Petabase-scale Data Mining Identifies Novel Clostridial Species and Neurotoxins Associated with Ancient Human DNA

by

Harold Paul Hodgins

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2022

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see the Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The work presented in this thesis has been submitted for publication and deposited as a pre-print (see below):

> Ancient *Clostridium* DNA and variants of tetanus neurotoxins associated with human archaeological remains. Harold P. Hodgins, Pengsheng Chen, Briallen Lobb, Benjamin JM Tremblay, Michael J. Mansfield, Victoria CY Lee, Pyung-Gang Lee, Jeffrey Coffin, Xin Wei, Ana T. Duggan, Alexis E. Dolphin, Gabriel Renaud, Min Dong, Andrew C. Doxey. *bioRxiv* 2022.06.30.498301; doi: https://doi.org/10.1101/2022.06.30.498301

I would like to thank the following individuals for their collaboration and contributions to this work:

- Dr. Michael Mansfield, who performed the variant calling pipeline for the tetanus neurotoxin gene sequences recovered from ancient DNA samples.
- Dr. Briallen Lobb, who performed the community analysis, phylogenetic analysis, average nucleotide analysis, and CheckM analysis of *C. tetani* related metagenome-assembled genomes recovered from ancient DNA samples.
- Dr. Pengsheng Chen and Dr. Pyung-Gang Lee, who performed the experimental testing of the TeNT/Chinchorro toxin, in the lab of Dr. Min Dong (Boston Children's Hospital, Harvard Medical School).
- Benjamin Tremblay, who developed scripts for the genome coverage and alignment visualization
- Dr. Alexis Dolphin and Jeffrey Coffin, who assisted with the analysis of archaeological samples and their historical and geographical contexts

Abstract

Analyzing microbial genomes found in archaeological samples can provide insights into the origins of modern infectious diseases. A large-scale metagenomic analysis of archeological samples discovered bacterial species related to modern-day *Clostridium tetani* (which produces the tetanus neurotoxin (TeNT) and causes the disease tetanus). Draft genomes were assembled from 38 distinct human archeological samples (which came from five continents with the oldest sample estimated to be ~6000 years old) and which displayed hallmarks of ancient DNA damage to varying degrees. A phylogenetic analysis of the draft genomes found several which fall into existing *C. tetani* clades, several potentially novel *C. tetani* lineages, and a potentially novel *Clostridium* species related to modern *C. tetani*. Fifteen TeNT variants were found, including a unique variant found exclusively in ancient samples from South America. A TeNT variant associated with a ~6,000-year-old Chilean mummy sample was experimentally tested and was found to induce tetanus like muscle paralysis in mice with a potency similar to modern TeNT. This work provides the first identification of neurotoxigenic *C. tetani* in ancient DNA, the discovery of a potentially new *Clostridium* species, and the discovery of a tetanus like neurotoxin which is functionally active and able to cause disease in mice.

Acknowledgements

I would like to thank my supervisor, Dr. Andrew Doxey, for supporting my research and for giving me the opportunity to work on such an interesting project. I also wish to thank my committee members, Dr. Trevor Charles and Dr. Brendan McConkey and the many collaborators that contributed to the research presented in this thesis.

I gratefully acknowledge funding and resources provided by NSERC, the University of Waterloo and the Digital Research Alliance of Canada (formerly Compute Canada).

Lastly, I wish to thank past and present member of the Doxey lab for their stimulating conversations and my parents for instilling in me a love of science.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
aDNA	ancient DNA
ANI	average nucleotide identity
BoNT	botulinum neurotoxin
bp	base pairs
DNA	deoxyribonucleic acid
Endo VIII	endonuclease VIII
Kb	kilo base
kDa	kilo dalton
LD_{50}	lethal dose 50%
Mb	mega base
mtDNA	mitochondrial DNA
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
PCR	polymerase chain reaction
SNARE	soluble N-ethylmaleimide-sensitive factor attachment protein receptor
SRA	Sequence Read Archive
STAT	SRA taxonomy analysis tool
TB	tuberculosis
TeNT	tetanus neurotoxin
UDG	uracil-DNA-glycosylase
UNG	uracil N-glycosylase

Chapter 1
Literature review: Clostridial neurotoxins and ancient DNA

Introduction

Stepping on a rusty nail in North America often results in a trip to the hospital where the patient will be given a booster dose of the tetanus vaccine and/or a preventative dose of tetanus anti-serum depending on their vaccination history. This is because the bacterium *Clostridium tetani* (which is ubiquitous in soil) creates spores which germinate in oxygen deprived tissues, such as those created by deep puncture wounds. Once established in a wound, the newly revived bacteria begin to produce one of the most potent neurotoxins currently known (the tetanus neurotoxin), which migrates to the central nervous system and begins to block inhibitory nervous signals, resulting in spastic paralysis. Without proper medical intervention the patient may stop breathing due to paralysis of the respiratory system.

Although tetanus has plagued humans for thousands of years, nothing is known about its early history. Was the tetanus toxin more or less potent than modern versions? Did it bind to other targets? Using DNA recovered from archeological samples this research begins to answer these questions by analyzing several *Clostridium tetani* draft genomes and by comparing a potentially ancient neurotoxin with modern tetanus neurotoxin.

1.1.1 *Clostridium tetani* and toxigenic clostridia

The genus *Clostridium* is comprised of anaerobic, spore-forming, gram-positive, rod-shaped bacteria which can be found in wide range of environments including soil, marine sediments, and human gastrointestinal tracts (Zaragoza *et al.* 2019). While most clostridial species (currently estimated to be approximately 311 (Parte *et al.* 2022), with 98 genomes in the NCBI database) are thought to be saprophytic decomposers, at least twenty are pathogenic and produce toxins which infect humans or animals (Hatheway 1990; Carter *et al.* 2014). The toxins produced by clostridial species affect the gastrointestinal tract, soft-tissues, organs, and neurons and cause damage ranging from mild to fatal (Carter *et al.* 2014). Of these, the neurotoxins produced by *Clostridium botulinum* and *Clostridium*

tetani are currently the most potent toxins known, as measured by their LD_{50} in mice (Rossetto and Montecucco 2019).

Botulinum neurotoxin (BoNT) and tetanus neurotoxin (TeNT) are produced by *C. botulinum* and *C. tetani* respectively and are responsible for the diseases known as botulism and tetanus in humans and animals. These neurotoxins have a similar structure and initial mode of action but differ in several important ways which fundamentally change their associated diseases.

Evidence of tetanus like symptoms go back to the time of Hippocrates in 4th century BC (Pappas *et al.* 2008). For many years tetanus was considered to be an untreatable syndrome until it was demonstrated to be caused by an infectious agent by Carle and Rattone in 1884. In 1889 Kitasato demonstrated that tetanus was caused by *C. tetani* with the TeNT being specifically implicated the following year by Tizzoni and Cattani. In contrast the earliest historical records of botulism like symptoms are from eighteenth century Württemberg in Southwestern Germany where there was an increase in deaths following the consumption of improperly cooked blood sausages (Zhang *et al.* 2010).

Although *C. tetani* can be found in the gastrointestinal tract of humans (Cook *et al.* 2001) and is ubiquitous in soil (Popoff 2020), tetanus infections only occur after its spores germinate in oxygen-depleted and necrotic tissue (Popoff 2020). In contrast botulism in humans is predominately caused by the ingestion of pre-formed toxin from improperly cooked or preserved food (food-borne botulism) or by *C. botulinum* bacteria colonizing the gut which then produce the toxin in situ (infant and adult intestinal botulism) (Sobel 2005).

A tetanus toxoid vaccine was developed by Ramon and collaborators in the early 1920s (Smith 1969) with the same general formulation still being used today. Recent work developing a recombinant vaccine has shown that the immunogenicity of the tetanus toxin C-fragment is similar to that of the full native toxin (Yu *et al.* 2018).

1.1.2 Tetanus neurotoxin

Bacterial toxins can be classified as exotoxins or endotoxins depending on how they are released into their local environment. Exotoxins are secreted out of the bacteria (or released during cell lysis) whereas endotoxins are embedded in the bacterial cell wall and only released during cell lysis (Cai *et*

al. 2021). While exotoxins are selective and generally only bind to specific cell types and/or receptors, endotoxins can directly trigger a general host response (Peterson 1996). Bacterial toxins are further separated into three types based on their interactions with host cells. Type I toxins interact with the host without entering the host cells. Type II toxins create pores in host cell membranes and then enter host cells. Type III toxins (also known as A-B toxins) are composed of two components and often target tissues far away from site of the original bacterial infection. The B component facilitates binding and cell entry which is followed by the A component enzymatically damaging the cell (Cai *et al.* 2021).

Seven main serotypes of BoNT, labeled A-G, have been previously categorized (Popoff and Bouvet 2009; Smith *et al.* 2015; Dong *et al.* 2019; Kumar *et al.* 2019) More recently, BoNT/H (a hydrid of BoNT/A and BoNT/F) was discovered by (Barash and Arnon 2014) with BoNT/X, an entirely new BoNT, being discovered by Zhang, Berntsson, *et al.* in 2017. In addition, several BoNT-related toxins have been identified outside of the genus *Clostridium* including "BoNT/Wo" in *Weissella oryzae* (Mansfield *et al.* 2015) *and* "BoNT/En" in *Enterococcus faecium* (Zhang *et al.* 2018). In contrast there is only one known serotype of the tetanus toxin (Dong *et al.* 2019).

The BoNT and TeNT are synthesized as a single-chain (~150kDa) precursor protein which has weak or no activity (Singh 2006). The precursor protein is released via cell wall exfoliation (Call *et al.* 1995) and is then proteolytically activated by clostridial or host proteases resulting in a di-chain form which is linked via an interchain disulfide bond. The heavy chain (~100 kDa) binds to nerve cells at the presynaptic nerve terminal and helps transfer the light chain (~ 50 kDa) into the cells using receptor-mediated endocytosis. Once inside, the light chain functions as an endopeptidase and targets specific soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins and inhibits neurotransmitter release (Cai *et al.* 2021). BoNTs are co-produced with several nontoxic-associated proteins which are thought to protect the toxin in the gut and enable absorption (Sakaguchi 1982). TeNT does not have any associated proteins and thus is not toxic when ingested or produced in the gut (Cai *et al.* 2021).

While both the BoNT and TeNT cause neurological damage, BoNTs cause flaccid paralysis and TeNTs cause spastic paralysis (Singh 2006; Popoff and Bouvet 2009). This is due to differences in toxin translocation following the initial neuron infection. Whereas BoNTs target motor neuron termini, TeNTs enter motor and sensory neurons and then move retrogradely along the axis ultimately

targeting the central inhibitory neurons (Cai *et al.* 2021). Without treatment both toxins can cause respiratory failure and death (Cai *et al.* 2021).

1.1.3 *Clostridium tetani* genome

An early isolate of *C. tetani* (the 1920 Harvard E88 strain) is still widely used as a reference today. Bruggemann et al. (2003) sequenced this strain and revealed a genome consisting of a single ~2.8 Mb chromosome and a ~74 Kb plasmid. This genomic organization is maintained among all known strains of *C. tetani* (Brüggemann *et al.* 2015; Cohen *et al.* 2017; Chapeton-Montes *et al.* 2019). The plasmid is critical to pathogenicity as it encodes the key virulence genes including the *tent* gene which encodes the neurotoxin and the *colT* gene which encodes a collagenase enzyme involved in tissue degradation (Bruggemann *et al.* 2003). To date, no pathogenic strain of *C. tetani* has been found that lacks the *tent* gene encoded on the plasmid.

Forty-three strains of *C. tetani* have been sequenced to date with strains being classified as Harvard strains (from the original isolate) or wild type (from clinical cases) (Garrigues *et al.* 2022). Based on a comparative genomic analysis, modern *C. tetani* strains cluster into two phylogenetically distinct clades (Chapeton-Montes *et al.* 2019), but are closely related and exhibit low genetic variation with average nucleotide identities of 96-99%. Similarly, the *tent* gene is extremely conserved and exhibits 99% to 100% amino acid identity across all strains. Modern *C. tetani* genomes therefore offer a limited perspective on the full diversity of *C. tetani* and its evolutionary history as a human disease-causing bacterium.

Ancient DNA

In a living cell, DNA is constantly being damaged and subsequently enzymatically repaired. Following cell death, the cell's endogenous DNA begins to degrade via enzymatic and chemical processes, while at the same time exogenous DNA from the surrounding environment begins to infiltrate. Eventually the DNA present becomes too short or chemically damaged to be sequenced and the information it held is lost. However, under favourable conditions (such as frozen or rapidly desiccated tissues) DNA can remain sequenceable for thousands (but not millions) of years (Dabney *et al.* 2013b). DNA has been successfully extracted from a wide range of substrates ranging from

bones to animal skin parchments and has been used to study various topics including human genetics, ancient environments and ancient pathogens (Orlando *et al.* 2021). Although there is no exact definition of what is ancient, microbiological samples older than 100 years are generally considered to be the domain of paleomicrobiology with DNA extracted from ancient sources being called ancient DNA (aDNA) (Drancourt 2016).

In 1984, Higuchi *et al*. successfully extracted a small fragment of DNA from a museum quagga specimen (an extinct relative of the modern Zebra) and amplified it using bacterial cloning techniques. This was the first evidence that DNA was more stable than previously assumed and offered scientists another tool for studying the past beyond the fossil record. The following year Pääbo recovered human DNA from an ancient Egyptian mummy and the hunt began for the world's oldest DNA.

The number of studies amplifying DNA from ever older sources began to rapidly increase with the advent of PCR in 1987. However, early studies were plagued by the question of authenticity (i.e. was the DNA present really ancient or just modern contamination) and guidelines for PCR studies of aDNA were introduced (Cooper and Poinar 2000). The arrival of next generation sequencing (NGS) in the early 2000's finally provided a way to authenticate DNA from ancient samples by checking the raw reads for age associated signatures of damage found in aDNA (Margulies *et al.* 2005; Arning and Wilson 2020).

Although DNA has been extracted from several samples with ages estimated to be greater than a million years, including remains found in amber by DeSalle *et al.* (1992) and in plant fossils by Golenberg *et al.* (1990), Pääbo and Wilson (1991), Lindahl (1993), and Pääbo *et al.* (2004) argue that these are more likely to be modern DNA contamination. While early studies often focused on obtaining the oldest DNA possible, more recent studies have shifted the focus towards reliably extracting DNA for a diverse set of sample types (Jones and Bösl 2021).

1.1.4 Sources of ancient DNA

Environmental conditions play a critical role in the amount and quality of DNA that can be successfully extracted from any sample, with DNA being more successfully extracted from frozen or rapidly desiccated samples (Campos *et al.* 2012; Dabney *et al.* 2013b). Frozen human tissues are an

excellent source of DNA as freezing rapidly preserves the tissue and immediately slows the degradation, but they are rare and typically only discovered by accident (Aboudharam 2016).

Although human DNA was successfully extracted from bones by Hagelberg *et al.* in 1989, skepticism was expressed by early pioneers in the field, including Svante Pääbo who stated that 'Of course you can't get DNA from bone!' (because of trouble with contamination and authentication (Jones and Bösl 2021)) at the Biomolecular Palaeontology Community Meeting the following year. Although extracting DNA from bones is destructive and often unsuccessful (Arning and Wilson 2020), bones have become one of the main sources of aDNA, as they are readily available and often have a known provenance (Aboudharam 2016). Some chronic diseases (such as leprosy caused by *Mycobacterium leprae* (Schuenemann *et al.* 2013; Mendum *et al.* 2014) and tuberculosis caused by *Mycobacterium tuberculosis* (Müller *et al.* 2014)) leave skeletal lesions, which although insufficient for a retrospective disease diagnosis can be used to determine which bones should be to choose for further analysis. In general, only a small amount DNA is successfully extracted from skeletons found in graves, with any DNA extracted often being highly damaged (likely due their contact with soil). In contrast, the DNA extracted from skeletons found in vaults is generally more abundant and less damaged (Kay *et al.* 2015; Aboudharam 2016).

Dental calculus is a calcified biofilm which forms on teeth and which can be used to investigate changes in the oral microbiome through time as it contains whole bacteria (including oral and respiratory pathogens) along with their DNA (Warinner *et al.* 2014; Aboudharam 2016; Davenport *et al.* 2017). The rate of formation is influenced by oral hygiene and exposure to things which influence the rate of saliva production such as smoking/radiation/drugs. A study by Velsko *et al.* (2019) compared modern dental plaque and modern dental calculus and found a systematic bias in the microbial composition during calcification (Arning and Wilson 2020). aDNA studies using dental calculus are thus limited to human oral microbiome species (although theoretically they could include respiratory organisms) and likely present a skewed representation of past microbial diversity.

In 1998 and 2000, Drancourt *et al.* and Raoult *et al*. successfully extracted and amplified *Yersinia pestis* DNA from the dental pulp of plague victims. However, in 2004 Gilbert *et al.* were unable to replicate these findings with an alternative set of samples. This was challenged by Drancourt and Raoult (2004) who note that Gilbert *et al.* (2004) did not properly replicate the original methods.

Since then, dental pulp has become the preferred source of DNA for detecting several ancient pathogens, with all *Y. pestis* paleogenomes to date coming from ancient teeth. The dental pulp is highly vascularized meaning that pathogens present in the bloodstream at the time of death are often found in the dental pulp (Drancourt *et al.* 1998; Aboudharam 2016). Teeth are also relatively abundant, survive longer than bones, and have a core which is better protected from external contamination (Drancourt *et al.* 1998; Bos *et al.* 2011, 2019; Aboudharam 2016; Vågene *et al.* 2018; Mühlemann *et al.* 2018). This is thought to be because teeth are 70-75% dry weight mineral compared to 6% for bones (Kirsanow and Burger 2012). Also, dental pulp is easier to extract DNA from as there is no need to demineralize the sample and the taxonomic composition is not as skewed as that of dental calculus (Aboudharam 2016; Arning and Wilson 2020). X-rays can be used to help choose teeth for further analysis, with single rooted teeth which have a large pulp cavity (incisors, canines, premolars) from young adults or adolescents considered to be the best. However, if the apical end is open it may have been exposed to soil and become contaminated (Aboudharam 2016).

Other calcified structures from chronic infections (e.g. calcified plurea) are another good source of ancient pathogen DNA as they are very resistant to environmental contamination with levels of preservation approaching that of mineralized dental calculus (Donoghue *et al.* 1998; Kay *et al.* 2014, 2015; Devault *et al.* 2017; Mann *et al.* 2018).

Coprolites (fossilized feaces) are a good source of aDNA for human microbiome studies as they provide a snapshot (although likely skewed) of the intestinal contents of the host. They can be found in dry, cold, or tropical environments but are best preserved in extremely dry or frozen environment. However, they can be hard to distinguish from rocks (Cano *et al.* 2000, 2014; Poinar *et al.* 2003; Santiago-Rodriguez *et al.* 2013).

For ancient environmental bacterial DNA, Arning and Wilson (2020) consider ice cores to be the only acceptable source. However, Linderholm (2021) describes several papers which successfully retrieved whole genomes (although not bacterial) from soil. There is some debate regarding the migration of DNA in soil. Andersen *et al.* (2012) state that it is stable, however Haile *et al.* (2007) found sheep DNA in a layer of soil in New Zealand estimated to be older then when sheep were first introduced to New Zealand.

1.1.5 Ancient DNA damage

DNA damage determines if the DNA can be sequenced and how the resulting reads are interpreted. DNA damage can also be used to "authenticate" aDNA as certain patterns of damage can be used to distinguish damaged DNA from modern or undamaged DNA. As described below, there are three types of damage commonly found with aDNA. 1) Shorter fragment lengths, 2) damage which changes what nucleotides are incorporated during replication, and 3) damage which blocks replication (Dabney *et al.* 2013b).

1.1.5.1 Ancient DNA fragment sizes

When Pääbo (1989) extracted DNA from a variety of samples (age 4 to 13,000 years old) they found the DNA fragment size varied between 40 and 500 base pairs (bp). This is considered to be a common feature of aDNA, with aDNA generally having fragments shorter than 100 bp (Duchêne *et al.* 2020). Additionally, the number of reads as a function of length drops rapidly with increasing length (Warinner *et al.* 2017).

The steady decrease in fragment size is thought to be due to hydrolytic depurination which results in an abasic site which is then followed by β-elimination resulting in single strand breaks (Lindahl 1993). Supporting evidence for this was found in invitro experiments done by Lindahl and Andersson (1972) and Lindahl and Nyberg (1972) which used radioactively labeled purine residues.

High throughput sequencing has provided additional evidence for this mechanism of DNA fragmentation. Using reference genomes, Briggs *et al.* (2007) found that purines (adenine and guanine) were more likely to be found adjacent to strand breaks than pyrimidines (thymine and cytosine) at the 5′ ends of DNA fragments from Neanderthal, mammoth, and cave bear remains (~ 40,000 years old). However, because of the sequencing protocol that was used, they were unable to determine if the pattern also occurred at the 3′ end of the DNA fragments.

Overballe-Petersen *et al.* (2012) tried adding a poly A tail to the 3′ end of permafrost Pleistocene horse DNA. They did not find a preference for strand breaks at any particular base at the 3′ end but noted that this was likely due to inefficient ligation of the poly A tail to aldehydic 3′ ends which are a predicted byproduct of β-elimination. Dabney and Meyer (2012) used a protocol which preserves the original 3′ and 5′ ends and found that purines and especially guanines were

overrepresented at both ends of the DNA molecules in DNA extracted from bones which were tens of thousands of years old.

1.1.5.2 Ancient DNA nucleotide misincorporations

Nucleotide bases can be hydrolytically deaminated resulting in them being misread by polymerases. This often occurs with cytosine bases turning them into uracil bases. If this is not enzymatically repaired, the uracils appear in the final sequence as a C → T transition on the forward strand or a G → A transition on the complementary strand.

Pääbo (1989) inferred that aDNA contains uracil bases because it is sensitive to uracil-N-glycosylase (UNG) treatment (also known as UDG treatment). Additionally, Hofreiter *et al.* (2001) PCR amplified aDNA and found that the majority of substitutions in ancient DNA are C → T and that this dramatically decreased following UNG treatment.

Using high throughput sequencing Stiller *et al.* (2006) and Gilbert *et al.* (2007) found that C → T substitutions are the most common nucleotide misincorporations found in ancient DNA. Subsequently, Briggs *et al.* (2007) and Brotherton *et al.* (2007) found that these substitutions are concentrated at the ends of the molecules, where up to 40% of cytosines appear as thymines. This is followed by an exponential decrease along the molecule. A newer protocol which does not remove 3′ overhangs was used by Meyer *et al.* (2012) to confirm that C → T substitutions occur at both ends of ancient DNA molecules at an elevated rate.

Because of their concentration at the ends of molecules, it is thought that C → T substitutions are due to single stranded overhangs resulting from β-eliminations (Briggs *et al.* 2007). This is because the rate of cytosine deamination is estimated to be 2 times higher in single stranded DNA than in double stranded DNA (Lindahl 1993) and Dabney *et al.* (2013) found that uracils rarely occur in the middle of aDNA molecules.

1.1.5.3 Replication blocking modifications to DNA

DNA polymerases can be hindered or stopped by modified DNA molecules. These include modified bases and the DNA molecule being bound to itself, other DNA molecules, or other macro molecules

such as proteins (Dabney *et al.* 2013b). Hoss *et al.* (1996) analyzed DNA extracted from permafrost and non-permafrost samples and found that all 11 samples contained 5-hydroxy-5-methylhydantoin and 5-hydroxyhdantoin, which are oxidation products of pyrimidines. They were only able to amplify DNA from the samples which had lower levels of hydantoins present. Poinar (2002) found evidence of Maillard reaction products (which can bind DNA to proteins) using gas chromatography and mass spectrometry in sloth coprolites estimated to be 20,000-year-old. Only after the coprolites were treated with with N-phenacylthiazolium bromide (N-PTB) (which cleaves Maillard products) were they able to amplify the DNA. Hansen *et al.* (2006) estimate that cross links accumulate 100 times faster than single strand breaks in permafrost derived ancient DNA. However, Heyn *et al.* (2010) say that all blocking lesions (cross linked or otherwise) are present in no more than 40% of the molecules. Further studies are needed to resolve this issue.

1.1.5.4 DNA damage over time

Based on invitro experiments Lindahl (1993) estimates that DNA cannot survive more than a few hundred thousand years making any claim of multi-million year DNA suspect. However Allentoft *et al.* (2012) predict that DNA from frozen samples might be able to last more than a million years based on estimates of the half-life of mitochondrial DNA fragments extracted from bird bones in New Zealand (which were estimated to be 500 years old). Sawyer *et al.* (2012) tried to find a correlation between DNA damage and its age using DNA from animal remains which were between 18 and 60,000 years old. They found that fragment length was not a good indicator of age while also finding that strand breaks are concentrated adjacent to purine residues and that C → T substitutions at the 5′ ends of the molecules had a strong positive correlation with age, regardless of site and burial conditions.

It is unlikely that a precise rate for DNA degradation will ever be determined as this is entirely dependent on environmental factors including temperature, free water, oxygen, pH, salt, and radiation exposure (Campos *et al.* 2012). Thus, DNA from younger samples can have more damage than DNA from older samples (Orlando *et al.* 2021) making estimating the age of a sample directly from the amount of DNA damage present impossible. More studies from varying environments may help estimate the likelihood of DNA surviving in a specific environment but given the micro-environmental differences within any specific environment, even this is unlikely.

DNA from different species of bacteria is predicted to degrade at different rates due to differences in cell wall composition. Gram-positive bacteria have cell walls which are 2-10 times thicker than those of gram-negative bacteria which is predicted to protect their DNA from environmental assault postmortem. Additionally, bacterial structures such as spores are likely to provide protection from degradation (Setlow 2007). However, no systematic studies comparing the DNA degradation over time versus microbial composition have yet been done.

Unfortunately, many pathogenic bacteria are Gram-negative making their DNA more likely to be degraded and thus harder to detect in ancient samples. *M. tuberculosis* and *M. leprae* (two heavily studied ancient pathogens) have high levels of mycolic acid in their cell walls and produce lipid exudates which are thought to help preserve their DNA. This may account for the lower amounts of deamination seen in ancient *M. leprae* samples (Duchêne *et al.* 2020).

1.1.6 Analysis of ancient DNA

The steps used for sequencing and analyzing modern DNA also apply to sequencing and analyzing aDNA, with the additional step of authenticating which reads are from aDNA versus modern DNA. However, because aDNA has shorter fragments, is damaged, and easy to contaminate the specifics of the individual steps differ from those used for modern DNA. For many steps choices must be made which affect the quality of the final data and any interpretations.

1.1.6.1 Biochemical analysis of ancient DNA

Ancient DNA facilities

To avoid contamination, samples used for aDNA analysis should be handled as little as possible between when they are collected and when they are processed for DNA extraction and sequencing. This includes during excavation, transport, and storage. Of course, this is not possible for all samples, especially those which come from archives or museum collections (Orlando *et al.* 2021).

Extraction and processing of aDNA needs to be done in a dedicated facility to reduce the risk of contamination as this will heavily influence the interpretation of the results. These labs are kept sterile with HEPA-filtered positive air pressure and daily UV/bleach decontamination of work surfaces with

workers dressed in apparel similar to that used in semiconductor clean rooms (Orlando *et al.* 2021). Early aDNA studies were done before facilities like these existed and thus their results are mostly unsubstantiated.

DNA extraction

This step is the most critical in the entire aDNA analysis workflow. Because of the irreproducible nature of aDNA samples it is desirable to use the least amount of source material to obtain the maximum amount of aDNA possible while also leaving enough material for further analysis in the future (Orlando *et al.* 2021).

Early methods required the destruction of the entire sample in order to obtain enough DNA for subsequent analysis (Scarsbrook *et al.* 2022). Various protocols have been proposed and tested (with varying levels of success) which reduce the amount of sample material needed and which allow for the coextraction of proteins (Hofreiter 2012; Gomes *et al.* 2015; Sirak *et al.* 2017; Korlević *et al.* 2018; Fagernäs *et al.* 2020; Harney *et al.* 2021). The varying levels of success are unsurprising given the range of sample materials and that good physical preservation does not guarantee good DNA preservation (Martínez-Delclòs *et al.* 2004).

In general, samples are converted to a fine powder which is then added to series of buffers which decalcify mineral matrices, break down any proteins and lipids present, and release the DNA from the organic and inorganic molecules which it is bound to (Orlando *et al.* 2021). Bleach or an enzyme cocktail can also be used to destroy chemical inhibitors, but this comes at the cost of also destroying some of the DNA present (Damgaard *et al.* 2015; Korlević *et al.* 2015; Gamba *et al.* 2016; Orlando *et al.* 2021). Some protocols include discarding (or setting aside) the first extraction fraction which is predicted to contain the highest concentration of contaminates (Damgaard *et al.* 2015; Gamba *et al.* 2016).

Standard extraction protocols designed for modern undamaged, unfragmented DNA do not work well with aDNA and custom protocols have been developed (Dabney *et al.* 2013a) with silica particles in solution (Höss and Pääbo 1993), on a column (Damgaard *et al.* 2015; Gamba *et al.* 2016; Rohland *et al.* 2018), or attached to magnetic beads (Glocke and Meyer 2017; Rohland *et al.* 2018)

being the current standard. The DNA can then be washed with ethanol and eluted using a low-salt buffer (Orlando *et al.* 2021).

Removing DNA damage

The abundance of uracil molecules in aDNA can bias sequence analyses (Ho *et al.* 2007; Axelsson *et al.* 2008). To reduce the amount of damage induced sequencing errors, some protocols include a step where the extracted DNA is treated with the USER regent from New England Biolabs. This is a commercial mix which includes Uracil-DNA-glycosylase (UDG) and endonuclease VIII (Endo VIII). UDG removes uracil residues and then Endo VIII cleaves the abasic site that was created (Orlando *et al.* 2021).

Although this reduces sequencing errors it also removes the primary signal used to authenticate DNA as being ancient and results in even shorter fragments which can result in less efficient amplification during later steps. For mammalian DNA, CpG dinucleotides can still be used as an alternative damage signal with USER treated samples and for non-mammalian DNA, some protocols split the sample into UDG and non-UDG samples which can then be compared separately. This keeps the damage signal intact but increases the overall cost and complexity of the experiment (Orlando *et al.* 2021).

An alternative protocol known as UDG-half has been developed with removes most of the damage but leaves uracil molecules at the ends of the DNA strands (Rohland *et al.* 2015). This leaves some of the damage signal available for authentication and minimizes damage induced sequencing errors. In general, the decision to apply USER treatment to samples is made on a per study and per sample basis depending on the research question being asked (Orlando *et al.* 2021).

Next generation sequencing library construction

The original double stranded protocols used to create NGS libraries were not optimized for aDNA with most protocols using the T4 DNA polymerase which removes 3′ overhangs and fills in 5′ overhangs before adapters and indexes were ligated to the freshly created double stranded ends (Meyer and Kircher 2010). In contrast newer, single stranded protocols ligate the adapters directly to

individual single stranded molecules (Gansauge and Meyer 2013; Gansauge *et al.* 2020) which results in overhanging 3′ ends and nicked molecules being preserved (Orlando *et al.* 2021). Interestingly, single strand protocols have been extended to enable direct capture of DNA molecules containing uracil bases (Gansauge and Meyer 2014).

Although single stranded protocols have been shown to reduce the amount of DNA lost during the library preparation (Orlando *et al.* 2021) they are not yet the default protocol used in aDNA studies. However, with the development of cheaper protocols which use double stranded polymerases this is likely to change in the near future (Gansauge *et al.* 2017; Harkins *et al.* 2020). Single tube based protocols are likely to help reduce the overall cost even further (Carøe *et al.* 2018). Alternatively, single-molecule sequencing (e.g. Helicos (Milos 2010), Pacific Biosciences (Quail *et al.* 2012) and Oxford Nanopore (Howorka *et al.* 2001)) have the potential to further reduce the cost and complexity of aDNA sequencing.

A standard step in the preparation of NGS DNA libraries is the inclusion of unique identifiers (indexes) as part of the adaptors used for each sample. This enables multiple samples to be sequenced at the same time and while also reducing contamination due to the amplification of DNA which does not have the correct index (Orlando *et al.* 2021). Using indexes on both of the adapters used enables the detection of chimeric DNA templates formed through jumping PCR (Kircher *et al.* 2012) or through index hopping during the cluster generation (van der Valk *et al.* 2020).

Library amplification

In order to have enough DNA, most NGS aDNA libraries need a PCR amplification step before they can be sequenced. However, PCR amplification can skew the final DNA library complexity due to preferential template binding (Dabney and Meyer 2012) and differences in how specific DNA polymerases handle damaged DNA (Seguin-Orlando *et al.* 2015). Polymerases which are commonly used (as they do not substantially skew the NGS DNA library complexity) include Pfu Turbo Cx, Herculase II, and Accuprime Pfx. Of these Pfu Turbo Cx is a non-proof reading polymerase which is able to amplify damaged, uracil containing templates and which is often used during NGS DNA library construction. The other two enzymes are proofreading and often used to amplify NGS DNA libraries or during target enrichment (Orlando *et al.* 2021).

The optimal number of PCR cycles is dependent on the specific sample and should be determined using real-time PCR before the final amplification is performed (Meyer *et al.* 2008). Too many PCR cycles generates PCR duplicates which results in clonality and saturation during the sequencing step (Orlando *et al.* 2021).

Target enrichment

The DNA present in ancient samples is a mixture of DNA from endogenous (target DNA of interest) and exogenous (non-target contaminating DNA) sources (some of which may have come from the individuals processing the samples) (Warinner *et al.* 2017) with the amount of aDNA extracted usually ranging from 1-10% of the overall amount of DNA extracted (Duchêne *et al.* 2020). For bacterial pathogen DNA this can be less than 0.1% of the whole DNA content and is generally proportionally far less than the total DNA extracted (Devault *et al.* 2014b; Rasmussen *et al.* 2015; Andrades Valtueña *et al.* 2017; Vågene *et al.* 2018; Schuenemann *et al.* 2018; Zhou *et al.* 2018; Guellil *et al.* 2018).

Thus it of considerable interest to increase the amount of a target DNA present before sequencing to minimize the costs and to increase the chances of successfully finding DNA relevant to the research question. Early work used probes bound to microarrays while current research predominately uses short DNA or RNA oligonucleotides in solution with probes of varying lengths designed to target specific loci, whole genomes, or somewhere in between (Orlando *et al.* 2021). Probe design needs to take base composition into account to avoid biasing the recovered DNA sequences which can affect downstream analyses (Cruz-Dávalos *et al.* 2017). Additional rounds of enrichment can be useful to a certain extent. However, too many rounds of enrichment can reduce overall library complexity (Orlando *et al.* 2021).

An alternative to sequence specific targeted capture are protocols which enable the direct capture of DNA molecules which contain uracil bases and thus are likely of an ancient origin (Gansauge and Meyer 2014; Weiß *et al.* 2020). Combining both of these approaches could be useful for studying mixed populations of ancient and modern DNA (e.g. *C. tetani* in soil).

Sequencing

Various NGS sequencers have been used to sequence aDNA. From 2006 to 2010 the Roche 454 system was popular, but it has since been replaced by Illumina sequencers due to them being readily available and producing large amounts of data with low error rates. Conveniently Illumina sequencers work well with short (<300bp) fragments of DNA and are thus well suited for sequencing shorter aDNA fragments (Orlando *et al.* 2021).

However, Illumina sequencers are not perfect, with batch affects being possible. These can be partially mitigated by calibrating individual runs using the addition of PhiX DNA as a control (Kircher *et al.* 2009; Renaud *et al.* 2013). To avoid index hopping (when indices bind to the wrong sample during sequencing) heteroduplexes and unbound adapters should be removed before sequencing and chimeric sequences removed computationally after sequencing (Kircher *et al.* 2012; van der Valk *et al.* 2020; Orlando *et al.* 2021).

1.1.6.2 Computational analysis of ancient DNA

Read processing and alignment

DNA can be sequenced in one or in both directions resulting in single reads (SE) or paired-end reads (PE) data. The sequencer software takes the raw florescence values and converts them into fastq files which include Phred encoded quality scores (Ewing and Green 1998; Ewing *et al.* 1998). These scores represent the confidence that the assigned base is the correct one, with p representing the probability that the assigned base is incorrect such that

$$p = 10^{\frac{-Q}{10}}$$

The Q value can be included in the fastq file using several encodings with Phred+33 being the most common, and the older Phred+64 still found in some older datasets. Several other encoding schemes were developed by specific companies for their specific sequencers, but they are rarely seen in modern datasets. Phred+33 and Phred+64 both store a Q value from 0 to 93 as an ascii character starting at 33 (ascii !) for Phred+33 or 64 (ascii @) for Phred+64. As the @ symbol is used to label the lines containing read meta-data in fastq files, badly written parsers are more likely to incorrectly parse Phred+64 encoded files and Phred+33 has become the community standard.

Unfortunately, the Phred encoding scheme used in a specific fastq file is not explicitly defined, leaving it up to the user to try and predict the encoding based on a subset of the file (e.g. the first 1000 reads). This is an important step in any bioinformatic analysis as many bioinformatics programs assume the input data is Phred+33 encoded unless they are explicitly told otherwise. Since the Phred+33 and Phred+64 encodings overlap in their expected range of ascii values a Q value of zero in Phred+64 (p=1) would be interpreted as a Q value of 31 in the Phred+33 encoding ($p=\frac{1}{10^{3.2}}$). Databases such as the National Center for Biotechnology Information (NCBI) sequence read archive (SRA) try to ensure all datasets stored in their collection are Phred+33 encoded but this does not always happen, especially with datasets mirrored from other organizations.

After the fastq files have been checked to ensure they have a consistent encoding they can be processed to remove index and adapter sequences. At the same time reads can be filtered based on their Q values and length, with lower quality bases on the ends removed and PE reads merged into a single sequence. Since the ends of aDNA molecules have an increased concentration of C → T transitions, removing the last few bases on both ends can sometimes improve downstream read mapping at the expensive of potentially removing some of the signal used for authentication that the reads are ancient (Schubert *et al.* 2012). Common programs for doing this include leeHom (Renaud *et al.* 2014), AdapterRemoval v2 (Schubert *et al.* 2016), and fastp (Chen *et al.* 2018).

Read Mapping

Reads can be mapped to reference sequences using a variety of programs. Currently BWA (Li and Durbin 2009) and Bowtie2 (Langmead and Salzberg 2012) are the most commonly used. The accuracy and sensitivity of both programs is dependent on the how close the reads are the reference sequence, the amount of damage present in the reads, the type of DNA library preparation used, whether samples were USER treated, and the specific program parameters used (Orlando *et al.* 2021). The program gargammel (Renaud *et al.* 2017) can be used to generate pseudo aDNA reads from a reference sequence which can then be used to benchmark programs in silico. A few studies have compared the accuracy and specificity of various read mapping programs when used with aDNA (Schubert *et al.* 2012; Martiniano *et al.* 2020; Oliva *et al.* 2021), but more studies are needed.

Competitive mapping should be used when using read mapping to screen metagenomic data for specific organisms. Mapping reads against several related sequences and only keeping those that map to one specific sequence can help reduce the number of false positive matches if the reference sequences are phylogenetically informative (Key *et al.* 2017; Warinner *et al.* 2017). The distribution of reads across the reference sequence should also be checked to determine if there are multiple sources of DNA present. If the DNA is from a single source the coverage is expected have a random even distribution. If multiple sources are present some regions will have a concentration of reads piling up (Warinner *et al.* 2017).

De novo genome assembly

De novo genome assembly of aDNA is controversial. Orlando *et al.* (2021) say that *de novo* assembly of aDNA is uncommon due to environmental contamination, shorter read lengths, and damage. As an outlier they mention the study by Schuenemann et al. (2013) which used *de novo* assembly to obtain a high coverage genome of *M. leprae* from a well preserved sample. However, Bos *et al.* (2019) suggest that de novo assembly can be useful when aDNA comes from organisms with no close reference genome. They also list several de novo assemblers which were specifically developed to work with shorter reads (although not specifically for aDNA) including Velvet (Zerbino and Birney 2008), SPAdes (Bankevich *et al.* 2012) and SOAPdenovo (Luo *et al.* 2012), along with several meta-genomic assemblers including Ray Meta (Boisvert *et al.* 2012), MetaVelvet-SL (Namiki *et al.* 2012), MEGAHIT (Li *et al.* 2016), and metaSPAdes (Nurk *et al.* 2017). They note that how well each program does is dependent on the input data and users should experiment (van der Walt *et al.* 2017; Sczyrba *et al.* 2017) and compare the various assemblies using program such as QUAST (Gurevich *et al.* 2013). Of particular interest, *de novo* assembly of mixtures of ancient and modern DNA should be tested.

Authentication

After reads have been mapped to reference sequences there are four key criteria that can be used to try to separate ancient endogenous DNA from exogenous (presumed modern) DNA.

1. Checking for an increase in cystine deamination at the ends of molecules as this is considered to be the primary signature of ancient DNA molecules (Jónsson *et al.* 2013). Rohland *et al.* (2015) suggest a cutoff of at least 3% damage for partially UDG treated samples and 10% for untreated samples.
2. Checking for a disproportionate frequency of purines next to the end of fragments as this is unique to aDNA molecules (Arning and Wilson 2020).
3. Checking for a log-normal tailing off of the number of reads versus fragment length (Warinner *et al.* 2017).
4. Checking the distribution of reads across the reference genome along with their edit distances (the minimum number of operations required to transform one sequence into another). aDNA should be randomly distributed with minimal edit distances. A concentration of reads could be reads from multiple similar sequences or PCR duplicates (Hübler *et al.* 2019).

Several programs can be used to estimate how damaged DNA fragments after they have been mapped to a reference. mapDamage2 (Jónsson *et al.* 2013) uses a bayesian framework to estimate several damage parameters (on a per sample basis) and generates several plots (and associated data files) including misincorporations and read length distribution. It does not use terminal purine frequency or edit distance in its calculations. It also does not estimate how contaminated a sample is.

In contrast PMDtools (Skoglund *et al.* 2014; pontussk 2021) estimates damage on a per read basis and can filter reads based on the estimated amount of damage present. More recently developed programs include AuthentiCT (Peyrégne and Peter 2020), pyDamage (Borry *et al.* 2021), and DamageProfiler (Neukamm *et al.* 2021). These programs should be tested with a variety of simulated datasets to determine how well they perform with different types of data.

However, regardless of the underlying model used, the output from these programs can be misleading. Damaged DNA is not guaranteed to be ancient, and undamaged DNA is not guaranteed to be modern. If the source material is particularly old there will be some exogenous DNA with similar amounts of damage as the endogenous DNA of interest (Weiß *et al.* 2020). Additionally whether samples were fully or partially UDG treated will change the damage estimates (Rohland *et al.* 2015).

Microbiome profiling

For microbiome studies, a secondary challenge (beyond determining if the DNA is damaged and/or contaminated) is identifying which organisms were present. This is done using programs which compare raw reads, or *de novo* assembled contigs, to a reference database built from specific loci (e.g. 16S rRNA), multiple loci (e.g. housekeeping genes) or whole genomes (Warinner *et al.* 2017). Querying these databases is then done by aligning query sequences to the reference sequences in the database or by using k-mers to match query sequences to reference sequences in the databases.

Alignment based programs include mothur (Schloss *et al.* 2009) and QIIME (Caporaso *et al.* 2010) (database of 16S rRNA), Metagenomic Phylogenetic Analysis (METAPHLAN) (Segata *et al.* 2012) (database of marker genes) and MEGAN Alignment Tool (MALT) (Herbig *et al.* 2016) and MIDAS (Nayfach *et al.* 2016) (whole genome database). K-mer based programs include Kraken (Wood and Salzberg 2014) and Kaiju (Menzel *et al.* 2016). Each program has specific strengths and weaknesses and have been compared in several studies (Warinner *et al.* 2017; Sczyrba *et al.* 2017; Velsko *et al.* 2018; Eisenhofer and Weyrich 2019).

Although none of these programs were developed to specifically work with aDNA, METAPHLAN and MALT are used by metaBIT (Louvel *et al.* 2016) and HOPS (Hübler *et al.* 2019) respectively as part of a pipeline for metagenomic analyses. Although metaBIT has been tested with ancient DNA samples, it was not explicitly designed for aDNA (Louvel *et al.* 2016). In contrast HOPS was explicitly designed to work with and authenticate aDNA (Hübler *et al.* 2019).

Regardless of the algorithms used, these programs are only as good as the databases they use, which are only as good as the reference data they were built from. Many organisms have not been cultured or had their genome sequenced and thus missing from any databases (Warinner *et al.* 2017). Additionally with ancient pathogens there can be many false positives, as many pathogenic bacteria are from the same genre as environmental bacteria (Campana *et al.* 2014; Warinner *et al.* 2017).

An additional complication is that bacterial communities shift over time, especially post mortem, and a species DNA survival is predicted to be heavily dependent on the GC content and cell membrane composition (Rollo *et al.* 2007; Arning and Wilson 2020). Clostridial species are often found in ancient microbiome studies but this is thought to likely be due to post death colonization by soil species which then acquired an ancient signature (Philips *et al.* 2017; Arning and Wilson 2020) or from clostridial species in the gut spreading throughout the host post-mortem (Javan *et al.* 2017).

aDNA Analysis Pipelines

The wide range of sample types and research questions, combined with an ever-increasing number of bioinformatic tools has resulted in a lack of standards for analyzing aDNA. This has been exacerbated by differences in how each paper describes its methods making it hard to replicate or extend previous work (Orlando *et al.* 2021). Analysis pipelines like nf-core/eager (Fellows Yates *et al.* 2021) help make aDNA analyses scalable and reproducible, but more work is needed with individual programs to determine how well they work with uncommon, but interesting, datasets (such as mixtures of ancient and modern *C. tetani* sequences).

1.1.7 Ancient Pathogens

DNA from ancient pathogens is used to study two main questions: How has a particular pathogen changed over time and when/where has it crossed paths with humans? The DNA of Gram-negative organisms is thought to be less protected and thus less likely to be found in ancient samples. However DNA from Gram-negative pathogens has been successfully extracted from teeth (*Y. pestis* (Bos *et al.* 2011), *Salmonella enterica* serovar Paratyphy C (Vågene *et al.* 2018)), an alcohol preserved colon (*Vibrio cholerae* (Devault *et al.* 2014a)) and mummified tissues (*Helicobacter pylori* (Castillo-Rojas *et al.* 2008; Swanston *et al.* 2011; Maixner *et al.* 2016)). For Gram-positive pathogens, *M. tuberculosis* and *M. leprae* have been the focus, aided by the existence of skeletal lesions which help determine which bones to further investigate. Interestingly, *Clostridium* species are Gram-positive which may account for their persistent association with archeological samples (Philips *et al.* 2017), but to date they have not been the focus of an aDNA study.

1.1.8 Sources of ancient DNA datasets

Thus far, studies of aDNA have used a very targeted approach to decide what DNA to look for in specific samples. A few studies have used metagenomic methods in their analyses, but those were still focused on specific samples predicted to contain aDNA relevant to the specific research question.

What is needed is a publicly available repository which contains DNA extracted and sequenced from ancient samples which can then be queried for datasets of interest.

The SRA is the world's largest publicly available repository of genomic sequencing data (Katz *et al.* 2022). The SRA contains publicly deposited datasets from genome sequencing, shotgun metagenomic, amplicon sequencing, transcriptomics, and virtually any other application of DNA and RNA sequencing, including the study of aDNA. The NCBI makes the datasets in the SRA freely available to the public as part of their "findable, accessible, interoperable and reusable (FAIR)" policy with copies of the data stored on NCBI servers and mirrored to Amazon and Google cloud servers (Katz *et al.* 2022). Given the global use and availability of the SRA, there is potential to use the SRA for genomic biomonitoring of pathogens, the analysis of publicly available environmental samples for metagenomic applications, and the large-scale analysis of ancient DNA.

The volume of data in the SRA has grown exponentially since its inception in 2009 (Katz *et al.* 2022). It currently contains more then 36 petabytes of data and is predicted to contain over 43 petabytes by 2023 (2020). Because of its vast size and continuous growth, searching the SRA for a specific gene or organism is not possible within a reasonable time frame using traditional bioinformatic workflows (e.g. BLAST (Basic Local Alignment Search Tool) (Altschul *et al.* 1990) searches). Although BLAST was designed to be a rapid sequence similarity search algorithm, pre-computing the BLAST database for the entire SRA would be unfeasible for storage reasons alone. Therefore, the NCBI does not offer a BLAST service for the entire SRA, and only permits dataset specific BLAST searches.

To help researchers search the SRA for datasets with specific characteristics, the NCBI recently developed the SRA Taxonomy Analysis Tool (STAT) (Katz *et al.* 2021). STAT uses a precomputed k-mer (short sequence fragment of length k) index (k=32) to provide a rough estimate of the taxonomic composition of every sequencing run in the SRA. The STAT results and sequencing meta-data can be queried via Google's Big Query or Amazon's Athena cloud resources (Katz *et al.* 2022). Although this makes it possible to search the SRA for datasets containing specific organisms, it is not currently feasible to directly search the entire SRA for specific sequences of interest or for organisms that were not in the original set of reference genomes. Since the SRA potentially contains data for most sequencing studies since 2007, it likely includes many sequences from ancient DNA samples to date.

1.1.9 Thesis Hypothesis and Objectives

This thesis explores the hypothesis that the SRA is a valuable resource for the discovery of novel pathogenic clostridial genomes. Key objectives include:

- Application of a data-mining approach using the STAT to identify which datasets (among the millions present in the SRA) are estimated to contain non-trivial levels of *C. tetani* DNA
- Recovery and analysis of *C. tetani* draft genomes, including damage analysis and a phylogenetic comparison with modern *C. tetani* strains
- Analysis of tetanus neurotoxin genes from these samples in collaboration with experimentalists and experts in clostridial neurotoxin biology

Through these objectives, this thesis provides insights into the diversity of *C. tetani* and clostridial neurotoxins in general, which are still poorly understood in terms of their ecology and evolutionary history (Montecucco and Rasotto 2015; Mansfield and Doxey 2018).

Chapter 2
Searching the SRA for *Clostridium tetani*

Material in this chapter has been prepared for publication and is available as a pre-print in *bioRxiv* accessible at the following DOI: 10.1101/2022.06.30.498301

Introduction

To explore the genomic diversity of *C. tetani*, STAT analysis results were used to search the entire NCBI Sequence Read Archive (10,432,849 datasets from 291,458 studies totaling ~18 petabytes of compressed data June 8, 2021) for datasets predicted to contain *C. tetani* DNA. Of the top 136 datasets predicted to contain *C. tetani* DNA, 76 were from human archeological remains and were chosen for further study. Draft genomes were assembled for each sample and compared with modern strains. From the draft genomes several novel *Clostridial* species and TeNT like toxins were discovered. One TeNT like toxin was experimentally tested and demonstrated to have similar properties to modern TeNT.

Methods

All major calculations were run on Compute Canada servers to obtain results in a timely manner. GNU Parallel (Tange 2011) was also used to efficiently use computational resources.

NCBI STAT analysis

The NCBI developed the SRA taxonomy analysis tool (STAT) to supplement and enhance the existing meta-data that is included with each SRA dataset when it is submitted (Katz *et al.* 2021). STAT k-mers were chosen using an iterative min-hash algorithm were each segment of a reference genome of length L was split into L - 64 + 1 k-mers of length 64. Each potential k-mer was converted to a binary representation and then to a 64-bit hash. The k-mer with the smallest 64-bit hash was then chosen as the representative k-mer for that segment of the reference sequence. Any representative k-mers that appeared in multiple sequences were mapped to the lowest common taxonomic level they shared.

In the SRA, every read has been mapped to taxonomic labels using these representative k-mers. For ever dataset there is an additional table that can be queried which contains the run ID, the predicted taxonomic label, a k-mer self-count, and a k-mer total-count. The self-count represents only k-mers that mapped to that specific taxonomic label, whereas the total-count represents k-mers from that label or lower in the taxonomic tree (Katz *et al.* 2021, 2022).

Identification of sequencing runs predicted to contain C. tetani DNA in the NCBI sequence read archive

To identify datasets within the SRA predicted to contain *C. tetani* DNA the NCBI-STAT database was queried (March 15, 2021) via Google's Big Query API using

```
google-cloud-sdk/bin/bq --format=csv query --nouse_legacy_sql --max_rows
20000000 'SELECT m.* EXCEPT (attributes, biosamplemodel_sam,
geo_loc_name_sam, ena_first_public_run, ena_last_update_run,
sample_name_sam,
jattr,datastore_filetype,datastore_provider,datastore_region),
tax.total_count, tax.self_count FROM nih-sra-datastore.sra.metadata as m,
nih-sra-datastore.sra_tax_analysis_tool.tax_analysis as tax WHERE
m.acc=tax.acc and tax_id=1513  ORDER BY tax.total_count' >
Clostridium_tetani.txt
```

STAT estimation of taxonomic abundance for specific sequencing runs in the NCBI sequence read archive

STAT results for datasets chosen for further analysis were obtained (June 11, 2021) using the query

```
SELECT * FROM nih-sra-datastore.sra_tax_analysis_tool.tax_analysis AS tax
WHERE tax.acc IN ("DRR046402","DRR046405",…)
```

where "…" represents the remaining 74 sequencing runs. Total counts for each mapped bacterial and archaeal taxon at the species level were extracted, were converted to proportional values and subsequently visualized in R v4.0.4.

Downloading datasets and checking Phred encodings

FASTQ files of identified sequencing runs were downloaded using the `fasterq-dump` program from the sra-toolkit v2.9.6 The quality encodings of all runs were checked using the first 10000 lines of each fastq file using awk and od. Eight runs were Phred+64 encoded and were converted to Phred+33 using seqtk v1.3. Twenty-three runs (from 9 different BioSamples) had an unknown encoding and were assumed to be Phred+33 encoded based on the range of the quality scores.

Measurement and visualization of genome coverage

Bowtie2 v2.4.2 (Langmead and Salzberg 2012) was used to map reads from individual runs to the *C. tetani* E88 strain chromosome (NCBI accession : NC_004557.1) and plasmid (NCBI accession : NC_004565.1) which were then converted to a bam file using samtools v1.12. The bam file was then sorted, indexed, and merged with other bam files which had the same BioSample ID.

The total (average # of reads per base) and percent (number of bases with 1 or more reads divided by total number of bases) coverage was calculated for the entire chromosome and plasmid as well as the *tent* (68640 - 72587) and *colT* (39438 - 42413) regions. Coverage was visualized using Python v3.8.5 and matplotlib v3.3.2. Circular plots were created using R and a custom script.

Circular coverage plots were generated by loading the BAM files into R v4.1.0 with the Rsamtools library v2.8.0 and plotted as area plots using functions from the circulize library v0.4.12. Coverage was calculated by averaging the number of reads per base in 300bp bins for the plasmid sequences, and 11,250bp bins for the chromosome sequences. Values were capped to the 90th percentile to prevent high coverage regions from obscuring other regions. Genes were plotted as black bars using RefSeq annotations. For the plasmid plots, the *tent* (68,640-72,587) and *colT* (39,438-42,413) genes were also coloured red and blue, respectively.

Genome reconstruction

Reads were pre-processed using fastp v0.20.1 (Chen *et al.* 2018) with default settings to perform quality filtering and remove potential adapters. FASTQ pre-processing statistics are included in the Supplementary data. Metagenome co-assembly, using all reads with the same BioSample ID, was

performed using megahit v1.2.9 with default parameters (Li *et al.* 2014). Contigs were then taxonomically classified using Kaiju v1.7.4 (Menzel *et al.* 2016) against the Kaiju database nr 2021-02-24 with default settings. Any contigs mapped to *C. tetani* (NCBI taxonomy ID 1513) or any of its strains (NCBI taxonomy IDs 1231072, 212717, 1172202, and 1172203) were selected for further analyses. The length of total *C. tetani* contigs was compared to the mapped read coverage with cor.test() in R v4.0.4. The average nucleotide identity (ANI) was calculated using fastANI v1.33 (Jain *et al.* 2018; 'FastANI' 2022). CheckM v1.0.18 (Parks *et al.* 2015) was used on the contigs identified as *C. tetani* with the pre-built set of *Clostridium* markers supplied with the tool to calculate completeness, contamination, and strain heterogeneity. Contigs are available with the Supplementary data.

Analysis of ancient DNA damage

Fastq files were pre-processed using leeHom v1.2.15 (Renaud *et al.* 2014) to remove adapters and to perform Bayesian reconstruction of aDNA. The `--ancientdna` flag was applied only to paired end datasets. The leeHom output was then merged by bioSample ID (concatenated sequentially into one file per bioSample ID). Individual and merged results were then processed using seqtk v1.3 to remove sequences < 30 bp in length. For each bioSample, trimmed reads were then mapped using bwa v0.7.17 to the contigs that were classified as *C. tetani* using Kaiju, and separately to the human mitochondrial reference genome (NCBI accession # NC_012920.1) with parameters (`-n 0.01 -o 2 -l 16500`) and converted to a sorted bam file using samtools v1.12. Misincorporation rates were measured for all samples using mapDamage2 v2.2.1 (Jónsson *et al.* 2013) with parameters (`--merge-reference-sequences` and `--no-stats`).

Whole genome SNP-based phylogenetic reconstruction

Single nucleotide polymorphisms within the assembled *C. tetani* contigs were identified using snippy-multi from the Snippy package v4.6.0 (https://github.com/tseemann/snippy) using the *C. tetani* E88 strain as the reference genome (GCA_000007625.1_ASM762v1_genomic.gbff). A genome-wide core SNP alignment was constructed using snippy-core. Five aDNA samples (SAMEA103957995, SAMEA103971604, SAMEA3486793, SAMEA104402285, SAMEA3937653)

were removed due to very poor alignment coverage (<1%). Using the resulting alignment, phylogeny was built using FastTree (Price *et al.* 2010) v2.1.10 with the GTR model and aLRT metric for assessment of clade support. The alignment and tree can be found in the Supplementary data.

Sequence analysis of ancient tetanus neurotoxins

From the plasmid read alignments used earlier, reads aligning to the *tent* region were extracted, and re-aligned using BWA mem v0.7.17-r1188 using default parameters. Read alignments were manipulated with samtools v1.12 and htslib v1.12. The read alignment was restricted to the *tent* gene locus for variant calling (using the reverse complement of NC_004565.1, bases 1496-5443). Variants were called on each individual sample using the Octopus variant caller v0.7.4 (Cooke *et al.* 2021) with stringent parameters (`--mask-low-quality-tails 5 --min-mapping-quality 10 --min-variant posterior 0.95 --min-pileup-base-quality 35 --min-good-base-fraction 0.75`). This combination of parameters reports only variants with very high confidence and read mapping quality, minimizing identification of false positive variant calls. Consensus sequences of *tent* genes were built from each sample using the bcftools consensus tool v1.12, and htslib v1.12, replacing positions with 0 coverage with a gap character. MAFFT v7.4.80 (Katoh and Standley 2013) was used to realign fragments against the reference sequence using the `--keeplength` option, which keeps the length of the reference unchanged and therefore ignores the possibility of unique insertions. The final *tent* alignments are available in the Supplementary data.

Phylogenetic analysis of ancient tetanus neurotoxins

The *tent* consensus alignment generated as described earlier was processed to keep only sequences (N = 20) with alignment coverage exceeding 80%. The following BioSamples were removed: SAMEA104402285, SAMEA104281225, SAMEA104281219, SAMEA5054093, SAMN02799091, SAMEA103971604, SAMN02799089, SAMN12394113, SAMN06046901, SAMEA104233049, SAMEA6502100, SAMEA3486793, SAMEA3713711. The 20 ancient *tent* gene sequences were aligned with 30 *tent* sequences from modern *C. tetani* strains, which reduced to 12 representative modern *tent* sequences after duplicates were removed using Jalview v2.9.0b2 (Waterhouse *et al.* 2009). *tent*/E88 was identical with *tent* from 11 strains (1586-U1, CN655, 641.84, C2, Strain_3,

75.97, 89.12, 46.1.08, A, 4784A, Harvard), *tent*/132CV with 1 other (Mfbjulcb2), *tent*/63.05 with 2 others (3483, 184.08), *tent*/1337 with 2 others (B4, 1240), *tent*/ATCC_453 with 1 other (3582), and *tent*/202.15 with 1 other (358.99). A phylogeny was constructed using PhyML v3.1 (Guindon and Gascuel 2003) with GTR model, empirical nucleotide equilibrium frequencies, invariable sites = none, across site rate variation optimized, NNI tree search, and BioNJ as the starting tree. PhyML analysis identified 362 patterns, and aLRT (SH-like) branch supports were calculated. The final newick tree is available in the Supplementary Data.

TeNT alignment visualization

The *tent* MSA was loaded into R v4.1.0 using the Biostrings library v2.60.1 (Pagès *et al.* 2022). An equal dimension matrix was created and each position assigned the colour black if the DNA letter matched the reference sequence, red if it did not, yellow if it did not and additionally was not present in any of the known modern *tent* sequences, and none if a gap was present. This matrix was plotted as a tile plot using the ggplot2 library v3.3.3 .

Structural analysis of ancient tetanus neurotoxin

A structural model of TeNT/Chinchorro was generated by automated homology modeling using the SWISSMODEL server (Waterhouse *et al.* 2018). Modeling was performed using two top-scoring homologous template structures of tetanus neurotoxins: PDB IDs 7BY5.1.A (97.18% identity), 5N0C.1.A (97.34% identity). 7BY5.1.A was selected as the best template based on the QMEAN quality estimate (Benkert *et al.* 2011). The model was visualized using PyMOL v2.4.1 (Schrödinger, LLC 2015) and unique substitutions (present in TeNT/Chinchorro but absent in modern TeNT sequences) were highlighted.

Experimental testing of TeNT***/Chinochorro (ch***TeNT***)***

The following work was done by collaborators at Boston Children's Hospital (Harvard Medical School) in the lab of Dr. Min Dong.

Antibodies and constructs: Antibodies for Syntaxin-1 (HPC-1), SNAP25 (C171.2), VAMP1/2/3 (104102) were purchased from Synaptic Systems. Antibody against actin (AC-15) was purchased from Sigma. The cDNAs encoding ch-LC-H_N (the N-terminal fragment, residues 1-870) and ch-H_C (the C-terminal fragment, residues 875-1315) were synthesized by Twist Bioscience (South San Francisco, CA). The cDNA encoding TeNT-LC-H_N (residues 1-870) and TeNT-H_C were synthesized by GenScript (Piscataway, NJ). A thrombin protease cleavage site was inserted between I448 and A457 in both TeNT-LC-H_N and ch-LC-H_N. LC-H_N fragments were cloned into pET28a vector, with peptide sequence LPETGG fused to their C-termini, followed by a His6-tag. H_C fragments were cloned into pET28a vectors with a His6-tag and thrombin recognition site on their N-termini.

Protein purification: E. coli BL21 (DE3) was utilized for protein expression. In general, transformed bacteria were cultured in LB medium using an orbital shaker at 37 °C until OD600 reached 0.6. Induction of protein expression was carried out with 0.1 mM IPTG at 18 °C overnight. Bacterial pellets were collected by centrifugation at 4,000 g for 10 min and disrupted by sonication in lysis buffer (50 mM Tris pH 7.5, 250 mM NaCl, 1 mM PMSF, 0.4 mM lysozyme), and supernatants were collected after centrifugation at 20,000 g for 30 min at 4 °C. Protein purification was carried out using a gravity nickel column, then purified proteins were desalted with PD-10 columns (GE, 17-0851-01) and concentrated using Centrifugal Filter Units (EMD Millipore, UFC803008).

Sortase ligation: H_C protein fragments were cleaved by thrombin (40 mU/μL) (EMD Millipore, 605157-1KU) overnight at 4 °C. Ligation reaction was set up in 100 μL TBS buffer with LC-H_N (8 μM), H_C (5 μM), Ca2+ (10mM) and sortase (1.5 μM), for 1 hour at room temperature. Then full-length proteins were activated by thrombin (40 mU/μL) at room temperature for 1 hour. Sortase ligation reaction mixtures were analyzed by Coomassie blue staining and quantified by BSA reference standards.

Neuron culture and immunoblot analysis: Primary rat cortical neurons were prepared from E18-19 embryos using a papain dissociation kit (Worthington Biochemical) following the manufacturer's instruction. Neurons were exposed to sortase ligation mixtures in culture medium for 12 hrs. Cells were then lysed with RIPA buffer with protease inhibitor cocktail (Sigma-Aldrich). Lysates were centrifuged at 12000 g at 4 °C for 10 min. Supernatants were subjected to SDS–PAGE and immunoblot analysis.

Animal study: All animal studies were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee (Protocol Number: 18-10-3794R). Toxins were diluted using phosphate buffer (pH 6.3) containing 0.2% gelatin. Mice (CD-1 strain, female, purchased from Envigo, 6-7 weeks old, 25–28 g, n=3) were anesthetized with isoflurane (3–4%) and injected with toxin (10 μL) using a 30-gauge needle attached to a sterile Hamilton syringe, into the gastrocnemius muscles of the right hind limb, and the left leg served as negative control. Muscle paralysis was observed for 4 days. The severity of spastic paralysis was scored with a numerical scale modified from a previous report (0, no symptoms; 4, injected limb and toes are fully rigid) (Mellanby *et al.* 1968).

Biosafety and biosecurity: Procedures were approved by the Institute of Biosafety Committees at Boston Children's Hospital (Protocol Number: IBC-P00000501). To ensure biosafety and biosecurity, no active full-length *tent*/Chinchorro toxin gene was produced in any form. The amount of sortase ligation reaction was strictly controlled to ensure that only a minimal amount of full-length toxins was produced, which was immediately utilized for functional studies.

Results

2.1.1 Identification and assembly of draft *C. tetani* genomes from archeological samples

Searching the SRA for datasets predicted to contain *C. tetani* DNA returned 43,620 hits, of which 42719 (98%) had a k-mer total-count <= 1000 and were ignored as they were likely false positives or would have insufficient coverage to be useful. The meta-data for top 25 datasets (sorted by the number of k-mer total-count) listed *C. tetani* as the target organism. Of the 136 sequencing datasets possessing the highest predicted *C. tetani* DNA (k-mer total-count >23,000), 28 were previously sequenced *C. tetani* genomes, along with 108 uncharacterized sequencing runs with high levels of *C. tetani* DNA content, of which 79 were labeled as being human samples. Manual curation determined that 76 out of these 79 human samples were collected from archeological human bone and tissue specimens, with the remaining three from modern human gut microbiome samples. Based on the associated publications, 31 of the 38 (82%) were found to be from teeth, with only 1 of the 31 being from dentine and the remaining 30 assumed to be from dental pulp. Of the six non-teeth samples, one appears to be from a mummy chest extract with the remaining five being from various bones.

These 76 ancient DNA datasets are from 38 distinct archeological samples (bioSample IDs), spanning a timeframe of ~6,000 years. Although these archeological samples are of a human origin, STAT analysis of the 38 DNA samples predicted a predominantly microbial composition (Figure 2). *C. tetani*-related DNA was consistently abundant among predicted microbial communities, detected at 13.82% average relative abundance (Figure 2, Supplementary Data).

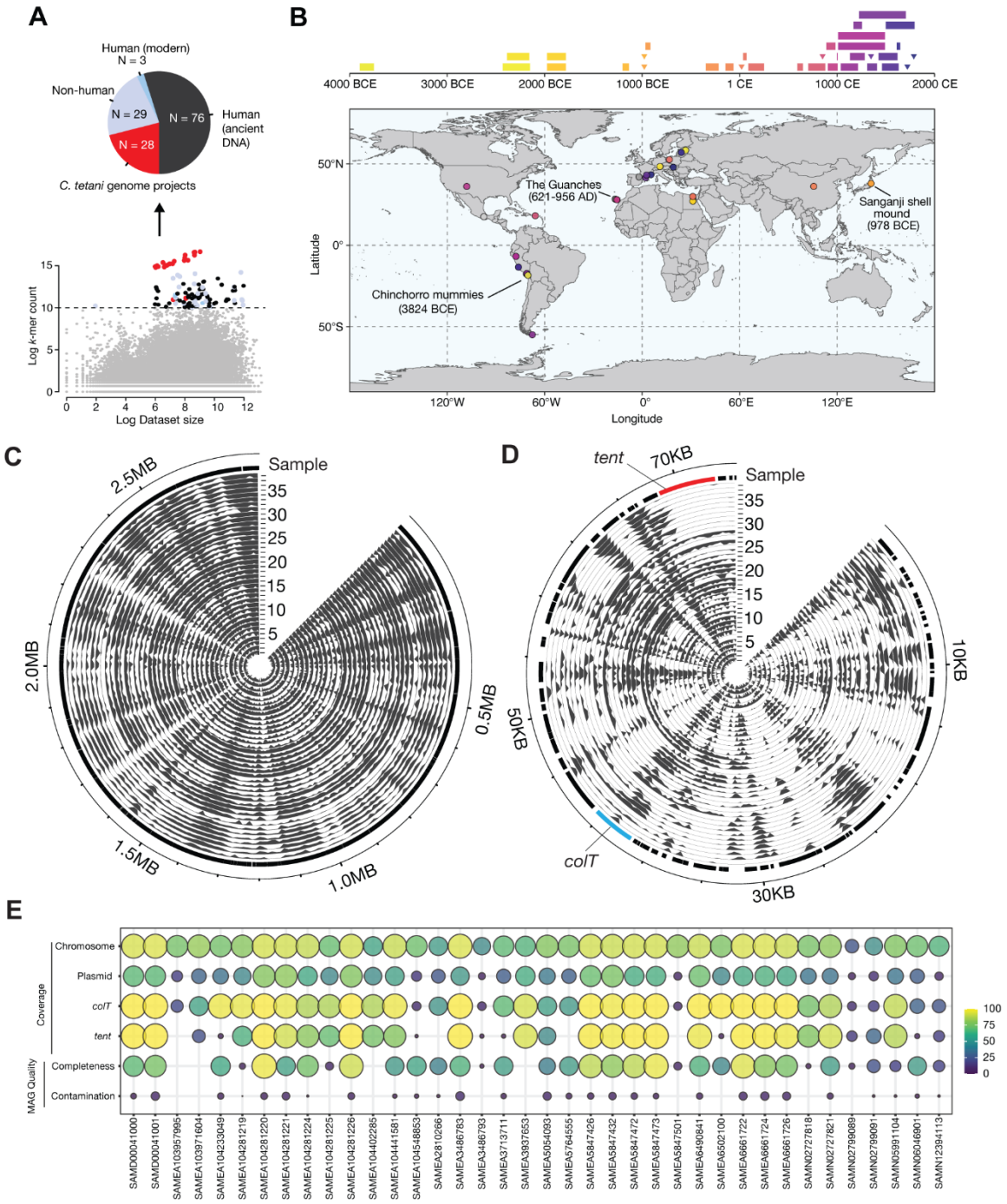

Figure 1 : Petabase-scale screen of the NCBI sequence read archive predicts the presence of *C. tetani* DNA in ancient human archeological samples.

(A) Analysis of 43,620 samples from the NCBI sequence read archive. Each sample is depicted according to its *C. tetani* k-mer abundance (y-axis) versus the overall dataset size (x-axis). An arbitrary threshold, based on computational resources available, was used to distinguish samples with high detected *C. tetani* DNA content. These data points are colored by sample origin: modern *C. tetani* genomes (red), non-human (light blue), modern human (blue), ancient human (black). The pie chart displays a breakdown of identified SRA samples with a high abundance of *C. tetani* DNA signatures.

(B) Geographical locations and timeline of ancient DNA samples. The 76 ancient DNA datasets are associated with 38 distinct samples (bioSample IDs), which are represented as individual data points. Four samples lack date information and are absent from (B).

(C) *C. tetani* chromosomal percent coverage and; (D) plasmid percent coverage detected for reads from archeological samples using the *C. tetani* E88 genome as a reference. The *tent* and *colT* genes are indicated on the plasmid in red and blue, respectively.

(E) Average per-sample coverage of *C. tetani* chromosome, plasmid, and key virulence genes, *tent* and *colT*. Also shown is the estimated completeness and contamination of *C. tetani* draft genomes assembled from archeological samples as calculated by CheckM.

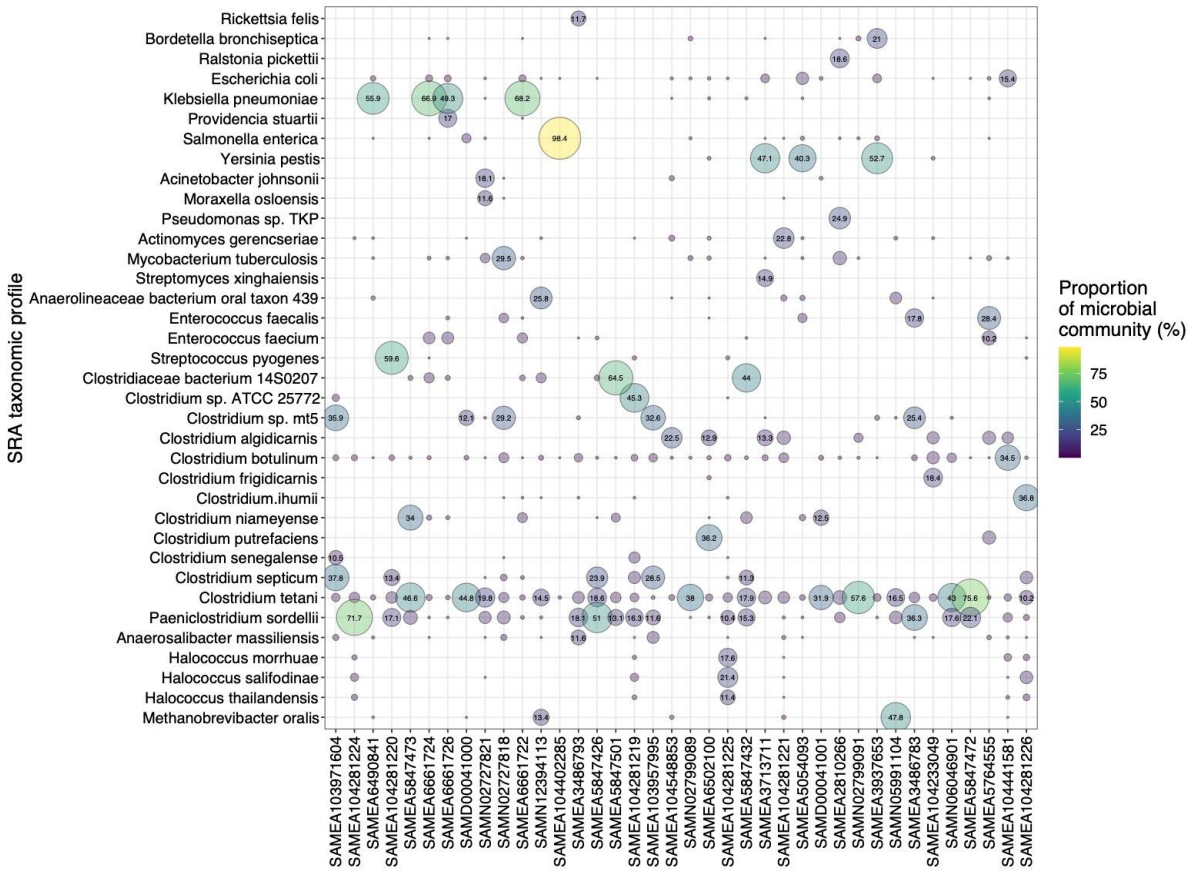

Figure 2 : Predicted proportional abundance of microbial taxa detected.

Abundance values are based on NCBI sequence read archive taxonomic profiles including all identified bacterial and archaeal species. Only species with >10% abundance in at least one sample have been plotted. Only values greater than 10% are depicted in the figure. For bioSample IDs associated with more than one SRA ID, the sample with a median count of *C. tetani* was chosen. In the case of a choice between two, a random SRA ID was chosen.

To further verify the presence of *C. tetani* DNA in the archeological samples, reads from each sample were mapped to the modern *C. tetani* reference (Harvard E88 strain) chromosome and plasmid. Percent coverage was evenly distributed across the chromosome for most samples (Figure 1C/D) whereas coverage across the plasmid was more variable, with some samples lacking coverage

for specific plasmid regions and genes (Figure 1C/D). Sequencing reads mapping to the *tent* gene were detected in 34/38 (89%) samples, whereas reads mapping to a second plasmid-encoded virulence gene, *colT*, were detected in all samples (Supplementary Data). Thus, all detected *C. tetani*-like genomes from ancient samples likely possess a chromosome and plasmid, and all but four are likely toxigenic.

Reads were metagenomically assembled using megahit (Li *et al.* 2016) for each sample and the resulting contigs were taxonomically classified using Kaiju (Menzel *et al.* 2016) to identify those mapping to *C. tetani* and not to other bacterial species (Supplementary Data). *C. tetani* contigs comprised an average of 4.12% of total assembly size in the archeological samples, reaching as high as 28.20% (Supplementary Data). All *C. tetani* contigs from each sample were binned together resulting in 38 *C. tetani* draft genomes, which were further assessed using CheckM (Parks *et al.* 2015) for percentage completion, contamination resulting from genomic fragments of divergent taxa, and strain heterogeneity estimated based on fragments from different strains of the same species (Figure 1E, Supplementary Data).

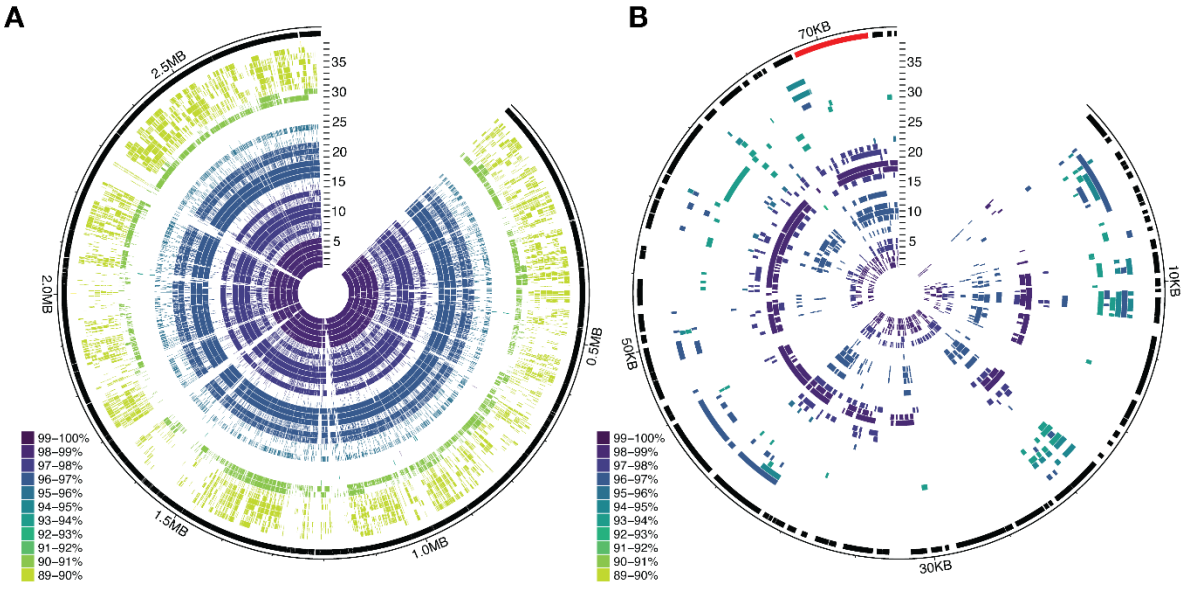

Figure 3 : Alignment of *C. tetani* draft genomes from 38 ancient samples with the reference *C. tetani* E88 chromosome (A) and plasmid (B).

Samples are labeled 1-38 and have been sorted and colored based on their average percentage identity to the reference. See Appendix: Table 1 for the ring versus bioSample ID mapping.

2.1.2 A subset of *C. tetani* draft genomes show signs of age associated DNA damage

The *C. tetani* draft genomes were examined for characteristic patterns of ancient DNA damage using mapDamage2 (Jónsson *et al.* 2013). Seven of the reconstructed *C. tetani* draft genomes exhibited patterns of ancient DNA damage, with a damage rate greater than 10% (the recommended threshold for non-UDG treated samples (Rohland *et al.* 2015)). The highest damage rate (19%) occurred in the draft genome from a *Y. pestis* tooth sample from France (circa 1348 CE) (Namouchi *et al.* 2018) (Figure 5). As controls, the corresponding human mitochondrial DNA (mtDNA) from the same samples (Figure 5), and 21 modern *C. tetani* samples (not shown) was examined. A significant correlation between damage rates of draft *C. tetani* genome DNA and human mtDNA ($R^2 = 0.46$, $p < 0.01$, two-sided Pearson) (Figure 4) was observed, although human mtDNA rates were generally of higher magnitude (Figure 5). Finally, draft *C. tetani* genomes from the archeological samples had significantly shorter fragment lengths ($p = < 0.01$, Wilcoxon test) than those obtained from 21 sequencing datasets of modern *C. tetani* genomes (not shown) with the 21 modern *C. tetani* draft genomes showing no evidence of DNA damage. Thus, some of the *C. tetani* draft genomes from the archeological samples display evidence of ancient DNA damage and are possibly of an ancient origin.

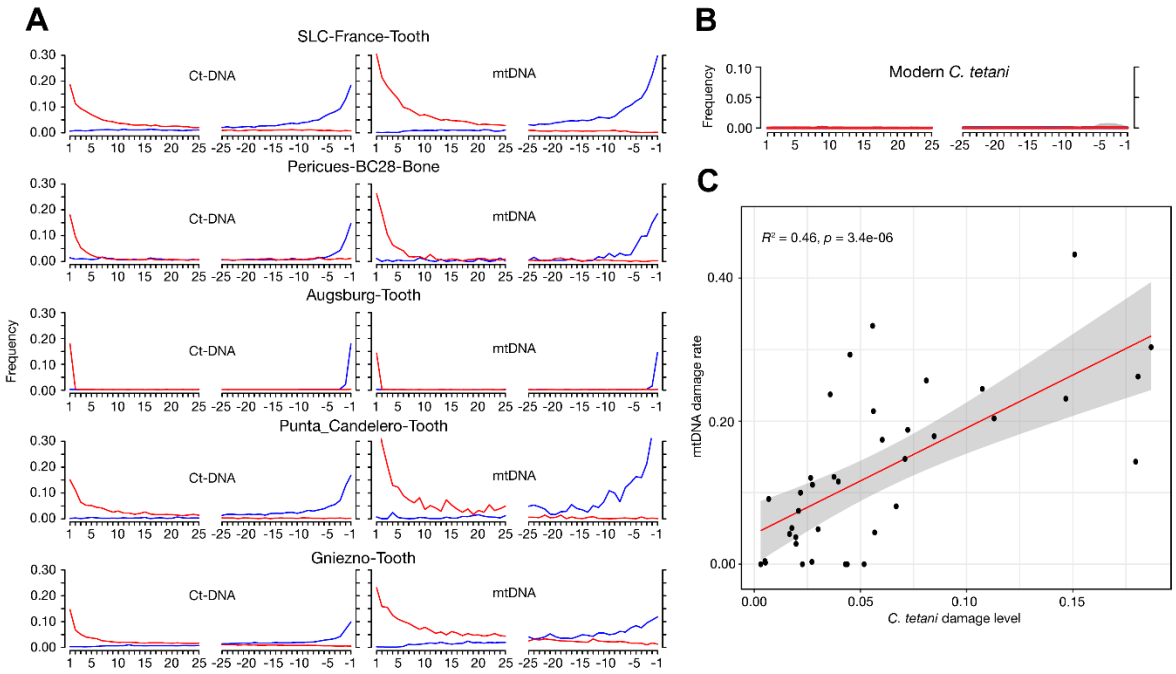

Figure 4 : *C. tetani* DNA from a subset of ancient samples show hallmarks of ancient DNA.

(A) MapDamage misincorporation plots for five *C. tetani draft genomes* displaying the highest damage levels. The plot shows the frequency of C→T (red) and G→A (blue) misincorporations at the first and last 25 bases of sequence fragments. Increased misincorporation frequency at the edges of reads is characteristic of ancient DNA.

(B) This pattern is not observed in a representative modern *C. tetani* genomic dataset.

(C) Correlation between damage levels of *C. tetani draft genomes* and corresponding human mtDNA from the same sample.

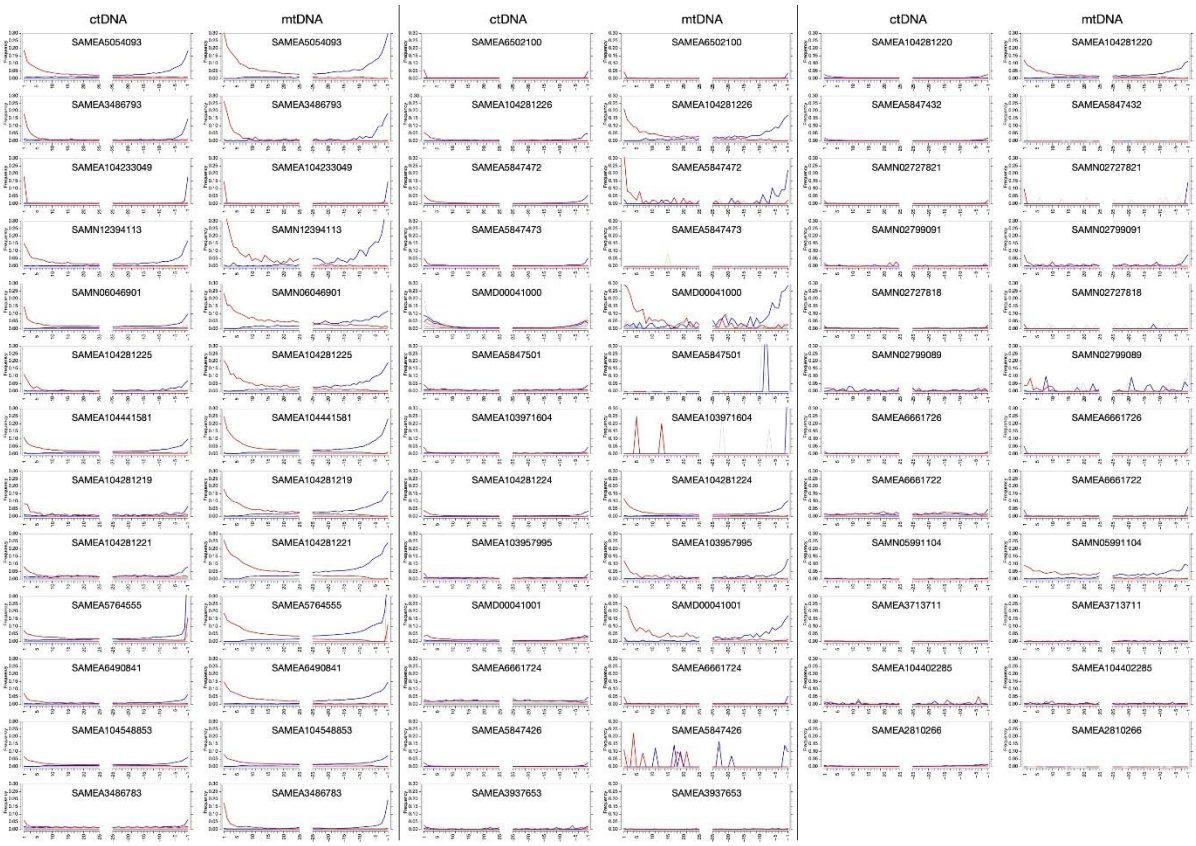

Figure 5 : MapDamage profiles depicting misincorporation levels for the first and last 25 bases of *C. tetani* and human mtDNA fragments from 38 ancient DNA samples.

G-to-A misincorporations (blue); C-to-T misincorporations (red); nucleotide-to-gap misincorporations (green). The top of each column is labeled by DNA type (columns 1,3,5 – *C. tetani*; columns 2,4,6 – human mtDNA) and each plot has been labeled according to its BioSample ID. Many ancient samples show a characteristic pattern of increased C-to-T mutations at the 5′ end and complementary G-to-A mutations at the 3′ end of sequence fragments due to cytosine deamination of 5′ overhanging ends.

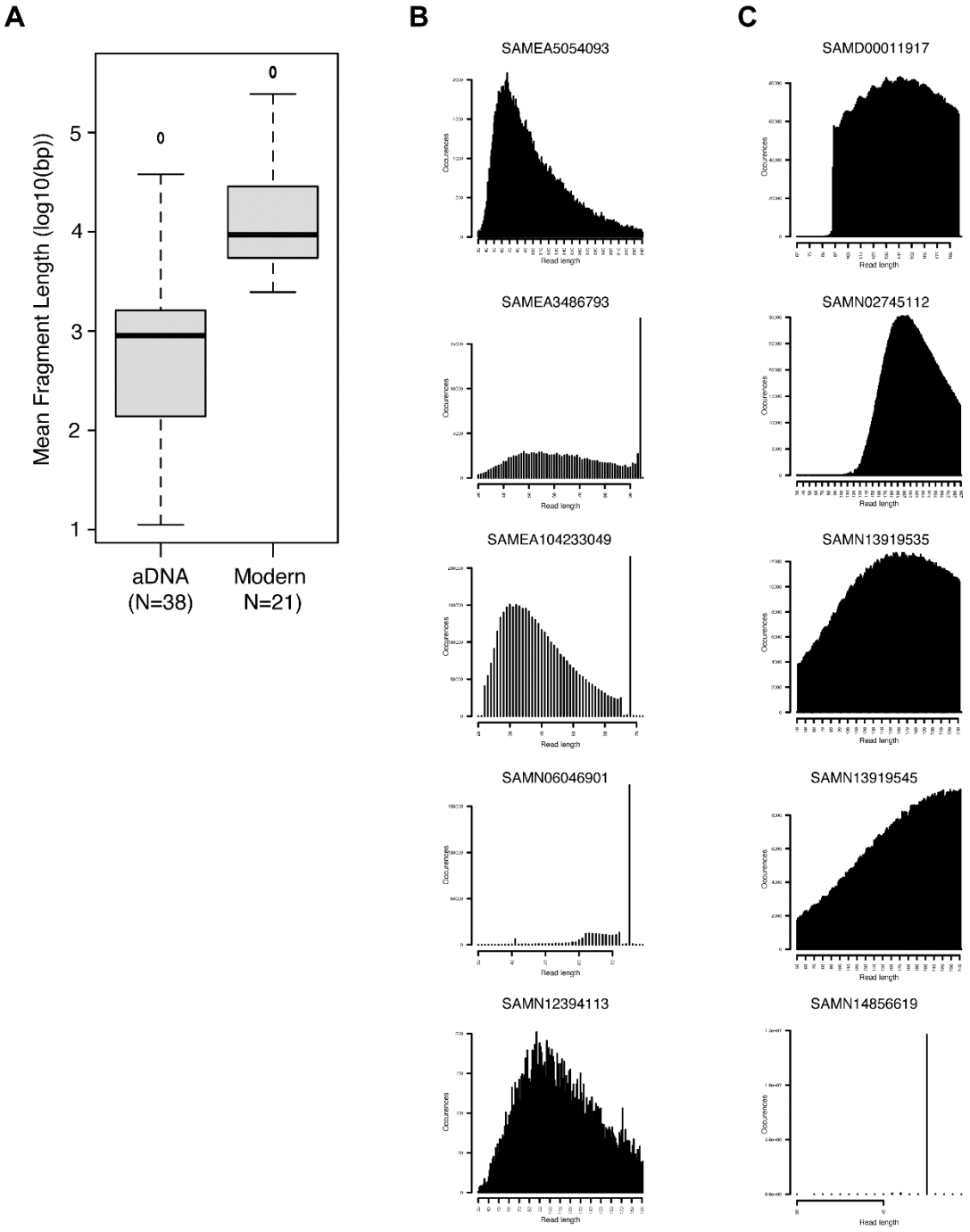

Figure 6 : Comparison of fragment length distributions for *C. tetani* contigs from ancient DNA datasets and modern datasets.

(A) boxplot depicting fragment length distributions for reads mapped to 38 *C. tetani draft genomes* and 21 modern *C. tetani* draft genomes. Fragment lengths were computed for each sample, combining information from both strands, using MapDamage2.

(B) Fragment length distributions of reads mapped to *C. tetani* draft genomes for the top five ancient samples based on damage level (misincorporation frequency).

(C) Fragment length distributions for five random modern *C. tetani* draft genomes metagenomically assembled from modern *C. tetani* sequencing datasets from the SRA.

2.1.3 Identification of novel *C. tetani* lineages and Clostridium species

A whole genome single nucleotide polymorphism (SNP) based phylogeny was constructed using fasttree (Price *et al.* 2010) using 33 *C. tetani* draft genomes and 37 known modern *C. tetani* genomes (Chapeton-Montes *et al.* 2019) (Figure 7A). Five *C. tetani* draft genomes were omitted due to extremely low (<1%) genome coverage, which could result in phylogenetic artifacts. *C. cochlearium* was included as a phylogenetic outgroup, as it is the closest known genomic relative to *C. tetani* (Rainey *et al.* 2015). The genome-based phylogeny which was produced (Figure 7A) is consistent with the expected phylogenetic structure, and contains all previously established *C. tetani* lineages (Chapeton-Montes *et al.* 2019). Twenty of the *C. tetani* draft genomes were assigned to existing *C. tetani* lineages (Figure 7A), including new members of clades 1B (N = 1), 1F (N = 1), 1H (N = 9), and 2 (N = 9), greatly expanding the known genomic diversity of clade 1H which previously contained a single strain and clade 2 which previously contained five strains (Figure 7A).

Four *C. tetani* draft genomes clustered within clade 1 but fell outside of established sublineages (Figure 7A). The remaining nine *C. tetani* draft genomes could not be assigned to any existing clade, and clustered as novel lineages (Figure 7A). One sample from the Canary Islands (Rodríguez-Varela *et al.* 2017) (dated to 936 BCE) forms a highly divergent lineage (labeled "Y") clustering outside all other *C. tetani* genomes. Based on the CheckM analysis, this *C. tetani* draft genome is of high quality with 74% completeness, 0.47% contamination and 0% strain heterogeneity (Supplementary Data). It exhibits an average nucleotide identity (ANI) of 87.5% to *C. tetani* E88, and 85.1% to *C. cochlearium*, below the 95% threshold typically used for species assignment (Warinner *et al.* 2017) (Supplementary Data). Eight *C. tetani* draft genomes form a novel clade (labeled lineage "X"), which

clustered outside of the entire *C. tetani* tree. These samples are exclusively of European origin and span a timeframe of 2290BC to 1722AD. The highest quality *C. tetani* draft genome for this clade is from another *Y. pestis* tooth sample from Germany (Andrades Valtueña *et al.* 2017) (circa 4203 BP) and has 59.8% completeness, and 5.56% contamination and strain heterogeneity as calculated by CheckM (Supplementary Data). Further comparison of clade X *C. tetani* draft genomes to other Clostridium species revealed that they are closest to *C. tetani* (86.33 +- 1.78 average nucleotide identity to E88 strain) and *C. cochlearium* (ANI = 85.16 +- 1.61) (Supplementary Data, Figure 7D). Therefore, clade X is potentially a novel species of *Clostridium* related to *C. tetani* which includes potentially neurotoxigenic strains (Figure 7A).

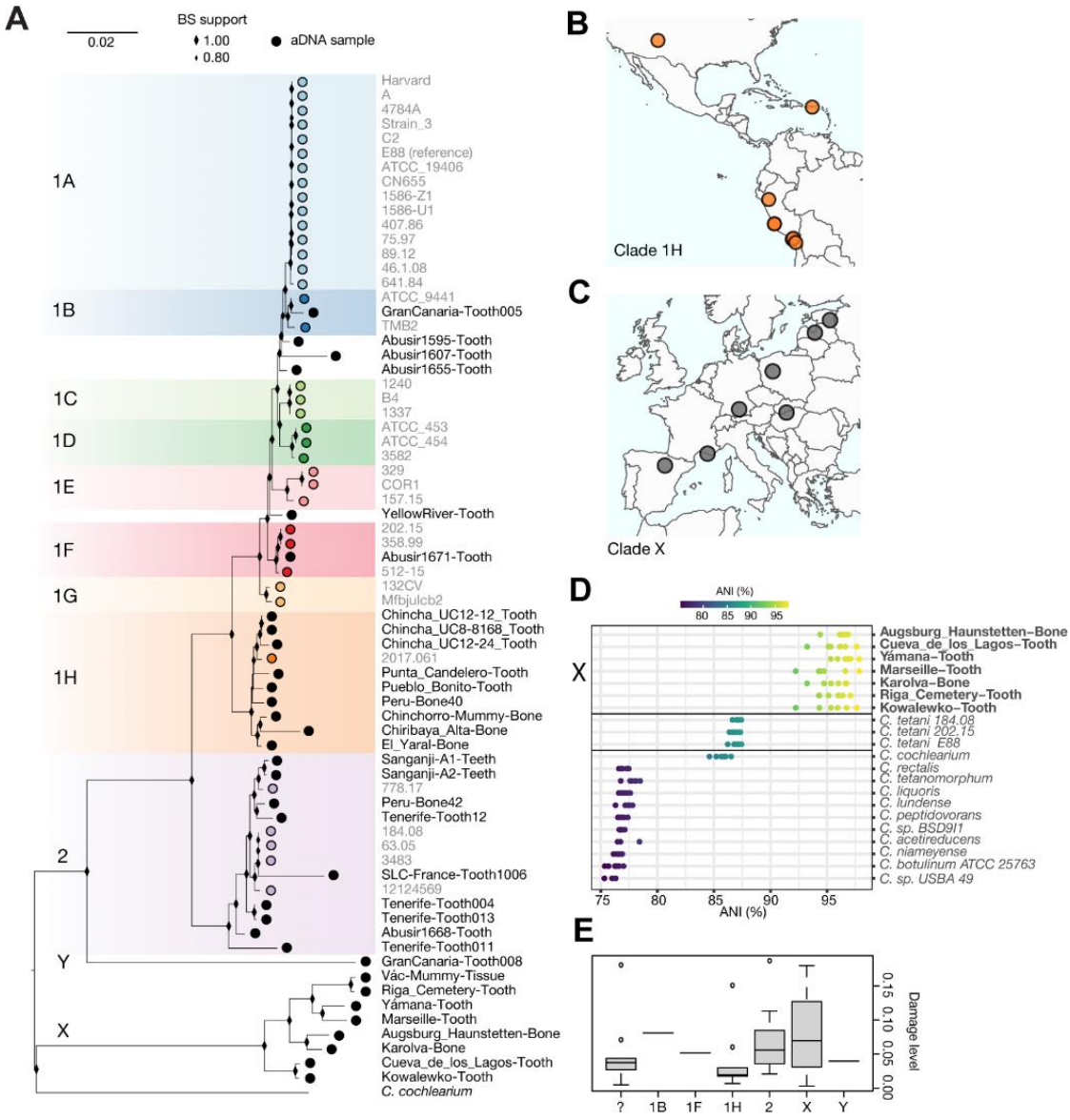

Figure 7 : Phylogenetic analysis reveals known and novel lineages of *C. tetani* in ancient DNA

(A) Whole genome phylogenetic tree of draft *C. tetani* genomes from ancient samples and modern *C. tetani* genomes along with previously labeled phylogenetic lineages. Novel lineages are labeled "X" and "Y", which are phylogenetically distinct from existing *C. tetani* genomes.

(B) Geographic clustering of newly identified lineage 1H for *C. tetani* draft genomes in ancient samples from the Americas.

(C) Geographic clustering of newly identified clade X species in ancient archaeological samples from Europe.

(D) ANI between clade X *C. tetani* draft genomes recovered from archeological samples and genomes of modern *Clostridium* species. Clade X *C. tetani* draft genomes show the highest ANI to *C. tetani* and *C. cochlearium* at a level that is sufficient to classify them as a novel *Clostridium* species. Note that one sample (from a mummy in Hungary (circa 1787 CE) (Kay *et al.* 2015)) was removed due to insufficient data required for fastANI. See Supplementary Data for ANI values and genome IDs.

(E) Distributions of damage levels for draft *C. tetani* genomes from each phylogenetic group

Individual maximum-likelihood phylogenies were also built using the ribosomal marker genes *rpsL*, *rpsG* and *recA*. As in the genome-wide tree, the gene phylogenies subdivided into two major clades (1 and 2) and the newly discovered lineage X and Y clustered as divergent lineages. The individual gene phylogenies therefore support the topology of the genome-based tree and reinforce clades X and Y as divergent *C. tetani*-related lineages.

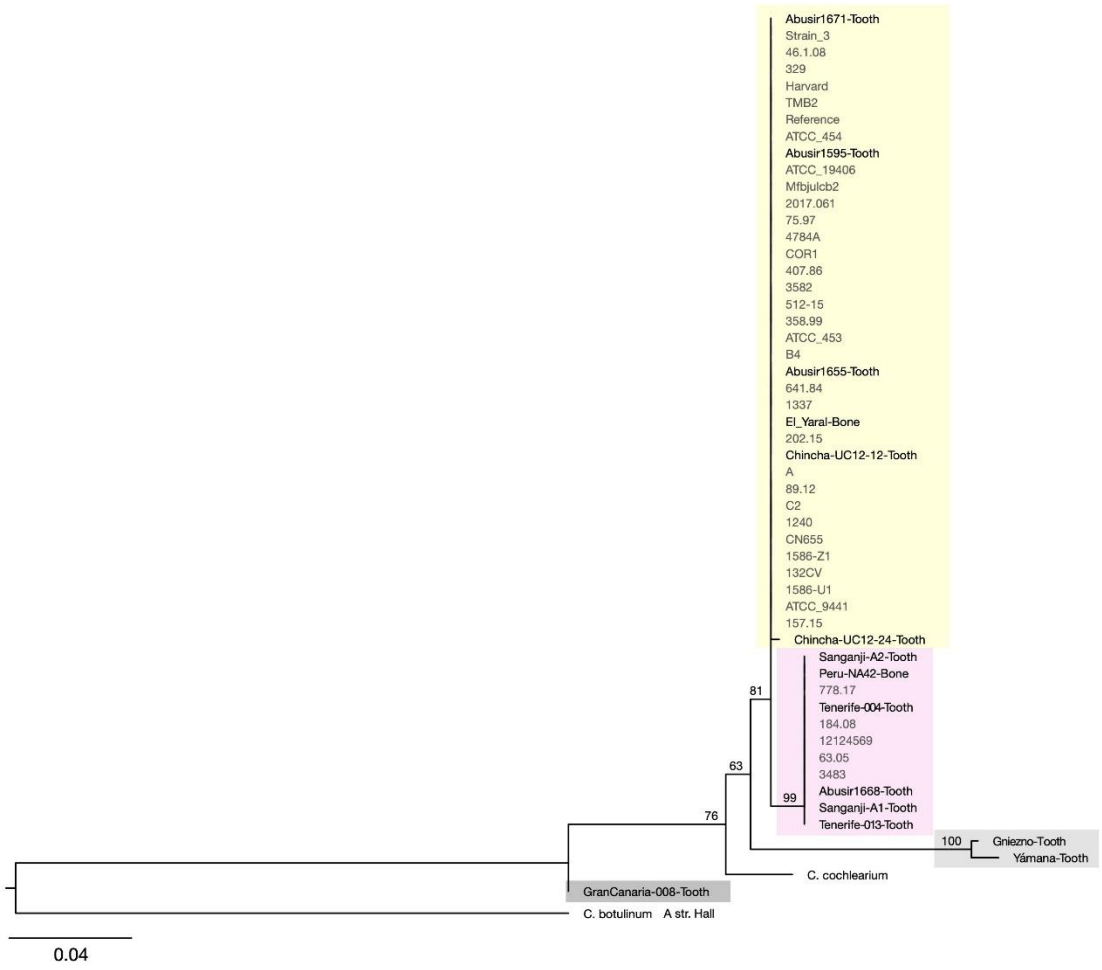

Figure 8 : Phylogenetic tree of *rpsL* coding sequences. Trees are based on a blastn search with *Clostridium tetani* E88 sequences. The phylogeny is based on a multiple alignment of *rpsL* (AE015927.1:c2752816-2752442) sequences identifed from aDNA *C. tetani* contigs and modern *C. tetani* strains. Genome IDs of modern *C. tetani* strains are listed in Chapeton-Montes *et al*. (2019). Sequences with 80% or greater coverage of the *C. tetani* E88 query sequences were aligned with MUSCLE v3.8.31, and RAxML (v8.2.4) trees using the GTR+GAMMA model were created. Bootstrap values (based on 100 runs) are displayed for major clades. The tree is highlighted based on Figure 7A as follows: GranCanaria-Tooth008 (dark grey) and clades 1 (yellow), 2 (purple), and X (light grey).

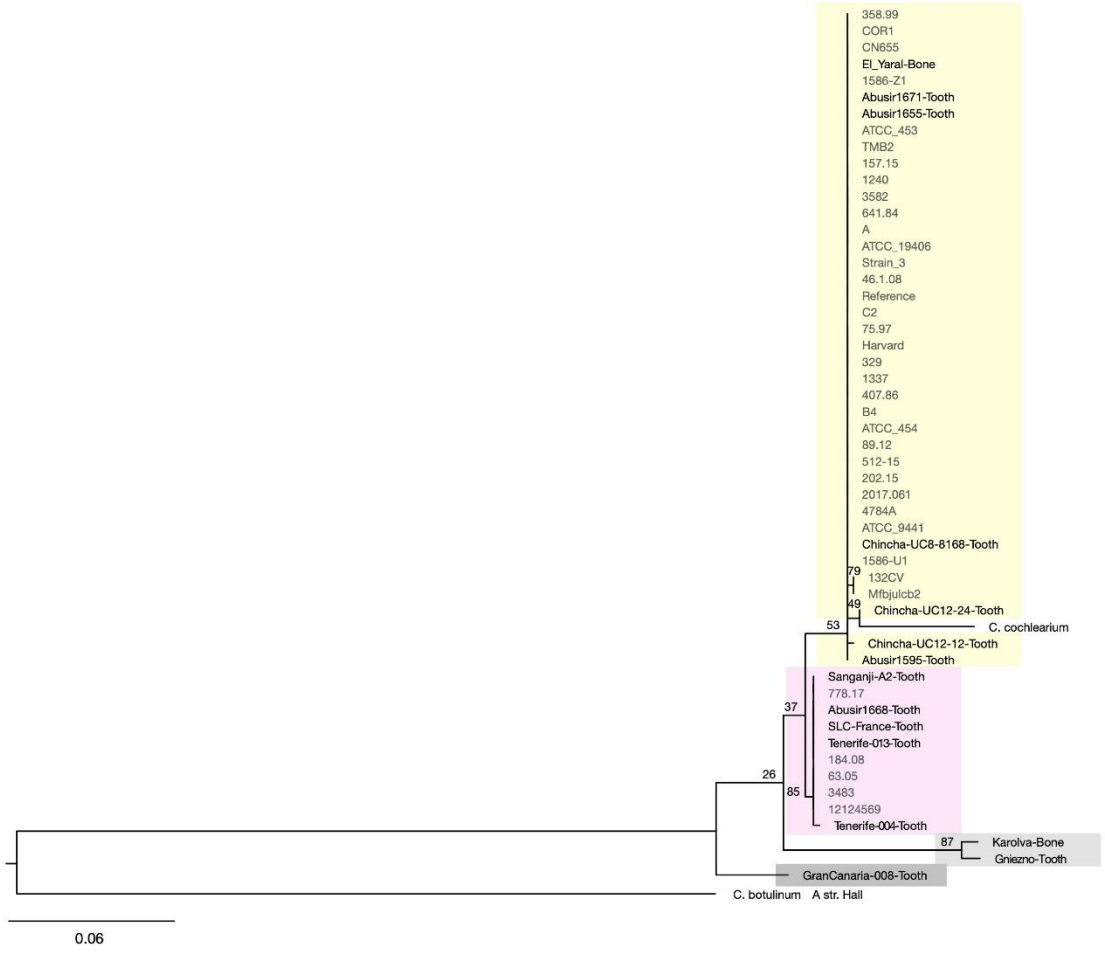

Figure 9 : Phylogenetic trees of *rpsG* coding sequences. Trees are based on a blastn search with *Clostridium tetani* E88 sequences. The phylogeny is based on a multiple alignment of *rpsG* (AE015927.1:c2752268-2751801) sequences identified from aDNA *C. tetani* contigs and modern *C. tetani* strains. Genome IDs of modern *C. tetani* strains are listed in Chapeton-Montes *et al*. (2019). Sequences with 80% or greater coverage of the *C. tetani* E88 query sequences were aligned with MUSCLE v3.8.31, and RAxML (v8.2.4) trees using the GTR+GAMMA model were created. Bootstrap values (based on 100 runs) are displayed for major clades. The tree is highlighted based on Figure 7A as follows: GranCanaria-Tooth008 (dark grey) and clades 1 (yellow), 2 (purple), and X (light grey).

Figure 10 : Phylogenetic trees of recA coding sequences. Trees are based on a blastn search with *Clostridium tetani* E88 sequences. The phylogeny is based on a multiple alignment of *recA* (AE015927.1:1383544-1384548) sequences identified from aDNA *C. tetani* contigs and modern *C. tetani* strains. Genome IDs of modern *C. tetani* strains are listed in Chapeton-Montes *et al*. (2019). Sequences with 80% or greater coverage of the *C. tetani* E88 query sequences were aligned with MUSCLE v3.8.31, and RAxML (v8.2.4) trees using the GTR+GAMMA model were created. Bootstrap values (based on 100 runs) are displayed for major clades. The tree is highlighted based on Figure 7A as follows: GranCanaria-Tooth008 (dark grey) and clades 1 (yellow), 2 (purple), and X (light grey).

2.1.4 Identification and experimental testing of a novel tetanus neurotoxin

A total of 20 draft *tent* gene sequences were assembled: six with complete coverage, and fourteen with 75-99.9% coverage (Supplementary Data). Four are identical to modern *tent* sequences, while 16 (including two identical sequences) are novel *tent* variants with 99.1-99.9% nucleotide identity to the E88 *tent*, which is comparable to the variation seen among modern *tent* genes (98.6-100%). A phylogeny was built using Fastree (Price *et al.* 2010) using, the 20 *tent* draft genes and all 12 modern *tent* sequences, which clustered sequences into four distinct subgroups (Figure 11A). Consistent with the novel clostridial lineages in the SNP-based phylogeny, the *tent* gene phylogeny revealed novel lineages of *tent* that are exclusive to the archeological samples. The *tent* genes clustered into four subgroups (Figure 11A) with both modern *tent* and several draft *tent* genes found in subgroups 1 and 3, and the remaining draft *tent* genes forming novel subgroups '2' and '4'. The *tent* sequence from an early Neolithic tooth from Spain (Valdiosera *et al.* 2018) (*C. tetani* clade "X") is the exclusive member of *tent* subgroup '4', and three *tent* sequences from clade 1H aDNA strains form the novel *tent* subgroup '2'.

The uniqueness of aDNA-associated *tent* genes was visualized by mapping nucleotide substitutions onto the phylogeny (Figure 11B) and focusing on "unique" *tent* substitutions found only in ancient samples and not in modern *tent* sequences. A total of 54 such substitutions were identified that are completely unique to one or more aDNA-associated *tent* genes (Figure 11B), which were statistically supported by a stringent variant calling pipeline (Supplementary Data). Interestingly, the largest number of unique substitutions occurred in *tent* subgroup '2'. *tent*/Chinchorro, from the oldest sample in the dataset (from a Chilean mummy bone sample, cira 3,889 BCE (Raghavan *et al.* 2015)), possesses 18 unique substitutions not found in modern *tent*, and 12 of these are shared with *tent*/El-Yaral and 10 with *tent*/Chiribaya (Figure 11B). The three associated draft *C. tetani* genomes also cluster as neighbors in the phylogenomic tree (Figure 7A), and the three associated archaeological samples are from a similar geographic region. These shared patterns suggest a common origin for these *C. tetani* strains and their unique neurotoxin genes and highlights *tent* subgroup 2 as a distinct group of *tent* variants exclusive to ancient samples (Figure 7A).

tent/Chinchorro was used as a representative sequence of this group as its full-length gene sequence could be completely assembled. The 18 unique substitutions present in the *tent*/Chinchorro gene result in 12 unique amino acid substitutions, absent from modern TeNT protein sequences (L140S, E141K, P144T, S145N, A147T, T148P, T149I, P445T, P531Q, V653I, V806I, H924R) (table S14).

Seven of these substitutions are spatially clustered within a surface loop on the TeNT structure and represent a potential mutation “hot spot” (Figure 11C). Interestingly, 7/12 amino acid substitutions found in TeNT/Chinchorro are also shared with TeNT/El-Yaral and 5/12 are shared with TeNT/Chiribaya (table S14). As highlighted in Figure 11C, TeNT/Chinchorro and TeNT/El-Yaral share a divergent 9-aa segment (amino acids 141-149 in TeNT, P04958) that is distinct from all other TeNT sequences. Reads mapping to the *tent*/Chinchorro gene show a low damage level similar to that seen in the *C. tetani* contigs, and it is weaker than the corresponding damage pattern from the associated human mitochondrial DNA (Figure 11D).

Given the phylogenetic novelty and unique pattern of substitutions observed for the *tent*/Chinchorro gene, it was of interest to determine whether it encodes an active tetanus neurotoxin. For biosafety reasons, the production of a *tent*/Chinchorro gene construct was avoided and instead sortase-mediated ligation was used to produce limited quantities of full-length protein toxin, as was done previously for other neurotoxins (Zhang *et al.* 2017b, 2018). The resulting full-length TeNT/Chinchorro protein cleaved the canonical substrate, VAMP2, in cultured rat cortical neurons, and can be neutralized with anti-TeNT anti-sera (Figure 11E, fig. S13). TeNT/Chinchorro induced spastic paralysis in vivo in mice when injected to the hind leg muscle and displayed a classic tetanus-like phenotype similar to that seen for wild-type TeNT (Figure 11F). Quantification of muscle rigidity following TeNT and TeNT/Chinchorro exposure demonstrated that TeNT/Chinchorro exhibits a level of potency that is indistinguishable from the TeNT (Figure 11G).

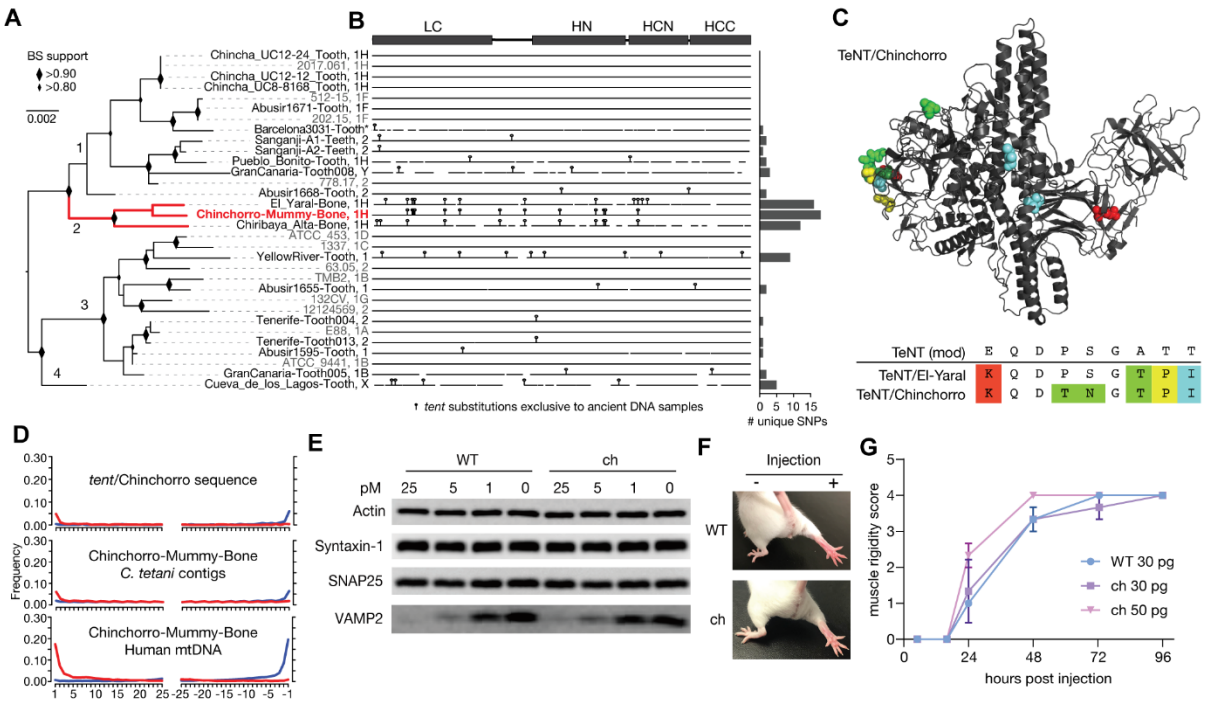

Figure 11 : Analysis and experimental testing of a novel TeNT **lineage identified from ancient DNA.**

(A) Maximum-likelihood phylogenetic tree of *tent* genes including novel *tent* sequences assembled from ancient DNA samples and a non-redundant set of *tent* sequences from existing strains in which duplicates have been removed (see Methods for details). The phylogeny has been subdivided into four subgroups. Sequences are labeled according to sample followed by their associated clade in the genome-based tree (Fig. 2), except for the Barcelona3031-Tooth sequence (*) as it fell below the coverage threshold.

(B) Visualization of *tent* sequence variation, with vertical bars representing nucleotide substitutions found uniquely in *tent* sequences from ancient DNA samples. On the right, a barplot is shown that indicates the number of unique substitutions found in each sequence, highlighting the uniqueness of subgroup 2.

(C) Structural model of TeNT/Chinchorro indicating all of its unique amino acid substitutions, which are not observed in modern TeNT sequences. Also shown is a segment of the translated alignment for a specific N-terminal region of the TeNT protein (residues 141-149, uniprot ID P04958). This sub-

alignment illustrates a segment containing a high density of unique amino acid substitutions, four of which are shared in TeNT/El-Yaral and TeNT/Chinchorro.

(D) MapDamage2 analysis of the *tent*/Chinchorro gene, and associated *C. tetani* contigs and mtDNA from the Chinchorro-Mummy-Bone sample.

(E) Cultured rat cortical neurons were exposed to full-length toxins in culture medium at indicated concentration for 12 hrs. Cell lysates were analyzed by immunoblot. WT TeNT and TeNT/Chinchorro ("ch") showed similar levels of activity in cleaving VAMP2 in neurons.

(F-G) Full-length toxins ligated by sortase reaction were injected into the gastrocnemius muscles of the right hind limb of mice. Extent of muscle rigidity was monitored and scored for 4 days (means ± se; n=3). TeNT/Chinchorro

Chapter 3
Discussion

In this work, large-scale data mining of millions of existing genomic datasets revealed the occurrence of neurotoxigenic *C. tetani* and related lineages of *Clostridium* in aDNA samples from human archaeological remains. This study has three main findings: 1) the first identification of neurotoxigenic *C. tetani* in archaeological samples including several *C. tetani* draft genomes of a potentially ancient origin; 2) the discovery of potentially novel lineages of *C. tetani* as well as a potentially new species of *Clostridium* (clade X); and 3) the identification of novel variants of a TeNT like toxin including TeNT/Chinchorro which was demonstrated to be an active neurotoxin with a potency comparable to modern TeNT.

This work is unique in several respects. Importantly, the recently developed STAT method (Katz *et al.* 2021) enabled a large-scale survey of all available DNA samples in the NCBI SRA, demonstrating the feasibility of discovering patterns across spatially and temporally diverse datasets. This study did not specifically target *C. tetani* in ancient samples, but rather this came as an unexpected finding from the results of the large-scale screen.

Also unexpected was the considerable diversity of ancient samples in which neurotoxigenic *C. tetani* was identified. This suggests that there may be an association between this organism (and related species) and human archaeological samples. Although 242 of the 43,620 datasets listed in the primary search results have ‘fossil metagenome’ as their target organism, only 10 of the final 76 datasets list their target organism as ‘fossil metagenome’, with many having a k-mer total-count under 100 indicating that they are either false positives or too fragmented properly assemble. Thus, it is not clear how many of these datasets with large amounts of predicted *C. tetani* DNA present in the SRA are from archeological samples and further study is needed.

Despite the abundance of environmental (e.g., soil metagenomic) samples in the SRA, these samples did not come to the surface of our genomic screen for *C. tetani*. This is consistent with the idea that, although *C. tetani* spores may be ubiquitous in terrestrial environments such as soil (Popoff 2020), these spores may be rare and so *C. tetani* DNA may not regularly appear at appreciable levels in shotgun metagenomes. However, the top 5 target organisms found in our search are ‘human gut metagenome’ (6744 datasets), ‘gut metagenome’ (3618 datasets), ‘Homo sapiens’ (3412 datasets),

‘soil metagenome’ (2833 datasets), and ‘metagenome’ (2456 datasets) accounting for 44% of the all the datasets predicted to contain *C. tetani* DNA. Conversely, many of these are likely false positives or too fragmented to assemble as only including datasets with a k-mer total-count >= 1000 results in the top 5 target organisms being ‘Homo sapiens’ (289), ‘Equus caballus’ (40), ‘gut metagenome’ (35), ‘Yersinia pestis’ (28), and ‘Clostridium tetani’ (26) accounting for 46% of the datasets with a k-mer total-count >= 1000. Further study is needed to fully determine how likely *C. tetani*, and in particular, neurotoxogenic *C. tetani* is in various environments.

The majority (31/38) of the aDNA in this current study was extracted from teeth. This is expected as teeth are commonly used in aDNA studies due to the survival and concentration of endogenous aDNA content (Adler *et al.* 2011). However, finding high levels of *C. tetani* DNA in teeth is unexpected should be explored further.

It is important to point out, that unlike other examples of ancient pathogens such as *M. tuberculosis* or *Y. pestis*, the identification of neurotoxigenic *C. tetani* in aDNA samples alone is insufficient to implicate tetanus as the cause of death or even to suggest that the corresponding *C. tetani* strains are contemporaneous with the archaeological samples. A variety of environmental factors and mechanisms may account for the presence of neurotoxigenic clostridia in aDNA samples, including the possibility of post-mortem colonization by environmental clostridia (Philips *et al.* 2017; Arning and Wilson 2020). This explanation may account for the observation of low *C. tetani* damage rates but high human mtDNA rates in some samples. For other samples, the *C. tetani* damage levels (>10%) are indicative of an ancient origin, but it is unknown whether these strains are the result of ancient sample colonization, or whether they are as old as the archaeological samples themselves.

Several of the samples display signs of damage consistent with UDG or partial UDG treatment. Additionally, although the length of the reads mapped to the *C. tetani* draft contigs are statistically shorter than those from modern *C. tetani*, several of them do not follow a gaussian or log tailing distribution of length versus the number of reads, with several having a spike in the number of longer reads. This should be further investigated to rule out assembly or mapping errors.

Regardless of whether the identified *C. tetani* genomes are contemporaneous with the archaeological samples, an important finding of this work is the substantial expansion of the genomic knowledge surrounding *C. tetani* and its relatives, such as the expansion of clade 2 and clade 1H, as well as the discovery of lineages X and Y. Lineage 1H in particular has undergone the greatest

expansion through the newly identified draft *C. tetani* genomes, from one known sample derived from a patient in France in 2016 (Chapeton-Montes *et al.* 2019), to 9 additional draft genomes assembled from ancient DNA. This may indicate that a broader diversity of 1H strains exists in under sampled environments. Interestingly, these newly identified lineage 1H strains share a common pattern of originating from the Americas, perhaps a common region-specific (or regionally abundant) environmental *C. tetani* strain colonized these samples at some point in the past. However, no *C. tetani* draft genomes were assembled with 100% coverage of the E88 reference chromosome and plasmid and thus errors in the phylogenetic analysis may have occurred.

In addition to the expansion of existing lineages, the genomic analysis revealed two highly unique lineages of *Clostridium* that are closely related to, but distinct from, *C. tetani* and its nearest genomic neighbour *C. cochlearium*. One of these novel lineages ("Y") was assembled from an aDNA sample from the Canary Islands taken from an archeological specimen dated to 936 CE. As it clusters outside of the entire *C. tetani* tree based on three phylogenetic analyses, this may be a lineage derived from an ancient lineage of *C. tetani* that predates the emergence of clade 1 and 2 genomes. Lineage Y also appears to be toxigenic, possessing a *tent* variant that has a unique substitution profile including unique substitutions not observed in any other *tent* sequences (modern or otherwise). However, as only 98 out of the approximately 311 member of the genus *Clostridium* have genomes available for comparison, it is possible that this a previously known but unsequenced member of the genus.

Perhaps even more intriguing is clade "X", a group of closely related *Clostridium* strains that also formed a sister lineage to *C. tetani* and yet resemble no other species that has been sequenced to date. This clade is unlikely to have arisen by errors in genome sequencing or assembly as it is supported by the co-clustering of multiple genomes as well as the consistently divergent placement of clade X species in ribosomal gene phylogenies. While the *C. tetani* draft genomes contain a *C. tetani*-like plasmid, and some strains (e.g., the sample from an early Neolithic tooth from Spain (Valdiosera *et al.* 2018)) appear to be toxin-encoding, it is important to note that *tent* gene was only recovered from this single clade X-associated sample, and with lower coverage relative to other plasmid-associated genes (*colT*). It is therefore possible that the apparent presence of the *tent* gene is due to contamination by other *C. tetani* (or unsequenced *Clostridium*) strains in this sample. Indeed, CheckM estimated 2.51% contamination, 12.5% of which was estimated to be due to strain variation. However, *de novo* assembly of DNA with a varying amount of damage has not yet been studied and thus it is possible that the draft genomes were incorrectly assembled from a mixture of *C. tetani* DNA

from several sources. Additionally, because individual datasets were combined by their associated bioSampleID, it is possible that some samples were incorrectly assembled. Further understanding of clade X may be addressed through future efforts to sequence more microbiomes associated with archaeological samples, as well as environmental *Clostridium* isolates.

Beyond expanding *C. tetani* and *Clostridium* genomic diversity, this work also expands the known diversity of clostridial neurotoxins which are the most potent family of toxins known to science (Rossetto and Montecucco 2019). Analysis of DNA from archeological samples revealed potential variants and lineages of TeNT, including the newly identified "subgroup 2" toxins: TeNT/Chinchorro, TeNT/El-Yaral toxins, and TeNT/Chiribaya-Alta. Not only do these toxins share a similar SNP profile, but they are derived from a similar geographic area (regions of Peru and Chile in South America) and their associated draft *C. tetani* genomes also cluster phylogenetically as the closest neighbors. Of the three subgroup 2 *tent* sequences identified, one of them (*tent*/Chinchorro) had sufficient coverage to be fully assembled and was also the most divergent from modern *tent* sequences with the greatest number of unique substitutions. Despite being the most divergent *tent*, reads mapping to the *tent*/Chinchorro gene, along with the associated *C. tetani* draft genome did not show strong patterns of DNA damage, and the damage level was weaker than that for human mtDNA. This suggests that, despite originating from the oldest sample in this study and possessing a unique *tent* variant, it is possible that the Chinchorro mummy associated *C. tetani* DNA is from a relatively "newer" strain that colonized or contaminated the sample post-mortem. Or they may from an older *C. tetani* strain which had its DNA protected and thus has less damage.

Due to the uniqueness of TeNT/Chinchorro, and its collection of amino acid substitutions that were not observed in any modern TeNT variants, it was of interest to determine whether this TeNT variant is a functional neurotoxin. A lack of toxicity might indicate a sequencing or assembly artifact or even a TeNT variant that targets other species. Therefore, a previous approach based on sortase-mediated ligation was used to produce small quantities of the full-length protein toxin (Zhang *et al.* 2017b, 2018). The recombinant protein produced a classic tetanus phenotype in mouse assays and exhibited a potency comparable to modern TeNT while also cleaving VAMP2, the canonical substrate of the TeNT. This suggests that the recombinant protein is neurotoxic and that its multiple unique amino acid substitutions have a limited impact on its potency and neurotoxicity. Such substitutions may alter yet-to-be identified TeNT protein-protein interactions.

Conclusions

Using large-scale data mining, evidence of neurotoxigenic *Clostridium* was identified in archeological samples. This resulted in a substantial expansion of the known genomic diversity and occurrence of *C. tetani* and led to the discovery of potentially novel *C. tetani* lineages, and *Clostridium* species, and tetanus like neurotoxins with functional activity. The discovery of neurotoxigenic clostridial genomes in such a wide diversity of ancient samples, both geographically and temporally, is unexpected, but perhaps not inconsistent with prior hypotheses about the role of these organisms in the natural decomposition process (Montecucco and Rasotto 2015; Javan *et al.* 2017; Mansfield and Doxey 2018). Although the precise origin of this DNA in ancient samples remains difficult to determine, future exploration of these and additional ancient archaeological samples will shed further light on the genomic and functional diversity of these fascinating organisms, as well as the ecological origins of their remarkably potent neurotoxins.

Open Problems and Future Work

This work has demonstrated the potential for using the STAT analysis results in combination with high performance computing to explore the genomic diversity of previously unstudied species. A future study should be done with additional *C. tetani* datasets (by lowering the k-mer total-count threshold) and using SRA BLAST to predetermine if additional samples have sufficient coverage to warrant further analysis. All datasets analyzed should be tested to see if DNA damage signals are present which may indicate the presence of ancient DNA.

A future study should also be performed to test the effect of mixed modern and ancient DNA on read mapping and *de novo* assembly. This will help determine the limits of these algorithms when used to study ubiquitous but less frequently studied organisms such as *C. tetani.*

References

Aboudharam G. 2016. Sources of Materials for Paleomicrobiology. *Microbiology Spectrum* 4: 4.4.52.

Adler C.J., Haak W., Donlon D. & Cooper A. 2011. Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science* 38: 956–964.

Allentoft M.E., Collins M., Harker D., Haile J., Oskam C.L., Hale M.L., Campos P.F., Samaniego J.A., Gilbert M.T.P., Willerslev E., Zhang G., Scofield R.P., Holdaway R.N. & Bunce M. 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* 279: 4724–4733.

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Andersen K., Bird K.L., Rasmussen M., Haile J., Breuning-Madsen H., Kjaer K.H., Orlando L., Gilbert M.T.P. & Willerslev E. 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology* 21: 1966–1979.

Andrades Valtueña A., Mittnik A., Key F.M., Haak W., Allmäe R., Belinskij A., Daubaras M., Feldman M., Jankauskas R., Janković I., Massy K., Novak M., Pfrengle S., Reinhold S., Šlaus M., Spyrou M.A., Szécsényi-Nagy A., Tõrv M., Hansen S., Bos K.I., Stockhammer P.W., Herbig A. & Krause J. 2017. The Stone Age Plague and Its Persistence in Eurasia. *Current Biology* 27: 3683-3691.e8.

Arning N. & Wilson D.J. 2020. The past, present and future of ancient bacterial DNA. *Microbial Genomics* 6.

Axelsson E., Willerslev E., Gilbert M.T.P. & Nielsen R. 2008. The Effect of Ancient DNA Damage on Inferences of Demographic Histories. *Molecular Biology and Evolution* 25: 2181–2187.

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A. & Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19: 455–477.

Barash J.R. & Arnon S.S. 2014. A Novel Strain of Clostridium botulinum That Produces Type B and Type H Botulinum Toxins. *The Journal of Infectious Diseases* 209: 183–191.

Benkert P., Biasini M. & Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics (Oxford, England)* 27: 343–350.

Boisvert S., Raymond F., Godzaridis É., Laviolette F. & Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology* 13: R122.

Borry M., Hübner A., Rohrlach A.B. & Warinner C. 2021. PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA *de novo* assembly. *PeerJ* 9: e11845.

Bos K., Kühnert D., Herbig A., Esquivel-Gomez L., Andrades Valtueña A., Barquera R., Giffin K., Kumar Lankapalli A., Nelson E., Sabin S., Spyrou M. & Krause J. 2019. Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. *Annual review of microbiology* 73: 639–666.

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

Briggs A., Stenzel U., Johnson P., Green R., Kelso J., Prüfer K., Meyer M., Krause J., Ronan M., Lachmann M. & Pääbo S. 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.

Brotherton P., Endicott P., Sanchez J.J., Beaumont M., Barnett R., Austin J. & Cooper A. 2007. Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research* 35: 5717–5728.

Bruggemann H., Baumer S., Fricke W., Wiezer A., Liesegang H., Decker I., Herzberg C., Martinez-Arias R., Merkl R., Henne A. & Gottschalk G. 2003. The genome sequence of Clostridium tetani, the causative agent of tetanus disease. *Proceedings of the National Academy of Sciences of the United States of America* 100: 1316–1321.

Brüggemann H., Brzuszkiewicz E., Chapeton-Montes D., Plourde L., Speck D. & Popoff M.R. 2015. Genomics of Clostridium tetani. *Research in Microbiology* 166: 326–331.

Cai S., Kumar R. & Singh B.R. 2021. Clostridial Neurotoxins: Structure, Function and Implications to Other Bacterial Toxins. *Microorganisms* 9: 2206.

Call J. e., Cooke P. h. & Miller A. j. 1995. In situ characterization of Clostridium botulinum neurotoxin synthesis and export. *Journal of Applied Bacteriology* 79: 257–263.

Campana M.G., Robles García N., Rühli F.J. & Tuross N. 2014. False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. *BMC Research Notes* 7: 111.

Campos P.F., Craig O.E., Turner-Walker G., Peacock E., Willerslev E. & Gilbert M.T.P. 2012. DNA in ancient bone – Where is it located and how should we extract it? *Annals of Anatomy - Anatomischer Anzeiger* 194: 7–16.

Cano R.J., Rivera-Perez J., Toranzos G.A., Santiago-Rodriguez T.M., Narganes-Storde Y.M., Chanlatte-Baik L., García-Roldán E., Bunkley-Williams L. & Massey S.E. 2014.

Paleomicrobiology: Revealing Fecal Microbiomes of Ancient Indigenous Cultures White B.A. (ed.). *PLoS ONE* 9: e106833.

Cano R.J., Tiefenbrunner F., Ubaldi M., Del Cueto C., Luciani S., Cox T., Orkand P., Künzel K.H. & Rollo F. 2000. Sequence analysis of bacterial DNA in the colon and stomach of the Tyrolean Iceman. *American Journal of Physical Anthropology* 112: 297–309.

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Peña A.G., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., Muegge B.D., Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters W.A., Widmann J., Yatsunenko T., Zaneveld J. & Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335–336.

Carle A. & Rattone G. 1884. Studio sperimentale sull'eziologia del tetano (Experimental studies of the etiology of tetanus). *Giorn. Accad. Med. Torino* 32: 174–179.

Carøe C., Gopalakrishnan S., Vinner L., Mak S.S.T., Sinding M.H.S., Samaniego J.A., Wales N., Sicheritz-Pontén T. & Gilbert M.T.P. 2018. Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution* 9: 410–419.

Carter G.P., Cheung J.K., Larcombe S. & Lyras D. 2014. Regulation of toxin production in the pathogenic clostridia. *Molecular Microbiology* 91: 221–231.

Castillo-Rojas G., Cerbón M.A. & López-Vidal Y. 2008. Presence of Helicobacter pylori in a Mexican Pre-Columbian Mummy. *BMC Microbiology* 8: 119.

Chapeton-Montes D., Plourde L., Bouchier C., Ma L., Diancourt L., Criscuolo A., Popoff M. & Brüggemann H. 2019. The population structure of Clostridium tetani deduced from its pan-genome. *Scientific reports* 9.

Chen S., Zhou Y., Chen Y. & Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34: i884–i890.

Cohen J.E., Wang R., Shen R.F., Wu W.W. & Keller J.E. 2017. Comparative pathogenomics of Clostridium tetani. *PLoS ONE* 12.

Cook T.M., Protheroe R.T. & Handel J.M. 2001. Tetanus: a review of the literature. *British Journal of Anaesthesia* 87: 477–487.

Cooke D., Wedge D. & Lunter G. 2021. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature biotechnology* 39: 885–892.

Cooper A. & Poinar H.N. 2000. Ancient DNA: Do It Right or Not at All. *Science* 289: 1139.

Cruz-Dávalos D.I., Llamas B., Gaunitz C., Fages A., Gamba C., Soubrier J., Librado P., Seguin-Orlando A., Pruvost M., Alfarhan A.H., Alquraishi S.A., Al-Rasheid K.A.S., Scheu A.,

Beneke N., Ludwig A., Cooper A., Willerslev E. & Orlando L. 2017. Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources* 17: 508–522.

Dabney J., Knapp M., Glocke I., Gansauge M.-T., Weihmann A., Nickel B., Valdiosera C., García N., Pääbo S., Arsuaga J.-L. & Meyer M. 2013a. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences* 110: 15758–15763.

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: 87–94.

Dabney J., Meyer M. & Paabo S. 2013b. Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology* 5: a012567–a012567.

Damgaard P.B., Margaryan A., Schroeder H., Orlando L., Willerslev E. & Allentoft M.E. 2015. Improving access to endogenous DNA in ancient bones and teeth. *Scientific Reports* 5: 11184.

Davenport E.R., Sanders J.G., Song S.J., Amato K.R., Clark A.G. & Knight R. 2017. The human microbiome in evolution. *BMC Biology* 15: 127.

DeSalle R., Gatesy J., Wheeler W. & Grimaldi D. 1992. DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. *Science* 257: 1933–1936.

Devault A.M., Golding G.B., Waglechner N., Enk J.M., Kuch M., Tien J.H., Shi M., Fisman D.N., Dhody A.N., Forrest S., Bos K.I., Earn D.J.D., Holmes E.C. & Poinar H.N. 2014a. Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849. *New England Journal of Medicine* 370: 334–340.

Devault A.M., McLoughlin K., Jaing C., Gardner S., Porter T.M., Enk J.M., Thissen J., Allen J., Borucki M., DeWitte S.N., Dhody A.N. & Poinar H.N. 2014b. Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. *Scientific Reports* 4: 4245.

Devault A.M., Mortimer T.D., Kitchen A., Kiesewetter H., Enk J.M., Golding G.B., Southon J., Kuch M., Duggan A.T., Aylward W., Gardner S.N., Allen J.E., King A.M., Wright G., Kuroda M., Kato K., Briggs D.E., Fornaciari G., Holmes E.C., Poinar H.N. & Pepperell C.S. 2017. A molecular portrait of maternal sepsis from Byzantine Troy. *eLife* 6: e20983.

Dong M., Masuyer G. & Stenmark P. 2019. Botulinum and Tetanus Neurotoxins. : 30.

Donoghue H.D., Spigelman M., Zias J., Gernaey-Child A.M. & Minnikin D.E. 1998. Mycobacterium tuberculosis complex DNA in calcified pleura from remains 1400 years old. *Letters in Applied Microbiology* 27: 265–269.

Drancourt M. 2016. Paleomicrobiology Data: Authentification and Interpretation. *Microbiology Spectrum* 4: 4.3.33.

Drancourt M., Aboudharam G. & Raoult D. 1998. Detection of 400-year-old Yersinia pestis DNA in human dental pulp: An approach to the diagnosis of ancient septicemia. *Proceedings of the National Academy of Sciences of the United States of America.* 95: 12637–12640.

Drancourt M. & Raoult D. 2004. Molecular detection of Yersinia pestis in dental pulp. *Microbiology* 150: 263–264.

Duchêne S., Ho S.Y.W., Carmichael A.G., Holmes E.C. & Poinar H. 2020. The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Current Biology* 30: R1215–R1231.

Eisenhofer R. & Weyrich L.S. 2019. Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ* 7: e6594.

Ewing B. & Green P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research* 8: 186–194.

Ewing B., Hillier L., Wendl M.C. & Green P. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research* 8: 175–185.

Fagernäs Z., García-Collado M.I., Hendy J., Hofman C.A., Speller C., Velsko I. & Warinner C. 2020. A unified protocol for simultaneous extraction of DNA and proteins from archaeological dental calculus. *Journal of Archaeological Science* 118: 105135.

FastANI. 2022.

Fellows Yates J.A., Lamnidis T.C., Borry M., Andrades Valtueña A., Fagernäs Z., Clayton S., Garcia M.U., Neukamm J. & Peltzer A. 2021. Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ* 9: e10947.

Gamba C., Hanghøj K., Gaunitz C., Alfarhan A.H., Alquraishi S.A., Al-Rasheid K.A.S., Bradley D.G. & Orlando L. 2016. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources* 16: 459–469.

Gansauge M.-T., Aximu-Petri A., Nagel S. & Meyer M. 2020. Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. *Nature Protocols* 15: 2279–2300.

Gansauge M.-T., Gerber T., Glocke I., Korlević P., Lippik L., Nagel S., Riehl L.M., Schmidt A. & Meyer M. 2017. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research* 45: e79.

Gansauge M.-T. & Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols* 8: 737–748.

Gansauge M.-T. & Meyer M. 2014. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Research* 24: 1543–1549.

Garrigues L., Do T.D., Bideaux C., Guillouet S.E. & Meynial-Salles I. 2022. Insights into Clostridium tetani: From genome to bioreactors. *Biotechnology Advances* 54: 107781.

gcloud CLI overview | Google Cloud CLI Documentation. *Google Cloud*.

Gilbert M.T.P., Binladen J., Miller W., Wiuf C., Willerslev E., Poinar H., Carlson J.E., Leebens-Mack J.H. & Schuster S.C. 2007. Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Research* 35: 1–10.

Gilbert M.T.P., 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: 341–354.

Glocke I. & Meyer M. 2017. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Research* 27: 1230–1237.

Golenberg E.M., Giannasit D.E., Clegg M.T., Durbin M., Henderson D. & Zurawski G. 1990. Chloroplast DNA sequence from a Miocene Magnolia species. 344: 3.

Gomes C., Palomo-Díez S., Roig J., López-Parra A.M., Baeza-Richer C., Esparza-Arroyo A., Gibaja J. & Arroyo-Pardo E. 2015. Nondestructive extraction DNA method from bones or teeth, true or false? *Forensic Science International: Genetics Supplement Series* 5: e279–e282.

Guellil M., Kersten O., Namouchi A., Bauer E.L., Derrick M., Jensen A.Ø., Stenseth N.C. & Bramanti B. 2018. Genomic blueprint of a relapsing fever pathogen in 15th century Scandinavia. *Proceedings of the National Academy of Sciences* 115: 10422–10427.

Guindon S. & Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52: 696–704.

Gurevich A., Saveliev V., Vyahhi N. & Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.

Hagelberg E., Sykes B. & Hedges R. 1989. Ancient bone DNA amplified. *Nature* 342: 485–485.

Haile J., Holdaway R., Oliver K., Bunce M., Gilbert M.T.P., Nielsen R., Munch K., Ho S.Y.W., Shapiro B. & Willerslev E. 2007. Ancient DNA Chronology within Sediment Deposits: Are Paleobiological Reconstructions Possible and Is DNA Leaching a Factor? *Molecular Biology and Evolution* 24: 982–989.

Hansen A.J., Mitchell D.L., Wiuf C., Paniker L., Brand T.B., Binladen J., Gilichinsky D.A., Rønn R. & Willerslev E. 2006. Crosslinks Rather Than Strand Breaks Determine Access to Ancient DNA Sequences From Frozen Sediments. *Genetics* 173: 1175–1179.

Harkins K.M., Schaefer N.K., Troll C.J., Rao V., Kapp J., Naughton C., Shapiro B. & Green R.E. 2020. A novel NGS library preparation method to characterize native termini of fragmented DNA. *Nucleic Acids Research* 48: e47.

Harney É., Cheronet O., Fernandes D.M., Sirak K., Mah M., Bernardos R., Adamski N., Broomandkhoshbacht N., Callan K., Lawson A.M., Oppenheimer J., Stewardson K., Zalzala F., Anders A., Candilio F., Constantinescu M., Coppa A., Ciobanu I., Dani J., Gallina Z., Genchi F., Nagy E.G., Hajdu T., Hellebrandt M., Horváth A., Király Á., Kiss K., Kolozsi B., Kovács P., Köhler K., Lucci M., Pap I., Popovici S., Raczky P., Simalcsik A., Szeniczey T., Vasilyev S., Virag C., Rohland N., Reich D. & Pinhasi R. 2021. A minimally destructive protocol for DNA extraction from ancient teeth. *Genome Research* 31: 472–483.

Hatheway C.L. 1990. Toxigenic clostridia. *Clinical Microbiology Reviews* 3: 66–98.

Herbig A., Maixner F., Bos K.I., Zink A., Krause J. & Huson D.H. 2016. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. : 050559.

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 Research* 38: e161.

Ho S.Y.W., Heupink T.H., Rambaut A. & Shapiro B. 2007. Bayesian Estimation of Sequence Damage in Ancient DNA. *Molecular Biology and Evolution* 24: 1416–1422.

Hofreiter M. 2012. Nondestructive DNA Extraction from Museum Specimens. In: Shapiro B. & Hofreiter M. (eds.), *Ancient DNA: Methods and Protocols*, Humana Press, Totowa, NJ, pp. 93–100.

Hofreiter M., Jaenicke V., Serre D., Haeseler A. von & Pääbo S. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* 29: 4793–4799.

Hoss M., Jaruga P., Zastawny T.H., Dizdaroglu M. & Paabo S. 1996. DNA Damage and DNA Sequence Retrieval from Ancient Tissues. *Nucleic Acids Research* 24: 1304–1307.

Höss M. & Pääbo S. 1993. DNA extraction from Pleistocene bones by a silica-based purification method. *Nucleic Acids Research* 21: 3913–3914.

Howorka S., Cheley S. & Bayley H. 2001. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature Biotechnology* 19: 636–639.

Hübler R., Key F.M., Warinner C., Bos K.I., Krause J. & Herbig A. 2019. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology* 20: 280.

Jain C., Rodriguez-R L.M., Phillippy A.M., Konstantinidis K.T. & Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 9: 5114.

Javan G.T., Finley S.J., Smith T., Miller J. & Wilkinson J.E. 2017. Cadaver Thanatomicrobiome Signatures: The Ubiquitous Nature of Clostridium Species in Human Decomposition. *Frontiers in Microbiology* 8.

Jones E.D. & Bösl E. 2021. Ancient human DNA: A history of hype (then and now). *Journal of Social Archaeology* 21: 236–255.

Jónsson H., Ginolhac A., Schubert M., Johnson P.L.F. & Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29: 1682–1684.

Katoh K. & Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30: 772–780.

Katz K., Shutov O., Lapoint R., Kimelman M., Brister J. & O'Sullivan C. 2021. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome biology* 22.

Katz K., Shutov O., Lapoint R., Kimelman M., Brister J.R. & O'Sullivan C. 2022. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Research* 50: D387–D390.

Kay G.L., Sergeant M.J., Giuffra V., Bandiera P., Milanese M., Bramanti B., Bianucci R. & Pallen M.J. 2014. Recovery of a Medieval Brucella melitensis Genome Using Shotgun Metagenomics Keim P.S. (ed.). *mBio* 5: e01337-14.

Kay G.L., Sergeant M.J., Zhou Z., Chan J.Z.-M., Millard A., Quick J., Szikossy I., Pap I., Spigelman M., Loman N.J., Achtman M., Donoghue H.D. & Pallen M.J. 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature Communications* 6: 6717.

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 in Genetics* 33: 508–520.

Kircher M., Sawyer S. & Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* 40: e3–e3.

Kircher M., Stenzel U. & Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biology* 10: R83.

Kirsanow K. & Burger J. 2012. Ancient human DNA. *Annals of Anatomy - Anatomischer Anzeiger* 194: 121–132.

Kitasato S. 1889. Ueber den Tetanusbacillus. *Z Hyg* 7: 225–234.

Korlević P., Gerber T., Gansauge M.-T., Hajdinjak M., Nagel S., Aximu-Petri A. & Meyer M. 2015. Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *BioTechniques* 59: 87–93.

Korlević P., Talamo S. & Meyer M. 2018. A combined method for DNA analysis and radiocarbon dating from a single sample. *Scientific Reports* 8: 4127.

Kumar R., Feltrup T.M., Kukreja R.V., Patel K.B., Cai S. & Singh B.R. 2019. Evolutionary Features in the Structure and Function of Bacterial Toxins. *Toxins* 11: 15.

Langmead B. & Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Li H. & Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li D., Liu C.M., Luo R., Sadakane K. & Lam T.W. 2014. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674–1676.

Li D., Luo R., Liu C.-M., Leung C.-M., Ting H.-F., Sadakane K., Yamashita H. & Lam T.-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102: 3–11.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. *Nature* 362: 709–715.

Lindahl T. & Andersson A. 1972. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry* 11: 3618–3623.

Lindahl T. & Nyberg B. 1972. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* 11: 3610–3618.

Linderholm A. 2021. Palaeogenetics: Dirt, what is it good for? Everything. *Current Biology* 31: R993–R995.

Louvel G., Der Sarkissian C., Hanghøj K. & Orlando L. 2016. metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data. *Molecular Ecology Resources* 16: 1415–1427.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.-W. & Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 2047-217X-1–18.

Maixner F., Krause-Kyora B., Turaev D., Herbig A., Hoopmann M.R., Hallows J.L., Kusebauch U., Vigl E.E., Malfertheiner P., Megraud F., O'Sullivan N., Cipollini G., Coia V., Samadelli M.,

Engstrand L., Linz B., Moritz R.L., Grimm R., Krause J., Nebel A., Moodley Y., Rattei T. & Zink A. 2016. The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* 351: 162–165.

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

Mansfield M.J., Adams J.B. & Doxey A.C. 2015. Botulinum neurotoxin homologs in non-Clostridium species. *FEBS letters* 589: 342–348.

Mansfield M.J. & Doxey A.C. 2018. Genomic insights into the evolution and ecology of botulinum neurotoxins. *Pathogens and Disease* 76: fty040.

Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.-J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., Irzyk G.P., Jando S.C., Alenquer M.L.I., Jarvie T.P., Jirage K.B., Kim J.-B., Knight J.R., Lanza J.R., Leamon J.H., Lefkowitz S.M., Lei M., Li J., Lohman K.L., Lu H., Makhijani V.B., McDade K.E., McKenna M.P., Myers E.W., Nickerson E., Nobile J.R., Plant R., Puc B.P., Ronan M.T., Roth G.T., Sarkis G.J., Simons J.F., Simpson J.W., Srinivasan M., Tartaro K.R., Tomasz A., Vogt K.A., Volkmer G.A., Wang S.H., Wang Y., Weiner M.P., Yu P., Begley R.F. & Rothberg J.M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380.

Martiniano R., Garrison E., Jones E.R., Manica A. & Durbin R. 2020. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology* 21: 250.

Martínez-Delclòs X., Briggs D.E.G. & Peñalver E. 2004. Taphonomy of insects in carbonates and amber. *Palaeogeography, Palaeoclimatology, Palaeoecology* 203: 19–64.

Mellanby J., Mellanby H., Pope D. & Heyningen W. Van. 1968. Ganglioside as a prophylactic agent in experimental tetanus in mice. *J Gen Microbiol* 54: 161–168.

Mendum T.A., Schuenemann V.J., Roffey S., Taylor G., Wu H., Singh P., Tucker K., Hinds J., Cole S.T., Kierzek A.M., Nieselt K., Krause J. & Stewart G.R. 2014. Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics* 15: 270.

Menzel P., Ng K. & Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications* 7.

Meyer M., Briggs A.W., Maricic T., Höber B., Höffner B., Krause J., Weihmann A., Pääbo S. & Hofreiter M. 2008. From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Research* 36: e5.

Meyer M. & Kircher M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols* 2010: pdb.prot5448.

Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., Filippo C. de, Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J. & Pääbo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338: 222–226.

Milos P.M. 2010. Helicos single molecule sequencing: unique capabilities and importance for molecular diagnostics. *Genome Biology* 11: I14.

Montecucco C. & Rasotto M.B. 2015. On Botulinum Neurotoxin Variability. *mBio* 6: e02131-14.

Mühlemann B., Jones T., Damgaard P., Allentoft M., Shevnina I., Logvin A., Usmanova E., Panyushkina I., Boldgiv B., Bazartseren T., Tashbaeva K., Merz V., Lau N., Smrčka V., Voyakin D., Kitov E., Epimakhov A., Pokutta D., Vicze M., Price T., Moiseyev V., Hansen A., Orlando L., Rasmussen S., Sikora M., Vinner L., Osterhaus A., Smith D., Glebe D., Fouchier R., Drosten C., Sjögren K., Kristiansen K. & Willerslev E. 2018. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* 557: 418–423.

Müller 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: 20133236.

Namiki T., Hachiya T., Tanaka H. & Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* 40: e155.

Namouchi A., Guellil M., Kersten O., Hänsch S., Ottoni C., Schmid B.V., Pacciani E., Quaglia L., Vermunt M., Bauer E.L., Derrick M., Jensen A.Ø., Kacki S., Cohn S.K., Stenseth N.C. & Bramanti B. 2018. Integrative approach using Yersinia pestis genomes to revisit the historical landscape of plague during the Medieval Period. *Proceedings of the National Academy of Sciences of the United States of America* 115: E11790–E11797.

Nayfach S., Rodriguez-Mueller B., Garud N. & Pollard K.S. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* 26: 1612–1625.

Neukamm J., Peltzer A. & Nieselt K. 2021. DamageProfiler: fast damage pattern calculation for ancient DNA. *Bioinformatics* 37: 3652–3653.

Nurk S., Meleshko D., Korobeynikov A. & Pevzner P.A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27: 824–834.

Oliva A., Tobler R., Cooper A., Llamas B. & Souilmi Y. 2021. Systematic benchmark of ancient DNA read mapping. *Briefings in Bioinformatics* 22: bbab076.

Orlando L., Allaby R., Skoglund P., Der Sarkissian C., Stockhammer P.W., Ávila-Arcos M.C., Fu Q., Krause J., Willerslev E., Stone A.C. & Warinner C. 2021. Ancient DNA analysis. *Nature Reviews Methods Primers* 1: 1–26.

Overballe-Petersen S., Orlando L. & Willerslev E. 2012. Next-generation sequencing offers new insights into DNA degradation. *Trends in Biotechnology* 30: 364–368.

Pääbo S. 1985. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* 314: 644–645.

Pääbo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences* 86: 1939–1943.

Pääbo S., Poinar H., Serre D., Jaenicke-Després V., Hebler J., Rohland N., Kuch M., Krause J., Vigilant L. & Hofreiter M. 2004. Genetic Analyses from Ancient DNA. *Annual Review of Genetics* 38: 645–679.

Pääbo S. & Wilson A.C. 1991. Miocene DNA sequences — a dream come true? *Current Biology* 1: 45–46.

Pagès H., Aboyoun P., Gentleman R. & DebRoy S. 2022. Biostrings: Efficient manipulation of biological strings.

Pappas G., Kiriaze I.J. & Falagas M.E. 2008. Insights into infectious disease in the era of Hippocrates. *International Journal of Infectious Diseases* 12: 347–350.

Parks D., Imelfort M., Skennerton C., Hugenholtz P. & Tyson G. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* 25: 1043–1055.

Parte A.C., Sardà Carbasse J., Meier-Kolthoff J.P., Reimer L.C. & Göker M. 2020. 2022. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology* 70: 5607–5612.

Peterson J.W. 1996. Bacterial Pathogenesis. In: Baron S. (ed.), *Medical Microbiology*, University of Texas Medical Branch at Galveston, Galveston (TX).

Peyrégne S. & Peter B.M. 2020. AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome Biology* 21: 246.

Philips A., Stolarek I., Kuczkowska B., Juras A., Handschuh L., Piontek J., Kozlowski P. & Figlerowicz M. 2017. Comprehensive analysis of microorganisms accompanying human archaeological remains. *GigaScience* 6: 1–13.

Poinar H.N. 2002. The Genetic Secrets Some Fossils Hold. *Accounts of Chemical Research* 35: 676–684.

Poinar H., Kuch M., McDonald G., Martin P. & Pääbo S. 2003. Nuclear Gene Sequences from a Late Pleistocene Sloth Coprolite. *Current Biology* 13: 1150–1152.

pontussk. 2021. PMDtools.

Popoff M. 2020. Tetanus in animals. *J Vet Diagn Invest* 32: 184–191.

Popoff M.R. & Bouvet P. 2009. Clostridial toxins. *Future Microbiology* 4: 1021–1064.

Price M., Dehal P. & Arkin A. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* 5.

Quail M.A., Smith M., Coupland P., Otto T.D., Harris S.R., Connor T.R., Bertoni A., Swerdlow H.P. & Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.

Raghavan M., Steinrücken M., Harris K., Schiffels S., Rasmussen S., DeGiorgio M., Albrechtsen A., Valdiosera C., Ávila-Arcos M.C., Malaspinas A.-S., Eriksson A., Moltke I., Metspalu M., Homburger J.R., Wall J., Cornejo O.E., Moreno-Mayar J.V., Korneliussen T.S., Pierre T., Rasmussen M., Campos P.F., Damgaard P. de B., Allentoft M.E., Lindo J., Metspalu E., Rodríguez-Varela R., Mansilla J., Henrickson C., Seguin-Orlando A., Malmström H., Stafford T., Shringarpure S.S., Moreno-Estrada A., Karmin M., Tambets K., Bergström A., Xue Y., Warmuth V., Friend A.D., Singarayer J., Valdes P., Balloux F., Leboreiro I., Vera J.L., Rangel-Villalobos H., Pettener D., Luiselli D., Davis L.G., Heyer E., Zollikofer C.P.E., Ponce de León M.S., Smith C.I., Grimes V., Pike K.-A., Deal M., Fuller B.T., Arriaza B., Standen V., Luz M.F., Ricaut F., Guidon N., Osipova L., Voevoda M.I., Posukh O.L., Balanovsky O., Lavryashina M., Bogunov Y., Khusnutdinova E., Gubina M., Balanovska E., Fedorova S., Litvinov S., Malyarchuk B., Derenko M., Mosher M.J., Archer D., Cybulski J., Petzelt B., Mitchell J., Worl R., Norman P.J., Parham P., Kemp B.M., Kivisild T., Tyler-Smith C., Sandhu M.S., Crawford M., Villems R., Smith D.G., Waters M.R., Goebel T., Johnson J.R., Malhi R.S., Jakobsson M., Meltzer D.J., Manica A., Durbin R., Bustamante C.D., et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349: aab3884.

Rainey F.A., Hollen B.J. & Small A.M. 2015. Clostridium. In: *Bergey's Manual of Systematics of Archaea and Bacteria*, pp. 1–122.

Raoult D., Aboudharam G., Crubézy E., Larrouy G., Ludes B. & Drancourt M. 2000. Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of Medieval Black Death. *Proceedings of the National Academy of Sciences* 97: 12800–12803.

Rasmussen S., Allentoft M.E., Nielsen K., Orlando L., Sikora M., Sjögren K.-G., Pedersen A.G., Schubert M., Van Dam A., Kapel C.M.O., Nielsen H.B., Brunak S., Avetisyan P., Epimakhov A., Khalyapin M.V., Gnuni A., Kriiska A., Lasak I., Metspalu M., Moiseyev V., Gromov A.,

Pokutta D., Saag L., Varul L., Yepiskoposyan L., Sicheritz-Pontén T., Foley R.A., Lahr M.M., Nielsen R., Kristiansen K. & Willerslev E. 2015. Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* 163: 571–582.

Renaud G., Hanghøj K., Willerslev E. & Orlando L. 2017. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33: 577–579.

Renaud G., Kircher M., Stenzel U. & Kelso J. 2013. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics* 29: 1208–1209.

Renaud G., Stenzel U. & Kelso J. 2014. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic acids research* 42: e141.

Rodríguez-Varela R., Günther T., Krzewińska M., Storå J., Gillingwater T.H., MacCallum M., Arsuaga J.L., Dobney K., Valdiosera C., Jakobsson M., Götherström A. & Girdland-Flink L. 2017. Genomic Analyses of Pre-European Conquest Human Remains from the Canary Islands Reveal Close Affinity to Modern North Africans. *Current Biology* 27: 3396-3402.e5.

Rohland N., Glocke I., Aximu-Petri A. & Meyer M. 2018. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nature Protocols* 13: 2447–2461.

Rohland N., Harney E., Mallick S., Nordenfelt S. & Reich D. 2015. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370: 20130624.

Rollo F., Luciani S., Marota I., Olivieri C. & Ermini L. 2007. Persistence and decay of the intestinal microbiota's DNA in glacier mummies from the Alps. *Journal of Archaeological Science* 34: 1294–1305.

Rossetto O. & Montecucco C. 2019. Tables of Toxicity of Botulinum and Tetanus Neurotoxins. *Toxins* 11: 686.

Sakaguchi G. 1982. Clostridium botulinum toxins. *Pharmacology & Therapeutics* 19: 165–194.

Santiago-Rodriguez T.M., Narganes-Storde Y.M., Chanlatte L., Crespo-Torres E., Toranzos G.A., Jimenez-Flores R., Hamrick A. & Cano R.J. 2013. Microbial Communities in Pre-Columbian Coprolites Hawks J. (ed.). *PLoS ONE* 8: e65191.

Sawyer S., Krause J., Guschanski K., Savolainen V. & Pääbo S. 2012. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* 7.

Scarsbrook L., Verry A.J.F., Walton K., Hitchmough R.A. & Rawlence N.J. 2022. Ancient mitochondrial genomes recovered from small vertebrate bones through minimally destructive DNA extraction: Phylogeography of the New Zealand gecko genus Hoplodactylus. *Molecular Ecology* n/a.

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J. & Weber C.F. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* 75: 7537–7541.

Schrödinger, LLC. 2015. The PyMOL Molecular Graphics System, Version 2.4.1.

Schubert M., Ginolhac A., Lindgreen S., Thompson J.F., AL-Rasheid K.A., Willerslev E., Krogh A. & Orlando L. 2012. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 13: 178.

Schubert M., Lindgreen S. & Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* 9: 88.

Schuenemann V.J., Lankapalli A.K., Barquera R., Nelson E.A., Hernández D.I., Alonzo V.A., Bos K.I., Morfín L.M., Herbig A. & Krause J. 2018. Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. *PLOS Neglected Tropical Diseases* 12: e0006447.

Schuenemann V., Singh P., Mendum T., Krause-Kyora B., Jäger G., Bos K., Herbig A., Economou C., Benjak A., Busso P., Nebel A., Boldsen J., Kjellström A., Wu H., Stewart G., Taylor G., Bauer P., Lee O., Wu H., Minnikin D., Besra G., Tucker K., Roffey S., Sow S., Cole S., Nieselt K. & Krause J. 2013. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science (New York, N.Y.)* 341: 179–183.

Sczyrba A., Hofmann P., Belmann P., Koslicki D., Janssen S., Dröge J., Gregor I., Majda S., Fiedler J., Dahms E., Bremges A., Fritz A., Garrido-Oter R., Jørgensen T.S., Shapiro N., Blood P.D., Gurevich A., Bai Y., Turaev D., DeMaere M.Z., Chikhi R., Nagarajan N., Quince C., Meyer F., Balvočiūtė M., Hansen L.H., Sørensen S.J., Chia B.K.H., Denis B., Froula J.L., Wang Z., Egan R., Don Kang D., Cook J.J., Deltel C., Beckstette M., Lemaitre C., Peterlongo P., Rizk G., Lavenier D., Wu Y.-W., Singer S.W., Jain C., Strous M., Klingenberg H., Meinicke P., Barton M.D., Lingner T., Lin H.-H., Liao Y.-C., Silva G.G.Z., Cuevas D.A., Edwards R.A., Saha S., Piro V.C., Renard B.Y., Pop M., Klenk H.-P., Göker M., Kyrpides N.C., Woyke T., Vorholt J.A., Schulze-Lefert P., Rubin E.M., Darling A.E., Rattei T. & McHardy A.C. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods* 14: 1063–1071.

Seemann T. 2022. Snippy.

Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O. & Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9: 811–814.

Seguin-Orlando A., Hoover C.A., Vasiliev S.K., Ovodov N.D., Shapiro B., Cooper A., Rubin E.M., Willerslev E. & Orlando L. 2015. Amplification of TruSeq ancient DNA libraries with

AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns. *STAR: Science & Technology of Archaeological Research* 1: 1–9.

Setlow P. 2007. I will survive: DNA protection in bacterial spores. *Trends in Microbiology* 15: 172–180.

Singh B.R. 2006. Botulinum neurotoxin structure, engineering, and novel cellular trafficking and targeting. *Neurotoxicity Research* 9: 73–92.

Sirak K.A., Fernandes D.M., Cheronet O., Novak M., Gamarra B., Balassa T., Bernert Z., Cséki A., Dani J., Gallina J.Z., Kocsis-Buruzs G., Kővári I., László O., Pap I., Patay R., Petkes Z., Szenthe G., Szeniczey T., Hajdu T. & Pinhasi R. 2017. A minimally-invasive method for sampling human petrous bones from the cranial base for ancient DNA analysis. *BioTechniques* 62: 283–289.

Skoglund P., Northoff B.H., Shunkov M.V., Derevianko A.P., Pääbo S., Krause J. & Jakobsson M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences* 111: 2229–2234.

Smith J.W.G. 1969. Diphtheria and tetanus toxoids. *British Medical Bulletin* 25: 177–182.

Smith T.J., Hill K.K. & Raphael B.H. 2015. Historical and current perspectives on Clostridium botulinum diversity. *Research in Microbiology* 166: 290–302.

Sobel J. 2005. Botulism. *Clinical Infectious Diseases* 41: 1167–1173.

Stiller M., Green R.E., Ronan M., Simons J.F., Du L., He W., Egholm M., Rothberg J.M., Keates S.G., Ovodov N.D., Antipina E.E., Baryshnikov G.F., Kuzmin Y.V., Vasilevski A.A., Wuenschell G.E., Termini J., Hofreiter M., Jaenicke-Després V. & Pääbo S. 2006. Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proceedings of the National Academy of Sciences* 103: 13578–13584.

Swanston T., Haakensen M., Deneer H. & Walker E.G. 2011. The Characterization of Helicobacter pylori DNA Associated with Ancient Human Remains Recovered from a Canadian Glacier. *PLOS ONE* 6: e16864.

Tange O. 2011. GNU Parallel - The Command-Line Power Tool. ;*login: The USENIX Magazine* 36: 42–47.

Tizzoni G. & Cattani G. 1890. Uber das Tetanusgift. *Zentralbl. Bakt.* 8: 69–73.

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

Valdiosera C., Günther T., Vera-Rodríguez J.C., Ureña I., Iriarte E., Rodríguez-Varela R., Simões L.G., Martínez-Sánchez R.M., Svensson E.M., Malmström H., Rodríguez L., Bermúdez de Castro J.-M., Carbonell E., Alday A., Hernández Vera J.A., Götherström A., Carretero J.-M., Arsuaga J.L., Smith C.I. & Jakobsson M. 2018. Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. *Proceedings of the National Academy of Sciences of the United States of America* 115: 3428–3433.

Valk T. van der, Vezzi F., Ormestad M., Dalén L. & Guschanski K. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources* 20: 1171–1181.

Velsko I.M., Fellows Yates J.A., Aron F., Hagan R.W., Frantz L.A.F., Loe L., Martinez J.B.R., Chaves E., Gosden C., Larson G. & Warinner C. 2019. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* 7: 102.

Velsko I.M., Frantz L.A.F., Herbig A., Larson G. & Warinner C. 2018. Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. 3: 23.

Walt A.J. van der, Goethem M.W. van, Ramond J.-B., Makhalanyane T.P., Reva O. & Cowan D.A. 2017. Assembling metagenomes, one community at a time. *BMC Genomics* 18: 521.

Warinner C., Herbig A., Mann A., Fellows Yates J.A., Weiß C.L., Burbano H.A., Orlando L. & Krause J. 2017. A Robust Framework for Microbial Archaeology. *Annual Review of Genomics and Human Genetics* 18: 321–356.

Warinner C., Rodrigues J.F.M., Vyas R., Trachsel C., Shved N., Grossmann J., Radini A., Hancock Y., Tito R.Y., Fiddyment S., Speller C., Hendy J., Charlton S., Luder H.U., Salazar-García D.C., Eppler E., Seiler R., Hansen L.H., Castruita J.A.S., Barkow-Oesterreicher S., Teoh K.Y., Kelstrup C.D., Olsen J.V., Nanni P., Kawai T., Willerslev E., Mering C. von, Lewis C.M., Collins M.J., Gilbert M.T.P., Rühli F. & Cappellini E. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* 46: 336–344.

Waterhouse A., Bertoni M., Bienert S., Studer G., Tauriello G., Gumienny R., Heer F., Beer T. de, Rempfer C., Bordoli L., Lepore R. & Schwede T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* 46: W296–W303.

Waterhouse A.M., Procter J.B., Martin D.M.A., Clamp M. & Barton G.J. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.

Weiß C.L., Gansauge M.-T., Aximu-Petri A., Meyer M. & Burbano H.A. 2020. Mining ancient microbiomes using selective enrichment of damaged DNA molecules. *BMC Genomics* 21: 432.

Wood D.E. & Salzberg S.L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15: R46.

Yu R., Ji C., Xu J., Wang D., Fang T., Jing Y., Kwang-Fu Shen C. & Chen W. 2018. The Immunogenicity of the C Fragment of Tetanus Neurotoxin in Production of Tetanus Antitoxin. *BioMed Research International* 2018: 6057348.

Zaragoza N.E., Orellana C.A., Moonen G.A., Moutafis G. & Marcellin E. 2019. Vaccine Production to Protect Animals Against Pathogenic Clostridia. *Toxins* 11: 525.

Zerbino D.R. & Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

Zhang S., Berntsson R.P.-A., Tepp W.H., Tao L., Johnson E.A., Stenmark P. & Dong M. 2017a. Structural basis for the unique ganglioside and cell membrane recognition mechanism of botulinum neurotoxin DC. *Nature Communications* 8: 1637.

Zhang S., Lebreton F., Mansfield M.J., Miyashita S.-I., Zhang J., Schwartzman J.A., Tao L., Masuyer G., Martínez-Carranza M., Stenmark P., Gilmore M.S., Doxey A.C. & Dong M. 2018. Identification of a Botulinum Neurotoxin-like Toxin in a Commensal Strain of Enterococcus faecium. *Cell Host & Microbe* 23: 169-176.e6.

Zhang S., Masuyer G., Zhang J., Shen Y., Lundin D., Henriksson L., Miyashita S.-I., Martínez-Carranza M., Dong M. & Stenmark P. 2017b. Identification and characterization of a novel botulinum neurotoxin. *Nature Communications* 8: 14130.

Zhang J.-C., Sun L. & Nie Q.-H. 2010. Botulism, where are we now? *Clinical Toxicology* 48: 867–879.

Zhou Z., Lundstrøm I., Tran-Dien A., Duchêne S., Alikhan N.-F., Sergeant M.J., Langridge G., Fotakis A.K., Nair S., Stenøien H.K., Hamre S.S., Casjens S., Christophersen A., Quince C., Thomson N.R., Weill F.-X., Ho S.Y.W., Gilbert M.T.P. & Achtman M. 2018. Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Current Biology* 28: 2420-2428.e10.

NCBI staff. 2020. We want to hear from you about changes to NIH's Sequence Read Archive data format and storage. *NCBI Insights*.

Appendix

Table 1 : Labels for Figure 3

1-SAMEA5847432	21-SAMD00041000
2-SAMEA5847426	22-SAMEA104281225
3-SAMEA104281221	23-SAMEA104281219
4-SAMEA5847473	24-SAMEA5054093
5-SAMEA6490841	25-SAMEA103957995
6-SAMEA5847501	26-SAMEA3937653
7-SAMN02727821	27-SAMEA103971604
8-SAMN05991104	28-SAMEA104402285
9-SAMEA6661726	29-SAMEA104281224
10-SAMEA6661722	30-SAMEA104441581
11-SAMEA6661724	31-SAMEA2810266
12-SAMN12394113	32-SAMEA5764555
13-SAMEA3486783	33-SAMEA104233049
14-SAMN02799089	34-SAMEA6502100
15-SAMN02727818	35-SAMN06046901
16-SAMEA5847472	36-SAMEA3713711
17-SAMEA104281226	37-SAMEA104548853
18-SAMEA104281220	38-SAMEA3486793
19-SAMN02799091	
20-SAMD00041001	

Supplementary Data

Supplementary data can be found at https://github.com/harohodg/aDNA-tetanus-analysis/