). Both runs used 'semi-global' alignment and a minimum percent identity of 95% (Supplementary Methods 3). The shotgun data was also mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1) and the human genome (hg19) and damage plots were generated (Supplementary Methods 3, 4; Supplementary Table 4, 5; Supplementary Figure 3).

Runtime comparison

The programs MALT (version 0.3.8) and BLAST[25] (version 2.6.0+) were applied to the shotgun screening data of Tepos_35 consisting of altogether 952,511 reads. For both programs the DNA alignment mode (blastn) was chosen. The maximal E-value was set to 1.0. The maximal number of alignments for each query was set to 100. The minimal per cent identity was set to 95. The number of threads was set to 16. The alignment type of MALT was set to 'Local' in order to be comparable to BLAST. The total amount of RAM required by MALT during this run was 252.7 GB.

For MALT the runtime was measured excluding the initial loading of our reference database, which happens only once when screening multiple samples. The loading of the database takes 27.27 minutes. Including taxonomic binning the application of MALT to our complete shotgun screening data took 123.36 minutes. As a comparison, processing only the screening data of a single sample (Tepos_35) with BLASTn took 1420.58 minutes without any taxonomic analysis. Processing of this sample alone with MALT, including taxonomic binning, took 6.48 minutes, which constitutes a 200-fold improvement in terms of computation time.

The computations were performed on a Dell PowerEdge R820 with four Intel Xeon E5-4620 2.2 GHz CPUs und 768 GB RAM.

Probe design and whole-genome capture

Array probes were designed based on 67 publicly available *S. enterica* chromosomes/assemblies and 45 associated plasmid sequences (Supplementary Methods 5; Supplementary Table 6). Extracts from samples deemed to be positive for *S. enterica* Paratyphi C were converted into additional rich UDG-treated libraries[69] for whole-genome capture (Supplementary Methods 6, 7). Pre-contact and post-contact samples were serially captured using our custom probe design, according to two established methods[35,70]. The eluate from both array and in-solution capture was

sequenced to a sufficient depth to allow high coverage genome reconstruction (Supplementary Methods 6, 7).

Sequence data processing, initial phylogenetic assessment and authenticity

The sequence data was adapter clipped and quality filtered before being mapped to the *S.* Paratyphi C reference (NC_012125.1) (Supplementary Methods 8; Supplementary Table 7). Deamination patterns for the DNA were generated to assess the authenticity of the ancient *S.* Paratyphi C DNA using mapDamage[23] (Supplementary Methods 8; Supplementary Table 7; Supplementary Figure 3). Artificial read data was generated for a dataset of 23 genomes selected for comparative phylogenetic analysis; this data was also mapped to the *S.* Paratyphi C reference (Supplementary Methods 8). SNP calling was carried out with the Genome Analysis Toolkit (GATK) using a quality score of ≥30 for the five *S.* Paratyphi C genomes and the artificial read dataset. A neighbor-joining tree was constructed using MEGA6[71], based on a homozygous SNPs called at a minimum of 3-fold coverage where at least 90% of reads are in agreement (Supplementary Methods 8; Supplementary Figure 4). In order to exclude a reference bias in the ascertainment of the phylogenetic positioning of the five ancient genomes (Tepos_10, Tepos_14 Tepos_20, Tepos_35 and Tepos_37), mapping, SNP calling and tree construction was repeated for the *S.* Typhi CT18 reference (NC_003198.1) (Supplementary Methods 8; Supplementary Table 8; Supplementary Figure 5).

SNP typing and phylogenetic analysis

Homozygous SNPs were called from the complete dataset (5 ancient and 23 modern) based on our criteria using a tool called MultiVCFAnalyzer (Supplementary Methods 9). Repetitive and highly conserved regions of the *S.* Paratyphi C genome (NC_012125.1) were excluded from SNP calling to avoid spurious mapping reads. Maximum Parsimony[71] and a Maximum Likelihood[72] trees were made including the five genomes (Fig. 3; Supplementary Figure 6). Heterozygous positions were also called and their allele frequency distributions plotted using R[73] (Supplementary Methods 9; Supplementary Figure 7). SNP calling and phylogenetic tree construction was repeated excluding the Tepos_10, Tepos_20 and Tepos_37 genomes (Supplementary Methods 10; Fig. 3; Supplementary Figure 8).

The five weak-positive samples: Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41, that did not yield enough data for genome reconstruction, were investigated for 46 SNPs unique to the ancient genomes to verify that the captured reads for these samples are true ancient *S.* Paratyphi C reads (Supplementary Methods 11; Supplementary Table 11).

SNP effect and Indel analyses

SNP effect analysis was carried out for the two ancient genomes (Tepos_14 and Tepos_35) alongside the modern dataset (see Supplementary Methods 10). SNPs unique to the ancient genomes, pseudogenes and homoplastic positions were investigated (Supplementary Methods 10; Supplementary Tables 9, 10). Insertions and

deletions (Indels), 700bp or larger, in the two ancient genomes were identified through two approaches. Deletions were visually detected by mapping the ancient data to the *S.* Paratyphi C reference using a mapping quality threshold (-q) of 0 and manually viewing the genome alignment in the Integrative Genomics Viewer (IGV) (Supplementary Methods 12). In order to detect insertions, or regions present in the ancient genomes that are missing the modern reference, the ancient data was mapped to concatenated reference pairs. Where one reference was in all cases the *S.* Paratyphi C RKS4594 reference (NC_012125.1) and the other was one of four *S. enterica* genomes of interest. A mapping quality threshold (-q) of 37 was used, thus allowing only regions unique to one or the other genome in the pair to map (Supplementary Methods 12; Supplementary Table 12).

Virulence factor analysis

43 effector genes identified within *Salmonella enterica subsp. enterica*[74] were investigated using the BEDTools suite[75]. The percentage of each gene that was covered at least 1-fold in the ancient and modern genomes in our dataset was plotted using the ggplot2package[76] in R[73] (Supplementary Methods 13; Supplementary Figure 9).

Plasmid analysis

The ancient data was mapped to the *S.* Paratyphi C virulence plasmid, pSPCV. SNP effect analysis was carried out in comparison to three other similar plasmid references (Supplementary Methods 14; Supplementary Tables 14, 15).

Data Availability

Sequence data that support the findings of this study have been submitted to the European Nucleotide Archive under accession number PRJEB23438 (https://www.ebi.ac.uk/ena/data/view/PRJEB23438). MALT is open source and freely available from: http://ab.inf.uni-tuebingen.de/software/malt. The program MultiVCFAnalyzer is available on GitHub: https://github.com/alexherbig/MultiVCFAnalyzer. Source data for figures are available upon request.

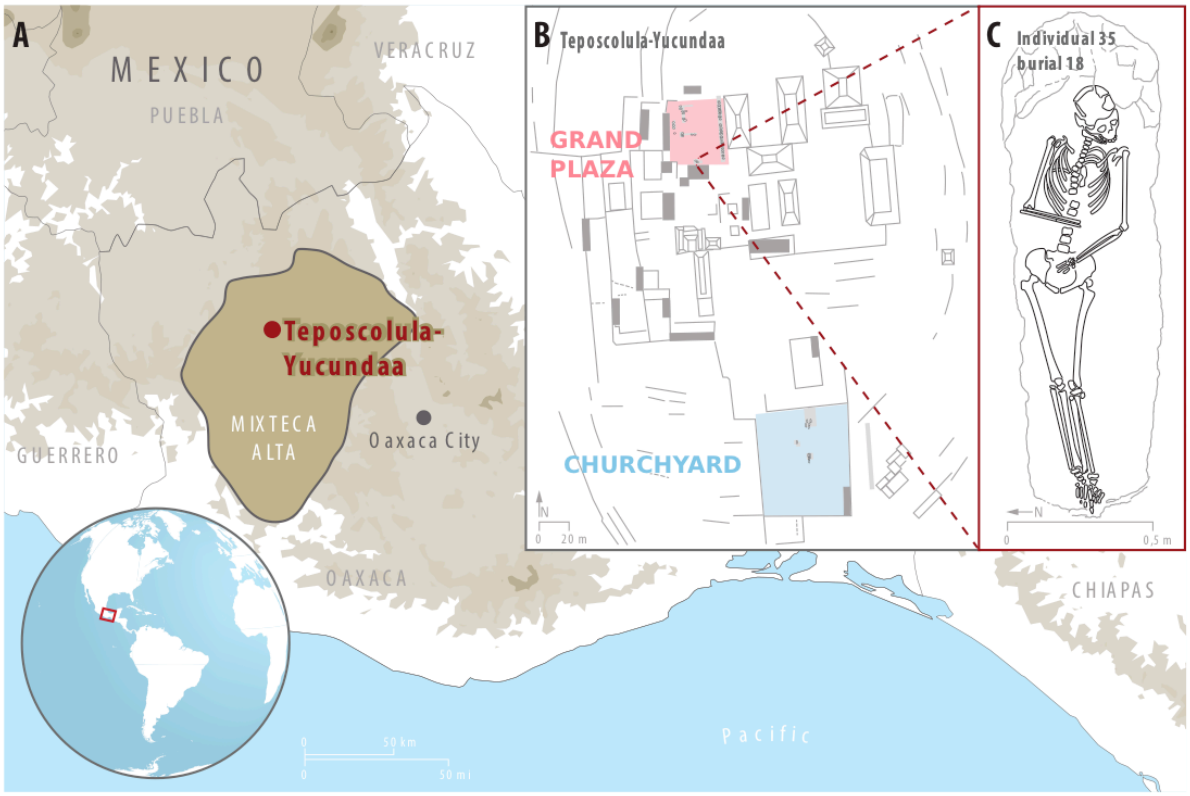


Figure 1 | Overview of Teposcolula-Yucundaa. A) The location of the Teposcolula-Yucundaa (15.502500°N, 97.467493°W) site in the Mixteca Alta region of Oaxaca, Mexico; B) The central administrative area of Teposcolula-Yucundaa showing the relative positioning of the Grand Plaza and churchyard cemetery sites. Burials within each cemetery are indicated with dark grey outlines, and the excavation area is shaded in grey; C) Drawing of individual 35 from which the Tepos_35 *S.* Paratyphi C genome was isolated. Panels B and C are adapted from drawings provided by the Teposcolula-Yucundaa archaeological project archives-INAH and Christina Warinner. The figure was produced by A. Günzel.

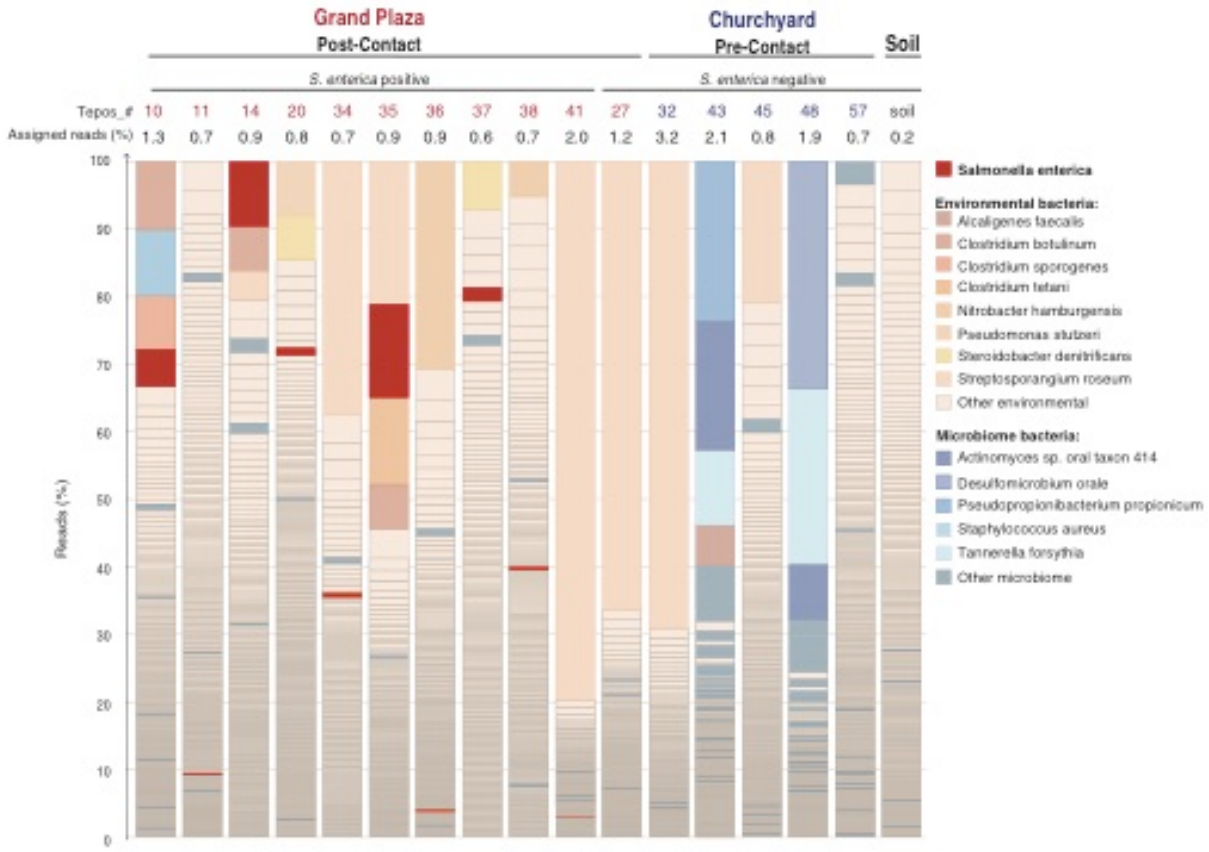


Figure 2 | MALT analysis and pathogen screening of shotgun data. Shotgun data were analysed with MALT using a database constructed from all bacterial genomes available through NCBI RefSeq (December 2016). MALT results were visualized using MEGAN6 (ref.[28]). The bar chart was constructed from the MEGAN6 output and is based on the per cent reads assigned to bacterial species when using a 95% identity filter. Reads assigned to *Salmonella enterica* are coloured red regardless. Other taxa to which ≥3% reads, per sample, were assigned are colour-coded depending on whether they are 'environmental' or 'human oral microbiome' bacteria. Remaining taxa are sorted into two categories: 'other environmental' or 'other microbiome' (Supplementary Methods 3). Samples from the post-contact Grand Plaza epidemic cemetery containing *S. enterica* reads, pre-contact era samples from the churchyard cemetery and the soil sample are illustrated. Additionally, a sample negative for *S. enterica* from the Grand Plaza cemetery (Tepos_27) is shown. Samples whose names are coloured in red are from the Grand Plaza and those in blue from the churchyard. The percentage of reads in the shotgun data assigned by MALT per sample is indicated at the top of each column. Only taxa with ≥4 reads assigned are visualized.

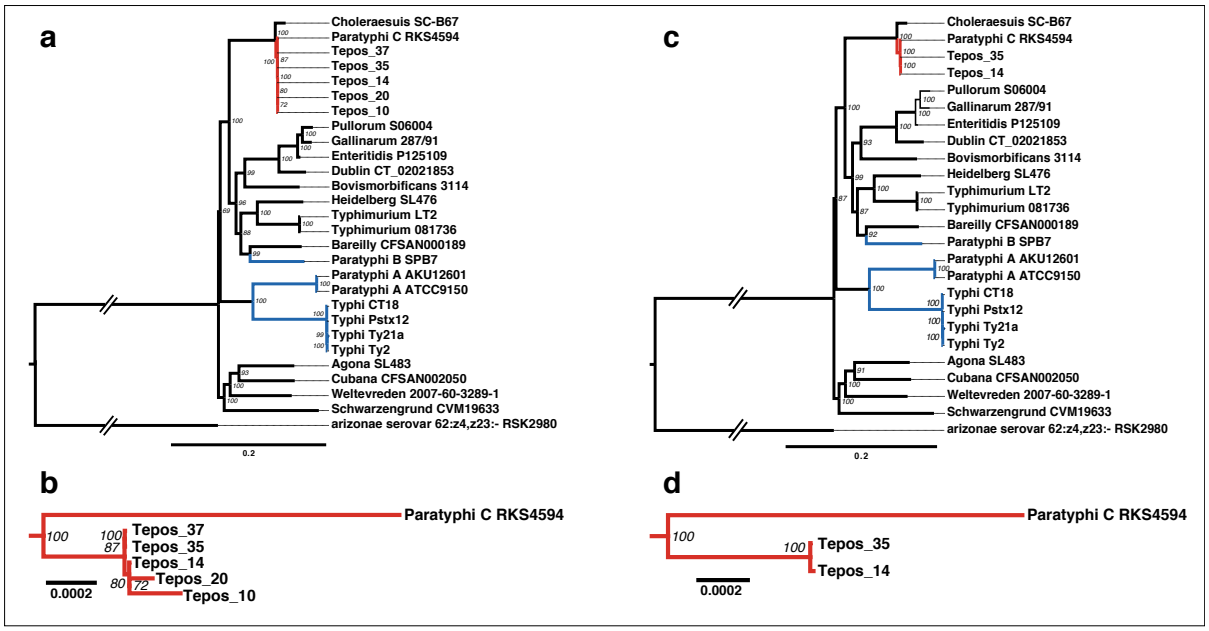


Figure 3 | Maximum Likelihood trees for *S. enterica* subsp. *enterica* phylogeny. Two maximum likelihood trees were produced. Positions with missing data were excluded in both cases. A) The tree includes all five ancient genomes and is based on 3-fold SNP calls and 51,456 variant positions. B) A zoomed in view of the *S.* Paratyphi C genomes. C) The tree includes two high-coverage ancient genomes and is based on 5-fold SNP calls and 81,474 variant positions. D) A zoomed in view of the *S.* Paratyphi C genomes, illustrating the branch shortening of the two ancient genomes (Tepos_14 and Tepos_35). Both trees were built with RAxML[72]. Branches coloured red indicate *S.* Paratyphi C genomes, and branches in blue indicate other genomes that are human-specific and cause enteric (typhoid/paratyphoid) fever.

Table 1 | Overview of mapping statistics of captured sample libraries from the Grand Plaza (contact era) and churchyard (pre-contact)

Sample ID	Cemetery Site	Library treatment	# processed reads before mapping	# Unique mapped reads	Endogenous DNA (%) - quality filtered reads	Mean Fold Coverage	% of Genome Covered at least 3-fold
Tepos_10	Grand Plaza	non-UDG	16,945,834	399,561	20.56	4.35	52.17
		UDG	68,628,270	2,903,258	16.30	32.84	95.49
Tepos_14	Grand Plaza	non-UDG	20,559,478	1,222,402	23.51	14.41	95.77
		UDG	73,204,225	3,410,610	18.62	36.44	97.67
Tepos_35	Grand Plaza	non-UDG	27,248,720	1,803,043	31.37	25.50	97.67
		UDG	90,815,050	7,025,774	30.00	96.43	98.06
Tepos_11	Grand Plaza	non-UDG	21,941,119	19,576	0.87	0.21	0.93
		UDG	48,959,732	103,492	0.75	1.21	14.56
Tepos_20	Grand Plaza	non-UDG	771,431	15,236	6.94	0.15	0.26
		UDG	20,123,713	427,781	4.75	4.59	67.53
Tepos_34	Grand Plaza	non-UDG	18,934,710	123,307	2.55	1.35	14.65
		UDG	26,284,766	157,930	2.05	1.74	21.67
Tepos_36	Grand Plaza	non-UDG	23,147,904	36,224	0.75	0.40	1.76
		UDG	21,910,196	33,327	0.42	0.36	1.4
Tepos_37	Grand Plaza	non-UDG	5,223,138	218,874	9.28	2.55	42.12
		UDG	9,603,890	416,449	7.71	5.49	74.48
Tepos_38	Grand Plaza	non-UDG	8,280,412	18,308	0.91	0.19	0.97
		UDG	47,835,731	65,812	0.54	0.67	4.22
Tepos_41	Grand Plaza	non-UDG	17,608,445	33,664	0.72	0.37	1.47
		UDG	19,966,958	36,208	0.48	0.40	1.34
Tepos_27	Grand Plaza	non-UDG	17,931,300	4,778	0.07	0.04	0.27
Tepos_32	Churchyard	non-UDG	25,721,427	6,665	0.08	0.06	0.47
Tepos_43	Churchyard	non-UDG	31,129,662	3,426	0.05	0.03	0.25
Tepos_45	Churchyard	non-UDG	18,027,289	6,879	0.12	0.06	0.34
Tepos_48	Churchyard	non-UDG	17,915,341	4,312	0.06	0.04	0.25
Tepos_57	Churchyard	non-UDG	24,478,844	5,527	0.07	0.05	0.28
Soil	Grand Plaza & churchyard	non-UDG	10,875,300	796	0.02	0.01	0.07

References:

1 Ubelaker, D. H. Prehistoric New World population size: Historical review and current appraisal of North American estimates. *Am J Phys Anthropol* **45**, 661-665, doi:10.1002/ajpa.1330450332 (1976).

2 Crosby, A. W. Virgin soil epidemics as a factor in the aboriginal depopulation in America. *William Mary Q* **33**, 289-299 (1976).

3 Dobyns, H. F. Disease Transfer at Contact. *Annu Rev Anthropol* **22**, 273-291 (1993).

4 Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Griffin, R. D. & Cleaveland, M. K. When half of the population died: the epidemic of hemorrhagic fevers of 1576 in Mexico. *FEMS Microbiol Lett* **240**, 1-5, doi:10.1016/j.femsle.2004.09.011 (2004).

5 Llamas, B. *et al*. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv* **2**, e1501385, doi:10.1126/sciadv.1501385 (2016).

6 Lindo, J. *et al*. A time transect of exomes from a Native American population before and after European contact. *Nat Commun* **7**, 13175, doi:10.1038/ncomms13175 (2016).

7 Cook, N. D. & Lovell, W. G. *Secret Judgments of God: Old World Disease in Colonial Spanish America*. (University of Oklahoma Press, 2001).

8 Fields, S. L. *Pestilence and Headcolds: Encountering Illness in Colonial Mexico*. (Columbia University Press, 2008).

9 Ortner, D. J. *Identification of pathological conditions in human skeletal remains*. 2 edn, (Academic Press, 2003).

10 Walker, R. S., Sattenspiel, L. & Hill, K. R. Mortality from contact-related epidemics among indigenous populations in Greater Amazonia. *Sci Rep* **5**, 14032, doi:10.1038/srep14032 (2015).

11 Joralemon, D. New World Depopulation and the Case of Disease. *J Anthropol Res* **38**, 108-127 (1982).

12 Larsen, C. S. In the wake of Columbus: Native population biology in the postcontact Americas. *Am J Phys Anthropol* **37**, 109-154, doi:10.1002/ajpa.1330370606 (1994).

13 Bos, K. I. *et al*. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494-497, doi:10.1038/nature13591 (2014).

14 Schuenemann, V. J. *et al*. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science* **341**, 179-183, doi:10.1126/science.1238286 (2013).

15 Warinner, C. *et al*. Pathogens and host immunity in the ancient human oral cavity. *Nat Genet* **46**, 336-344, doi:10.1038/ng.2906 (2014).

16 Bos, K. I. *et al*. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506-510, doi:10.1038/nature10549 (2011).

17 Maixner, F. *et al*. The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**, 162-165, doi:10.1126/science.aad2545 (2016).

18 Devault, A. M. *et al*. Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. *Sci Rep* **4**, 4245, doi:10.1038/srep04245 (2014).

19 Bos, K. I. *et al*. Parallel detection of ancient pathogens via array-based DNA capture. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370**, 20130375, doi:10.1098/rstb.2013.0375 (2015).

20 Devault, A. M. *et al*. A molecular portrait of maternal sepsis from Byzantine Troy. *Elife* **6**, doi:10.7554/eLife.20983 (2017).

21 Warinner, C. *et al*. A Robust Framework for Microbial Archaeology. *Annu Rev Genomics Hum Genet*, doi:10.1146/annurev-genom-091416-035526 (2017).

22 Key, F. M., Posth, C., Krause, J., Herbig, A. & Bos, K. I. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet* **33**, 508-520, doi:10.1016/j.tig.2017.05.005 (2017).

23 Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682-1684, doi:10.1093/bioinformatics/btt193 (2013).

24 Prufer, K. *et al*. Computational challenges in the analysis of ancient DNA. *Genome Biol* **11**, R47, doi:10.1186/gb-2010-11-5-r47 (2010).

25 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

26 Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* **16**, 363, doi:10.1186/s12859-015-0788-5 (2015).

27 Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci Rep* **6**, 19233, doi:10.1038/srep19233 (2016).
28 Huson, D. H. *et al*. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol* **12**, e1004957, doi:10.1371/journal.pcbi.1004957 (2016).
29 Spores, R. & Robles García, N. A prehispanic (postclassic) capital center in colonial transition: excavations at Yucundaa Pueblo Viejo de Teposcolula, Oaxaca, Mexico. *Lat Am Antiq* **18**, 33-353 (2007).
30 Warinner, C., Robles García, N., Spores, R. & Tuross, N. Disease, Demography, and Diet in Early Colonial New Spain: Investigation of a Sixteenth-Century Mixtec Cemetery at Teposcolula Yucundaa. *Lat Am Antiq* **23**, 467-489 (2012).
31 Tuross, N., Warinner, C. & Robles García, N. in *Yucundaa: La cuidad mixteca Yucundaa-Pueblo Viejo de Teposcolula y su transformación prehispánica-colonial* Vol. vol. 2 (eds Ronald Spores & Robles García Nelly) 541-546 (Instituto Nacional de Antropología e Historia, 2014).
32 Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and megadeath in 16th century Mexico. *Emerg Infect Dis* **8**, 360-362 (2002).
33 Pickard, D. *et al*. Molecular characterization of the Salmonella enterica serovar Typhi Vi-typing bacteriophage E1. *Journal of bacteriology* **190**, 2580-2587, doi:10.1128/JB.01654-07 (2008).
34 Burbano, H. A. *et al*. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**, 723-725, doi:10.1126/science.1188046 (2010).
35 Fu, Q. *et al*. DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2223-2227, doi:10.1073/pnas.1221359110 (2013).
36 Campbell, J. W., Morgan-Kiss, R. M. & Cronan, J. E., Jr. A new Escherichia coli metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. *Molecular microbiology* **47**, 793-805 (2003).
37 Rivera-Chavez, F. *et al*. Salmonella uses energy taxis to benefit from intestinal inflammation. *PLoS Pathog* **9**, e1003267, doi:10.1371/journal.ppat.1003267 (2013).
38 Liu, W. Q. *et al*. Salmonella paratyphi C: genetic divergence from Salmonella choleraesuis and pathogenic convergence with Salmonella typhi. *PLoS One* **4**, e4510, doi:10.1371/journal.pone.0004510 (2009).
39 Tam, C. K., Morris, C. & Hackett, J. The Salmonella enterica serovar Typhi type IVB self-association pili are detached from the bacterial cell by the PilV minor pilus proteins. *Infect Immun* **74**, 5414-5418, doi:10.1128/IAI.00172-06 (2006).
40 Tam, C. K., Hackett, J. & Morris, C. Salmonella enterica serovar Paratyphi C carries an inactive shufflon. *Infect Immun* **72**, 22-28 (2004).
41 Campana, M. G., Robles Garcia, N., Ruhli, F. J. & Tuross, N. False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Res Notes* **7**, 111, doi:10.1186/1756-0500-7-111 (2014).
42 Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46, doi:10.1186/gb-2014-15-3-r46 (2014).
43 Segata, N. *et al*. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811-814, doi:10.1038/nmeth.2066 (2012).
44 Singer, M. & Clair, S. Syndemics and public health: reconceptualizing disease in bio-social context. *Med Anthropol Q* **17**, 423-441 (2003).
45 Herring, D. A. & Sattenspiel, L. Social Contexts, Syndemics, and Infectious Disease in Northern Aboriginal Populations. *American Journal of Human Biology* **19**, 190-202, doi:10.1002/ajhb (2007).
46 Guy, P. L. Prospects for analyzing ancient RNA in preserved materials. *Wiley Interdiscip Rev RNA* **5**, 87-94, doi:10.1002/wrna.1199 (2014).
47 Gal-Mor, O., Boyle, E. C. & Grassl, G. A. Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ. *Front Microbiol* **5**, 391, doi:10.3389/fmicb.2014.00391 (2014).
48 Wain, J., Hendriksen, R. S., Mikoleit, M. L., Keddy, K. H. & Ochiai, R. L. Typhoid fever. *Lancet* **385**, 1136-1145, doi:10.1016/S0140-6736(13)62708-7 (2015).
49 Monack, D. M., Mueller, A. & Falkow, S. Persistent bacterial infections: the interface of the pathogen and the host immune system. *Nat Rev Micro* **2**, 747-765 (2004).
50 Sahagún, B. d. *Florentine Codex: general history of the things of New Spain*. (1950-1982).
51 Zhou, Z. *et al*. Millennia of genomic stability within the invasive Para C Lineage of *Salmonella enterica*. *bioRxiv*, doi:10.1101/105759 (2017).

52 Achtman, M. *et al*. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. *PLoS Pathog* **8**, e1002776, doi:10.1371/journal.ppat.1002776 (2012).

53 Centers for Disease Control and Prevention (CDC). National Typhoid and Paratyphoid Fever Surveillance Annual Summary, 2014. (Atlanta, Georgia, 2016).

54 Acuna-Soto, R., Romero, L. C. & Maguire, J. H. Large epidemics of hemorrhagic fevers in Mexico 1545-1815. *Am J Trop Med Hyg* **62**, 733-739 (2000).

55 Smith, D. C. Gerhard's distinction between typhoid and typhus and its reception in America, 1833-1860. *Bull Hist Med* **54**, 368-385 (1980).

56 Crump, J. A., Luby, S. P. & Mintz, E. D. The global burden of typhoid fever. *Bull World Health Organ* **82**, 346-353 (2004).

57 World Health Organization. *Typhoid fever – Uganda*, <http://www.who.int/csr/don/17-march-2015-uganda/en/> (2015).

58 Burkhardt, S. & Kärkkäinen, J. in *Combinatorial Pattern Matching: 12th Annual Symposium, CPM 2001 Jerusalem, Israel, July 1–4, 2001 Proceedings* (ed Amihood Amir) 73-85 (Springer Berlin Heidelberg, 2001).

59 Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440-445 (2002).

60 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

61 Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-1729, doi:10.1101/gr.194201 (2001).

62 Chao, K. M., Pearson, W. R. & Miller, W. Aligning two sequences within a specified diagonal band. *Comput Appl Biosci* **8**, 481-487 (1992).

63 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).

64 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).

65 Benson, D. A. *et al*. GenBank. *Nucleic Acids Res* **41**, D36-42, doi:10.1093/nar/gks1195 (2013).

66 Dabney, J. *et al*. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15758-15763, doi:10.1073/pnas.1314445110 (2013).

67 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

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

69 Briggs, A. W. *et al*. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* **38**, e87, doi:10.1093/nar/gkp1163 (2010).

70 Hodges, E. *et al*. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974, doi:10.1038/nprot.2009.68 (2009).

71 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).

72 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

73 R Development Core Team. (The R Foundation for Statistical Computing, Vienna, Austria, 2011).

74 Connor, T. R. *et al*. What's in a Name? Species-Wide Whole-Genome Sequencing Resolves Invasive and Noninvasive Lineages of Salmonella enterica Serotype Paratyphi B. *MBio* **7**, doi:10.1128/mBio.00527-16 (2016).

75 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

76 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

Acknowledgements: This work was supported by the Max Planck Society (J.K.), the European Research Council (ERC) starting grant APGREID (to J.K.), Social Sciences and Humanities Research Council of Canada postdoctoral fellowship grant 756-2011-501 (to K.I.B.), and the Mäxi Foundation (M.G.C.). We thank the Archaeology Council at Mexico's National Institute of Anthropology and History (INAH) and the Teposcolula-Yucundaa Archaeological Project for sampling permissions. We are grateful to Antje Wissgott, Guido Brandt and Verena Schuenemann for assistance with laboratory work, Annette Günzel for providing graphic support, and Rodrigo Barquera, Jim Hackett and Menaka Pi for thoughts and discussion on the manuscript. Part of the data storage and analysis was performed on the computational resource bwGRiD Cluster Tübingen funded by the Ministry of Science, Research and the Arts Baden-Württemberg, and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC. We thank the MALT user community for helpful comments and bug reports.

Author Contributions: K.I.B., M.G.C., A.H., N.T. and J.K. conceived the investigation. K.I.B., A.H. Å.J.V., M.G.C. and J.K. designed experiments. N.M.R.G. provided archaeological information and drawings, submitted INAH permits and assisted in the sampling processes. Å.J.V., M.G.C., S.S., M.A.S. and K.I.B. performed laboratory work. Å.J.V., A.H., K.I.B., C.W. and A.A.V. performed analyses. D.H. implemented the MALT algorithm. A.H., J.K. and D.H. designed and set up the MALT ancient DNA analysis pipeline. C.W. performed ethnohistorical analyses. Å.J.V. and K.I.B. wrote the manuscript with contributions from all authors.

Competing financial interests: The authors declare no competing financial interests.

Salmonella enterica genomes from victims of a major 16th century epidemic in Mexico

Supplementary Information

Åshild J. Vågene[1], Alexander Herbig[1], Michael G. Campana[2, 3]†, Nelly M. Robles García[4], Christina Warinner[1], Susanna Sabin[1], Maria A. Spyrou[1], Aida Andrades Valtueña[1], Daniel Huson[5], Noreen Tuross[2], Kirsten I. Bos[1] and Johannes Krause[1]

[1]Max Planck Institute for the Science of Human History, Jena, Germany.

[2]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA.

[3]Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland.

[4]INAH, National Institute of Anthropology and History, Mexico, Teposcolula-Yucundaa Archaeological Project.

[5]Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Tübingen, Germany.

†Current address: M.G.C.: Smithsonian Conservation Biology Institute, Center for Conservation Genomics, 3001 Connecticut Avenue NW, Washington, DC 20008, USA.

CONTENTS:

Refer to CD for Supplementary Tables 2-5, 7, 9, and 11

Supplementary Methods

1. Archaeological context

The site of Teposcolula-Yucundaa sits on a mountain ridge in the Mixteca Alta, northwest of the city of Oaxaca, Mexico. Prior to the arrival of the Spanish, the *señorío* (*yuhuitayu*, or local state) of Teposcolula-Yucundaa controlled a large Mixtec territory with an estimated population of 60,000 (7,000-8,000 in the urban core)[1-3], and was well connected to regional and long-distance trade routes[4]. From 1458-1520 CE, Teposcolula-Yucundaa was a subject of the Aztec Triple Alliance, paying tribute in gold, cochineal, textiles, jade beads, and quetzal feathers[4-6]. Following the conquest of Mexico Tenochtitlan in 1521 CE, Teposcolula became a subject of the Spanish crown, and at least six parties of early Spanish explorers and conquistadors passed through the region from 1520-1525 CE[7]. Teposcolula was held briefly in encomienda from 1527-1531 CE, after which it re-established as a corregimiento[8]. The site became the focus of Dominican evangelism in the 1530s[9,10] and at least fifteen Dominican friars were assigned to Teposcolula from 1538-1552 CE[11]. From 1544-1546 CE, Teposcolula-Yucundaa hosted an inquisition trial against the ruler of the neighbouring *yuhuitayu* of Yanhuitlan, at which the *yya toniñe* (native lords) of several local Mixtec states provided testimony[12,13]. It is noted in this testimony that the vicar of Teposcolula, Friar Bernardino de Santa María, fell ill while performing translation at the trial, but the nature of this illness is unknown.

Few records survive for Teposcolula-Yucundaa's early colonial history, but historical documents from the neighbouring *yuhuitayu* of Coixtlahuaca (located 30 km to the northeast of Teposcolula) record a disease outbreak in 1545 CE, in which it is lamented that they could not keep pace with the burial of the 30-40 people dying each day at the height of the epidemic[14]. In official questionnaires dating to the late 1570s, four other neighbouring Mixtec communities (Mitlatongo, Tamazola, Justlahuaca, and Tecomaxtlahuaca) reported that during the period 1530-1577 CE the only epidemic to reach the Mixteca Alta was the 1545-1548 CE pestilence[15]. At Teposcolula-Yucundaa, the toll taken by disease is evidenced by a large plaster covered cemetery located in the Grand Plaza (administrative square), which is estimated to contain as many as 800 individuals[5,16,17]. In 1552 CE the Yucundaa site was abandoned and its people were relocated to a new location 2 km south that exists today as the modern town of San Pedro y San Pablo Teposcolula[2,5].

During 2004-2010, archaeologists led by Dr. Ronald Spores and Dr. Nelly Robles García (INAH) excavated the ancient town of Teposcolula-Yucundaa, including burials encountered in the churchyard and the Grand Plaza. A detailed description of the excavation is found in Spores and Robles García (2007)[2]. Multiple occupation periods were

noted, including a Postclassic phase that predates the arrival of the Spanish. In addition, an early Convento phase was associated with the Dominican presence[18]. Direct radiocarbon dating confirmed the presence of pre-contact individuals in the churchyard, while the majority of the radiocarbon dates from individuals excavated from the Grand Plaza overlap with the arrival of the Spanish[19] (see Supplementary Table 1). Multiple simultaneous burials found in the Grand Plaza, as well as disproportional mortality of young adults in the absence of skeletal trauma, suggests a high rate of disease causing death among the indigenous Mixtec peoples[5].

2. DNA extraction and library preparation

Teeth were collected from 29 individuals (one from each individual) excavated at the Grand Plaza (n=24) and churchyard (n=5) cemeteries at Teposcolula-Yucundaa (Supplementary Table 1). Each tooth was sectioned at the cementoenamel junction and a sample was drilled from the crown pulp chamber. Samples were processed according to an established protocol tailored for extracting DNA from archaeological bone[20], samples were rotated during the lysis step for at least 16 hours. One extraction blank was added for every ten samples processed per batch and a positive control (bone powder from an ancient cave bear) was included in every batch. DNA extracts were eluted in 100µl of TET (10 mM Tris-Cl, pH 8.0; 1 mM EDTA, pH 8.0; 0.05% Tween-20).

Libraries for screening were generated using 10µl of extract[21]. One library blank was added for every ten samples processed. All libraries were double-indexed[22] using a combination of custom 8nt P5 and P7 Illumina indexes in a 10-cycle PCR reaction. Indexed libraries were further amplified using AccuPrime (Thermo Scientific) enzyme with IS5/IS6 primers to reach a concentration of 1x10^13 copies per reaction. All steps from sampling to setting up the indexing reactions were carried out in facilities dedicated to ancient DNA work at the University of Tübingen. Libraries were diluted and pooled into an equimolar solution of 10nmol/l and shotgun sequenced on either a HiSeq 2000 or NextSeq 500. Sequencing yielded between 1,708,868-4,457,666 paired-end reads for the samples and 276,176-1,446,088 for the negative controls (see Supplementary Table 4).

Additionally, an aggregate soil sample consisting of soil taken near three skeletons – two in the Grand Plaza (individuals 2 (not investigated in this study) and 26) and one in the churchyard (individual 32) – was collected for DNA screening (Supplementary Table 1). The aggregate soil sample was extracted at Harvard University using the PowerMax Soil Maxi Kit, concentrated via ultracentrifugation (30 kDA) and sheared using a Covaris (430 sec, PeakPower 175.0, Duty Factor 10.0, cycles/burst 200). A library was generated using 10µl of extract and indexed in the modern laboratory facilities at the University of Tübingen, using the same method as above. Shotgun sequencing was carried out on a HiSeq

2500 yielding 12,100,244 reads for the soil sample and 748,026 reads for its associated library blank (Supplementary Table 4).

3. Screening with MALT

The shotgun data generated for all tooth pulp chamber samples, the soil sample and negative controls were screened for ancient bacterial pathogen DNA using the bioinformatics tool MALT (Megan ALignment Tool), a tool specifically designed for the analysis of metagenomic data. MALT is a rapid sequence alignment tool that uses a reference database. For the construction of the reference database all complete bacterial genomes and plasmids were downloaded from the NCBI FTP server on the 1st of December 2016. The following entries were excluded:
Burkholderia multivorans ATCC,
Nitrosospira multiformis ATCC 25196ATCC 2519619375,
Nakamurella multipartita DSM 44233DSM 4423320804,
Acidiphilium multivorum AIU301AIU30119783,
Sulfurospirillum multivorans DSM 12446DSM 1244651002,
Chania multitudinisentens RB-25RB-2535133,
Burkholderia pseudomultivorans SUB-INT23-BP219466,
Desulfococcus multivorans DSM 205921754,
Burkholderia multivorans ATCC 17616ATCC 1761619466,
Burkholderia multivoransDDS 15A-119466,
Burkholderia multivorans ATCC BAA-247ATCC BAA-24719466,
Burkholderia multivorans AU118519466,
Burkholderia multivorans MSMB1640WGS19466.

malt-build (version 0.3.8) was used to create the MALT target database using default parameters. The inclusion of all complete bacterial genomes in the reference database allows for the identification of both pathogenic and non-pathogenic bacteria. MALT uses a taxonomic binning approach (LCA) to assign reads to their taxonomic node of best fit based on the alignment of each read against every reference genome in the database (see Methods).

The data from all samples was de-indexed and the EAGER pipeline[23] was used to perform adapter clipping and paired-end read merging. Only merged reads were used as input for MALT (version 0.3.8). The MALT run was performed using 95 as the "minimum percent identity" parameter (--minPercentIdentity). The minimum support parameter (--minSupport) was set to 1, i.e. only nodes with a minimum support of 1 read are kept. BlastN mode and SemiGlobal alignment were applied and a top per cent value (--topPercent) of 1

was set. All other parameters were set to default. MALT results were viewed in MEGAN6[24].

Using this approach we identified samples from ten individuals (Tepos_10, Tepos_11, Tepos_14, Tepos_20, Tepos_34, Tepos_35, Tepos_36, Tepos_37, Tepos_38 and Tepos_41) and one extraction blank (EB2-091013) that contained reads ranging from 4 to 659 assigned to *Salmonella enterica* (Supplementary Table 2). The 15 reads assigned to *S. enterica* in EB2-091013 likely arose from contaminants stemming from the lab or elsewhere. We considered three samples to be strong positives: Tepos_10, Tepos_14 and Tepos_35, respectively harbouring 365, 384 and 659 assigned reads. And the remaining seven sample libraries that harboured between 4 and 51 assigned reads to be weak positives. For all samples the majority of reads cluster at the downstream *S. enterica* subsp. *enterica* node, and for samples Tepos_10, Tepos_14 and Tepos_35: 10, 5 and 20 reads were specifically assigned to *S.* Paratyphi C (NC_012125.1), respectively. All ten *S. enterica* positive individuals were excavated from the Grand Plaza epidemic cemetery.

A further four negative controls and nine samples had 1 to 2 reads assigned to *S. enterica*, including two samples from the pre-contact churchyard cemetery (Tepos_32 and Tepos_48) (Supplementary Table 2). However, we did not qualify these samples as 'positive' for *S. enterica*. Upon closer inspection of the reads in MEGAN it was observed that these reads were largely assigned to *S. enterica* species that inhabit the soil. After capture these negative controls and churchyard samples were considered to be negative for *S. enterica* (see Supplementary methods 8). All other samples investigated from the Grand Plaza and churchyard cemeteries, the soil sample and negative controls were negative for *S. enterica* DNA in the MALT screening.

A taxon table of the MALT results for shotgun data from all samples and blanks is shown in Supplementary Table 2. The metagenomic profiles of selected samples are visualized in Figure 2. For Figure 2, the Human Oral Microbiome Database (HOMD)[25] was consulted in order to classify the microbial taxa as likely members of the 'human oral microbiome' or as 'environmental'. This list of oral taxa was manually validated using articles available in NCBI's PubMed database, and bacterial taxa that are included in the HOMD, but taxa that predominantly occur in the environment were counted as exceptions and were classified as 'environmental'. In most cases, these taxa are common food contaminants found in soil. These species include: *Achromobacter xylosoxidans*, *Acinetobacter baumannii*, *Agrobacterium tumefaciens*, *Burkholderia cepacia*, *Comamonas testosterone*, *Kytococcus sedentarius*, *Mesorhizobium loti*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, *Pseudomonas stutzeri*, *Ralstonia pickettii*, *Rhodobacter capsulatus*, *Sanguibacter keddieii* and *Variovorax paradoxus*.

Visual inspection of the reads assigned within *S. enterica* by MALT for samples Tepos_10, Tepos_14 and Tepos_35 revealed mismatches consistent with an ancient DNA deamination damage pattern, evidenced by C to T transitions on the 5-prime end and G to A transitions on the 3-prime end of fragments. In order to confirm the visually observed deamination pattern the merged reads used as MALT input were mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1) using the Burrows-Wheeler-Aligner (BWA, v. 0.7.12)[26] with parameters adjusted to accommodate deaminated bases (-l 16, -n 0.01, -q 37). Mapping statistics are shown for all samples and blanks in Supplementary Table 4. mapDamage 2.0[27] plots were subsequently generated from the mapping reads, where the percentage of reads with deamination of the first bases on the 5-prime end was 22.86%, 21.62% and 17.13% of reads respectively for Tepos_10, Tepos_14 and Tepos_35 (Supplementary Figure 3; Supplementary Table 4). Deamination plots could not be reliably generated for the weak positive samples based on the shotgun data due to the low number of mapping reads (Supplementary Table 4).

The shotgun data was additionally screened with MALT using all complete genomes in the NCBI Nucleotide (nt) database downloaded from ftp://ftp-trace.ncbi.nih.gov/blast/db/FASTA/ on the 7th December 2016. The purpose of this screening was to see if we could detect DNA from any DNA viruses, or RNA viruses that have DNA life stages, which may have been circulating and acting synergistically in the Teposcolula-Yucundaa population, in addition to *S. enterica* as indicated by our bacterial MALT screening results, during the 1545-1550 CE epidemic.

Only one notable viral taxon hit was detected that was not related to environmental soil organisms (e.g. *Clostridium*). Eight of the sample libraries from Grand Plaza individuals had reads (1 to 40) assigned to '*Salmonella* phage Vi II-E1'. This phage is associated with *Salmonella enterica* subsp. *enterica* serovars that produce the Vi capsule, a virulence associated polysaccharide capsular antigen, carried by three known serovars: *S.* Typhi, *S.* Dublin and *S.* Paratyphi C[28,29]. DNA virus screening results are presented in Supplementary Table 3.

4. Human DNA analysis

The non-UDG shotgun data for samples and negative controls were mapped to the human genome (hg19) using non-UDG parameters. Endogenous human DNA ranged from 0.021 to 36.8% for the tooth samples, 0.19 to 46.7% for the blanks and was 0.008% for the soil sample (Supplementary Table 5). Damage patterns were estimated using mapDamage 2.0[27]. The human DNA in the blanks does not have a damage pattern and is likely stemming from human DNA contamination introduced from reagents and plastic ware during extraction and/or library preparation. The characteristic ancient DNA damage pattern is

present in 25 of the 29 human tooth pulp chamber samples (Supplementary Table 5). The remaining four samples (Tepos_9, Tepos_12, Tepos_13 and Tepos_19) do not exhibit a damage pattern, likely due to the low amount of preserved human DNA. It may also be that no human DNA was preserved in these four samples and the DNA that is present stems from modern contaminant sources.

5. Probe design for capture

To further verify the finding of *S. enterica* DNA in samples Tepos_10, Tepos_14 and Tepos_35 via MALT based screening and to attempt whole-genome reconstruction, we designed a set of probes for whole-genome enrichment of *S. enterica* for array-based hybridization capture.

Probes were designed based on 112 publicly available reference sequences (67 chromosomes/assemblies and 45 plasmids; see Supplementary Table 6 for details). These reference genomes were selected based on modern strain diversity within the species *Salmonella enterica*. The probes were designed to be 60bp long and have a tiling density of 7bp across the template. This was achieved by generating two different sets of probes with 15bp tiling density each, which differ from each other by a coordinate offset of 7bp (versions A and B). Low complexity repetitive and duplicate probes were excluded from the final probe set. This produced 928,395 and 928,078 unique probes for versions A and B respectively. By randomly sampling probes each probe set was enlarged to 968,000 probes, as this is the maximum number of probes that can be included on an Agilent One-million feature array.

We decided to capture the seven *S. enterica* weak positive samples at a later date. At this time the established protocol in the lab for genome capture was in-solution capture according to Fu *et al.* (2013)[30]. The probe set was generated using the same genome dataset as described above. The probes were designed with a tiling density of 7bp across the template with a probe length of 52bp with an additional 8bp 3-prime linker sequence as described in Fu *et al.* (2013)[30]. Two probe sets were generated using 15bp tiling density as before, which yield a 7bp tiling density when combined. Two probe sets were generated, consisting of 931,667 and 931,866 unique probes, which were respectively printed on two Agilent One-million feature arrays. As before, each probe set was raised to 968,000 probes by randomly sampling probes. The two arrays were turned into in-solution DNA capture libraries as described elsewhere[30].

6. Array capture

Concentrated libraries were made using 30-40µl extract from samples Tepos_10, Tepos_14 and Tepos_35. Prior to library preparation the DNA extracts were pre-treated with USER enzyme (New England BioLabs), which contains Uracil DNA glycosylase (UDG) and endonuclease VIII (endoVIII). UDG removes uracil residues, which are most commonly located in 5-prime and 3-prime overhangs in ancient DNA, creating an abasic site that is cleaved and removed by endoVIII. This is done to avoid the incorporation of incorrect bases during amplification[31]. Allowing for more stringent mapping parameters to be used in the reconstruction of full genomes and excludes erroneous nucleotide substitutions from interfering with downstream data analyses. Indexing and further amplification was done as described above using the enzyme Herculase II Fusion DNA Polymerase (Agilent).

UDG and non-UDG libraries for samples Tepos_10, Tepos_14 and Tepos_35 were amplified to make two pools of 20µg, where 75% of each pool was dedicated to equimolar quantities of the UDG treated libraries and the remaining 25% to equimolar quantities of the non-UDG libraries used for screening. Each pool was serially captured using versions A and B of the array; together we term these the 'MALT-positives' array.

The five non-UDG screening libraries made from the pre-contact churchyard samples (Tepos_32, Tepos_43, Tepos_45, Tepos_48, Tepos_57), one sample (Tepos_27) from the Grand Plaza cemetery negative for *S. enterica* in the MALT screening and the soil sample were amplified and pooled in equimolar amounts to make a 20µg pool. This pool was serially captured on a version A array; we term this the 'MALT-negatives' array. A fourth 10µg equimolar pool was made consisting of negative controls carried along during extraction and library preparation. The 'negative controls' pool was captured in a single round on a version A array.

Array capture was performed according to an established method[32]. The eluate from the first round of capture performed for all arrays was quantified on the qPCR using IS5/IS6 primers and amplified using Herculase II Fusion DNA Polymerase. The 'MALT-positives' and 'MALT-negatives' array eluate was further amplified up to 17µg and serially captured on identical arrays to those used in the first round. The eluted product was quantified as above and re-amplified using IS5/IS6 primers. The product from the 'MALT-positives' and 'MALT-negatives' arrays were diluted and pooled in equimolar amounts to create a 10nmol/l sequencing pool. The pool was paired-end sequenced (2x75bp cycles) on a NextSeq 500. The capture product for the 'negative controls' array was sequenced separately on part of a HiSeq 4000 paired-end run (2x75bp cycles). This yielded between 19,758,038 and 192,213,420 reads per sample library (non-UDG and UDG) and 670,824 and 3,859,834 reads for the negative controls (Supplementary Table 7).

7. In-solution capture

UDG treated indexed libraries were prepared, according to the protocols outlined above, for the seven weak-positive samples: Tepos_11, Tepos_20, Tepos_34, Tepos_36, Tepos_37, Tepos_38 and Tepos_41 using between 40-45µl extract. Two library blanks were carried along.

In-solution capture was carried out according to an established protocol[30]. All samples previously captured on the array were re-captured with in-solution capture, alongside the weak positive samples. Each sample was captured in a separate well. The input DNA for the positive samples consisted of 25% non-UDG treated library and 75% UDG treated library, mirroring the array capture approach in this respect. Only non-UDG libraries were captured for the pre-contact samples, the Grand Plaza negative (Tepos_27) and the soil sample. The extraction blank that had 15 reads assigned to *S. enterica* in MALT for the shotgun screening was captured separately (EB2-091013), while all other negative controls were pooled and captured together in one well. The in-solution capture eluate received two rounds of sequencing. The first was a shallow single-end sequencing run on part of a HiSeq 4000 (1x75bp cycles), yielding between 236,434 and 7,088,118 raw reads per sample library (non-UDG and UDG) and 87,341 and 2,316,301 for the negative controls, used to judge the outcome of the capture, using the mapping approach described below in Supplementary methods 8. Based on these results we conducted a deeper round of sequencing for all samples on the NextSeq 500 (2x75bp cycles) producing between 1,154,734 and 99,573,998 paired-end reads per sample library (non-UDG and UDG) and between 208,090 and 6,031,192 for the negative controls.

Preparation of the weak positive libraries, the in-solution capture and subsequent sequencing was carried out in the clean room and modern lab facilities of the Max Planck Institute for the Science of Human History Department of Archaeogenetics in Jena, Germany.

8. Read processing, mapping and ascertainment of phylogenetic positioning

All sequenced capture data were de-indexed using bcl2fastq (Illumina; http://support.illumina.com/downloads/bcl2fastq-conversion-software-v217.html) and further processed using the EAGER pipeline[23] (v.1.92.54) to clip adapters, merge paired-end reads, map the data using BWA (v. 0.7.12)[26], remove duplicates, execute mapDamage 2.0[27] and carry out SNP calling with the GATK UnifiedGenotyper[33]. Only merged reads were used in all analyses of paired-end data.

All data were adapter clipped, and paired-end data were merged with AdapterRemoval v.2[34].
Merged paired-end data from samples that were captured via both array and in-solution capture was combined for further processing. For these samples the low coverage single-end data, produced from the in-solution capture, were not used for analyses. This allowed us to apply DeDup, the duplicate removal tool implemented in EAGER, which considers both 5-prime and 3-prime ends of fragments. Keeping only merged paired-end data allowed us to gain higher coverage as we could observe the real 3-prime end of the sequenced molecule.

For the weak positives, which were not previously array captured, both the single-end and the paired-end data generated from the in-solution capture were combined after adapter clipping and merging of the paired-end data. MarkDuplicates, which only considers the 5-prime end, was used to remove duplicates. For the negative controls both sets of paired-end data and the single-end data were merged prior to mapping.

All data were subsequently mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1). BWA mapping parameters were adjusted depending on whether the library was pre-treated with UDG or not. UDG treated libraries were mapped with more stringent parameters (BWA parameters: -l 32; -n 0.1; -q 37) than non-UDG libraries (BWA parameters: -l 16; -n 0.01; -q 37). Mapping results show that non-UDG libraries for Tepos_27, the pre-contact samples and the soil sample contained 6,879 or fewer unique mapping reads after capture (see Supplementary Table 7). Whereas non-UDG libraries for the three strong positives contain 399,561 to 1,803,043 unique mapping reads, and the seven weak positives contain 15,236 to 218,874 unique mapping reads. UDG treated libraries for the strong positives contain 2,903,258 to 7,025,774 unique mapping reads and the weak positives contain 33,327 to 427,781 unique mapping reads (Supplementary Table 7). The UDG treated data used for downstream genome analyses yielded average coverages of 33-, 36- and 96-fold respectively for the strong positives: Tepos_10, Tepos_14 and Tepos_35. Two lower coverage genomes were also captured from the weak positives for Tepos_20 and Tepos_37 that had average genome coverages of 4.6- and 5.5-fold, respectively. The remaining weak positives had low coverage genome-wide data resulting in average coverages ranging from 0.35- to 1.75-fold.

Deamination patterns generated with mapDamage[27] for the non-UDG capture data for Tepos_10, Tepos_14 and Tepos_35 yielded 19,83%, 28,31% and 19,7% deamination of the first base on the 5-prime ends of the reads. These numbers differ from the deamination values yielded by the shotgun data, likely because they are based on a higher number of reads than previously estimated using the shotgun reads (Supplementary Figure 3; Supplementary Tables 4, 7). Deamination rates, after capture, for the seven weak positive

non-UDG libraries ranged from 12.32% to 23.8% of the first base on the 5-prime ends of the reads.

All negative controls, regardless of whether they were non-UDG or UDG treated, were mapped to the S. *paratyphi C* reference using non-UDG parameters (-l 16; -n 0.01; -q 37) with BWA (v. 0.7.12). The negative controls contained between 174 and 2,911 unique mapping reads after capture. The EB2-091013 extraction blank that had 15 reads assigned to *S. enterica* in the MALT screening had 691 unique mapping reads after capture with an extremely high read duplication factor of 137. These reads exhibited a damage pattern of 11.67% C to T transitions on the first base of the 5-prime end (Supplementary Table 7), likely due to cross contamination during processing before the addition of unique indices. We also observed an "ancient DNA"-like damage pattern for Tepos_32, one of the pre-contact samples, corresponding to 8.12% C to T changes in the 5-prime ends for the 6,665 mapping reads (Supplementary Table 7). To better determine the origin of these reads, we performed an additional MALT analysis for the captured data of all pre-contact samples, Tepos_27, the soil sample and all negative controls. The MALT analysis was performed as outlined in Supplementary Methods 3 using the database of 6,427 complete bacterial genomes from the NCBI FTP server (downloaded 1st December 2016). Subsequently, the MALT output was visualized in MEGAN and all reads summarized at the *Salmonella enterica* subsp. *enterica* node for each library were extracted. The extracted reads, numbering between 363 and 34,002, were subsequently mapped to the *S.* Paratyphi C genome (NC_012125.1) using non-UDG parameters (see above). Lower numbers of mapping reads were reported from the extracted reads than from the direct mapping of the capture data. Of the extracted reads, the pre-contact samples had between 61 and 255 unique mapping reads, where Tepos_32 had 61. Tepos_27 and the soil sample had 127 and 27 unique mapping reads, respectively. The negative controls had between 39 and 1,270 unique mapping reads. Deamination patterns were investigated, and none were observed for Tepos_32, nor for any of the remaining libraries with the exception of EB2-091013 that retained a pattern of 17.54% damage at the first base from the 5-prime end based on its 184 unique mapping reads. From the low number of mapping reads and the lack of molecular damage expected of ancient DNA, we determined that the pre-contact samples, Tepos_27, the soil sample and all negative controls (except for EB2-091013), were indeed negative for *S. enterica* DNA even after capture.

Artificial read data (100bp reads with 1bp tiling density) was generated from a subset of the genomes used in the array design consisting of 23 fully scaffolded or assembled *S. enterica* genomes using an in-house script. The artificial read data was mapped against the *S.* Paratyphi C RKS4594 reference (NC_012125.1) using stringent UDG mapping parameters (see above).

The dataset used for initial phylogenetic analyses consisted of the genomes reconstructed from the UDG treated capture data for samples Tepos_10, Tepos_14, Tepos_20, Tepos_35 and Tepos_37, in addition to the 23 genomes reconstructed from the artificial read data mapped against *S.* Paratyphi C RKS4594. SNP calling was carried out for all dataset genomes using the 'EMIT_ALL_SITES' function, providing a call for all variant or non-variant bases in the *vcf* file output.

We used an in-house tool (MultiVCFanalyzer) to collate homozygous SNPs (90% of reads covering a position must be in agreement) called at a minimum of 3X coverage against the *S.* Paratyphi C RKS4594 reference for the three captured UDG treated positive samples and the artificial reference genome dataset. The collated SNP alignment was used to generate a neighbor-joining tree in MEGA6[35] (Supplementary Figure 4). This tree shows that the five captured genomes cluster with *S.* Paratyphi C with a high bootstrap support of 100, confirming the initial taxonomic indication provided by MALT. Oddly, the Tepos_10 genome has a much longer branch length compared to the two other high-coverage ancient genomes (Tepos_14 and Tepos_35) (see also Supplementary methods 9).

In order to exclude the possibility of a reference bias in the ascertainment of the phylogenetic positioning, we mapped the UDG data for the five ancient genomes using the approach outlined above, but this time against the *S.* Typhi CT18 reference (NC_003198.1). All captured genomes show a decrease in percentage of the reference genome covered and a decrease in average coverage from *S.* Paratyphi C (Supplementary Table 7) to *S.* Typhi (Supplementary Table 8). SNP calling was also repeated for the genome dataset with the parameters described above. A neighbor-joining tree was constructed, confirming the positioning of the ancient genomes with *S.* Paratyphi C (Supplementary Figure 5).

Capture efficiency was calculated using the *S.* Paratyphi C mapping results for the paired–end non-UDG shotgun and paired-end capture data for the ten positive samples, for the array capture and the in-solution capture separately. Capture efficiencies were calculated based on the quality filtered mapping reads before duplicate removal. For the three strong positives capture efficiency was estimated to 386-, 400- and 356-fold increase for the array and 523-, 596- and 400-fold increase for in-solution capture for Tepos_10, Tepos_14 and Tepos_35 respectively. For the weak positive samples, in-solution capture efficiency was estimated to 552-, 974-, 1104-, 529-, 1230-, 383- and 702-fold increase for Tepos_11, Tepos_20, Tepos_34, Tepos_36, Tepos_37, tepos_38 and Tepos_41 respectively.

9. SNP typing and phylogenetic analysis

SNP calls generated for all 28 genomes in the dataset were compared in parallel using an in-house Java tool (MultiVCFanalyzer). MultiVCFanalyzer outputs a multi-genome SNP-alignment with an entry for all genomes where at least one variant position is called within the dataset. Homozygous positions were called where 90% or more of the reads covering a position were in agreement, a minimum of 3 reads were covering the position and the position's GATK quality score was a minimum of 30. Homozygous calls were also made in cases where GATK had called a heterozygous position, but the above requirements were still met. Non-variant positions meeting the above criteria were called as the reference base and positions with missing data or those that did not meet the above requirements were inserted as an 'N' in the SNP-alignment. Positions called in repetitive regions, phage-related regions, recombination-related regions or regions prone to cross-mapping from other organisms were excluded. These regions were identified based on two genome annotation (*gff*) files for the *S.* Paratyphi C RKS4595 reference genome (NC_012125.1). The two *gff* files can be found respectively through the current version (ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Salmonella_enterica/) and the archived older version (ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Salmonella_enterica_serovar_Paratyphi_C_RKS4594_uid59063/) of the FTP archive of publicly available bacterial genomes. Regions excluded were identified in the *gff* files as: mobile elements, phages or phage-related, transposase, resolvase, rRNA, tRNA, repeat-region, insertion sequence or as a recombination protein.

208,411 variant positions were called from the total dataset. The five ancient genomes had a respective number of SNP calls of 702, 676, 446, 684 and 496 for Tepos_10, Tepos_14, Tepos_20, Tepos_35 and Tepos_37. The multi-genome SNP alignment consisting of homozygous calls for the dataset was used to construct a Maximum Parsimony tree in MEGA6[35] and a Maximum Likelihood tree in RAxML[36] using the GTR gamma model (Fig. 3; Supplementary Figure 6). Complete deletion was used, restricting phylogenetic analysis to the core genome.

Heterozygous positions in the five ancient genomes were investigated due to the extensively long branch observed for the Tepos_10 genome compared to the two other ancient genomes (Fig. 3; Supplementary Figures 4, 5, 6). Heterozygous positions were called in MultiVCFanalyzer using the parameters described above (3X SNP calling), where additionally all positions with a SNP allele frequency between 10-90% were typed as heterozygous. 5,519, 985, 175, 395 and 130 heterozygous positions were called respectively for Tepos_10, Tepos_14, Tepos_20, Tepos_35 and Tepos_37. The distribution of SNP allele frequencies for each of the five ancient samples is shown in Supplementary

Figure 7. The tail of the distribution of heterozygous positions in the Tepos_10 genome infringes upon the threshold of 90% set for calling homozygous positions. This indicates that the homozygous calls for this genome cannot be considered to be reliable. In light of this, Tepos_10 was excluded from further analyses.

Tepos_20 and Tepos_37 were also excluded from further analyses due to their lower coverage, making them unsuitable for 5X SNP calling. Only the two clean, high coverage genomes, Tepos_14 and Tepos_35, were used for in-depth genome analyses.

A new SNP alignment was generated for the dataset (Tepos_14, Tepos_35 and 23 modern genomes), comprising 203,256 variant positions. A Maximum Parsimony tree was generated in MEGA[35] and a Maximum Likelihood tree in RAxML[36], using the GTR gamma model and complete deletion, based on homozygous positions called from the dataset (Fig. 3; Supplementary Figure 8). All trees show a bootstrap support of 100 for the phylogenetic positioning of the ancient genomes with *S.* Paratyphi C. The two ancient genomes, Tepos_14 and Tepos_35, exhibit branch shortening in comparison to the modern strain.

10. SNP analysis of protein coding genes

A SNP table of variant positions occurring in at least one strain within the dataset comprised of Tepos_14, Tepos_35 and the 23 reference genomes, was generated using MultiVCFanalyzer applying the same parameters as described above for generating the SNP alignment; with 5X SNP calling. The SNP table was annotated using the *gff* file for the *S.* Paratyphi C RKS4594 genome (NC_012125.1) located in the archived version of the FTP archive.

The annotated SNP table of variant positions was used as input for the bioinformatics tool snpEff[37], which predicts the amino acid changes and effects of the variant SNP positions on genes. MultiVCFanalyzer was subsequently used to collate the output from snpEff with the SNP table dataset. In total 203,256 homozygous SNPs are present in at least one genome in the dataset. 681 variant SNP positions are present in one or both of the two ancient genomes, Tepos_14 and Tepos_35. 339 of these are non-synonymous SNPs (nsSNPs): with 326 nsSNPs leading to an amino-acid codon change, 9 cause the loss of a stop-codon (STOP_LOST) and 4 are stop mutations (STOP_GAINED). 211 are synonymous changes and 13 are considered as non-coding because they occur in RNA related genes.

133 variant positions are unique to one or both of the ancient strains. Tepos_14 has two unique nsSNPs occurring in the *yfbU* and *yhcK* genes. The *yfbU* gene is annotated as

encoding a 'hypothetical protein'. Whilst the *yhcK* gene, or *nanR,* is involved in sialic acid utilization and regulates the *nan* operon by repressing it in the absence of sialic acid[38]. It is worth noting that Tepos_10 and Tepos_20 (and potentially Tepos_36) also share the particular SNP in the *yhcK* gene (Supplementary methods 11; Supplementary Table 11). Only one SNP is unique to Tepos_35 and it is intergenic. A table comprising all SNPs occurring in one or both of the ancient strains is shown in Supplementary Table 9.

46 genes contain two or more variant SNPs (Supplementary Table 9). Two of these genes are of particular note with regard to the ancient genomes. In the *ydiD* gene two nsSNPS and one sSNP occur, the sSNP and one of the nsSNPs are unique to the ancient strains. The *ydiD* gene encodes for a putative acyl-CoA synthetase involved in the breaking down of fatty acids[39]. In the *tsr* gene there are two SNPs, one non-synonymous and one synonymous, which are uniquely shared by both the ancient strains within the dataset. *Tsr* codes for a methyl-accepting chemotaxis protein involved in serine sensing, steering the bacterium towards host-sources of nitrate[40]. In *S.* Typhimurium, the *tsr* gene is associated with enhanced rates of infection in the mouse intestine[40,41].

Nine 'STOP_LOST' mutations were identified in the ancient genomes in comparison to the *S.* Paratyphi C reference (Supplementary Table 9). However, these nine pseudogenes present in the *S.* Paratyphi C reference are active genes or missing in the other 22 modern genomes. This suggests that the modern *S.* Paratyphi C RKS4594 strain is the odd one out based on the diversity of *Salmonella enterica* strains included in this analysis. Three of the four 'STOP_GAINED' mutations are unique to the two ancient strains within the dataset. Two affect genes coding for hypothetical proteins (SPC_0289 and SPC_4426), the third affects the *rbsA* gene, coding for part of an ATP-binding cassette, which is involved in nutrient uptake[42].

Seven homoplastic SNP positions were detected (Supplementary Table 10a). Only one is non-synonymous and is located in the *phsC* gene. Within the dataset this homoplasy is shared with strains *S.* Typhimurium 08-1736 and *S.* Dublin CT-02021853. The *phsC* gene is part of the *phsABC* operon in *S. enterica* that encodes a thiosulfate reductase, which metabolizes, or reduces, thiosulfate into hydrogen sulfide[43] providing an alternate energy source in anaerobic environments. Thiosulfate is readily available in the mammalian-gut, and is suggested to support bacterial growth of *S. enterica* during gut colonization and potentially contributing to pathogenesis[44]. Additionally, the ancient genomes share two nsSNPs, one each in the *phsA* and *phsB* genes which make up the other two genes in the *phsABC* operon with all other strains included in the analysis (excluding *S. arizonae* where it is absent). This may be some indication that the ancient *S.* Paratyphi C strains metabolized thiosulfate differently than the modern *S.* Paratyphi C RKS4594 reference strain. Four SNP positions were also found to be tri-allelic within the analysed dataset

(Supplementary Table 10b). Two tri-allelic SNP positions are non-synonymous. One is located in the *bah* gene, which putatively codes for acetyl esterase (http://www.uniprot.org/uniprot/Q57HI6). The second occurs in SPC_3759, which codes for a hypothetical protein.

11. SNP analysis of weak positive samples

The following five weak positive samples, Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41, did not yield sufficient data to allow genome reconstruction and phylogenetic placement. Non-UDG capture data for these samples display clear deamination patterns characteristic of ancient DNA (Supplementary Figure 3; Supplementary Table 7). To further verify the captured reads for these five weak positive samples as true ancient *S.* Paratyphi C reads, 46 SNPs unique to the ancient genomes were investigated from the UDG treated data for these samples. We chose SNPs of phylogenetic significance, i.e. that are displayed in the Maximum Parsimony tree after complete deletion. These SNPs are comprised of 17 shared by all five ancient genomes (with 3X SNP calling) and a further 28 SNPs shared by the two high-coverage genomes Tepos_14 and Tepos_35 (with 5X SNP calling), see Supplementary Figures 6 and 8. A single additional SNP shared by Tepos_10, Tepos_14 and Tepos_20, illustrated in Supplementary Figure 6, was also investigated. The SNP positions were manually assessed by viewing the data in IGV[45]. All reads that covered an investigated SNP position were in agreement.

Of the 45 SNPs placed along the main branch shared by the ancient genomes, 24, 29, 6, 14 and 7 positions for Tepos_11, Tepos_34, Tepos_36, Tepos_38 and Tepos_41, respectively, were in agreement with the SNP calls for the other ancient genomes. In all other cases, no reads covered the position. No SNPs were in agreement with the reference call. The remaining SNP position 3,442,532 shared by Tepos_10, Tepos_14 and Tepos_20 was also present in Tepos_36. Indicating that Tepos_36 might belong within this cluster. Tepos_11, Tepos_38 and Tepos_41 display the reference call at this position, in agreement with Tepos_35 and Tepos_37. No reads covered this position in Tepos_34. Results are shown in Supplementary Table 11.

12. Indel analysis

Insertions and deletions (indels) in the Tepos_14 and Tepos_35 genomes were identified through two different approaches. Deletions in the ancient genomes larger than 700bp in comparison to the *S.* Paratyphi C RKS4594 reference were identified based on visual inspection using the IGV browser[45]. The ancient data was mapped to the reference using UDG parameters with a mapping quality (-q) of 0. Thus, reads that map equally well at more than one point in the genome are kept in the alignment. Only one region absent in both of the ancient genomes (Tepos_14 and Tepos_35) was identified. This region is

~1,841bp in length, spanning positions 1,355,411 to 1,357,252 in the *S.* Paratyphi C RKS4594 genome (NC_012125.1). This region contains the prophage related gene SPC_1297, which encodes for the terminase large subunit (TerL). TerL is part of the machinery that prophages use to translocate DNA and package it into empty capsid heads[46]. This absent region was a part of the probe design and its absence is not due to a capture bias. An additional mapping was performed where the less stringent (non-UDG) parameters (with –q 0) were used to verify the absence of the region in the case that it should be present, but with a high number of mismatches to the reference. However, it was still not detected as part of the ancient genomes.

In order to investigate regions present in the ancient genomes that are absent in the *S.* Paratyphi C RKS4594 genome, the ancient UDG treated genome data was mapped to four concatenated reference pairs, where one genome in every pair was the *S.* Paratyphi C RKS4594 (NC_012125.1) and the other was one of the following: *S.* Choleraesuis SC-B67 (NC_006905.1), *S.* Paratyphi A ATCC-9150 (NC_006511.1), *S.* Paratyphi B SGSC-4150 (NC_010102.1) or *S.* Typhi CT18 (NC_003198.1). These four genomes were chosen because *S.* Typhi and *S.* Paratyphi A, B cause human specific enteric fever, while *S.* Choleraesuis is the closest relative of *S.* Paratyphi C and can cause disease in both pigs and humans. Mapping was done with standard UDG mapping parameters, and because the parameter for mapping quality (-q) was set to 37, reads that map equally well at more than one position across the two concatenated references were discarded, meaning only reads unique to either the *S.* Paratyphi C or its paired genomes were mapped. Through this method several genes were determined to be present in the ancient genome capture data that were not present in the *S.* Paratyphi C RKS4594 strain. Mapping to the concatenated pair with *S.* Typhi CT18 yielded a number of regions, including a portion of the Salmonella Pathogenicity Island 7 (SPI-7) containing five genes, which are absent in the *S.* Paratyphi C RKS4594 reference (NC_012125.1). Notably, SPI-7 also contains the *viaB*-locus, an important *Salmonella* virulence factor[47]. An overview of the regions identified is shown in Supplementary Table 12. These extra regions were captured due to the *S. enterica* genome diversity included in the probe design.

The five additional SPI-7 genes (*pilS*, *pilT*, *pilU*, *pilV*, *rci*) present in our ancient genomes, which are absent in the Paratyphi C RKS4594 strain, form part of the *pil* operon that encodes type IVB pili[47] (Supplementary Table 12). The *pil* operon is concluded with a shufflon that codes for Rci recombinase. Rci recombinase acts upon the two 19bp C termini of the PilV protein and shuffles rapidly between its two protein states (PilV1 and PilV2), thereby preventing PilV production[48-50]. PilV is produced in strains where the *rci* gene is locked by mutation due to the insertion of an extra adenine base in each of the, now 20bp, C termini, thus disabling its shuffling function[49]. The production of PilV detaches type IVB pili located on the outer membrane[50].

In *S.* Typhi and some strains of *S.* Dublin the 19bp C termini in PilV remain intact, thus maintaining the inversion activity of Rci, whereby the pili also remain intact to facilitate bacterial self-aggregation[50]. *S.* Typhi bacteria have been found to vary the rate at which the Rci acts upon two 19bp repeats in the *pilV* gene depending on the growth environment. In low oxygen environments, such as the gut, the inversion activity of the *rci* shufflon increases[48]. Modern *S.* Paratyphi C strains have been found to lack these five genes, such as in Paratyphi C RKS4594, or to carry an intact version where *rci* is locked by mutation[49]. The *pilV* gene in the two ancient genomes carries the 19bp inverted repeats present in the *S.* Typhi CT18 genome and all other *S.* Typhi strains included in our modern dataset. When the ancient genomes are mapped to the intact *pil* operon sequence of *S.* Paratyphi C SGSC 2712 (AY249242.1) we observe a SNP deletion in both the 20bp inverted repeats of the *pilV* gene. This indicates that the *rci* in our ancient strains were 19bp and may have had full functionality. It is thought that the ability for the bacteria to self-aggregate is an early step in *S.* Typhi pathogenesis and that it aids in the invasion of epithelial cells in the human gut via the cystic fibrosis transmembrane conductance regulator (CFTR)[47,50]. The absence of, or inactive versions of, the *pilV* and *rci* genes carried by modern *S.* Paratyphi C strains has been suggested to account for their inability to cause epidemic scale outbreaks[47,49,50]. Finding the active versions of these genes in our ancient genomes may indicate an increased capacity for these strains to cause an epidemic scale outbreak, potentially supported by the fact that they were isolated from the Grand Plaza epidemic cemetery.

13. Presence/absence analysis of virulence factors

A set of 43 effector genes identified within the *Salmonella enterica* subsp. *enterica* serovars presented in a recent paper by Connor *et al.*[51] were used to make a concatenated reference file (Supplementary Table 13). All genomes (including plasmids) in our dataset were mapped to this reference using BWA with the following parameters –l 32, -n 0.1 and –q 0. When the mapping quality is reduced to zero, reads that map equally well to two or more genes will be kept and randomly mapped to one of the positions. The BEDTools bioinformatics suite[52] was used to generate the percentage of each gene covered at least 1-fold in each genome in the dataset. This information was plotted using the ggplot2 package[53] in R[54] and is shown in Supplementary Figure 9. Notably, in comparison to the modern *S.* Paratyphi C RKS4594 reference, the *pipB2* gene has a higher percentage of the gene covered in the two ancient genomes with 94% coverage in both Tepos_14 and Tepos_35 and only 77% covered in the *S.* Paratyphi C RKS4594 genome. *pipB2* is an effector protein secreted via the type III secretion system and it is involved in regulating the recruitment of kinesin-1[55].

14. Plasmid analysis

S. Paratyphi C strains harbour a virulence plasmid called pSPCV, which was included in the design of our capture probes. To investigate the presence of this plasmid in relation to our captured ancient strains we mapped our UDG treated data with the corresponding parameters to the pSPCV reference sequence (NC_012124.1). pSPCV is present in the five assembled genomes between average coverages of 10- to 224-fold (Supplementary Table 14). The pSPCV plasmid is present at 62-fold average coverage in the Tepos_10 genome, indicating that this sample is indeed positive for *S*. Paratyphi C (Supplementary Table 14), despite the chromosome containing too many heterozygous positions to allow SNP analysis. This plasmid is estimated to be present in 1-2 copy numbers per bacterial cell[56], which may account for the near double coverage of the pSPCV in comparison to the rest of the genome for all samples. The plasmid is present at between 0.35- to 3.6-fold in the remaining five weak positive samples (Supplementary Table 14).

SNP analysis of the pSPCV plasmids present in our two ancient strains (Tepos_14 and Tepos_35) was carried out in comparison to three virulence plasmids present in other *S. enterica* subsp. *enterica* strains that share a high degree of sequence similarity with pSPCV[57,58]. Artificial sequencing data (100bp reads with 1bp tiling density) was generated for these three plasmids and mapped to the pSPCV reference. The additional plasmids comprise pSCV50 (NC_006855.1) present in *S.* Choleraesuis SC-B67, pKDSC50 (NC_002638.1) present in *S.* Choleraesuis RF-1 and pSLT (NC_003277.1) present in *S.* Typhimurium LT2.

411 homozygous SNP positions were called from our dataset using the same parameters as outlined previously in Supplementary methods 9. Ten SNPs were shared by or unique to the ancient pSPCV plasmids (Supplementary Table 15). Two non-synonymous SNPs were identified to be specific to one or both of the ancient pSPCV plasmids. One nsSNP specific to Tepos_14 occurs in the *pefD* gene that is part of the fimbrial *pef* operon involved in bacterial adherence to the intestinal epithelium, where *pefD* specifically encodes for the periplasmic chaperone[58]. The second nsSNP occurs in the replication related *parA* gene carried by both Tepos_14 and Tepos_35[56].

Supplementary discussion

1. Summary of 16th century accounts of the 1545 epidemic and medical historical analyses

Descriptions and depictions of the 1545 epidemic appear in more than a dozen early colonial sources and include both Spanish and indigenous accounts. For example, the Codex en Cruz records an indigenous depiction of the disease in which a man covered in a full body rash writhes on the floor while liquid flows from his face – either blood or vomit (Codex en Cruz[59]; Supplementary Figure 10a). Other indigenous depictions include clear-skinned victims bleeding from the face (Codex Mexicanus[60]; Supplementary Figure 10b; Tira de Tepechpan[61]; Supplementary Figure 10c), corpses (Codex Aubin[62]; Supplementary Figure 10d), and stacks of shrouded corpses (Codex Telleriano-Remensis[63]; Supplementary Figure 10e).

The Anales de San Gregorio Acapulco[64] report that in 1545 there was much death resulting from a "pestilence in which blood poured from the nose"[i], and the Tira de Tepechpan[61] glosses a picture of a victim dying in the year 1 House (1545) with the Nahuatl words for "sickness" and "bleeding"[ii] .The Anales de Tecamachalco[65], written in the 1590s, states that the victims bled "from their mouth, from their nose, from their teeth" and that "ten, fifteen, thirty, forty people were buried on one day" during the *huey cocoliztli* epidemic of 1545[iii], and that among those who died were "very many children" including the children of great lords, and "more and more children"[iv]. Later, the same account reports that the victims of the 1576 epidemic died after haemorrhaging blood "from the nose, from the ears, from the eyes, and from the anus," and women were said to have bled from their vaginas and men from their penises[v].

---

[i] Spanish: "En este mismo año ocurrió la gran epidemia, "huey cocoliztli". Salió sangre por la boca, por la nariz y por los dientes de la gente. Aquí vino a propagarse en el tiempo de la siembra, en el mes de mayo. Fue espantosa la mortandad; al comienzo de la epidemia en un día se enterraban diez, quince, veinte, treinta, cuarenta, en un día..."

[ii] Nahuatl: "1 calli Y coco [illegible] tztica on [illegible]."

[iii] Spanish: "En este mismo año ocurrió la gran epidemia, "huey cocoliztli". Salió sangre por la boca, por la nariz y por los dientes de la gente. Aquí vino a propagarse en el tiempo de la siembra, en el mes de mayo. Fue espantosa la mortandad; al comienzo de la epidemia en un día se enterraban diez, quince, veinte, treinta, cuarenta, en un día..." Nahuatl: "1 calli: ynipan xihuitl mill e qujs. Y quarenta y cinco…Zan no ypan xihuitl yn huey cocoliztli mo chiuh yn eztli tecamac paquiz teyaca paquiz tetlam paquiz y nican callaquico ypan metztli mayo tequizpan temamauhti ynic micotlac yn cem ilhuitl mo tocaya y zan oquic peuhqui X, XV, XX, XXX, XL ynin cem ilhuitl."

[iv] Spanish: "Y los niños muchísimos murieron en vino á desparecer la epidemia allá murieron niños gran señor era y más y más uiños etc." Nahuatl: "Auh yn pipiltzitzintin cenca miyequintin yn micque xiuhtica yn poliuco in cocoliztli y noncan micqz. Pipiltin huey teuctli catca auh occequintin pipiltin ets."

[v] Spanish: "Cuando llegamos al mes de octubre se habían enterrado treinta personas. De tres o en dos días morían por hemorragia; les salía sangre por la nariz, por las orejas, por los ojos, por el ano. A las mujeres les salía sangre por sus entrepiernas. A nosotros los varones nos salía sangre por el miembro." Nahuatl: "yn

The Franciscan friar and eyewitness Gerónimo de Mendieta (1525-1604)[66] called it *pujamiento de sangre* and reported that it was characterized by *calenturas de sangre* ("blood fevers") and bleeding from the nostrils[vi]. He later went on to refer to the second *pujamiento de sangre* epidemic in 1576 as *tabardillo*, a Spanish term for typhus[vii].

The Franciscan scholar Fray Bernardino de Sahagún (1499-1590)[67] likewise witnessed the *pestilencia grandísima* of 1545 and in Tlateloco claimed to have personally supervised the burial of 10,000 people during the epidemic, when he himself came down with the illness, and nearly died[viii]. When disease broke out again thirty years later in 1576, he likened it to the earlier 1545 epidemic. Sahagun later died in 1590 during a third *cocoliztli* epidemic[68], which is believed to have originated in the Mixteca Alta[69].

In the Historia Tlaxcala (composed ca. 1580-1595)[70], the mestizo historian Diego Muñoz Camargo briefly reported on the 1545 epidemic, noting that it lasted more than six months and "ruined and depopulated" most of the country[ix]. A decade later, writing in the early 17th century[71], the indigenous chronicler Chimalpahin Quauhtlehuanitzin reported that during the 1545 *cocoliztli* "blood flowed from the mouth, eyes, nose, and anus" of both nobles and commoners, and that after they had perished "the dogs and coyotes ate the people in Chalco"[x].

---

yhquac ypan meztli otobre tahçico yn ye mo toca centecpantli omahtlactli zan ei'lhuitl zan omilhuitl yn micoua. Eztli te yacacpa te nacazco te yxco te zinco qui caya. Auh yn çiuah yn maxac quiztia yn eztli tequichtin yn techquizaya yn eztli yn totonh."

[vi] Spanish: "La tercera pestilencia grande y general vino en el año de cuarenta y cinco, que de rliquia de las pasadas debió de retoñecer. Este fué de pujamiento de sangre, y juntamente calenturas, y era tanta la sangre, que les reventaba por las narices. De esta pestilencia muriero en ciento y cincuenta mil indios, y en Cholula cien mil, y conforme á esto en los demas pueblos, segun la poblacion de cada *uno."*

[vii] Spanish: "El año de setenta y seis vino otra general pestilencia, de que murió grandísima suma de gente por todas partes, y fué de pujamiento de sangre, como las demas, y daba en tabardillo."

[viii] Spanish: "…el año de 1545, hubo una pestilencia grandísima y universal, donde en toda esta Nueva España, murió la mayor parte de la gente que en ella había. Yo me hallé en el tiempo de esta pestilencia en esta ciudad de México, en la parte de Tlatiluco, y enterré más de diez mil cuerpos, y al cabo de la pestilencia dióme a mí la enfermedad y estuve muy al cabo."

[ix] Spanish: "Obo en su tiempo una muy gran pestilencia y mortandad en los naturales desta Nueva España el año de 1545, que duró mas de seis meses. Arruinó y despobló la mayor parte de todo lo poblado de la tiera."

[x] French: "Année 1 maison, 1545. Alors il 1545 y eut mortalité; le sang coulait par la bouche, parles yeux, le nez et le fondement ; il périt extrê mement de nobles, hommes et femmes, ainsi que des gens du peuple. Alors les chiens et les chacals mangèrent des personnes à Chalco." Nahuatl: "I calli xihuitl, 1545 anos.Ypan in yhcuac yc micohuac; yn eztli cocoliztli tocamac, tixco1, toyacacpa, totzinco quiz ; cenca yquixpoliuhque' yn pipiltin, yn cihua- pipiltin, yhuan yn macehualtin. Yhcuac tecuaque chichime yhuan cocoyo yn Chalco."

Numerous references to the 1545 epidemic appear in the Relaciones Geográficas, a compilation of responses to a questionnaire circulated by King Philip II of Spain between 1578 and 1586[15]. The purpose of the questionnaire was to collect basic information about the Spanish held territories in the Americas. One hundred ninety-one communities responded to the questionnaire between 1578 and 1586, and the present location of 167 responses is known. Question 17 relates to the health of the towns and their history of illness, and the answers given by the communities provides detailed information about the extent and regional severity of the 16th century epidemics, as well as details on local climate and water quality[15]. In the Mixteca region of Oaxaca, Mexico, for example, the town of Tilantongo[15], located just "seven leagues" to the west of Teposcolula, reported that there is now "much disease and pestilence, which did not exist in the past"[xi]. Tamazola, another local town, reported three previous epidemics in this otherwise "healthy" land, in which the first two were "in the manner of smallpox" and the third *pujamiento de sangre*[xii]. Interestingly here the second epidemic is associated with a rash rather than bleeding[15]. A third local town, Mitlantongo[15], likewise complained of three previous *cocolistles o pestilencias* in which the first two were "in the manner of smallpox" and the third was *pujamiento de sangre* [xiii]. Another town, Justlahuaca[15], reported having endured three *pestilencias* that occurred after the arrival of the Spaniards and noted that they continue to suffer from diseases of fevers, *cámeras de sangre* ("bloody diarrhoea"), and pain in the body and head[xiv].

[xi] Spanish: "…En el pueblo de Tilantongo, en cinco días del mes de noviembre de mil y quinientos y setenta y nueve años…Del pueblo de Teposcolula, a éste de Tilantongo, hay siete leguas de mal camino y áspero, por ser toda esta provincia serranías…Y que antiguamente vivían más que ahora y vivían más sanos, porque, ahora, dicen que les han sucedido muchas enfermedades y pestilencias, y que, antiguamente no las tenían.

[xii] Spanish: "Es tierra sana por ser tierra fría, más que cuantas en esta provincia hay, y dicen que nunca han tenido género [alguno] de enfermedad, si no han sido tres pestes generales que [ha] habido desde que la tierra se ganó. Y [dicen] que las dos pestes primeras fue a manera de viruelas y, esotra, pujamiento de sangre, y que no han tenido otra enfermedad; y que no entienden ni saben curarse y, a esta causa, cuando las pestilencias dio, murió mucha cantidad de gente."

[xiii] Spanish: "Es tierra sana y lo ha sido, que si no han sido tres cocolistles o pestilencias que [ha] habido en esta tierra generales, que el primero habrá cincuenta años, recién ganada la tierra, y otra enfermedad que habrá treinta años, que fue a manera de viruelas, y otra de pujamiento de sangre, que habrá tres años que dio esta enfermedad generalmente en todas las Indias, no ha [ha]bido otro género de enfermedad; de las cuales enfermedades murió, de cuatro partes, las tres de la gente que solía haber antiguamente. Y, ansí, es tierra muy sana, y [dicen] que la gente que falleció [fue], toda la más, por no tener remedios con que atajar las enfermedades."

[xiv] Spanish: "El puesto destos dos pu[ebl]os parece sano, porque está asentado en buena parte y no es húmedo; y es tierra de buen temple, más frío que caliente, y corre todos los días desta vida, desde las dos de la tarde hasta las seis de la noche, una marea muy recia, que es tenida por sana. Las enfermedades que los indios destos d[ich]os dos pu[ebl]os padecen son calenturas y cámaras de sangre, y dolores de cuerpo y de cabeza..."

Stomach pain and diarrhoea are frequently mentioned in the Relaciones Geográficas, including in relation to *cocoliztli*. For example, a town in the province of Ocopetlayuca[72] noted with respect to the 1576 epidemic that "…cocoliztle, from which many die…the nature of this illness that it causes great pain at the 'mouth of the stomach' and is accompanied by a high fever in all parts of the body. Death sets in after six or seven days…the sick who survive this time become healthy. At the same time, there are cases of relapse with deadly consequences"[xv]. Additional important observations are made in the Relación de La Villa de Tepoztlán[74], which records that the most ordinary illness before the Spanish arrived was fever, which was called *tlacacocoliste*, but that now they are afflicted with "a thousand kinds of illnesses" such as *matlaltotonque*, which is translated as *tabardete*, "after the rash that covers the body like tabard cloak", and also *matlalcagua*, which is said to "resemble measles and bloody diarrhea and bleeding from the nose"[xvi]. Likewise, the town of Tepeapulco[74] notes that "the diseases from which people die" since the arrival of the Spanish are *tavardete* and stomach pain and "for this reason they lose much blood from their face and nostrils" [xvii].

Overall, *cámeras de sangre* ("bloody diarrhoea") is the most prevalent illness reported in the Relaciones Geográficas[15,73], including in both tropical lowland and cold highland contexts. In the Mixteca Baja region of Oaxaca, Mexico, for example, the town of Acatlán reports that there are "many diseases of chancres that no one can cure, and they suffer many other diseases of fever and bloody diarrhea"[xviii]. The people of Chila are said to have "various diseases" of which "bloody diarrhea is the most common"[xix]. In Petlaltzingo, people are said to have "chancres and bloody diarrhea" because of "bad water"[xx]. The town

[xv] Spanish: "…y que agora de presente, y desde que entro el Marques en la tierra, biben con munchas enfermedades agudas de cocoliste, en que mueren mucha cantidad dellos, como de presente se bee, y acabar toda la gente deste dicho pueblo y su juridicion; y la calidad de la dicha enfermedad es que da gran dolor en la boca del estomago con grandisimo acidente de calor en todo el cuerpo, y cor responde en la cabeca, y el que muere es en seis, siete dias, y de los que de aqui pasan escapan casi todos avnque les suele voluer y morir…". English translation by Fields73 Fields, S. L. *Pestilence and Headcolds: Encountering Illness in Colonial Mexico*. (Columbia University Press, 2008)..

[xvi] Spanish: "…y que antiguamente la mas hordinaria enfermedad que les perseguia hera vna que llamavan tlacacocoliste que es como dezir de calenturas y que les durava mucho y se secavan hasta que se morian, y que no sabian que cosa era sangrias mas de tomar gumo de yerbas que conoscian, e que a algunos les hera provechoso pero otros no sanaban, y que agora en estos tienpos les persiguen myll generos de enferme dades como son: matlaltotonque ques lo que dezimos tabardete, llamanle ansi por las manchas que descubren en el cuerpo; y otra que se dize en la lengua matlalcagua ques lo mesmo que «sarampion y camaras de sangre y fluxo de sangre por las narizes."

[xvii] Spanish: "…y las enfermedades de que mueren, despues quel Marques vino a esta tierra, es tavardete y dolor destomago, por donde procede echar mucha sangre por la boca y narizes"

[xviii] Spanish: "Hay muchos enfermos de bubas, de las cuales ellos no se saben curar, y padecen otras muchas enfermedades de calenturas y cámaras de sangre."

[xix] Spanish: "Las enfermedades son varias, y lo más común es cámaras de sangre."

[xx] Spanish: "Es de malas aguas, y los naturales padecen de bubas y de cámaras de sangre."

of Piaztla is also said to "not have any good water" and the townspeople consequently suffer from diseases of chancres, mange, measles, and bloody diarrhea," and in Icxitlan, where the water quality is not specified, the townspeople suffer from "chancres, mange, measles, and bloody diarrhea"[xxi].

The etiological cause of the 1545-1548 *cocoliztli* epidemic is difficult to determine on the basis of historical records alone, but several possible bacterial and viral diseases have been proposed, including: epidemic typhus (*Rickettsia prowazekii*)[73,75-80]; enteric/typhoid fever (*Salmonella enterica*)[76]; pneumonic plague (*Yersinia pestis*)[75,81-83]; bartonelosis (*Bartonella bacilliformis*)[84,85]; leptospirosis (*Leptospira* spp.)[84,85]; smallpox (variola virus)[76]; measles[76]; a variety of tropical viruses such as dengue fever virus, arenavirus, hantavirus, and yellow fever virus[84,85], as well as an unspecified viral hemorrhagic fever[84,86,87]; and even malaria (*Plasmodium vivax*)[84,85].

Until the early 2000s, there was a general consensus among medical historians that *cocoliztli* was a form of epidemic typhus, a disease with similar symptoms to enteric fever, leading one historian to state, "It is not possible to say definitely that the *cocoliztli* of Sahagun and of Cárdenas was typhus, but it is almost a certainty"[88]. This assessment was made on the basis of a wide range of evidence, including the fact that the second *cocoliztli* epidemic in in 1576 was described by some eyewitnesses as *tabardillo* ("tabard cloak-like rash")[66,73] and *matlazáhuatl* ("net-like rash"), Spanish and Nahuatl words that increasingly came to refer to the rash characteristic of epidemic typhus, but which also resembles that of enteric fever[75,89].

More recently, it was hypothesized that *cocoliztli* may have been a viral hemorrhagic fever[86], an idea that was later expanded in the work of microbiologist Rodolfo Acuna-Soto, primarily on the basis that the disease broke out in late summer during a brief wet period following a long drought, and the fact that "[t]hese symptoms are not consistent with known European or African diseases present in Mexico during the 16th century"[87]. To be fair, however, the historical symptoms of *cocoliztli* are also not consistent with any known viral disease, a point also conceded by Acuna-Soto that "[d]espite some similarities, there is not a perfect match between *cocoliztli* and any other specific form of hemorrhagic fever"[84]. Additionally, the pattern of disease outbreak following sudden heavy rainfall has been not only documented for rodent-borne viruses, such as hantavirus, as proposed by Acuna-Soto *et al*.[84,87], but also for waterborne bacterial diseases, such as enteric fever and cholera[90].

[xxi] Spanish: "El dicho pueblo de Piaztla, como dicho es, de agradables aires, excepto que no hay agua buena, que es gorda y hace mal si no la quebrantan. Y los naturales padecen enfermedades de bubas, sarna, y sarampión y cámaras de sangre."

Early accounts of typhus/enteric fever, which was at times called *tabardillo* or *tabardete* after the "tabard cloak-like" rash pattern that it produced, can be traced in Spain to the late 14th century, when Juan de Avinon described *tabardillo* as being very prevalent in Seville in 1393, along with smallpox, measles, and other fevers[88,91]. The physician Fracastorius described the disease as producing a severe nosebleed: "I have often seen cases where three pounds of blood burst from the nostrils, yet the patients died soon after"[88,92]. In 1489, soldiers from Cyprus introduced an outbreak of what was almost certainly typhus or enteric fever during the Spanish siege of Granada, but interestingly it was described not with the common names *tabardillo* or *tabardete*, but rather as *calenture maligna puncticular*, a malignant fever with punctate eruption[88,93]. Thus, there was historical precedent for the Spanish to adopt new names for what was essentially an old disease.

In the context of the Americas, the source of the epidemics was likely the ships that regularly arrived at the ports along the Gulf Coast, on the order of hundreds per year. Hospitals were founded along the royal road connecting Veracruz to México City, and the 1580 Relación of Xalapa de Veracruz[74] describes a hospital in which "they treat the Spanish passengers who come from Spain in fleets of ships"[xxii].

A detailed account of Spanish ship traffic and the diseases that followed is provided in a personal diary recorded from 1544-1545 by Friar Tomás de la Torre[75,94]. Torre and his fellow travellers (comprising twenty-six ships) journeyed from Salamanca, Spain to the southern Mexican city of Ciudad Real de Chiapa with brief stops in the Caribbean and at the port of Campeche. Disease and death are frequently mentioned during the journey. For example, during a layover in Puerto Rico, Torre noted "many people had already died." After recuperating for several months in Santo Domingo on the island of Hispañola, the group reached Campeche in December of 1544 and immediately fell ill. Torre complained bitterly of the mosquitos, which left one traveller incapacitated "from pains in the bowels that doubled him up and tied him into knots, from which we thought he could never be cured…" They were then plagued by a series of fevers as they travelled inland, and at one point several of the party were reduced to crawling because they could no longer walk. A sick servant was said to issue an "infinite number of worms" from his nose, and during one particularly severe bout of fever, Torre recalled that "I do not know how many there were of us, except that there was no one left to serve anyone. I think that there were thirteen of us strewn about the hut; three on the mattress of the cleric, and the rest on the ground; all

[xxii] Spanish: "En este pueblo ay vn hospital…en el se curan los pasajeros spañoles que vienen de Spaña en las flotas de los nautios, por que este pueblo esta media legua del camyno Real que viene de la çiudad de la Veracruz a Mexico, y los que pasan enfermos se curan en este hospital: tanbien se curan en el los indios que enferman en el seruiçio de la carreterio y harrias…"

with fever in a forsaken place." A fellow traveller, Pedro Calvo, suffered from severe gastrointestinal pain and could not raise his head, and one priest had to be carried on a litter. They finally arrived in Ciudad Real de Chiapa in March 1545, coinciding with the outbreak of the *pujamiento de sangre* epidemic across New Spain. It is unknown if any of the many fevers endured by Torres and his party were related to *pujamiento de sangre*, but the account nevertheless paints a picture of the overall disease burden of the land and the likelihood of mixed infections and compounded symptoms once the epidemic emerged.

A large part of the controversy surrounding the diagnosis of *cocoliztli* revolves around the descriptions of nosebleeds, in combination with fever and body rash. For some, this has been taken as evidence of a viral haemorrhagic fever; however, the symptoms also fall within the range of those described for epidemic typhus/enteric fever in Spain during the 14th and 15th centuries. Moreover, symptoms were likely exacerbated by a lack of palliative care. Sahagun[67] noted that "[m]any died of hunger and because no one was able to care for them; in many cases every member of a household fell ill, without a single person left to give even a cup of water to the sick"[xxiii].

The compounding effect of a lack of basic care on morbidity was likely related to the particular way in which household economy was organized in ancient Mesoamerica, an observation first made by the conquistador López de Gómara, who noted that during the first smallpox epidemic in 1520, "[f]amine came upon them because, having no mills, the women were unable to grind the grain between stones and many died of hunger"[88]. As such, disease manifestation during the 1545 epidemic was likely not limited to that of the infectious agent alone, but also to the additional physiological consequences induced by dehydration and a lack of food over the week-long course of the disease.

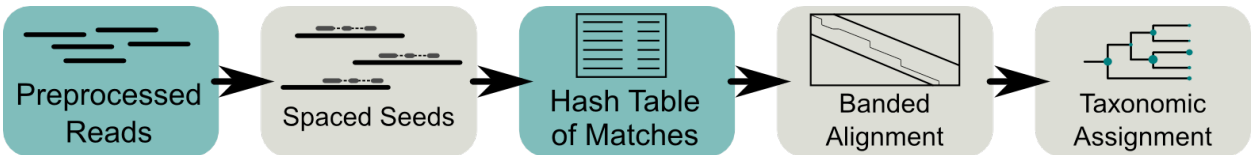
One problem with the accounts of the 1545 epidemic is that few were actually written at the time the epidemic occurred. The Codex Aubin[62], for example, records events occurring between 1520 and 1607, and is believed to have been started in 1576[95]. The Tira de Tepechpan, records the period of 1298-1596[61,95], and the Codex Mexicanus[60] records the period of 1168-1571 in one artist's hand, followed by additions until 1583 by two further artists[95]. Additionally, although some of the descriptions are known to have been provided by first-hand witnesses (e.g., Mendieta and Sahagun)[66,67], most accounts were recorded decades after the event, and many were written after the 1576 epidemic, which may have influenced memories of the 1545 epidemic. Two exceptions are the Codex Telleriano Remensis[63], which was likely written ca. 1553-1555, but perhaps as late as 1563[96], and the

[xxiii] Spanish: "Como también en la otra arriba dicha [1545 epidemic], muchos murieron de hambre, y de no tener quién los cuidase, ni los diese lo necesario; aconteció y acontece en muchas casas caer todos los de la casa enfermos, sin haber quién los pudiese dar un jarro de agua."

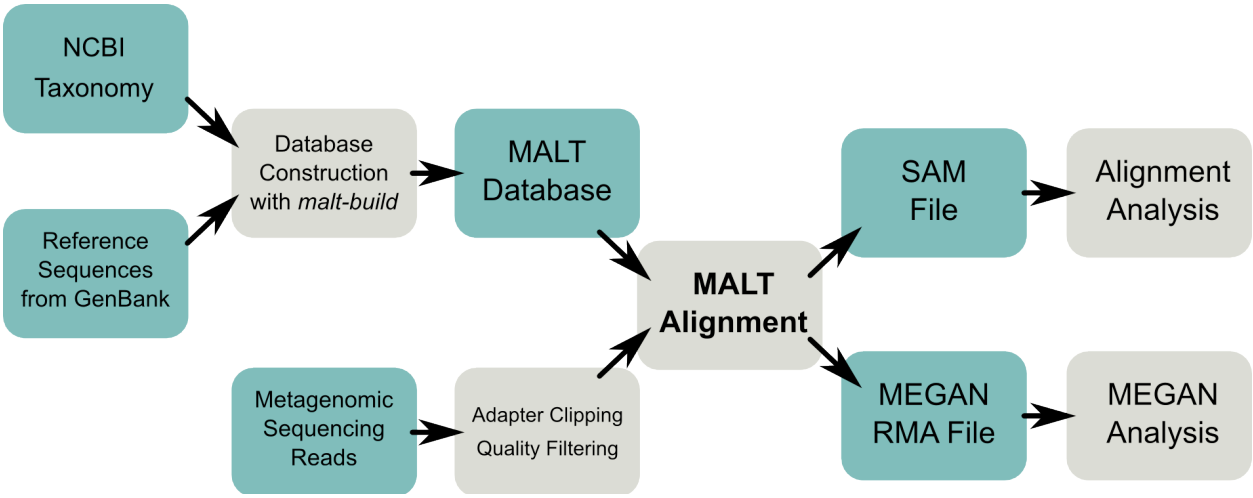
Codex en Cruz[59], which records events from 1402-1553 and which was likely completed in the mid 1550s, although subsequent additions were made through 1603[95].

Taken as a whole, medical historians have greatly advanced our understanding of the 1545 epidemic, but the ultimate cause of the disease is still far from clear. Although some physicians have been persuaded to accept the endogenous haemorrhagic fever hypothesis with as much certainty (e.g.,[97]) as medical historians once had for epidemic typhus (e.g., [75,88,98]), it is important to note that making confident diagnoses on the basis of such limited information is fraught with difficulty. As Acuna-Soto himself has noted "[t]oday, making a precise diagnosis based exclusively on clinical manifestations without geographical or laboratory data in a patient with severe hemorrhagic fever carries a high probability of error"[84].

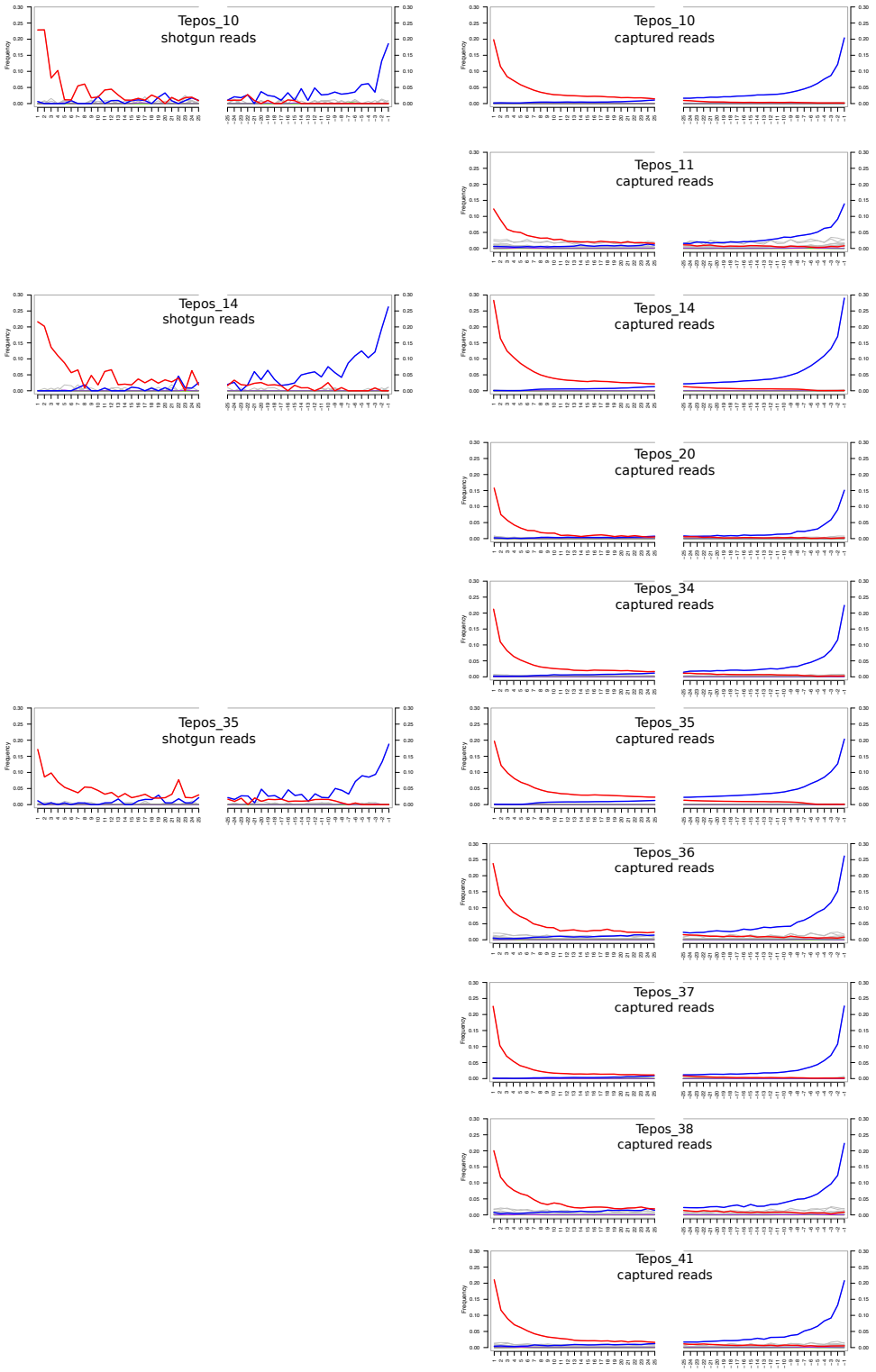
Rather than a straightforward path from symptom description to diagnosis, the historical record attesting to the 1545 epidemic is fraught with variability, contradiction, incomplete information, and perhaps even compromised memories. At its core, we can assert that the disease was a systemic blood-borne fever that likely produced a rash and frequently resulted in bleeding from the facial orifices. It likely had many other symptoms that were not recorded, either because they were not recognized or because they were thought not worthy of mention because they were so common, such as bloody diarrhoea. What is needed is independent biological evidence of the disease, such as that provided by microbial paleogenomics.



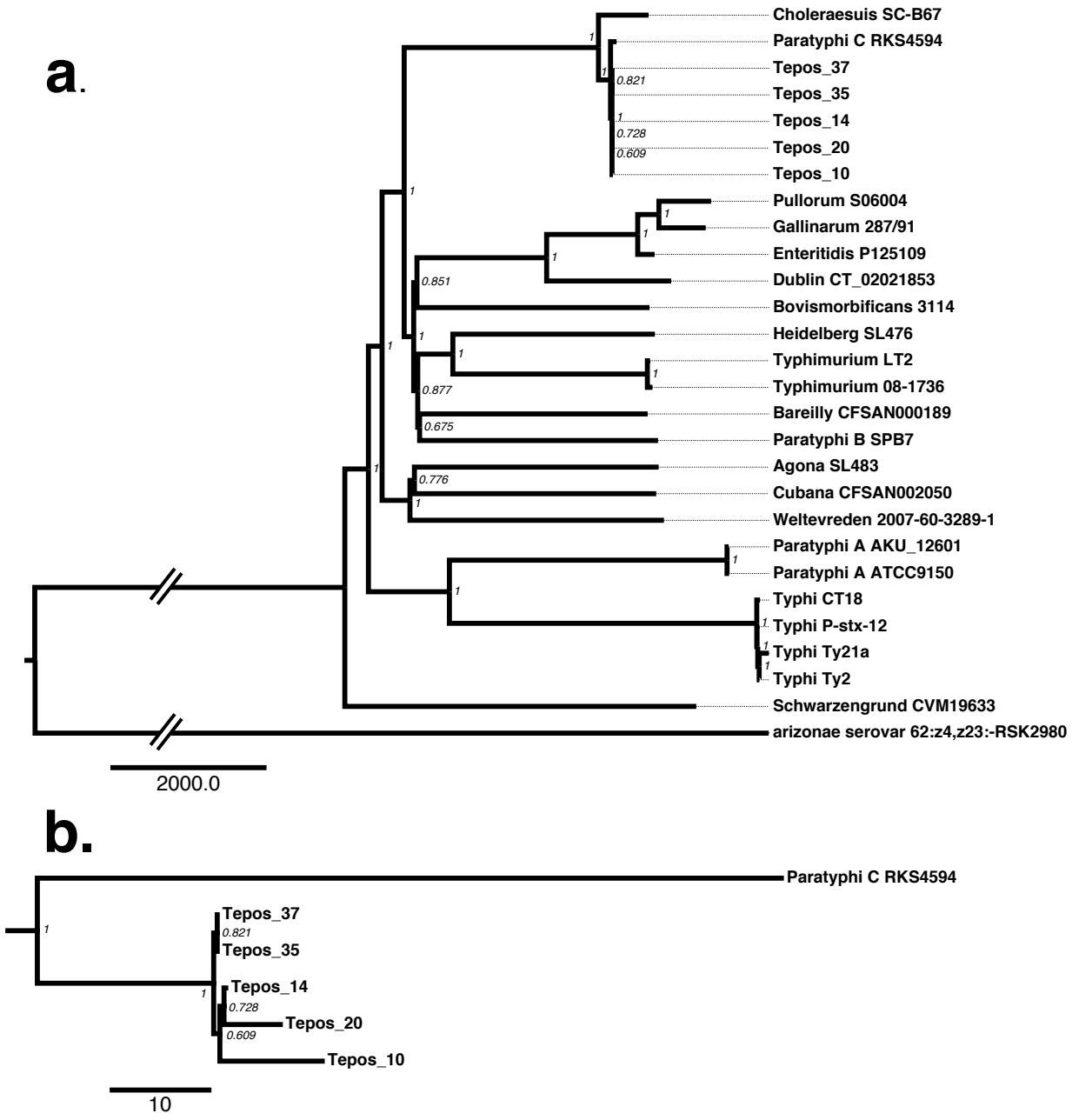
Supplementary Figure 1 | Schematic overview of MALT. For each pre-processed metagenomic sequencing read, the algorithm generates all contained spaced seeds and looks them up in a hash table of spaced seeds representing the reference database. A banded alignment is calculated for each match of seeds. Once all alignments for a given read have been calculated taxonomic binning of the read is performed using the LCA algorithm



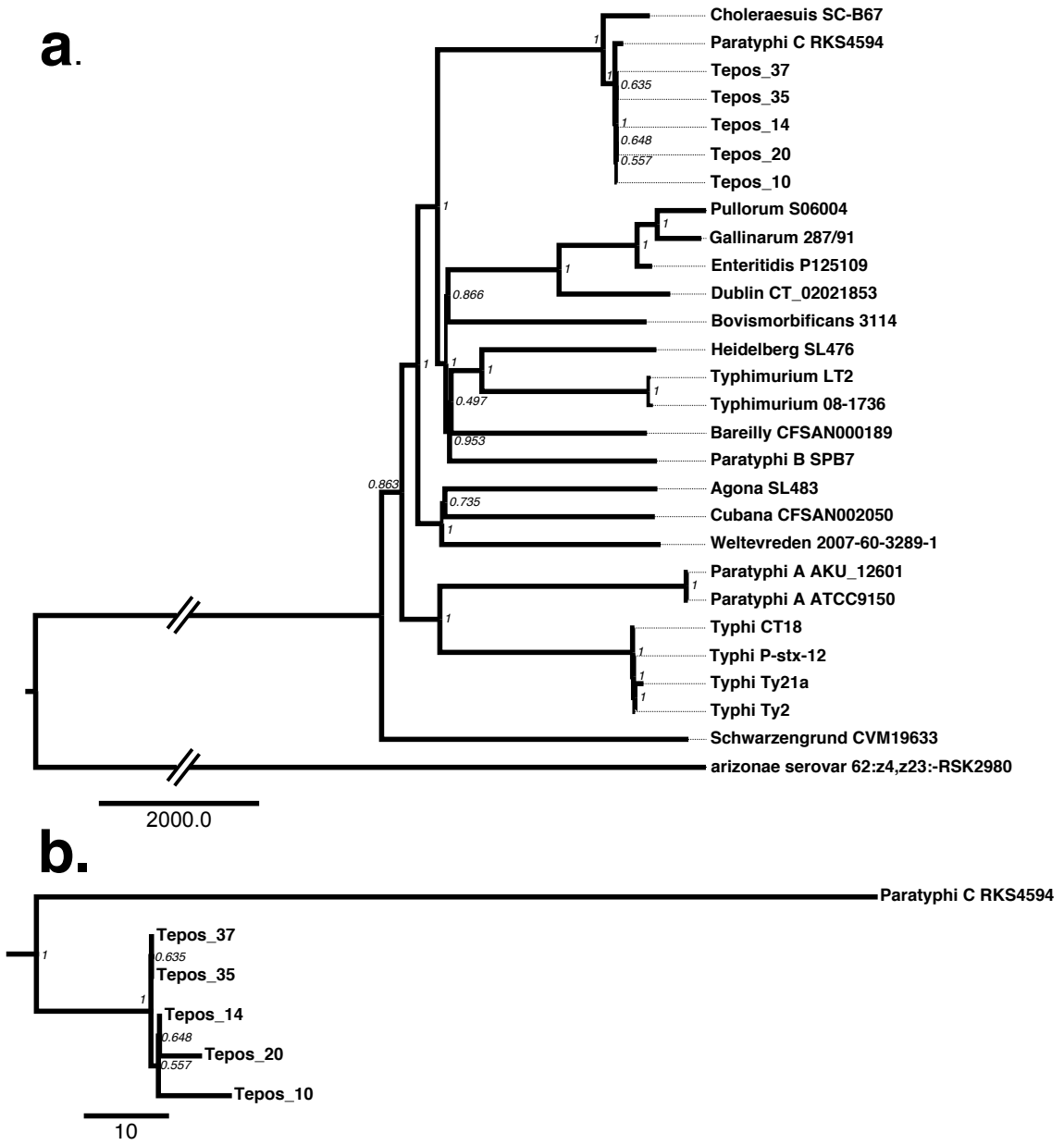
Supplementary Figure 2 | Schematic overview of our MALT-based analysis workflow. A reference index is generated for all bacterial genomes from GenBank. Pre-processed metagenomic sequencing reads are aligned against the reference sequences. An RMA result file is produced for further analysis in MEGAN. Alignments are also stored in SAM format.



Supplementary Figure 3 | Damage plots. Comparison of damage plots generated based on shotgun data versus capture data for the ten samples when mapped to *S.* Paratyphi C (NC_012125.1). The damage plots are more uneven for the shotgun data due to the low number mapping reads, compared to the smoother curves yielded by the capture data consisting of a much higher number of reads (see Supplementary Tables 4, 7).

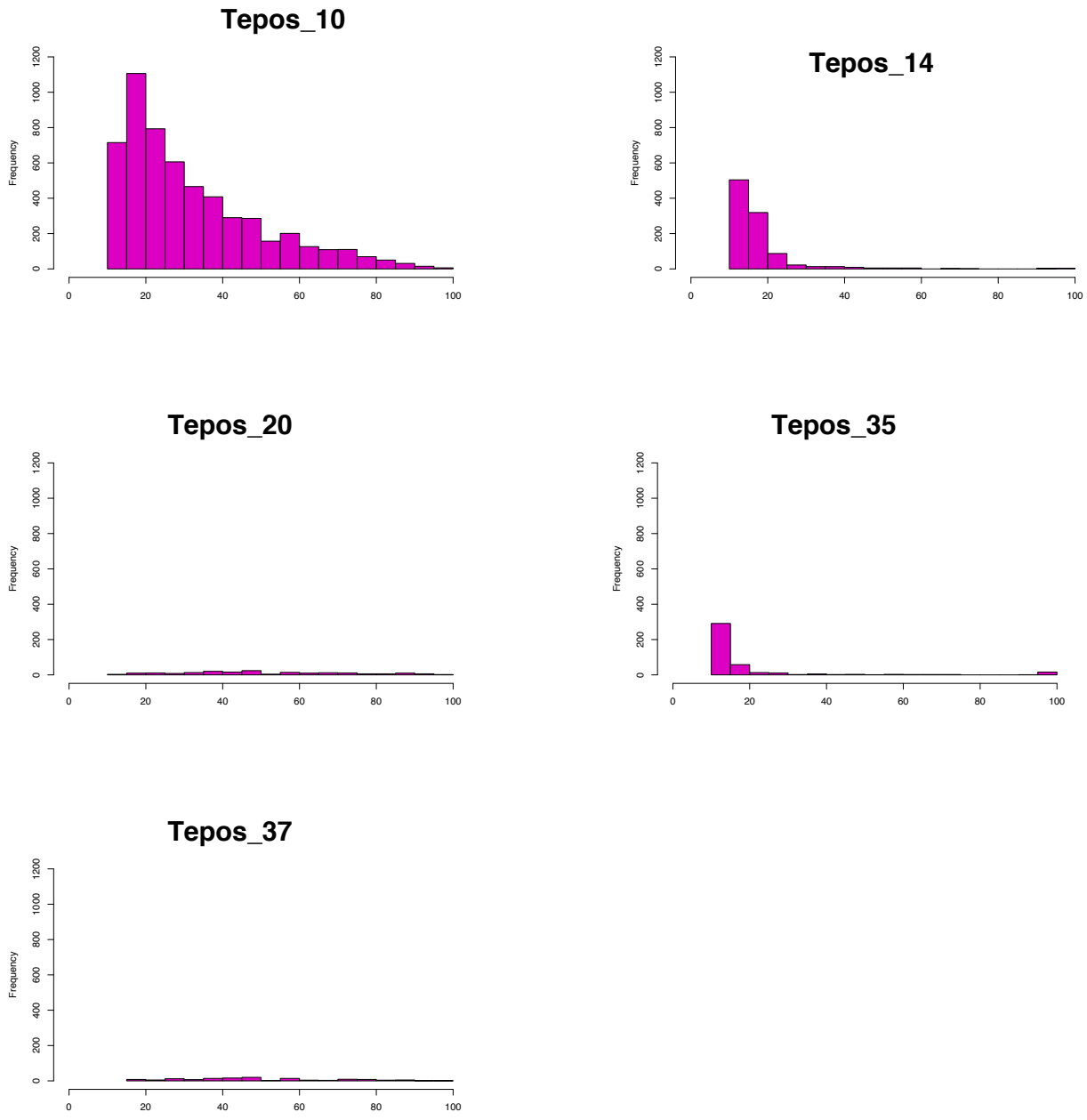


Supplementary Figure 4 | *S.* Paratyphi C reference based Neighbor-joining *S. enterica* phylogeny including the five ancient genomes. a, Neighbor-joining tree constructed using the dataset including the five ancient genomes, when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). No regions were excluded from SNP calling. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 51,602 positions. The five ancient genomes cluster with *S.* Paratyphi C, with high bootstrap support. The tree was constructed using MEGA6[35]. **b**, An enlarged view of the *S.* Paratyphi C clade.

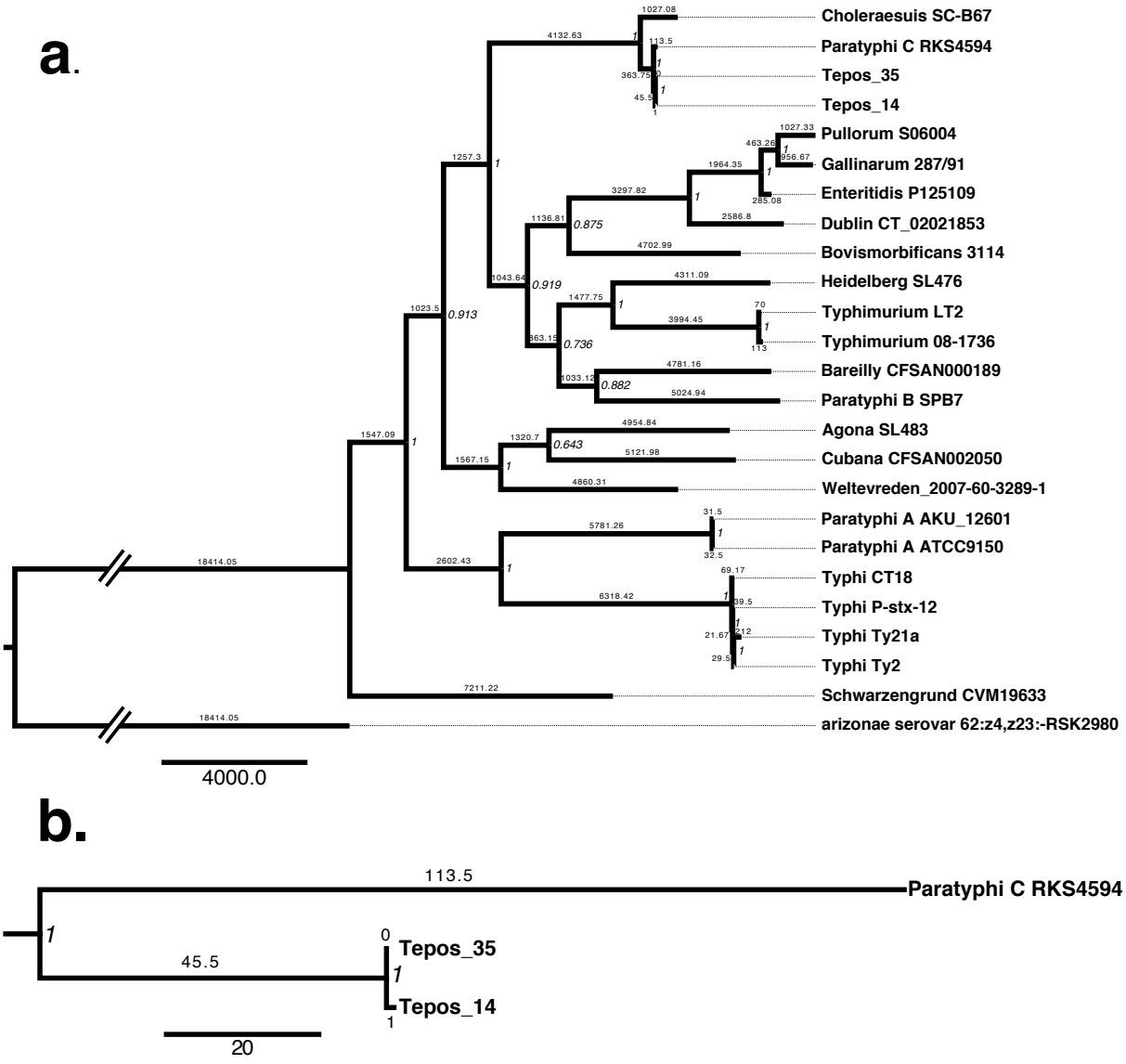


Supplementary Figure 5 | *S.* Typhi reference based Neighbor-joining *S.* enterica phylogeny including the five ancient genomes. a, Neighbor-joining tree constructed using the dataset including the five ancient genomes, when mapped to the *S.* Typhi CT18 genome (NC_003198.1). No regions were excluded from SNP calling. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 45,305 positions. The five ancient genomes cluster with *S.* Paratyphi C, with high bootstrap support. The tree was constructed using MEGA6[35]. **b**, An enlarged view of the *S.* Paratyphi C clade.

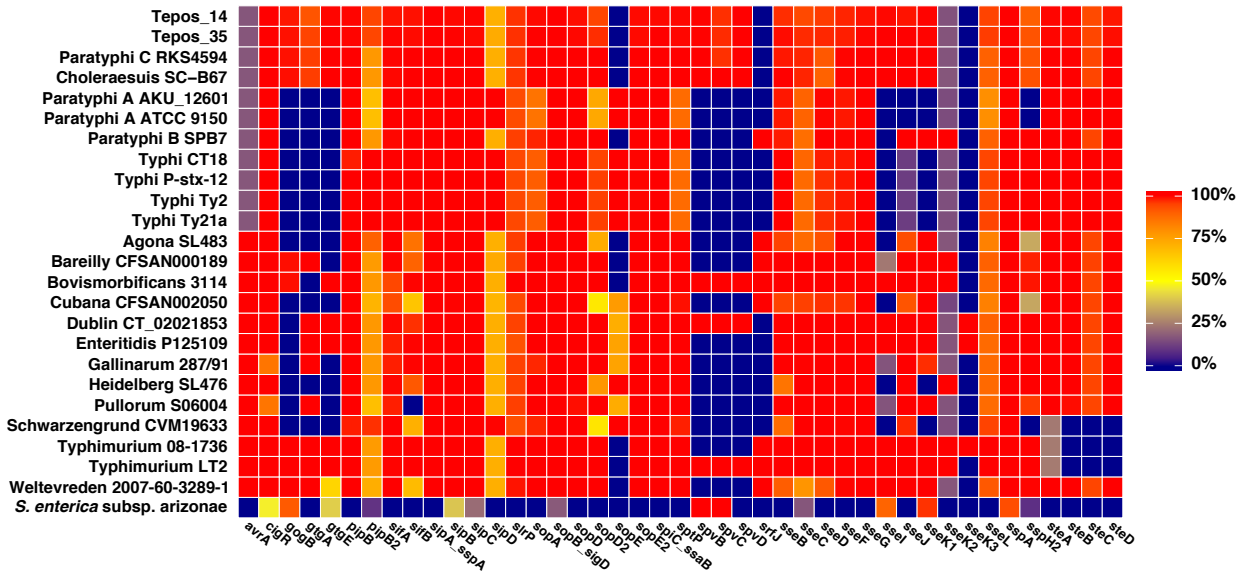
Supplementary Figure 6 | Maximum Parsimony *S. enterica* phylogeny. a, Maximum Parsimony tree constructed using the dataset, including the five ancient genomes, when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). Regions were excluded from SNP calling. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 51,456 positions. The five ancient genomes cluster with *S.* Paratyphi C, with a bootstrap support of 100%. The tree was constructed using MEGA6[35]. Branch lengths are displayed. **b**, An enlarged view of the *S.* Paratyphi C clade.



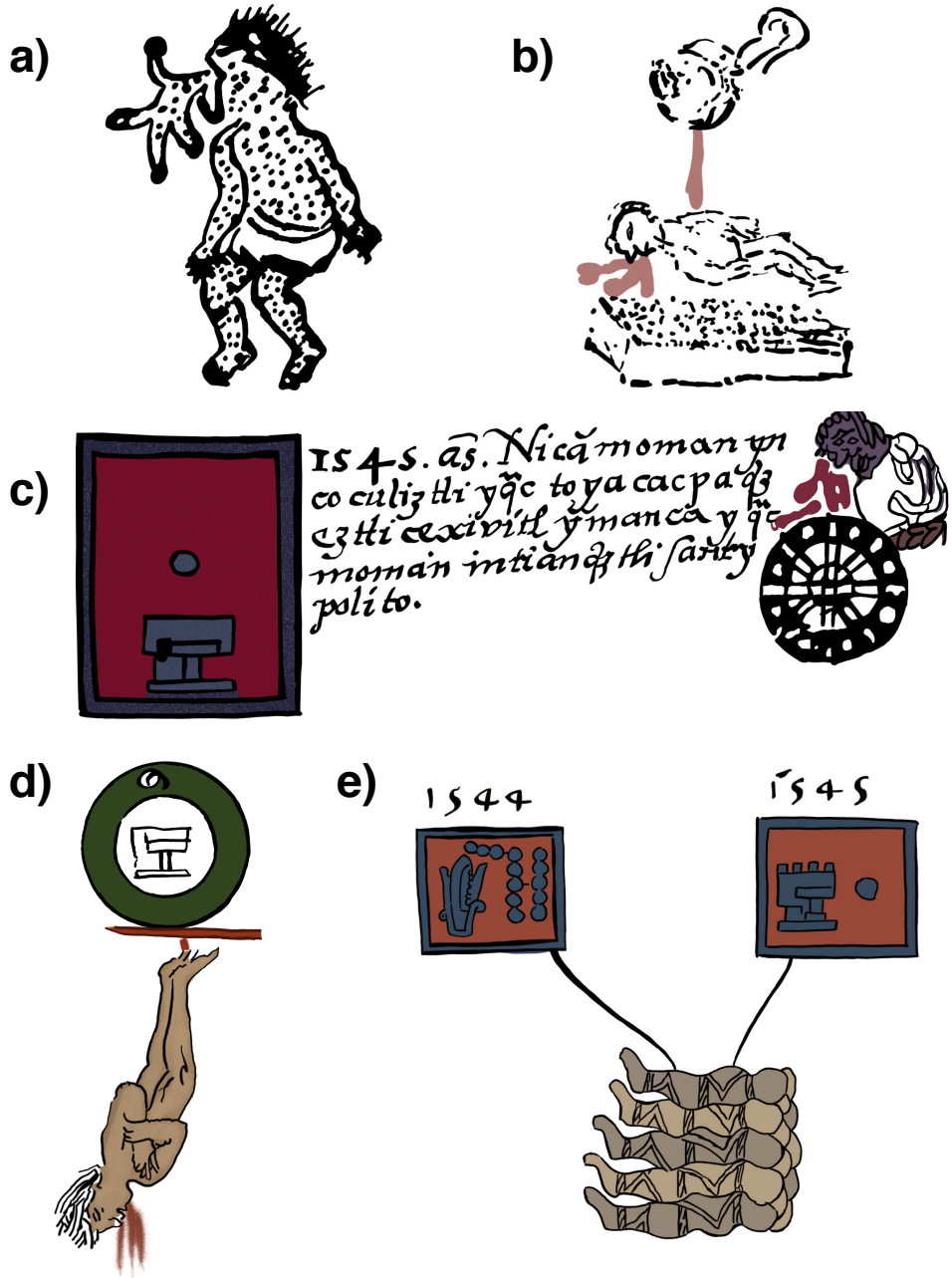
Supplementary Figure 7 | Histograms of SNP allele frequency distributions for the five ancient *S.* Paratyphi C genomes. The *x*-axis shows the SNP allele frequencies as a percentage. All variants where the SNP allele frequency is higher than 10% and lower than 100% are shown.



Supplementary Figure 8 | Maximum Parsimony *S. enterica* phylogeny. a, Maximum Parsimony tree constructed using the dataset, including Tepos_14 and Tepos_35, when mapped to the *S.* Paratyphi C RKS4594 reference (NC_012125.1). Regions were excluded from SNP calling. The tree was constructed by excluding all missing and ambiguous data, using 1000 bootstrap replicates and is based on 81,474 positions. The two ancient genomes cluster with *S.* Paratyphi C, with a bootstrap support of 100%. The tree was constructed using MEGA6[35]. Branch lengths are displayed. **b**, An enlarged view of the *S.* Paratyphi C clade.



Supplementary Figure 9 | Heatmap visualizing the presence/absence of effector protein coding genes. The colour scale signifies the percentage of each gene covered at least 1-fold.



Supplementary Figure 10 | Indigenous depictions of the 1545 epidemic in 16th century documents. (**a**) Codex en Cruz, (**b**) Codex Mexicanus, (**c**) Codex Aubin, (**d**) Tira de Tepechpan, (**e**) Codex Telleriano-Remensis. The Codex en Cruz[59] was written only a few years after the epidemic and is the only illustration to clearly depict a body rash. All of the images depict bleeding from the face, with the exception of the Codex Telleriano-Remensis[63] in which the corpses are shrouded. The Codex en Cruz[59] is painted in black ink only, but the Codex Mexicanus[60], Codex Aubin[62], and Tira de Tepechpan[61] are painted in colour and depict the fluid emerging from the face as red, the colour of blood. The Nahuatl gloss beneath the Codex Aubin states: *1 Calli 1545 Años. Nicān moman incocoliztli ìcuāc toyacacpa quīz eztli cē xihuitl in manca ìcuāc moman in tiyanquiztli San Hipolito*. ("The year 1545. At this time an epidemic spread so that everyone's nose bled. It had prevailed for a year when the market at San Hipólito opened"). The Spanish gloss beneath the corpses in the Codex Telleriano-Remensis (not pictured) states: *Año de 1544 y de mil y quinientos y quarenta y cinco buo vna gran mortandad entre los yndios* ("In the year 1544 and fifteen hundred and forty-five there was a great mortality among the Indians"). All images were re-drawn after the original illustrations, except for (a) Codex en Cruz, which was re-drawn from a facsimile produced by Dibble[59]. The figure was kindly provided by Annette Günzel.

Supplementary Table 1. Overview of the individuals from which samples were analyzed in this study, their archaeological burial context and associated radiocarbon dates.

					Radiocarbon Dates for Teposcolula-Yucundaa individuals from Tuross *et al.* 2014 (Supplemental ref. 19)			
Sample ID	**Individual number**	**Burial number**	**Cemetery Site**	**Contact or Pre-Contact era**	**Laboratory number**	**14C Age (years B.P.)**	**Calibarted Age (C.E.)**	**Confidence interval**
Tepos_1	1	1	Grand Plaza	Contact	OS-86842	470 ± 25	1415-1451	95,4
Tepos_4	4	3	Grand Plaza	Contact				
Tepos_9	9	5	Grand Plaza	Contact	OS-90540	420 ± 25	1430-1615	95,4
Tepos_10	10	6	Grand Plaza	Contact	OS-86838	400 ± 25	1438-1619	95,4
Tepos_11	11	6	Grand Plaza	Contact				
Tepos_12	12	6	Grand Plaza	Contact	OS-86840	365 ± 25	1430-1615	95,4
Tepos_13	13	7	Grand Plaza	Contact	OS-90571	375 ± 30	1446-1633	95,4
Tepos_14	14	8	Grand Plaza	Contact				
Tepos_15	15	8	Grand Plaza	Contact	OS-90569	370 ± 25	1449-1632	95,4
Tepos_16	16	9	Grand Plaza	Contact				
Tepos_17	17	9	Grand Plaza	Contact				
Tepos_18	18	9	Grand Plaza	Contact				
Tepos_19	19	9	Grand Plaza	Contact				
Tepos_20	20	9	Grand Plaza	Contact				
Tepos_24	24	10	Grand Plaza	Contact				
Tepos_26	26	12	Grand Plaza	Contact	OS-90534	395 ± 40	1435-1634	95,4
Tepos_27	27	13	Grand Plaza	Contact				
Tepos_32	32	17	Churchyard	Pre-Contact	OS-80882	530 ± 30	1320-1440	95,4
Tepos_34	34	18	Grand Plaza	Contact	OS-90572	355 ± 25	1454-1634	95,4
Tepos_35	35	18	Grand Plaza	Contact				
Tepos_36	36	19	Grand Plaza	Contact				
Tepos_37	37	19	Grand Plaza	Contact				
Tepos_38	38	19	Grand Plaza	Contact				
Tepos_41	41	21	Grand Plaza	Contact	OS-90570	365 ± 25	1446-1633	95,4
Tepos_43	43	23	Churchyard	Pre-Contact				
Tepos_45	45	25	Churchyard	Pre-Contact	OS-80887	615 ± 25	1295-1400	95,4
					OS-83897	545 ± 40	1305-1440	95,4
Tepos_46	46	26	Grand Plaza	Contact	OS-86839	475 ± 25	1413-1450	95,4
Tepos_48	48	27	Churchyard	Pre-Contact	OS-80888	490 ± 30	1403-1450	95,4
					OS-83883	560 ± 30	1306-1430	95,4
Tepos_57	57	32	Churchyard	Pre-Contact				

Supplementary Table 6. Publicly available *Salmonella* genomes used for array probe design.
Genomes in bold text were included in the modern strain dataset used in genome analyses and phylogenetic tree construction.

Identifier(s)	Name
NC_010067.1	**Salmonella enterica subsp. arizonae serovar 62:z4,z23:- str. RSK2980 chromosome, complete genome**
NC_021820.1	**Salmonella enterica subsp. enterica serovar Typhimurium str. 08-1736, complete genome**
NC_011149.1	**Salmonella enterica subsp. enterica serovar Agona str. SL483 chromosome, complete genome**
NC_021844.1	**Salmonella enterica subsp. enterica serovar Bareilly str. CFSAN000189, complete genome**
NC_022241.1	**Salmonella enterica subsp. enterica serovar Bovismorbificans str. 3114 complete genome**
NC_006905.1	**Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67 chromosome, complete genome**
NC_021818.1	**Salmonella enterica subsp. enterica Serovar Cubana str. CFSAN002050, complete genome**
NC_011205.1	**Salmonella enterica subsp. enterica serovar Dublin str. CT_02021853 chromosome, complete genome**
NC_011294.1	**Salmonella enterica subsp. enterica serovar Enteritidis str. P125109 chromosome, complete genome**
NC_011274.1	**Salmonella enterica subsp. enterica serovar Gallinarum str. 287/91 chromosome, complete genome**
NC_011083.1	**Salmonella enterica subsp. enterica serovar Heidelberg str. SL476 chromosome, complete genome**
NC_020307.1	Salmonella enterica subsp. enterica serovar Javiana str. CFSAN001992, complete genome
NC_011080.1	Salmonella enterica subsp. enterica serovar Newport str. SL254 chromosome, complete genome
NC_011147.1	**Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601 chromosome, complete genome**
NC_006511.1	**Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 chromosome, complete genome**
NC_010102.1	**Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7 chromosome, complete genome**
NC_012125.1	**Salmonella enterica subsp. enterica serovar Paratyphi C strain RKS4594 chromosome, complete genome**
NC_021984.1	**Salmonella enterica subsp. enterica serovar Pullorum str. S06004, complete genome**
NC_011094.1	**Salmonella enterica subsp. enterica serovar Schwarzengrund str. CVM19633 chromosome, complete genome**
NC_022525.1	Salmonella enterica subsp. enterica serovar Thompson str. RM6836, complete genome
NC_003198.1	**Salmonella enterica subsp. enterica serovar Typhi str. CT18, complete genome**
NC_016856.1	Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S chromosome, complete genome
NC_003197.1	**Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 chromosome, complete genome**
NC_016832.1	**Salmonella enterica subsp. enterica serovar Typhi str. P-stx-12, complete genome**
NC_021176.1	**Salmonella enterica subsp. enterica serovar Typhi str. Ty21a, complete genome**
NC_004631.1	**Salmonella enterica subsp. enterica serovar Typhi str. Ty2 chromosome, complete genome**
NT_187069.1-NT_187134.1	**Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1**
NZ_ABAK02000001.1	Salmonella enterica subsp. enterica serovar Kentucky str. CVM29188, whole genome shotgun sequence
NZ_ABAM02000001.1-NZ_ABAM02000002.1	Salmonella enterica subsp. enterica serovar Saintpaul str. SARA23, whole genome shotgun sequence
NZ_ABAN01000001.1-NZ_ABAN01000182.1	Salmonella enterica subsp. enterica serovar Saintpaul str. SARA29, whole genome shotgun sequence
NZ_APWL01000001.1-NZ_APWL01000004.1	Salmonella enterica subsp. enterica serovar Tennessee str. CDC07-0191, whole genome shotgun sequence
NZ_ABFG01000001.1-NZ_ABFG01000050.1	Salmonella enterica subsp. enterica serovar Hadar str. RI_05P066, whole genome shotgun sequence
NZ_ABFH02000001.1-NZ_ABFH02000003.1	Salmonella enterica subsp. enterica serovar Virchow str. SL491, whole genome shotgun sequence
NZ_AESM01000001.1-NZ_AESM01000034.1	Salmonella enterica subsp. enterica serovar Montevideo str. 515920-2 SEEM202, whole genome shotgun sequence
NZ_AETU01000001.1-NZ_AETU01000123.1	Salmonella enterica subsp. enterica serovar Montevideo str. IA_2010008284 SEEM8284, whole genome shotgun sequence
NZ_AFCI01000001.1-NZ_AFCI01002076.1	Salmonella enterica subsp. enterica serovar Adelaide str. A4-669, whole genome shotgun sequence
NZ_AFCJ01000001.1-NZ_AFCJ01002683.1	Salmonella enterica subsp. enterica serovar Alachua str. R6-377, whole genome shotgun sequence
NZ_AFCK01000001.1-NZ_AFCK01001762.1	Salmonella enterica subsp. enterica serovar Baildon str. R6-199, whole genome shotgun sequence
NZ_AFCL01000001.1-NZ_AFCL01002028.1	Salmonella enterica subsp. enterica serovar Gaminara str. A4-567, whole genome shotgun sequence
NZ_AFCM01000001.1-NZ_AFCM01001599.1	Salmonella enterica subsp. enterica serovar Give str. S5-487, whole genome shotgun sequence
NZ_AFCN01000001.1-NZ_AFCN01001812.1	Salmonella enterica subsp. enterica serovar Hvittingfoss str. A4-620, whole genome shotgun sequence
NZ_AFCO01000001.1-NZ_AFCO01002138.1	Salmonella enterica subsp. enterica serovar Inverness str. R8-3668, whole genome shotgun sequence
NZ_AFCP01000001.1-NZ_AFCP01002111.1	Salmonella enterica subsp. enterica serovar Johannesburg str. S5-703, whole genome shotgun sequence
NZ_AFCQ01000001.1-NZ_AFCQ01002468.1	Salmonella enterica subsp. enterica serovar Minnesota str. A4-603, whole genome shotgun sequence
NZ_AFCT01000001.1-NZ_AFCT01002416.1	Salmonella enterica subsp. enterica serovar Rubislaw str. A4-653, whole genome shotgun sequence
NZ_AFCV01000001.1-NZ_AFCV01001543.1	Salmonella enterica subsp. enterica serovar Uganda str. R8-3404, whole genome shotgun sequence
NZ_AFCW01000001.1-NZ_AFCW01002504.1	Salmonella enterica subsp. enterica serovar Urbana str. R8-2977, whole genome shotgun sequence
NZ_AFCX01000001.1-NZ_AFCX01002114.1	Salmonella enterica subsp. enterica serovar Wandsworth str. A4-580, whole genome shotgun sequence
NZ_AHIA01000001.1-NZ_AHIA01000094.1	Salmonella enterica subsp. enterica serovar Pomona str. ATCC 10729 SEEPO729, whole genome shotgun sequence
NZ_AJGK01000001.1-NZ_AJGK01000191.1	Salmonella enterica subsp. enterica serovar Typhi str. BL196, whole genome shotgun sequence
NZ_AJTD01000001.1-NZ_AJTD01000405.1	Salmonella enterica subsp. enterica serovar Typhi str. UJ308A, whole genome shotgun sequence
NZ_AJTE01000001.1-NZ_AJTE01000334.1	Salmonella enterica subsp. enterica serovar Typhi str. UJ816A, whole genome shotgun sequence
NZ_AJXA01000001.1-NZ_AJXA01000222.1	Salmonella enterica subsp. enterica serovar Typhi str. ST0208, whole genome shotgun sequence
NZ_AKIC01000001.1-NZ_AKIC01000538.1	Salmonella enterica subsp. enterica serovar Typhi str. CR0063, whole genome shotgun sequence
NZ_AKZO01000001.1-NZ_AKZO01000201.1	Salmonella enterica subsp. enterica serovar Typhi str. CR0044, whole genome shotgun sequence
NZ_ALPO01000001.1-NZ_ALPO01000047.1	Salmonella enterica subsp. enterica serovar Hartford str. 06-0676, whole genome shotgun sequence
NZ_ALPQ01000001.1-NZ_ALPQ01000142.1	Salmonella enterica subsp. enterica serovar Mississippi str. 2010K-1406, whole genome shotgun sequence
NZ_AMRS01000001.1-NZ_AMRS01000053.1	Salmonella enterica subsp. enterica serovar Mbandaka str. 2009K-0807, whole genome shotgun sequence
NZ_AMSN01000001.1-NZ_AMSN01000110.1	Salmonella enterica subsp. enterica serovar Thompson str. 2010K-1863, whole genome shotgun sequence
NZ_CAAR01000001.1-NZ_CAAR01001445.1	Salmonella enterica subsp. enterica serovar Typhi str. E00-7866, whole genome shotgun sequence
NZ_CAAT01000001.1-NZ_CAAT01000422.1	Salmonella enterica subsp. enterica serovar Typhi str. E02-1180, whole genome shotgun sequence
NZ_CAAV01000001.1-NZ_CAAV01003682.1	Salmonella enterica subsp. enterica serovar Typhi str. E98-2068, whole genome shotgun sequence
NZ_CAAW01000001.1-NZ_CAAW01001065.1	Salmonella enterica subsp. enterica serovar Typhi str. J185, whole genome shotgun sequence
NZ_CAAX01000001.1-NZ_CAAX01003024.1	Salmonella enterica subsp. enterica serovar Typhi str. M223, whole genome shotgun sequence
NZ_CAAZ01000001.1-NZ_CAAZ01000415.1	Salmonella enterica subsp. enterica serovar Typhi str. E98-3139, whole genome shotgun sequence
NZ_CAGQ01000001.1-NZ_CAGQ01000109.1	Salmonella enterica subsp. enterica serovar Senftenberg str. SS209, whole genome shotgun sequence
NZ_CM001274.1	Salmonella enterica subsp. enterica serovar Infantis str. SARB27 chromosome, whole genome shotgun sequence

Supplementary Table 8. Mapping statistics for the ten *S. enterica* positive samples mapped against the *S.* Typhi CT18 reference genome (NC_003198.1).

Sample ID	Library treatment	Capture method	Data type - Paired end (PE) or Single end (SE)	# raw reads	Total # raw reads	# of reads after paired-end data has been merged (and combined with single end data)	# mapped reads prior to duplicate removal	# quality filtered mapped reads prior to duplicate removal	# of duplicates removed	Mapped reads after duplicate removal	Endogenous DNA (%)	Endogenous DNA (%) quality filtered reads	Duplication factor
Tepos_10		Array capture	PE	138357636	150455954	68628270	10206000	9174537	6784941	2389596	14.871	13.368	3.839
	UDG	In Solution capture	PE	12098318									
Tepos_14		Array capture	PE	148509156	160571850	73204225	12303069	11079291	8296225	2783066	16.807	15.135	3.981
	UDG	In Solution capture	PE	12062694									
Tepos_35		Array capture	PE	192213420	197940170	90815050	24162883	21998780	16304865	5693915	26.607	24.224	3.864
	UDG	In Solution capture	PE	5726750									
Tepos_11		In Solution capture	SE	2633406	102207404	48959732	702978	302283	217000	85283	1.436	0.617	3.544
	UDG	In Solution capture	PE	99573998									
Tepos_20		In Solution capture	SE	6370988	36366380	20123713	982938	768666	420382	348284	4.884	3.820	2.207
	UDG	In Solution capture	PE	29995392									
Tepos_34		In Solution capture	SE	3803639	52584885	26284766	574625	346313	228138	118175	2.186	1.318	2.931
	UDG	In Solution capture	PE	48781246									
Tepos_36		In Solution capture	SE	2746026	44039996	21910196	233144	75731	48295	27436	1.064	0.346	2.760
	UDG	In Solution capture	PE	41293970									
Tepos_37		In Solution capture	SE	4759992	15464690	9603890	673332	566079	235355	330724	7.011	5.894	1.712
	UDG	In Solution capture	PE	10704698									
Tepos_38		In Solution capture	SE	7088118	95447736	47835731	494470	174339	126398	47941	1.034	0.364	3.637
	UDG	In Solution capture	PE	88359618									
Tepos_41		In Solution capture	SE	2595217	40568305	19966958	189450	43755	24256	19499	0.949	0.219	2.244
	UDG	In Solution capture	PE	37973088									

Supplementary Table 8 continued…

Sample ID	Mean coverage	Standard deviation in coverage	Coverage >= 1-fold (%)	Coverage >= 2-fold (%)	Coverage >= 3-fold (%)	Coverage >= 4-fold (%)	Coverage >= 5-fold (%)	% damage of first 5-prime base	% damage of second 5-prime base	% damage of first 3-prime base	% damage of second 3-prime base	Average fragment length (bp)	median fragment length (bp)	GC content (%)
Tepos_10	27.0186	35.84	86.71	84.95	82.89	80.63	78.18	0.0124	0.0118	0.0125	0.0117	54.37	52.0	46.44
Tepos_14	29.7373	19.4816	87.62	86.99	86.39	85.81	85.23	0.0173	0.01	0.0176	0.0095	51.38	48.0	50.96
Tepos_35	78.1635	46.4079	88.19	87.91	87.67	87.46	87.28	0.0125	0.0106	0.0128	0.0101	66.02	61.0	50.82
Tepos_11	0.9957	1.4371	52.6	26.21	11.52	4.71	1.87	0.016	0.0103	0.0185	0.0119	56.15	52.0	52.15
Tepos_20	3.7357	3.5407	78.82	67.7	55.83	44.47	34.41	0.0115	0.0083	0.0136	0.0093	51.58	48.0	52.12
Tepos_34	1.2359	1.6217	56.87	32.45	17.08	8.5	4.07	0.0145	0.0086	0.0147	0.0112	50.3	46.0	52.68
Tepos_36	0.2918	0.8419	21.78	4.77	1.02	0.3	0.16	0.0187	0.0104	0.018	0.0113	51.15	47.0	52.74
Tepos_37	4.3207	3.8066	81.03	72.23	62.19	51.75	41.67	0.0149	0.0093	0.0183	0.0105	62.83	59.0	52.6
Tepos_38	0.4708	1.0173	31.74	9.87	2.84	0.87	0.34	0.0169	0.0102	0.0189	0.0111	47.23	43.0	52.84
Tepos_41	0.1929	0.6474	15.32	2.55	0.49	0.16	0.1	0.0169	0.0099	0.0181	0.009	47.58	44.0	52.77

Supplementary Table 10a. Homoplastic positions within the dataset.

Position	Ref	SNP	Tepos_14	Tepos_35	ParatyphiC_RKS4594	CholeraesuisSC_B67	ParatyphiA_AKU12601	ParatyphiA_ATCC9150	ParatyphiB_SPB7	Typhi_CT18	Typhi_Pstx12	Typhi_Ty2	Typhi_Ty21a	Agona_SL483	Bareilly_CFSAN000189	Bovismorbificans_3114	Cubana_CFSAN002050	Dublin_CT02021853	Enteritidis_P125109	Gallinarum_28791	Heidelberg_SL476	Pullorum_S06004	Schwarzengrund_CVM19633	Typhimurium_081736	Typhimurium_LT2	Weltevreden_2007-60-3289-1	*S. enterica* subsp. arizonae	SNP Effect	Gene ID	Gene name	Gene function	old_AA/new_AA	Old_codon/New_codon	Codon_Num(CDS)	CDS_size
1734165	T	G	G	G	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	G	.	.	N	NON_SYNONYMOUS_CODING	SPC_1651	phsC	thiosulfate reductase cytochrome b subunit	V/G	gTc/gGc	110	765
2036957	G	A	A	A	.	.	A	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	N	INTERGENIC							
2333212	C	T	T	T	.	.	.	.	.	T	T	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	N	SYNONYMOUS_CODING	SPC_2274	rnfC	electron transport complex protein RnfC	H/H	caC/caT	232	2115
2337991	T	C	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	.	.	.	N	INTERGENIC							
2647235	C	T	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	N	SYNONYMOUS_CODING	SPC_2611	csgG	assembly/transport component in curli production	N/N	aaC/aaT	148	834
4024083	G	T	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	SYNONYMOUS_CODING	SPC_3967	rbsA	D-ribose transporter ATP binding protein	S/S	tcG/tcT	75	1551
4092379	G	A	A	A	.	.	N	N	.	N	N	N	N	.	.	N	A	.	.	.	.	.	.	N	N	.	N	SYNONYMOUS_CODING	SPC_4022	ilvC	ketol-acid reductoisomerase	V/V	gtG/gtA	171	1476

Supplementary Table 10b. Positions with Tri-allelic states within the dataset.

Position	Ref	SNP	Tepos_14	Tepos_35	Paratyphi C RKS4594	Choleraesuis SC-B67	Paratyphi A AKU_12601	Paratyphi A ATCC 9150	Paratyphi B SPB7	Typhi CT18	Typhi P-stx-12	Typhi Ty2	Typhi Ty21a	Agona SL483	Bareilly CFSAN000189	Bovismorbificans 3114	Cubana CFSAN002050	Dublin CT_02021853	Enteritidis P125109	Gallinarum 287/91	Heidelberg SL476	Pullorum S06004	Schwarzengrund CVM19633	Typhimurium 08-1736	Typhimurium LT2	Weltevreden 2007-60-3289-1	*S. enterica* subsp. arizonae	SNP Effect	Gene ID	Gene name	Gene function	old_AA/new_AA	Old_codon/New_codon	Codon_Num(CDS)	CDS_size
3210239	G	A	A	A	.	.	T	T	.	T	T	T	T	.	T	.	.	T	.	.	.	.	.	.	.	.	N	SYNONYMOUS_CODING	SPC_3180	nupG	nucleoside permease	L/L	ctG/ctA	360	1257
3210239	G	T	A	A	.	.	T	T	.	T	T	T	T	.	T	.	.	T	.	.	.	.	.	.	.	.	N	SYNONYMOUS_CODING	SPC_3180	nupG	nucleoside permease	L/L	ctG/ctT	360	1257
3531094	G	A	A	A	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	T	.	.	.	.	N	INTERGENIC							
3531094	G	A	A	A	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	T	.	.	.	.	N	UPSTREAM: 59 bases	SPC_3513	yheA	bacterioferritin-associated ferredoxin				195
3531094	G	T	A	A	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	T	.	.	.	.	N	INTERGENIC							
3531094	G	T	A	A	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	T	.	.	.	.	N	UPSTREAM: 59 bases	SPC_3513	yheA	bacterioferritin-associated ferredoxin				195
3804669	G	A	A	A	.	T	.	.	N	N	N	N	N	N	.	N	N	.	.	.	.	.	.	N	N	A	N	NON_SYNONYMOUS_CODING	SPC_3759	SPC_3759	hypothetical protein	G/S	Ggc/Agc	462	1956
3804669	G	T	A	A	.	T	.	.	N	N	N	N	N	N	.	N	N	.	.	.	.	.	.	N	N	A	N	NON_SYNONYMOUS_CODING	SPC_3759	SPC_3759	hypothetical protein	G/C	Ggc/Tgc	462	1956
4213532	C	T	G	G	.	.	N	N	N	N	N	N	N	T	.	N	T	N	N	N	N	N	T	N	N	N	T	NON_SYNONYMOUS_CODING	SPC_4134	bah	acetyl esterase	R/K	aGg/aAg	105	678
4213532	C	G	G	G	.	.	N	N	N	N	N	N	N	T	.	N	T	N	N	N	N	N	T	N	N	N	T	NON_SYNONYMOUS_CODING	SPC_4134	bah	acetyl esterase	R/T	aGg/aCg	105	678

Supplementary Table 12. Overview of the regions present in the high-coverage ancient genomes (Tepos_14 and Tepos_35) that are present in the other genomes of interest, but absent in the *S.* Paratyphi C RKS4594 reference. Only regions equal to or larger than 700bp in one or both of the ancient genomes are reported.
No regions larger than 700bp were identified to be mapping to *S.* Paratyphi A in either of the two high-coverage ancient genomes.

	Tepos_14			Tepos_35						
Reference strain	start position	end position	region size (bp)	start position	end position	region size (bp)	Extra info about region	Name of affected genes	Locus tag of affected genes	Gene product annotation
Typhi CT18 (NC_003198.1)	1473611	1475563	1952	1473625	1475686	2061		pqaA	STY1518	PhoPQ-activated pathogenicity-related protein PqaA
								STY1519	STY1519	membrane transport protein
	1476961	1478406	1445	1476963	1478425	1462		STY1520	STY1520	alcohol dehydrogenase
								STY1521	STY1521	regulatory protein
	1538733	1546701	7968	1538732	1546701	7969	bacteriophage	STY1591	STY1591	bacteriophage transciptional regulator
								STY1592	STY1592	hypothetical protein
								STY1594	STY1594	hypothetical protein
								STY1595	STY1595	hypothetical protein
								STY1598	STY1598	hypothetical protein
								gam	STY1601	bacteriophage host-nuclease inhibitor protein
								eha	STY1603	ExeA protein
								STY1604	STY1604	hypothetical protein
	1572249	1573001	752	1572244	1573001	757	bacteriophage	STY1643	STY1643	DNA-invertase
	2046859	2047499	640	2046798	2047530	732		STYt026	STYt026	tRNA-Asn
								STYt027	STYt027	tRNA-Asn
	4432019	4436407	4388	4432031	4436412	4381	Salmonella Pathogenicity island 7 (SPI-7)	pilS	STY4547	prepilin
								pilT	STY4548	hypothetical protein
								pilU	STY4549	prepilin pepetidase
								pilV	STY4550	prepilin
								rci	STY4552	shufflon-specific DNA recombinase
Paratyphi B (NC_010102.1)	1003612	1004556	944	1003553	1004563	1010		SPAB_01108	SPAB_01108	tRNA-Asn
								SPAB_01110	SPAB_01110	hypothetical protein
								SPAB_01109	SPAB_01109	hypothetical protein
								SPAB_01111	SPAB_01111	hypothetical protein
								SPAB_01112	SPAB_01112	hypothetical protein
								SPAB_01113	SPAB_01113	hypothetical protein
	1499879	1501982	2103	1499879	1502011	2132		SPAB_01757	SPAB_01757	hypothetical protein
								SPAB_01758	SPAB_01758	hypothetical protein
								SPAB_01759	SPAB_01759	hypothetical protein
								SPAB_01760	SPAB_01760	hypothetical protein
	1503286	1504731	1445	1503288	1504750	1462		SPAB_01761	SPAB_01761	hypothetical protein
								SPAB_01762	SPAB_01762	hypothetical protein
Choleraesuis (NC_006905.1)	1651241	1651975	734	1651237	1651976	739		pqaA	SC1557	PhoPQ-regulated protein
	2105005	2105992	987	2104944	2106051	1107		SCTRNA34	SCTRNA34	tRNA-Asn
								SC2004	SC2004	hypothetical protein
								int	SC2005	P4-type integrase
								SC2006	SC2006	hypothetical protein
								SCTRNA35	SCTRNA35	tRNA-Asn

Supplementary Table 13. Source information for the effector genes

Gene name	Gene/genome sequence identifier	Range (bp)
avrA	AF013573.1	1-1388
cigR	16763390	c3961227-3960748
gogB	169257208	425-1918
gtgA	169257267	c16435-15749
gtgE	661245796	1265203-1265889
pipB	16763390	c1177325-1176450
pipB2	16758993	c2780918-2779866
sifA	16758993	c1220339-1219329
sifB	16758993	c1413513-1412563
sipA/sspA	16758993	c2878746-2876689
sipB	16758993	c2882895-2881114
sipC/sspC	16758993	c2881086-2879857
sipD/sspD	16758993	c2879787-2878765
slrP	16763390	866944-869241
sopA	16763390	2141570-2143918
sopB/sigD	16758993	c1090372-1088687
sopD	16763390	3087148-3088101
sopD2	16763390	1054061-1055020
sopE	16758993	4481801-4482523
sopE2	16763390	c1952734-1952012
spiC/ssaB	16758993	c1646652-1646269
sptP	16763390	c3023702-3022071
spvB	521233317	c45843-44068
spvC	410654531	c63954-63229
spvD	261888681	c23960-23310
srfJ	16763390	4669016-4670359
sseB	16763390	1483934-1484524
sseC	16763390	1484997-1486451
sseD	16763390	1486467-1487054
sseF	16763390	1487975-1488757
sseG	16763390	1488754-1489443
sseI	16763390	1139971-1140939
sseJ	16763390	1721293-1722519
sseK1	16763390	4375350-4376360
sseK2	16763390	2231496-2232542
sseK3	AY055382.1	19334-20341
sseL	16763390	2394726-2395748
sspA	16763390	c3511191-3510553
sspH2	16763390	c2343251-2340885
steA	16758993	2948970-2949557
steB	16758993	2949638-2952337
steC	16758993	2952350-2953123
steD	16758993	2953143-2953649

Supplementary Table 14. Mapping statistics for the UDG treated data for the ten positive samples mapped against the pSPCV reference (NC_012124.1).

Sample ID	Library treatment	Capture method	Data type - Paired end (PE) or Single end (SE)	# raw reads	Total # raw reads	# of reads after paired-end data has been merged (and combined with single-end data)	# mapped reads prior to duplicate removal	# quality filtered mapped reads prior to duplicate removal	# of duplicates removed	Mapped reads after duplicate removal	Endogenous DNA (%)	Endogenous DNA (%) quality filtered reads	Duplication factor
Tepos_10		Array capture	PE	138357636	150455954	68628270	258495	256763	194338	62425	0.377	0.374	4.113
	UDG	In Solution capture	PE	12098318									
Tepos_14		Array capture	PE	148509156	160571850	73204225	343876	339990	255128	84862	0.470	0.464	4.006
	UDG	In Solution capture	PE	12062694									
Tepos_35		Array capture	PE	192213420	197940170	90815050	749785	743265	553809	189456	0.826	0.818	3.923
	UDG	In Solution capture	PE	5726750									
Tepos_11		In Solution capture	SE	2633406	102207404	48959732	7440	7394	5287	2107	0.015	0.015	3.509
	UDG	In Solution capture	PE	99573998									
Tepos_20		In Solution capture	SE	6370988	36366380	20123713	27184	26853	15658	11195	0.135	0.133	2.399
	UDG	In Solution capture	PE	29995392									
Tepos_34		In Solution capture	SE	3803639	52584885	26284766	11739	11660	7747	3913	0.045	0.044	2.980
	UDG	In Solution capture	PE	48781246									
Tepos_36		In Solution capture	SE	2746026	44039996	21910196	2303	2285	1504	781	0.011	0.010	2.926
	UDG	In Solution capture	PE	41293970									
Tepos_37		In Solution capture	SE	4759992	15464690	9603890	20763	20581	9421	11160	0.216	0.214	1.844
	UDG	In Solution capture	PE	10704698									
Tepos_38		In Solution capture	SE	7088118	95447736	47835731	5684	5567	4158	1409	0.012	0.012	3.951
	UDG	In Solution capture	PE	88359618									
Tepos_41		In Solution capture	SE	2595217	40568305	19966958	807	799	393	406	0.004	0.004	1.968
	UDG	In Solution capture	PE	37973088									

Supplementary Table 14 continued...

Sample ID	Mean coverage	Standard deviation in coverage	Coverage >= 1-fold (%)	Coverage >= 2-fold (%)	Coverage >= 3-fold (%)	Coverage >= 4-fold (%)	Coverage >= 5-fold (%)	% damage of first 5-prime base	% damage of second 5-prime base	% damage of first 3-prime base	% damage of second 3-prime base	Average fragment length (bp)	median fragment length (bp)	GC content (%)
Tepos_10	62.3917	68.3378	99.81	99.46	98.75	97.9	96.8	0.0084	0.002	0.0081	0.0024	55.38	53.0	45.23
Tepos_14	79.0277	35.7556	99.96	99.92	99.86	99.83	99.71	0.0111	0.003	0.0113	0.0028	51.6	49.0	51.07
Tepos_35	224.4831	88.337	100	99.99	99.98	99.95	99.92	0.0065	0.0025	0.0066	0.0022	65.66	61.0	50.7
Tepos_11	2.1575	1.8473	81.19	56.14	35.61	20.89	11.63	0.0063	0.0063	0.0063	0.0047	56.74	53.0	52.42
Tepos_20	10.5534	7.241	96.92	93.5	89.44	84.52	78.77	0.0068	0.0034	0.0102	0.0038	52.24	49.0	52.52
Tepos_34	3.5827	3.0656	86.7	71.18	55.69	42.34	30.8	0.0086	0.0067	0.0196	0.0049	50.74	46.0	52.88
Tepos_36	0.7473	1.0229	46.55	18.67	6.15	2.13	0.78	0.0058	0.0061	0.0	0.0116	53.02	49.0	52.48
Tepos_37	12.7835	8.2865	97.63	94.93	91.54	88.21	84.17	0.0072	0.0031	0.0158	0.0045	63.48	59.0	53.24
Tepos_38	1.2335	1.3455	62.98	33.8	15.45	6.85	2.7	0.0142	0.0034	0.0086	0.0109	48.51	44.0	53.25
Tepos_41	0.3503	0.637	27.86	5.77	1.14	0.2	0.07	0.0104	0.0	0.011	0.0133	47.81	45.0	51.9

Supplementary Table 15. SNP effect analysis of the ancient *S.* Paratyphi C pSPCV virulence plasmid in comparison to closely related virulence plasmids.

Position	Ref	SNP	Tepos_14	Tepos_35	pSPCV	pKDSC50	pSLT	pSCV50	SNP Effect	Gene ID	Gene name	Gene function	old_AA/ new_AA	Old_codon/ New_codon	Codon_Num(CDS)	CDS_size
11956	G	A	A	A	.	A	A	A	SYNONYMOUS_CODING	SPC_p016	ccdB	plasmid maintenance protein CcdB	S/S	agC/agT	12	306
11956	G	A	A	A	.	A	A	A	DOWNSTREAM: 37 bases	SPC_p017	ccdA	plasmid maintenance protein CcdA				219
12805	A	G	G	G	.	G	N	G	INTERGENIC							
13219	G	A	A	A	.	A	N	A	NON_SYNONYMOUS_CODING	SPC_p018	SPC_p018	hypothetical protein	G/R	Ggg/Agg	101	489
21099	A	T	T	.	.	.	.	.	NON_SYNONYMOUS_CODING	SPC_p028	pefD	PefD	S/C	Agc/Tgc	95	693
26503	A	G	G	G	.	G	.	G	NON_SYNONYMOUS_CODING	SPC_p037	repA3	DNA replication protein	S/P	Tct/Cct	2	123
36266	A	G	G	G	.	G	G	G	NON_SYNONYMOUS_CODING	SPC_p045	traD	conjugal transfer protein TraD	V/A	gTc/gCc	350	2244
43352	C	A	A	A	.	.	.	.	NON_SYNONYMOUS_CODING	SPC_p053	parA	ParA	D/Y	Gat/Tat	190	1206
47604	A	G	G	G	.	G	G	G	NON_SYNONYMOUS_CODING	SPC_p059	tlpA	TlpA	C/R	Tgt/Cgt	155	1116
47679	A	G	G	G	.	G	G	G	SYNONYMOUS_CODING	SPC_p059	tlpA	TlpA	L/L	Ttg/Ctg	130	1116

References:

1 Balkansky, A. K. *et al*. Archaeological survey in the Mixteca Alta of Oaxaca, Mexico. *Journal of Field Archaeology* **27**, 365-389 (2000).

2 Spores, R. & Robles García, N. A prehispanic (postclassic) capital center in colonial transition: excavations at Yucundaa Pueblo Viejo de Teposcolula, Oaxaca, Mexico. *Lat Am Antiq* **18**, 33-353 (2007).

3 Stiver, L. *Prehispanic Mixtec settlement and state in the Teposcolula Valley of Oaxaca, Mexico*. (Doctoral dissertation, Vanderbilt University, Nashville, 2001).

4 Warinner, C. G. *Life and death at Teposcolula Yucundaa: Mortuary, archaeogenetic, and isotopic investigations of the early colonial period in Mexico*. (Ph.D. dissertation, Department of Anthropology, Harvard University, Cambridge, 2010).

5 Warinner, C., Robles García, N., Spores, R. & Tuross, N. Disease, Demography, and Diet in Early Colonial New Spain: Investigation of a Sixteenth-Century Mixtec Cemetery at Teposcolula Yucundaa. *Lat Am Antiq* **23**, 467-489 (2012).

6 Chimalpahin Cuauhtlehuanitzin, D. F. d. S. *Codex Chimalpahin: society and politics in Mexico Tenochtitlan, Tlateloco, Texcoco, Culhuacan, and other Nahua altepetl in Central Mexico*. (University of Oklahoma Press, 1997).

7 Oudijk, M. R. & Restall, M. in *Indian conquistadors: indigenous allies in the conquest of Mesoamerica* (University of Oklahoma Press, 2007).

8 Calderón Galván, E. *Teposcolula: breve ensayo monográfico*., (Sectaría de Desarollo Económico y Social, Dirección general de Educación, Cultura y Bienestar Social del Gobierno del Estado de Oaxaca, 1988).

9 Robles García, N. & Spores, R. Teposcolula, Oaxaca. *Arqueología Mexicana* **15**, 42-43 (2008).

10 Spores, R. & Robles García, N. M. *Yucundaa: la ciudad Mixteca y su transformación prehispánica-colonial México*. (Instituto Nacional de Antropología e Historia-Fundación Alfredo Harp Helú, 2014).

11 Vences Vidal, M. *Evangelización y arquitectura dominicana en Coixtlahuaca (Oaxaca) México*. (Editorial San Esteban, 2000).

12 Jimenez Moreno, W. & Mateos Higuera, S. *Codice de Yanhuitlan*. (Museo Nacional, Secretaria de Educacion Publica, Instituto Nacional de Antropología e Historia, 1940).

13 Pérez Ortiz, A. *Tierra de Brumas: Conflictos en la Mixteca Alta, 1523-1550*. (Plaza y Valdés, 2003).

14 Terraciano, K. *The Mixtecs of colonial Oaxaca: Ñudzahui history, sixteenth through eighteenth centuries*. 362 (Stanford University Press, 2001).

15 Acuña, R. *Relaciones geográficas del siglo XVI*. Vol. Vol. 2-3, Antequera (Universidad Nacional Autónoma de México, 1984).

16 Spores, R. A., Robles García, N., Diego Luna, L., Roldán López, L. L. & Ruiz Ríos, I. G. Avances de investigación de los entierros humanos del sitio Pueblo Viejo de Teposcolula y su contexto arqueológico. *Estudios de antropología biológica* **13** (2007).

17 Spores, R., Robles García, N., Diego, L. L. & Ixchel, R. Epidemias en la Mixteca Alta durante el siglo XVI: evidence arqueológica en el Pueblo Viejo de Teposcolula. *Quaderni di Thule: Rivista italiana di studi americanístici : Actas del XXVIII congreso internacional de Americanística*, 869-876 (2006).

18 Spores, R. & Robles Garcia, N. A prehispanic (postclassic) capital center in colonial transition: excavations at Yucundaa Pueblo Viejo de Teposcolula, Oaxaca, Mexico. *Latin American Antiquity* **18**, 33-353 (2007).

19 Tuross, N., Warinner, C. & Robles García, N. in *Yucundaa: La cuidad mixteca Yucundaa-Pueblo Viejo de Teposcolula y su transformación prehispánica-colonial* Vol. vol. 2 (eds Ronald Spores & Robles García Nelly) 541-546 (Instituto Nacional de Antropología e Historia, 2014).

20 Dabney, J. *et al*. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15758-15763, doi:10.1073/pnas.1314445110 (2013).

21 Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010**, pdb prot5448, doi:10.1101/pdb.prot5448 (2010).

22 Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40**, e3, doi:10.1093/nar/gkr771 (2012).
23 Peltzer, A. *et al*. EAGER: efficient ancient genome reconstruction. *Genome Biol* **17**, 60, doi:10.1186/s13059-016-0918-z (2016).
24 Huson, D. H. *et al*. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol* **12**, e1004957, doi:10.1371/journal.pcbi.1004957 (2016).
25 Chen, T. *et al*. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**, baq013, doi:10.1093/database/baq013 (2010).
26 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
27 Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682-1684, doi:10.1093/bioinformatics/btt193 (2013).
28 Hu, X. *et al*. Vi capsular polysaccharide: Synthesis, virulence, and application. *Crit Rev Microbiol* **43**, 440-452, doi:10.1080/1040841X.2016.1249335 (2017).
29 Pickard, D. *et al*. Molecular characterization of the Salmonella enterica serovar Typhi Vi-typing bacteriophage E1. *Journal of bacteriology* **190**, 2580-2587, doi:10.1128/JB.01654-07 (2008).
30 Fu, Q. *et al*. DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2223-2227, doi:10.1073/pnas.1221359110 (2013).
31 Briggs, A. W. *et al*. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* **38**, e87, doi:10.1093/nar/gkp1163 (2010).
32 Hodges, E. *et al*. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**, 960-974, doi:10.1038/nprot.2009.68 (2009).
33 DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).
34 Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* **9**, 88, doi:10.1186/s13104-016-1900-2 (2016).
35 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
36 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
37 Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
38 Vimr, E. R., Kalivoda, K. A., Deszo, E. L. & Steenbergen, S. M. Diversity of microbial sialic acid metabolism. *Microbiol Mol Biol Rev* **68**, 132-153 (2004).
39 Campbell, J. W., Morgan-Kiss, R. M. & Cronan, J. E., Jr. A new Escherichia coli metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. *Molecular microbiology* **47**, 793-805 (2003).
40 Rivera-Chavez, F. *et al*. Salmonella uses energy taxis to benefit from intestinal inflammation. *PLoS Pathog* **9**, e1003267, doi:10.1371/journal.ppat.1003267 (2013).
41 Rivera-Chavez, F. *et al*. Energy Taxis toward Host-Derived Nitrate Supports a Salmonella Pathogenicity Island 1-Independent Mechanism of Invasion. *MBio* **7**, doi:10.1128/mBio.00960-16 (2016).
42 Clifton, M. C. *et al*. In vitro reassembly of the ribose ATP-binding cassette transporter reveals a distinct set of transport complexes. *J Biol Chem* **290**, 5555-5565, doi:10.1074/jbc.M114.621573 (2015).
43 Heinzinger, N. K., Fujimoto, S. Y., Clark, M. A., Moreno, M. S. & Barrett, E. L. Sequence analysis of the phs operon in Salmonella typhimurium and the contribution of thiosulfate reduction to anaerobic energy metabolism. *Journal of bacteriology* **177**, 2813-2820 (1995).

44 Stoffels, L., Krehenbrink, M., Berks, B. C. & Unden, G. Thiosulfate reduction in Salmonella enterica is driven by the proton motive force. *Journal of bacteriology* **194**, 475-485, doi:10.1128/JB.06014-11 (2012).

45 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

46 Black, L. W. Old, new, and widely true: The bacteriophage T4 DNA packaging mechanism. *Virology* **479-480**, 650-656, doi:10.1016/j.virol.2015.01.015 (2015).

47 Seth-Smith, H. M. SPI-7: Salmonella's Vi-encoding Pathogenicity Island. *J Infect Dev Ctries* **2**, 267-271 (2008).

48 Morris, C., Yip, C. M., Tsui, I. S., Wong, D. K. & Hackett, J. The shufflon of Salmonella enterica serovar Typhi regulates type IVB pilus-mediated bacterial self-association. *Infect Immun* **71**, 1141-1146 (2003).

49 Tam, C. K., Hackett, J. & Morris, C. Salmonella enterica serovar Paratyphi C carries an inactive shufflon. *Infect Immun* **72**, 22-28 (2004).

50 Tam, C. K., Morris, C. & Hackett, J. The Salmonella enterica serovar Typhi type IVB self-association pili are detached from the bacterial cell by the PilV minor pilus proteins. *Infect Immun* **74**, 5414-5418, doi:10.1128/IAI.00172-06 (2006).

51 Connor, T. R. *et al*. What's in a Name? Species-Wide Whole-Genome Sequencing Resolves Invasive and Noninvasive Lineages of Salmonella enterica Serotype Paratyphi B. *MBio* **7**, doi:10.1128/mBio.00527-16 (2016).

52 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

53 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

54 R Development Core Team. (The R Foundation for Statistical Computing, Vienna, Austria, 2011).

55 Baison-Olmo, F., Cardenal-Munoz, E. & Ramos-Morales, F. PipB2 is a substrate of the Salmonella pathogenicity island 1-encoded type III secretion system. *Biochem Biophys Res Commun* **423**, 240-246, doi:10.1016/j.bbrc.2012.05.095 (2012).

56 Rotger, R. & Casadesus, J. The virulence plasmids of Salmonella. *Int Microbiol* **2**, 177-184 (1999).

57 Liu, W. Q. *et al*. Salmonella paratyphi C: genetic divergence from Salmonella choleraesuis and pathogenic convergence with Salmonella typhi. *PLoS One* **4**, e4510, doi:10.1371/journal.pone.0004510 (2009).

58 Haneda, T., Okada, N., Nakazawa, N., Kawakami, T. & Danbara, H. Complete DNA sequence and comparative analysis of the 50-kilobase virulence plasmid of Salmonella enterica serovar Choleraesuis. *Infect Immun* **69**, 2612-2620, doi:10.1128/IAI.69.4.2612-2620.2001 (2001).

59 Dibble, C. E. *Codex en Cruz*. Vol. 2 (University of Utah Press, 1981).

60 *Codex Mexicanus*. Vol. planch 81 (Bibliothèque nationale de Paris).

61 *Tira de Tepechpan : códice colonial procedente del Valle de México*. (Biblioteca del Estado, 1978).

62 Codex Aubin. in *Quellenwerke zur alten Geschichte Amerikas 13* (ed Walter Lehmann and Gerdt Kutscher) (Mann Verlag, 1981).

63 *Telleriano-Remensis, C. Pictografia mexicana del siglo XVI*. (Echaniz, 1963).

64 *Anales de San Gregorio Acapulco 1520-1606. Tlalocan*. Vol. 3 110 (1952).

65 Anales de Tecamachalco. in *Colección de documentos para la historia mexicana* (ed Antonio Peñafiel) (Secretaría de Fomento, 1897-1903).

66 Mendieta, G. d. *Historia eclesiastica indiana obra escrita a fines del siglo XVI / por fray Geronimo de Mendieta*. (Antigua Libreria, 1870).

67 Sahagún, B. d. *Florentine Codex: general history of the things of New Spain*. (1950-1982).

68 Fomento, S. M. d. *Cartas de Indias*. 1v (Hernandez Madrid 1877).

69 Gay, J. A. *Historia de Oaxaca*. (Impr. del Comercio, de Dublan y ca, 1881).

70 Muñoz Camargo, D. *Historia de Tlaxcala*. (Dastin Export S.L., 2002).

71 Chimalpahin, D. F. d. S. A. M. in *Teil 2: Das Jahrhundert nach der Conquista* 13, 76 (Cram, de Gruyter, 1965).

72 Troncoso, F. P. in *Papeles de Nueva España publicados de orden y con fondos del gobierno mexicano por Francisco del Paso y Troncoso Geográfica y Estadística* (Editoral Cosmos, 1979).

73 Fields, S. L. *Pestilence and Headcolds: Encountering Illness in Colonial Mexico*. (Columbia University Press, 2008).

74 Real Academia de la Historia & Troncoso, F. P. *Relaciones geograficas de la diocesis de Mexico: Manuscritos de la Real Academia de la historia de Madrid y del Archivo de Indias en Sevilla. Anos 1579-1582*. (Est. tipografico "Sucesores de Rivadeneyra," 1905).
75 Cook, N. D. *Born to Die: Disease and New World Conquest, 1492-1650*. (Cambridge University Press, 1998).
76 Gibson, C. *The Aztecs Under Spanish Rule: A History of the Indians of the Valley of Mexico, 1519-1810*. (Stanford University Press, 1964).
77 Humboldt, A. v. *Political essay on the Kingdom of new Spain*. (Longman, Hurst, Rees, Orme, and Brown, 1811).
78 Sticker, G. DIE EINSCHLEPPUNG EUROPÄISCHER KRANKHEITEN IN AMERIKA WÄHREND DER ENTDECKUNGSZEIT; IHR EINFLUSS AUF DEN RÜCKGANG DER BEVÖLKERUNG. *Ibero-amerikanisches Archiv* **6**, 194-224 (1932).
79 Sticker, G. DIE EINSCHLEPPUNG EUROPÄISCHER KRANKHEITEN IN AMERIKA WÄHREND DER ENTDECKUNGSZEIT; IHR EINFLUSS AUF DEN RÜCKGANG DER BEVÖLKERUNG. *Ibero-amerikanisches Archiv* **6**, 62-83 (1932).
80 Zinsser, H. *Rats, lice and history : being a study in biography, which, after twelve preliminary chapters indispensable for the preparation of the lay reader, deals with the life history of typhus fever* (Little, Brown, and Co., 1935).
81 MacLeod, M. J. *Spanish Central America: A Socioeconomic History, 1520-1720*. (University of California Press, 1973).
82 Malvido, E. & Viesca, C. La epidemia de cocoliztli de 1576. *Historias (México, D.F.)* **11**, 27-33 (1985).
83 Orellana, S. L. *Indian medicine in highland Guatemala : the pre-Hispanic and colonial periods*. (University of New Mexico Press, 1987).
84 Acuna-Soto, R., Romero, L. C. & Maguire, J. H. Large epidemics of hemorrhagic fevers in Mexico 1545-1815. *Am J Trop Med Hyg* **62**, 733-739 (2000).
85 Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Griffin, R. D. & Cleaveland, M. K. When half of the population died: the epidemic of hemorrhagic fevers of 1576 in Mexico. *FEMS Microbiol Lett* **240**, 1-5, doi:10.1016/j.femsle.2004.09.011 (2004).
86 Marr, J. S. & Kiracofe, J. B. Was the huey cocoliztli a haemorrhagic fever? *Medical History* **44**, 341-362 (2000).
87 Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and megadeath in 16th century Mexico. *Emerg Infect Dis* **8**, 360-362 (2002).
88 Ashburn, P. M. *The ranks of death: a medical history of the conquest of Americа*. (Coward-McCann, 1947).
89 Cook, N. D. & Lovell, W. G. *Secret Judgments of God: Old World Disease in Colonial Spanish America*. (University of Oklahoma Press, 2001).
90 Wu, X., Lu, Y., Zhou, S., Chen, L. & Xu, B. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environ Int* **86**, 14-23, doi:10.1016/j.envint.2015.09.007 (2016).
91 Ashburn, P. M. Smallpox, the Plague of Athens. *Military Surgeon* **LXIX**, 188-190 (1931).
92 Lind, J. *An Essay On Diseases Incidental to Europeans: In Hot Climates, With the Method of Preventing Their Fatal Consequences* (1758).
93 Sawyer, W. A. Recent progress in yellow fever research. *Medicine* **X** (1931).
94 Torre, T. d. l. *Diario de viaje de Salamanca a Ciudad Real de Chiapa, 1544-1545*. (Editorial OPE, 1985).
95 Aguilar-Moreno, M. *Handbook to Life in the Aztec World*. (Oxford University Press, 2007).
96 Keber, E. Q. *Codex Telleriano-Remensis: Ritual, Divination, and History in a Pictorial Aztec Manuscript*. (University of Texas Press, 1995).
97 Rappuoli, R., Pizza, M., Del Giudice, G. & De Gregorio, E. Vaccines, new opportunities for a new society. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12288-12293, doi:10.1073/pnas.1402981111 (2014).
98 Prem, H. J. in *Secret Judgements of God: Old World disease in Colonial Spanish America* (eds Noble David Cook & W. George Lovell) 20-48 (University of Oklahoma Press, 1992).

Paper III

Å. J. Vågene, J. Krause and K. I. Bos. (2016).
Metagenomic analysis and mitochondrial genome reconstruction of the post-medieval individual from Moneen Cave.
In: Dowd, M. (Ed.), Archaeological Excavations in Moneen Cave, The Burren, Co. Clare. Oxford, England: Archaeopress Publishing Ltd. (pp. 49-52).

Metagenomic analysis and mitochondrial genome reconstruction of the post-medieval individual from Moneen Cave

Åshild J. Vågene, Johannes Krause and Kirsten I. Bos

In the context of post-medieval Christian burial practices in Ireland, the internment of an adolescent individual in Moneen Cave is considered to be non-normative (Garattini 2007; Murphy 2011). Whether this individual was purposefully 'buried' in Moneen Cave or died there circumstantially is not clear from the archaeological evidence (Dowd 2013a). Both scenarios raise questions relating to how this individual's remains came to be interned in such a hidden place. One potential reason for the sequestering of this individual's remains could have been infectious disease. This unusual burial placement may have resulted from an intention to distance this individual from local inhabitants.

Osteological analyses determined this individual to be between 14-16 years of age at the time of death (Section 15). Due to the individual's young age, sex could not be determined using traditional osteological methods. However, previous ancient DNA analyses found this individual to be male (Section 16). No skeletal changes consistent with infectious disease were found on the remains (Section 15); however, the majority of infectious diseases affect soft tissue and do not cause changes to skeletal elements.

In August 2014 a right first maxillary incisor from the Moneen individual was collected for ancient DNA analyses. The purpose of these investigations was two-fold. First, the microbial DNA content was evaluated and screened for molecular traces of ancient pathogenic microbes that may have infected this individual. Secondly, the mitochondrial DNA (mtDNA) of this individual was reconstructed to determine the mitochondrial lineage, or haplogroup, to which this individual belongs. Mitochondria are multi-copy cellular organelles that have their own genetic makeup, termed mtDNA. Their small circular genome is 16,569bp in length and is only inherited through the maternal lineage. Haplogroups represent the nomenclature used to refer to subsets of mutational variation that exists amongst human mtDNAs.

Methods and results

Sample preparation and sequencing

The incisor was horizontally cross-sectioned at the cementoenamel junction, and 34mg of dentin was drilled from the pulp chamber. The sample was extracted using an established protocol tailored for the extraction of ancient DNA (Dabney *et al*. 2013). The extract was eluted in 100ul TET (10mM Tris, 1mM EDTA, and 0.05% Tween), 10ul of which were converted into a double-stranded DNA-library (Meyer and Kircher 2010), followed by double indexing using a library specific barcode combination (Kircher *et al*. 2012). A portion of the indexed library was further amplified using Herculase II Fusion DNA Polymerase (Agilent) and reactions were suspended before reaching saturation. The amplified indexed library was shotgun sequenced on a HiSeq 4000 lane, producing 21,667,649 single-end reads. Extraction and library blanks were carried along in the experiments, and were also shotgun sequenced.

Data pre-processing

Raw sequencing reads were de-indexed using Illumina's Bcl2fastQ (Illumina). The data from the Moneen individual and the blanks were subsequently processed using a subset of bioinformatics tools integrated in the EAGER pipeline (version 1.92.7) (Peltzer *et al*. 2016). Adapter clipping was done using Clip&Merge, discarding all reads shorter than 30bp (Peltzer *et al*. 2016). The reads were subsequently mapped to the human reference genome (hg19/GRC37) using the Burrows-Wheeler Aligner (BWA) version 0.7.12 (Li and Durbin 2009) to estimate the amount of human endogenous DNA in the library. BWA mapping parameters were customised (seeding turned off, -l 1000; mapping stringency, –n 0.01; quality filter, –q 30) to accommodate the characteristic deaminated bases that occur towards the ends of ancient DNA fragments. Duplicate removal was done using MarkDuplicates from the Picard toolkit (http://broadinstitute.github.io/picard/). In order to separately analyse mitochondrial reads and to determine the individual's haplogroup, clipped reads were mapped separately to the human mitochondrial genome reference (rCRS) (Andrews *et al*. 1999). The same method and parameters as outlined above were applied, with the exception that Circularmapper (Peltzer *et al*. 2016) was used in place of BWA, in order to retain

the genetic information at the ends of the reference that would otherwise be lost in linear mapping.

Given the number of mapping reads using the above approach, it is estimated that the library has approximately 8% human endogenous DNA, and ~0.027% endogenous mtDNA when mapped only to the human mitochondrial genome. Human mitochondria exist in high copy number and have very small genomes compared to the nuclear genome. Even though there is only 0.027% endogenous mtDNA in the sequenced reads, this yields a mean mitochondrial coverage of 14.5-fold, with 99.11% of the reference covered at least 5-fold (see Figure 80). The blanks are negative for both modern and ancient human DNA, indicating that the reagents and work environment used during extraction and library preparation did not contribute a significant amount of human contaminant DNA to interfere with downstream bioinformatics analyses.

Reference	Total raw reads	Mapped reads after duplicate removal	Endogenous DNA (%)	Mean coverage (fold; X)	Average fragment length (bp)	Reads with damage on 1st base 5-prime (%)	Reads with damage on 2nd base 5-prime (%)	Initial contamination estimate (%) (low, high)	Final contamination estimate (%) (low, high)
HG19 (complete human genome)	21667649	1543204	8.213	0.0255	51.12	0.1086	0.0648	n.a.	n.a.
rCRS	21667649	4223	0.027	14.5	56.83	0.1314	0.0935	0 (0, 0.5)	1(0, 2)

Figure 80 Mapping statistics (Peltzer *et al.* 2016), mapDamage (Ginolhac *et al.* 2011) deamination values and Schmutzi (Renaud *et al.* 2015) contamination values for the mtDNA.

Deamination pattern and ancient DNA authenticity

Ancient DNA degrades over time into increasingly shorter fragments. The terminal ends of the resulting fragments are vulnerable to chemical damage, and deamination of cytosines accumulates at the 5-prime end of the molecules. In the sequencing data such damage is evidenced by an accumulation of C to T transitions at the ends of molecules, which can in turn be used to authenticate molecules as ancient as opposed to modern contamination (Sawyer *et al*. 2012).

The bioinformatic tool mapDamage (Ginolhac *et al*. 2011) was used to calculate the deamination pattern of the mapping reads to both hg19 and the human mitochondrial reference. As single-end data were produced, only the deamination pattern reported for the 5-prime ends of the reads can be reliably estimated. Deamination rates for the 5-prime ends of the mapped reads were 10.8% and 13% for reads mapped to the human

reference genome and the mitochondrial reference sequence, respectively (see Figures 80 and 81). Radiocarbon dates indicate that the Moneen individual lived during the 1500-1600s (Dowd 2013a). The deamination rates are relatively low, indicating good DNA preservation. A contributing factor may be the lower temperatures in Ireland, as heat is known to increase the degradation of ancient DNA (Sawyer *et al*. 2012), and also the fact that cave environments are usually quite stable.

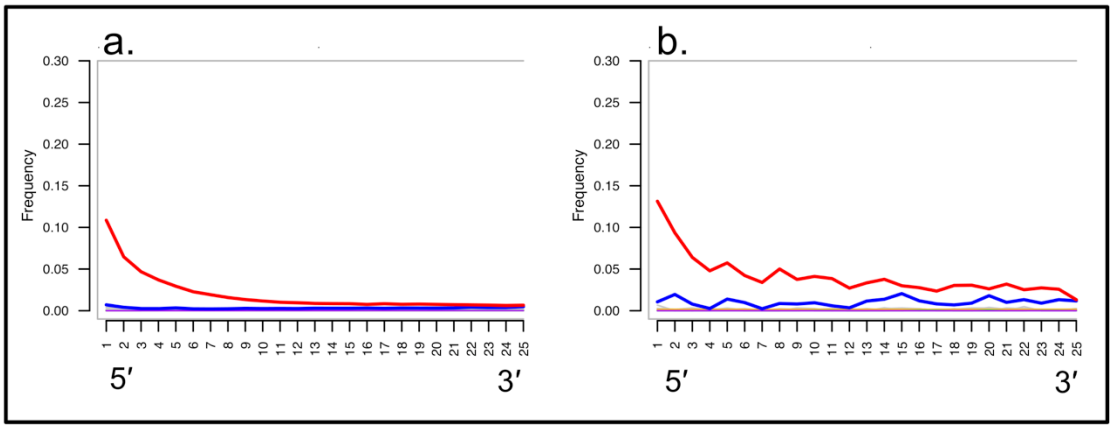


Figure 81 mapDamage (Ginolhac *et al*. 2011) curves depicting the deamination pattern of reads produced from sequencing the Moneen individual, mapped to a) the human genome, b) the rCRS (mtDNA). The curves depict the frequency of deaminated bases occurring towards the terminal 5-prime ends of mapped reads.

Contamination estimate and consensus calling

Schmutzi (Renaud *et al*. 2015) is a tool that uses an iterative approach to estimate the degree of mitochondrial contamination within a sample. Putative contaminant reads are identified and removed, and consensus sequence of the endogenous mitochondrial genome is produced. Reads mapping to the mitochondrial genome were predicted to have a contamination estimate of 0% after the first iteration. The final iteration gave a contamination estimate of around 1% after removal of mapping reads that deviate from the expected length, deamination pattern, and base calls (Renaud *et al*. 2015). This is a low level of contamination. A database of present-day Eurasian mitochondrial genomes accompanying the Schmutzi software was used as a reference for identifying contaminant sequences.

The reads remaining after the final iteration of Schmutzi were used to generate a consensus sequence based on the endogenous DNA of the Moneen individual. A quality

filter of -q 20 was applied to the final endogenous consensus sequence output from Schmutzi, resulting in 34 out of 16,569 positions not being sufficiently covered to facilitate a consensus call. The quality-filtered consensus was used to determine the Moneen individual's mitochondrial haplogroup.

Haplogroup assignment

Using Haplogrep 2.0 (van Oven and Kayser 2009; Weissensteiner *et al*. 2016) it was determined, based on the quality filtered Schmutzi consensus sequence that the mitochondrial genome of the Moneen individual belongs to haplogroup J2b1b1. The following positions in the mitochondrial genome of the Moneen individual contribute to its haplogroup assignment: 73G, 263G, 489C, 750G, 1438G, 2404C, 2706G, 4216C, 4769G, 5633T, 6962T, 7028T, 7211A, 7476T, 8860G, 10172A, 10389C 10398G, 11251G, 11719A, 12612G, 13708A, 14766T, 15257A, 15326G, 15452A 15812A, 16069T, 16126C and 16193T. This individual also had one hotspot mutation located in the control region of the mtDNA: 16519C.

Sex determination

The reads mapped to the human reference genome were used to determine the sex of the Moneen individual using an established method (Skoglund *et al*. 2013), which compares the ratio of reads mapped to the X chromosome versus the Y chromosome. The result indicates that the Moneen individual is male with a 95% confidence interval for R_Y of 0.0844-0.0896, thus confirming the result of previous ancient DNA analyses (Section 16).

Metagenomic analysis using MALT

MALT (**M**egan **AL**ignment **T**ool) (Herbig *et al*. 2016) is a rapid sequence alignment tool based on a binning algorithm that uses a reference database – in this case one consisting of all publicly available complete bacterial genomes and plasmids (NCBI RefSeq, December 2015) and eukaryotic organelles (NCBI RefSeq Organelle Genome Resources, February 2016) – to assign sequence reads to the taxa/genus/species where they align best. MALT is specifically designed to produce an output compatible with

visualisation in the metagenomic analysis software MEGAN (Huson *et al*. 2007; Herbig *et al*. 2016). MALT (version 0.1.2) was run twice using 'SemiGlobal' alignment, with 100 set as the maximum number of alignments for each read and allowing 64 threads/CPU cores. The following parameters were set for the taxonomic assignment: *top percent* was set at 1.0 and *min support percent* at 0.01. The minimal percent identity of aligned reads included in the output of the analysis was set to 85% and 95% respectively for each run. In both cases ancient oral and environmental bacteria dominated the microbial content of the sample.

No traces of DNA belonging to pathogens that cause systemic disease were identified. Several species of bacteria belonging to the oral microbiome were identified from the MALT results. In the MALT output run with the minimal percent identity score set to 95%, the five most numerous bacterial species were: *Capnocytophaga ochracea*, *Streptococcus gordonii*, *Tannerella forsythia*, *Rothia dentocariosa* and *Streptococcus sanguinis*. These bacteria are commonly known to be part of the human oral microbiome (Chen *et al*. 2010).

Discussion and concluding remarks

Mitochondrial haplogroup J is found to be present in Europe and the Near East (Logan 2008; Pierron *et al*. 2011). In Ireland alone, based on modern data, haplogroup J occurs at a frequency of around 10.7% (mtdna.eu). Haplogroup J has two major sub-haplogroups, J1 and J2, where J1 is more commonly represented in Europe and the Near East than J2 (Pierron *et al*. 2011). There is limited information about the specific sub-haplogroup of J2b1b1 that the Moneen individual belongs to. At least four individuals belonging to mitochondrial haplogroup J2b1b have been entered into the Family Tree DNA database (https://www.familytreedna.com/public/J-mtDNA/default.aspx?section=mtmap), one of which is from Northern Ireland, two from Scotland and one from England. In all, the haplogroup assignment of J2b1b1 indicates that the Moneen boy has European maternal ancestry. Further genetic analysis concentrating on autosomal loci would be required to gain a more detailed picture of this individual's genetic ancestry.

Individuals with Leber's hereditary optic neuropathy (LHON), a maternally inherited genetic disease that causes blindness and deafness, has been linked to individuals belonging to haplogroup J, specifically J1c and J2b (Achilli *et al*. 2012). There are three major point mutations that are highly associated with this disease that occur in the mitochondrial genome (Achilli *et al*. 2012). The mtDNA of the Moneen individual was checked for the presence of these LHON associated point mutations, and none were present.

The location and preservation of ancient pathogen DNA in skeletal remains is highly variable depending on the pathogen in question (Bos *et al*. 2014; Kay *et al*. 2014; Schuenemann *et al*. 2013). As the skeletal remains of this individual did not display any changes consistent with known diseases, the best chance of recovering ancient pathogen DNA was the pulp chamber of the tooth in the hope of finding blood borne pathogens whose DNA may have been preferentially protected by the hard enamel coating. Based on the DNA that was extracted, we found no trace of bacterial or eukaryotic pathogens that could have caused a lethal infectious disease. The MALT database, however, does not include viruses. Although responsible for many diseases, a large number of viruses are made up of molecularly unstable RNA, and are thus highly unlikely to be preserved in skeletal remains found in the archaeological record. The viruses that cause smallpox and chickenpox, however, have genetic material made of double stranded DNA. The presence of DNA from either of these viruses was investigated in the sequencing data using the mapping approach outlined above. Zero reads mapped to either virus.

We were not able to detect any pathogens that could be responsible for the death of this boy. However, the preservation of ancient pathogen DNA is not guaranteed, and therefore our negative findings do not wholly exclude the scenario that this individual may have died due to an infection.

Acknowledgements

The authors would like to thank Cosimo Posth, Alissa Mittnik, Alexander Peltzer and Wolfgang Haak for fruitful discussions and advice regarding data analysis.

References

Achilli, A., Iommarini, L., Olivieri, A., Pala, M., Hooshiar Kashani, B., Reynier, P., La Morgia, C., Valentino, M. L., Liguori, R., Pizza, F., Barboni, P., Sadun, F., De Negri, A. M., Zeviani, M., Dollfus, H., Moulignier, A., Ducos, G., Orssaud, C., Bonneau, D., Procaccio, V., Leo-Kottler, B., Fauser, S., Wissinger, B., Amati-Bonneau, P., Torroni, A., and Carelli, V. (2012), 'Rare primary mitochondrial DNA mutations and probable synergistic variants in Leber's hereditary optic neuropathy', *PLoS One,* 7 (8), e42242.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999), 'Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA', *Nat Genet,* 23 (2), 147.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J., Campbell, T. J., Majander, K., Wilbur, A. K., Guichon, R. A., Wolfe Steadman, D. L., Cook, D. C., Niemann, S., Behr, M. A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J. E., Gagneux, S., Stone, A. C., and Krause, J. (2014), 'Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis', *Nature,* 514 (7523), 494-7.

Chen, T., Yu, W. H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010), 'The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information', *Database (Oxford),* 2010, baq013.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., Valdiosera, C., Garcia, N., Paabo, S., Arsuaga, J. L., and Meyer, M. (2013), 'Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments', *Proc Natl Acad Sci U S A,* 110 (39), 15758-63.

Dowd, Marion (2013), 'About a boy: excavations at Moneen Cave', *Archaeology Ireland,* 27 (1), 9-12.

Garattini, Chiara (2007), 'Creating memories: material culture and infantile death in contemporary Ireland', *Mortality,* 12 (2), 193-206.

Ginolhac, A., Rasmussen, M., Gilbert, M. T., Willerslev, E., and Orlando, L. (2011), 'mapDamage: testing for damage patterns in ancient DNA sequences', *Bioinformatics,* 27 (15), 2153-5.

Herbig, Alexander, Maixner, Frank, Bos, Kirsten I., Zink, Albert, Krause, Johannes, and Huson, Daniel H. (2016), 'MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman', *bioRxiv*

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007), 'MEGAN analysis of metagenomic data', *Genome Res,* 17 (3), 377-86.

Kay, G. L., Sergeant, M. J., Giuffra, V., Bandiera, P., Milanese, M., Bramanti, B., Bianucci, R., and Pallen, M. J. (2014), 'Recovery of a medieval Brucella melitensis genome using shotgun metagenomics', *MBio,* 5 (4), e01337-14.

Kircher, M., Sawyer, S., and Meyer, M. (2012), 'Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform', *Nucleic Acids Res,* 40 (1), e3.

Li, H. and Durbin, R. (2009), 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics,* 25 (14), 1754-60.

Logan, Jim (2008), 'A Comprehensive Analysis of mtDNA Haplogroup J', *Journal ofGenetic Genealogy,* 4 (2), 104-24.

Meyer, M. and Kircher, M. (2010), 'Illumina sequencing library preparation for highly multiplexed target capture and sequencing', *Cold Spring Harb Protoc,* 2010 (6), pdb prot5448.

Murphy, Eileen M. (2011), 'Children's Burial Grounds in Ireland (Cilliní) and Parental Emotions Toward Infant Death', *International Journal of Historical Archaeology,* 15 (3), 409-28.

Peltzer, A., Jager, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016), 'EAGER: efficient ancient genome reconstruction', *Genome Biol,* 17 (1), 60.

Pierron, D., Chang, I., Arachiche, A., Heiske, M., Thomas, O., Borlin, M., Pennarun, E., Murail, P., Thoraval, D., Rocher, C., and Letellier, T. (2011), 'Mutation rate switch inside Eurasian mitochondrial haplogroups: impact of selection and consequences for dating settlement in Europe', *PLoS One,* 6 (6), e21543.

Renaud, G., Slon, V., Duggan, A. T., and Kelso, J. (2015), 'Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA', *Genome Biol,* 16, 224.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Paabo, S. (2012), 'Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA', *PLoS One,* 7 (3), e34131.

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jager, G., Bos, K. I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J. L., Kjellstrom, A., Wu, H., Stewart, G. R., Taylor, G. M., Bauer, P., Lee, O. Y., Wu, H. H., Minnikin, D. E., Besra, G. S., Tucker, K., Roffey, S., Sow, S. O., Cole, S. T., Nieselt, K., and Krause, J. (2013), 'Genome-wide comparison of medieval and modern Mycobacterium leprae', *Science,* 341 (6142), 179-83.

Skoglund, P., Stora, J., Gotherstrom, A., and Jakobsson, M. (2013), 'Accurate sex identification of ancient human remains using DNA shotgun sequencing', *Journal of Archaeological Science,* 40 (12), 4477-82.

van Oven, M. and Kayser, M. (2009), 'Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation', *Hum Mutat,* 30 (2), E386-94.

Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H. J., Kronenberg, F., Salas, A., and Schonherr, S. (2016), 'HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing', *Nucleic Acids Res*.