Beyond phylogenies: advancing analytical approaches for the field of ancient pathogenomics

Dissertation

To Fulfill the

Requirements for the Degree of

„doctor rerum naturalium“ (Dr. rer. nat.)

Submitted to the Council of the Faculty

of Biological Sciences

of the Friedrich Schiller University Jena

by B.Sc. M.Sc. Aida Andrades Valtueña **born on** 03 May 1991 **in** Barcelona

Gutachter:

1. Prof. Dr. Johannes Krause (Max Planck Institute for Evolutionary Anthropology, Jena)
2. Prof. Dr. Christina Warinner (Max Planck Institute for the Science of Human History, Jena)
3. Prof. Dr. Javier Garaizar Candina (Faculty of Pharmacy, University of the Basque Country UPV/EHU, Vitoria, Spain)

Beginn der Promotion: 22 November 2018
Dissertation eingereicht am: 17 Dezember 2020
Tag der öffentlichen Verteidigung: 08 Juli 2021

TABLE OF CONTENTS

Introduction

Health and disease: connecting the past to the present

Disease is and has been present throughout all human populations, independent of social or economic backgrounds. The ongoing COVID-19 pandemic in 2020 has shown again the importance in understanding the factors and mechanisms that contribute to the emergence, evolution and spread of infectious diseases. Infectious diseases, particularly zoonoses, represent one of the major risks for human populations. Ongoing processes such as climate change, habitat loss and species extinction are factors that contribute to the likelihood of emergence of new pathogens (Wilkinson et al. 2018; Khan et al. 2019). It is then prime to understand how and under which conditions microorganisms evolve and adapt to become pathogenic to humans and how they adapt to changes in human behaviour. However, infectious diseases are not the only concern for human populations. Recent advances in the study of the human microbiome, composed of the microorganisms that live within and on us, has shown the importance of the microbiome in health and disease (Pflughoeft and Versalovic, 2012). Of particular interests are pathobionts: microorganisms that form part of the natural microbiome but given certain circumstances may become pathogenic. Examples of pathobionts are *Helicobacter pylori, Staphylococcus aureus* and *Streptococcus mutans*. Microbiome composition has been linked to metabolic diseases, such as obesity (Maruvada et al. 2017), or oral diseases (Wade, 2013), such as caries or periodontal disease, all of which pose a huge burden in the health system (Vos et al. 2017). These observations have led to the hypothesis that many diseases are not the result of aggressive 'attack' of one pathogen, but rather a result of dysbiosis, a disequilibrated state between commensal and pathogenic microorganisms (Belizário and Napolitano, 2015). In order to respond and adapt to disease caused by both infectious pathogens and members of the microbiome, one needs to understand their co-evolution with the human host. In this regard, the past represents a well of knowledge that can be used to understand not only which diseases have been plaguing human populations throughout our history, but also how past societies understood and responded to disease. This knowledge can help inform modern management of pandemics. Quarantine, a commonly used method to isolate infectious patients from healthy individuals, has been in practise since the medieval ages (Tognotti, 2013), thus highlighting the importance of the past events to deal with current pan/epidemics. Furthermore, the study of ancient individuals can also provide information on the change in the human microbiome through time and contextualise the potential

shifts with specific changes in human behaviour, which could help to inform disease prevention and treatment.

How can we learn from past disease? Texts, mummies, bones and their tales

There are various sources of information one can study in order to understand past disease and health: written records, art of the time, archaeological sites and human remains. Historians have been addressing this question by analysing the written and artistic record. Studies of these resources have allowed to identify past pan/epidemics, as well as the social, economic and cultural consequences of past epidemics. These have also been employed to perform retrospective diagnostics to determine the responsible disease and agent of those past pandemics. Based on depictions and written records from physicians and historians at the time, it was hypothesised that the Black Death (1348) and subsequent outbreaks that lasted until the 18th century were the consequences of plague (Benedictow, 2004). However, diagnosis based on the historical record alone is challenging due to overlapping symptoms between different diseases and incomplete descriptions of the disease by past clinicians. Take for example the *cocoliztli*, an epidemic that broke out in 'New Spain' in 1545 and led to a population decline in Mesoamerica. The disease was depicted in the Codex en Cruz (currently held in Bibliothèque Nationale in Paris, France) - an Aztec pictorial book recoding historical events between the 15-16th century - as producing body rash and severe blood vomiting. Since the described symptoms and depictions in the Codices can be indicative of multiple diseases, debate is on-going regarding which disease was responsible for this epidemic, with viral haemorrhagic fever, typhus or even pneumonic plague as the proposed potential culprits (Warinner et al. 2012).

Skeletons and preserved bodies, such as mummies, from past individuals also represent a source for the study of past disease. The study of mummified bodies by palaeopathologists using both invasive and non-invasive methods on soft tissues has led to the understanding of diseases thought to be consequences of modern life such as cancer (Nerlich and Bianucci 2020) or arteriosclerosis (Allam 2009; Allam et al. 2011), and showed that these are not only common in modern times but they were actually already present in past societies. Furthermore, palaeopathologists have been learning about disease and health in the past by analysing skeletal remains, where bone malformation or modifications can be indicative of the health of the individual or traces of disease. Long-lasting disease with bone implication, such as tuberculosis (*Mycobacterium tuberculosis*), leprosy (*Mycobacterium leprae*) and syphilis (*Treponema pallidum*

pallidum), have been documented in skeletal remains (See reviews by Harper et al. 2011; Barberis et al. 2017; Roberts 2018). These methods are limited to the presence of physical signs of the disease, however not all the infectious disease results in characteristic bone lesions (Ortner 2003).

Ancient DNA

The retrieval of DNA from skeletons of ancient individuals can also be used to identify past pathogens, as well as host-related microorganism (Bos et al. 2019; Warinner et al. 2014; Spyrou, Bos, et al. 2019). On the contrary to the previous two mentioned resources, ancient DNA (aDNA) is not limited to documented epidemics nor to osteological evidence of disease in the tissues, thus allowing to explore both diseases where there are indications of its presence, but also diseases that would have remained invisible otherwise. aDNA provides molecular fossils for past pathogens and microbiomes allowing to explore past diversity; date the emergence of these species; and track their virulence evolution and adaptation through time.

Despite its advantages, aDNA research faces some challenges due to the intrinsic characteristics of its source. After the death of the individual, the body becomes colonised not only by its own microorganisms but also from environmental microorganisms, some of which are involved in body decomposition. These environmental taxa tend to represent the majority of the DNA present in an ancient sample, while the host and microorganism of interest, such as pathogens or members of the human microbiome, represent just a small amount of the DNA extracted (Bos et al. 2019). Additionally, aDNA goes through the process of degradation after death. Without repair mechanisms, DNA accumulates damage over time. One of the processes that contributes to damage accumulation in DNA is hydrolysis, which results in depurination and deamination of cytosines. Depurination is one of the processes that contribute to the fragmentation of DNA through time (Lindahl, 1993). The fragmented DNA is further affected by hydrolytic damage, which results in the deamination of cytosines (C) to uracil (U) that is then interpreted as thymine (T) by the DNA polymerase during library construction. Since the end of the fragmented DNA molecules are single-stranded they are more susceptible to the deamination, resulting in an increased frequency of misincorporated C→T towards the end of the molecule (Briggs et al. 2007). DNA degradation can lead to the loss of the DNA of interest, or the presence of small and damaged fragments that can make difficult their retrieval, by hindering Polymerase Chain Reaction (PCR) amplification during library construction (Höss et al. 1996), as well as pose a challenge for analysis due to misincorporated bases. Despite that all of these characteristics make it difficult to easily

retrieve aDNA, they offer a means to authenticate the ancient origin of the retrieved DNA molecules.

The development of PCR, a technique that allows for the amplification of DNA by utilising designed priming sites, opened the possibility to detect small amounts of DNA present in a sample. PCR techniques revolutionised the field of molecular diagnostics allowing for the detection of pathogens in clinical settings by targeting small stretches of DNA on the genome (Zauli, 2019). These molecular diagnostics methods were adapted to detect pathogenic DNA in ancient individuals to answer questions about disease in the past (Drancourt et al. 1998; Raoult et al. 2000; Arriaza et al. 1995; Haas et al. 2000; Kolman et al. 1999; Papagrigorakis et al. 2006), marking the start of the field of palaeomicrobiology. However, these early attempts to recover pathogenic DNA were met with scepticism from the scientific community, since contamination with modern sources, such as from modern work with pathogenic species in the same lab, unspecific amplification and lack of reproducibility by these studies were of major concern (see for example McHugh, Newport, and Gillespie 1997; Gilbert et al. 2004; Wilbur et al. 2009; Willerslev and Cooper, 2005; Shapiro, Rambaut, and Gilbert, 2006). The realisation of a lack of quality standards in the field of aDNA, led to the establishment of laboratory and analytical guidelines to ensure minimisation of modern contamination, as well as to ensure reproducibility in the field (Cooper and Poinar, 2000). Among other requirements, these guidelines obliged researchers to perform laboratory manipulations of ancient specimens such as extraction in a dedicated clean-lab, to use protective gear and the utilisation of positive and negative controls (modern guidelines can be seen in Llamas et al. 2017).

The establishment of guidelines to ensure minimisation of modern contamination in the lab, in combination with the advent of Next Generation Sequencing (NGS) techniques that allow for the untargeted sequencing of all the DNA present in a library at a high throughput, revolutionised the aDNA field. NGS not only open the possibility to retrieve full genomes due to the high throughput but more importantly it allowed for the clear establishment of authentication criteria for aDNA. The high throughput of NGS allowed recovery of enough ancient molecules to check the molecular characteristics of aDNA. By analysing NGS data, Briggs et al. (2007) showed the shift of length distribution towards short fragments was indicative of aDNA. This study also characterised the damage patterns of aDNA that are observed as C → T misincorporations with an increasing frequency towards the end of the molecules, and which can be approximately correlated to time. The presence of short fragments as well of damage patterns have become one of the standard measures to show the authentic ancient origin of the extracted DNA. Furthermore, NGS

techniques allowed for the establishment of additional criteria to ensure the correct identification of the pathogen. Coverage of the reference genome, which is calculated based on the number of reads covering a position (depth) and proportion of the reference containing mapping reads (breadth), is expected to be even across of the reference (Warinner et al. 2017). High coverage in specific regions (compared to the rest of the genome) is indicative of contaminants present in the sample and can lead to false positives. Another measurement often used to evaluate the correct assignment of the reads is edit distance. Edit distance is calculated based on the number of mismatches found in the reads in comparison to the reference, and it is expected that the majority of reads do not have any mismatches, with a continuous drop in the number of the reads in relation to increasing number of mismatches. Deviations from this pattern can point towards a false identification of the given taxa. The fact that ancient pathogenic DNA normally represents only a small fraction of the DNA in the sample makes their recovery from shotgun sequencing cost intensive by requiring the sequencing of billions of reads (e.g. Rasmussen et al. 2015; Manuscript A (Andrades Valtueña et al. 2017)). To address this challenge, capture-based techniques have been developed to enrich DNA libraries with the DNA of interest for multiple pathogens (see for example Bos et al. 2011; 2014; Schuenemann et al. 2013; 2011; Manuscript A (Andrades Valtueña et al. 2017) and Manuscript E) for the cost effective recovery of whole genomes from aDNA samples.

Since the development of these authentication criteria that help to distinguish authentic DNA molecules from environmental and modern contaminants, there has been the successful recovery of a range of human pathogenic bacteria DNA such as: *Yersinia pestis* (Spyrou et al. 2016; Spyrou, Keller, et al. 2019; Giffin et al. 2020; Bos et al. 2011, 20; 2016; Feldman et al. 2016; Wagner et al. 2014; Keller et al. 2019; Namouchi et al. 2018), *Mycobacterium leprae* (Schuenemann et al. 2013; Neukamm et al. 2020; Schuenemann, Avanzi, et al. 2018), *Mycobacterium tuberculosis* (Bos et al. 2014), *Salmonella enterica* (Key et al. 2020; Vågene et al. 2018; Zhou et al. 2018), *Gardnerella vaginalis* and *Staphylococcus saprophyticus* (Devault et al. 2017), *Treponema pallidum pallidum* (Giffin et al. 2020; Barquera et al. 2020; Majander et al. 2020; Schuenemann, Lankapalli, et al. 2018), *Borrelia recurrentis* (Guellil et al. 2018), and *Tannerella forsythia* (Warinner et al. 2014). This success has not only been limited to bacteria and now we have ancient genomes from various viral taxa, for example smallpox (Ferrari et al. 2020; Mühlemann et al. 2020) or Hepatitis B virus (Mühlemann et al. 2018; Krause-Kyora et al. 2018; Neukamm et al. 2020), as well as eukaryotic parasites, such as *Plasmodium falciparum* (Gelabert et al. 2016) *and Plasmodium vivax* (van Dorp et al. 2020). Furthermore, skeletal

remains are not the only source of aDNA. Medical collections are now providing a new avenue to retrieve ancient pathogen DNA, as for example they have allowed the recovery of smallpox from an 18th century infant leg (Ferrari et al. 2020) or *Plasmodium vivax* and *Plasmodium falciparum* from blood staining from the last malaria outbreak in Spain (van Dorp et al. 2020; Gelabert et al. 2016). Since NGS allows for the sequencing of the whole DNA present in the samples, it has allowed for the study of ancient human microbiome in an unprecedented detail. The human gut microbiome has been analysed from palaeofaeces (Maixner et al. 2016; Hagan et al. 2020; Tett et al. 2019; Rifkin et al. 2020). Additionally, dental calculus has been used to study the human oral microbiome due to its excellent preservation (Mann et al. 2018; Weyrich et al. 2017; Warinner et al. 2014; Velsko et al. 2019; Neukamm et al. 2020).

To showcase the advantages of aDNA, this thesis will be exploring two bacterial species associated with humans: *Yersinia pestis* and *Streptococcus mutans*. While *Y. pestis* is a pathogenic bacterium that has been subject of multiple aDNA studies due to its involvement with three documented pandemics in historical times, *S. mutans* is a typical microbiome-associated pathobiont that has not been studied before with aDNA data at whole genome level. In the following sections, I will describe these bacterial species and the diseases they are responsible for.

Plague

Plague is a disease that has been the focus of intense research due to the devastating effects of its pandemics. Rodents are the main host of plague, and represent the main reservoirs of human disease (Stenseth et al. 2008). Plague is maintained in the rodent populations by the transmission from rodent to rodent via the flea vector. Zoonotic outbreaks of plague tend to occur when the enzootic rodent populations experience a reduction in population. It is then that infected fleas leave their rodent host to look for new potential hosts, such as domesticated animals and humans. However, this is not the only transmission route by which humans can acquire plague. There has been reports of infection due to skinning (Wong et al. 2009) or ingestion (Arbaji et al. 2005; Christie, Chen, and Elberg 1980; Kehrmann et al. 2020) of animals infected by plague.

Plague can clinically manifest in three different forms, depending on the entry point of the bacterium in the host. In the bubonic form, the bacterium enters the body of the host via subcutaneous injection. The most common means of injection is through the bite of an infected flea, which acts as a vector for the disease, however contact with contaminated fluid or tissue

through injection or wounds has also been described (Stenseth et al. 2008). Once the bacteria have entered the body, it enters the lymphatic system where it is transported to the closest lymphatic nodes, where it starts reproducing. This leads to the immune response of the body, which manifests as the enlargement of the lymphatic node (Sebbane et al. 2005). The enlarged nodes are called buboes, hence the naming of bubonic form. The pneumonic form can occur via direct inhalation of infectious droplets, also known as primary pneumonic plague. Primary pneumonic plague is characterised by two phases: a pre-inflammatory phase, where the bacterium reproduces at high numbers by impairing the immune system of the host, and a proinflammatory phase where a full immune response is triggered by the host, however the immune response is typically unable to clear the infection due to the high bacterial titer and results into the death of the host within days after the symptoms onset (Lathem et al. 2005; Bubeck, Cantwell, and Dube 2007; Koster et al. 2010). Delay in treatment of bubonic plague usually derives into secondary pneumonic plague. Cats are particularly susceptible to pneumonic plague and become infected via the consumption of infected rodents, and pose a risk for their owners and veterinarian workers (Gage et al. 2000). Primary pneumonic plague is the rarest form of the diseases and there is only a few reported outbreaks (Begier et al. 2006; Ratsitorahina et al. 2000; Richard et al. 2015; Lien-teh et al. 1936; Bertherat et al. 2011). Finally, the septicaemic form occurs when the bacterium is introduced directly into the bloodstream. Secondary septicaemia can also occur from untreated bubonic or pneumonic plague.

Yersinia pestis

The causative agent of plague is the gram-negative bacterium *Yersinia pestis*. This bacterium belongs to the genera *Yersinia*, of which three species are known to infect humans: *Yersinia enterocolitica, Yersinia pseudotuberculosis* and *Y. pestis*. While *Y. enterocolitica* and *Y. pseudotuberculosis* are responsible for food-borne enteric diseases that are rarely life-threatening to humans (unless underlying causes are present, such as low iron or immunodepression), *Y. pestis* is responsible for plague, associated with high mortality as already described. The three pathogenic species share a common plasmid, named pCD1 or pYV. This plasmid contains pathogenicity-related genes such as the *Yop* operon, which codes for immunity down-regulators that allow the bacteria to survive in the host (Cornelis and Wolf-Watz, 1997; Pha and Navarro, 2016; Cornelis, 2002). While *Y. enterocolitica* does not form a monophyletic group with the other human pathogens, it has been shown that *Y. pestis* has evolved from *Y. pseudotuberculosis* (Achtman et al. 1999). The emergence of *Y. pestis* from its ancestor is characterised by the loss

of function of genes due to pseudogenisation as well as gene loss, genomic rearrangements and acquisition of new genetic material (Figure 1, Hinnebusch, Chouikha, and Sun 2016), however open questions remain about which conditions led to the divergence of these two species.

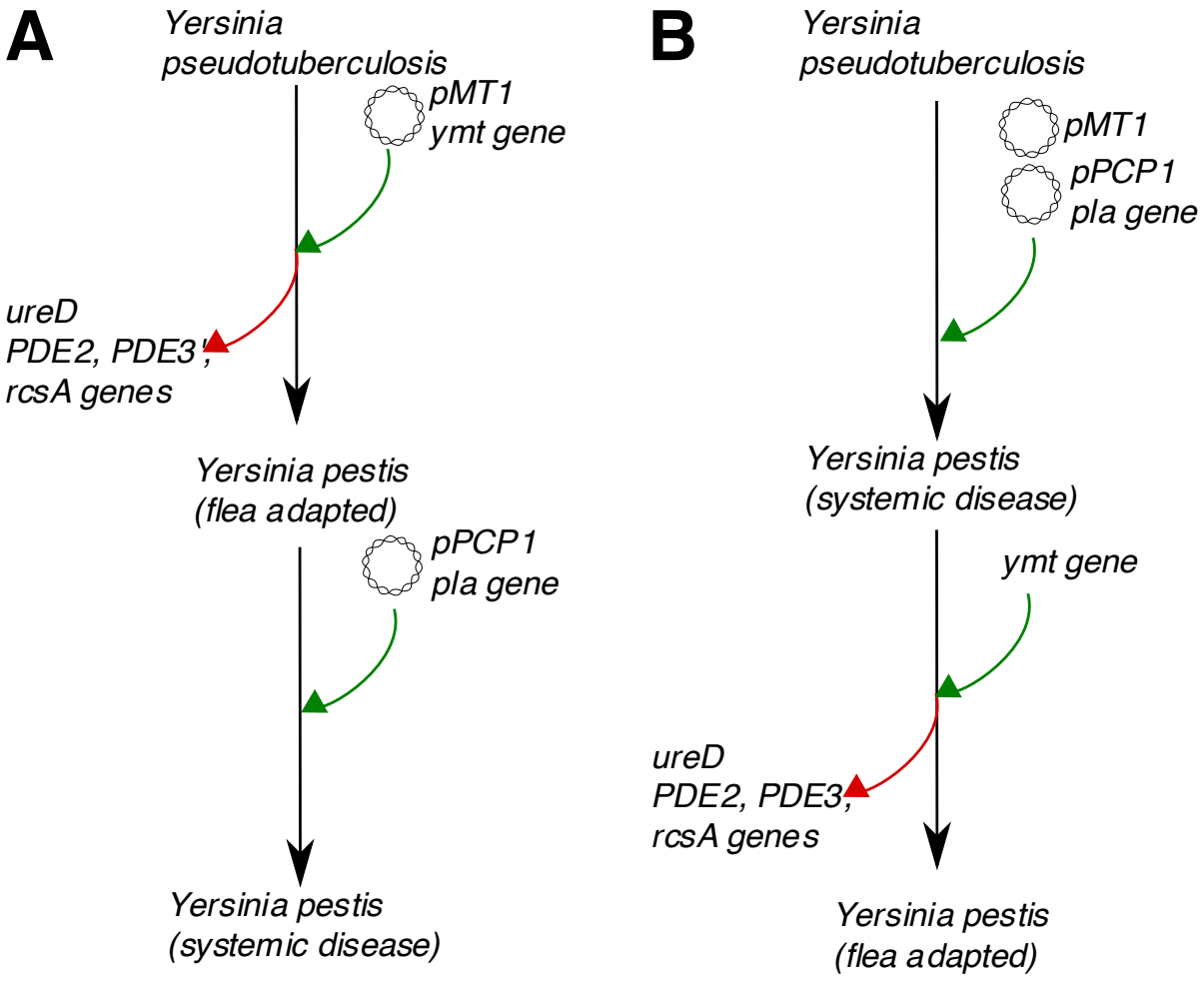


FIGURE 1: GENETIC ADAPTATIONS THAT LEAD TO THE EMERGENCE OF *Y. PESTIS* FROM *Y. PSEUDOTUBERCULOSIS*. A) INFERENCE BASED ON MODERN GENOMES AFTER SUN ET AL. (2014). B) KNOWLEDGE GAINED AFTER THE RECOVERY OF PREHISTORIC *Y. PESTIS* GENOMES (RASMUSSEN ET AL. (2015) AND MANUSCRIPT A (ANDRADES VALTUEÑA, 2017))

In terms of genetic acquisition, *Y. pestis* possess two unique plasmids: pMT1 and pPCP1. The adaptations of *Y. pestis* can be summarised as adaptations to a new means of transmission, the flea vector, and a new disease manifestation characterised by the invasiveness in the mammalian host. For the flea adaptation, various genes have been regarded as essential for an efficient transmission via the 'blocked flea' model (Hinnebusch 2012). The *ymt* gene, which is encoded in the pMT1 plasmid, allows the bacterium to survive in the midgut of the flea vector (Hinnebusch et

al. 2002). The presence of this gene, together with the silencing of biofilm regulators allows the bacteria to form a biofilm that blocks the proventriculus (entrance of the stomach) of the flea (Sun, Hinnebusch, and Darby, 2008). This leads to an increase feeding frequency by the starving flea, thus increasing the chances of *Y. pestis* to get transmitted (Chouikha and Hinnebusch 2012). This model also requires the silencing of *ureD* since its expression causes toxicity and morbidity of the flea vector (Chouikha and Hinnebusch 2014). However, recent studies have shown that none of these genes are in fact required for the transmission of *Y. pestis* by unblocked fleas (Scott and Duncan 2005), and that the 'early phase transmission' model can be as effective at transmitting the disease as blocked fleas (Eisen et al. 2007). This means of transmission has been proposed as the main mode in the rapid spread of the disease during epizootic outbreaks (Eisen et al. 2006; Eisen and Gage 2009, 201). Additionally to the adaptation to the flea vector, *Y. pestis* has also adapted to cause a systemic disease on the contrary to its ancestor *Y. pseudotuberculosis*, where cases of septicaemia are rare (Deacon, Hay, and Duncan 2003). The *pla* gene, which is contained in the pPCP1 plasmid, is a gene that confers *Y. pestis* with the ability to survive and invade the mammalian host (Hinnebusch, Fischer, and Schwan 1998; Sebbane et al. 2006). Despite the deep understanding on the virulence factors, the timing on the acquisition of those changes is still unclear. By utilising modern genomes, it was inferred that the adaptation to the flea vector predated the acquisition of the *pla* gene, which allows for rapid dissemination in the mammalian host (Figure 1A, Sun et al. 2014). However, the study of the gene coverage of the virulence factors of *Y. pestis* in recent aDNA studies (Rasmussen et al. 2015; Manuscript A (Andrades Valtueña et al. 2017), Figure 1B) have shown that *pla* was acquired first, prior to the adaptations required for the blocked flea transmission. These findings highlight the potential of ancient genomes to inform on the timing of crucial events in plague evolution.

History of plague and contributions of ancient DNA

Plague has been linked to three pandemics based on the written record, which had important consequences on the shaping of human history. The first pandemic represents the first documented plague pandemic that started with the plague of Justinian, named after emperor of the west-Roman empire, in the 6th century and lasted until the 8th century with recurrent outbreaks. This pandemic, together with an already weakened empire due to war and starvation, has been linked to the decline of the Roman empire (Russell 1968). Afterwards, plague seemed to disappeared from Europe, until the 14th century. The infamous Black Death (1348) marked the start of the second pandemic in Europe (Benedictow 2004). This pandemic represents the longest

lasting plague pandemic recorded in history, with recurrent outbreaks until the 18th century. Numerous medical measures started during this epidemic to combat the disease, such as the establishment of quarantine as a mean to stop the spreading of the disease, with its first instance of its imposition dating from 1377 in the city of Dubrovnik, Croatia (Tognotti 2013). The third pandemic is the most recent plague pandemic and started in the Yunnan province (China) in the 19th century. From there, it spread world-wide via sea-faring; creating endemic foci around the world that are still active today. It was also during this pandemic when Alexandre Yersin isolated the agent responsible for plague, *Y. pestis*, in 1894 (Butler 2014). Although some may believe plague is a disease from the past, new cases of plague are reported yearly, with the most recent cases reported in China (Kehrmann et al. 2020) and the Democratic Republic of the Congo (World Health Organisation 2020) during 2020. With numerous still active rodent reservoirs, plague still poses a risk to modern human populations.

Plague was linked to the first and second pandemics by historians, based on similarity of the symptoms reported by the clinicians or historians at the time, as well as depictions of the diseases in primary sources. However, other viral or bacterial diseases had also been argued to be responsible for these pandemics (See discussions in Cohn JR 2008; Sallares 2006). In order to provide a definitive answer, whether *Y. pestis* was indeed responsible for the second pandemic, PCR studies targeting this bacterium were performed on DNA extracts from individuals that had died during the second pandemic (Drancourt et al. 1998; Raoult et al. 2000). Despite the reported detection of *Y. pestis* by these studies, these early attempts were met with scepticism by the scientific community as they could not be reproduced (Gilbert et al. 2004). With the introduction of NGS-techniques and authentication criteria, Schuenemann et al. (2011) and Bos et al. (2011) were able to recover the first pPCP1 plasmids and *Y. pestis* genomes, respectively, from individuals interred in the East Smithfield cemetery (London) documented to have been in use during the Black Death outbreak (Cowal et al. 2008). These studies, together with the previous PCR results, showed that *Y. pestis* was indeed one of the causes of the second pandemic. Since then, multiple aDNA studies have contributed tens of genomes from both the second and the first pandemic, thus confirming *Y. pestis* involvement in both of these pandemics, and enriching our understanding of the dynamics of those major historical events.

The first ancient genomes from the first pandemic were recovered from individuals from Germany by Wagner et al. (2014) and Feldman et al. (2016) and confirmed that plague was responsible for this pandemic as previously suggested by PCR studies conducted in those sites (Harbeck et al. 2013). Further evidence of plague being responsible for this pandemic was recently provided by

Keller et al. (2019). In this research, it was shown that plague was widespread across Europe, by recovering *Y. pestis* genomes from regions with and without historical accounts of the disease. However, the authors could not provide an answer on how the disease was introduced to Europe. Historical accounts place the start of the pandemic in Egypt (Russell 1968), thus sampling individuals from the Eastern Mediterranean, as well as from Asia, from the 6th century could provide insights into the route of entry and origin of the strains responsible for this pandemic.

Two competing hypotheses have been proposed to explain how plague affected Europe during the second pandemic. On one hand, some scholars supported the hypothesis of a single introduction into Europe during the Black Death and establishment of local reservoirs that were responsible for the recurring epidemics. On the other hand, it had also been proposed that plague was introduced multiple times into Europe. The reintroductions would then explain the multiple waves during the second pandemic. Climatic indicators from the past have supported the second explanation (Schmid et al. 2015). Contrastingly, aDNA studies have provided strong support for the single introduction hypothesis. Firstly, it has been shown that the Black Death outbreak was likely caused by a single clone that spread quickly across the whole of Europe, given that all the genomes dating from that time are identical (Spyrou et al. 2016; Spyrou, Keller, et al. 2019). Based on historical records, it was hypothesised that plague was likely introduced into Europe from via the Black Sea (Benedictow 2004), with supporting evidence coming from a genome recovered prior the Black Death from the Samara region (Spyrou et al. 2019). The identification of genomes dating from post-Black Death forming a single independent and extinct lineage that descends from the recovered genomes dating to the Black Death (Spyrou 2016, 2019), strongly suggested a single introduction. With the addition of more genomes, Spyrou et al (2019) argued for the presence of two different reservoirs - one responsible for the outbreaks in central European and one for outbreaks related to port cites, which has been also suggested by Guellil et al. (2020). It has also been shown that the Black Death is the ancestor of the ‘Branch 1’ genomes that are also responsible for the third pandemic (Spyrou 2016, Namouchi 2018). Open questions still remain, however, such as where were the reservoirs of plague located and why did the disease disappear from Europe after the 18th century.

The study of ancient plague has been mainly focused on time periods where historical accounts are available, and particularly targeting individuals found in mass-graves, which are indicative of a catastrophic event such as an epidemic. However, one can ask whether plague did not also affect human populations before those pandemics, given these repeated pandemic events. In order to explore this question, studies should be performed in individuals predating those

pandemics, where no historical records are available. The advancements in aDNA described in the section above, have allowed not only for the recovery of ancient pathogens but also for the generation of large ancient human DNA datasets (See for example Allentoft et al. 2015; Damgaard et al. 2018; Haak et al. 2015). Those datasets represented a unique opportunity to explore disease in the past, without the need for historical or epidemiological indicators of disease, such as mass-graves. By analysing the Allentoft et al. (2015) dataset, Rasmussen et al. (2015) made the surprising discovery that plague had been affecting human populations at least since 5000 years ago. They recovered two complete genomes from the Altai region and found indications of the bacterium in individuals from Europe. Whether the European strains were part of the same ancestral lineage formed by the Altai strains was an open question. This question was tackled in Manuscript A (Andrades Valtueña et al. 2017) were we show that indeed those European strains were part of the already described lineage, and link human mobility with the spread of *Y. pestis* into Europe. Since then, further genomes from prehistorical times have shown the presence of multiple ancient extinct lineages and showed that *Y. pestis* experienced a quick diversification during this period (Rascovan et al. 2019; Spyrou et al. 2018, Manuscript B).

Caries

Oral disease, and specially caries, represents one of the major health burdens worldwide (Vos et al. 2017). In contrast to *Y. pestis* that aggressively attacks the human immune system, caries is a tooth lesion that is characterised by the demineralisation of the enamel due to 'endogenous' acid producing microorganisms. Untreated caries can lead to tooth loss, excruciating pain and malnutrition due to painful mastication, all of which can impair normal life of the host. It can also evolve to a tooth abscess, inflammation and infection of the tissue surrounding the tooth, which can result in serious and even life-threatening infections. Early studies on the causes of caries linked this disease to single bacterial species (Loesche 1986), however it has now been shown that caries has a multifactorial and polymicrobial nature (Marsh 1994). Human behaviour, such as the consumption of sugary drinks or food (Moynihan and Petersen 2004; Peres et al. 2016), as well as microbial background play a major role in the development of carious lesions (Aas et al. 2008). Furthermore, social background and lack of oral health education have also been hypothesised to be predictors of caries development (Sälzer et al. 2017).

There are three proposed theories to explain the underlaying microbial influence in the formation of caries: the specific plaque, the non-specific plaque and the ecological plaque hypotheses. The specific plaque hypothesis proposes that the formation of caries is driven by specific

microorganisms, such as *Streptococcus mutans* or *Streptococcus sobrinus*. It was supported by early studies where the isolation or detection from carious lesion of single bacterial species was observed (Loesche 1986). In contrast, the non-specific plaque hypothesis suggests that oral diseases is the result of the action of the overall plaque microbiota rather than single 'pathogenic' species, thus implying that multiple bacterial species are the aetiological agents of caries (Theilade 1986). Finally, the ecological hypothesis describes caries formation as a result of changes in the local environment that causes a shift in the microbial communities resulting in imbalanced relationships between the microbiota and the host; a concept that is currently the most accepted hypothesis (Kleinberg 2002; Aas et al. 2008). These hypotheses have been inferred from studying living individuals, and little is known about whether these hypotheses also applied to past populations. By studying the archaeological and historical records, we can start to understand how the oral microbiota has changed overtime and whether the underlaying causes of caries have changed. Understanding the mechanisms and long-term evolution of the microorganisms associated with carious lesions could help inform the development of new treatments and measures to prevent the disease.

Evidence of caries in the archaeological record

The presence of caries is ubiquitous throughout the archaeological record (Pezo Lanfranco and Eggers 2012). While prevalence of caries in past human populations is highly variable, various studies have shown an increase in carious lesions in populations associated with the introduction of agriculture (See for example Cohen and Armelagos 1984; Formicola 1987; Lukacs 1992; Nicklisch et al. 2016; or the review by Warinner 2016). Currently, the earliest evidence of caries has been found in a hunter-gatherer human population dating to 15,000 and 13,700 calibrated years Before Present (calBP) from Morocco (Humphrey et al. 2014). The authors hypothesised that the carious lesions were the results of the consumption of wild plants that contained fermentable carbohydrates by this human group. This challenged the common belief that an increase of prevalence of the disease was due to the adoption of agriculture and consumption of highly processed foods, thus highlighting the importance on the need of multidisciplinary research to understand caries formation in the past.

Despite caries being easily identified in the archaeological record, there is little information on the microorganisms that were associated with them. By recovering microorganism from carious lesions, we could start to understand how the ecology, composition of the microbiota has changed over time. Furthermore, coupling the microbial work with other methods such as stable isotopes,

osteological research or residues analysis could help to infer diets and other possible behaviours that put the host at risk of developing caries.

In order to start addressing those questions, I have focused on one of the best studied bacteria that has been highly associated with caries, *Streptococcus mutans* (Manuscript E).

Streptococcus mutans: A commensal with disease potential

S. mutans is a gram-positive bacterium from the genus *Streptococcus*. It is a normal resident of the human oral microbiota; however, given the right conditions, it can become pathogenic (also known as pathobiont). It is one of the species that can contribute to dental caries (Loesche 1986). In addition, more recently, *S. mutans* of the k serotype have been shown to play a role in infective endocarditis (Nagata et al. 2006; Nakano et al. 2007; Nomura et al. 2006).

S. mutans was first isolated by Clarke from teeth in advanced states of caries in 1924 (Clarke 1924). *S. mutans* can be classified into 4 serotypes depending on their sequence of the rhamnose-glucose polymers (RGP), cell wall antigens: c, e, f and k, which are found in the oral cavity with a proportion of 70-80%, 20%, <5% f and <2% respectively (Nakano and Ooshima 2009). Although all serotypes are able to migrate to the bloodstream and cause endocarditis in coronary arteries, serotype k is the most commonly found in the endothelial cells of coronary arteries. This is hypothesised to be due its low cariogenic and antigenic potential, which allows it to survive for longer periods in the blood and extend its virulence (Bedoya-Correa, Rincón Rodríguez, and Parada-Sanchez 2018).

S. mutans has four phenotypic traits which confer it cariogenic potential: biofilm formation, acidogenicity, acidurity, and bacteriocin production. The capacity of producing biofilms allows *S. mutans* to colonise the tooth, and endure changing environmental conditions (Krzyściak et al. 2014). Acidogenicity is the capacity of producing acid, which *S. mutans* produces in the form of lactic acid as a by-product of carbohydrates fermentation (Clarke 1924; Ajdić et al. 2002). Lactic acid is then secreted to the environment and demineralises the tooth enamel, thus leading to carious lesions (Featherstone and Rodgers 1981). Acidic environments normally lead to slow growth and death in bacteria, however *S. mutans* has adapted to survive extremely low pH (acidurity). It has been shown that *S. mutans* can survive in lower pH than other streptococci, which allows this species to outcompete other bacteria in low pH environments (Bender, Thibodeau, and Marquis 2016). Furthermore, *S. mutans* produces and secretes bacteriocins that

can destroy other *S. mutans* as well as other competitor species, thus increasing the chances of becoming the dominants species in the carious lesions (Ajdić et al. 2002). These bacteriocins are coupled with a competence pathway in *S. mutans* that allows it to incorporate environmental DNA into its chromosome (Shanker and Federle 2017). The imbalance from the overgrowth of this taxa versus others is what makes this species to become dominant in carious lesions.

The first genome of *S. mutans* was sequenced in 2002 (Ajdić et al. 2002). It consists of a circular chromosome of around 2.1Mb, and revealed the set of genes that confer *S. mutans* its cariogenic potential. The sequencing of further genomes provided evidence of recombination and lateral gene transmission (LGT) between *S. mutans* and other *Streptococcus* species, due to its capability of becoming competent (Hagen and Son 2017). This complicated subsequent phylogenetic analysis of this bacteria, since these regions show 'non-tree' like patterns, making it difficult to infer the genetic relationship among strains of this species. The analysis of all the genes present in *S. mutans,* what is called the pangenome, has shown that strains of this bacterium differ greatly in their gene content (an open pangenome) and has been linked to the capability of *S. mutans* to adapt to changing environmental conditions and its ability to colonise other body parts (Meng et al. 2017). The study of the pangenome formation of *S. mutans* could inform us about changes in the *S. mutans* that can be linked to human behaviour. Major shifts in human subsistence strategies, such as the introduction of agriculture (Neolithization), may have resulted in the acquisition of new genes by *S. mutans* to adapt to a more carbohydrate richer diet. The early antibiotic treatment or use of fluoride in toothpaste, used to prevent caries formation, could have also led to changes in the genomic content of cariogenic species such as *S. mutans,* for example acquiring genes to become resistance to those substances.

The transmission of *S. mutans* has been showed to typically be from care-givers to children (Alves et al. 2009; Douglass, Li, and Tinanoff 2008; Berkowitz and Jones 1985; da Silva Bastos et al. 2015; Lapirattanakul and Nakano 2014; Y. Li and Caufield 1995). The transmission window for *S. mutans* comprises the first years of life, after which it is less likely that new strains will colonise and become part of the oral microbiota (Caufield, Cutter, and Dasanayake 1993). The evidence for vertical transmission between mothers and children (Berkowitz and Jones 1985; da Silva Bastos et al. 2015; Lapirattanakul and Nakano 2014; Y. Li and Caufield 1995) has led to the assumption that *S. mutans* would provide similar phylogeographical information as other pathobionts, such as *H. pylori*, which mirrors the dispersal of humans across the world (Mégraud, Lehours, and Vale 2016), making it an ideal taxon to track human migrations in a finer scale since bacteria reproduce at a higher degree than the human host. However, whether *S. mutans* displays

a phylogeographical signal remains unclear: Cornejo et al. (2013) showed no such signal, while González-Ittig et al. (2016) argued for the presence of phylogeography in a smaller dataset. Manuscript E supports the lack of phylogeography in this species, however we propose that the lack of evidence arises from the currently available modern dataset. Recent globalisation could have contributed to the recombination of *S. mutans* strains from different continents, thus diluting the phylogeographic signal.

Despite the fact that techniques for aDNA are available and caries are found in individuals from the past, there has been limited work performed on ancient *S. mutans*. The taxon has been detected in a few studies either from PCR amplification of a gene specific to this species (Simón et al. 2014) or from amplicon data from dental calculus (De La Fuente, Flores, and Moraga 2013; Adler et al. 2013). These studies gave an indication that this pathobiont may have been present in past human populations, however the authentication of the fragments as ancient cannot be assessed due to the lack of ancient damage in PCR amplification sequencing reads – as occurred in early work in *Y. pestis*. This is particularly relevant given that the oral cavity contains multiple streptococcal species that could lead to false positives, and contamination from modern sources during the handling of the sample is likely. For these reasons, obtaining whole genomes employing NGS techniques for aDNA, will allow for the verification and confirmation of these studies.

Adler et al. (2013) proposed that there was an increase in the *S. mutans* population after the Neolithic transition and also during the industrialization. Although this study was based on a small number of samples and reads, Cornejo et al. (2013) also showed using a modern dataset that the population of *S. mutans* experienced a demographic expansion around 10,000 years ago, coinciding with the onset of agriculture in multiple locations of the world (Gepts 2010). Ancient *S. mutans* genomes could help support this hypothesis by providing a deep-time calibration point in this analysis. However, no ancient *S. mutans* genomes has been recovered to date, with the exception of the one presented in Manuscript E of this thesis.

Comparative ancient pathogenomics

Ancient pathogenomics research has relied heavily on the use of phylogenetic analysis for the inference on how ancient strains are related to the extant strains, and how the pathogen has spread in past time. However, as stated above for *S. mutans*, phylogenetic analyses are not suitable for all pathogenic species. For recombinant bacteria, this type of analysis can lead to

wrong inferences due to the presence of parts of the genome that do not evolve in a 'tree-like' manner, and results in phylogenies with an incorrect topology and low statistical support. Furthermore, phylogenetic analyses do not provide any information on changes in the virulence of past pathogens, which is more relevant to understand the evolution of pathogenicity.

In this thesis, I will attempt to move the attention from phylogenies, to new types of analysis that can help show how the genetic content of an ancient strain can shed light on the evolution of virulence and adaptation of pathogens to the human host. In the following paragraphs, I will present the basic concepts for the new analytical approaches that I developed over the course of this thesis, as well as, the current state of their application in the ancient pathogenomics field.

Comparative genomics is a field of biology that compares different genetic features, such as gene content or genomic structure, from different strains or microorganism in order to identify their similarities and differences (Binnewies et al. 2006). This can help researches to identify genetic features that explain the different phenotype of a species, as for example describe genes that may play a role in the virulence of a pathogen or identify species-specific genes that could be used for antibiotic design. This is normally achieved by aligning whole genomes, which are assembled *de novo*.

De novo assembly allows for the reconstruction of a genome without the need of a reference (Sohn and Nam 2018). The most commonly used algorithms to assemble genomes from sequencing reads are *de Bruijn* graph-based algorithms such as SOAPdenovo (Luo et al. 2012), velvet (Zerbino and Birney 2008) or SPAdes (Bankevich et al. 2012). Those algorithms reconstruct the genomes by finding overlaps of the sequencing reads to build a graph, where a read is represented by multiple k-mers, which are parts of the read of length k, and in turn those are connected to other reads through edges. Contigs represent continuous paths in the graph composed by multiple overlapping reads that represent a genomic region. Those contigs are then ordered to form scaffolds. A "perfect" assembly consists of a single scaffold for each genomic molecule present in a cell such as the chromosome, plasmid or organelle genome. The presence of repetitive regions complicates path finding and leads to the fragmentation of the final assembly. To resolve repetitive regions, long read data (such as from PacBio or Nanopore sequencing) is used in modern experiments to expand the repetitive regions, thus creating a contiguous assembly (Sohn and Nam 2018). However, *de novo* assembly has remained elusive in the context of aDNA due to its molecular characteristics. As explained above, aDNA is highly fragmented, which in combination with low amounts of pathogen DNA, makes it difficult to find overlapping

reads to build long contigs, thus resulting in fragmented assemblies. Furthermore, the metagenomic context of aDNA can lead to the formation of chimeric contigs composed by DNA of different closely related species present in the sample. This has limited the application of *de novo* assembly techniques to exceptionally well preserved samples such as for *Mycobacterium leprae* (Schuenemann et al. 2013), hepatitis B virus (Krause-Kyora et al. 2018) or *Gardnella vaginalis* and *Staphylococcus saprophyticus* recovered from a nodule (Devault et al. 2017). *De novo* assembly has also been attempted in less well preserved ancient samples resulting in more fragmented assemblies, such as those available for ancient *Y. pestis* (Bos et al. 2011; Luhmann, Doerr, and Chauve 2017).

Since *de novo* assembly remains challenging with aDNA, the field of ancient pathogenomics started by developing means to explore the gene content of ancient strains via a reference. The development of workflows to detect the presence or absence, as the one presented in manuscript A, of known virulence factors allows to explore the virulence potential of ancient strains. This type of approach has been applied for example to *Y. pestis* (Spyrou, Keller, et al. 2019; Spyrou et al. 2018; Keller et al. 2019; Rascovan et al. 2019; Rasmussen et al. 2015; Manuscript A (Andrades Valtueña et al. 2017); Manuscript B) or *Salmonella enterica* (Vågene et al. 2018; Key et al. 2020; Zhou et al. 2018). These studies have led to the discovery of particular genetic backgrounds of ancient strains, such as the lack of genetic adaptation necessary for the transmission via the blocked flea vector in the earlier forms of plague (Rasmussen et al. 2015; Rascovan et al. 2019; Manuscript A (Andrades Valtueña et al. 2017); Manuscript B). However, comparative genomics is not limited to the study of specific virulence factors.

Recently, the study of the pangenome has been proposed as a tool to study pathogenic bacteria (Rouli et al. 2015). The pangenome consist of all the genes present in a given taxa and can be divided in core and accessory genes. While the core genes are present in all the strains of a taxa and probably represent the essential genes for the bacterium to survive, the accessory part of the pangenome is formed by genes that are present in only one or few strains and probably explain the different phenotypes of the strains. Most of the modern pangenomic tools require well-annotated assemblies for the computation of the pangenome. Since *de novo* assembly is challenging in ancient samples as just discussed, there is currently no annotated ancient genome that could be included in the pangenome construction. For that reason, to date there is only a few studies that have looked at the pangenome of ancient strains. Zhou et al. (2018) recovered an *S. enterica* Paratyphi C genome from an ancient individual and by analysing the pangenome of *S. enterica,* they identified genomic islands that differentiated the Paratyphi C lineage from other *S.*

enterica strains since the past. Furthermore, they showed that pseudogenisation played a role in host specificity. This was further confirmed by Key et al. (2020), who by analysing the pangenome and pseudogenisation of this bacterium showed that ancient strains presented less pseudogenes and proposed that the neolithization, acquisition of agriculture and animal domestication by human populations, contributed to the host specialisation in *S. enterica*. Manuscript C represents the first study to apply pangenomics with the inclusion of the published ancient genomes to the study of the emergence of *Y. pestis*. The study of the pangenome of ancient strains can provide insights into the long-term evolution of microbial species by detecting genes that could be involved in the adaptation to a new host, a novel transmission or virulence potential of the microbe.

Aim

Current ancient pathogen research is based on phylogenetic analysis of the recovered strains. I want to increase the potential of the field of ancient pathogenomics by exploring not only the phylogenetic relationship of ancient pathogens with their modern counterparts but also gaining insights into changes in virulence, transmission mechanisms and ecology through time based on changes in the genomic content and structure. In order to answer those questions, there is a need to adapt the existing tools for the study of modern genomes to account for the highly degraded and fragmented DNA obtained from ancient individuals. By creating new workflows adapted to ancient DNA, we can obtain time transects in which to explore past diversity that can inform the timing of essential adaptations for pathogens, such as the acquisition of new genes or variants or the loss of genetic functions detrimental for pathogenicity. I demonstrate the advantage of ancient DNA by showcasing the application of modern genomic tools, such as SPAdes (a *de novo* assembler) or panX (*a pangenomic tool*), to two bacterial species that can cause disease in humans: the exclusively pathogenic *Y. pestis* and the pathobiont *S. mutans*.

For the case of *Y. pestis*, this thesis aims to shed light into the early evolution of this pathogen to understand its emergence from its rarely life-threatening ancestor *Y. pseudotuberculosis*. Furthermore, I aim to provide evidence on how its transmission, virulence and genomic content has changed over time and contextualised those findings in light of the historical and archaeological record. For *S. mutans*, I provide the first complete genome recovered from an ancient individual. *S. mutans* is one of the bacteria involved in the formation of caries hypothesised to be a consequence of the dysbiotic state of the modern oral microbiome. The recovery of *S. mutans* DNA from ancient individuals could provide a timing into the introduction of this taxa into the human oral microbiome and shed light into the evolution of this pathobiont in response to changes in human behaviour, subsistence strategies or cultural practises. A main motivation to recover genomes from this bacterium is to test whether this pathobiont displays a phylogeographic signal which could be utilised to track humans in the past, a useful proxy for human movements that does not require the study of the host. Overall, this thesis aims to showcase the potential of ancient DNA as a tool to understand pathogen emergence, evolution and adaptation and the context of those changes through time.

Manuscripts overview and author contribution

Manuscript A: The Stone Age Plague and Its Persistence in Eurasia.

Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, Michal Feldman, Rimantas Jankauskas, Ivor Janković, Ken Massy, Mario Novak, Saskia Pfrengle, Sabine Reinhold, Mario Šlaus, Maria A. Spyrou, Anna Szécsényi-Nagy, Mari Tõrv, Svend Hansen, Kirsten I. Bos, Philipp W. Stockhammer, Alexander Herbig, Johannes Krause

Published in Current Biology, 2017 Dec 4; 27(23):3683-3691.e8. doi: 10.1016/j.cub.2017.10.025. Epub 2017 Nov 22.

In Manuscript A, we present the first six European genomes of plague dating to the Stone Age (5000-3000 calBP). We show that those genomes are related to the previously published *Yersinia pestis* genomes from the Altai region. Combining the knowledge gained from archaeology and human aDNA studies, together with the phylogenetic analysis of plague, we propose that *Y. pestis* was introduced to Europe in a process related to the expansion of the human groups from the steppe, closely associated to the Yamnaya culture. We report that the European genomes are also missing key genes for its efficient transmission via a flea, considered to be the main vector of plague. We discuss what could be the implication of this regarding its transmission during the Stone Age.

AAV Contribution (80%):

- Planned the research
- Established a fast screening method to identify *Y. pestis* in metagenomic samples
- Performed the UDG treatment of the KunilaII sample
- Analysed the reconstructed genomes, specifically:
    - Performed the basic mapping and SNP calling
    - Performed phylogenetic analysis
    - Established a script to check for presence/absence of virulence genes in *Y. pestis*
    - Performed dating analysis with BEAST
- Generated all figures and tables

- Wrote the initial manuscript
- Refined the manuscript with the comments from all co-authors

Manuscript B: Stone Age *Yersinia pestis* genomes shed light into the early evolution, diversity and transmission ecology of plague

Aida Andrades Valtueña*, Gunnar U. Neumann*, Maria A. Spyrou*, Lyazzat Musralina*, Beisenov Arman, Alexandra Buzhilova, Matthias Conrad, Leyla B. Djansugurova, Miroslav Dobes, Michal Ernée, Javier Fernández-Eraso, Bruno Frohlich, Mirosław Furmanek, Agata Hałuszko, Svend Hansen, Éadaoin Harney, Felix M. Key, Elmira Khussainova, Yegor Kitov, Corina Knipper, Carles Lalueza Fox, Judith Liddleton, Ken Massy, Alissa Mittnik, José Antonio Mujika-Alustiza, Iñigo Olalde, Luka Papac, Sandra Penske, Ron Pinhasi, David Reich, Sabine Reinhold, Harald Stäuble, Christina Warinner, Philipp Stockhammer, Alexander Herbig, Wolfgang Haak, Johannes Krause
*Equal contribution
In preparation for submission to Nature Ecology and Evolution

In Manuscript B, we recover 15 new ancient *Y. pestis* genomes dating between ~5000-2400 years Before Present (yBP) in order to understand the plague dynamics during the Stone Age. 14 of the genomes fall in the previously described non-flea adapted LNBA lineage, thus expanding the geographical regions where the disease was found and also showing the long survival of this lineage until the Iron Age (2400yBP). We also recover a flea-adapted *Y. pestis* genome from Spain, which falls close to the RT5 genome recovered from the Samara regions in Russia. We discuss the different disease and ecology potential of both of this *Y. pestis* forms, which showcases the diversity of the bacterium early on in its evolution.

AAV Contribution (50%):

- Participated in the planning of the study
- Analysis
    - Performed genome reconstruction by mapping to the *Y. pestis* reference
    - Generated basic files for phylogenetic analysis
    - Performed phylogenetic analysis
    - Performed indel analysis
    - Implemented and run genetic vs geographic vs time analysis

    - Implemented a method to genotype single-stranded libraries to account for damage
- Writing:
    - Drafted the initial manuscript: introduction, most of the method section, most of the results section, and discussion
    - Created tables, figures and supplementary figures

Manuscript C: A Pangenome of the *Yersinia pseudotuberculosis* complex

Aida Andrades Valtueña[1], Alexander Herbig[1]

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

In preparation

In Manuscript C, we explore the pangenome of the *Yersinia pseudotuberculosis* complex. We utilise panX to create both a pangenome of the *Y. pseudotuberculosis* complex and a core genome of *Y. pestis*. We implement a new workflow to recover the present or absence of the genes in a pangenome which allows the use of other sources other than assemblies, such is the case of ancient DNA genomes or modern genomes where only sequencing data is available. Additionally, we present a workflow to detect pseudogenisation from a core genome. We use the available genomes from the *Y. pseudotuberculosis* complex to demonstrate the potential of the implemented methods to obtain information in the changes of the pangenome and pseudogenisation in this complex.

AAV Contributions (95%):

- Conceived the idea for the study
- Analysis
    - Created pan-genome for *Y. pseudotuberculosis* complex and a core-genome for *Y. pestis*
    - Compiled a metadata file for the *Y. pseudotuberculosis*, *Y. similis* and *Y. pestis* genomes used in the analysis

    - Implemented a pipeline for the detection of presence/absence of genes from the pan-genome
    - Implemented a pipeline for the detection of pseudogenes form a set of genes
    - Wrote a script for functional classification of the genes by extracting GO terms
    - Performed analysis in the pan-genome presence/absence of *Y. pseudotuberculosis* complex
    - Performed pseudogene analysis on the core genome of *Y. pestis*
    - Performed phylogenetic analysis
- Writing:
    - Wrote the initial draft
    - Created figures and tables

Manuscript D: *De novo* assembly of a 17th century *Yersinia pestis* genome from a plague victim buried in the New Churchyard burial ground in London

Aida Andrades Valtueña[1], Maria A. Spyrou[1], Elizabeth A. Nelson[1], Niamh Carty[2], Robert Hartle[2], Michael Henderson[2], Elizabeth L. Knox[2], Don Walker[2], Kirsten I. Bos[1], Alexander Herbig[1]

1 Max Planck Institute for the Science of Human History, Jena, Germany
2 Museum of London Archaeology (MOLA), London, United Kingdom

In preparation for submission

BED030 is an exceptionally well-preserved sample of a plague victim buried in the New Churchyard burial ground in London, which allowed for the *de novo* reconstruction of the genome of the plague causing agent bacterium *Yersinia pestis*. In this manuscript we evaluate different assembly tools and sequencing depths to maximise *de novo* assembly of an ancient plague genome in terms of cost and completeness of the genome by simulating ancient data. We found that the best performing strategy was using the SPAdes assembler with a depth of 200-fold coverage. We then re-sequenced the BED030 sample to achieve the desired coverage and performed SPAdes assembly, Ragout scaffolding and gap filling with GAPPadder. BED030 assembly had improved continuity than previously assembled ancient *Y. pestis* genomes. We observe some rearrangements in the BED030 compared to the reference. Furthermore, we

annotate the obtained genome with Prokka, which could be used as a resource for other studies. This study highlights the potential of obtaining *de novo* genomes from well-preserved ancient samples, allowing for the exploration of genomic architecture of genomes in the past.

AAV Contribution (85%):

- Planned the research
- Simulated ancient data and performed analysis to decide on the assembler and optimal sequencing depth to perform the assembly.
- Established the assembly pipeline.
- Performed comparative analysis with other assemblies and modern genomes.
- Generated all data tables and figures.
- Wrote the draft of the manuscript.

Manuscript E: Insights into the population structure and virulence of *Streptococcus mutans* from an ancient genome

Aida Andrades Valtueña, Alissa Mittnik, Saskia, Alexander Herbig, Johannes Krause

The chapter presented here will be published together with a currently ongoing project involving more ancient *S. mutans* genomes.

In Manuscript E, we explore the genetic composition of an ancient *Streptococcus mutans* obtained from a hunter-gathered individual that lived ~2000 years before present (calibrated). We retrieve a 156X fold shotgun *S. mutans* genome. We established an analytical framework for comparative genomics of ancient *S. mutans* genomes together with modern strains. This enabled us to explore the change in gene content, both in terms of presence and absence of known virulence factors and specific point-mutations; but also, to test for a phylogeographic signal by utilising phylogenetic and principal component analysis. During this work, we question how well suited is the modern *S. mutans* dataset for the testing of phylogeography and propose ancient DNA as a tool to retrieve a less biased dataset. Additionally, aDNA can provide temporal insights into the questions of phylogeography and genome content of *S. mutans*. To aid future retrieval of *S. mutans* in a cost-efficient manner, we design a capture-based on 191 modern genomes and test its efficiency and potential bias introduced by capturing the same individual. The main conclusions of this manuscript are: the ancient *S. mutans* cannot be distinguish from modern genomes, in terms of

absence of genes or genomic regions, pointing to no gain of new genetic components in *S. mutans* for the last 2000 years; we do not detect any phylogeography in *S. mutans*, however whether the phylogeographic signal is obscured by the current modern comparative dataset remains to be tested; the designed capture is efficient at retrieving ancient *S. mutans* DNA and does not introduce biases in the standard analysis performed to analyse ancient pathogens.

AAV Contribution (90%):

- Planned the research
- Bioinformatic analysis consisting of:
    - Basic mapping and SNP calling for the genome
    - Establishing an analytical framework for the study of *S. mutans*: scripts to check for heterozygosity, Principal Component Analysis, to check for presence/absence of virulence genes, phylogenetic analysis.
    - Gathered a modern reference panel for the generation of capture probes.
    - Evaluation of probe sets in terms of specificity using MALT
    - Analysed capture data and compared the results with the shotgun genome
- Generated all figures and tables in the manuscript.
- Wrote the draft of the manuscript.

Report

Current Biology

The Stone Age Plague and Its Persistence in Eurasia

Highlights

- Six Late Neolithic–Early Bronze Age European *Y. pestis* genomes were reconstructed
- All Late Neolithic and Early Bronze Age *Y. pestis* form a single phylogenetic branch

Authors

Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, ..., Philipp W. Stockhammer, Alexander Herbig, Johannes Krause

Correspondence

herbig@shh.mpg.de (A.H.), krause@shh.mpg.de (J.K.)

In Brief

Andrades Valtueña et al. present the first six European *Y. pestis* genomes dating from the Late Neolithic and the Early Bronze Age. These data suggest that *Y. pestis* entered Europe during a human migration around 4800 BP, persisted in Europe, and traveled back to Central Eurasia.

Andrades Valtueña et al., 2017, Current Biology *27*, 3683–3691
December 4, 2017
https://doi.org/10.1016/j.cub.2017.10.025

Current Biology
Report

The Stone Age Plague and Its Persistence in Eurasia

Aida Andrades Valtueña,[1] Alissa Mittnik,[1,2] Felix M. Key,[1] Wolfgang Haak,[1,3] Raili Allmäe,[4] Andrej Belinskij,[5] Mantas Daubaras,[6] Michal Feldman,[1,2] Rimantas Jankauskas,[7] Ivor Janković,[8,9] Ken Massy,[10,11] Mario Novak,[8] Saskia Pfrengle,[2] Sabine Reinhold,[12] Mario Šlaus,[13] Maria A. Spyrou,[1,2] Anna Szécsényi-Nagy,[14] Mari Tõrv,[15] Svend Hansen,[12] Kirsten I. Bos,[1,2] Philipp W. Stockhammer,[1,10] Alexander Herbig,[1,2,*] and Johannes Krause[1,2,16,*]
[1]Max Planck Institute for the Science of Human History, Jena, Germany
[2]Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, Tübingen, Germany
[3]School of Biological Sciences, The University of Adelaide, Adelaide SA 5005, South Australia, Australia
[4]Archaeological Research Collection, Tallinn University, Tallinn, Estonia
[5]"Nasledie" Cultural Heritage Unit, Stavropol, Russia
[6]Department of Archaeology, Lithuanian Institute of History, Vilnius, Lithuania
[7]Department of Anatomy, Histology and Anthropology, Vilnius University, Vilnius, Lithuania
[8]Institute for Anthropological Research, Zagreb, Croatia
[9]Department of Anthropology, University of Wyoming, Laramie, WY, USA
[10]Institute for Pre- and Protohistoric Archaeology and Archaeology of the Roman Provinces, Ludwig-Maximilians-University Munich, Munich, Germany
[11]Heidelberg Academy of Sciences, Heidelberg, Germany
[12]Eurasia Department, German Archaeological Institute, Berlin, Germany
[13]Anthropological Center, Croatian Academy of Sciences and Arts, Zagreb, Croatia
[14]Institute of Archaeology, Research Centre for the Humanities, Hungarian Academy of Sciences, Budapest 1097, Hungary
[15]Department of Archaeology, Institute of History and Archaeology, University of Tartu, Tartu, Estonia
[16]Lead Contact
*Correspondence: herbig@shh.mpg.de (A.H.), krause@shh.mpg.de (J.K.)
https://doi.org/10.1016/j.cub.2017.10.025

SUMMARY

Yersinia pestis, the etiologic agent of plague, is a bacterium associated with wild rodents and their fleas. Historically it was responsible for three pandemics: the Plague of Justinian in the 6th century AD, which persisted until the 8th century [1]; the renowned Black Death of the 14th century [2, 3], with recurrent outbreaks until the 18th century [4]; and the most recent 19th century pandemic, in which *Y. pestis* spread worldwide [5] and became endemic in several regions [6]. The discovery of molecular signatures of *Y. pestis* in prehistoric Eurasian individuals and two genomes from Southern Siberia suggest that *Y. pestis* caused some form of disease in humans prior to the first historically documented pandemic [7]. Here, we present six new European *Y. pestis* genomes spanning the Late Neolithic to the Bronze Age (LNBA; 4,800 to 3,700 calibrated years before present). This time period is characterized by major transformative cultural and social changes that led to cross-European networks of contact and exchange [8, 9]. We show that all known LNBA strains form a single putatively extinct clade in the *Y. pestis* phylogeny. Interpreting our data within the context of recent ancient human genomic evidence that suggests an increase in human mobility during the LNBA, we propose a possible scenario for the early spread of *Y. pestis*: the pathogen may have entered Europe from Central Eurasia following an expansion of people from the steppe, persisted within Europe until the mid-Bronze Age, and moved back toward Central Eurasia in parallel with human populations.

RESULTS

Screening

A total of 563 tooth and bone samples dating from the Late Neolithic to the Bronze Age (LNBA) from Russia (n = 122), Hungary and Croatia (n = 139), Lithuania (n = 27), Estonia (n = 45), Latvia (n = 10), and Germany (n = 220) were screened for *Y. pestis* by mapping reads ranging from 700,000 to 21,000,000 against a multi-fasta reference of 12 different *Yersinia* (Table 1).

To evaluate whether an individual was potentially *Y. pestis*-positive, we calculated a score based on the number of specific reads mapping to *Y. pestis* compared to other *Yersinia* (see STAR Methods). Individuals with a positive score were deemed potential candidates. Those with scores > 0.005 and reads mapping to all three *Y. pestis* plasmids were considered "strong" positives. We identified five "strong" candidates: one individual from Rasshevatskiy (RK1001; North Caucasus, Russia), one from Gyvakarai (Gyvakarai1; Lithuania), one from Kunila (Kunila II; Estonia), and two from Augsburg, Germany (Haunstetten, Unterer Talweg 85 Feature 1343 [1343UnTal85]; Haunstetten, Postillionstrasse Feature 6 [6Post]). One individual from Beli Manastir-Popova zemlja (GEN72; Croatia) did not pass the "strong" candidate threshold but was included by virtue of having the highest number of reads mapping to the *Y. pestis* chromosome and plasmids (chromosome = 993, pCD1 = 243,

Table 1. Genomes from the NCBI RefSeq/Nucleotide Database, Used in the Multi-species Reference Panel for Screening for *Y. pestis* aDNA

Species Name	Strain	NCBI Accession Number
Y. pestis	CO92	NC_003143.1
Y. pseudotuberculosis	IP 32953	NC_006155.1
Y. enterocolitca	subsp. enterocolitica 8081	NC_008800.1
Y. aldovae	ATCC 35236	NZ_ACCB01000210.1
Y. bercovieri	ATCC 43970	NZ_AALC02000229.1
Y. frederiksenii	ATCC 33641	NZ_AALE02000161.1
Y. intermedia	ATCC 29909	NZ_AALF02000123.1
Y. kristensenii	ATCC 33638	NZ_ACCA01000153.1
Y. mollaretii	ATCC 43969	NZ_AALD02000179.1
Y. rohdei	ATCC 43380	NZ_ACCD01000141.1
Y. ruckeri	ATCC 29473	NZ_ACCC01000174.1

pMT1 = 111, pPCP1 = 22). For additional archaeological information, see Table 2 and STAR Methods.

Genome Reconstruction

"Strong" positive individuals were shotgun sequenced to a depth of 379,155,741–1,529,935,532 reads. RK1001 and GEN72 were further enriched for *Y. pestis* DNA using in-solution capture (see STAR Methods). After mapping to the reference genome (*Y. pestis* CO92, NC_003143.1), we reconstructed genomes for all six candidates with a mean coverage between 3- and 12-fold, with 86%–94% of the reference covered 1-fold (Table 2, Figure S2). The reads were independently mapped to the three *Y. pestis* CO92 plasmids yielding mean coverages of 7- to 24-fold (pCD1), 3- to 14-fold (pMT1), and 18- to 43-fold (pPCP1; Table S1, Figure S2).

To authenticate the ancient origin of the bacterial genomes, we evaluated terminal deamination damage common to ancient DNA [14]. Our samples presented typical damage profiles similar to the corresponding associated human DNA (Figure S1), and multi-strain infection was not observed (Figure S1).

Phylogeny and Dating

To assess the phylogenetic position of the reconstructed genomes in comparison to modern and ancient *Y. pestis* genomes (see STAR Methods), we computed neighbor joining (Figure S3A), maximum parsimony (Figure S3B), and maximum likelihood (Figures 1 and S3C) trees. Our samples form a clade with the previously reported RISE509 and RISE505 strains [7], with a bootstrap support > 95% for all methods.

The branching point of the LNBA genomes and all other strains represents the most recent common ancestor (MRCA) of all currently available *Y. pestis* genomes, which was "tip-dated" using BEAST [15] to 6,078 years (95% highest posterior density interval: 5,036–7,494 years), in agreement with previous estimates [7]. The time to the MRCA of *Y. pestis* and *Y. pseudotuberculosis* strain IP32953 was estimated at 28,258 years (95% highest posterior density interval: 13,200–44,631 years). The maximum clade credibility tree (Figure S4A) supports the same topology as the methods described above, with high statistical support of the LNBA branching point.

Genetic Makeup

We identified 423 single-nucleotide polymorphisms (SNPs) on the LNBA branch (Data S1, sheets 1–4), including strain-specific and shared SNPs, of which 114 are synonymous and 202 are non-synonymous (see STAR Methods). The LNBA genomes share five SNPs (Data S1, sheet 2).

The percent coverage was calculated for genes related to virulence, flea transmission, colonization, and dissemination (Figure 1B). The YpfΦ prophage [16], integrated only into the chromosomes of the 1.ORI strains responsible for the third pandemic [17], was absent in all LNBA genomes. Additionally, *yapC*, possibly involved in the adhesion to mammalian cells, autoagglutination, and biofilm formation [18], was lost in the three youngest LNBA strains (1343UnTal85, Post6, and RISE505).

The only plasmid virulence factor missing in the LNBA strains is *ymt* (Figures 1B and S2). *ymt* codes for the *Yersinia* murine toxin, an important virulence factor in flea transmission [19, 20]. Expression of *ymt* protects against toxic blood digestion byproducts and permits colonization of the flea midgut [20]. Other plasmid virulence factors such as *pla* and *caf1*, absent in *Y. pseudotuberculosis*, were already present in the LNBA *Y. pestis* strains.

Urease D (*ureD*) plays an important role in flea transmission. *ureD* expression causes a toxic oral reaction killing 30%–40% [21] of infected fleas. *ureD* is a pseudogene in *Y. pestis* due to a frameshift mutation [22]. Close inspection of this gene revealed that the frameshift is not present, indicating that this gene was functional in the LNBA strains and possibly making them as toxic to fleas as their ancestor *Y. pseudotuberculosis*.

Furthermore, *Y. pseudotuberculosis*-specific regions that have been lost in *Y. pestis* were still present in the LNBA strains (Data S1, sheet 5). We also observed genome decay in the LNBA clade mostly affecting flagellin genes and membrane proteins (Figure S2; Data S1, sheet 5).

DISCUSSION

The prehistoric genomes presented here are the first to reveal *Y. pestis* diversity in the European LNBA. This complements contemporary *Y. pestis* genomes from Bronze Age individuals recovered from Southern Siberia [7] and offers higher resolution to evaluate the evolution and dissemination of prehistoric plague strains. All LNBA genomes, including those previously reconstructed from Southern Siberia [7], form a distinct clade. The strains RISE509 [7] and RK1001 occupy the most basal position of all *Y. pestis* genomes sequenced to date, formally tested with CONSEL [23] (Figure S4B). These data are compatible with two scenarios. In scenario 1, plague was introduced multiple times to Europe from a common reservoir between 5,000 and 3,000 BP. The bacterium spread to Europe from a source most likely located in Central Eurasia at least four times during a period of over 1,000 years: once to Lithuania and Croatia, once to Estonia, and twice to Germany. In this model, the phylogeny of the LNBA lineages results exclusively from their temporal relationship. In scenario 2, plague entered Europe from Central Eurasia once during the Neolithic. A reservoir was established within or close

Table 2. Statistics of the *Y. pestis* Genome Reconstruction

Individual	Tissue Sampled	Site	Country	Radiocarbon Date (14dC)	Dating (Median cal BP)	2-Sigma Interval [cal BP]; [cal BC]	In-Solution Capture	Clipped, Merged, and Quality-Filtered Reads before Mapping	Unique Reads Mapping to *Y. pestis* Reference	Endogenous DNA (%)	Mean Coverage	Coverage (%)		Publication
												≥1×	≥3×	
RK1001	Tooth	Rasshevatskiy	Russia	4,171 ± 22	4720	[4,828–4,622]; [2,879–2,673]	no	1,529,935,532	119,540	0.01	1.0213	58.11	10.87	This study
RK1001	Tooth	Rasshevatskiy	Russia	4,171 ± 22	4720	[4,828–4,622]; [2,879–2,673]	yes	303,148,884	383,900	0.85	3.3984	82.83	55.3	This study
RK1001	Tooth	Rasshevatskiy	Russia	4,171 ± 22	4720	[4,828–4,622]; [2,879–2,673]	combined shotgun/capture	1,833,084,416	418,581	0.17	3.6816	86.16	59.63	This study
GEN72	Tooth	Beli Manastir-Popova zemlja	Croatia	4,176 ± 28	4721	[4,833–4,592]; [2,884–2,640]	yes	19,777,683	1,321,320	24.36	12.6549	91.65	86.61	[10]
Gyvakarai1	Tooth	Gyvakarai	Lithuania	4,030 ± 30	4485	[4,571–4,422]; [2,578–2,491]	no	1,021,452,137	473,207	0.05	5.2245	94.07	84.12	[11]
KunilaII	Tooth	Kunila	Estonia	3,960 ± 40	4427	[4,524–4,290]; [2,576–2,340]	no	379,155,741	546,243	0.16	5.5418	92.48	77.58	[12]
1343UnTal85	Tooth	Augsburg	Germany	3,819 ± 24	4203	[4,346–4,098]; [2,397–2,149]	no	1,174,989,269	1,165,435	0.14	10.5745	93.69	92.59	[13]
6Post	Tooth	Augsburg	Germany	3,574 ± 19	3873	[3,957–3,832]; [2,009–1,883]	no	419,717,299	598,030	0.17	5.3062	89.71	71.14	[13]

The radiocarbon dates were calibrated with Calib 7.1. calBP = calibrated years Before Present; cal BC = calibrated years Before Christ. All individuals were directly radiocarbon dated. See also Table S1.

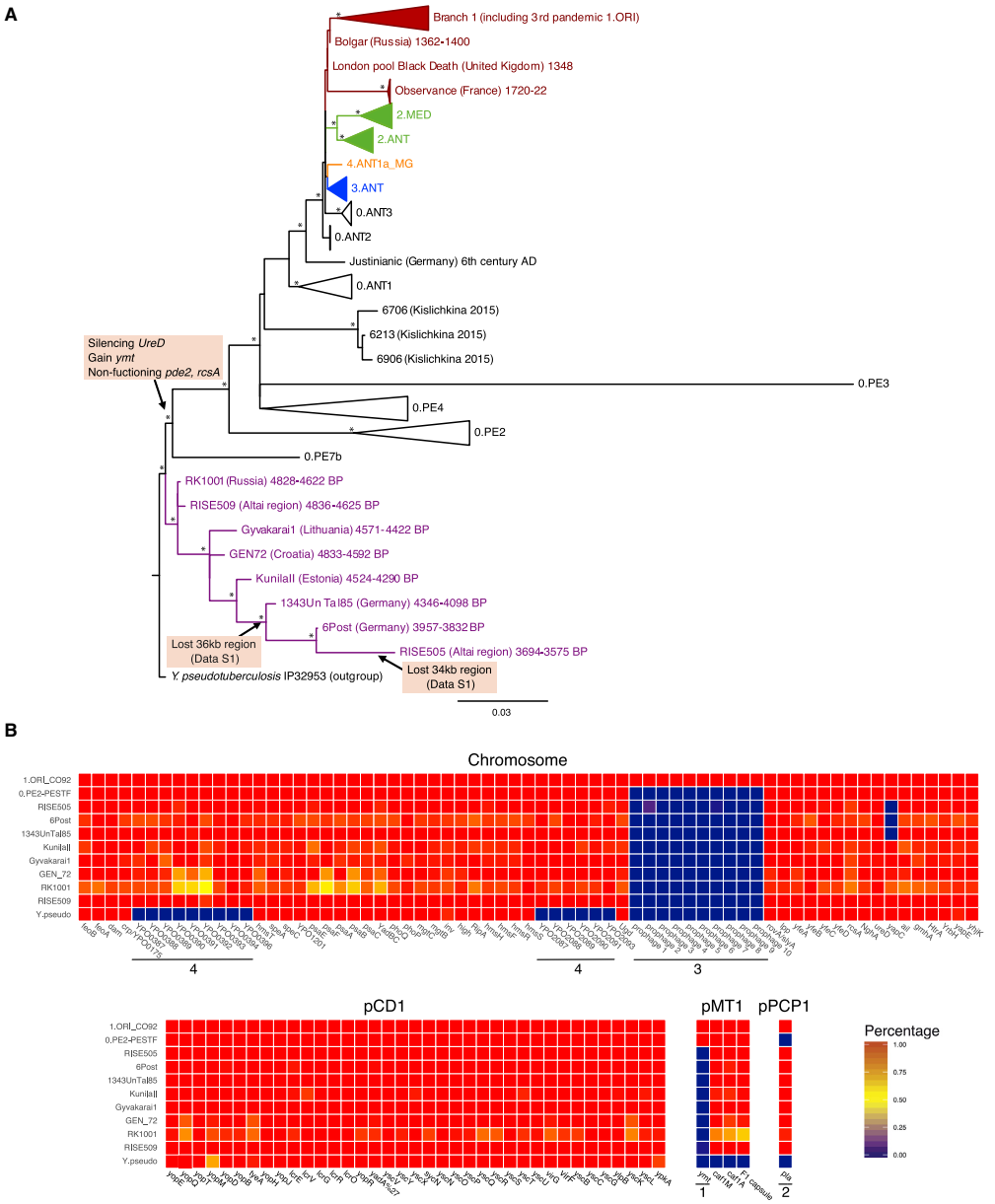


Figure 1. Maximum-Likelihood Tree and Percent Coverage Plot of Virulence Factors of *Yersinia pestis*

(A) Maximum-likelihood tree of all *Yersinia pestis* genomes, including 1,265 SNP positions with complete deletion. Nodes with support ≥ 95% are marked with an asterisk. The colors represent different branches in the *Y. pestis* phylogeny: branch 0 (black), branch 1 (red), branch 2 (green), branch 3 (blue), branch 4 (orange),

(legend continued on next page)

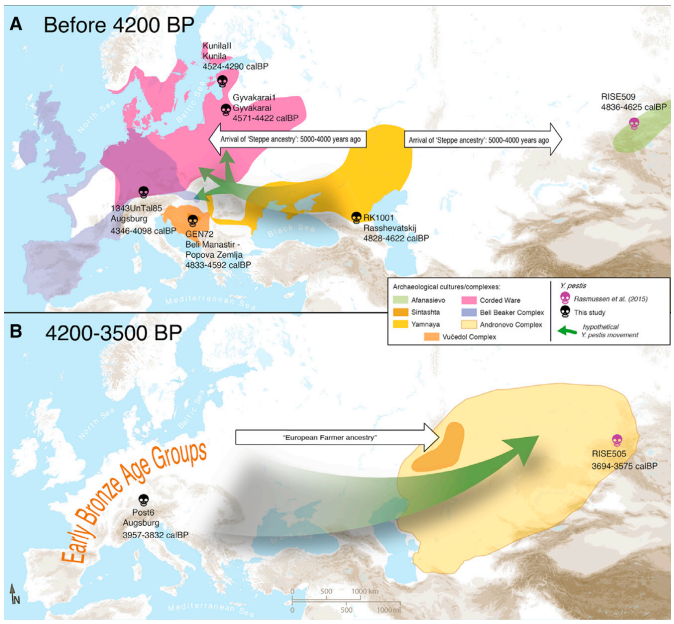


Figure 2. Map of Proposed *Yersinia pestis* Circulation throughout Eurasia

(A) Entrance of *Y. pestis* into Europe from Central Eurasia with the expansion of Yamnaya pastoralists around 4,800 years ago.

(B) Circulation of *Y. pestis* to Southern Siberia from Europe. Only complete genomes are shown.

to Europe from which it circulated, and ultimately it moved back to Central Eurasia during the Bronze Age (Figure 2).

With few genomes available, it is difficult to disentangle the two hypotheses; however, interpreting our data in the context of human genetics and archaeological data can offer some resolution. Ancient human genomic data point to a change in mobility and large-scale expansion of people from the Caspian-Pontic Steppe associated with the Yamnaya complex, both east and west starting around 4,800 BP. These people carried a distinct genetic component that is also seen in highly mobile groups associated with the Southern Siberian Afanasievo complex, the Yamnaya complex, and the Central and Eastern European Corded Ware complex [24]. In Central European individuals, it is first observed in the Corded Ware complex and then becomes part of the genetic composition of most subsequent and all modern-day European populations [24, 25].

Our earliest indication of plague in Europe is found in Croatia and the Baltic, coinciding with the arrival of "steppe ancestry" [24, 25] in human populations. The Baltic Late Neolithic *Y. pestis* genomes (Gyvakarai1 and KunilaII) were reconstructed from individuals associated with the Corded Ware complex. Along with the Croatian *Y. pestis* genome (Vučedol complex), these are derived from a common ancestor shared with the Yamnaya-derived RK1001 and Afanasievo-derived RISE509. This supports the notion of the pathogen spreading in the context of the large-scale expansion of steppe peoples from Central Eurasia to Eastern and Central Europe. Furthermore, human genomic analyses indicate that RISE509, Gyvakarai1, KunilaII, and GEN72 carry "steppe ancestry" [10, 11, 24]. Evidence for such long-distance contact is also present in the archaeological record. For example, the Gyvakarai1 burial is characterized by a specific inventory of grave items (e.g. hammer-headed pins) and distinct skeletal morphology that have no analogs in earlier local populations [26].

The younger Late Neolithic *Y. pestis* genomes from southern Germany are derived from the Baltic strains, and one of these is found in an individual associated with the Bell Beaker complex. Previous analyses have shown that Bell Beaker individuals from Germany also carry "steppe ancestry" [24, 25, 27]. This suggests that *Y. pestis* may have spread further southwest, analogous to the human "steppe" component. The youngest of the LNBA *Y. pestis* genomes (RISE505, Southern Siberia) associated with the Central Eurasian Andronovo complex, descends from the Central European strains, suggesting a spread into Southern Siberia. Interestingly, genome-wide human data show that individuals associated with the Sintashta, Srubnaya, and Andronovo cultural complexes in the Eurasian steppes (dating around 3,700–3,300 calibrated years before present) carried mixed ancestry of middle Neolithic European farmers and Bronze Age steppe people, suggesting a backflow of human genes from Europe to Central Eurasia [24]. Archaeologically, there seems to be a close connection between the Eastern European Abashevo cultural complex and Sintashta that might have included population shifts from west to east. In particular, the post-Sintashta Andronovo complex is an epoch of population shifts affecting all areas east of the Urals to the western borders of China, including populations with European origin [28, 29]. The steppe, a natural corridor connecting people and their livestock throughout Central and Western Eurasia, might have facilitated the spread of strains related to the European Early Bronze Age *Y. pestis* to Southern Siberia, where RISE505 was found. In our view, human genetic ancestry and admixture, in combination with the temporal series within the

and LNBA *Y. pestis* branch (purple). *Y. pseudotuberculosis*-specific SNPs were excluded from the tree for clarity of representation. In the light-colored boxes, discussed losses and gains of genomic regions and genes are indicated. Related to Figure S3.

(B) Percent coverage of virulence factors located on the *Yersinia pestis* chromosome and plasmids, plotted in R using the ggplot2 package. The numbers represent specific genes: (1) *ymt* gene, (2) *pla* gene, (3) filamentous prophage YpfΦ, (4) *Y. pestis*-specific genes. Related to Figure S2. See also Data S1.

LNBA *Y. pestis* branch, supports scenario 2. *Y. pestis* was possibly introduced to Europe from the steppe around 4,800 BP. Thereafter, a local reservoir was established within or in close proximity to Europe. The European *Y. pestis* strain was disseminated to Southern Siberia potentially through anthropogenic processes connected to the backflow of human genetic ancestry from Western Eurasia into Southern Siberia. The pathogen diversity mirrors the archaeological evidence, which indicates intensification of Eurasian trade networks from the beginning of the Bronze Age [8, 9].

Even though *Y. pestis* seems to have spread in patterns strikingly similar to human movements (Figure 2), the mode of transmission during this early phase of its evolution cannot be easily determined. Most contemporary cases of *Y. pestis* infection occur via a flea vector and stem from sylvatic rodent populations with resistance to the bacterium. Flea transmission is accomplished by one of two mechanisms [30]: blockage-dependent transmission [31] or early-phase transmission (EPT) [32]. In the former, *Y. pestis* obstructs the flea digestive system by producing a biofilm that blocks the flea's foregut within 1–2 weeks post-infection. This blockage prevents a blood meal from reaching the flea's midgut, and blood regurgitation during failed feeding sheds live bacteria into the host [31, 33]. The blockage-dependent transmission requires a functional *ymt* gene and *hms* locus, and non-functional *rcsA*, *pde2*, and *pde3* genes [34]. *ymt* protects *Y. pestis* within the digestive system of the flea, allowing colonization of the flea midgut. The *hms* locus is involved in biofilm formation, and *rcsA*, *pde2*, and *pde3* are biofilm downregulators. However, *Y. pestis* can be transmitted within the first 1–4 days after entering the flea prior to colonization of the midgut and biofilm formation [32, 35] (the EPT model). This model is currently less understood than blockage-dependent transmission but has been shown to be both biofilm [36] and *ymt* independent [37].

The genetic characteristics of the LNBA genomes (i.e., lack of *ymt*, functional *pde2*, and *rcsA*) were previously interpreted as evidence that early forms of *Y. pestis* were unable to cause blockage in the flea gut, thus suggesting that the bubonic form of the disease evolved later [7]. However, as none of these genes seem to be required for EPT, one cannot exclude that LNBA *Y. pestis* infections could have been acquired from fleas via this transmission mode. Under this model, transmission would have been less efficient since a functional UreD would have reduced the number of flea vectors by 30%–40%.

The presence of mammalian virulence-related genes such as *pla* and *caf1* indicates that LNBA *Y. pestis* was to some extent adapted to these hosts. The LNBA *pla* presents the ancestral I259 variant, shown to be less efficient in infiltrating the host [38, 39] than the derived T259 form [40]. Strains carrying the ancestral variant can cause pneumonic disease but are less efficient in colonizing other tissues [38]. This indicates that LNBA *Y. pestis* could have caused a pneumonic or less severe bubonic form. The genome decay we detected, affecting membrane and flagellar proteins possibly involved in interactions with the host's immune system, could indicate adaptation to new hosts or pathogenic lifestyles [41].

Modern plague is a rodent-adapted disease, in which commensal species such as *Rattus rattus* and their fleas play a central role as disease vectors for humans [42]. Although rodent-flea transmission is compatible with the genomic makeup of the LNBA strains, disease dynamics may have differed in the past. The Neolithic is considered to be a time period in which new diseases were introduced into human groups during transition from a mobile to sedentary lifestyle. Adoption of agriculture and increased population density are thought to have acted synergistically to change the disease landscape [43]. Whether commensal rodent populations were large enough to function as a plague reservoir during human migrations at this time is unknown. In Central Eurasian Bronze Age cultures, agriculture (i.e., large-scale food storage) was mostly absent [44]. However, contact between sylvatic rodents in the steppe, pastoralists, and their herds might have been frequent in these environments. Alternative models of transmission involving different host species, perhaps even humans or their livestock, might carry some traction, as the ancient disease may have behaved differently from what we know today.

The LNBA was a time of increased mobility and cultural change. The threat of *Y. pestis* infections may have been one of the causes for this increased mobility [7]. Further sampling of skeletal material could provide much-needed details about the range and frequency of *Y. pestis* infections during this transformative period. Presence of the disease in Europe could have played a role in the processes that led to the genetic turnover observed in European human populations, who may have harbored different levels of immunity against this disease. Testing these hypotheses will require extensive assessment of both human and *Y. pestis* genomes from the steppes and from Europe before and after the steppe migration.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Ethical approvals
  - Description of samples and archaeological sites
- METHOD DETAILS
  - Sampling and extraction
  - Shotgun screening sequencing
  - *In silico* screening
  - Deep shotgun sequencing
  - *Y. pestis* in-solution capture
  - Genome reconstruction and authentication
  - Individual sample treatment due to laboratory preparation and sequencing strategies
  - SNP calling, heterozygosity, and phylogenetic analysis
  - Dating analysis
  - SNP effect analysis and virulence factors analysis
  - Indel analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - *In silico* screening
  - Phylogenetic analysis
  - Tree topology test
  - Molecular clock test
  - Dating analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, one table, and one dataset and can be found with this article online at https://doi.org/10.1016/j.cub.2017.10.025.

AUTHOR CONTRIBUTIONS

J.K., A.H., and A.A.V. conceived the study. K.M., R.A., M.D., R.J., M.T., P.W.S., A.B., I.J., M.N., S.R., M.S., A.S.-N., and S.H. provided samples and performed archaeological assessment. A.M., S.P., M.F., A.A.V., and A.S.-N. performed laboratory work. A.A.V., A.H., M.A.S., F.M.K., and J.K. analyzed the data. A.A.V., A.H., J.K., P.W.S., K.I.B., W.H., and A.M. wrote the manuscript with contributions from all authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Corina Knipper, Ernst Pernicka, Stephanie Metz, Fabian Wittenborn, Stephan Schiffels, Joris Peters, Michaela Harbeck, and Archaeogenetics Department members of the MPI-SHH for helpful discussion. We thank Annette Günzel for graphical support. We thank Isil Kucukkalipci, Antje Wissgott, Marta Burri, and Franziska Göhringer for technical support in the lab. We thank Joachim Wahl, Josip Burmaz and Dženi Los, and Gunita Zatina and Andrejs Vasks for kindly providing the Althausen, Croatian, and Latvian samples, respectively. We thank Natalia Berezina and Julia Gresky for the anthropological assessment of RK1001. We thank James A. Fellows Yates for proofreading. The Heidelberg Academy of Science financed the genetic and archaeological research on human individuals from the Augsburg region within the project WIN Kolleg: "Times of Upheaval: Changes of Society and Landscape at the Beginning of the Bronze Age." This work was financially supported by the Max Planck Society, European Research Council starting grant APGREID (to J.K.), and Croatian Science Foundation grant 1450 (to M.N. and I.J.).

Received: May 16, 2017
Revised: July 31, 2017
Accepted: October 9, 2017
Published: November 22, 2017

REFERENCES

1. Russell, J.C. (1968). That earlier plague. Demography *5*, 174–184.
2. Zietz, B.P., and Dunkelberg, H. (2004). The history of the plague and the research on the causative agent *Yersinia pestis*. Int. J. Hyg. Environ. Health *207*, 165–178.
3. Benedictow, O.J. (2004). The Black Death, 1346–1353: The Complete History (Boydell Press).
4. Cohn, S.K., Jr. (2008). Epidemiology of the Black Death and successive waves of plague. Med. Hist. Suppl. *2008*, 74–100.
5. Stenseth, N.C., Atshabar, B.B., Begon, M., Belmain, S.R., Bertherat, E., Carniel, E., Gage, K.L., Leirs, H., and Rahalison, L. (2008). Plague: past, present, and future. PLoS Med. *5*, e3.
6. World Health Organization (2017). Plague fact sheet. http://www.who.int/mediacentre/factsheets/fs267/en/.
7. 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., et al. (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. Cell *163*, 571–582.
8. Vandkilde, H. (2016). Bronzization: the Bronze Age as pre-modern globalization. Praehist. Z. *91*, 103–123.
9. Hansen, S. (2014). The 4th Millennium: a watershed in European prehistory. In Western Anatolia before Troy: Proto-Urbanisation in the 4th Millennium BC?, B. Horejs, and M. Mehofer, eds. (Vienna: Österreichische Akademie der Wissenschaften), pp. 243–260.
10. Mathieson, I., Roodenberg, S.A., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olade, I., Broomandkhoshbacht, N., Cheronet, O., Fernandes, D., et al. (2017). The Genomic History Of Southeastern Europe. bioRxiv. https://doi.org/10.1101/135616.
11. Mittnik, A., Wang, C.-C., Pfrengle, S., Daubaras, M., Zariņa, G., Hallgren, F., Allmäe, R., Khartanovich, V., Moiseyev, V., Furtwängler, A., et al. (2017). The Genetic History of Northern Europe. bioRxiv. https://doi.org/10.1101/113241.
12. Kriiska, A., Lõugas, L., Lõhmus, M., Mannermaa, K., and Johanson, K. (2007). New AMS dates from Estonian Stone Age burials sites. Estonian J. Archaeol. *11*, 83–121.
13. Stockhammer, P.W., Massy, K., Knipper, C., Friedrich, R., Kromer, B., Lindauer, S., Radosavljević, J., Wittenborn, F., and Krause, J. (2015). Rewriting the Central European Early Bronze Age Chronology: Evidence from Large-Scale Radiocarbon Dating. PLoS ONE *10*, e0139705.
14. Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. USA *104*, 14616–14621.
15. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.
16. Derbise, A., Chenal-Francisque, V., Pouillot, F., Fayolle, C., Prévost, M.-C., Médigue, C., Hinnebusch, B.J., and Carniel, E. (2007). A horizontally acquired filamentous phage contributes to the pathogenicity of the plague bacillus. Mol. Microbiol. *63*, 1145–1157.
17. Derbise, A., and Carniel, E. (2014). YpfΦ: a filamentous phage acquired by *Yersinia pestis*. Front. Microbiol. *5*, 701.
18. Felek, S., Lawrenz, M.B., and Krukonis, E.S. (2008). The *Yersinia pestis* autotransporter YapC mediates host cell binding, autoaggregation and biofilm formation. Microbiology *154*, 1802–1812.
19. Hinnebusch, J., Cherepanov, P., Du, Y., Rudolph, A., Dixon, J.D., Schwan, T., and Forsberg, A. (2000). Murine toxin of *Yersinia pestis* shows phospholipase D activity but is not required for virulence in mice. Int. J. Med. Microbiol. *290*, 483–487.
20. Hinnebusch, B.J., Rudolph, A.E., Cherepanov, P., Dixon, J.E., Schwan, T.G., and Forsberg, A. (2002). Role of Yersinia murine toxin in survival of *Yersinia pestis* in the midgut of the flea vector. Science *296*, 733–735.
21. Chouikha, I., and Hinnebusch, B.J. (2014). Silencing urease: a key evolutionary step that facilitated the adaptation of *Yersinia pestis* to the flea-borne transmission route. Proc. Natl. Acad. Sci. USA *111*, 18709–18714.
22. Sebbane, F., Devalckenaere, A., Foulon, J., Carniel, E., and Simonet, M. (2001). Silencing and reactivation of urease in *Yersinia pestis* is determined by one G residue at a specific position in the ureD gene. Infect. Immun. *69*, 170–176.
23. Shimodaira, H. (2001). Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. Comm. Stat. A Theory Methods *30*, 1751–1772.
24. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.
25. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211.
26. Tebelškis, P., and Jankauskas, R. (2006). The Late Neolithic grave at Gyvakarai in Lithuania in the context of current archaeological and anthropological knowledge. Archaeol. Baltica, 8–20.
27. Olalde, I., Brace, S., Allentoft, M.E., Armit, I., Kristiansen, K., Rohland, N., Mallick, S., Booth, T., Szécsényi-Nagy, A., Mittnik, A., et al. (2017). The Beaker Phenomenon and the Genomic Transformation of Northwest Europe. bioRxiv. https://doi.org/10.1101/135962.

28. Kuzmina, E.E. (2008). The Prehistory of the Silk Road (University of Pennsylvania Press).
29. Koryakova, L., and Epimakhov, A.V. (2007). The Urals and Western Siberia in the Bronze and Iron Ages (Cambridge University Press).
30. Hinnebusch, B.J. (2012). Biofilm-dependent and biofilm-independent mechanisms of transmission of *Yersinia pestis* by fleas. In Advances in Yersinia Research, A.M.P. de Almeida, and N.C. Leal, eds. (New York: Springer), pp. 237–243.
31. Hinnebusch, B.J., Fischer, E.R., and Schwan, T.G. (1998). Evaluation of the role of the *Yersinia pestis* plasminogen activator and other plasmid-encoded factors in temperature-dependent blockage of the flea. J. Infect. Dis. *178*, 1406–1415.
32. Eisen, R.J., Bearden, S.W., Wilder, A.P., Montenieri, J.A., Antolin, M.F., and Gage, K.L. (2006). Early-phase transmission of *Yersinia pestis* by unblocked fleas as a mechanism explaining rapidly spreading plague epizootics. Proc. Natl. Acad. Sci. USA *103*, 15380–15385.
33. Chouikha, I., and Hinnebusch, B.J. (2012). Yersinia–flea interactions and the evolution of the arthropod-borne transmission route of plague. Curr. Opin. Microbiol. *15*, 239–246.
34. Sun, Y.-C., Jarrett, C.O., Bosio, C.F., and Hinnebusch, B.J. (2014). Retracing the evolutionary path that led to flea-borne transmission of *Yersinia pestis*. Cell Host Microbe *15*, 578–586.
35. Eisen, R.J., Dennis, D.T., and Gage, K.L. (2015). The Role of Early-Phase Transmission in the Spread of *Yersinia pestis*. J. Med. Entomol. *52*, 1183–1192.
36. Vetter, S.M., Eisen, R.J., Schotthoefer, A.M., Montenieri, J.A., Holmes, J.L., Bobrov, A.G., Bearden, S.W., Perry, R.D., and Gage, K.L. (2010). Biofilm formation is not required for early-phase transmission of *Yersinia pestis*. Microbiology *156*, 2216–2225.
37. Johnson, T.L., Hinnebusch, B.J., Boegler, K.A., Graham, C.B., MacMillan, K., Montenieri, J.A., Bearden, S.W., Gage, K.L., and Eisen, R.J. (2014). Yersinia murine toxin is not required for early-phase transmission of *Yersinia pestis* by Oropsylla montana (Siphonaptera: Ceratophyllidae) or Xenopsylla cheopis (Siphonaptera: Pulicidae). Microbiology *160*, 2517–2525.
38. Lathem, W.W., Price, P.A., Miller, V.L., and Goldman, W.E. (2007). A plasminogen-activating protease specifically controls the development of primary pneumonic plague. Science *315*, 509–513.
39. Sebbane, F., Jarrett, C.O., Gardner, D., Long, D., and Hinnebusch, B.J. (2006). Role of the *Yersinia pestis* plasminogen activator in the incidence of distinct septicemic and bubonic forms of flea-borne plague. Proc. Natl. Acad. Sci. USA *103*, 5526–5530.
40. Zimbler, D.L., Schroeder, J.A., Eddy, J.L., and Lathem, W.W. (2015). Early emergence of *Yersinia pestis* as a severe respiratory pathogen. Nat. Commun. *6*, 7487.
41. Ochman, H., and Moran, N.A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. Science *292*, 1096–1099.
42. Perry, R.D., and Fetherston, J.D. (1997). *Yersinia pestis*–etiologic agent of plague. Clin. Microbiol. Rev. *10*, 35–66.
43. Barrett, Ronald, Kuzawa, Christopher W., McDade, Thomas, and Armelagos, G.J. (1998). Emerging and re-emerging infectious diseases: the third epidemiologic transition. Annu. Rev. Anthropol. *27*, 247–271.
44. Ryabogina, N.E., and Ivanov, S.N. (2011). Ancient agriculture in Western Siberia: problems of argumentation, paleoethnobotanic methods, and analysis of data. Archaeol. Ethnol. Anthropol. Eurasia *39*, 96–106.
45. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, t5448.
46. Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. Genome Biol. *17*, 60.
47. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
48. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
49. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.
50. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.
51. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294.
52. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinformatics *11*, 11.10.1–11.10.33.
53. Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature *514*, 494–497.
54. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. *30*, 2725–2729.
55. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.
56. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.
57. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and Mclnerney, J.O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. *6*, 29.
58. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
59. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
60. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer).
61. R Development Core Team (2008). R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing), http://www.R-project.org.
62. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.
63. Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L.A., Wang, Z., Guo, Z., Xu, L., et al. (2013). Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. Proc. Natl. Acad. Sci. USA *110*, 577–582.
64. Zhgenti, E., Johnson, S.L., Davenport, K.W., Chanturia, G., Daligault, H.E., Chain, P.S., and Nikolich, M.P. (2015). Genome Assemblies for 11 *Yersinia pestis* Strains Isolated in the Caucasus Region. Genome Announc. *3*, e01030-15.
65. Kislichkina, A.A., Bogun, A.G., Kadnikova, L.A., Maiskaya, N.V., Platonov, M.E., Anisimov, N.V., Galkina, E.V., Dentovskaya, S.V., and Anisimov, A.P. (2015). Nineteen Whole-Genome Assemblies of *Yersinia pestis* subsp. *microtus*, Including Representatives of Biovars caucasica, talassica,

hissarica, altaica, xilingolensis, and ulegeica. Genome Announc. *3*, e01342-15.

66. 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., et al. (2011). A draft genome of *Yersinia pestis* from victims of the Black Death. Nature *478*, 506–510.
67. Feldman, M., Harbeck, M., Keller, M., Spyrou, M.A., Rott, A., Trautmann, B., Scholz, H.C., Päffgen, B., Peters, J., McCormick, M., et al. (2016). A High-Coverage *Yersinia pestis* Genome from a Sixth-Century Justinianic Plague Victim. Mol. Biol. Evol. *33*, 2911–2923.
68. Spyrou, M.A., Tukhbatova, R.I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A.G., Castex, D., Wahl, J., et al. (2016). Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. Cell Host Microbe *19*, 874–881.
69. Bos, K.I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S.A., Klunk, J., Schuenemann, V.J., Poinar, D., Kuch, M., et al. (2016). Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus. eLife *5*, e12994.
70. Rostunov, V.L. (2006). Opyt prekonstrukcii sakral'nogo postranstva rannykh kurganov Evropy I Severogo Kavkaza (Vladikavkaz: Severo-Oseinskiy institute gumannitarnykh I sozial'nykh issledovanny).
71. Rostunov, V.L. (2007). Epokha eneolita - sredney bronzy Zentralnogo Kavkaza III Opyt prekonstrukcii sakral'nogo postranstva rannykh kurganov Evropy (Vladikavkaz: Severo-Oseinskiy institute gumannitarnykh I sozial'nykh issledovanny).
72. Jaanits, L. (1948). Aruanne kaevamistest Kursi khk-s ja vallas Kunila külas Mäe-Jaaniantsu e. Keldri talu piirides asuval Jaaniantsu mäel 5.-10. (Tallinn: Institute of History of Tallinn University).
73. Jaanits, L., Laul, S., Lõugas, V., and Tõnisson, E. (1982). Eesti esiajalugu (Tallinn: Eesti Raamat).
74. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. USA *110*, 15758–15763.
75. Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.
76. Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest, S., Coombes, B.K., Wood, J.W., Earn, D.J.D., et al. (2011). Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. Proc. Natl. Acad. Sci. USA *108*, E746–E752.
77. Briggs, A., and Heyn, P. (2012). Preparation of next-generation sequencing libraries from damaged DNA. In Ancient DNA Methods in Molecular Biology, B. Shapiro, and M. Hofreiter, eds. (Humana Press), pp. 143–154.
78. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20130624.
79. Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., and Pääbo, S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proc. Natl. Acad. Sci. USA *110*, 2223–2227.
80. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. Nature *528*, 499–503.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human archaeological remains	This paper	ENA: PRJEB19335, https://www.ebi.ac.uk/ena/data/view/PRJEB19335
Chemicals, Peptides, and Recombinant Proteins		
0.5 M EDTA pH 8.0	Life Technologies	Cat No./ID: AM9261
1x Tris-EDTA pH 8.0	AppliChem	Cat No./ID: A8569,0500
Proteinase K	Sigma-Aldrich	Cat No./ID: P2308-100MG
Guanidine hydrochloride	Sigma-Aldrich	Cat No./ID: G3272-500 g
3M Sodium Acetate pH 5.5	Sigma-Aldrich	Cat No./ID: S7899-500ML
Ethanol	Merck	Cat No./ID: 1009832511
Isopropanol	Merck	Cat No./ID: 1070222511
ATP	New England Biosciences	Cat No./ID: P0756 S
BSA 20mg/ml	New England Biosciences	Cat No./ID: B9000 S
Bst 2.0 DNA Polymerase	New England Biosciences	Cat No./ID: M0537 S
Buffer Tango	Life Technologies	Cat No./ID: BY5
dNTPs 25 mM	Thermo Scientific	Cat No./ID: R1121
Ethanol	Merck	Cat No./ID: 1009832511
NEBuffer 2 10x	New England Biosciences	Cat No./ID: B7002 S
T4 DNA Polymerase	New England Biosciences	Cat No./ID: M0203 L
T4 Polynucleotide Kinase	New England Biosciences	Cat No./ID: M0201 L
Pfu Turbo Cx Hotstart DNA Polymerase	Agilent Technologies	Cat No./ID: 600412
Tween 20	Sigma-Aldrich	Cat No./ID: P9416-50ML
Uracil Glycosylase inhibitor (UGI)	New England Biosciences	Cat No./ID: M0281 S
User Enzyme	New England Biosciences	Cat No./ID: M5505 L
Water Chromasolv Plus	Sigma-Aldrich	Cat No./ID: 34877-2.5L
Critical Commercial Assays		
Min Elute PCR Purification Kit	QIAGEN	Cat No./ID:28006
Quick Ligation Kit	New England Biosciences	Cat No./ID: M2200 L
DyNAmo Flash SYBR Green qPCR Kit	Life Technologies	Cat No./ID: F-415L
SureSelect DNA Capture Arrays 1M	Agilent Technologies	Cat No./ID:G3358A
High Pure Viral Nucleic Acid Large Volume Kit	Roche	Cat No./ID: 5114403001
Oligonucleotides		
IS1_adapter.P5 A*C*A*C*TCTTTCCCTACACG ACGCTCTTCCG*A*T*C*T	[45]	Sigma Aldrich
IS2_adapter.P7 G*T*G*A*CTGGAGTTCAGAC GTGTGCTCTTCCG*A*T*C*T	[45]	Sigma Aldrich
IS3_adapter.P5+P7 A*G*A*T*CGGAA*G*A*G*C	[45]	Sigma Aldrich
P5 Indexing 5′-AATGATACGGCGACCACCGAGATCT ACACxxxxxxxxACACTCTTTCCCTACACGACGC-3′	[45]	Sigma Aldrich
P7 Indexing 5′-CAAGCAGAAGACGGCATACGAGA TxxxxxxxxGTGACTGGAGTTCAGACGTGTGC-3′	[45]	Sigma Aldrich
Deposited Data		
Y. pestis LNBA aDNA data	This study	ENA: PRJEB19335, https://www.ebi.ac.uk/ena/data/view/PRJEB19335

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
EAGER	[46]	https://github.com/apeltzer/EAGER-GUI
Burrows-Wheeler Aligner (BWA)	[47]	http://bio-bwa.sourceforge.net/; RRID: SCR_010910
samtools	[48]	http://samtools.sourceforge.net/; RRID: SCR_002105
MarkDuplicates (Picard)	http://broadinstitute.github.io/picard/	http://broadinstitute.github.io/picard/; RRID: SCR_006525
Dustmasker (BLAST+)	[49]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
MapDamage	[50]	https://ginolhac.github.io/mapDamage/; RRID: SCR_001240
Qualimap	[51]	http://qualimap.bioinfo.cipf.es/; RRID: SCR_001209
GATK UnifiedGenotyper	[52]	https://software.broadinstitute.org/gatk/; RRID: SCR_001876
MultiVCFAnalyzer	[53]	https://github.com/alexherbig/MultiVCFAnalyzer
MEGA6	[54]	http://megasoftware.net/; RRID: SCR_000667
PhyML 3.0	[55]	http://www.atgc-montpellier.fr/phyml/; RRID: SCR_014629
TREE-PUZZLE	[56]	http://www.tree-puzzle.de/
CONSEL	[23]	http://stat.sys.i.kyoto-u.ac.jp/prog/consel/
BEAST	[15]	http://beast.community/; RRID: SCR_010228
Calib 7.1	http://calib.qub.ac.uk/calib/	http://calib.qub.ac.uk/calib/
ModelGenerator	[57]	http://mcinerneylab.com/software/modelgenerator/#
SnpEff	[58]	http://snpeff.sourceforge.net/; RRID: SCR_005191
BEDtools	[59]	http://bedtools.readthedocs.io/en/latest/; RRID: SCR_006646
ggplot2	[60]	http://ggplot2.org/; RRID: SCR_014601
R Project for Statistical Computing	[61]	http://www.r-project.org/; RRID: SCR_001905
Integrative Genomics Viewer (IGV)	[62]	https://www.broadinstitute.org/igv/; RRID: SCR_011793
Other		
Y. pestis Bronze Age aDNA	[7]	ENA: PRJEB10885, https://www.ebi.ac.uk/ena/data/view/PRJEB10885
130 genomes modern comparison dataset	[63]	SRA: SRA010790, https://www.ncbi.nlm.nih.gov/sra/?term=SRA010790
Georgia_1412	[64]	GenBank: CP006783, https://www.ncbi.nlm.nih.gov/nuccore/CP006783
Georgia_1413	[64]	GenBank: CP006762, https://www.ncbi.nlm.nih.gov/nuccore/CP006762
Georgia_1670	[64]	GenBank: AYLR00000000, https://www.ncbi.nlm.nih.gov/nuccore/AYLR00000000
Kyrgyzstan_790	[64]	GenBank: CP006806, https://www.ncbi.nlm.nih.gov/nuccore?term=CP006806
Armenia_1522	[64]	GenBank: CP006758, https://www.ncbi.nlm.nih.gov/nuccore/CP006758
Russia_Federation_2944	[64]	GenBank: CP006792, https://www.ncbi.nlm.nih.gov/nuccore/CP006792

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Georgia_3067	[64]	GenBank: CP006754, https://www.ncbi.nlm.nih.gov/nuccore?term=CP006754
Georgia_8787	[64]	GenBank: CP006748, https://www.ncbi.nlm.nih.gov/nuccore/CP006748
Georgia_3770	[64]	GenBank: CP006751, https://www.ncbi.nlm.nih.gov/nuccore?term=CP006751
Armenia_14735	[64]	GenBank: AYLS00000000, https://www.ncbi.nlm.nih.gov/nuccore?term=AYLS00000000
Azerbaijan_1045	[64]	GenBank: CP006794, https://www.ncbi.nlm.nih.gov/nuccore?term=CP006794
19 draft genomes *Y. pestis* subsp. *microtus* strains	[65]	BioProject: PRJNA269675, https://www.ncbi.nlm.nih.gov//bioproject/PRJNA269675
Black Death *Y. pestis* genome	[66]	SRA: SRA045745, https://www.ncbi.nlm.nih.gov/sra/?term=SRA045745
Justinian *Y. pestis* genome	[67]	ENA: PRJEB14851, https://www.ebi.ac.uk/ena/data/view/PRJEB14851
Bolgar *Y. pestis* genome	[68]	ENA: PRJEB13664, https://www.ebi.ac.uk/ena/data/view/PRJEB13664
Observance *Y. pestis* genomes	[69]	ENA: PRJEB12163, https://www.ebi.ac.uk/ena/data/view/PRJEB12163

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Johannes Krause (krause@shh.mpg.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Ethical approvals

The scope of the study was limited to prehistoric archaeological material. Therefore, no ethics approvals were required.

Description of samples and archaeological sites

Rasshevatskiy, Russia, RK1001, Yamnaya Complex

The Rasshevatskiy site is part of an agglomeration of burial mound cemeteries in linear structure situated not far from river Kuban. Here the river turns from a south-north direction to the West. The cemetery 1 consists of 27 burial mounds, with mound 21 being the largest, and stretching over approximately 2 km. Nearby mound 20 was another large mound, all others were moderate in size. With the exception of mound 24, all features were excavated during construction work in two field seasons 1998 and 2000 by J.B. Berezin and V.L. Rostunov [70, 71].

The big mound 21 had an oval shape with an annex and a size of 85 × 110 m and was 6.2 m in height. The mound was built in five construction phases of which the first is associated with the Maikop epoch, the second was built by Yamnaya groups, and the third and fourth are related to the Novotitorovskaya culture, which is a local variant succeeding Yamnaya in the Northwest Caucasus. The last mound-shell was added by groups of the Catacomb grave complex and North Caucassian cultural complexes. Thus, the mound included 22 graves in total with different cultural affiliations and the dates of the Bronze Age individuals span from before 4560 ± 60 BP to at least 3960 BP, i.e., approximately 500-600 years. Novotitorovskaya grave 7 as well as Catacomb graves 8 and 20 contained wooden wagons or wagon parts.

Grave 11 (RK1001) is one of four Yamnaya burials in the mound and was dug into the second mound shells. The grave is directly built on top of grave 13 dated to 4,447 ± 22 uncalBP (MAMS-29818). The second shell was constructed over grave 3, another probable Yamnaya grave-pit but without a skeleton or inventory. Grave 11 is a typical Yamnaya inhumation in an oval-rectangular pit lying in 5.22 m deep slightly off-center. The individual was placed in a supine position on the back with the head in a western direction, legs crouched and slightly tilted to the north. The skeleton was placed on organic bedding and remains of dark red ochre were found below the feet. Likewise, small pieces of chalk were found below the skeleton. As inventory a small silex instrument was found. The individual is a male, 30-39 years of age. He presents enamel hypoplasia and cribra orbitalia; both signs of childhood stress. There are signs of chronic inflammation of the endocrania, plural prints of small blood vessels, and newly built bone on the inner surface of the frontal and occipital bones and inside the frontal and maxillary sinuses. Furthermore, there are signs of chronic inflammation present on the legs of the individual: slight periostitis and blood vessel imprints visible on the femurs, tibias, and fibulas, and the

metatarsals are also affected by inflammation. Pathologies such as interproximal grooving, chipping, calculus, and abnormal wear of the front teeth in comparison with the posterior teeth are observed in the dentition of the individual. In addition, numerous fractures of the foot bones were recorded. The individual was directly radiocarbon dated 4171 ± 22 uncalBP (MAMS-29816).

The principle stratigraphy of mound 21 is published [70, 71], the comprehensive publication including anthropology and isotope-studies, as well as a re-evaluation of the stratigraphy of the sites is forthcoming.

Beli Manastir - Popova zemlja, Croatia, GEN72, Vučedol Complex

This site is located in eastern Croatia, approximately 2 km West from the town Beli Manastir. Two rescue excavations took place in 2014 and 2015 unearthing approximately 37,000 m^2, where two main layers were identified: a prehistoric layer with several strata from Neolithic and Chalcolithic periods; and a Roman layer where two rectangular brick furnaces were uncovered.

In the prehistoric layer, a total of 28 dwelling pits and 35 inhumations were recovered. Some of the prehistoric burials (21 in total) were found within the dwelling pits either at the bottom or at the top of the backfills of the pits. The rest of the inhumations were located at the bottom of either waste pits or a large canal at the eastern side of the settlement. The Neolithic burials were found in a crouched position on either the right or the left side with different orientation in most of the cases, and some had one or more ceramic vessels placed by the head of the deceased.

The individual GEN72 was found in the grave number 17. The skeleton was in a crouched lying on its belly on its left side. A well-healed ante-mortem depression fracture was located on the posterior part of the left parietal bone. It also exhibited mild, healed ectocranial porosity on the occipital and parietal bones, healed cribra orbitalia on its superior orbits, and mild, healed periostitis on the right tibia and fibula. The individual was recently dated by AMS dating to 4176 ± 28 uncal BP (Labno. BRAMS-1304).

Gyvakarai, Lithuania, Gyvakarai1, Corded Ware/Boat Axe Complex

This site is located in the northern part of Lithuania on the steep gravelly bank (elevation up to 79 m a. s. l.) of the rivulet Žvikė. The burial was accidently discovered in 2000 by locals and subsequently rescue excavations were conducted in the same year in the surrounding area of the highly disturbed grave resulting in discovery of a single grave, Gyvakarai 1, containing a fragmentary skeleton belonging to an adult man, 35-45 years of age, and C14 dated to the Late Neolithic (3,745 ± 70 uncal BP (right tibia, Ki-9470) and 3,710 ± 80 uncal BP (left ulna, Ki-9471) [26]. The dating, the fact that it is a singular grave of an adult individual and the set of grave goods, including a boat-shaped polished stone axe, led to the association of this individual with the Late Neolithic Corded Ware/Boat Axe cultural complex. The individual was recently dated by AMS dating to 4030 ± 30 uncal BP (Labno. Poz-61584).

Kunila, Estonia, Kunila II, Corded Ware Complex

This burial site is situated in Central Estonia, 4 km south-west of Puurmani on the western side of a small drumlin on Jaaniantsu Hill. The burial site was discovered in 1938 during gravel digging, revealing a stone axe and loose human bones. During archaeological excavations in 1948 the skeletal remains of two adult males (Kunila I and II) were uncovered (AI 3723) [72]. Kunila I was buried with a stone adze and a battle-axe; Kunila II with an adze of white flint and a grinding stone. The burials are attributed to Corded Ware Culture [73] and Kunila II has been directly dated to 3960 ± 40 uncal BP (mandible; Poz-10825) [12].

Haunstetten Unterer Talweg 85, Feature 1343 = Grave I/3, Augsburg, Germany, Bell Beaker Complex

The site of "Unterer Talweg 85" (due to a change of the street numbers after the excavation, the site is nowadays also known as "Unterer Talweg 49") is situated in Haunstetten, a quarter of Augsburg to the very south of the city and only 300 m north of Unterer Talweg 58-62, from which samples are also included in this study (see above). The cemetery consists of two small groups of burials, group I with 5 graves and group II with 2 graves, both situated roughly 20 m apart from each other. Group I, the so-called northern group, was excavated in 2001. Three single burials were radiocarbon dated and their 2 sigma ranges fall between 2,465 and 2,152 cal BCE [13].

Palaeogenetic data from the dentine of the individual from grave I/3 (Feat.no. 1343) was included in this study. Grave I/3 (feat. 1343) is dated to 2,397–2,149 cal BCE (2 sigma) and contained a male individual in contracted position with an arrowhead and several pieces of flint as grave goods.

Haunstetten Postillionstrasse 6, Feature 6 = Grave 36, Augsburg, Germany, Early Bronze Age

The site of "Postillionstrasse" is also situated in Haunstetten ca. 3.2 km south of the site "Unterer Talweg 85" and was excavated in 19. The cemetery consists at least of 41 graves, all dating to the Early Bronze Age (total span of radiocarbon dates: 2,198-1,772 BCE). Three graves were covered with a burial mound and surrounded by a ring ditch.

Grave 36 (feat. 6) is dated to 2,009–1,883 cal BCE (2 sigma) [13]. It contained a male individual in crouched position. A copper dagger and a bangle, two flint arrowheads were part of his grave goods. The burial was placed in the southern part of the cemetery surrounded by other graves.

METHOD DETAILS

Sampling and extraction

Sampling of a total of 563 tooth and bone samples (Russia (n = 122), Hungary and Croatia (n = 139), Lithuania (n = 27), Estonia (n = 45), Latvia (n = 10), and Germany (Althausen n = 4, Augsburg n = 83, Mittelelbe-Saale n = 133)) took place in the clean room facilities of the Institute for Archaeological Sciences at the University of Tübingen, the Institute of Archaeology RCH HAS in Budapest and of the MPI-SHH in Jena. After irradiation with UV light to remove surface contamination, teeth were sawed apart transversally at the border of crown and root, and dentine from inside the crown was sampled and powdered using a sterile dentistry drill. For the samples processed in Budapest, whole teeth were powdered. For bone samples, the surface layer from the sampling area was removed with a

dentistry drill prior to obtaining bone powder from the inside of the bone by drilling. For each specimen we gathered between ~30 and 120 mg of powder to be used for DNA extraction.

Extraction was performed following a protocol optimized for the recovery of small ancient DNA molecules [74] in the following manner: around 30-50 mg of powder (bone or teeth) for each individual was combined with the extraction buffer (EDTA,1.04ml; Proteinase K, 0.03ml; and UV-water 0.9ml), and rotated over night at 37°C, then it was centrifuged at 14000 rpm to pellet the bone powder and the supernatant was transferred to a 50ml Falcon tube (containing funnel and purification column) already containing 10 mL of binding buffer and 400 μg sodium acetate. The Falcon tube was then spun at 1500 rpm for 8 min with slow acceleration and fast deceleration. After which, the column was transferred into a new collection tube and liquid remaining in the funnel was transferred to the column. This was followed by 2 washing steps consisting of: adding 450 μL of wash buffer (40ml Ethanol plus 22ml wash buffer from the High Pure Viral Nucleic Acid Large Volume Kit) to the column and spun it at 8000 rpm for 1 min; plus two dry spins at 14000 rpm for 1 min. Then the DNA was eluted in a new tube by doing two elution steps by adding into the purification column 50 μL of TET and spinning for 1 min at 14000 rpm. The extraction resulted in 100 μL of DNA extract per sample. An aliquot of 20 μL of extract was used to generate double-indexed libraries as described elsewhere [45, 75]. In brief, samples were blunt-end repaired by adding the extract to 60 μL of Blunt-End Repair mastermix (NEB Buffer 2, 10.50 μl; ATP, 10.50 μl; BSA, 4.20 μl; dNTPs, 0.42 μl; T4 PNK, 4.20 μl; T4 Polymerase, 0.84 μl; and UV water, 53.34 μl); the reactions were incubated in a thermocycler at 15°C for 15 min, followed by 15 min at 25°C and thereafter purified with a MinElute kit. The blunt-end repaired extract was then ligated to Illumina adapters by adding 21 μL of Quick Ligase Buffer, 1.05 μL of adaptor mix (0.25 μM concentration) and 18 μL of the purified blunt-end repaired extract, and finally adding 1 μL of Quick ligase (0.125 U concentration) and incubated for 20 min at 20°C, followed by purification with a MinElute kit. Finally, 20 μL of each sample was adaptor filled by adding 20 μL mastermix (Themopol Buffer, 4 μl; dNTPs 0.20 μl; Bst, 2 μl; and UV-Water 13.80 μl) and incubated at 37°C for 30 min followed by 10 min at 80°C. Then, each sample was quantified by qPCR. Each sample was then indexed with dual sample-specific indexes. For each indexing reaction, each library was split to ensure that there were less than 1.5e+10 copies of DNA in each indexing reaction. For example, if the library had to be split into two, we added 18 μL of the prepared libraries and combine it with the indexing mastermix (Pfu Turbo Buffer, 10 μl; BSA, 1.5 μl; dNTPs, 1 μl, Pfu Turbo Polymerase, 1 μl; P5 and P7 sample-specific indexes, 2 μL for each; and UV-water 64.5 μl) and if it had to be split into four reactions, we would add 9 μL of the library into the indexing mastermix, only changing the amount of UV-water to 73.5 μl. Once the reactions were set up, these were transferred to a modern DNA lab for indexing PCR: initial denaturation at 95°C for 2 min and 10 cycles of: 30 s at 95°C, 30 s at 58°C, and 1 min at 72°C; and a elongation phase of 10 min at 72°C. The amplified double-indexed libraries were then purified with a MinElute kit. Negative controls were included in the extraction and library preparation and taken along for all further processing steps. Negative controls were included in the extraction and library preparation and taken along for all further processing steps.

Shotgun screening sequencing

Libraries were PCR-amplified and quantified using an Agilent 2100 Bioanalyzer DNA 1000 chip and pooled at equimolar concentrations prior to paired-end sequencing on a NextSeq500 with 2x101+8+8 and a HiSeq2500 with 2x101+8+8 cycles according to the manufacturer's instructions to a depth of ~1.5 million reads per library.

In silico screening

The sequencing data for the 170 samples was preprocessed with ClipAndMerge [46] to remove adaptors, base quality-trim (20) and merging and filtering for only merged reads. Reads were mapped using the BWA aln algorithm [47] to a multi-species reference panel, containing various representatives of the genus *Yersinia* (Table 1) and the plasmids of *Yersinia pestis*: pCD1, pMT1 and pPCP1 from *Y. pestis* CO92. The region comprising 3000-4200bp of the *Y. pestis* specific plasmid pPCP1 was masked in the reference, since it is highly similar to an expression vector used during the production of enzyme reagents [76].

Mapped files were then filtered for reads with a mapping quality higher than 20 with Samtools [48]. PCR duplicates were removed using the MarkDuplicates tool in Picard (1.140, http://broadinstitute.github.io/picard/). The number of reads mapping specifically to each genome and to the plasmids were retrieved from the bam files using Samtools [48] idxstats. An endogenous based score was used to assess the potential of the sample being 'positive' for *Y. pestis*. It was calculated as follows:

$$\frac{(YPS - \max(YS))}{M} \times 1,000$$

where YPS is the number of reads specifically mapping to *Y. pestis*; YS is the maximum number of reads mapping specifically to a *Yersinia* species with the exception of *Y. pestis* and M is the total number of merged reads in the sample. By using the maximum number of reads mapping to another species of the genus *Yersinia*, the score takes in account different source of contamination other than *Y. pseudotuberculosis*. Five samples (RK1001, Gyvakarai1, KunilaII, 6Post and 1343UnTal85) fulfilled the criteria for being considered strong candidates (score higher than 0.005 (Figure S4C) and reads mapping to all plasmids). Another samples, GEN72, was also included in further processing and analysis since it had higher numbers mapping the *Y. pestis* chromosome and plasmids even though it did not full-fill the score requirements. For a detailed description of the archaeological sites and individuals see the next section.

Deep shotgun sequencing

The five strong candidate samples detected in screening of the shotgun data were processed for deep shotgun sequencing as following: For Gyvakarai1 the screening library described above was pair-end sequenced on two lanes of a HiSeq4000 for 100 cycles, and on a full run of a NextSeq500 for 75 cycles. The screening library for KunilaII was pair-end sequenced deeper on 80% of one lane of a HiSeq4000 for 100 cycles. Additionally, 40 μL of DNA extract of KunilaII was converted in to a library treated with UDG and endonuclease VIII to remove deaminated bases [77], and pair-end sequenced on one lane of a HiSeq4000 for 75 cycles.

For RK1001, Post6 and 1343UnTal85, 60 μL of DNA extract each were converted into DNA libraries using so-called UDG-half treatment, whereby deaminated bases are partially removed and retained mostly at the ends of the molecule [78]. The library of RK1001 was deep shotgun pair-end sequenced in 8 lanes of a HiSeq4000 for 55 cycles. The libraries of 6Post and 1343UnTal85 were deep shotgun single-end sequenced on 2 and a half lanes of a HiSeq4000 for 75 cycles. Post6 was additionally pair-end sequenced on a full run of a NextSeq500 for 75 cycles.

Y. pestis in-solution capture

Y. pestis whole-genome DNA capture probes were designed using as template sequences the *Y. pestis* CO92 chromosome (NC_003143.1), *Y. pestis* CO92 plasmid pMT1 (NC_003134.1), *Y. pestis* CO92 plasmid pCD1 (NC_003131.1), *Y. pestis* KIM 10 chromosome (NC_004088.1), *Y. pestis* Pestoides F chromosome (NC_009381.1) and *Y. pseudotuberculosis* IP 32953 chromosome (NC_006155.1). We used a 6 bp tiling with a probe length of 52 bp with an additional 8 bp 3′ linker sequence as described in [79]. Low complexity regions were masked using dustmasker [49] (version 2.2.32+). Redundant probes as well as probes with more than 20% masked nucleotides were discarded. The procedure resulted in 816,413 unique probe sequences. A second probe set was created with a coordinate offset of 3 bp resulting in 827,438 unique probe sequences. In combination the two probe sets represent an effective tiling density of 3 bp. The two probe sets were ordered on two 1 million feature Agilent SureSelect DNA Capture Arrays. The full capacity of the arrays was filled up with randomly selected probes. The two arrays were turned into in-solution DNA capture libraries as described elsewhere [79].

For GEN72, 25 μL of DNA extract was converted into DNA libraries using so-called UDG-half treatment as described above [78]. The UDG-half libraries of RK1001 and GEN72 were enriched for *Y. pestis* DNA using in-solution DNA capture probes (see above) as described elsewhere [25, 79, 80]. The capture products of RK1001 and GEN72, were sequenced on 1 and 0.6 of the lane, respectevely, of the HiSeq4000 for 75 cycles.

Genome reconstruction and authentication

All samples were processed with the EAGER pipeline [46]. Sequencing quality for each sample was evaluated with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and adaptors clipped using the ClipAndMerge module in EAGER. For paired-end data, the reads were also merged with ClipAndMerge and only the merged reads were kept for further analysis. Furthermore, 7 bp in the 5 prime end were clipped in GEN72 due to the presence of an extra barcode in the DNA library.

Individual sample treatment due to laboratory preparation and sequencing strategies

Gyvakarai1: two HiSeq lanes and one Next-Seq run paired-end of the non-UDG treated library were combined and reads mapped to *Y. pestis* CO92 reference with BWA aln (-l 16, -n 0.01, hereby referred to as non-UDG parameters). Reads with mapping quality scores lower than 37 were filtered out. PCR duplicates were removed with MarkDuplicates. MapDamage (v2.0) [50] was used to calculate damage plots (Figure S1). Coverage was calculated with Qualimap (v2.2) [51].

KunilaII: UDG and the non-UDG libraries were sequenced in 2 HiSeq pair-end lanes and processed separately until calculation of the coverage. The non-UDG treated libraries were mapped with non-UDG parameters while the UDG treated library reads were mapped with more stringent parameters (-l 32, -n 0.1, referred to as UDG parameters). Reads with mapping qualities less than 37 were filtered out and duplicates were removed with MarkDuplicates as before. The non-UDG bam file was used to calculate damage plots as indicated above (Figure S1). After duplicate removal, the UDG- and non-UDG treated BAM files were merged together and used to calculate the coverage as above.

GEN72, Post6 and 1343UnTal85: the UDG-half treated libraries were sequenced in two HiSeq lanes for Post6 and 1343UnTal85 and 19,777,683 reads were generated in the HiSeq for GEN72, and two different runs were performed. For the first run, reads without clipping were used to retain miscoding lesions indicative of aDNA. BWA aln was used for mapping with non-UDG parameters (-l 16 and -n 0.01). Reads with mapping qualities lower than 37 were filtered and PCR duplicates were removed with MarkDuplicates as described above. Coverage and damage plots were calculated as above (Figure S1). After clipping the last two bases with the module ClipAndMerge in eager, potentially affected by damage, the samples were mapped with UDG parameters.

RK1001: UDG-half library was shotgun sequenced pair-end in 8 HiSeq lanes and in-solution captured and sequenced single end to a depth of 303,148,884 reads sequenced in the HiSeq. Shotgun and captured data were combined in a fastq file and processed as described above for GEN72, Post6 and 1343UnTal85.

SNP calling, heterozygosity, and phylogenetic analysis

Prior to SNP calling in order to avoid false SNP calling due to aDNA damage, the quality scores of damaged sites in the non-UDG treated samples were downscaled using MapDamage [50] (v2.0), as performed in previous analysis [7]. For the UDG-half data, the files with the two last bases clipped, hence removing the damage signal, and mapped with UDG parameters were used for SNP calling (see above).

SNP calling was performed with GATK UnifiedGenotyper [52] in EAGER[43] with default parameters and the 'EMIT_ALL_SITES' output mode.

VCF files were generated for the two complete genomes from Rasmussen et al. (2015) [7], the Black Death [66], Justinianic Plague [67], Bolgar [68] and Observance [69] genomes and a curated dataset of 130 modern genomes [63] in addition to 11 samples from the Former Soviet Union [64] and 19 draft genomes of *Y. pestis* subsp. *microtus* strains [65] in the following manner: All the modern genomes were cut into 100bp reads and map to the reference *Y. pestis* CO92 with UDG parameters as described above, including the reference genome. The raw reads from the previous Bronze Age *Y. pestis* samples [7], were mapped to the reference with non-UDG parameters and downscaled using MapDamage [50] (v2.0), as above. The rest of the ancient genomes, were mapped to the reference with UDG parameters. The SNP calling for all the modern and ancient genomes was performed as described above.

All the VCF files from the modern and ancient dataset and the samples produced in this study were combined and processed with *MultiVCFAnalyzer* (v0.85, https://github.com/alexherbig/MultiVCFAnalyzer) [53] that produced a SNP table and a fasta alignment file containing the concatenation of all variable positions in the dataset (SNPs alignment), in respect to the reference *Y. pestis* CO92. In order to call a SNP a minimum genotyping quality (GATK) of 30 was required, with a minimum coverage of 3X, and with a minimal allele frequency of 90% for a homozygous call. No heterozygous calls were included in the output files.

The SNP alignment was curated by removing all alignment columns with missing data (complete deletion). The curated SNP alignment was then used to compute Neighbor Joining (NJ) and Maximum Parsimony (MP) trees with MEGA6 [54] and a Maximum Likelihood (ML) tree using PhyML 3.0 [55] with the GTR model used in previous *Y. pestis* work [7, 63], with 4 gamma categories and the best of NNI and SPR as tree branch optimization. The specific variants of *Y. pseudotuberculosis* were removed from the analysis to improve the visual resolution of the tree.

To check for infections with multiple strains, all the VCF files from modern and ancient dataset including the samples produced in this study were combined and processed with *MultiVCFAnalyzer* (v0.85, https://github.com/alexherbig/MultiVCFAnalyzer) [53] as described above, with the exception that heterozygous calls were included in the output files. The only parameters changed were minimal allele frequency of 90% for a homozygous and of 10% for the heterozygous calls (Figure S1).

Dating analysis

The SNP alignment after complete deletion was used for molecular dating using BEAST 1.8.2 [15]. The modern sample 0.PE3, also called Angola, was removed from the dataset due to its long branch.

For tip dating, all modern genomes were set to an age of 0. The dates of the ancient samples presented in this study plus the two complete genomes from Rasmussen et al. (2015) [7] were recalibrated with Calib 7.1 (http://calib.qub.ac.uk/calib/) to the IntCal13 calibration curve. The ancient samples were given the median calibrated probability as their age, and the 2 sigma interval was used as the boundaries for a uniform prior sampling (Table 2). The dates published for previous historical genomes were transformed to cal BP assuming 1950 as age 0 and given the mean as the age with the interval as the boundaries of a prior uniform distribution: RISE509 4729 (4625-4836) [15], RISE505 3635 (3575-3694) [15], Black Death 603 (602-604) [66]; Observance 229 (228-230) [69], Bolgar 569 (550-588) [68] and Justinian 1453 (1382-1524) [67].

The molecular clock was tested and rejected using MEGA6 (See Molecular Clock Test in QUANTIFICATION AND STATISTICAL ANALYSIS section below). Therefore, we followed previous work and used an uncorrelated relaxed clock with lognormal distribution [7, 63] with the substitution model GTR+G4, selected using ModelGenerator [57]. Tree model was set up to coalesce assuming a constant population size and a rooted ML tree was provided as a starting tree. Two independent 1,000,000,000 MCMC chains were computed sampling every 5,000 steps. The two chains were then combined using LogCombiner from BEAST 1.8.2 [15] with a 10 percent burn-in (100,000,000 steps per chain). The Effective Sample Size (ESS) of the posterior, prior, treeModel.rootHeight, tMRCA_allpestis are 4,589, 4,087, 1,054 and 7,571 respectively. The trees files for the 2 chains were combined with LogCombiner with 100,000,000 of burning and resampled every 20,000 steps giving a total number of 90,000 trees, that were used to produce a Maximum Clade Credibility tree using TreeAnnotator from BEAST 1.8.2 [15].

SNP effect analysis and virulence factors analysis

The SNP table from *MultiVCFAnalyzer* was provided to *SnpEff* [58] and the effect of the SNPs within genes present in the dataset was evaluated. Additionally the SNP table was manually assessed for possible homoplasies.

For the virulence factors, the samples were mapped as indicated above but without applying quality filtering and the percentage of coverage was calculated for each region using BEDtools [59] and plotted using the package ggplot2 [60] in R [61]. Additionally, *ureD* was manually explored for SNPs using Integrative Genomics Viewer (IGV) [62].

Indel analysis

The samples including the two complete Bronze Age genomes [7] were mapped against *Y. pseudotuberculosis* IP 32953 with bwa with non-UDG parameters (-n 0.01, -l 16), except for RK1001, GEN72, 1343UnTal85 and 6Post that were mapped with bwa with UDG

parameters (-n 0.1, -l 32),. The modern genomes from branch 0 (0.PE7, 0.PE2, 0.PE3 and 0.PE4), *Y. pestis* CO92 and *Y. pestis* KIM10 were *in-silico* cut in 100 bp fragments with 1bp tiling and mapped to *Y. pseudotuberculosis* reference using bwa with UDG parameters (-n 0.1, -l 32). The non-covered regions were extracted using the BEDtools genomecov function [59]. Missing regions larger than 1kb were comparatively explored in order to identify indels. Using the BEDtools intersect function [59], we extracted regions missing in the Neolithic genomes and present in the modern ones and also the regions missing in the modern ones but still present in the Neolithic genomes. The results were check by manual inspection in IGV [62].

QUANTIFICATION AND STATISTICAL ANALYSIS

In silico screening

A total of n = 563 samples were screened for the presence of *Y. pestis* and we determined candidates for further analysis by calculating a score as indicated above (see *In silico* screening in METHODS DETAILS).

Phylogenetic analysis

Phylogenetic analyses were performed with the NJ, MP and ML algorithms in MEGA6 [54] and PhyML 3.0 [55] (see SNP calling, heterozygocity, and phylogenetic analysis in METHODS DETAILS section above). To test the statistical support of the phylogenetic trees we bootstrapped each tree with 1,000 bootstrap replicates.

Tree topology test

To statistically test the obtained topology (see SNP calling, heterozygocity, and phylogenetic analysis in METHODS DETAILS section above), we performed statistical tests with TREE-PUZZLE(v5.3) [56] and CONSEL(v1.2) [23]. Four different topologies, described in Figure S4B, were provided together with the SNP alignment file to TREE-PUZZLE which evaluated these user defined trees with the following parameters: Neighbor joining + quartet sampling for parameter estimation uses, GTR model with 4 gamma categories. The output .sitelh file was used to perform the statistical test with CONSEL with 1,000 bootstraps.

Molecular clock test

In order to test the molecular clock hypothesis, we performed a molecular clock test (ML) in MEGA6, by providing the alignment used for phylogenetic analysis and the topology obtained with the Neighbout Joining algorithm (see Dating analysis in METHODS DETAILS above) with default parameters.

Dating analysis

Dating analysis was performed with BEAST 1.8.2 [15] as indicated above (see Dating analysis in METHODS DETAILS). We run two independent MCMC chains with 1,000,000,000 steps each.

DATA AND SOFTWARE AVAILABILITY

Raw sequencing data have been deposited at the European Nucleotide Archive under accession number ENA: PRJEB19335 (https://www.ebi.ac.uk/ena/data/view/PRJEB19335).

Current Biology, Volume *27*

Supplemental Information

The Stone Age Plague and Its Persistence in Eurasia

Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, Michal Feldman, Rimantas Jankauskas, Ivor̆ Janković, Ken Massy, Mario Novak, Saskia Pfrengle, Sabine Reinhold, Mario Šlaus, Maria A. Spyrou, Anna Szécsényi-Nagy, Mari Tõrv, Svend Hansen, Kirsten I. Bos, Philipp W. Stockhammer, Alexander Herbig, and Johannes Krause

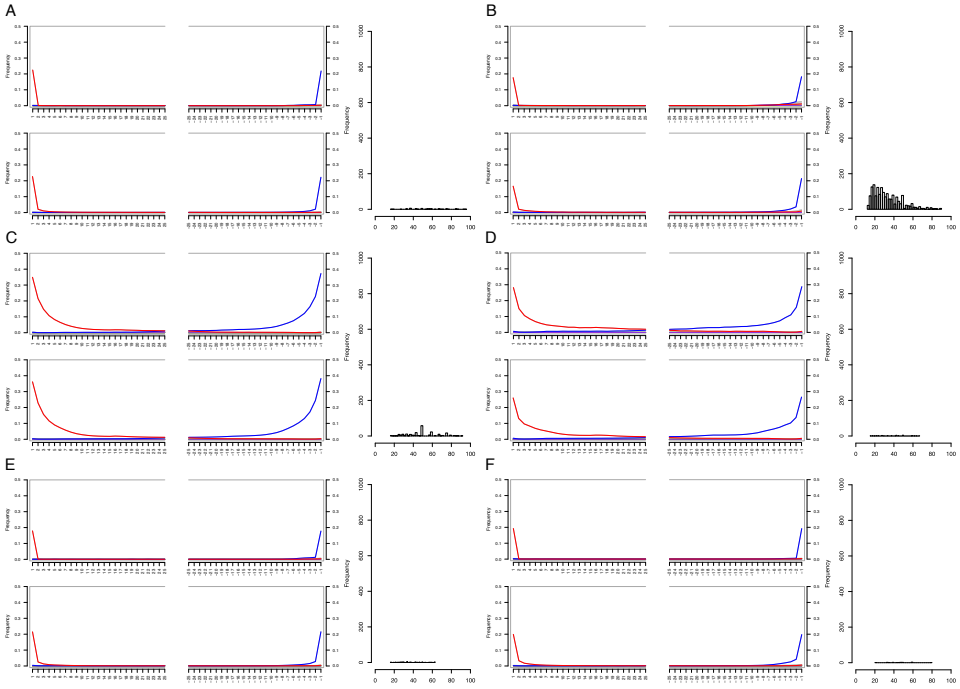


Figure S1: Ancient DNA damage plots for *Y. pestis* chromosome (top) and human HG19 (bottom) generated by MapDamage2.0[S1] and heterozygosity plots (right, see STAR Methods) from RK1001 (A), GEN72 (B), Gyvakarai1 (C), KunilaII (D), 1343UnTal85 (E) and Post6 (F). The damage plots of GEN72, RK1001, Post6 and 1343UnTal85 only retain damage in the last two bases as these libraries were prepared using an 'UDG-half' protocol[S2] (see STAR Methods). For the SNP heterozygosity plots, the y axis is percentage of allele frequency, the reference calls (0%) and alternative calls (100%) were excluded for representative purposes. Heterozygosity plots were generated in R[S3] with ggplot2[S4]

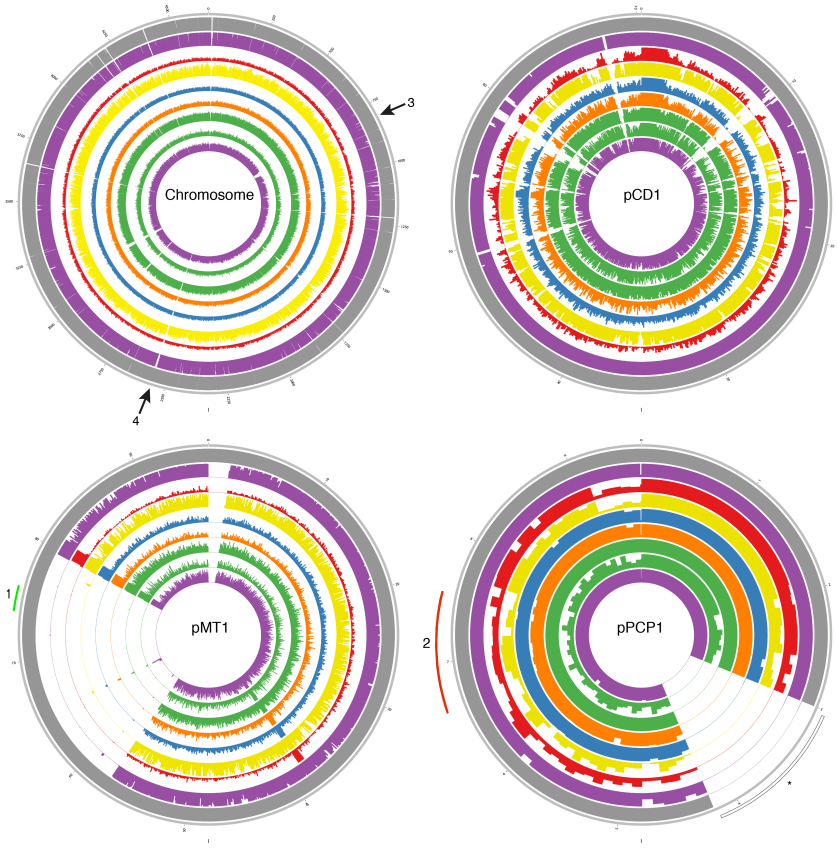


Figure S2: Average coverage plot for the chromosome and plasmids of *Yersinia pestis*, from the outer ring to the inner ring: *Y. pestis* CO92 (NC_003143.1, reference), RISE509, RK1001, GEN72, Gyvakarai1, KunilaII, 6Post, 1343UnTal85 and RISE505. Related to Figue 1B. Colours correspond to the regions where the genomes were recovered from: Altai region (purple), Russia (red), Croatia (dark yellow), Gyvakarai, Lithuania (blue), Kunila, Estonia (orange), Augsburg, Germany (green). The average depth of coverage was calculated for 1kb regions for the chromosome and 100bp for the plasmids, each ring represents a maximum of 20X coverage. The figure was generated with Circos[78]. (1) *ymt* gene, (2) *pla*, (3) deletion of

flagelin genes, (4) filamentous prophage YpfΦ, (*) region mask in pPCP1 due to high similarity to expression vectors during enzyme production[S5]. Image generated with circos[S6].


A
B
C

Figure S3: A) Neighbor Joining tree B) Maximum Parsimony tree C) Maximum Likelihood complete tree. Related to Figure 1A. Only bootstrap values higher than 95% are shown in the figure. Trees were rooted using *Y. pseudotuberculosis* IP 32953 (*Y.pseudotuberculosis*). *Y. pseudotuberculosis*-specific SNPs were excluded from the tree for representative matters, which lead to the observed branch shortening.

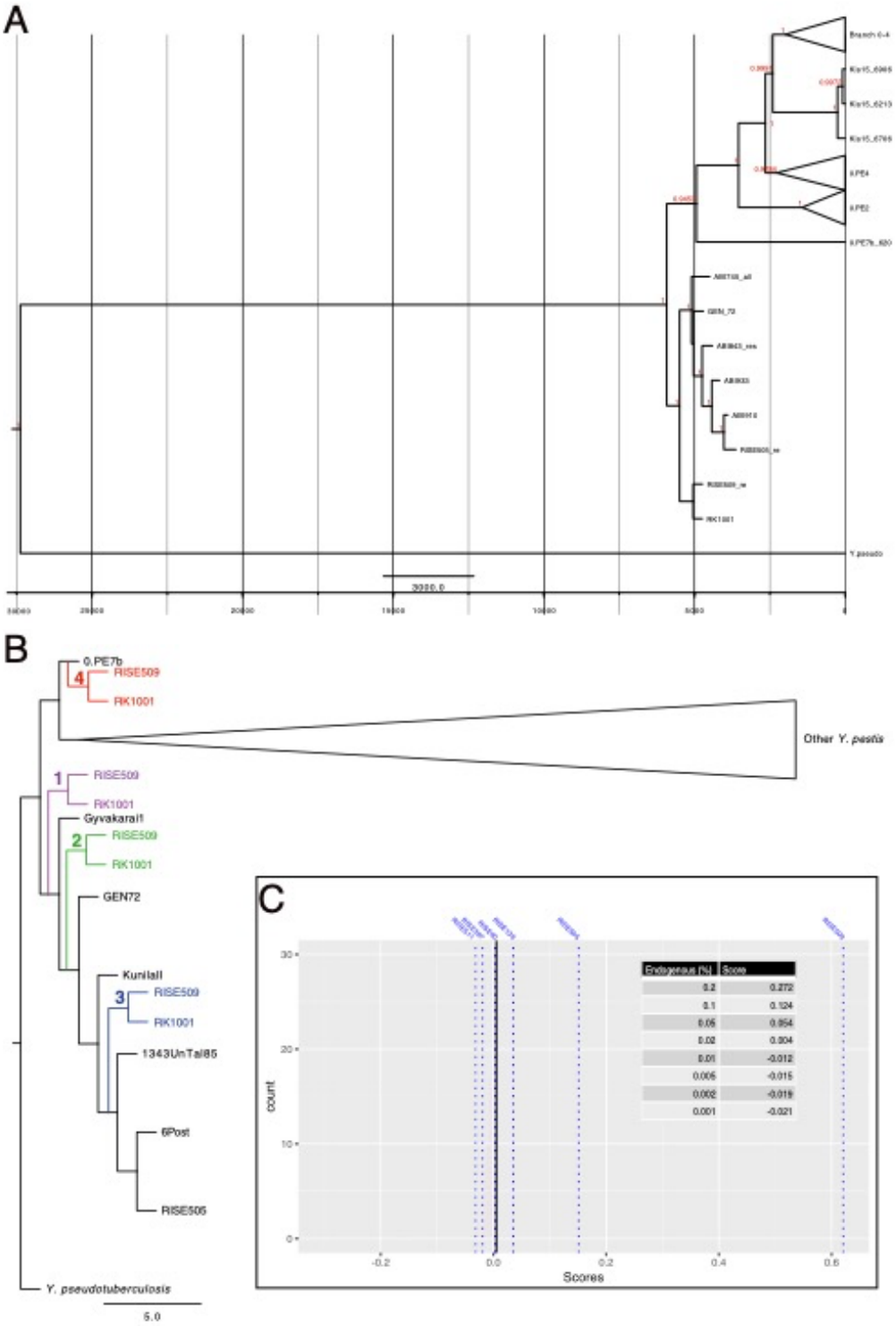

A
Branch 0-4
Kim95_6065
Kim95_6213
Kim95_6765
0.PE4
0.PE2
0.PE7b_620
Y.pseudo
2000.0
B
0.PE7b
4
RISE509
RK1001
Other Y. pestis
1
RISE509
RK1001
Gyvakarai1
2
RISE509
RK1001
GEN72
KunIa1l
3
RISE509
RK1001
1343UnTal85
6Post
RISE505
Y. pseudotuberculosis
5.0
C
count
Scores
Endogenous (%)
Score
0.2
0.272
0.1
0.124
0.05
0.054
0.02
0.004
0.01
-0.012
0.005
-0.015
0.002
-0.019
0.001
-0.021

Figure S4: (A) Maximum Clade Credibility tree generated using TreeAnnotator from the BEAST v1.8.2 package. Posterior probabilities higher than 0.90 are shown in red. The outgroup is *Y. pseudotuberculosis* (Y. pseudo). Related to STAR Methods (B) Statistical testing of the tree topology. Related to STAR Methods. Each colour represents one tree model. AU value represents the p-values for the different topologies. A tree topology is rejected with p-value<0.05. The AU values for each topology are: topology 1=1.000; topology 2=2e-04; topology 3=1e-08; topology 4=2e-11. Only topology 1 is not rejected, and corresponds to the one inferred by NJ, MP and ML phylogenetic algorithms. **(C) Distribution of scores after screening (red). Related to STAR Methods.** The black line represents the threshold for strong positive samples at 0.005. Blue dashed lines represent the scores for the previous generated data for *Y. pestis* by Rasmussen et al.[S7]: the samples were down-sampled to 6,000,000 for comparability. The table represents the scores of datasets (5,000,000 reads) generated with gargammel[S8] with decreasing percentages of *Y. pestis* endogenous DNA in the Clovis bacterial background[S9] provided with the program.

Sample	Clipped, merged and quality-filtered reads before mapping	Plasmid	Unique reads mapping to *Y. pestis* reference	Endogenous DNA (%)	Mean Coverage	Coverage (%)				
						>=1X	>=2X	>=3X	>=4X	>=5X
RK1001	1,833,084,416	pCD1	12,405	0.006	7.42	89.27	82.28	74.43	66.47	58.80
		pMT1	7,262	0.003	3.10	66.65	54.63	42.41	30.41	20.68
		pPCP1	4,236	0.002	19.28	87	85.99	83.73	81.30	79.21
GEN72	19,777,683	pCD1	36,172	0.918	23.94	92.63	89.72	87.99	86.42	84.73
		pMT1	29,919	0.649	13.82	74.35	72.84	71.65	70.43	69.01
		pPCP1	4,818	0.243	24.42	83.15	79.80	77.24	73.24	70.15
Gyvakarai1	1,021,452,137	pCD1	15,805	0.002	11.83	95.87	94.92	94.28	92.92	90.85
		pMT1	11,198	0.002	6.07	74.73	73.23	70.15	64.87	56.88
		pPCP1	5,083	0.001	28.50	87.60	87.54	87.40	87.33	87.10
KunilaII	379,155,741	pCD1	20,431	0.008	14.03	94.48	92.72	91.1	88.79	85.61
		pMT1	12,696	0.006	6.29	73.48	69.72	63.78	56.31	47.97
		pPCP1	7,033	0.003	36.31	87.48	87.38	87.32	87.15	86.99
1343UnTal85	1,174,989,269	pCD1	33,103	0.005	20.65	94.84	94.29	93.90	93.65	93.36
		pMT1	22,631	0.004	10.04	75.02	74.80	74.28	73.38	71.87
		pPCP1	9,066	0.002	43,60	87.56	87.52	87.47	87.44	87.40
Post6	419,717,299	pCD1	28,223	0.01	17.29	94.05	92.96	91.88	90.56	88.73
		pMT1	14,313	0.006	6.22	73.63	69.51	63.97	57.30	50.56
		pPCP1	4,023	0.002	18.45	87.08	86.04	84.01	81.73	78.06

Table S1: Plasmids reconstruction statistics of the LNBA *Y. pestis*. Related to Table 2.

Supplemental References

S1. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

S2. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., and Reich, D. (2015). Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *370*, 20130624.

S3. R Development Core Team (2008). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing) Available at: http://www.R-project.org.

S4. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York) Available at: http://ggplot2.org.

S5. Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest, S., Coombes, B.K., Wood, J.W., Earn, D.J.D., *et al.* (2011). Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. PNAS *108*, E746–E752.

S6. Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. Available at: http://genome.cshlp.org/content/early/2009/06/15/gr.092759.109 [Accessed October 18, 2016].

S7. 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., *et al.* (2015). Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. Cell *163*, 571–582.

S8. Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). gargammel: a sequence simulator for ancient DNA. Bioinformatics *33*, 577–579.

S9. Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., *et al.* (2014). The genome of a late Pleistocene human from a Clovis burial site in western Montana. Nature *506*, 225–229.

Manuscript B

Stone Age *Yersinia pestis* genomes shed light into the early evolution, diversity and transmission ecology of plague

Aida Andrades Valtueña*, Gunnar U. Neumann*, Maria A. Spyrou*, Lyazzat Musralina*, Franziska Aron, Beisenov Arman, Alexandra Buzhilova, Matthias Conrad, Leyla B. Djansugurova, Miroslav Dobes, Michal Ernée, Javier Fernández-Eraso, Bruno Frohlich, Mirosław Furmanek, Agata Hałuszko, Svend Hansen, Éadaoin Harney, Felix M. Key, Elmira Khussainova, Yegor Kitov, Corina Knipper, Carles Lalueza Fox, Judith Liddleton, Ken Massy, Alissa Mittnik, José Antonio Mujika-Alustiza, Iñigo Olalde, Luka Papac, Sandra Penske, Ron Pinhasi, David Reich, Sabine Reinhold, Harald Stäuble, Christina Warinner, Philipp Stockhammer, Wolfgang Haak, Alexander Herbig, Johannes Krause
*Contributed equally to the work

Abstract

Plague, caused by the bacterium *Yersinia pestis*, has had important implications in past societies and it is known for its devastating effect on medieval Europe during the Black Death and subsequent outbreaks. Although often thought of as a historical disease, the bacterium is known to have affected human populations as early as 5000 years ago. Ancient DNA has contributed to the understanding of *Y. pestis* early evolution through the analysis of genomes from the Eurasian Late Neolithic/Early Bronze Age (LNBA). These data have helped in the understanding of key steps that contributed to the emergence of this pathogen from its close relative *Yersinia pseudotuberculosis*, but questions still remain on its transmission and its early diversification. Here we present 15 new ancient *Y. pestis* genomes dating from 5000-2000 years Before Present (BP) to tackle these questions. We show that during the LNBA period there are two forms of *Y. pestis* differing in their transmission and ecology.

Main

Plague is a zoonotic rodent disease caused by the Gram-negative bacterium *Yersinia pestis*. Plague has been shown to be the causative agent of three historically-recorded pandemics: the first pandemic or plague of Justinian (6-8th century), the second pandemic which started in Europe with the Black Death (14-18th century) and the third pandemic (19th century), during

which the disease spread world-wide leaving active remnants around the world[1]. Ancient DNA studies have recovered *Y. pestis* genomes from victims of the first and second pandemics[2–9], confirming the bacterium as a biological agent of the disease. Furthermore, recent studies have shown that *Y. pestis* has been affecting humans since the Late Neolithic[10–12], long before the first pandemic. The *Y. pestis* genomes from the Late Neolithic and Bronze Age (LNBA) have provided insights into the early evolution of this pathogen, and it has been hypothesised that the increased human mobility during this period has played a role in its early spread across Eurasia[10,12]. However, there are still open questions regarding the disease manifestation and ecology during that period. The LNBA plague presents a genetic background incompatible with the efficient transmission by the flea vector[10–12]. Current knowledge on *Y. pestis* ecology suggests that the pathogen relies, to a great extent, on the flea vector for its enzootic transmission across rodent populations, which represent its main reservoirs[13]. The earliest ancient *Y. pestis* genome containing all the adaptations required for the efficient transmission was found in an individual in the Samara region (Russia) dating to 3,800 years Before Present[14], suggesting that the Bronze Age was possibly a crucial period for the development of the epidemic pathogen we know today. However, this does not explain how the earlier forms were transmitted. Notably, flea-mediated transmission is not the only documented of plague spread: pneumonic plague is a disease form that can result from direct human-to-human contact with only a few reported outbreaks[15–19], and plague cases are known to have been caused by the handling or ingestion of infected animals[20–22]. Expanding the number of *Y. pestis* genomes from the LNBA period can offer a high-resolution snapshot of important stages of the evolution of the bacterium. Furthermore, by linking the genomic evidence with the available archaeological context, we can gain insights into the development of the transmission mechanisms of plague during its early evolution.

Results

Screening and genome reconstruction

We screened a total of 227 samples from 13 archaeological sites that span from Western Europe to central Asia, dating from the Late Neolithic until the Iron Age (~5,000-2,000 years ago, Figure 1A, Supplementary Table 1) using the HOPS pipeline[23] for the presence of *Yersinia pestis* DNA. Based on the output of HOPS, we selected potential candidates for capture enrichment for *Y. pestis* DNA. The criteria for selection were: reads aligning to *Y. pestis* or the *Y. pseudotuberculosis* complex, a decreasing edit distance distribution, presence of aDNA damage pattern, together with visual inspection of the alignments in MEGAN[24] to check for even distribution of the reads across the reference. After capture, a total of 15 ancient *Y. pestis* genomes were reconstructed with coverages ranging from 2.5-30.6, 6.4-66.3, 3.2-38.2- and 11.7-155-fold coverages for the chromosome, pCD1, pMT1 and pPCP1 plasmids, respectively (Table 1 for the chromosome, and Supplementary File 1 Table 1 for the plasmids). Within our selection, we included a sample that was originally published as RISE139 (CHC004 in this study) in [11] where 487,240,605 reads were sequenced resulting in 0.14-, 0.28-, 0.24- and 0.76-fold coverages for the chromosome, pCD1, pMT1 and pPCP1 plasmids, respectively. After capture enrichment, we sequenced 13,860,197 reads yielding a 4.4-, 11.4-, 5.3- and 29.6-fold coverages for the

chromosome, pCD1, pMT1 and pPCP1, respectively. This highlights the economic/resourceful use of applying capturing techniques to recover *Y. pestis* genomes even when low levels of the pathogen DNA are present in shotgun sequencing data.

Phylogenetic analysis

In order to understand the phylogenetic relationship of the newly recovered genomes in comparison to other *Y. pestis* strains, we computed a Maximum Likelihood (ML) phylogeny including modern representatives as well as previous ancient genomes from the Neolithic to the Bronze Age period[11,12,14,25], first pandemic[4,9] and second pandemic[2,3,6,7].

Fourteen of the fifteen newly reconstructed genomes fall into the previously reported LNBA lineage (Figure 1B). Genomes of this lineage have been reported from Russia, Germany, Poland, Croatia, Estonia and Lithuania[11,12,25]. We now report the presence of this pathogen also in the Czech Republic, Ukraine, Kazakhstan, and Mongolia, thus widening the geographical area where *Y. pestis* was found in the past. As previously shown [11,12,25], the genomes within this lineage branch in clocklike fashion in order of their mean calibrated radiocarbon date, with the exception of I5884 (Figure 1C, Material and Methods). Given that I5884 falls derived in the phylogeny from Gyvakarai1, which was dated to 4571-4422 calibrate Before Present (cal. BP), we will expect that the C14 dating range of I5884 overlapping with the one of Gyvakarai1 or being younger. However, we observe a non-overlapping range (4840-4646 cal. BP) and an older date for I5884. This unexpected age could be explained by the reservoir effect, which results in abnormally old C14 dates. This occurs with consumption of marine or freshwater resources, whereby, via deep geological filtering, carbon in these foodstuffs are derived from more ancient sources of carbonates than terrestrial sources of water. Given that this effect has been described in the Dereivka I site[26], this could explain the incongruence observed in of the dating of I5884. In order to address this, we took advantage of the high correlation between the age and the root-to-tip distance of the samples present in the LNBA lineage (Supplementary Figure 1) and estimated the molecular date of I5884 to 4413-4649 years BP (95 % HPD) using BEAST[27], which is in line with the expected calendar date based on the phylogenetic position (Figure 1B and C). Additionally, we recover a *Y. pestis* genome from an Iron Age individual (KZL002) from modern day Kazakhstan dated to 2736-2457 cal. BP. This is to our knowledge the youngest genome recovered from the LNBA lineage and provides evidence for the survival of this lineage until the Iron Age, and attests the presence of the LNBA lineage in Eurasia for at least 3,000 years. Despite the long survival of this lineage, there remains a lack of known modern representatives, and on the basis of the current data we assume the lineage to have gone extinct.

Intriguingly, we also were able to reconstruct a novel genome from an individual found in the dolmen “El Sotillo” (Spain, I2470). This represents the first evidence of prehistoric plague in the Iberian Peninsula. Despite the radiocarbon dating falling within the same interval of the other newly reported genomes, it does not fall with the rest of the LNBA lineage in the phylogeny. The I2470 genome branches off ancestral to the previously reported RT5 genome from Russia[14], which was the first genome identified as fully flea adapted. The I2470 genome represents another lineage of bubonic plague in Europe, highlighting the diversity of strains present in Eurasia shortly after the emergence of *Y. pestis*. The fact that we observe these two bubonic lineages in opposite

sides of Europe, raises questions on how widespread the flea adapted forms were during this period.

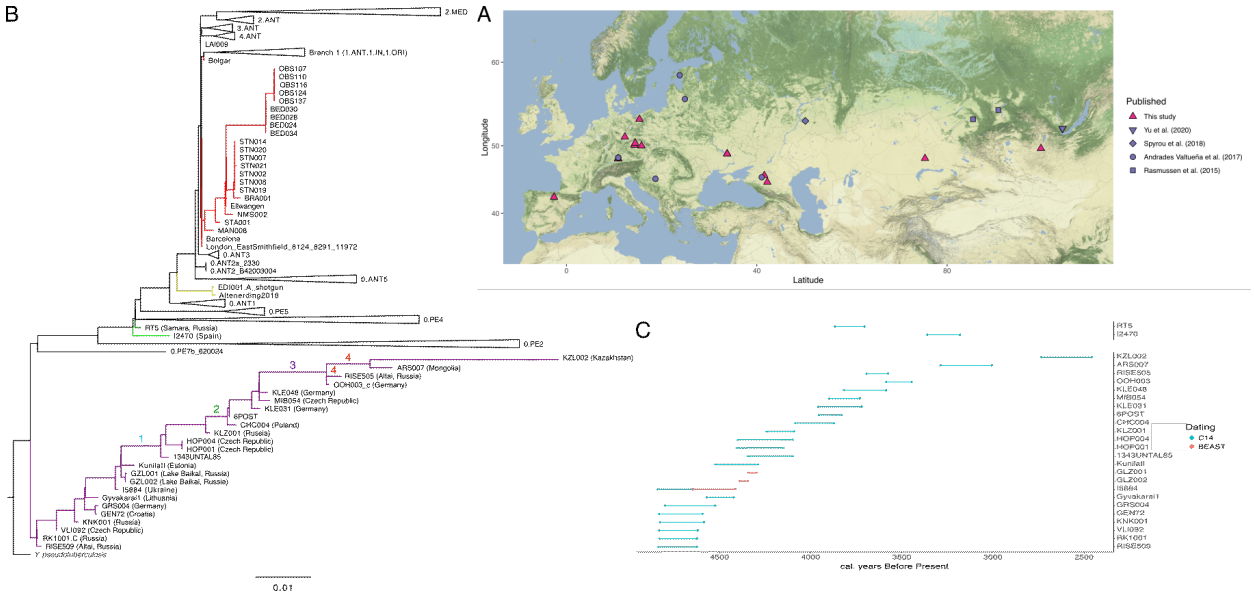


Figure 1: A) Archaeological sites where *Y. pestis* from the LNBA period have been recovered. B) Maximum Likelihood tree computed from a 98% partial deletion SNP alignment (n=5229). Numbers on the tree indicate the deletions detected in the genomes displayed in Supplementary Figure 4. Purple colour indicates the LNBA lineage, green colours the flea-adapted genomes from the Bronze Age, the yellow-green colour the first pandemic genomes and red the second pandemic genomes. C) Radiocarbon 2σ ranges (C14) or 95 % HPD BEAST intervals (BEAST) of the ages of the genomes from the LNBA period aligned to the corresponding tips in the Maximum Likelihood tree. Colours indicate if the date corresponds to the original radiocarbon date (C14) or if they were inferred with BEAST (BEAST), for details see Methods.

Genomic content of *Y. pestis* during the LNBA period

To evaluate the virulence potential of the newly recovered strains, we evaluated the status (presence/absence) of known *Y. pestis* virulence factors in these genomes (Figure 2). In the case of the I2470 genome we observe the presence of the complete set of virulence factors in the chromosome and *Y. pestis* specific plasmids, thus confirming that this genome is fully adapted to the flea vector, like RT5. We also confirm the presence in I2470 of the ancestral less-efficient *pla* variant after manual inspection (Supplementary Figure 3), previously reported in RT5[14] and all the other LNBA genomes[10–12,25]. On the other hand, all new genomes within the LNBA lineage also lack the ymt and YPMT1.66c genes, both of which are important for the efficient flea transmission. The lack of those genes and the presence of active ureD and biofilm regulators, which has been previously shown in [10–12], suggests a non-flea adapted form of plague. We also observe the absence of the *yapC* gene in the 1343UnTal85 genome and all subsequent genomes. In order to check if this phenomenon was related to a more substantial loss of genetic material described in [11,12], we performed a more in-depth analysis by detecting missing regions across the Y. pestis

CO92 chromosome. We detected multiple deletions bigger than 500bp in the LNBA lineage, which can be grouped into four events (Figure 1B, Supplementary Figure 4, Supplementary Table 2): the oldest event (1) occurred in the ancestor of 1343UnTal85 and involved the loss of a 35Kb region which contains, among others, the *yapC* gene; followed by the loss of a 1.5Kb region (2) in the ancestor of CHC004 (RISE139); a third region (3) of 2Kb was lost in the ancestor of OOH003 and RISE505, and finally, a larger deletion comprising 37Kb (4) was detected in the genomes RISE505, ARS007 and KZL002. Event 4 may provide further insights into the relationship of OOH003, RISE505, ARS007 and KZL002 genomes, where we observe a lack of resolution in the phylogeny. The phylogenetic algorithm used here groups OOH003 and RISE505 in a clade that is ancestral to ARS007 and KZL002. Based on this topology, the deletion event 4 would have occurred independently in the lineage leading to RISE505 and in the lineage that gave rise to ARS007 and KZL002. However, we can also use the deletion as supporting evidence for RISE505, ARS007 and KZL002 forming a clade, which had lost the 37Kb region after the split from the ancestor of OOH003. The latter requires a single event to describe the presence of the deletion and thus is a more parsimonious explanation. We also detect an ongoing gene loss after those four main events represented by five additional missing regions in the KZL002, one of them being 45 kb long.

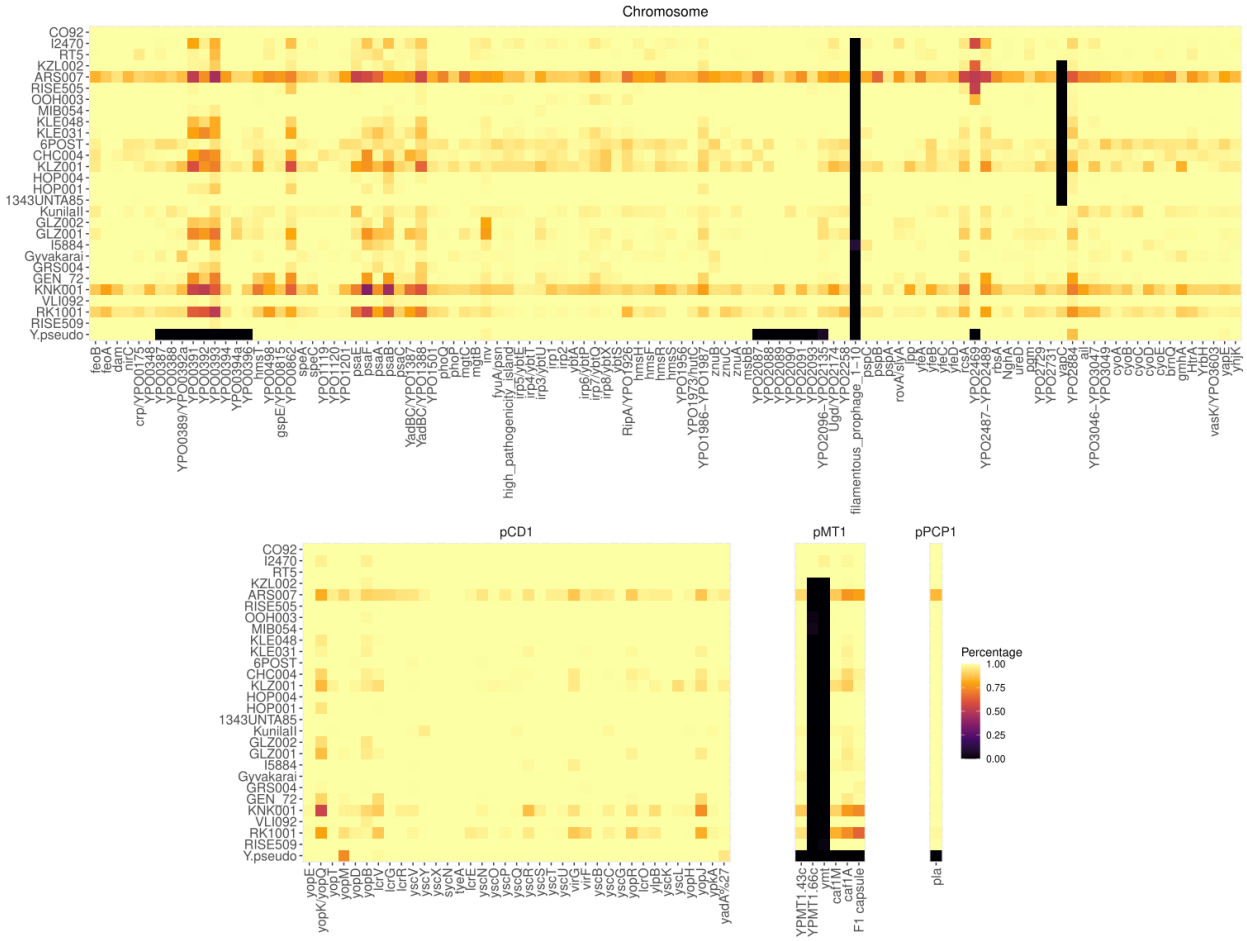


Figure 2: Heatmap displaying the presence or absence of 159 known virulence factors in *Yersinia pestis* genes of the chromosome (n=115), and the pCD1 (n=37), pMT1 (n=6), and pPCP1 (n=1)

plasmids. Blue represents 100 percent of the gene covered at least 1X while black represents 0 percent of the gene covered. Genomes are ordered based on their phylogenetic placement.

In order to check for the effect of single nucleotide polymorphisms (SNP) specific to the LNBA branch, we performed a SNP effect analysis with SNPEff. We detect a total number of 921 SNPs unique to the LNBA branch. Of those 465 SNPs are either intergenic (n=189) or synonymous (n=276) and probably represent neutral changes. In contrast, we observe the presence of 437 non-synonymous SNPs that could affect the protein function due to amino acid changes. However, the prediction of the effect is hard to infer with genetic information alone. We detect 20 substitutions that likely lead to pseudogenisation: 1 lost stop codon, 3 lost start codons and the gain of 16 stop codon. As with the deletions, we observe an accumulation of pseudogenes over time (Supplementary Figure 5). Interestingly, three of these affected genes (*flgB*, *flgF* and *fliZ*) are involved in flagella synthesis or are part of the flagellar system which is inactivated in modern *Y. pestis*, probably to evade the host's immune system[28].

LNBA genomes derive from a single lineage

The current diversity and genomic make-up of LNBA genomes appear to show different characteristics from those adapted to fleas and responsible for more recent plague epidemics. In order to test whether the genomes in the monophyletic LNBA branch evolved from a single population, which provided a perpetual source deme of the pathogen without parallel diversification, we explored the potential correlation between genetic versus geographical distance and genetic versus temporal distance. The rationale of this approach comes from the following three assumptions: a) we expect to see a correlation between geography and genetic affinity when genomes from the same location are closer to each other indicating the presence of multiple populations restricted to certain geographical areas; b) for a single population evolving through time we also expect a correlation between genetic affinity and temporal distance; c) if no correlation is observed between geography and genetic distance or between time and genetic distance, this suggests a globally distributed diversity of the bacteria from which we randomly sampled any given clade at any given time. To provide comparative datasets for this analysis, three more ancient bacterial datasets were included: (1) *Y. pestis* genomes dating to the second pandemic that form part of the European lineage that emerged from the Black Death clone[2,3,6,7], (2) *Salmonella enterica*[29–31] and (3) *Mycobacterium leprae*[32–34] (Supplementary Table 3).
Contrasting the results from all four cases under study, we observe a strong positive correlation between genetic distance and time (R2=0.95006) in the LNBA genomes, indicative that those arose from a single lineage. We also observe a minor contribution to the genetic distance explained by geography (R2=0.0344, Figure 3A). In the case of the second pandemic *Y. pestis* genomes, we expected to see a weak correlation between distance due to parallel lineages evolving through time and no correlation between geography and genetic distance, since we know that a single clone was responsible for the Black Death that spread across a large geographic area of Europe. We find this assumption confirmed by a weak but significant correlation between genetic distance and time (Figure 3B). We observe non-significant p-values in either of the correlations for *S. enterica* nor *M. leprae* (Figure 3C and D). This is probably due to the fact that

contemporaneous reconstructed genomes are distributed across the phylogeny of the species, independently of their location [29,32,33].

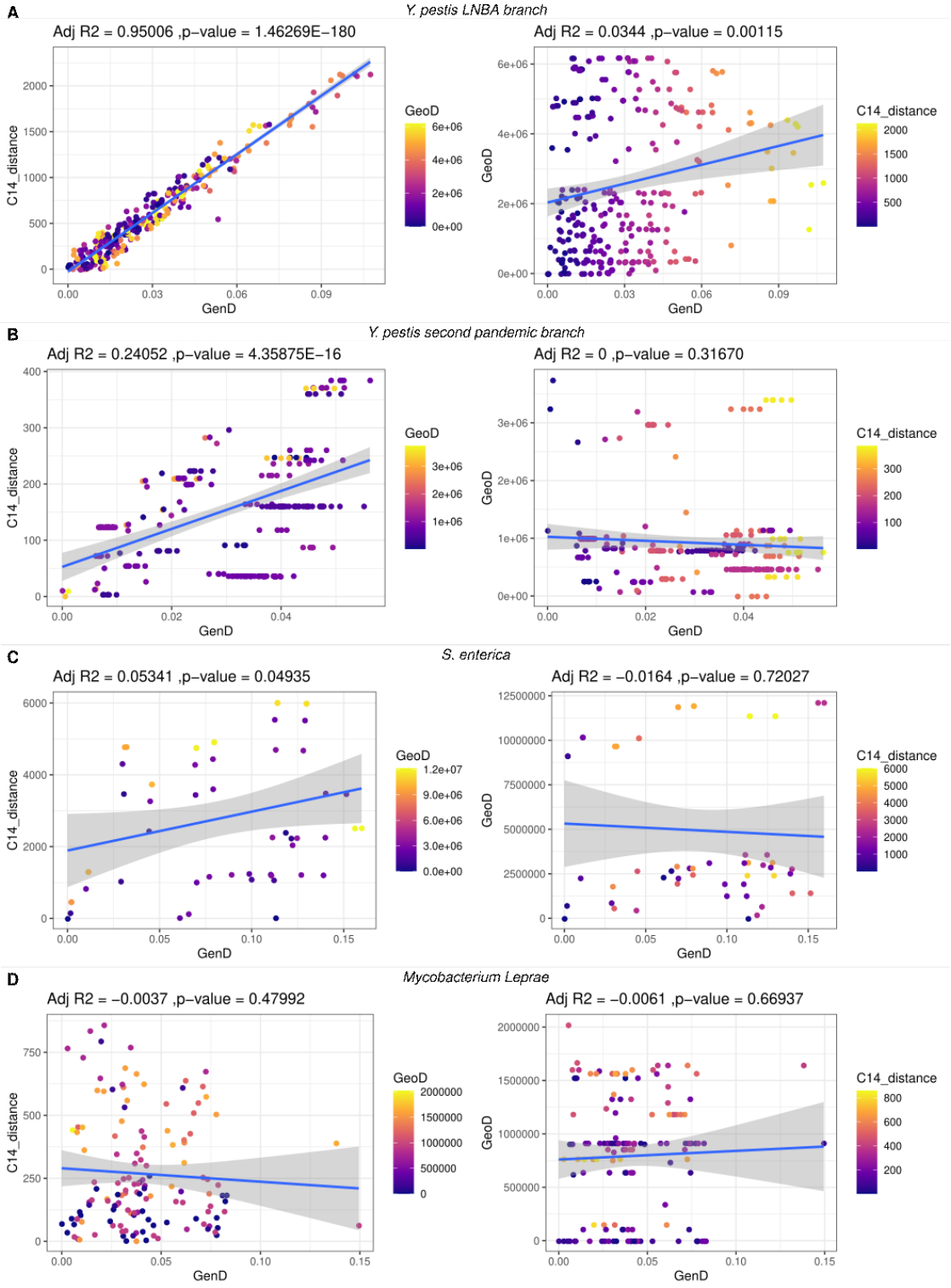


Figure 3: Correlations between time (C14_distance) and genetic (GenD) distance with colour indicating the geographical (GeoD) distance (left-hand side) and geographical and genetic distance with colour indicating time distance (right-hand side) for A) LNBA genomes, B) Second

pandemic genomes, C) *Salmonella enterica* ancient genomes, and D) *Mycobacterium leprae* ancient genomes.

Discussion

Ancient *Y. pestis* genomes recovered from humans who lived between 5000-2000 years Before Present (yBP) have contributed to the understanding of the early evolution of this pathogen[10–12,14]. The earliest evidence of plague in human to-date dates to around 5000 years ago, when two different strains have been observed in Eurasia: the most ancestral lineage recovered is a *Y. pestis* genome from Sweden[10], followed by the ancestral strains of the LNBA lineage[11,12]. Those basal lineages lacked genetic adaptations that have been shown to be essential for the efficient transmission of this bacterium via the blocked flea vector, namely the *ymt* gene[35], the silencing of biofilm regulators[36], and *ureD*[37]. Furthermore, in the period between 4000 to 2400 yBP, we observe the coexistence of the LNBA lineage with fully flea-adapted lineages, from which two ancient genomes have been reconstructed: RT5, recovered from an individual from the Samara region in Russia[14], and I2470, a representative of a newly discovered clade from an individual from Spain. These results show that a wide diversity of *Y. pestis* lineages was present in Eurasia shortly after the emergence of all known *Y. pestis* strains, with the most recent estimates dating to around 5,700 years BP (95% HPD interval: 5,250–6,364 BP)[10]. This time period is characterized by an intensification of human mobility across Eurasia, starting around 5000 years ago with the documented expansion of pastoralist groups, such as those associated with the Yamnaya and related cultural complexes, both eastward and westward from the Eurasian steppes[38,39]. The development of new forms of transport such as oxen-drawn carts and wagons and the domestication of horses will have aided in the rapid dispersal of humans[40]. The fact that we observe the LNBA lineage to be present across Eurasia poses the question whether humans were involved in the dispersal of plague during this period. Given that the Eurasian steppe served as a corridor for the connection between geographically distant human populations, especially in combination with intensified and expanding pastoralism during this period, suggests increased contact or habitat overlap between wild animals, such as rodents, and humans and their domesticates. In fact, modern *Y. pestis* studies have shown that Tibetan sheep can act as intermediate hosts between infected marmots and humans[41]. As such, a scenario with a similar transmission chain can be anticipated in the past, in which the pastoral grassland of the steppe zone served as an interaction zone between domesticates and sylvatic hosts. That, in combination with the increased human mobility could have facilitated the connection of ecosystems and habitats that otherwise would have not come into contact, therefore creating opportunities for the dispersal of diseases into and across new territories and hosts.

However, in order to understand the dispersal of the LNBA lineage, we need to understand its means of transmission. This is rather difficult to infer since no close modern relative with similar genetic characteristics has been isolated to date, strongly suggesting that this lineage has since become extinct. Based on the structure of the phylogenetic tree, we hypothesise that humans were not the only sustaining hosts of the disease, as we do not observe palaeoepidemiological patterns matching major human outbreaks nor a diversification into parallel evolving lineages. On the contrary, all individuals were buried with great care according to the respective local burial customs, indicating that the cause of death was not perceived as unusual. In addition, we have

confidently shown that the LNBA genomes form a single lineage that does not experience parallel diversification through time, potentially indicating a single or unified reservoir of the disease in an eco-geographic zone, in which zoonotic events must have occurred more frequently. We speculate that the Eurasian steppe belt could have been both a zone of habitat overlaps as well as corridor of mobility spanning vast geographic distances. The wide geographic spread of the LNBA and the fact that it also reached regions beyond the steppe (e.g. mixed temperate forest zones in Central Europe, the Altai and Lake Baikal regions) speaks for the involvement of wild animals and domesticates engaged in long-distance travel/droving.

We also observe an increase in the pseudogenisation and genetic loss during the evolution of the LNBA lineage starting around 4200 years ago. This could be an indication of strong selection pressure in the bacterium population[42] or a sign of an adaptation to a new host[43,44]. How this lineage was transmitted currently remains uncertain. Although lacking the genetic makeup, it could still have been inefficiently transmitted by fleas since it has been proven that no genetic changes are required for the 'early-phase transmission' mechanism[45–48]. Another potential route for transmission is the oral-faecal route as is the case for the *Y. pestis* ancestor *Y. pseudotuberculosis*. There have been reports of plague happening via this route by the consumption of animals infected by plague (camels[20,21], goats[20], marmots[22]). The LNBA *Y. pestis* strains would still have had a higher capacity of disseminating than its ancestor, *Y. pseudotuberculosis*, since the *pla* gene, involved in the dissemination of the bacterium in the mammalian host[49], had already been acquired at this point of evolution. Finally, it has been previously suggested that the initial form of plague was pneumonic in its nature[50]. This is the rarest form of plague nowadays with only a few reports[15–19], but it has been documented to be possible via the inhalation of blood droplets during the process of skinning animals that died of plague[51]. While all these transmission scenarios are possible, we suspect that the dissemination between humans was limited; supported by the observation of no genetic or archaeological evidence of major outbreaks caused by the LNBA lineage.

Overall, we observe the long co-existence of two forms of *Y. pestis* (a fully flea-adapted and a non-adapted form), which likely lasted for more than 2500 years of the bacterium's evolution. Whether these forms were competing in the same ecological niche, co-existed among the same hosts, or occupied entirely different niches is a research topic that requires further examination. For the flea-adapted forms (RT5 and I2470), we can assume a similar transmission cycle than that observed in modern contexts, where fleas serve as vectors of disease and maintain the transmission of the bacterium in the rodent host population[13]. For the non-adapted form, further ancient genomes from the LNBA period, particularly recovered from animal remains, combined with modern functional studies that reproduce the genetic characteristics of the early plague, would be interesting avenues of future research to shed light into the transmission mechanism of early forms of plague.

Methods

Data generation and screening

A total of 227 individuals from 13 sites in Eurasia dating between ~5000-2000 years Before Present (yBP) were screened for ancient DNA evidence suggestive of the presence of *Y. pestis* (Supplementary Table 1). Laboratory work was performed at the ancient DNA lab facilities of the Max Planck Institute for the Science of Human History in Jena, the Max Planck Institute of Evolutionary Anthropology in Leipzig, and the Department of Genetics, Harvard Medical School in Boston. For samples from the Dereivka I site, DNA extraction was performed in Earth Institute and School of Archaeology, University College Dublin, Belfield, Dublin 4, Republic of Ireland, and sent to Boston for further processing, as previously described[52].

In Jena, teeth were irradiated with UV light from two sides for 30 min each and then cut at the cemento-enamel junction. With a dentist drill, 30-60 mg of powder were drilled from the surface of the inner pulp chamber and the root canals. DNA from this powder was extracted with an established protocol[53,54]. The powder was incubated with 1 ml of extraction buffer (0.45 M EDTA, 0.25 mg/ml Proteinase K, pH 5-6) for at least 16 h at 37°C under rotation. After centrifugation for 2 min at 15,800 g, the supernatant was mixed with a 10 ml binding buffer (5 M guanidine hydrochloride, 40 % isopropanol) and 400 µl 3M sodium acetate. To bind the DNA, silica-based spin columns were used (High Pure Viral Nucleic Acid Large Volume Kit; Roche). After washing twice, DNA was eluted in two steps in TET (10 mM Tris, 1 mM EDTA, 0.05 % Tween-20, pH 8) to a final volume of 100 µl.

From these extracts, double-indexed double-stranded Illumina sequencing libraries were prepared[55] with initial USER enzyme (New England Biolabs (NEB)) treatment to reduce aDNA damage in form of deaminated cytosines[56]. 25 µl of extract were mixed with 0.072 U USER Enzyme in Tango buffer (Life technologies), 1.2 mM ATP, 0.2 mg/ml BBSA, 0.4 mM dNTPs and incubated for 30 min at 37˚C and 1 min at 12˚C. The reaction was stopped by adding 0.1343 U Uracil Glycosylase Inhibitor (UGI, NEB) and another incubation for 30 min at 37˚C and 1 min at 12˚C. After adding 0.515 U of T4 Polynucleotide Kinase (NEB) and 0.085 U T4 Polymerase (NEB), the mix was incubated at 25˚C for 20 min and 12˚C for 10 min. DNA was purified with the MinElute PCR Purification Kit (Qiagen) and eluted in 20 µl elution buffer EB containing 0.05 % Tween-20 (EBT). The eluate was mixed with 0.25 µM sequencing adapters and 0.125 U Quick Ligase (NEB) in Quick Ligase buffer and incubated for 20 min at room temperature. DNA was purified as described above and eluted in 22 µl EBT. DNA was then incubated with 0.4 U Bst Polymerase and 0.5 mM dNTPs in Isothermal buffer (NEB) at 37˚C for 30 min and 80˚C for 10 min. All libraries were double-indexed with a unique pair of 8 nt long indices in reactions of 0.025 U Pfu Turbo Polymerase (Agilent Technologies), 100 mM dNTPs, 0.3 mg/ml BSA, and 0.2 µM of indices in Pfu Turbo buffer in a thermocycler with the following program: 2 min at 95˚C, ten cycles of 30 sec at 95˚C, 30 sec at 58˚C and 1 min at 72˚C, and a final elongation with 10 min at 72˚C.

The laboratory process for the archaeological sites of Dereivka I and dolmen “El Sotillo” have been previously described in [52,57] respectively.

All libraries were shotgun sequenced to 5 million reads on an Illumina HiSeq 4000 with a single-end kit (1 x 76+8+8 cycles) in Jena or in Boston and screened for the presence of pathogen DNA as described below.
For samples OOH003, KNK001, KLZ001 and ARS007, additional single-stranded libraries (sslib) were prepared for in-solution capture from 30 µl DNA extract with an automated protocol[58]. All single-stranded libraries were prepared in the Jena laboratories with the exception of OOH003 which was prepared at the Leipzig facility.

Pathogen Screening

Pre-processed reads from shotgun sequencing data (227 samples in total) were screened for the presence of pathogen DNA using the screening pipeline HOPS[23]. In the first step, adapter-clipped reads were used as input for the MEGAN Alignment Tool (MALT) (v.0.4.0, [30]) and mapped against a custom RefSeq Genome database comprising all complete viral and bacterial genomes (as of 2017), a selection of eukaryotic pathogen genomes and the human reference sequence GRCHh38. Mapping parameters were set to a minimum of 90 % identity (--minPercentIdentity) and top percent value (--topPercent) as well as minimum support (--minSupport) to 1 with BlastN mode and semi-global alignment type. All other parameters were used in default settings. The output was then filtered as implemented in MaltExtract of HOPS with a predefined list of pathogens, and assigned reads were evaluated based on aDNA damage patterns and their edit distance to the reference genomes. Additionally, MALT mapping results were visually inspected in MEtaGenome Analyzer (MEGAN)[24].

Y. pestis enrichment

Libraries putatively positive for *Y. pestis* DNA were amplified to a concentration of 200-400 ng/µl with IS5/IS6 primers and enriched for *Y. pestis* DNA with in-solution whole genome capture as described before[12]. The probe set was designed with a combination of *Y. pestis* genomes including *Y. pestis* CO92 chromosome (NC_003143.1), CO92 plasmid pMT1 (NC_003134.1), CO92 plasmid pCD1 (NC_003131.1), KIM 10 chromosome (NC_004088.1), Pestoides F chromosome (NC_009381.1) and *Y. pseudotuberculosis* IP 32952 chromosome (NC_006155.1) as a template. The capture was performed on 96-well plates in two rounds.
After capture, samples were sequenced on an Illumina HiSeq 4000 platform with a 75 bp paired-end kit (2 x 76+8+8 cycles; samples ARS007, GRS004, HOP001, HOP004, KNK001, MIB054, I5884, I2470, KZL002, KLZ001) and/or a 75 bp single-end kit (sample KLZ001, OOH003ss). Additionally, the KLE031, KLE048, OOH003 samples were sequenced with a 75bp single-end kit and MIB054 with 75 bp paired-end on an Illumina NextSeq.

Data processing

Raw data was processed with nf-core/eager[59], with the exception of I5884 and I2470 samples that required a preprocessing step to remove the 7bp internal barcodes. These samples were preprocessed as follows: we ran AdapterRemoval v.2.3.1[60] to clip only adapters and extracted reads containing the barcodes using grep (v.3.1, http://www.gnu.org/software/grep/). We ran the

script fastq_trimming_barcodes.sh, which removed the barcodes by trimming 7 bp from each end using FASTX-trimmer v.0.0.14[61], and removed the reads without a pair (one of reads of the pair did not contain a barcode) with filterbyreadname.sh from BBmap from the tool suite BBTools[62]. The fastq files after removal of the barcodes were uploaded in the read repository. The preprocessed read pairs for the libraries were listed in a tab separated value (tsv) file used as input for nf-core/eager together with the rest of the raw data. nf-core/eager was run with the tsv input and the following processes were run: FASTQC v0.11.4[63] was run to evaluate the quality of the sequencing data. Adapter clipping, filtering of short (<30bp) or low-quality reads and merging of pair-end data were performed with AdapterRemoval. Prior to mapping, reads from the same libraries were merged. We mapped the reads against the *Y. pestis* CO92 reference (NC_003143.1) with bwa aln v.0.7.12[64] adapting the seed length (-l) to 16 and the mismatch allowance (-n) to 0.01, in order to increase the sensitivity of the mapping that can be compromised due to the highly fragmented nature of ancient DNA. To retain reads mapping uniquely to the reference, we filtered reads with mapping quality lower than 37 with samtools v1.3[65]. We then removed duplicates with Picard Tools v1.140 MarkDuplicates[66], merged different libraries from the same individual, and calculated the mapping statistics with Qualimap v. 2.2.1[67,68]. To authenticate the ancient origin of the molecules characterised by the deamination of C → T due to hydrolytic damage, we calculated the deamination patterns with DamageProfiler v.0.4.9[69]. To remove potential bias introduced due to the aforementioned deamination, we removed the damaged bases before further analysis. This was achieved by removing 1 bp from each end of the reads using FASTX-trimmer, except for the single-stranded libraries where no trimming nor additional mapping was applied. The resulting trimmed fastq files were processed as before with exception of the skipping of AdapterRemoval and setting the bwa aln parameters to -n 0.1 and -l 32 to be stricter during the mapping step. As before, duplicates were removed with Picard Tools MarkDuplicates. Libraries produced from the same individual were combined after the removal of duplicates and mapping statistics were then calculated with Qualimap as described above. The obtained bam files were then realigned with GATK v3.5 realigner and vcf files were obtained using GATK UnifiedGenotyper[70].

To explore the presence of the *Y. pestis* plasmids in the ancient samples, we repeated the mapping steps by independently mapping the preprocessed reads to the pCD1 (NC_003131.1), pMT1 (NC_003134.1) and pPCP1 (NC_003132.1) CO92 plasmids.

Variant calling, SNP effect and phylogenetic analysis

To produce the final SNP alignment containing all the variable positions to be used in the phylogenetic analysis, we ran MultiVCFAnalyzer[43] (https://github.com/alexherbig/MultiVCFAnalyzer) with the vcf files for the ancient genomes from this study but also including previous ancient samples from the first pandemic[4,9], second pandemic[2,3,6,7], prehistoric genomes[11,12,14,25] and modern genomes[71–78,78–82] summarised in Supplementary file 1, Table 2. In order for a SNP to be called, it must fulfil the following criteria: the allele must be supported by 90% of the reads, with a minimum of 3 reads supporting the call. For the single-stranded libraries an additional genotyping was performed, utilising the characteristic of the library construction for which we only have C → T observable damage. Here, damage will be observed as C → T change in forward mapping reads and as G → A in the reverse

mapping reads. We split the realigned bam file into forward and reverse mapping reads with samtools view, which are then ran together with the complete dataset in MultiVCFAnalyzer with the same parameters as above. We implemented the publicly available genoSL.R script (https://github.com/aidaanva/GenoSL) for the purpose of genotyping single-stranded libraries as follows: for all substitutions to T calls were drawn based on the reverse mapping reads in that position, for all A substitutions calls were drawn from the forward mapping reads and for all other substitution the call was drawn from the complete dataset. The resulting SNP Table and a corrected SNP alignment was then used for further processing.

Previous studies have shown that an abundance of environmental background in metagenomic datasets can result in the incorporation of false SNP calls during ancient bacterial genome reconstruction[7]. Most often, such erroneous calls manifest themselves in a phylogenetic analysis as private SNPs, as they are unlikely to be shared with the ingroup diversity[7,9]. As a consequence, erroneously calculated private branch lengths can interfere with evolutionary inferences and divergence date estimates[7]. Here, in order to filter out private SNP calls that result from environmental contaminants, we used the SNP Evaluation[9] (https://github.com/andreasKroepelin/SNP_Evaluation) to co-analyse all newly reported genomes in this study as well as previously published genomes RISE509, RISE505[11], Gyvakarai1, KunilaII, 6Post, 1343UnTal85, GEN72, RK1001[12], GZL001 and GZL002[25]. SNP calls were evaluated on the basis of sample laboratory processing, where UDG-treated data were assessed under different criteria from non-UDG-treated data to account for substitutions associated with aDNA damage in the latter.

For UDG-treated genomes, unique SNPs were assessed within a 50 bp window and were accepted as “TRUE” when:

1. A comparison between lenient mapping (BWA parameters -n 0.01, -l 16) and stringent mapping (BWA parameters -n 0.1, -l 32) resulted in <10% coverage increase around each SNP
2. No heterozygous SNP positions were identified within the evaluated region (50 bp window) around each private SNP
3. No gaps in genomic coverage were observed within the evaluated region around each private SNP

For non-UDG-treated genomes, unique SNPs were assessed within a 50 bp window using only lenient mapping parameters (BWA setting -n 0.01, -l 16) and were accepted as “TRUE” when:

1. The evaluated SNP position was not confined by aDNA damage
2. Heterozygous SNP positions within the evaluated region (50 bp window) are only permitted when consistent with aDNA damage (C-to-T or G-to-A substitutions)
3. No gaps in genomic coverage were observed within the evaluated region around each private SNP

All the positions assessed can be found in Supplementary File 1 Table 3-4. The false-positive SNPs were excluded with MultiVCFAnalyzer and genoSL.R was run to obtain the final SNP table and final SNP alignment containing all the variant sites in the dataset analysed. The effect of SNPs specific to the LNBA lineage was analysed with SnpEff v3.1[83]. The snpTableForSnpEff.tsv output from MultiVCFAnalyzer was used as input for SnpEff with a prebuilt SnpEff database based on the *Y. pestis* reference genome CO92 (NC_003143.1). The resulting annotated file was included in a second run of MultiVCFAnalyzer and genoSL.R as described above to obtain a

snpTable annotated with the effects which was then filtered for SNPs found exclusively in the LNBA lineage.
Finally, the SNP alignment file was filtered to contain sites where 98% of the genomes have a call (98% partial deletion) with MDF.R (https://github.com/aidaanva/MDF) and used to compute a Maximum Likelihood tree with RAxML-NG (v. 0.9.0, https://github.com/amkozlov/raxml-ng) with the following command:

```
raxml-ng --all --msa $CurrentAlignment --model GTR+G --seed 2 --threads 9 --bs-trees autoMRE --
                                        prefix $name
```

where $CurrentAlignment is the fasta alignment and $name is the output prefix for the current run.

Bayesian molecular dating of I5884

Given the incongruence between the phylogenetic positioning and radiocarbon date of I5884 (Figure 1B and C), we applied a molecular dating approach using the program BEAST v1.10[27] to re-evaluate the specimen's age. For this, we initially used the program TempEst v1.5 (http://tree.bio.ed.ac.uk/software/tempest/) to assess the temporal signal across the LNBA lineage, using all previously published and newly available genomes as well as their associated calibrated median radiocarbon dates (see Supplementary Figure 1). Overlapping variant positions across all LNBA isolates and the modern branch 0 strain 0.PE2 Pestoides F (used as outgroup) were used for the construction of a maximum parsimony tree (90% partial deletion retaining 766 SNPs) in MEGA7[84], which was used as input for TempEst in NEXUS format. Our analysis revealed a near perfect correlation between specimen ages and their tip distances from the root (R^2=0.958), which permitted us to pursue a tip dating analysis for I5884. Subsequently, the same SNP dataset was used as input for BEAST v1.10, including all available calibrated radiocarbon age ranges in years BP as uniform priors and the age of the 0.PE2 isolate set to 0 (See https://github.com/aidaanva/LNBAplague/blob/main/Data/2020-07-09_LNBA_leprosy_enterica_comp/LNBA_transect/Metadata_coordinates_dating_sex_updated_def.csv). Instead, the uniform prior used for the age of I5884 was set to span the entire currently known temporal range of the LNBA lineage, between 5000 and 2500 years BP. We used BEAUti v1.10 to set up two separate analyses using the coalescent constant size and coalescent skyline tree priors, both with a lognormal relaxed clock and a GTR substitution model (four gamma rate categories). For each of the two analyses, three independent chains of 50,000,000 states were run in BEAST v1.10. Subsequently, the three chains were combined using LogCombiner with a 10% burn-in. Finally, the runs were inspected in Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/) to ensure that the effective samples sizes (ESS) of all computed parameters are > 200. Evaluation of the posterior date of I5884 in Tracer v1.6 revealed overlapping ages ranges between both computed analyses spanning from 4413 to 4649 years BP, which is in line with the genome's phylogenetic positioning.

Virulence analysis and indel analysis

To assess the presence and absence of known virulence factors in *Y. pestis*, we compiled a bed file containing the coordinates for genes on the chromosome (n=115), and the pCD1 (n=37),

pMT1 (n=6), and pPCP1 (n=1) plasmids of *Y. pestis* CO92. In order to account for regions that may have mappability issues (e.g. duplicated regions), we mapped the trimmed reads and the sslib reads for OOH003 as above with the exception that no mapping quality filter was applied (--bam_mapping_quality_threshold 0). The output bam files were then used to calculate the percent of the gene covered using bedtools v2.25.0[85] and prepared the data for R using an Generate_bed_files.sh. The resulting bed files were concatenated together using the cat command and the final files can be found in https://github.com/aidaanva/LNBAplague/tree/main/Data/Virulence. The results were plotted in R[86] using the ggplot2 package[87].

Additionally, we used the resulting non-filtered bam files to explore the presence of chromosomal deletions using *Y. pestis* CO92 as reference. We recovered non-covered regions from bam files as follows: bedtools genomecov was used to calculate the non-covered regions per sample. Non-covered regions separated by less than 100 bp were then merged together and subsequently filtered to have a minimum size of 500bp. We also calculated the percentage of coverage for each missing window to account for sparse data in low coverage genomes. The resulting files per sample were then combined and analysed with R. Additionally, we extracted the genes affected by any deletion. All these steps were implemented in the script IndelCheck.sh. For the missing regions, we plotted deleted regions containing less than 15% of the region covered using the ggplot2 and ggalt[88] packages.

Phylogeography and temporal testing

To test whether the genomes in the LNBA lineage are indeed descendants of one another, we tested whether there is a linear correlation between either genetic and geographical distance or genetic and temporal distance. We performed this analysis in R by calculating the genetic distance as the pairwise distance using the dist.dna function of the ape package[89] and as input the filtered snpAlignment.fasta from MultiVCFAnalyzer to contain only the LNBA genomes and their variable sites. The geographic coordinates were collected from each of the archaeological sites used in this study (https://github.com/aidaanva/LNBAplague/blob/main/Data/2020-07-09_LNBA_leprosy_enterica_comp/LNBA_transect/Metadata_coordinates_dating_sex_updated_def.csv) and pairwise linear distances were calculated as the crow flies using the distm function of the geosphere package[90]. Finally, the median years Before Present (yBP) radiocarbon date was used to calculate the temporal pairwise distances using the outer function from base R. A linear model was fit for either genetic versus geographic distance or genetic versus temporal distance using the lm function in R. The correlations were plotted using ggplot2.

In order to provide comparative data for these correlations, we performed the same analysis using high coverage genomes from the second plague pandemics, ancient leprosy (*M. leprae*) and ancient *Salmonella* data (Supplementary Table 3, see https://github.com/aidaanva/LNBAplague/tree/main/Data/2020-07-09_LNBA_leprosy_enterica_comp subfolders for the data). The final figure was generated in R using the ggpubr package[94].

All the previously described R code can be found in the R notebook here: https://github.com/aidaanva/LNBAplague/blob/main/Stone_Age_Plague_v5.Rmd.

Code availability

All scripts and code mentioned can be found at https://github.com/aidaanva/LNBAplague

References

1. Perry, R. D. & Fetherston, J. D. Yersinia pestis--etiologic agent of plague. *Clin. Microbiol. Rev.* **10**, 35–66 (1997).
2. Bos, K. I. *et al.* Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *eLife* **5**, e12994 (2016).
3. Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506–510 (2011).
4. Feldman, M. *et al.* A high-coverage Yersinia pestis Genome from a 6th-century Justinianic Plague Victim. *Mol. Biol. Evol.* msw170 (2016) doi:10.1093/molbev/msw170.
5. Wagner, D. M. *et al.* Yersinia pestis and the Plague of Justinian 541–543 AD: a genomic analysis. *Lancet Infect. Dis.* **14**, 319–326 (2014).
6. Spyrou, M. A. *et al.* Historical Y. pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe* **19**, 874–881 (2016).
7. Spyrou, M. A. *et al.* Phylogeography of the second plague pandemic revealed through analysis of historical Yersinia pestis genomes. *Nat. Commun.* **10**, 1–13 (2019).
8. Namouchi, A. *et al.* Integrative approach using *Yersinia pestis* genomes to revisit the historical landscape of plague during the Medieval Period. *Proc. Natl. Acad. Sci.* **115**, E11790 (2018).
9. Keller, M. *et al.* Ancient Yersinia pestis genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). *Proc. Natl. Acad. Sci.* **116**, 12363–12372 (2019).
10. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* **176**, 295-305.e10 (2019).
11. Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago.

Cell **163**, 571–582 (2015).

12. Andrades Valtueña, A. *et al.* The Stone Age Plague and Its Persistence in Eurasia. *Curr. Biol.* **27**, 3683-3691.e8 (2017).
13. Stenseth, N. C. *et al.* Plague: Past, Present, and Future. *PLOS Med.* **5**, e3 (2008).
14. Spyrou, M. A. *et al.* Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. *Nat. Commun.* **9**, 2234 (2018).
15. Ratsitorahina, M., Chanteau, S., Rahalison, L., Ratsifasoamanana, L. & Boisier, P. Epidemiological and diagnostic aspects of the outbreak of pneumonic plague in Madagascar. *Lancet Lond. Engl.* **355**, 111–113 (2000).
16. Richard, V. *et al.* Pneumonic Plague Outbreak, Northern Madagascar, 2011. *Emerg. Infect. Dis.* **21**, 8–15 (2015).
17. Lien-teh, W., Chun, J. W. H., Pollitzer, R. & Wu, C. Y. Plague : a Manual for Medical and Public Health Workers. *Plague Man. Med. Public Health Work.* (1936).
18. Begier, E. M. *et al.* Pneumonic plague cluster, Uganda, 2004. *Emerg. Infect. Dis.* **12**, 460–467 (2006).
19. Bertherat, E. *et al.* Lessons learned about pneumonic plague diagnosis from two outbreaks, Democratic Republic of the Congo. *Emerg. Infect. Dis.* **17**, 778–784 (2011).
20. Christie, A. B., Chen, T. H. & Elberg, S. S. Plague in Camels and Goats: Their Role in Human Epidemics. *J. Infect. Dis.* **141**, 724–726 (1980).
21. Arbaji, A. *et al.* A 12-case outbreak of pharyngeal plague following the consumption of camel meat, in north–eastern Jordan. *Ann. Trop. Med. Parasitol.* **99**, 789–793 (2005).
22. Kehrmann, J. *et al.* Two fatal cases of plague after consumption of raw marmot organs. *Emerg. Microbes Infect.* **9**, 1878–1880 (2020).
23. Hübler, R. *et al.* HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biol.* **20**, 280 (2019).
24. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of

Large-Scale Microbiome Sequencing Data. *PLOS Comput. Biol.* **12**, e1004957 (2016).

25. Yu, H. *et al.* Paleolithic to Bronze Age Siberians Reveal Connections with First Americans and across Eurasia. *Cell* **181**, 1232-1245.e20 (2020).
26. Lillie, M., Budd, C., Potekhina, I. & Hedges, R. The radiocarbon reservoir effect: new evidence from the cemeteries of the middle and lower Dnieper basin, Ukraine. *J. Archaeol. Sci.* **36**, 256–264 (2009).
27. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, (2018).
28. Minnich, S. A. & Rohde, H. N. A Rationale for Repression and/or Loss of Motility by Pathogenic Yersinia in the Mammalian Host. in *The Genus Yersinia: From Genomics to Function* (eds. Perry, R. D. & Fetherston, J. D.) 298–311 (Springer, 2007). doi:10.1007/978-0-387-72124-8_27.
29. Key, F. M. *et al.* Emergence of human-adapted Salmonella enterica is linked to the Neolithization process. *Nat. Ecol. Evol.* **4**, 324–333 (2020).
30. Vågene, Å. J. *et al.* Salmonella enterica genomes recovered from victims of a major 16th century epidemic in Mexico. *bioRxiv* 106740 (2017) doi:10.1101/106740.
31. Zhou, Z. *et al.* Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* **28**, 2420-2428.e10 (2018).
32. Schuenemann, V. J. *et al.* Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. *Science* **341**, 179–183 (2013).
33. Schuenemann, V. J. *et al.* Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLOS Pathog.* **14**, e1006997 (2018).
34. Mendum, T. A. *et al.* Mycobacterium leprae genomes from a British medieval leprosy hospital: towards understanding an ancient epidemic. *BMC Genomics* **15**, 270 (2014).
35. Hinnebusch, B. J. *et al.* Role of Yersinia Murine Toxin in Survival of Yersinia pestis in the

Midgut of the Flea Vector. *Science* **296**, 733–735 (2002).

36. Sun, Y.-C., Hinnebusch, B. J. & Darby, C. Experimental evidence for negative selection in the evolution of a Yersinia pestis pseudogene. *Proc. Natl. Acad. Sci.* **105**, 8097–8101 (2008).
37. Chouikha, I. & Hinnebusch, B. J. Silencing urease: A key evolutionary step that facilitated the adaptation of Yersinia pestis to the flea-borne transmission route. *Proc. Natl. Acad. Sci.* **111**, 18709–18714 (2014).
38. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
39. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
40. Anthony, D. W. *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World.* (2007).
41. Dai, R. *et al.* Human plague associated with Tibetan sheep originates in marmots. *PLoS Negl. Trop. Dis.* **12**, e0006635 (2018).
42. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-Driven Gene Loss in Bacteria. *PLOS Genet.* **8**, e1002787 (2012).
43. Ochman, H. & Moran, N. A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
44. Sheppard, S. K., Guttman, D. S. & Fitzgerald, J. R. Population genomics of bacterial host adaptation. *Nat. Rev. Genet.* **19**, 549–565 (2018).
45. Vetter, S. M. *et al.* Biofilm formation is not required for early-phase transmission of Yersinia pestis. *Microbiology* **156**, 2216–2225 (2010).
46. Johnson, T. L. *et al.* Yersinia murine toxin is not required for early-phase transmission of *Yersinia pestis* by *Oropsylla montana* (Siphonaptera: Ceratophyllidae) or *Xenopsylla cheopis* (Siphonaptera: Pulicidae). *Microbiology* **160**, 2517–2525 (2014).
47. Eisen, R. J. *et al.* Early-phase transmission of *Yersinia pestis* by unblocked fleas as a

mechanism explaining rapidly spreading plague epizootics. *Proc. Natl. Acad. Sci.* **103**, 15380–15385 (2006).

48. Eisen, R. J., Dennis, D. T. & Gage, K. L. The Role of Early-Phase Transmission in the Spread of *Yersinia pestis*. *J. Med. Entomol.* **52**, 1183–1192 (2015).
49. Sebbane, F., Jarrett, C. O., Gardner, D., Long, D. & Hinnebusch, B. J. Role of the Yersinia pestis plasminogen activator in the incidence of distinct septicemic and bubonic forms of flea-borne plague. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5526–5530 (2006).
50. Zimbler, D. L., Schroeder, J. A., Eddy, J. L. & Lathem, W. W. Early emergence of Yersinia pestis as a severe respiratory pathogen. *Nat. Commun.* **6**, (2015).
51. Wong, D. *et al.* Primary Pneumonic Plague Contracted from a Mountain Lion Carcass. *Clin. Infect. Dis.* **49**, e33–e38 (2009).
52. Mathieson, I. *et al.* The Genomic History of Southeastern Europe. *Nature* **555**, 197–203 (2018).
53. Velsko, I., Skourtanioti, E. & Brandt, G. Ancient DNA Extraction from Skeletal Material. (2020) doi:10.17504/protocols.io.baksicwe.
54. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci.* **110**, 15758–15763 (2013).
55. Meyer, M. & Kircher, M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
56. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20130624 (2015).
57. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
58. Gansauge, M.-T., Aximu-Petri, A., Nagel, S. & Meyer, M. 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. *Nat. Protoc.* **15**, 2279–2300 (2020).

59. Fellows Yates, J. A. *et al.* Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *bioRxiv* 2020.06.11.145615 (2020) doi:10.1101/2020.06.11.145615.
60. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, (2016).
61. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/.
62. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/.
63. Andrews, S. *et al. FastQC*. (2015).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
65. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/index.html.
67. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
68. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
69. Neukamm, J. & Peltzer, A. *Integrative-Transcriptomics/DamageProfiler: DamageProfiler v0.4.9*. (Zenodo, 2019). doi:10.5281/zenodo.3557708.
70. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
71. Eppinger, M. *et al.* The complete genome sequence of Yersinia pseudotuberculosis IP31758, the causative agent of Far East scarlet-like fever. *PLoS Genet.* **3**, e142 (2007).
72. Garcia, E. *et al.* Pestoides F, an Atypical Yersinia pestis Strain from the Former Soviet Union. *Genus Yersinia* 17–22 (2007) doi:10.1007/978-0-387-72124-8_2.

73. Song, Y. *et al.* Complete Genome Sequence of Yersinia pestis Strain 91001, an Isolate Avirulent to Humans. *DNA Res.* **11**, 179–197 (2004).

74. Cui, Y. *et al.* Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc. Natl. Acad. Sci.* **110**, 577–582 (2013).

75. Deng, W. *et al.* Genome Sequence of Yersinia pestis KIM. *J. Bacteriol.* **184**, 4601–4611 (2002).

76. Chain, P. S. G. *et al.* Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13826–13831 (2004).

77. Kislichkina, A. A. *et al.* Nineteen Whole-Genome Assemblies of Yersinia pestis subsp. microtus, Including Representatives of Biovars caucasica, talassica, hissarica, altaica, xilingolensis, and ulegeica. *Genome Announc.* **3**, (2015).

78. Kislichkina, A. A. *et al.* Nine Whole-Genome Assemblies of Yersinia pestis subsp. microtus bv. Altaica Strains Isolated from the Altai Mountain Natural Plague Focus (No. 36) in Russia. *Genome Announc.* **6**, (2018).

79. Zhgenti, E. *et al.* Genome Assemblies for 11 Yersinia pestis Strains Isolated in the Caucasus Region. *Genome Announc.* **3**, e01030-15 (2015).

80. Kislichkina, A. A. *et al.* Eight Whole-Genome Assemblies of Yersinia pestis subsp. microtus bv. caucasica Isolated from the Common Vole (Microtus arvalis) Plague Focus in Dagestan, Russia. *Genome Announc* **5**, e00847-17 (2017).

81. Eroshenko, G. A. *et al.* Yersinia pestis strains of ancient phylogenetic branch 0.ANT are widely spread in the high-mountain plague foci of Kyrgyzstan. *PLOS ONE* **12**, e0187230 (2017).

82. Kutyrev, V. V. *et al.* Phylogeny and Classification of Yersinia pestis Through the Lens of Strains From the Plague Foci of Commonwealth of Independent States. *Front. Microbiol.* **9**, (2018).

83. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).

84. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

85. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

86. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008).

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

88. Rudis, B., Bolker, B. & Schulz, J. *ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'*. (2017).

89. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

90. Hijmans, R. J. *geosphere: Spherical Trigonometry*. (2019).

91. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

92. Mittnik, A. *et al.* Kinship-based social inequality in Bronze Age Europe. *Science* **366**, 731–734 (2019).

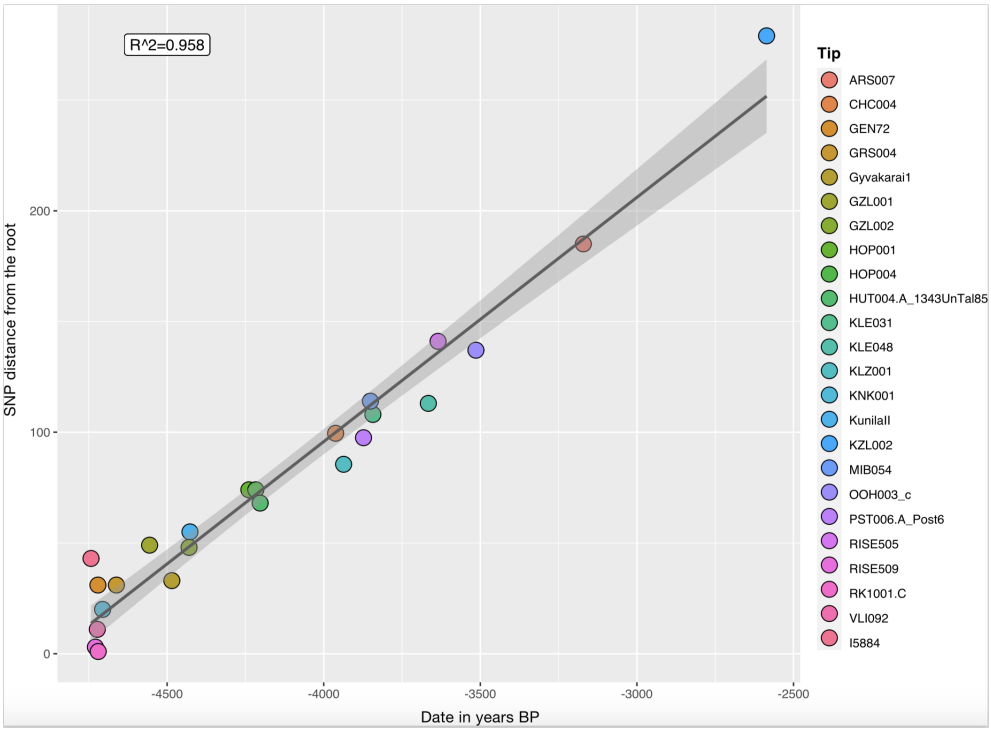
93. Jeong, C. *et al.* Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. *Proc. Natl. Acad. Sci.* **115**, E11248 (2018).

94. Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. (2020).

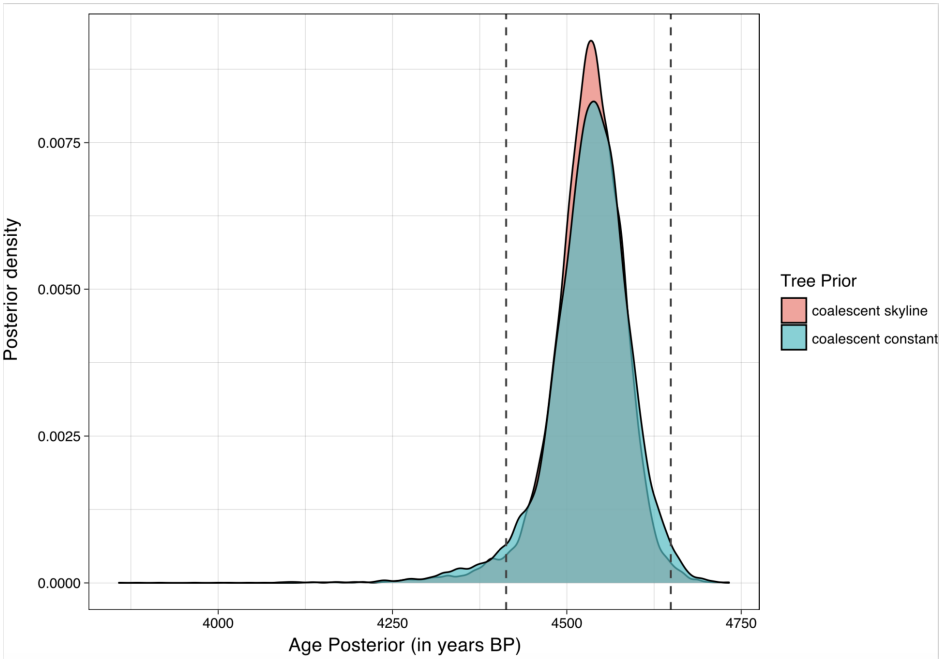
Table 1: Metadata and summary statistics for the *Y. pestis* chromosome reconstruction. * new data for the previously published genome RISE139, which is referred to as CHC004 in this study. calBP= calibrated Before Present

Sample name	Arch ID	Site name	Country	C14 ID	C14 dating (2σ calBP)	Mean Coverage	Coverage (%) >3X
I5884	I5884/Mos44, Grave 68, 294	Dereivka I	Ukraine	PSUAMS-2828	4840-4646	11	88
VLI092	P7A 41821	Vlineves	Czech Republic	MAMS-45801	4832-4619	17.6	92.5
KNK001	BZNK-1016/1, kurgan 1, grave 8	Krasnogvardeyskoe	Russia	MAMS-45959	4827-4585	2.7	46.5
GRS004	PEG-07/499	Großstorkwitz	Germany	Hd-21977	4800-4524	15.6	91.4
HOP001	Obj. 688 / Gr. 17	Hostivice-Palouky	Czech Republic	MAMS-30798	4405-4152	19	91.8
HOP004	Obj. 691 / Gr. 22	Hostivice-Palouky	Czech Republic	MAMS-38921	4400-4100	25	92.4
KLZ001	BZNK-1029/1, kurgan 1, grave 8	Kaluzhny 1	Russia	MAMS-45952	4243-4091	3.6	60
CHC004*	20	Chocivel	Poland	Ua-44034	4085-3873	4.4	71.5
KLE031	Bef.77A	Kleinaitingen	Germany	MAMS-21584	3959-3720	13.8	87.9
MIB054	42 (2040) NM Prague 43141	Mikulovice (big)	Czech Republic	MAMS-30479	3899-3730	30.6	92.8
KLE048	Bef. 120	Kleinaitingen	Germany	MAMS-21595	3818-3586	19.9	90.3
OOH003	Bef. 84	Oberottmarshausen	Germany	MAMS-21543	3586-3447	17.4	91.4
I2470	I2470/ES.3/4-1	El Sotillo	Spain	Beta-299307	3361-3181	11.5	87.3
ARS007	2007-9	Arbulag sum	Mongolia	No Reported	3288-3006	2.5	37.6
KZL002	Kyzyl (mound 4, right)	Kyzyl	Kazakhstan	UBA-25474	2736-2457	17	88.8

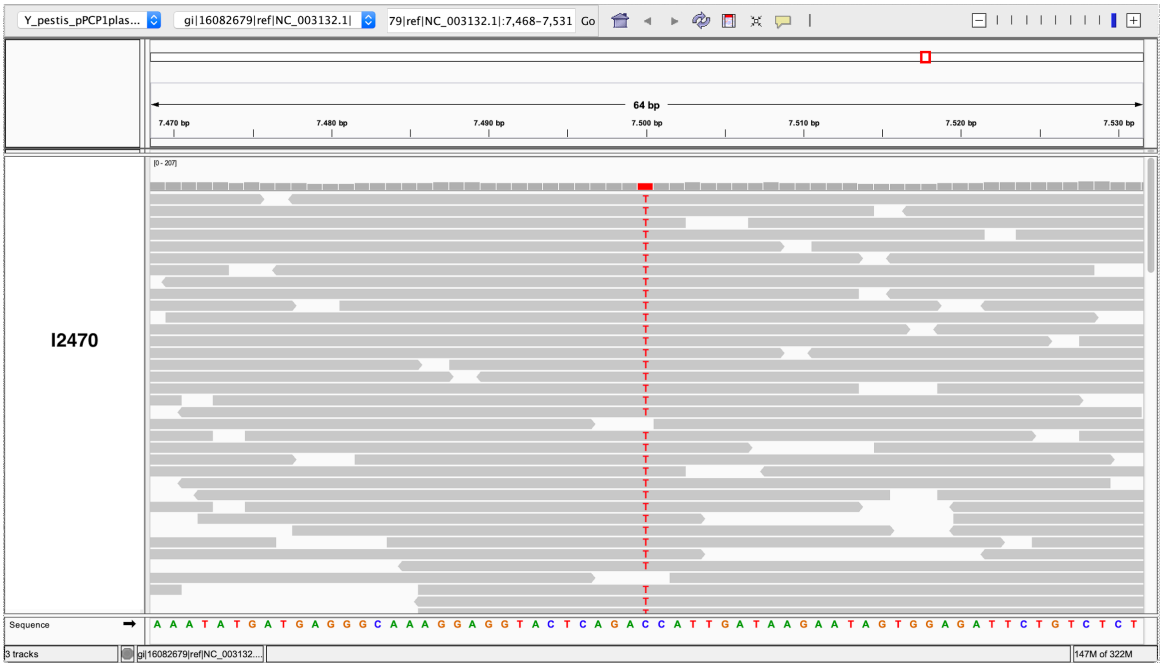
Supplementary Information



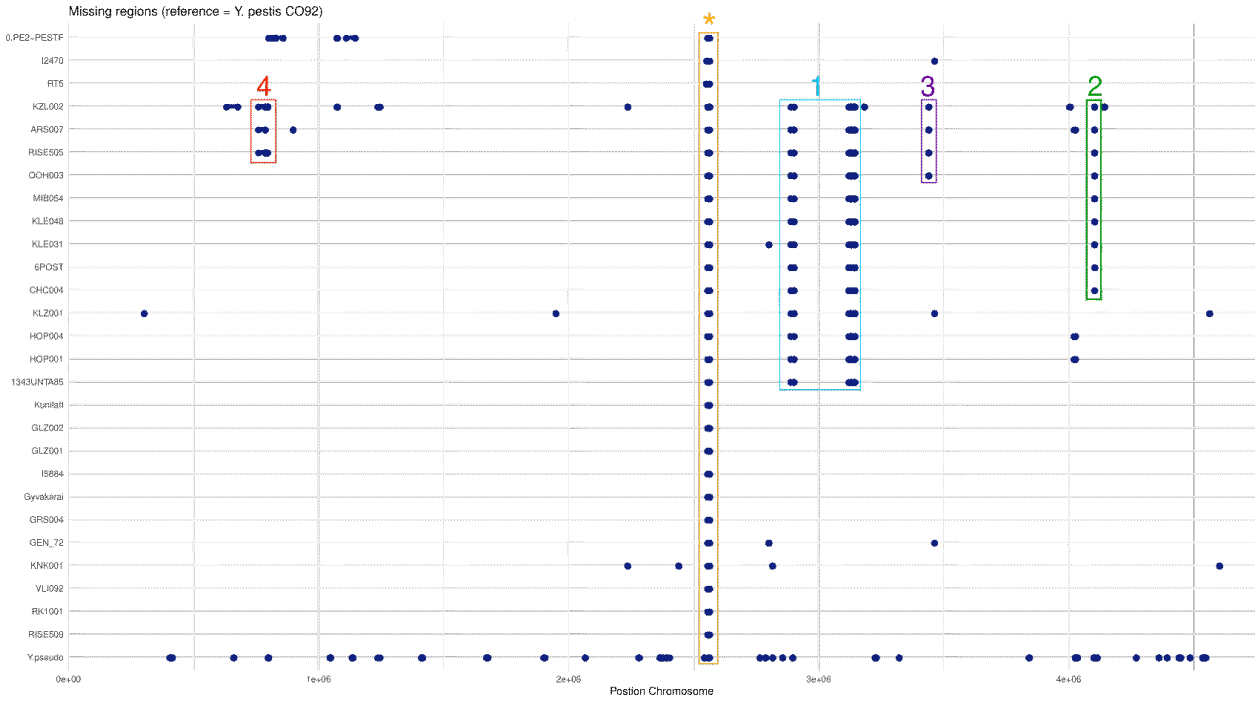
Supplementary Figure 1: tempest analysis for the root to tip regression in the LNBA lineage.



Supplementary Figure 2: distribution of the posterior probabilities of the inferred date for I5884 by BEAST. The colours indicate the demographic model used, red being coalescent skyline and blue coalescent constant.

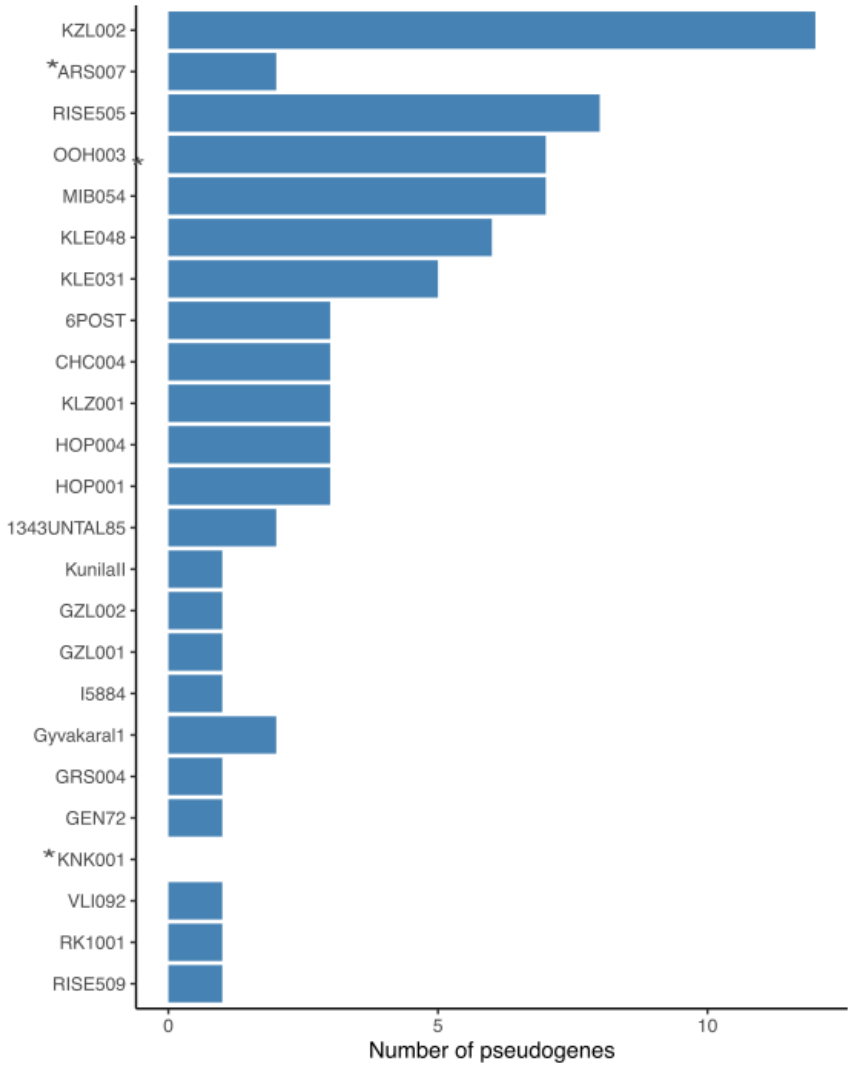


Supplementary Figure 3: Manual evaluation of the *pla* gene for the I2470 genome in the IGV genome browser[91]. We observe the 'T' variant in the position 7500 of the pPCP1 plasmid, which corresponds to the ancestral variant that results in the isoleucine amino acid in the position 259 of *pla*[50]



Supplementary figure 4: Dumbbell plot showing missing regions of at least 500bp detected in this study's dataset when using *Y. pestis* CO92 as reference. The numbered squares indicate the deletions in the LNBA lineage, and the numbers correspond to the ones displayed in Figure

1, and Supplementary Table 2. * represents the filamentous prophage YpfΦ, which is only consistently integrated in the chromosome of 1.ORI strains.



Supplementary Figure 5: Count plot of the pseudogenes present in the genomes in the LNBA branch. * Genome containing missing data for these positions due to low coverage.

Supplementary Table 1: Overview of archaeological sites and number of individuals per site screened in the course of this study

Site	Country	Number of Screened Individuals	Publication
Dereivka I	Ukraine	20	[52]
Vlineves	Czech Republic	50	This study
Krasnogvardeyskoe	Russia	6	This study
Großstorkwitz	Germany	4	This study
Hostivice-Palouky	Czech Republic	4	This study
Kaluzhny 1	Russia	1	This study
Chocivel	Poland	6	[11,38], this study*
Kleinaitingen	Germany	28	[92]
Mikulovice - big	Russia	70	This study
Oberottmarshausen	Germany	20	[92]
Dolmen “El Sotillo”	Spain	2	[57]
Arbulag sum, Khövsgöl	Mongolia	11	[93]
Kyzyl	Kazakhstan	5	This study
Total		227	

*New data generated, including RISE139.

Supplementary Table 2: Coordinates on the *Y. pestis* CO92 genome and total size of the missing regions in the LNBA lineage.

Event Number	Start	End	Size (bp)
1	2,887,052	2,900,727	13,675
	3,118,516	3,128,505	9,989
	3,142,367	3144351	10,736
	3,142,367	3,144,351	1,984
2	4,100,966	4,102,551	1,585
3	3,437,796	3,439,880	2,084
4	758,138	785,317	27,179
	786,017	795,670	9,653

Supplementary Table 3: Genomes used in the genetic versus time and genetic versus geography correlation analysis of *Y. pestis*, *S. enterica* and *M. leprae* and their corresponding publication

Name	**Species**	**Publication**
BED030	*Yersinia pestis*	7
BED028	*Yersinia pestis*	7
BED034	*Yersinia pestis*	7
BED024	*Yersinia pestis*	7
BRA001	*Yersinia pestis*	7
LAI009	*Yersinia pestis*	7
LBG002	*Yersinia pestis*	7
MAN008	*Yersinia pestis*	7
NAB003	*Yersinia pestis*	7
NMS002	*Yersinia pestis*	7
STA001	*Yersinia pestis*	7
STN014	*Yersinia pestis*	7

STN020	*Yersinia pestis*	7
STN021	*Yersinia pestis*	7
STN019	*Yersinia pestis*	7
STN007	*Yersinia pestis*	7
STN002	*Yersinia pestis*	7
STN008	*Yersinia pestis*	7
STN013	*Yersinia pestis*	7
London_BD	*Yersinia pestis*	3
ELW098	*Yersinia pestis*	6
Barcelona	*Yersinia pestis*	6
OBS137	*Yersinia pestis*	2
OBS116	*Yersinia pestis*	2
OBS107	*Yersinia pestis*	2
OBS110	*Yersinia pestis*	2
OBS124	*Yersinia pestis*	2
MUR009	*Salmonella enterica*	29
MUR019	*Salmonella enterica*	29
IV3002	*Salmonella enterica*	29
OBP001	*Salmonella enterica*	29
IKI003	*Salmonella enterica*	29
SUA004	*Salmonella enterica*	29
MK3001	*Salmonella enterica*	29
ETR001	*Salmonella enterica*	29
Ragna	*Salmonella enterica*	29
Tepos_14	*Salmonella enterica*	29

Tepos_35	*Salmonella enterica*	29
3077	*Mycobacterium leprae*	32
Jorgen_625	*Mycobacterium leprae*	32
Refshale_16	*Mycobacterium leprae*	32
SK2	*Mycobacterium leprae*	32
SK8	*Mycobacterium leprae*	32
Body 188	*Mycobacterium leprae*	33
GC96	*Mycobacterium leprae*	33
Jorgen_404	*Mycobacterium leprae*	33
Jorgen_427	*Mycobacterium leprae*	33
Jorgen_507	*Mycobacterium leprae*	33
Jorgen_533	*Mycobacterium leprae*	33
Jorgen_722	*Mycobacterium leprae*	33
Jorgen_749	*Mycobacterium leprae*	33
SK11	*Mycobacterium leprae*	33
T18	*Mycobacterium leprae*	33
SK14	*Mycobacterium leprae*	34
SK27	*Mycobacterium leprae*	34

Manuscript C

A pangenome of the *Yersinia pseudotuberculosis* complex

Aida Andrades Valtueña[1], Alexander Herbig[1]
[1] Max Planck Institute for the Science of Human History, Jena, Germany

Abstract

Pangenomics is an emerging tool to understand the genomic evolution in bacteria. A large focus in the study of the emergence of *Yersinia pestis* from *Yersinia pseudotuberculosis* has been on differentiating these species based on phenotypic traits, or the presence of known virulence genes. However, it is known that the acquisition of plasmids, gene loss and pseudogenisation in the *Y. pseudotuberculosis* complex may have played a role in the speciation of *Y. pestis*. Despite this, the pangenome of this complex of taxa remains underexplored. Here we present a workflow to explore the pangenome of the *Y. pseudotuberculosis* complex, not only in terms of presence and absence but also loss of function of genes due to pseudogenisation. Our workflow allows the incorporation of ancient DNA data that can aid in understanding the timing and anthropological and ecological context in which those changes happened. The results together with the compilation of the most comprehensive metadata of the available genomes for the *Y. pseudotuberculosis* complex to date will allow the community to explore and understand in finer-scale the evolution of this complex, and the emergence of *Y. pestis* from its much less life-threatening relative *Y. pseudotuberculosis*.

Introduction

The study of the pangenome - all the genes present in a taxonomic unit - has been proposed as a new tool to gain insights into bacterial genomics and evolution (Rouli et al., 2015). Despite the evolution of the pathogenic species in the genus *Yersinia* being characterised by the loss (physical loss or by pseudogenisation) and gain of genes (McNally et al., 2016), their pangenome remains largely unexplored. The *Yersinia* genus contains 19 species, the majority of which are environmental and non-pathogenic to mammals, with the exception of 3 species that can be pathogenic to mammals and humans: *Yersinia enterocolitica*, *Yersinia pseudotuberculosis* and *Yersinia pestis* (McNally et al., 2016). These pathogenic species share a common plasmid, known as pYV or pCD1, which has been acquired independently, thus showing the importance of parallel evolution in the emergence of these pathogens (Reuter et al., 2014). The disease manifestation and outcome varies between these species: while *Y. enterocolitica* and *Y. pseudotuberculosis* cause a self-limiting enteric disease, *Y. pestis* is responsible for the infamous bubonic plague (Stenseth et al., 2008). Furthermore, while *Y. enterocolitica* is not closely related to the other human pathogens, it has been shown that *Y. pestis* likely evolved from *Y. pseudotuberculosis*

(Achtman et al., 1999). *Y. pseudotuberculosis* and *Y. pestis* together with *Yersinia similis* form what is known as the *Y. pseudotuberculosis* complex (McNally et al., 2016). By comparing *Y. pestis* and *Y. pseudotuberculosis* genomes, it has been shown that *Y. pestis* emerged from *Y. pseudotuberculosis* via the acquisition and loss of genetic elements (Hinnebusch et al., 2016). *Y. pestis* differs from its ancestor by causing a more invasive and deadly disease and having a more efficient transmission via a flea vector. The acquisition of the *ymt* gene, which protects the bacteria from being digested in the midgut of the flea (Hinnebusch et al., 2002), in combination with the silencing of both biofilm regulators (Sun et al., 2008) and *ureD* (Sebbane et al., 2001) allowed for the adaptation of *Y. pestis* to the flea vector. The acquisition of the pPCP1 plasmid, which encodes the *pla* gene that is involved in dissemination within the mammalian host (Haiko et al., 2009; Sebbane et al., 2006), allows *Y. pestis* to cause a systemic disease. Previous attempts at exploring the pangenome of *Y. pseudotuberculosis* and *Y. pestis* has led to a finer-scale differentiation and the identification of genes unique to each of these species (Califf et al., 2015). However, previous pangenome studies have not incorporated ancient *Y. pestis* genomes, which could provide a more precise timing and nuanced understanding into the development of *Y. pestis* from *Y. pseudotuberculosis*. In particular, genomes recovered from individuals dating to 5,000-3,000 years ago (the Neolithic and Bronze Age period) have shown that *Y. pestis* did not have a genomic background necessary for the flea transmission (Andrades Valtueña et al., 2017; Rascovan et al., 2019; Rasmussen et al., 2015), and that this adaptation occurred shortly after the appearance of those lineages with the earliest evidence recovered from a 3,800 year old individual from Russia (Spyrou et al., 2018). Furthermore, we can contextualise the genomic findings with (pre-)historic evidence, thus allowing to correlate changes in the bacterial genome with changes in human behaviour or climate. In this study, we develop a workflow to compute a pangenome and incorporate ancient data into the analysis to explore both change in the presence and absence of genes as well as infer pseudogenisation of genes, which can inform us about the loss of function of a gene even if present. We applied the workflow to the largest *Y. pseudotuberculosis* complex dataset collected to date to gain insights into the genomic and pseudogenisation history of the emergence of the plague pathogen.

Methods

Dataset

We collected DNA sequencing data from published ancient and modern genomes of *Y. pestis*, *Y. pseudotuberculosis*, and *Y. similis* (Supplementary Table 1). This resulted in a comprehensive dataset containing: 3 *Y. similis* genomes, 71 *Y. pseudotuberculosis* genomes, 822 modern *Y. pestis* genomes and 69 ancient *Y. pestis* genomes spanning different time periods (49 from the second plague pandemic (650-284 years Before Present (yBP), Bos et al., 2016, 2011; Namouchi et al., 2018; Spyrou et al., 2019, 2016), 8 from the plague of Justinian (1,500-1,300 yBP, (Feldman et al., 2016; Keller et al., 2019), 2 genomes dating to 1,700 yBP and 1,200 yBP (Damgaard et al., 2018) and 9 from prehistoric individuals (5,000-3,500 yBP, Andrades Valtueña et al., 2017; Rascovan et al., 2019; Rasmussen et al., 2015). This dataset is referred to as “full dataset” in the following sections.

Yersinia pseudotuberculosis complex pangenome analysis

An overview of the workflow for pangenome generation and analysis can be found in Figure 1A. In brief, we compute a pangenome using panX (Ding et al., 2018) and use the consensus genes generated by panX as a reference for mapping the full dataset. We then generate a file containing the percentage of coverage for each gene and genome that would be then used for downstream presence-absence analysis in R.

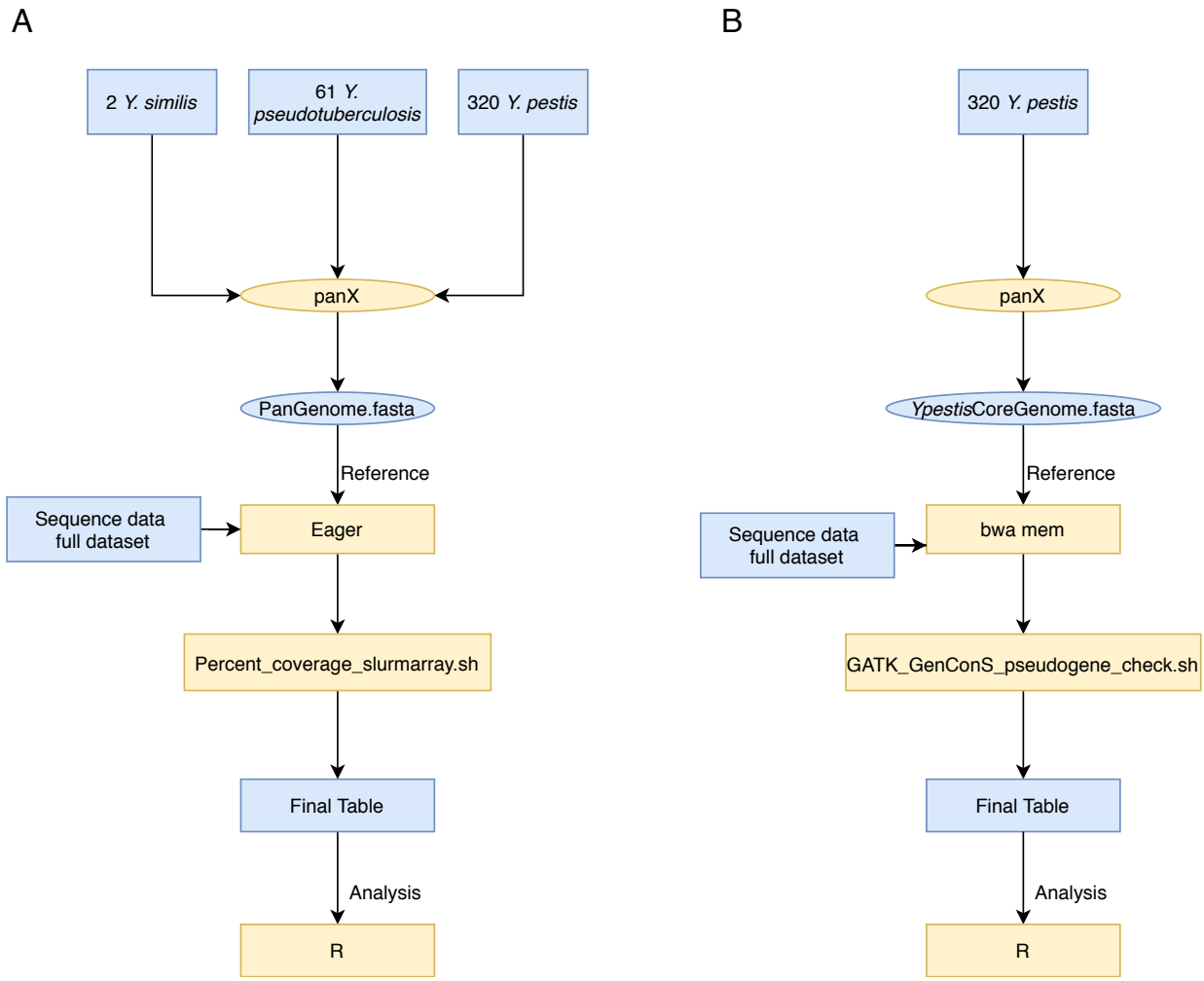


Figure 1: General overview of the presented workflows: (A) *Yersinia pseudotuberculosis* complex pangenome analysis workflow; (B) *Yersinia pestis* core genome analysis

In order to compute the pangenome for the *Y. pseudotuberculosis* complex we downloaded genome annotation files in GenBank (.gbk) format from the NCBI RefSeq database (O'Leary et al. 2016): 2 *Y. similis* (prior to 2020 these were classified as *Y. pseudotuberculosis* in the NCBI RefSeq database), 61 *Y. pseudotuberculosis* and 320 *Y. pestis* modern genomes (Supplementary Table 2). This subset was used for the pangenome computation because panX requires properly formatted GenBank files as input, which limited the analysis to well-annotated genomes. To incorporate ancient genomes and modern genomes where only the raw sequencing-read data is

available, or modern genomes that either lack or have a badly formatted annotation files in the NCBI RefSeq database, an additional mapping step is necessary. The subset of well-annotated genomes (Supplementary Table 2) were used as input to panX (v1.5.1, Ding et al., 2018) which was run with default parameters to generate a pangenome, except with -cg set to 0.8, to consider genes present in 80% of the strains as core genes:

```
panX.py -fn /path/to/folder -sl name_run -t 32 -cg 0.8 -dmdc -dcs 50 -sitr
```

In addition to the visualisation files, panX produces DNA and protein consensus files for each of the genes in the pangenome. To include the full dataset in the analysis, we extracted the DNA consensus for each gene in the pangenome from the panX output and concatenated them in a multi-fasta file, which was used as a reference to map the full dataset with the script Creating_fasta_from_Pangenome.sh (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scripts/Creating_fasta_from_Pangenome.sh). We observed that panX introduces gaps during consensus building to account for sites that do not have enough support by the genomes used in the pangenome building step. To increase the mappability of genomic sequencing reads against the gene consensus, we removed the gaps from the consensus fasta for each gene, with the assumption that reads will still map to the consensus even if a longer version of the gene is present. Prior to gap removal, we calculated the total length of the gaps and the percentage of the consensus of that particular gene that consisted of gaps. The gap removal and gap statistics were performed by the script Gap_removal_stats.sh (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scripts/Gap_removal_stats.sh).

For the ancient samples, we observed 3 different library-construction protocols that may affect mapping of the raw reads to the constructed pangenome: non-UDG, half-UDG and full-UDG. In non-UDG library construction there is no enzymatic treatment of the sample to remove uracils resulting from deaminated cytosines due to hydrolytic damage that accumulate overtime, typically observed in ancient DNA (Briggs et al 2007). These misread cytosines can result in an increase in mismatches from the original sequence, which makes mapping of authentic ancient reads difficult. For half-UDG (see Rohland et al., 2015) and full-UDG (published in (Briggs and Heyn, 2012) protocols, the samples have undergone a partial or complete removal of deaminated cytosines (uracil), respectively, during the library construction. We need to adapt the mapping parameters used depending on the library protocol used in order to take account the damage presence that can lead to false genotyping. The treatment of the samples can be found in Supplementary Table 1. For modern genomes where only assemblies were available, we generated reads from the fasta file computationally with an in-house script (Genome2Reads) that cuts the sequence in 100bp reads with a 1bp tiling from each other.

The data of all *Y. pestis* and *Y. pseudotuberculosis* was mapped to the pangenome multi-fasta reference using EAGER (v1.92.55, Peltzer et al., 2016) in the following manner: adapters were clipped from the reads; pair-end reads were merged and reads shorter than 30 base pairs and/or a quality lower than 20 were removed with AdapterRemoval (v2.2.0, Schubert, Lindgreen, and Orlando 2016). An additional step was performed for half-UDG treated data, where we clipped an additional 2 bases from each side to remove the typical deamination damage present in ancient DNA samples and were therefore treated downstream as full-UDG treated samples. We aligned

the reads using bwa aln (v.0.7.12, Li and Durbin 2009) with seed length (-l) 32 and mismatch parameter (-n) 0.1 (UDG-parameters) for the modern, clipped half-UDG and UDG treated data; and set -l 16 and -n 0.01 (non-UDG parameters) for non-UDG treated samples, which, since it is a more sensitive setting, will allow for the mapping of damaged reads containing mismatches. Finally we removed PCR duplicates using DeDup (v.0.12.2, Peltzer et al., 2016).

In order to estimate the presence or absence of genes in the genomes, we calculated the percentage of the gene covered using bedtools (v. 2.25.0, Quinlan and Hall, 2010) genomecov and coverage commands. The used script Percent_coverage_slurmarray.sh can be found in: https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scripts/Percent_coverage_slurmarray.sh.

In order to decide on the best aligning tools to detect presence and absence of genes in the pangenome, we mapped the modern samples used in the pangenome construction to the pangenome multi-fasta consensus with bwa mem (default parameters, Li, 2013) and bwa aln (see above). Bwa aln is better for short reads whereas bwa mem is designed for longer reads and allows for split mapping of reads, and we wanted to test which of these aligning tools will perform better in recovering the pangenome representation computed by panX. We then compared the recovery rate (prediction of the real state (present/absent) of the gene) between the bwa aln and mem algorithms. Based on this comparison, we filtered out genes with more than 50% gaps because of mappability issues and genes that contained the same WP code (an identifier for the NCBI identical protein database), since these genes probably represent a sub-optimal clustering by panX. We obtained the best recovery rate with bwa aln (78.69% vs. 75.56% mem). Based on the bwa aln results and the plotting of the percentage of the gene covered (Supplementary Figure 2), we decided to set the threshold for presence/absence to 85% of the gene covered to be considered present and 50% or less for it to be absent. Genes with intermediate values (51-84%) were labelled as unknowns. We excluded 14 ancient genomes that had information for less than 20% of the genes, and we removed genes that had either no data (meaning only absent or 'unknown' states) or with more than 20% of the genomes with unknown status for the gene. After applying all the filters, we retained a total of 923 genomes (indicated in Supplementary Table 1) and a set of 14,610 genes in the analysis.
We classified the genes in the following categories: core (present in more than 90% of the genomes), only *Y. similis*, only *Y. pseudotuberculosis*, only *Y. pestis*, and other. For species specific genes, we allowed only 5% of 'unknown' states in the other two species. The other category includes any genes that didn't fulfil any of the previous described criteria.
All the previous described steps were performed in R, the code can be found in: https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/PresenceAbsence/presence_absence_notebook_v4.Rmd.

To recover the function of the genes, we obtained Gene Ontology (GO) terms (Ashburner et al. 2000; The Gene Ontology Consortium 2019) for all the genes in the pangenome. This was done by extracting the WP code for each gene from the GenBank files used in pangenome generation with the script extractingpanXandWPfromGbk.py (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scr

ipts/extractingpanXandWPfromGbk.py). Since there is no *Y. pestis* GO database, we used the WP code to obtain a protein consensus fasta for each gene using NCBI entrez (Kans 2020), with the script WP_to_fasta.sh (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scripts/WP_to_fasta.sh). Finally, GO terms were obtained using InterProtScan5 (v. 5.45-80.0, Jones et al., 2014) with pangenome protein multi-fasta as an input (-i PanGenome.fasta):

```
interproscan-5.45-80.0/interproscan.sh -cpu 4 -d $OUTDIR -i PanGenome.fasta -f tsv -goterms
```

where $OUTDIR is the path to the output directory.

Yersinia pestis core genome analysis

The complete workflow for the core pseudogene analysis can be found in Figure 1B. A core genome is computed with panX, and the DNA consensus sequences of the core genes are then used as reference to map the full dataset, and the resulting gene consensus for each genome are checked for stop codons, frameshift, and amino acid changes. A detailed description of the workflow is as follows: To explore the amino acid changes and the pseudogenisation events in the evolution of the core genes of *Y. pestis*, we computed the core genome of *Y. pestis* using panX as described above, however the -cg parameter was set to 0.9:

```
panX.py -fn /path/to/folder -sl name_run -t 32 -cg 0.9 -dmdc -dcs 50 -sitr
```

As input, we used the same 320 *Y. pestis* genomes that were also used in the pangenome generation. For details on the strains, see Supplementary Table 2. We extracted the consensus sequences for the core genes from the panX output and concatenated them into a multi-fasta file using Creating_core_fasta.sh (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/new/master/General_scripts/Creating_core_fasta.sh). This multi-fasta file was used as a reference for sequence read alignment. The already processed fastq files with AdapterRemoval were mapped to the core genes reference with bwa mem (Li, 2013) with default parameters, and duplicates removed with DeDup.

To obtain a consensus for each gene and each genome, we called SNPs and INDELS for each individual gene using GATK UnifiedGenotyper with the mode EMIT_ALL_SITES and generate a consensus using GenConS (Fellows Yates et al., 2017) with the following parameters: a minimum of 3 reads must be covering a position and 90% of the reads must support an allele to be called, and 0.8 was used for the punishment_ratio for C->T and G->A changes in order to account for potential damage signal. When a position did not fulfil either of those criteria, no call was made by GenConS and a N was included in the reference to reflect missing data. Finally, we determined the pseudogenisation status for each gene. All previous steps where performed using the script GATK_GenConS_pseudogene_check.sh (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/General_scripts/GATK_GenConS_pseudogene_check.sh), which is optimised to run in a SLURM scheduler-based High-Performance-Cluster environment. This script checks for the presence of premature stops, change in the coding frame, and change in the amino acid sequence. To know the function

of the genes with pseudogenisation events, we again obtained GO terms using the same procedure as described in the *Yersinia pseudotuberculosis* complex pangenome analysis section. The results were visualised with R (https://github.com/aidaanva/PanGenomeComparativeAnalysisYpestis/blob/master/Pseudogene/pseudogene_analysis.Rmd).

Results

Pangenome analysis

To understand the distribution of genes in the *Y. pseudotuberculosis* complex, we computed a pangenome using panX (Ding et al., 2018) including 2 *Y. similis*, 61 *Y. pseudotuberculosis* genomes and 320 *Y. pestis* modern genomes from the RefSeq. 15/61 *Y. pseudotuberculosis* and 39/320 *Y. pestis* genomes represented resequencing of the same strain. The resulting pangenome comprises 15,631 genes which were classified by panX as: 3,181 core genes (present in at least 80% of the genomes) and 12,450 accessory genes.

In order to include in the pangenome analysis genomes not utilised during the pangenome calculation - such as ancient genomes where no assembly is available, modern genomes missing in the NCBI RefSeq database, or with a faulty annotation file - we mapped either the sequencing reads or the computationally generated reads (for assemblies of modern genomes) against a fasta file containing all genes of the computed pangenome (see methods). We observed that some genes had a considerable percentage of the panX generated consensus containing gaps (Supplementary Figure 1). We decided to exclude 137 genes that either had a length shorter than 100 bp (70 genes) or contained more than 50% gaps in the consensus (66 genes) since this could affect their mappability, and led to an incorrect assignment of their presence status. Additionally, we excluded 123 genes with identical protein group ID (WP NCBI code), indicating that they should have been clustered together during the reconstruction of the pangenome and affecting their mappability.

In order to set up a cut off for presence status (present or absent) for a given gene, we plotted the percentage of genes covered for all the 15,372 genes (Supplementary Figure 2). Based on the distribution, genes covered 50% or less were considered absent and genes covered 85% or more were considered present. Genes with an intermediate percent coverage between 51%-84% are considered as unknown status. Based on this classification we excluded genes with no data left (26 genes with only absent or unknown states) and genes that contained more that 20% of unknown states (731 genes). The final gene set included 14,610 genes which were used to compare the presence and absence profiles in 923 publicly available genomes (Supplementary Table 1). We excluded genomes that had information for less than 20% of the genes (>2,922 genes), which included 11 ancient genomes. The final genome dataset included 912 genomes distributed as follows: 56 ancient *Y. pestis*, 799 modern *Y. pestis*, 54 *Y. pseudotuberculosis* and 3 *Y. similis* genomes.

Pseudogene analysis

In order to check the pseudogenisation profile of *Y. pestis*, we generated a core genome using panX for the 320 *Y. pestis* genomes and assess the state in the complete dataset (959 genomes). The core genome of *Y. pestis* comprised 3,294 genes. We detected amino acid changes affecting 3,203 genes, however inference of the actual effect of those changes is highly dependent on the amino acid substitution. Further development of the implemented tool could allow to check if the amino acids are from different groups and help to predict effects in the folding of the protein itself. Changes in the protein that lead to a premature or shifting of the frame are more likely to cause the pseudogenisation of this gene. We detect a total of 171,011 pseudogenisations events affecting 2,930 of the core genes. The distribution of the number of genes affected by each pseudogenisation category can be seen in Figure 3: 625 genes with premature stop codon, 2,822 genes with frameshift and premature stop codon, 470 with frameshift and amino acid change and 1 with frameshift and amino acid change.

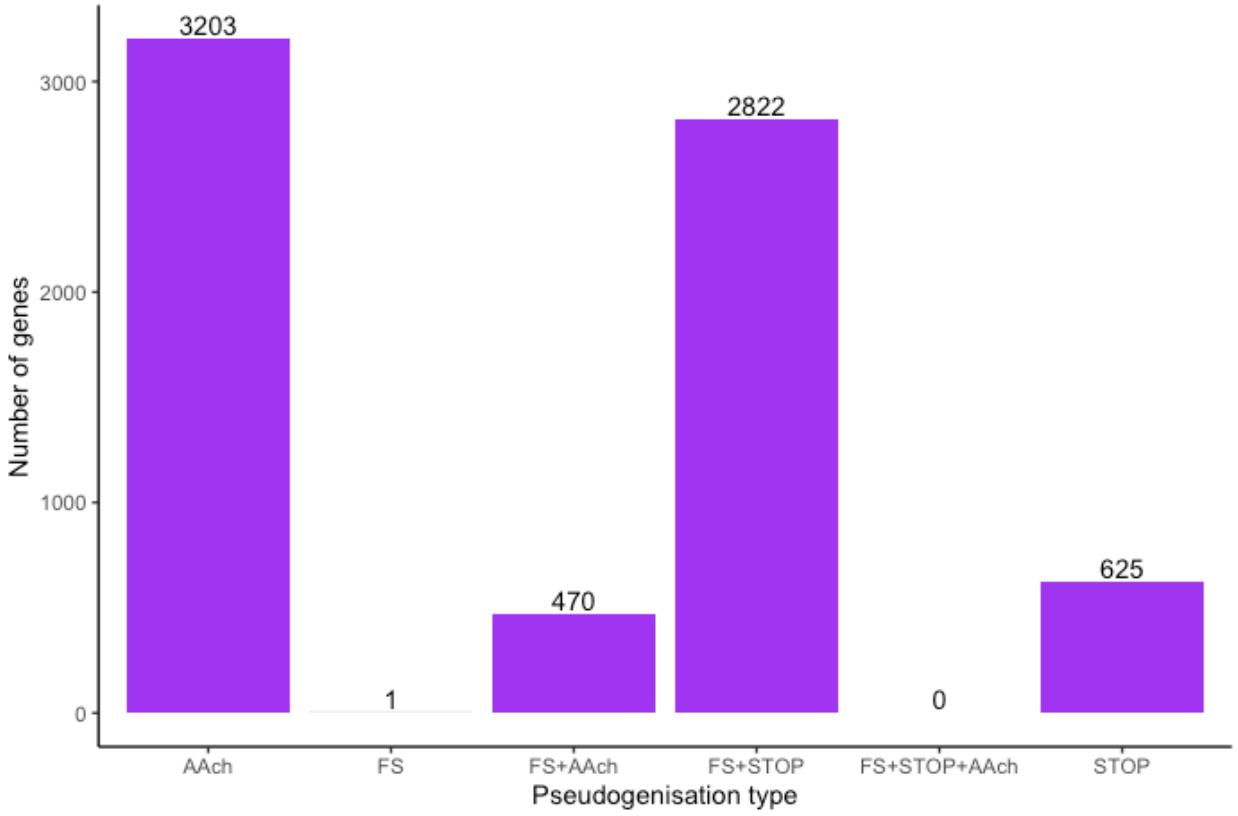


Figure 2: Summary plot of number of genes affected by pseudogenisations. The following pseudogenisation types are reported: Amino Acid Change (AAch), FrameShift (FS), Amino Acid Change and FrameShift (FS+Aach), FrameShift and premature STOP (FS+STOP), FrameShift and premature STOP and Amino Acid Change (FS + STOP + Aach), and premature STOP (STOP).

Discussion

The pangenome has been suggested as a new tool to gain insights not only in the gene content, but also to use it for taxonomic classification or lifestyle inference of the bacteria (Rouli et al., 2015). A recent study has analysed the pangenome of *Y. pseudotuberculosis* and *Y. pestis* and shown that *Y. pseudotuberculosis* has a more open pangenome than *Y. pestis,* and that there are genes outside classic virulence factors that differentiate these two species (Califf et al., 2015). However, no ancient *Y. pestis* genome has been included in this analysis. Key et al., (2020) showed the value of including ancient genomes into the exploring pangenomes of microbial taxa, which led to the hypothesis that pseudogenisation increases in the genus *Salmonella,* the more specialised the species becomes, and linked it to Neolithization - development and incorporation of farming practises - in Europe. However, this study based their analysis on an already existing pangenome for *Salmonella*. The last pangenome computed for *Y. pseudotuberculosis* and *Y. pestis* was in 2015, and included only 113 *Y. pestis* and 13 *Y. pseudotuberculosis* genomes (Califf et al., 2015). Since then, the number of genomes present in the NCBI and ENA (Leinonen et al. 2011) databases for these species has increased to 71 *Y. pseudotuberculosis* genomes and 822 *Y. pestis* genomes. Furthermore, there is now 3 genomes available for *Y. similis*; a species in the complex that was not taken into account in the previously published pangenome. In order to produce an updated pangenome, we present a new workflow (Figure 1) that allows for the exploration not only in terms of gene content but also the pseudogenisation events in the *Y. pseudotuberculosis* complex, together with the incorporation of ancient genomes. The incorporation of ancient genomes in the analysis will add a deep-time scale, thus allowing the exploration of the variation in gene content and pseudogenisation rate throughout the long-term evolution of the species. Furthermore, the transects created by ancient genomes from the second, first pandemic as well as the very early lineages from the Late Neolithic and Bronze Age periods allows for the exploration of the evolution of past plague diversity that does not exist anymore, and the effects of large-scale pandemic-like events. These extinct lineages can provide information on genes that are essential for the virulence for earlier forms of *Y. pestis*, thus making them targets for future research to combatting plague.

The *Y. pseudotuberculosis* complex pangenome contained 14,610 genes and included 912 genomes. We also applied the new pipeline to detect pseudogenes to the core genome of *Y. pestis*. A total of 171,011 pseudogenisations were detected in the 3,294 genes that form the core genome of *Y. pestis* across the entire phylogenetic tree, which could help identify novel virulence genes or defining characteristics for the different lineages in this species. It has been shown that pseudogenisation has played an important role in the development of *Y. pestis* (Hinnebusch et al., 2016), thus understanding the pseudogenisation history of this species can provide more insights into its evolution. Using archaeological and historical contexts, we could for example identify parallel pseudogenisation events which could be linked to specific animal hosts or virulence potential.

In this study we have provided a resource for the *Yersinia* community to explore the changes in genomic content and pseudogenisation history over time via the largest (923 genomes) dataset to date. Although we showcased our workflow by applying it to the *Y. pseudotuberculosis*

complex, it could be easily applied to other species, such as *Mycobacterium tuberculosis* or *Mycobacterium leprae* where ancient genomes have also been recovered from ancient individuals (See for example Bos et al., 2014; Schuenemann et al., 2013). Future improvements could be the automation of the workflow by integration into a pipeline framework such as snakemake (Köster and Rahmann 2012) or Nextflow (Di Tommaso et al. 2017).

References

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., Carniel, E., 1999. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. 96, 14043–14048. https://doi.org/10.1073/pnas.96.24.14043

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. Curr. Biol. 27, 3683-3691.e8. https://doi.org/10.1016/j.cub.2017.10.025

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. 'Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium'. *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., Campbell, T.J., Majander, K., Wilbur, A.K., Guichon, R.A., Wolfe Steadman, D.L., Cook, D.C., Niemann, S., Behr, M.A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J.E., Gagneux, S., Stone, A.C., Krause, J., 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514, 494–497. https://doi.org/10.1038/nature13591

Bos, K.I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S.A., Klunk, J., Schuenemann, V.J., Poinar, D., Kuch, M., Golding, G.B., Dutour, O., Keim, P., Wagner, D.M., Holmes, E.C., Krause, J., Poinar, H.N., 2016. Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus. eLife 5, e12994. https://doi.org/10.7554/eLife.12994

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. https://doi.org/10.1038/nature10549

Briggs, AdrianW., Heyn, P., 2012. Preparation of Next-Generation Sequencing Libraries from Damaged DNA, in: Shapiro, B., Hofreiter, M. (Eds.), Ancient DNA, Methods in Molecular Biology. Humana Press, pp. 143–154. https://doi.org/10.1007/978-1-61779-516-9_18

Califf, K.J., Keim, P.S., Wagner, D.M., Sahl, J.W., 2015. Redefining the differences in gene content between *Yersinia pestis* and *Yersinia pseudotuberculosis* using large-scale comparative genomics. Microb. Genomics 1. https://doi.org/10.1099/mgen.0.000028

Damgaard, P. de B., Marchi, N., Rasmussen, S., Peyrot, M., Renaud, G., Korneliussen, T., Moreno-Mayar, J.V., Pedersen, M.W., Goldberg, A., Usmanova, E., Baimukhanov, N.,

Loman, V., Hedeager, L., Pedersen, A.G., Nielsen, K., Afanasiev, G., Akmatov, K., Aldashev, A., Alpaslan, A., Baimbetov, G., Bazaliiskii, V.I., Beisenov, A., Boldbaatar, B., Boldgiv, B., Dorzhu, C., Ellingvag, S., Erdenebaatar, D., Dajani, R., Dmitriev, E., Evdokimov, V., Frei, K.M., Gromov, A., Goryachev, A., Hakonarson, H., Hegay, T., Khachatryan, Z., Khaskhanov, R., Kitov, E., Kolbina, A., Kubatbek, T., Kukushkin, A., Kukushkin, I., Lau, N., Margaryan, A., Merkyte, I., Mertz, I.V., Mertz, V.K., Mijiddorj, E., Moiyesev, V., Mukhtarova, G., Nurmukhanbetov, B., Orozbekova, Z., Panyushkina, I., Pieta, K., Smrčka, V., Shevnina, I., Logvin, A., Sjögren, K.-G., Štolcová, T., Taravella, A.M., Tashbaeva, K., Tkachev, A., Tulegenov, T., Voyakin, D., Yepiskoposyan, L., Undrakhbold, S., Varfolomeev, V., Weber, A., Sayres, M.A.W., Kradin, N., Allentoft, M.E., Orlando, L., Nielsen, R., Sikora, M., Heyer, E., Kristiansen, K., Willerslev, E., 2018. 137 ancient human genomes from across the Eurasian steppes. Nature 557, 369. https://doi.org/10.1038/s41586-018-0094-2

Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. 'Nextflow Enables Reproducible Computational Workflows'. *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Ding, W., Baumdicker, F., Neher, R.A., 2018. panX: pan-genome analysis and exploration. Nucleic Acids Res. 46, e5. https://doi.org/10.1093/nar/gkx977

Feldman, M., Harbeck, M., Keller, M., Spyrou, M.A., Rott, A., Trautmann, B., Scholz, H.C., Päffgen, B., Peters, J., McCormick, M., Bos, K., Herbig, A., Krause, J., 2016. A high-coverage *Yersinia pestis* Genome from a 6th-century Justinianic Plague Victim. Mol. Biol. Evol. msw170. https://doi.org/10.1093/molbev/msw170

Fellows Yates, J.A., Drucker, D.G., Reiter, E., Heumos, S., Welker, F., Münzel, S.C., Wojtal, P., Lázničková-Galetová, M., Conard, N.J., Herbig, A., Bocherens, H., Krause, J., 2017. Central European Woolly Mammoth Population Dynamics: Insights from Late Pleistocene Mitochondrial Genomes. Sci. Rep. 7, 17714. https://doi.org/10.1038/s41598-017-17723-1

Haiko, J., Kukkonen, M., Ravantti, J.J., Westerlund-Wikström, B., Korhonen, T.K., 2009. The Single Substitution I259T, Conserved in the Plasminogen Activator Pla of Pandemic *Yersinia pestis* Branches, Enhances Fibrinolytic Activity. J. Bacteriol. 191, 4758–4766. https://doi.org/10.1128/JB.00489-09

Hinnebusch, B.J., Chouikha, I., Sun, Y.-C., 2016. Ecological Opportunity, Evolution, and the Emergence of Flea-borne Plague. Infect. Immun. IAI.00188-16. https://doi.org/10.1128/IAI.00188-16

Hinnebusch, B.J., Rudolph, A.E., Cherepanov, P., Dixon, J.E., Schwan, T.G., Forsberg, Å., 2002. Role of Yersinia Murine Toxin in Survival of *Yersinia pestis* in the Midgut of the Flea Vector. Science 296, 733–735. https://doi.org/10.1126/science.1069972

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Kans, Jonathan. 2020. *Entrez Direct: E-Utilities on the Unix Command Line*. *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK179288/.

Keller, M., Spyrou, M.A., Scheib, C.L., Neumann, G.U., Kröpelin, A., Haas-Gebhard, B., Päffgen, B., Haberstroh, J., Lacomba, A.R. i, Raynaud, C., Cessford, C., Durand, R., Stadler, P., Nägele, K., Bates, J.S., Trautmann, B., Inskip, S.A., Peters, J., Robb, J.E., Kivisild, T., Castex, D., McCormick, M., Bos, K.I., Harbeck, M., Herbig, A., Krause, J., 2019. Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). Proc. Natl. Acad. Sci. 116, 12363–12372. https://doi.org/10.1073/pnas.1820447116

Key, F.M., Posth, C., Esquivel-Gomez, L.R., Hübler, R., Spyrou, M.A., Neumann, G.U., Furtwängler, A., Sabin, S., Burri, M., Wissgott, A., Lankapalli, A.K., Vågene, Å.J., Meyer, M., Nagel, S., Tukhbatova, R., Khokhlov, A., Chizhevsky, A., Hansen, S., Belinsky, A.B., Kalmykov, A., Kantorovich, A.R., Maslov, V.E., Stockhammer, P.W., Vai, S., Zavattaro, M., Riga, A., Caramelli, D., Skeates, R., Beckett, J., Gradoli, M.G., Steuri, N., Hafner, A., Ramstein, M., Siebke, I., Lösch, S., Erdal, Y.S., Alikhan, N.-F., Zhou, Z., Achtman, M., Bos, K., Reinhold, S., Haak, W., Kühnert, D., Herbig, A., Krause, J., 2020. Emergence of human-adapted *Salmonella enterica* is linked to the Neolithization process. Nat. Ecol. Evol. 4, 324–333. https://doi.org/10.1038/s41559-020-1106-9

Köster, Johannes, and Sven Rahmann. 2012. 'Snakemake—a Scalable Bioinformatics Workflow Engine'. *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Li, Heng, and Richard Durbin. 2009. 'Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform'. *Bioinformatics (Oxford, England)* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Leinonen, Rasko, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, et al. 2011. 'The European Nucleotide Archive'. *Nucleic Acids Research* 39 (Database issue): D28–31. https://doi.org/10.1093/nar/gkq967.

McNally, A., Thomson, N.R., Reuter, S., Wren, B.W., 2016. "Add, stir and reduce": *Yersinia* spp. as model bacteria for pathogen evolution. Nat. Rev. Microbiol. 14, 177–190. https://doi.org/10.1038/nrmicro.2015.29

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. Proc. Natl. Acad. Sci. 115, E11790. https://doi.org/10.1073/pnas.1812865115

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. 'Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation'. *Nucleic Acids Research* 44 (D1): D733-745. https://doi.org/10.1093/nar/gkv1189.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rascovan, N., Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C., Rasmussen, S., 2019. Emergence and Spread of Basal Lineages of *Yersinia pestis* during the Neolithic Decline. Cell 176, 295-305.e10. https://doi.org/10.1016/j.cell.2018.11.005

Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M.O., Nielsen, H.B., Brunak, S., Avetisyan, P.,

Epimakhov, A., Khalyapin, M.V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., 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. https://doi.org/10.1016/j.cell.2015.10.009

Reuter, S., Connor, T.R., Barquist, L., Walker, D., Feltwell, T., Harris, S.R., Fookes, M., Hall, M.E., Petty, N.K., Fuchs, T.M., Corander, J., Dufour, M., Ringwood, T., Savin, C., Bouchier, C., Martin, L., Miettinen, M., Shubin, M., Riehm, J.M., Laukkanen-Ninios, R., Sihvonen, L.M., Siitonen, A., Skurnik, M., Falcão, J.P., Fukushima, H., Scholz, H.C., Prentice, M.B., Wren, B.W., Parkhill, J., Carniel, E., Achtman, M., McNally, A., Thomson, N.R., 2014. Parallel independent evolution of pathogenicity within the genus *Yersinia*. Proc. Natl. Acad. Sci. 111, 6768–6773. https://doi.org/10.1073/pnas.1317161111

Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., Reich, D., 2015. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 370, 20130624. https://doi.org/10.1098/rstb.2013.0624

Rouli, L., Merhej, V., Fournier, P.-E., Raoult, D., 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 7, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. 'AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging'. *BMC Research Notes* 9 (February). https://doi.org/10.1186/s13104-016-1900-2.

Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jäger, G., Bos, K.I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J.L., Kjellström, A., Wu, H., Stewart, G.R., Taylor, G.M., Bauer, P., Lee, O.Y.-C., Wu, H.H.T., Minnikin, D.E., Besra, G.S., Tucker, K., Roffey, S., Sow, S.O., Cole, S.T., Nieselt, K., Krause, J., 2013. Genome-Wide Comparison of Medieval and Modern *Mycobacterium leprae*. Science 341, 179–183. https://doi.org/10.1126/science.1238286

Sebbane, F., Devalckenaere, A., Foulon, J., Carniel, E., Simonet, M., 2001. Silencing and Reactivation of Urease in *Yersinia pestis* Is Determined by One G Residue at a Specific Position in the *ureD* Gene. Infect. Immun. 69, 170–176. https://doi.org/10.1128/IAI.69.1.170-176.2001

Sebbane, F., Jarrett, C.O., Gardner, D., Long, D., Hinnebusch, B.J., 2006. Role of the *Yersinia pestis* plasminogen activator in the incidence of distinct septicemic and bubonic forms of flea-borne plague. Proc. Natl. Acad. Sci. U. S. A. 103, 5526–5530. https://doi.org/10.1073/pnas.0509544103

Spyrou, M.A., Keller, M., Tukhbatova, R.I., Scheib, C.L., Nelson, E.A., Valtueña, A.A., Neumann, G.U., Walker, D., Alterauge, A., Carty, N., Cessford, C., Fetz, H., Gourvennec, M., Hartle, R., Henderson, M., Heyking, K. von, Inskip, S.A., Kacki, S., Key, F.M., Knox, E.L., Later, C., Maheshwari-Aplin, P., Peters, J., Robb, J.E., Schreiber, J., Kivisild, T., Castex, D., Lösch, S., Harbeck, M., Herbig, A., Bos, K.I., Krause, J., 2019. Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. Nat. Commun. 10, 1–13. https://doi.org/10.1038/s41467-019-12154-0

Spyrou, M.A., Tukhbatova, R.I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A.G., Castex, D., Wahl, J., Gazimzyanov, I.R., Nurgaliev, D.K., Herbig, A., Bos, K.I., Krause, J., 2016. Historical *Y. pestis* Genomes Reveal the European Black

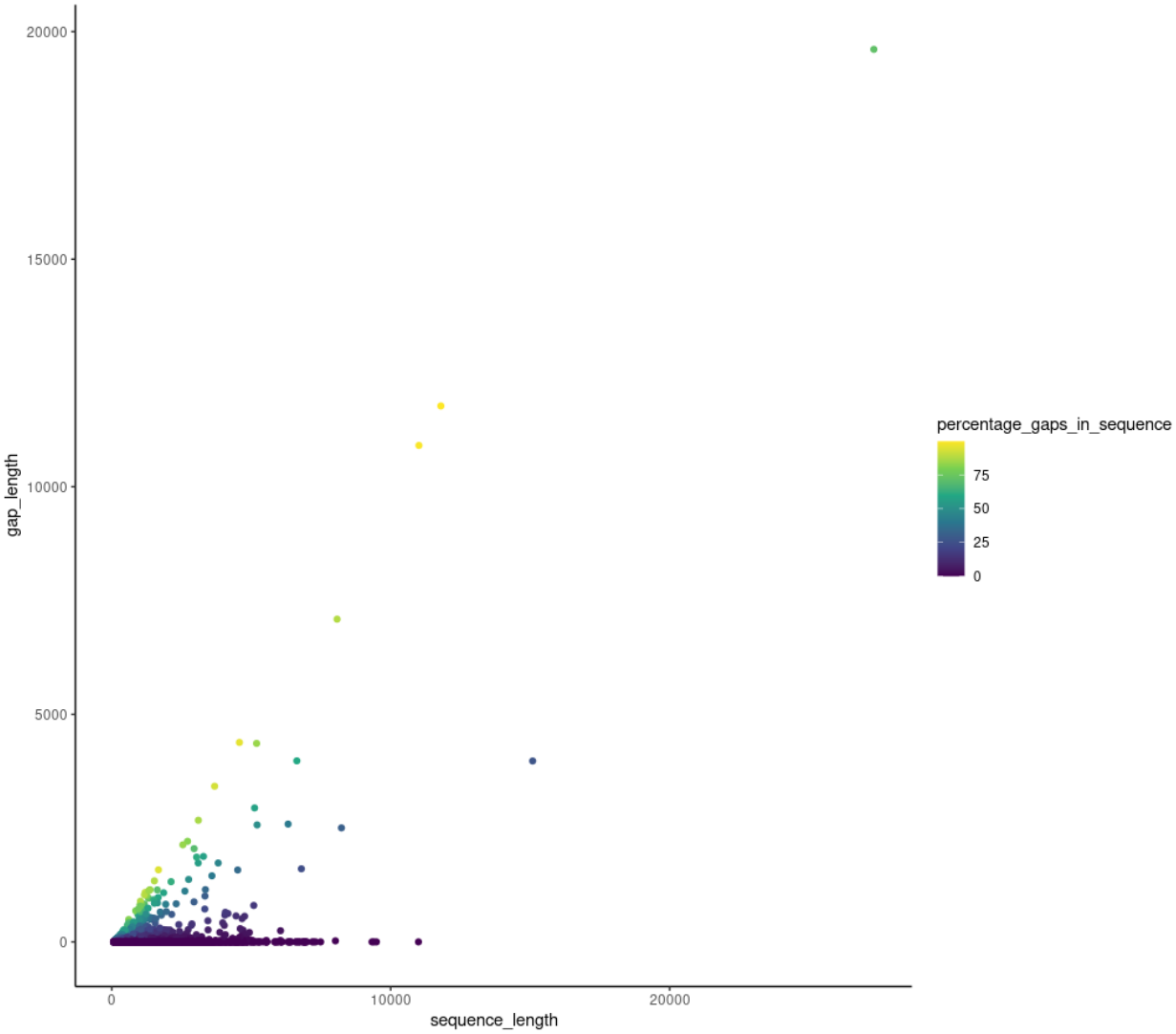
Death as the Source of Ancient and Modern Plague Pandemics. Cell Host Microbe 19, 874–881. https://doi.org/10.1016/j.chom.2016.05.012

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

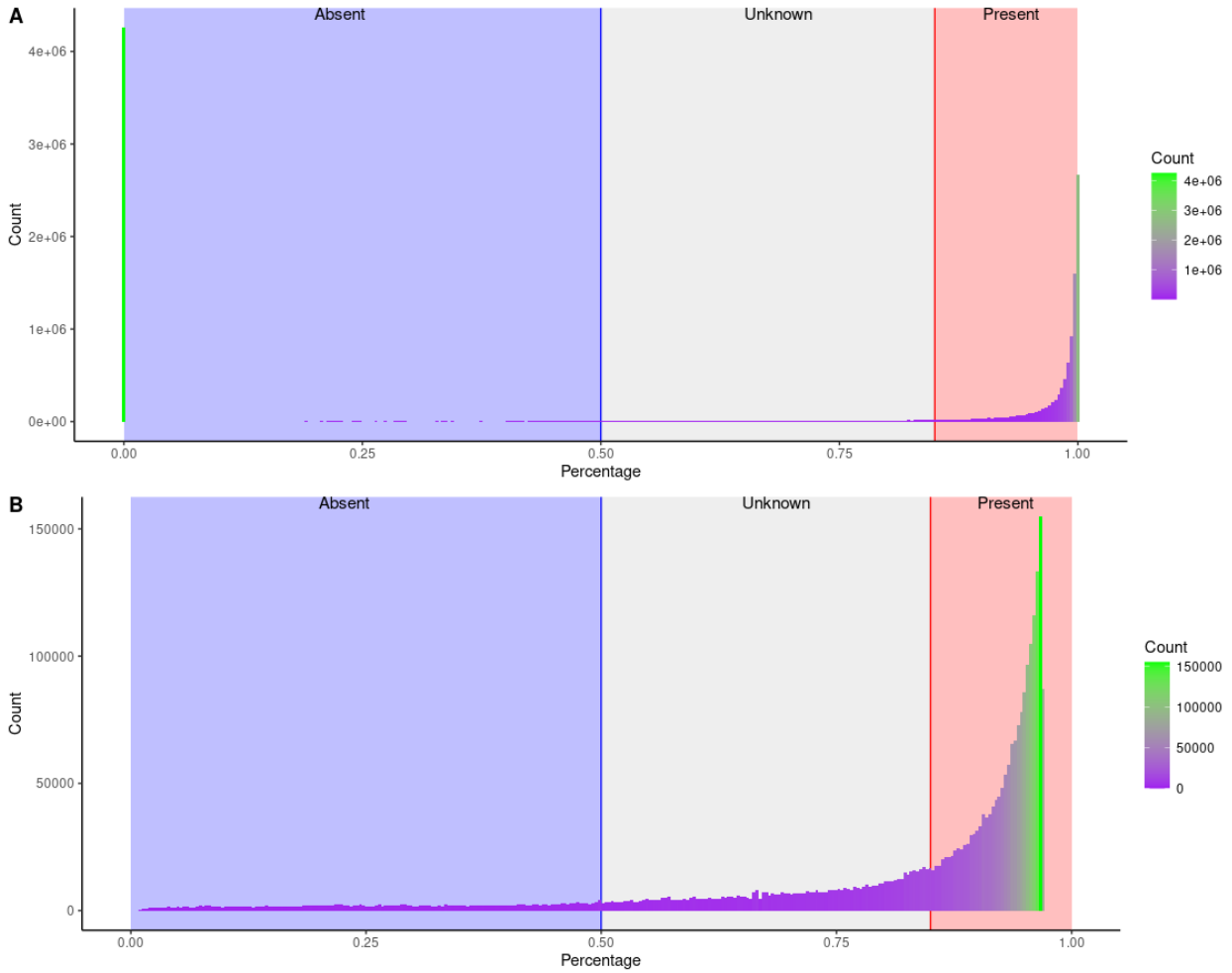
Stenseth, N.C., Atshabar, B.B., Begon, M., Belmain, S.R., Bertherat, E., Carniel, E., Gage, K.L., Leirs, H., Rahalison, L., 2008. Plague: Past, Present, and Future. PLOS Med. 5, e3. https://doi.org/10.1371/journal.pmed.0050003

Sun, Y.-C., Hinnebusch, B.J., Darby, C., 2008. Experimental evidence for negative selection in the evolution of a *Yersinia pestis* pseudogene. Proc. Natl. Acad. Sci. 105, 8097–8101. https://doi.org/10.1073/pnas.0803525105

The Gene Ontology Consortium. 2019. 'The Gene Ontology Resource: 20 Years and Still GOing Strong'. *Nucleic Acids Research* 47 (D1): D330–38. https://doi.org/10.1093/nar/gky1055.



Supplementary Figure 1: Scatter plot showing the distribution of sequence length and gap length in the sequence in the consensus genes generated by PanX. The points are coloured based on the percentage of the gene comprised by gaps.



Supplementary Figure 2: Count histogram of the percent coverage per gene (15372 in total) per genome (923 in total) (A) including all data and (B) excluding percent of coverage with less than 1% and more than 97%.

Manuscript D

De novo assembly of a 17th century *Yersinia pestis* genome from a plague victim buried in the New Churchyard burial ground in London

Aida Andrades Valtueña[1], Maria A. Spyrou[1], Elizabeth A. Nelson[1], Niamh Carty[2], Robert Hartle[2], Michael Henderson[2], Elizabeth L. Knox[2], Don Walker[2], Kirsten I. Bos[1], Alexander Herbig[1]
[1] Max Planck Institute for the Science of Human History, Jena, Germany
[2] Museum of London Archaeology (MOLA), London, United Kingdom

Introduction

Plague is a zoonotic rodent disease that can affect human populations (CDC, 2020). Based on historical documents (Zietz and Dunkelberg, 2004), this disease has been linked to three pandemics: the first pandemic or Plague of Justinian in the 6-8th century, the second plague pandemic 14-18th century, and the most recent pandemic in the 19th century. It was during the third pandemic that the causative agent of the disease was isolated: the gram-negative bacterium *Yersinia pestis* (Zietz and Dunkelberg, 2004). *Y. pestis* is one of the best studied pathogens in the field of ancient DNA (aDNA, Spyrou et al., 2019). aDNA has provided evidence to the hypothesis posed by historians that plague was responsible for the first and the second pandemic, by reconstructing *Y. pestis* genomes from the victims of those pandemics (Bos et al., 2016, 2011; Feldman et al., 2016; Keller et al., 2019; Namouchi et al., 2018; Spyrou et al., 2019b, 2016; Wagner et al., 2014). Furthermore, evidence for plague affecting human populations goes beyond the historical accounts, with the unexpected recovery of prehistoric ancient *Y. pestis* genomes, confirming that plague has been associated with humans at least since the Neolithic (Andrades Valtueña et al., 2017; Rascovan et al., 2019; Rasmussen et al., 2015; Spyrou et al., 2018). All the mentioned work has been performed by reconstructing *Y. pestis* genomes using a reference-based strategy, where reads from high-throughput sequencing are mapped against a modern reference genome. While this approach allows for the performance of multiple analyses to gain insights into the phylogenetic placement of the samples and gene variation, in terms of variant sites or loss of genes, it is limited to the variation known in the genome used as reference. In the case of ancient *Y. pestis* studies, the majority of the work has used the *Y. pestis* CO92 (NC_003143.1) as reference. An established way to circumvent this reference-bias is to perform *de novo* assembly. *De novo* assembly is a reference-free approach, where sequencing reads are combined together to create longer continuous fragments named contigs, which are then ordered creating scaffolds (Sohn and Nam, 2018). This method allows for the exploration of genome architecture and content, thus allowing the identification of new genetic elements or

rearrangements that may influence the virulence of the pathogen. Furthermore, by removing chimeric contigs one can avoid contamination from environmental organisms present in the sample, a phenomenon encountered when mapping aDNA reads to a reference with sensitive parameters (Keller et al., 2019; Vågene et al., 2018). Applying this technique in ancient DNA data will allow for a pangenomic exploration of ancient pathogens. Despite the advantages of *de novo* assembly, the currently available algorithms face a challenge when trying to piece together genomes from aDNA data. Short-fragment length and uneven coverage difficulties the construction of assembly graphs and the path-finding through the graph to reconstruct contigs and scaffolds, where repetitive elements are particularly challenging to recover. Relatively successful *de novo* assembly has been possible in cases of exceptionally well preserve samples with long-reads present, such in the case of leprosy (Schuenemann et al., 2013), genomes of *Staphylococcus saprophyticus* and *Gardnerella vaginalis* from calcified nodules (Devault et al., 2017) or small viral genomes (Krause-Kyora et al., 2018). In the case of *Y. pestis*, there have been previous attempts to perform de novo assembly from short aDNA fragments (Bos et al., 2011; Luhmann et al., 2017), resulting in rather fragmented assemblies. Furthermore, there is a lack of systematic testing and optimisation of the currently available assemblers for their application to aDNA. Information on the coverage required to circumvent the short fragments present in aDNA, it is essential for the generation of *de novo* assembled genomes from ancient DNA in a cost-effective manner. In order to address this, we present here the result from the de novo assembly of BED030, an exceptional well-preserved *Y. pestis* sample with an unusually long average fragment length for aDNA at 102bp, originally published by Spyrou et al. (2019). The BED030 sample is from a plague victim buried in the New Churchyard burial ground dating to the 17th century. We compare the current assembly with previously ancient *Y. pestis* assembled genomes from the second pandemic to gain insights into the structural and gene variation during the second plague pandemic.

Methods

Sequencing depth simulation and assembler choice

In order to maximise the performance of the assemblers, we decided to increase the coverage of the BED030 genome. By deep sequencing the sample and recovering as many fragments as possible from the genomic library, we try to circumvent the short-fragment length by providing as much overlap as possible between reads. Furthermore, it has been shown that different assemblers perform differently depending on the objective of the research (van der Walt et al., 2017). To then estimate the required sequencing depth and to select the best performing assembler for the reconstruction of the *Y. pestis* genome, we created in-silico datasets by simulating paired-end sequenced ancient reads using *Y. pestis* KIM10+ genome (NC_004088.1) as input to gargammel (Renaud et al., 2017). In order to simulate the length and the typical damage present in ancient samples, we provided gargammel with the length distribution and the damage profile from the RISE505 sample (Rasmussen et al., 2015) calculated with MapDamage2 (Jónsson et al., 2013). We used the provided Clovis dataset (Rasmussen et al., 2014) in gargammel to simulate a soil background. We produced 3 datasets with different mean coverage

for the *Y. pestis* genome (specified with the -c option in gargammel): 100X, 200X and 300X. We trimmed adapters and filtered reads with low quality and shorter than 30bp with ClipAndMerge using EAGER (v. 1.92.17, Peltzer et al., 2016). We then assembled the reads with 3 different de Bruijn graph assemblers: SOAPdenovo (v. 2.04-r240, Luo et al., 2012), velvet (v. 1.2.10, Zerbino and Birney, 2008) and SPAdes (v. 3.10.1, Bankevich et al., 2012). We generated assemblies for the different coverage datasets using k-mer sizes ranging from 47 to 67 with a 2 increment. We compared the results with QUAST (v 5.0.2, Gurevich et al., 2013) using *Yersinia pestis* KIM10+ as reference to evaluate the misassemblies introduced by the different assemblers. Based on the results, we decided to sequence the ancient BED030 sample to a coverage of 200X and use SPAdes for subsequent assembly.

Final assembly

The BED030 sample was processed in the lab as described in Spyrou et al. (2019). In short, the DNA extraction and UDG treated library for the BED030 sample was performed in the clean room facilities of the Max Planck Institute for the Science of Human History. The UDG treatment, where deaminated bases characteristic of ancient DNA have been removed, was performed in order to prevent false bases, due to deamination of cytosines, to be incorporated in the assembly. A UDG treated library was enriched for *Y. pestis* DNA via in-solution capture, previously described in Andrades Valtueña et al. (2017), to increase the amount of *Y. pestis* DNA present in the library. Based on the previously calculated genomic coverage for a more contiguous *de novo* assembly, we sequenced the BED030 enriched library to obtain a minimum of 200X mean fold coverage in a NextSeq500 machine with a paired-end 150bp kit for 300 cycles, resulting in 86,127,319 read pairs. We removed adapters and reads with less than 30 bp and minimum quality of 20 with AdapterRemoval (v. 2.2.0, Schubert et al., 2016). In order to test the performance of SPAdes with the paired-end and single-end mode, we ran AdapterRemoval twice: we used the --collapse flag to merge the read pairs and we refer to these as the single-end dataset; for the second dataset we refer to as paired-end dataset, we ran adapter removal without the --collapse flag to perform adapter trimming in each read pair separately, thus obtaining two fastq: one containing forward reads (R1) and the other containing reverse reads (R2). We reprocessed the data for BED030 published in Spyrou et al. (2019) in the same manner, however only single-end reads were present. We then assembled the two datasets with SPAdes by providing the reads either as single-end (all data was provided as -se) or as paired-end (deep sequenced data was provided as --pe for R1, R2 and --se for singletons, and previously published data was provided as an extra set of single end dataset with --se). We produced assemblies with k-mer sizes ranging from 47 to 127 with an increment of 2. The assembly quality was evaluated with QUAST using as reference *Y. pestis* CO92 chromosome (NC_003143.1). The best assembly was obtained with the paired-end configuration. We evaluated the assemblies with MetaQUAST (v5.0.2, Mikheenko et al., 2016), including as references those indicated in Table 1.

Table 1: Summary of modern genomes used in the course of analysis. For plasmids, the strain from which the plasmids were sequenced, if known, is indicated in the parenthesis.

Genome	NCBI Accession number	MetaQUAST evaluation	Ragout scaffolding	Complete genome alignment
Chromosome				
Y. pestis CO92	NC_003143.1	Yes	Yes	Yes
Y. pestis KIM10+	NC_004088.1	Yes	Yes	Yes
Y. pestis 0.PE2 Pestoides F	NC_009381	No	Yes	Yes
Y. pestis 0.PE4 Microtus 91001	NC_005810	No	Yes	Yes
Y. pestis 2ANT Nepal 561	NC_008149	No	Yes	Yes
Y. pseudotuberculosis IP32953	NC_006155	No	Yes	-
Plasmids				
pCD1 (*Y. pestis* CO92)	NC_003131.1	Yes	Yes	-
pMT1 (*Y. pestis* CO92)	NC_003134.1	Yes	Yes	-
pPCP1 (*Y. pestis* CO92)	NC_003132.1	Yes	Yes	-
pJARS35 (*Y. pestis* Java 9)	CP002179.1	Yes	No	-
pJARS36 (*Y. pestis* Java 9)	CP002181.1	Yes	No	-
pCRY (*Y. pestis* Microtus 91001)	NC_005814.1	Yes	No	-
pIP1202	CP000603.1	Yes	No	-
pYC	NC_002144.1	Yes	No	-

We scaffolded the contigs using the reference-based scaffolder Ragout (Kolmogorov et al., 2018). All genomes indicated in Table 1 were used to scaffold the chromosome. The unused reads from the chromosome scaffolding were used to scaffold the pCD1, pMT1 and pPCP1 plasmids from the *Y. pestis* CO92 strain. For the scaffolding of plasmids, we used the unplaced reads from the chromosome and any plasmids reconstructed before with the following order of scaffolding: pCD1, pMT1 and pPCP1. No other plasmids were scaffolded since there was no indication of them being present in the MetaQUAST analysis.

We ran GAPPadder (Chu et al., 2019, https://github.com/simoncchu/GAPPadder) on the resulting scaffolds to close any incorrectly inserted gaps. We further detected any incorrectly incorporated

ambiguous bases by estimating regions of the CO92 reference without any coverage, using bedtools genomecov with the following command:

```
bedtools genomecov -bga -ibam $FILE.bam | grep -w 0$ | awk 'BEGIN{FS=OFS="\t";} {print $0,($3-$2)}' > $name.missing.bed
```

Where $FILE.bam is the bam file containing all the reads for BED030 (merged-pairs, single end reads, and singletons all combined in a single fastq) with bwa mem (Li, 2013) against the reference. The output contains the regions missing and the length of these regions. We then produced a bed file containing the regions with ambiguous bases (BED030_ref_Ns.bed) in the scaffolds with a custom python script (countingNs.py, https://github.com/aidaanva/BED030_assembly/blob/master/Scripts/countingNs.py).
We aligned our scaffolds to the CO92 reference with Mauve (version snapshot_2015-02-25 build 0, Darling et al., 2010, 2007) and checked those specific regions for ambiguous bases. Any region detected during this step that was still contained in the assembly was removed using the custom python script removingNs.py (https://github.com/aidaanva/BED030_assembly/blob/master/Scripts/removingNs.py) by providing the corresponding coordinates from the BED030_ref_Ns.bed file. The output fasta file was then correctly formatted to have 60 bases per line with the following command:

```
fold -w 60 file.fasta > file_def.fasta
```

Finally, we annotated the scaffold using Prokka (Seemann, 2014):

```
prokka --compliant --centre MPISHH --outdir prokka --locustag BED030 --genus Yersinia --species pestis --strain BED030 --cpus 8 --proteins GCF_000009065.1_ASM906v1_protein.faa final_def.fasta
```

The GCF_000009065.1_ASM906v1_protein.faa file corresponds to the protein file for the *Y. pestis* CO92 reference and was downloaded from the NCBI FTP server.

Modern genomes and ancient assemblies' comparison

We downloaded the previous assemblies for the Black Death (BD, Bos et al., 2011) and Marseille (OBS, Bos et al., 2016) genomes performed by Luhmann et al., (2017) from http://paleogenomics.irmacs.sfu.ca/DOWNLOADS/AGAPES_data_results.zip. We then ordered the assemblies with the reorder contig option of mauve and using CO92 as reference to aid in the comparison with the BED030 assembly. In order to detect indels and rearrangement in the BED030 genome, we aligned the BED030 assembly, BD assembly and OBS assembly to the modern *Y. pestis* genomes indicated in Table 1 with Mauve. We detected indels by analysing the backbone file produced by Mauve in R.
To evaluate the correctness of the ancient assemblies, we mapped the original reads of each library to the references with bwa mem:

```
bwa mem -t 4 assembly.fasta original_reads.fastq.gz | samtools view -@ 4 -F 4 -bS - | samtools sort -@ 4 -m 32G - -o originial_reads.onlymapped.bam
```

Regions with no mapping reads were calculated with bedtools genomecov -bga as described above. The regions overlapping the 49kb region missing in the BED030 genome, described in Spyrou et al. (2019), were extracted with bedtools intersect and providing a bed file containing the coordinates for the deletion: 1,412,921-1,464,137 in BD and 1,419,095-1,468,514 in OBS and we calculated the percentage of non-covered bases in the region.

Results

Simulations and assembler selection

When working with aDNA resources are limited, not only economically but also in terms of samples. For that reason, maximizing the chances to obtain the best results in a cost-effective manner is fundamental. To find the best assembler and the most efficient depth of coverage to perform the assembly from the BED030 sample, we simulated 3 datasets using *Y. pestis* KIM10 as a reference with varying coverages of 100, 200 and 300 mean fold coverage using the software gargammel (Renaud et al., 2017). To evaluate the real coverage after the simulation, we use EAGER (Peltzer et al., 2016) to map and produce coverage statistics (Table 2).

Table 2: Mapping statistics for the gargammel datasets (100X, 200X and 300X) and the BED030 when aligned to the indicated reference genome with EAGER.

Sample	Reference genome	Number of reads after adapter removal and quality filtering	Uniquely mapped reads to *Y. pestis* CO92	Mean fold-coverage	Reads after duplicate removal	Mean fold coverage	Percentage of coverage at 5X
100X	KIM10	7,316,034	2,707,825	97.8961	2,333,098	86.1225	-
200X	KIM10	14,633,192	5,415,732	195.8455	4,054,882	152.7666	-
300X	KIM10	21,947,668	8,122,396	293.7906	5,329,887	204.693	-
BED030	CO92	92,442,230	31,266,205	683.3902	6,197,372	157.5644	93.1805

The final coverage for the datasets prior to duplicate removal (after duplicate removal) is 97.8961 (86.1225), 195.8455 (152.7666), and 293.7906 (204.693) for the 100, 200 and 300 datasets respectively. To determine the more suitable assembler for reconstructing this *Y. pestis* genome from ancient samples, we compared three de Bruijn graphs assemblers (SOAPdenovo (Luo et al., 2012), Velvet (Zerbino and Birney, 2008) and SPAdes (Bankevich et al., 2012) using the highest coverage dataset since we expected better performance with more data. We assembled the 300X dataset with each assembler with the following k-mer sizes: 47, 55, 57, 61, 63, 67. The best assembly was chosen with the following criteria: higher percentage of the genome covered by aligned contigs, largest alignment, highest NGA50, lowest LGA50, lower number of Ns, missassembles and mismatches. The best performing k-mer was 67 for SOAPdenovo and 63 for

Velvet. SPAdes makes use of all the k-mer provided and selects the final assembly, which can contain contigs assembled with different k-mers. The final SPAdes assembly was used for comparison with the other two *de novo* algorithms with the indicated k-mer sizes. Based on the QUAST (Gurevich et al., 2013) results, all three assemblers recovered around 96% of the genome and have comparable mis-matches and indels mistakes (Table 3). SPAdes had slightly higher mismatches per 100kbp (Table 3). Based on the largest alignment, NGA50 and LGA50 statistics, we decided to perform the assembly of the ancient genome with SPAdes. SPAdes has been previously recommended for the reconstruction of single genomes from metagenomic data (van der Walt et al., 2017). In our specific case, we gave priority to longer contigs versus correctness on the nucleotide levels, since the later could be corrected by aligning the reads to the obtained contigs. Furthermore, one of the goals of this study is to explore the genomic architecture of *Y. pestis* during one of the last plague outbreaks of the second pandemic for which longer contigs will be essential.

Table 3: Genomic statistics for the assemblers for the 200X dataset computed using QUAST and with *Y. pestis* KIM10 as a reference. The number after the _ indicates the k-mer size used in the assembly.

Assembly	**SOAPdenovo_67**	**SPAdes**	**Velvet_63**
Genome fraction (%)	96.309	96.558	96.175
# N's per 100 kbp	0.00	0.00	0.16
# mismatches per 100 kbp	0.95	1.49	1.22
# indels per 100 kbp	0.02	0.02	0.18
Largest alignment	99,072	120,038	99,061
Total aligned length	4,446,233	4,449,030	4,437,465
NGA50	29,014	47,002	27,577
NGA75	16,159	26,133	15,302
LGA50	52	34	53
LGA75	105	66	108

In order to determine the minimum mean fold coverage (depth of coverage), we compared the results of the SPAdes assemblies across the simulated databases. We observe that the assembly performance increases between 100X and 200X however there does not seem to be a great improvement between 200X and 300X. We decided to sequence our sample to a depth of 200X mean fold coverage.

Sequencing and assembly of BED030

Based on the results of the simulations, the BED030 sample was sequenced to a total of 86,127,319 read pairs. After AdapterRemoval, there were 85,494,794 usable read pairs and 42,420 singletons that were combined to the reads already published (Spyrou et al. 2019), totalling in 92,442,230 reads. To evaluate the coverage to the reference genome (*Y. pestis* CO92), we mapped the reads to the reference. We recovered a reference-based genome with an average mean fold coverage (depth of coverage) of 683 and 157 prior and after duplicate removal, respectively (Table 2).

We then assembled the pre-processed reads (92,442,230) with SPAdes and tested its performance by using the paired-end versus single-end data modes. In the paired data mode, SPAdes uses the positional data provided by the pairs during the assembly, and read pairs are provided as separate files (See methods). For the single end mode, the pairs were merged and concatenated with the truncated pairs and singletons and provided to SPAdes as a single fastq file. For both of the assemblies, the previously published reads were provided as an additional single-end experiment. We evaluated the scaffolds of the paired-end and single-end modes with QUAST and the results can be found in Supplementary Table 1. Both configurations (paired-end vs single-end) result in similar assembly statistics. We decided to use the assembly produced by the paired-end dataset, since SPAdes can use the pair-information during assembly, which could aid in the correct assembly and ordering of contigs during the scaffolding step. The final assembly consisted of a total of 78,024 contig from which 789 contigs aligned to the *Y. pestis* CO92 chromosome. The aligned contigs covered 94.498% of the reference, similarly to what we observed when we mapped the raw reads (Table 2). The largest contig aligning to the reference consisted of 81,789 bp. The assembly had an NG50 of 43,792 bp and a LG50 of 71 contigs. Based on the alignment of the contigs with MetaQUAST (Mikheenko et al., 2016) to the references indicated in Table 1, we confirm the presence of the plasmids pCD1, pMT1 and pPCP1 in the BED030, as already described in Spyrou et al. (2019). We did not detect any contig aligning to any other plasmids known to be present in other *Y. pestis* strains, suggesting that the BED030 strains did not have any additional known plasmids.

In order to improve the continuity of the scaffolds produced by SPAdes, we used the reference-based scaffolder Ragout (Kolmogorov et al., 2014). We provided the genomes indicated in Table 1 as reference for the chromosome, which were chosen for its completeness and for their phylogenetic position in order to include a wide range of diversity present in *Y. pestis*. Furthermore, *Y. pseudotuberculosis* IP32953 was included as an outgroup representative. To reconstruct the detected plasmids during the assembly evaluation, we used the pCD1, pMT1 and pPCP1 plasmids from *Y. pestis* CO92 (see methods). We used GAPPadder (Chu et al., 2019) to close gaps in the sequence, which was successful for the chromosome and pMT1 but failed for the pCD1 and pPCP1 plasmids due to very small gaps introduced during the scaffolding process. The results are summarised in Table 4.

Table 4: Assembly statistics for the *Y. pestis* BED030 *de novo* reconstruction after Ragout scaffolding (all) and gap closing with GAPadder (with exception of pCD1 and pPCP1). Ref.=Reference, in this case *Y. pestis* CO92.

Element	Size Ref.	Number Scaffolds	Total Scaffold length	Number of Used Contigs	Percent introduced Ns	Assembly N50	Length difference with Ref.
Chromosome	4,653,728	2	4,519,711	316	2.44	4488791	-134,017
pCD1	70,305	1	68034	8	0.43	68034	-2271
pMT1	96,210	1	95533	7	0.01	95533	-677
pPCP1	9,612	1	8109	2	0.14	8109	-1503

In order to remove potentially misincorporated Ns during the Ragout scaffolding, we detected missing regions when mapping the reads to *Y. pestis* CO92. We detected 8 regions with less than 10% coverage missing (Table 5). Since we have no evidence for these regions to be present, we aligned the scaffold to *Y. pestis* CO92 and removed Ns present in these regions.

Table 5: Missing regions of the BED030 genome when mapped to the CO92 reference. Note that no mapping quality was applied. In bold is indicated the 49Kb deletion detected in Spyrou et al. 2019 and removed from the current scaffold.

Start Position	End Position	Window size	Percentage covered	Present assembly as Ns
1,879,480	**1904389**	**24909**	**0.007427**	**Yes (1834382-1885429)**
1,906,601	**1928856**	**22255**	**0.004044**	
4,101,311	4102545	1234	0.0818477	No
1,780,139	1781238	1099	0	No
2,376,978	2377535	557	0	No
2,540,520	2542500	1980	0	No
2,551,771	2553158	1387	0	No
2,554,185	2562891	8706	0	No

All of the detected missing regions are also absent in the scaffold, with the exception of the 49kb deletion (1,879,480-1,904,389 and 1,906,601-1,928,856) described in Spyrou et al. 2019. In this case, Ragout incorporated ambiguous bases (N) since this region is present in all of the other genomes that were used as references, and GAPPadder failed to close the gap, potentially due to a rearrangement in this region. Despite this, we decided to remove the ambiguous bases since there is no indication (no coverage of the region by reads and no contigs confirming this region present in the assembly) that they are correct and would lead to an overestimation of the real size of the chromosome. The final size of the BED030 *Y. pestis* chromosome after the removal of the deletion is 4,468,664bp, consisting of two scaffolds with 4,437,744bp and 30,920bp. We

annotated the scaffolds with Prokka (Seemann, 2014) resulting in 3907 CDS, 3976 genes, 5 rRNA, 3 repeat_region, 63 tRNA and 1 tmRNA, which with the exception of fewer RNA genes detected are in line with that observed in *Y. pestis* CO92 with 4,252 CDS, 19 rRNA, 69 tRNA, 12 other RNAs.

Comparison to other ancient assemblies and modern genomes

We compared our assembly with previously assembled genomes of the Black Death (BD, Bos et al., 2011) and Marseille (OBS, Bos et al., 2016) by Luhmann et al., (2017). The results of the assemblies comparison can be seen in Table 6.

Table 6: Comparison with ancient assemblies indicating the scaffolding strategy and basic assembly statistics.

Assembly	Method	Number of scaffolds	Total length (bp)	N50	Number Ns
BED030	SPAdes + Ragout + GAPadder	2	4,468,664	4,437,744	59,307 (1.33%)
BD	Minia + AGapEs	5	4,441,104	3,511,710	0 (0%)
OBS	Minia + AGapEs	6	4,350,872	3,459,919	0 (0%)

We observe that our assembly strategy contains larger and fewer scaffolds but still contains 1.33% of ambiguous calls (Ns). The assemblies from Luhmann et al., (2017) do not have any ambiguous call at the cost of a more fragmented assembly. This is due to the gap filling strategy implemented in AGapEs in which missing regions are inferred by ancestral reconstruction of the gaps using extant genomes. We compared the three assemblies by aligning them with mauve to the modern *Y. pestis* genomes from the CO92, KIM 10, Microtus 91001 strains. Prior to alignment we reordered the scaffolds of BD and OBS based on *Y. pestis* CO92 (see methods). *Y. pestis* is known for its plastic genome, and this is well displayed in the alignment produced by mauve (Supplementary Figure 1). We identify 4 inverted regions in BED030 in comparison to the reference genome (*Y. pestis* CO92, Table 7). With the exception of the inversion located in 2,766,751-2,836,482, the other three inversions are supported by the presence of contigs containing inverted regions at least at one end of the inversion. We did not identify any region present exclusively in the BED030 genome that was present neither in second pandemic genomes (BED030, OBS, BD), nor second pandemic genomes and CO92. We detected 9 regions only missing on the second pandemic *de novo* assembled genomes, however they all correspond to either repeated regions or insertion elements (Supplementary File 1).
We noticed some difference between the BED030 assembly and the other previously assembled ancient *Y. pestis* genomes. There are 71 regions that are bigger than 500bp present in the BD and OBS assemblies that are missing in the BED030 assembly (Supplementary File 2). With the

exception of the 49Kb deletion already described above, the rest of the missing regions detected in BED030 are either IS related (63.5%), duplicated genes (19%), rRNA or tRNA clusters (7.9%). There were some genes (7.9%) that seem to be absent from the mauve alignment, however they were present in the gene annotation of BED030, suggesting that those genes are present in the genome. We also detected the 49Kb region being present in BD and OBS, however mauve aligned it to the version of the genes present in KIM10+ and 0.PE2 Pestoides F. This was rather surprising since this deletion was also seen in the OBS genomes (Spyrou et al., 2019b). In order to assess the correctness of this region in the previously assembled genomes BD and OBS, we mapped the original reads to the assemblies and calculated the coverage in the regions that correspond to the missing region in BED030 being 1,412,921-1,464,137 and 1,419,095-1,468,514 for BD and OBS respectively. As expected, the region only contains 1.2% of unmapped bases in the BD genome, while we observe that 90% of the region is not covered in OBS (Supplementary File 2). This further supports the presence of this deletion also in the OBS genome, which will have occurred in the ancestor of the OBS and BED030 genomes (Spyrou et al., 2019b). We suspect that this region was wrongly assembled in the OBS genome in Luhmann et al. (2017), probably due to the inferred ancestral reconstruction performed by AGapEs.
We also compared the BED030 reconstructed plasmids to those of *Y. pestis* CO92. We do not detect any rearrangements, but regions comprising mobile elements or transposons explain the difference in size with respect to the reference.

Table 7: Inverted regions identified in the BED030 *de novo* assembly.

Start	End	Size (bp)	Contig support
846,038	1,113,657	267,619	Right side
2,766,751	2,836,482	69,731	No
2,838,116	2,997,831	159,715	Both
3,847,188	3,975,231	128,043	Left side

Discussion

Ancient DNA is characterised by short fragments lengths, deamination due to hydrolytic damage (typical aDNA damage) and an uneven coverage across the genome due to differential preservation of genomic regions. Furthermore, ancient DNA samples are often composed not only by the host DNA and its associated microorganisms, such as representatives of the microbiome or potential pathogens that infected the individual, as they also contain environmental contaminants (decomposing microbes or organisms present in the environment where the individual was deposited), thus creating a metagenomic context. Those characteristics pose a challenge to the currently available *de novo* assembly algorithms and is one of the reasons why routine *de novo* assembly of ancient genomes has remained elusive. The BED030 sample is exceptionally well preserved, shown by the presence of longer fragments, which allowed us to perform a *de novo* reconstruction of the *Y. pestis* genome infecting the individual that has been

dated to the 17th century. After comparison of various assemblers and data inputs, we decided that the best *de novo* strategy was using SPAdes by providing paired-end reads separately followed by reference-based scaffolding with Ragout, and gap closing with GAPadder. This strategy is different to the one previously employed to reconstruct less well-preserved assemblies from the East Smithfield Black Death (BD) and Marseille (OBS) genomes (Luhmann et al., 2017). Luhmann et al., (2017) used the assembler Minia in combination with their gap closing algorithm AGapEs, which was developed for ancestral reconstruction of the gaps based on extant genomes and with special emphasis on the recovery of IS elements. The assembly of the BED030 genome is more continuous at the expense of incorporating ambiguous calls, largely due to repetitive elements such as IS elements, while the BD and OBS assemblies are more fragmented, but they recover more efficiently the repetitive elements which was part of the reason why AGapEs was developed. However, we detected a wrongly incorporated region in the OBS genome consisting of a 49kb region that has been previously reported to be missing in this genome (Spyrou et al., 2019b) and has been correctly assembled here with BED030. We observe differences in the genomic organisation between BED030, BD and the OBS assembled genomes. While the BD and OBS seem to be more similarly arranged to the *Y. pestis* KIM10 genome (used as one of the modern representatives in the AGapEs step), BED030 is closer in its genomic organisation to the *Y. pestis* CO92 strain, despite the fact that KIM10 was included in the set of references for Ragout scaffolding. These differences in genomic organisations are likely due to differing strategies in the scaffolding step. More assemblies from strains dating to the second pandemic will be needed to understand the genomic rearrangements during the second pandemic.

In addition to producing a more contiguous assembly, we went an additional step further over Luhmann et al (2017), and performed gene annotation. The number of CDS present in the BED030 strain (3,907) is within the range of the observed genes in other *Y. pestis* genomes (ranging between 3,056-5,351 CDS), however BED030 has fewer rRNA and tRNAs than reported in other *Y. pestis*, highlighting the challenged to assemble and annotate repeated regions.

A shared disadvantage of the assembled ancient *Y. pestis* genomes presented in here, is the fact that they all have been recovered from enriched libraries which can lead to capture bias (See Bos et al., 2016, 2011; Spyrou et al., 2019 for description of the designs). Regions not included in the capture design will have no coverage or substantially lower coverage than those enriched by the probes, thus leading to difficulties to assemble sample unique regions. With dropping sequencing costs, future shotgun sequencing of unenriched samples should be preferred over enriched libraries. This will aid in improving our understanding of the gene and structural content in ancient strains and their functional implications; which will likely be the next stage in the development of ancient pathogenomics.

Summary

Here we present a *de novo Y. pestis* genome from an individual who likely died of plague during an outbreak in London in the 17th century. The *Y. pestis* BED030 strain is a representative of the bacterium circulating during the second plague pandemic. We observed genomic rearrangements in this bacterium, thus allowing for the exploration of genomic structure of *Y. pestis* in the past. We detect 3,907 genes in the BED030 strain and confirm the absence of the 49kb region previously described (Spyrou et al., 2019b), both by the lack of contigs in that region but also by the absence of those genes in the annotated genome.

References

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. Curr. Biol. 27, 3683-3691.e8. https://doi.org/10.1016/j.cub.2017.10.025

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. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bos, K.I., Herbig, A., Sahl, J., Waglechner, N., Fourment, M., Forrest, S.A., Klunk, J., Schuenemann, V.J., Poinar, D., Kuch, M., Golding, G.B., Dutour, O., Keim, P., Wagner, D.M., Holmes, E.C., Krause, J., Poinar, H.N., 2016. Eighteenth century *Yersinia pestis* genomes reveal the long-term persistence of an historical plague focus. eLife 5, e12994. https://doi.org/10.7554/eLife.12994

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. https://doi.org/10.1038/nature10549

CDC, 2020. Plague home | CDC [WWW Document]. Cent. Dis. Control Prev. URL /plague/index.html (accessed 9.28.20).

Chu, C., Li, X., Wu, Y., 2019. GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads. BMC Genomics 20, 426. https://doi.org/10.1186/s12864-019-5703-4

Darling, A.E., Mau, B., Perna, N.T., 2010. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE 5, e11147. https://doi.org/10.1371/journal.pone.0011147

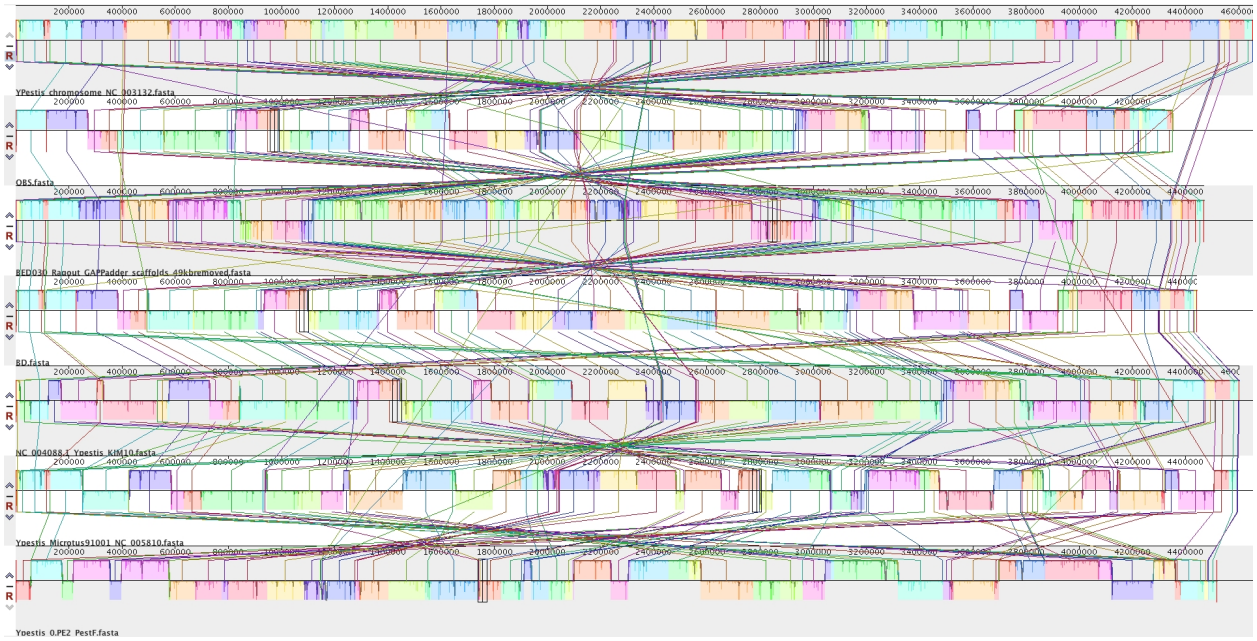
Darling, A.E., Treangen, T.J., Messeguer, X., Perna, N.T., 2007. Analyzing Patterns of Microbial Evolution Using the Mauve Genome Alignment System, in: Comparative Genomics, Methods In Molecular Biology™. Humana Press, pp. 135–152. https://doi.org/10.1007/978-1-59745-515-2_10

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.

https://doi.org/10.7554/eLife.20983
Feldman, M., Harbeck, M., Keller, M., Spyrou, M.A., Rott, A., Trautmann, B., Scholz, H.C., Päffgen, B., Peters, J., McCormick, M., Bos, K., Herbig, A., Krause, J., 2016. A high-coverage *Yersinia pestis* Genome from a 6th-century Justinianic Plague Victim. Mol. Biol. Evol. msw170. https://doi.org/10.1093/molbev/msw170
Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086
Keller, M., Spyrou, M.A., Scheib, C.L., Neumann, G.U., Kröpelin, A., Haas-Gebhard, B., Päffgen, B., Haberstroh, J., Lacomba, A.R. i, Raynaud, C., Cessford, C., Durand, R., Stadler, P., Nägele, K., Bates, J.S., Trautmann, B., Inskip, S.A., Peters, J., Robb, J.E., Kivisild, T., Castex, D., McCormick, M., Bos, K.I., Harbeck, M., Herbig, A., Krause, J., 2019. Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). Proc. Natl. Acad. Sci. 116, 12363–12372. https://doi.org/10.1073/pnas.1820447116
Kolmogorov, M., Raney, B., Paten, B., Pham, S., 2014. Ragout—a reference-assisted assembly tool for bacterial genomes. Bioinformatics 30, i302–i309. https://doi.org/10.1093/bioinformatics/btu280
Krause-Kyora, B., Susat, J., Key, F.M., Kühnert, D., Bosse, E., Immel, A., Rinne, C., Kornell, S.-C., Yepes, D., Franzenburg, S., Heyne, H.O., Meier, T., Lösch, S., Meller, H., Friederich, S., Nicklisch, N., Alt, K.W., Schreiber, S., Tholey, A., Herbig, A., Nebel, A., Krause, J., 2018. Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. eLife 7, e36666. https://doi.org/10.7554/eLife.36666
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.
Luhmann, N., Doerr, D., Chauve, C., 2017. Comparative scaffolding and gap filling of ancient bacterial genomes applied to two ancient Yersinia pestis genomes. Microb. Genomics 3. https://doi.org/10.1099/mgen.0.000123
Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, 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, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18. https://doi.org/10.1186/2047-217X-1-18
Mikheenko, A., Saveliev, V., Gurevich, A., 2016. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088–1090. https://doi.org/10.1093/bioinformatics/btv697
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. Proc. Natl. Acad. Sci. 115, E11790. https://doi.org/10.1073/pnas.1812865115
Rascovan, N., Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C., Rasmussen, S., 2019. Emergence and Spread of Basal Lineages of *Yersinia pestis* during the Neolithic Decline. Cell 176, 295-305.e10. https://doi.org/10.1016/j.cell.2018.11.005
Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., Poznik, G.D., Gudmundsdottir, V., Yadav, R., Malaspinas, A.-S., White, S.S., Allentoft, M.E., Cornejo, O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S., Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V., Lopes, M.C., Malhi, R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F., Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R., Willerslev, E., 2014. The genome of a late Pleistocene human from a Clovis burial site in western Montana. Nature

506, 225–229. https://doi.org/10.1038/nature13025
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. https://doi.org/10.1016/j.cell.2015.10.009
Renaud, G., Hanghøj, K., Willerslev, E., Orlando, L., 2017. gargammel: a sequence simulator for ancient DNA. Bioinformatics 33, 577–579. https://doi.org/10.1093/bioinformatics/btw670
Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jäger, G., Bos, K.I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J.L., Kjellström, A., Wu, H., Stewart, G.R., Taylor, G.M., Bauer, P., Lee, O.Y.-C., Wu, H.H.T., Minnikin, D.E., Besra, G.S., Tucker, K., Roffey, S., Sow, S.O., Cole, S.T., Nieselt, K., Krause, J., 2013. Genome-Wide Comparison of Medieval and Modern *Mycobacterium leprae*. Science 341, 179–183. https://doi.org/10.1126/science.1238286
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153
Sohn, J., Nam, J.-W., 2018. The present and future of *de novo* whole-genome assembly. Brief. Bioinform. 19, 23–40. https://doi.org/10.1093/bib/bbw096
Spyrou, M.A., Bos, K.I., Herbig, A., Krause, J., 2019a. Ancient pathogen genomics as an emerging tool for infectious disease research. Nat. Rev. Genet. 20, 323–340. https://doi.org/10.1038/s41576-019-0119-1
Spyrou, M.A., Keller, M., Tukhbatova, R.I., Scheib, C.L., Nelson, E.A., Andrades Valtueña, A., Neumann, G.U., Walker, D., Alterauge, A., Carty, N., Cessford, C., Fetz, H., Gourvennec, M., Hartle, R., Henderson, M., Heyking, K. von, Inskip, S.A., Kacki, S., Key, F.M., Knox, E.L., Later, C., Maheshwari-Aplin, P., Peters, J., Robb, J.E., Schreiber, J., Kivisild, T., Castex, D., Lösch, S., Harbeck, M., Herbig, A., Bos, K.I., Krause, J., 2019b. Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. Nat. Commun. 10, 1–13. https://doi.org/10.1038/s41467-019-12154-0
Spyrou, M.A., Tukhbatova, R.I., Feldman, M., Drath, J., Kacki, S., Beltrán de Heredia, J., Arnold, S., Sitdikov, A.G., Castex, D., Wahl, J., Gazimzyanov, I.R., Nurgaliev, D.K., Herbig, A., Bos, K.I., Krause, J., 2016. Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. Cell Host Microbe 19, 874–881. https://doi.org/10.1016/j.chom.2016.05.012
Spyrou, M.A., Tukhbatova, R.I., Wang, C.-C., Andrades Valtueña, A., Lankapalli, A.K., Kondrashin, V.V., Tsybin, V.A., Khokhlov, A., Kühnert, D., Herbig, A., Bos, K.I., Krause, J., 2018. Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. Nat. Commun. 9, 2234. https://doi.org/10.1038/s41467-018-04550-9
Wagner, D.M., Klunk, J., Harbeck, M., Devault, A., Waglechner, N., Sahl, J.W., Enk, J., Birdsell, D.N., Kuch, M., Lumibao, C., Poinar, D., Pearson, T., Fourment, M., Golding, B., Riehm, J.M., Earn, D.J.D., DeWitte, S., Rouillard, J.-M., Grupe, G., Wiechmann, I., Bliska, J.B., Keim, P.S., Scholz, H.C., Holmes, E.C., Poinar, H., 2014. *Yersinia pestis* and the Plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect. Dis. 14, 319–326. https://doi.org/10.1016/S1473-3099(13)70323-2
Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. https://doi.org/10.1101/gr.074492.107
Zietz, B.P., Dunkelberg, H., 2004. The history of the plague and the research on the causative agent *Yersinia pestis*. Int. J. Hyg. Environ. Health 207, 165–178. https://doi.org/10.1078/1438-4639-00259

Supplementary material



Supplementary Figure 1: Mauve alignment view. The Locally Collinear Blocks (LCB), representing alignment shared by different genomes, are coloured in the same shade in all the genomes. From top to bottom: *Y. pestis* CO92, OBS, BED030, BD, *Y. pestis* KIM10, *Y. pestis* Microtus 91001 and *Y. pestis* 0.PE2 Pestoides F.

Supplementary Table 1: Assembly statistics for Paired-end and Single-end modes assembled with SPAdes. The statistics were calculated with QUAST, using Yersinia pestis CO92 as reference.

Assembly	Paired-end mode	Single-end mode
Reference free statistics		
# contigs/total length (>= 1000 bp)	11,109/41,069,522	10,619/40,640,177
# contigs/total length (>= 5000 bp)	1735/23,012,425	1,735/23,369,046
# contigs/total length (>= 10000 bp)	818/16,587,847	825/17,035,419
# contigs/total length (>= 25000 bp)	190/7,026,814	206/7,726,797
# contigs/total length (>= 50000 bp)	21/1,410,220	25/1,708,144
# contigs	78,024	35,142
Largest contig	12,6625	147,632
Total length	69,781,808	54,983,912
Assembly GC (%)	57.73	58.68
N50	1629	3274
N75	481	947
L50	6200	2770
L75	28238	11234
Reference statistics (*Y. pestis* CO92)		
Genome fraction (%)	94.705	94.574
Reference GC (%)	47.64	47.64
Largest alignment.	81,789	81,789
Total aligned length	4,463,736	4,439,983
NG50	43,792	45,791
NG75	36,196	37,988
LG50	41	39
LG75	71	66
Duplication ratio	1.017	1.012
# N's per 100 kbp	2.56	0.81
# unaligned contigs	77235 + 319 part	34470 + 327 part
Unaligned length	65,298,222	50,528,300
NGA50	21,446	23,220
NGA75	11,031	11,172
LGA50	72	66
LGA75	146	136

Manuscript E

Insights into the population structure and virulence of *Streptococcus mutans* from an ancient genome

Aida Andrades Valtuena[1], Alissa Mittnik[1], Saskia Pfrengle[2], Alexander Herbig[1], Johannes Krause[1]

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

[2] Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, 72070, Tübingen, Germany

Abstract

Streptococcus mutans (S. mutans) is a bacterium renowned for its association to dental caries. Its ability to metabolise a wide variety of sugars, producing lactic acid as by-product (acidogenicity), to survive in low pH (acidurity), and to form biofilms contribute to its cariogenic potential. Dental caries poses a major burden on the health system in modern societies, and they have afflicted humans for millennia. Despite the fact that caries are ubiquitous in the archaeological record, it has been proposed that they became more prevalent and abundant in human populations after the introduction of agriculture due to the increased availability of dietary carbohydrates, including sugars, that can be metabolized by cariogenic bacteria, such as *S. mutans*. The causes of archaeological caries and how cariogenic bacteria have co-evolved over deep time scales with humans remain uncertain. Here, we present the first fully sequenced ancient genome of *S. mutans* from a hunter-gatherer individual with evidence of caries who lived around ~2000 years ago in South Africa. It represents the first sequenced ancient bacterial genome from Africa. Its gene content is indistinguishable from that of modern strains based on the presence/absence pattern of virulence factors. As observed in some modern genomes, it lacks elements from the two-component systems related to acid resistance and the competence pathway. Based on modern strain data, it had been proposed that *S. mutans* dispersed with ancient humans and developed phylogeographic structure due to its vertical transmission. However, we do not detect any phylogeographical signal when analysing our ancient strain together with 193 modern *S. mutans* isolates. We propose that the lack of signal is due in part to the available modern comparative dataset not being suitable to test this hypothesis. A large ancient *S. mutans* dataset is more appropriate for testing hypotheses related to *S. mutans* phylogeography, demographic expansion and virulence adaptation in response to human behavioural changes, possibly hidden by recent globalisation. In order to facilitate the recovery of *S. mutans* from metagenomic samples, we present here an in-solution capture assay for the efficient retrieval of *S. mutans*.

Introduction

Streptococcus mutans is a gram-positive bacterium that resides primarily in the human oral cavity. It was originally identified by a culture-based approach from teeth with advanced caries (Clarke, 1924), and it has been strongly associated with dental decay, particularly to the formation of caries (Loesche, 1986). Despite modern dental healthcare, dental caries remains a major health problem globally, being ranked as the most prevalent disease in the Global Burden of Disease Study, with an estimated 2.3 billion people affected by cavities in permanent teeth (Vos et al., 2017). Even though *S. mutans* was originally proposed as the primary cause of caries formation, it has been since shown that that caries is a polymicrobial disease, that other bacteria have also cariogenic potential, and that *S. mutans* is not solely required for caries initiation (Aas et al., 2008).

The cariogenic potential of *S. mutans* is associated with its ability to metabolise a wide range of sugars, which leads to the production of lactic acid that can demineralize the tooth (Loesche, 1986). Furthermore, *S. mutans* is renowned for its ability to survive in extremely low pH environments (Bender et al., 2016). In 2002, the first complete genome of *S. mutans* was sequenced showing the multiple pathways for the transport and metabolism of sugars, pathways for biofilm formation, two competence pathways by which *S. mutans* can acquire foreign DNA and virulence genes, which contribute to the cariogenic potential of the bacterium (Ajdić et al., 2002). *S. mutans* has a dynamic genome shaped by recombination, gene shuffling and acquisition of genetic material through horizontal gene transfer (Maruyama et al., 2009). Furthermore, the analysis of its pan-genome highlighted a large gene diversity present in the species, indicating an open pangenome with a smaller core genome than other streptococci (Meng et al., 2017). Comparison between strains that present different caries manifestations (defined by number of caries, onset and number of teeth affected), have shown that there are no clear underlying bacterial genetic differences among *S. mutans* that explain the different manifestations of caries in humans (Argimón et al., 2014; Argimón and Caufield, 2011). This highlights the important role that behavioural and social factors, as well as dietary habits, play in the development of caries. For example, in developed countries, consumption patterns of free sugars and fermentable carbohydrates explain trends of increasing caries prevalence (Moynihan and Petersen, 2004; Peres et al., 2016). There is also evidence correlating carbohydrate restricted diets with decreased levels of *S. mutans* (Stoppelaar et al., 1970). Finally, it has been shown that preventive interventions, such as tooth brushing and raising awareness among parents of the relationship between sugar and caries, can reduce the incidence of early childhood caries (Sälzer et al., 2017).

Despite a longstanding clinical research focus on understanding *S. mutans*, little is known about its evolution or long-term adaptation to the human host. Dental caries are present throughout the archaeological record, although some studies have proposed an increase of caries prevalence after the acquisition of agriculture (Cohen and Armelagos, 1984; Formicola, 1987; Lukacs, 1992; Nicklisch et al., 2016; Warinner, 2016). It has been argued that cultivation of starchy crops increased the consumption of carbohydrates thus providing an environment favourable to microbial agents associated with dental caries (Cohen and Armelagos, 1984; Lukacs, 1992), although caries are also prevalent at hunter-gatherer sites where carbohydrate-rich foods were consumed (Humphrey et al., 2014). Previous studies have shown the presence of *S. mutans* in

archaeological material by PCR amplification of either 16S rRNA (Adler et al., 2013) or dextranase, a gene specific to *S. mutans* (De La Fuente et al., 2013; Simón et al., 2014), or by shotgun sequencing (Warinner et al., 2014; Willmann et al., 2018). In concordance with some archaeological studies, Adler et al. (2013) suggested that *S. mutans* became more prevalent after the introduction of agriculture, and further increased after the introduction of sugar during the industrial revolution. Similarly, Cornejo et al. (2013) proposed a demographic expansion of *S. mutans* dating roughly with the start of agriculture. However, no ancient genome from *S. mutans* has been reconstructed to date, hindering the testing of these hypotheses. Reconstructing ancient *S. mutans* genomes would provide a temporal evidence on how this bacterium evolved in response to lifestyle and cultural transition of the human host, particularly given its dynamic genome, which can help inform future strategies to combat such prevalent disease in modern populations.

S. mutans is predominantly transmitted vertically, from caregiver to offspring (da Silva Bastos et al., 2015; Lapirattanakul and Nakano, 2014; Li and Caufield, 1995). Initially it was hypothesised that *S. mutans* genomes may retain phylogeographic signatures in a manner similar to *Helicobacter pylori*, preserving genetic patterns reflecting the migration histories of their human hosts (Caufield et al., 2007). However, whether there is a phylogeographic signal resembling the dispersal of modern humans is still debated. While Cornejo et al. (2013) studied 57 *S. mutans* isolates and did not detect any phylogeographic signal, González-Ittig et al. (2016) described some stratification of *S. mutans* strains that they correlated to the sample location. A potential explanation for the lack of phylogeographic structure in the modern dataset utilised in Cornejo et al. (2013) is the recent globalisation of modern human populations, which could dilute or hide previous population structure. The presence of a phylogeographic signal in *S. mutans* would open the possibility to track human movements at a finer time scale than that allowed by the human genome because of the faster generation time of the bacteria compared to the human host. Additionally, *S. mutans* is potentially a better candidate to track human movements in the past than *H. pylori* because the former has been detected in caries and also in dental calculus, both easily found in the archaeological record, while the latter is recovered from the stomach, and the only instance of a successful ancient recovery of this taxon has been from the unusually well-preserved mummy of Ötzi (Maixner et al., 2016).

Here we retrieved the first ancient DNA genome of *S. mutans* from a permanent tooth with caries from a San hunter-gatherer individual dating to ~2000 years ago (cal BP). We test whether there is a phylogeographic signal present in *S. mutans* strains and identify potential challenges to testing this hypothesis with currently available datasets. Furthermore, we assess the virulence potential of this ancient strain and evaluate changes in genes that may be related to host subsistence strategy, such as the adoption of agriculture. We also present a new genome-wide in-solution capture assay to retrieve *S. mutans* genomes from archaeological material. The newly designed capture assay allows for cost-effective recovery and reconstruction of ancient *S. mutans* genomes in a scalable manner. Generating a large-scale dataset of ancient *S. mutans* genomes and analysing them using the presented analytical framework here will shed light on the history and evolution of one of humanity's most prevalent and pernicious pathogens.

Methods

Archaeological information

The FAR004 individual (previously published in Skoglund et al. (2017) and referred to as either UTC-386 or I9133) was excavated from the rock shelter site of Faraoskop under controlled conditions in 1987 and 1988 (Manhire, 1993). A total of 12 individuals were found in this site (7 of which were retrieved prior to the excavation by the owner of the land where the site is located and 5 during excavation). The Faraoskop site is located about 30 km inland from Elands Bay on the west coast of the Western Cape Province of South Africa. Despite not containing any rock paintings, a rich archaeological assemblage from the Later Stone Age artifacts was recovered during the excavation (See Manhire, 1993 for more details). The independent radiocarbon dating from six individuals dates the site to 2300 to 1900 calibrated years before present (cal BP) (Manhire, 1993; Sealy et al., 1992). It is also to be noticed that all the available dentitions (5 individuals) display carious lesions, including the individual analysed here (Sealy et al., 1992).
FAR004 individual is a male that died between 40-50 years of age and it was dated to 2017-1748 calibrated BP (cal BP). Two carious lesions were detected in the 30 teeth available for this individual. The tooth used here for extraction had a carious lesion. The human remains of the Faroskop site are stored in the Department of Human Biology at the University of Cape Town.

Data production

The ancient DNA extraction and library preparation for FAR004 was performed at the University of Tübingen in a laboratory dedicated to ancient DNA work as previously described in (Skoglund et al., 2017). In short, the tooth of a South African individual (FAR004), radiocarbon dated to 2017-1748 cal BP, was cut transversely at the border of the crown and the root, and a total of 56 mg of dentine was sampled from the pulp chamber using a sterile dentistry drill. The dentine was extracted using a protocol optimized for the recovery of highly fragmented ancient DNA using the full-digestion method as described in Dabney et al., (2013), resulting in 100 mL of DNA extract. A double-indexed library was produced with a 20 mL aliquot of the extract as described in (Kircher et al., 2012; Meyer and Kircher, 2010). An additional, 20mL of extract was used to produce four UDG treated libraries (FAR004.A0101, FAR004.A0102, FAR004.A0103 and FAR004.A0104), which removed deaminated sites from the DNA template molecules (Briggs and Heyn, 2012). For all the described steps, positive and negatives controls were included. The non-UDG library was sequenced in a HiSeq2500 with 2x101+8 cycles as described in (Skoglund et al., 2017). Additional sequencing for the four UDG treated libraries was performed in 3 and a half HiSeq4000 runs in the Max Planck for the Science of Human History and 1 full HiSeq2500 at the University of Kiel.

Screening and authentication

The FAR004 non-UDG and one of the FAR004 UDG-treated libraries were screened with MALT v0.3.6 (Vågene et al., 2018) with --minPercentIdentity 85.0, --minSupportPercent 0.01 and --

topPercent 1.0, using a database built on the basis of the full-nt NCBI database in April 2016. The results were manually inspected in MEGAN6 (Huson et al., 2016), where a large number of hits to the *S. mutans* node were observed. The reads aligning to the node of *S. mutans* were extracted and a histogram of identity was computed with an R script (https://github.com/aidaanva/SmutansThesis). The FAR004 non-UDG treated library was mapped to the *S. mutans* reference UA159 (NC_004350.2) using the EAGER pipeline (Peltzer et al., 2016). Adapters were trimmed and low quality and/or short (<30bp) reads removed with ClipAndMerge v.1.7.6 (Peltzer et al., 2016). The resulting fastq file was mapped to the reference using bwa aln (v0.7.12, Li and Durbin, 2009) with the following parameters: seed length (-l) 16 and a mismatch allowance (-n) of 0.01. Non mapped reads and/or reads with mapping quality lower than 37 were removed using samtools (Li et al., 2009). The duplicates in the bam files were removed using DeDup v.0.12.2 (Peltzer et al., 2016). Coverage statistics were calculated with Qualimap (García-Alcalde et al., 2012). To authenticate the ancient origin of the DNA, we used MapDamage v2.0.6 (Jónsson et al., 2013) to check for the typical deamination patterns present in ancient DNA molecules.

Genome mapping

We combined the raw reads from different sequencing runs per library into a fastq file. After this procedure we had a fastq for each of the 4 UDG libraries: FAR004.A0101, FAR004.A0102, FAR004.A0103 and FAR004.A0104. Each fastq was processed with EAGER the pipeline as described above with the following changes: bwa aln was run with seeding set to 32 and the mismatch to 0.1, to allow less mismatches in the reads since expected mismatches due to damaged bases have been removed. As before, reads not mapped or with mapping quality less than 37 were removed with samtools and duplicated reads were removed with DeDup. Since the reads in the bam files come from different libraries, the fragments contained in the bam files cannot be PCR duplicates, for that reason we merged the resulting bam files after deduplication using samtools merge and we modified the read groups to be the same independently from which library they came from with Picard Tools AddOrReplaceReadGroups (“Picard Tools - By Broad Institute”, https://broadinstitute.github.io/picard/). Mapping statistics were calculated with Qualimap. GATK UnifiedGenotyper (Van der Auwera et al., 2013) was used to call SNPs with the option to ‘emit all sites’ applied.

Modern dataset

We downloaded 194 *S. mutans* genomes present in the NCBI Genome as of 2019-05-29 (See Supplementary Table 1). The modern genomes were sheared in silico into 100 bp reads with a sliding window of 1 bp. The reads for each modern genome were mapped to the *S. mutans* UA159 reference using the EAGER pipeline like the UDG treated libraries of FAR004. In short, bwa aln with a seeding of 32 and a mismatch allowance of 0.1. Reads with mapping quality below 37 were removed and the duplicates were removed using DeDup. A vcf file per modern genome containing SNP calls was produced with GATK UnifiedGenotyper, with the ‘emit all sites’ set, as above.

Phylogenetic, heterozygosity check and principal component analysis

To obtain the final SNP calls for the modern and the ancient strains, we ran MultiVCFAnalyzer v0.85 (https://github.com/alexherbig/MultiVCFAnalyzer, Bos et al., 2014) using the vcf files an input and with the following settings: a minimum of 5 reads should support the call at the position in consideration together with a 90% of allele frequency to be called as a SNP. The snpAlignment.fasta output, containing all variable positions, was used to compute a Maximum Likelihood (ML) tree with RAxML (Stamatakis, 2014) to explore both the phylogenetic placement of the ancient strain in respect to its modern counterparts and the phylogeography of the species. For the tree computation, the General Time Reversible (GTR) model together with six gamma rate categories were used to model the nucleotide change evolution. The tree with the best likelihood was then bootstrapped 1,000 times. To check for potential recombination, we computed a NeighborNet with SplitsTree (Huson and Bryant, 2006). To identify the unique positions for the FAR004 ancient strains, we filtered the snpTable produced by MultiVCFAnalyzer in R (R Development Core Team, 2008), and predicted the effect of the SNPs using SnpEff version 3.1i (Cingolani et al., 2012). We compute a histogram of allele frequencies to assess: (1) potential mismapping from other bacteria present in the sample or (2) the presence of multiple strains in the sample. To do so we calculated the allele frequency by rerunning MultiVCFAnalyzer allowing heterozygous SNPs calls with a frequency between 10-90% (Minimal allele frequency for homozygous call = 0.9 and minimal allele frequency for heterozygous call =0.1) with a minimum of 5 reads covering the position. We plotted the histogram of the SNP frequencies using R and checked the distribution of the allele frequencies to determine if our sample is affected by the two phenomena described above. We performed a principal component analysis (PCA) to check for the presence of a phylogeographic pattern in the *S. mutans* data. The PCA was performed and plotted with R (https://github.com/aidaanva/SmutansThesis) using as input data the genotype matrix produced by MultiVCFAnalyzer. Missing data was imputed based on the mean of all the modern genomes in a specific position.

Virulence genes and indel analysis

All modern and the ancient FAR004 (UDG treated shotgun and capture) *S. mutans* genomes were mapped to the reference genome as mentioned above, however reads with mapping quality equal to 0 were kept for this analysis to account for potential duplication of genes. We compiled a file containing the metadata for the virulence factors of *S. mutans* described in Meng et al. (2017). We generated a bed file containing all the genes of *S. mutans* UA159 and their coordinates. We obtained the bedgraph containing the coverage features from the bam file for each sample using bedtools genomecov with the option -bga (Quinlan and Hall, 2010) to include the missing regions and filtered for those using grep. We then use bedtools intersect with the all genes as targets to obtain any gene with missing data. We computed the percentage of the gene covered using R, selected the virulence factors and plotted a histogram with the percentage of the gene covered for all the samples. For the indel analysis, we filtered the bed file containing all the missing intervals and selected deletions of at least 500bp and plotted those in R. The code can be found in https://github.com/aidaanva/SmutansThesis.

Capture design and comparison with shotgun sequenced genome

The generation of the genome using a shotgun sequencing approach would require a substantial financial investment, which is infeasible when trying to generate multiple genomes for *S. mutans* from a metagenomic context. To provide a cost-effective method to recover *S. mutans* DNA from metagenomic samples, we design an in-solution capture assay for targeted genome enrichment. For the design, a total of 191 modern genomes found in the NCBI Genome database (date: 2018/11/02, described in Supplementary Table 1) were used to compute probes for two arrays for the in-solution capture array as described in Andrades Valtueña et al. (2017). In short, the probes were designed with a 6 bp tiling and a length of 52 bp containing an additional 8 bp 3' linker that has been previously described in (Fu et al., 2013) and regions of low complexity were masked with dustmasker v.2.2.32+ (Camacho et al., 2009). We filtered out probes that were redundant and that had more than 20% of masked regions, resulting in a total of 926,935 unique probes. We designed a second probe set with a coordinate offset of 3 bp in relation to the one described that contained 926,903 unique probes. By combining these two probe sets we obtained an effective tiling density of 3bp. We ordered two 1 million feature Agilent SureSelect DNA Capture Arrays, one for each probe set, filling up the full capacity (968,000 probes) with randomly selected probes. By following the procedure described in Fu et al. (2013), we converted the arrays into in-solution DNA capture libraries to be used to enrich *S. mutans* molecules from DNA extracts. We enriched the UDG-treated library FAR004.A0104 (now referred to as FAR004cap) for *S. mutans* DNA using the in-solution capture probes, following previously described protocols (Fu et al., 2013; Haak et al., 2015; Mathieson et al., 2015). FAR004cap was paired-end sequenced on a HiSeq4000, and obtained a total of 18,230,682 reads that resulted in 8,793,624 reads after adapter clipping and read merging as above. The sample was mapped and variants called as described above for the UDG-treated FAR004 shotgun. We calculated the enrichment factor as:

$$\frac{endogenous\ DNA\ postcapture}{endogenous\ DNA\ precapture}$$

where endogenous DNA is the percentage of the reads mapping to *S. mutans* in the sample calculates as:

$$\frac{number\ of\ reads\ mappint\ to\ S.mutans}{total\ sequenced\ reads} x100$$

To evaluate potential capture bias, we compare the coverage across the reference. We calculated a bedgraph using bedtools genomecov -d, to include coverage for all positions, for the shotgun and the capture genomes. We plotted the coverage across the genome normalised by the mean coverage with the ggplot package in R (See R Notebook in Supplementary material). Additionally, we use bedtools genomecov -bg, to include only covered regions together with bedtools intersect -v to detect any regions without coverage in the capture data compared to the shotgun data and vice versa. We plotted a histogram of the length of the missing fragments with R. We also assessed the profile of present virulence factors to check whether there was any change in the percentage of the gene covered between the shotgun and the capture data (as before, this was performed keeping reads with mapping quality 0). We included the capture data within the previously described run of MultiVCFAnalyzer. Then, we filtered the snpTable to obtain the unique SNPs for the capture genome and looked for inconsistencies in SNP calling between the shotgun

and capture reconstructed genomes. We generated the tree again using RAxML as explained above, and compared the positioning of the two samples in the tree. Additionally, we performed a principal component analysis (PCA) including the capture data to compare the positions in the PC space.

Results

Screening and authentication

We screened two libraries (one non-UDG and one UDG-treated) from a hunter-gatherer individual (FAR004) dating to ~2000 cal BP (Skoglund et al., 2017) using MALT with the NCBI nt database and manual inspection in MEGAN for the presence of pathogens. The metagenomic composition can be seen in Figure 1. The majority of the reads are classified as *Homo sapiens* (Fig. 1A), illustrating the exceptional DNA preservation of the host DNA as already shown in Skoglund et al. (2017). When excluding eukaryotic sequences, we observe a signal that corresponds to a typical human oral flora (Fig. 1B), including species such as *Actinomyces* sp. oral taxon 414, *S. mutans* and *Porphyromonas gingivalis*. Among the oral bacteria, the second most abundant hit is *S. mutans* with a total of 1,844 (0.3% assigned reads) and 80,925 (0.4% assigned reads) assigned reads in the FAR004 nonUDG and FAR004 UDG libraries respectively. To verify the presence of *S. mutans* in the sample, we first calculated the percent identity histogram for the reads assigned to the *S. mutans* node (Fig. 1C). We observed that the majority of the reads have a 100% identity and a decreasing edit distance towards the 85% identity. The shape of the distribution shows that the majority of reads are almost identical to the reference set of *S. mutans* strains in our MALT database. Such a declining edit distance is indicative of a correct taxonomic assignment in ancient microbial studies (Warinner et al., 2017). In order to check if the *S. mutans* reads were of genuine ancient origin, we mapped the reads from the non-UDG treated library, which still contains damaged (deaminated) cytosines, to the reference genome *S. mutans* UA159 (NC_004350). After clipping the adapters, merging and quality filtering, we had a total of 1,503,374 reads from which we recovered an *S. mutans* genome with a mean depth coverage of 0.04 (Table 1). Using mapDamage, we calculated the proportion of the substitutions along the positions in the read. We observed a damage pattern characteristic of ancient data (Briggs et al., 2007), with an increase C->T substitution in the 5’ end and G->A substitutions in the 3’ end towards the end of the fragment that are due to cytosine deamination from hydrolytic damage (Fig1 D).

Genomic reconstruction

To reconstruct the genome of *S. mutans*, we generated a total of four UDG-treated libraries for the FAR004 individual. This process repairs damaged fragments by trimming off sequences containing deaminated sites (See methods), which reduces the probability of potential false SNP calls due to damaged bases. The results from the mapping (see methods for a detailed description of the procedure) are summarised in Table 1. We reconstructed a *S. mutans* genome with a mean fold coverage of 175.51 with 90.07% of the reference genome (i.e. *S. mutans* UA159, NC_004350.2) covered at least 5-fold. In order to detect cross-mapping from related taxa, or

strain mixing, we called single nucleotide polymorphisms (SNPs, see methods) and plotted the allele frequency per SNP (Figure 1E). We observed that in the FAR004 genome, the majority of called SNPs were reference calls (0), followed by alternative calls (90-100). We observed a bimodal distribution of the calls with an allele frequency between 10-99.99. This distribution has two peaks: one around 90% and another one at around 10%, mirroring each other. This indicates that we have a mixture of strains with a major strain representing 90% of our sample and a minor strain representing around 10%. In order to reconstruct only the major strain, we did not allow for heterozygous calls and called SNPs with a minimum of 90% of the reads confirming the base call. In the PCA and phylogenetic analysis used to test for phylogeography, we will be only considering the major strain.

Virulence potential of the FAR004 strain

To investigate the virulence potential of FAR004 in comparison with its modern counterparts, we looked for the presence of 93 known virulence genes based on Meng et al. (2017). The results from the evaluation of presence/absence of the 93 virulence genes is represented in Figure 2. The FAR004 strain lacks genes related to the two-component system (TCS) which are involved in the acid tolerance (*scnK* and its regulator *scnR*) or damage recognition (*SMU45* and *SMU46*). scnK has been associated with a resistance to hydrogen peroxide and phagocytosis by macrophages (Chen et al., 2008). *SMU45/46* is a recently discovered two component-system system that seems to be involved in recognition of damaging agents thus aiding the bacteria during the stress response (Biswas et al., 2008). Additionally, FAR004 is also missing the *SMU1913c* and *SMU1914c* genes which are involved in the mutacin and mutacin immunity pathway. *SMU1914c* or *cipB* encodes the mutacin-V while *SMU1913c* encodes for the mutacin immunity protein of *cipB*. *S. mutans* possess an additional immunity protein for *SMU1914c*, *cipI* or *SMU925*, which seems to be the main immunity protein in some strains. It has been proposed that *SMU925* has emerged as a duplication of *SMU1913c* and later became the main immunity protein (Song et al., 2013). However, we observe the presence of *SMU925* in our ancient strain and other modern strains where *SMU1913c* and *SMU1914c* are not present, thus suggesting that *SMU925* is present in the genome without the presence of the mutacin-V gene. Further ancient genomes could help elucidate whether *SMU1913c* is ancestral to *SMU925*. We observe no structure in the presence/absence of virulence factors related to the isolation place of the *S. mutans* strains. Our ancient individual is indistinguishable from existing diversity in modern strains based on the presence or absence of these virulence genes (Figure 2), indicating that the ancient strain has the same virulence potential as some modern strains such as *S. mutans* GS 5, which has been isolated from a caries lesion and shown to cause caries and alveolar bone loss in gnotobiotic rats (Gibbons et al., 1966).

We also checked for insertions/deletions (indels) of a minimum size of 500 bp and observed no relationship between geographic provenance and regions missing (Supplementary Figure 1). We observe a multitude of the indels differing in size and location across the reference highlighting the great diversity of *S. mutans* and its small core genome as previously described (Meng et al., 2017). The ancient genome does not have any unique indel when compared with the modern genomes. These observations together with the virulence factors results is indicative of a very old

diversity in *S. mutans*. Further ancient genomes would be needed to determine how old the modern diversity is.

In order to look more into the potential differences in phenotype of the ancient genome, we detect a total of 1,501 unique SNPs in our strain FAR004. For those SNPs, we predict a total of 1,840 potential effects affecting 905 genes using SnpEff. The predicted effects were distributed as follows (Figure 3A): synonymous (36.58%), non-synonymous coding (30.60%), intragenic (13.97%), downstream (9.62%), upstream (8.53%), stop gain (0.54%) and synonymous stop (0.16%). From these effects, the most drastic changes would be in the genes where a 'stop' is gained. Stop gained effects are predicted for 9 genes and are summarised in Figure 3B. Most of these genes are annotated as hypothetical proteins or have not been included in phenotypic studies, and thus we are unable to predict their effect in the virulence of the strain. We detect a stop mutation in *SMU423* (*nlmD*) which encodes for the bacteriocin mutacin VI. The FAR004 strain could be less competitive than other *S. mutans* strains with other species in the biofilm since it is missing two bacteriocins, mutacin V (*SMU1914c*) and mutacin IV (*nlmD*). However, physiological tests are necessary to confirm this hypothesis.

Testing the phylogeographic signal in *S. mutans*

In order to assess if there is a correlation between *S. mutans* strain genetic relationships and their geographic location, we performed two analyses: a phylogenetic reconstruction and PCA. We analysed 193 modern *S. mutans* genomes together with the reconstructed ancient genome FAR004, and we included an additional 1 genome from the closest relative species - *S. troglodytae,* described in Okamoto et al., (2016) - to act as an outgroup (See Supplementary Table 1 for additional information on the strains used). We excluded the modern *S. mutans* NCTC10920 genome from the analyses, denoted as an *S. mutans* strain isolated from a rat in 1960, due to low coverage when mapped to the reference with 0.2746X coverage and only 0.9% of the genome coverage at 1X. This genome has now been removed from the NCBI RefSeq and labelled as an anomalous assembly. Furthermore, the culture is labelled as *Streptococcus ratti* in the culture collection of the Public Health England (https://www.phe-culturecollections.org.uk/products/bacteria/detail.jsp?collection=nctc&refId=NCTC+10920) which confirms that this genome is not an *S. mutans*, but rather its close relative *S. ratti*, which was mislabelled when uploaded into the NCBI RefSeq database.
We computed a Maximum Likelihood (ML) tree using RAxML, excluding all sites with missing data. We observed very low bootstrap support for the internal branches of the tree (Figure 4A) which is likely due to extensive recombination that has been described for *S. mutans* (Maruyama et al., 2009). We computed a NeighborNet network and observed a strong reticulation in the area close to the root (in the centre of the graph, Supplementary Figure 2) which supports the exchange of material. With the exception of a few clades, little reticulation was observed in the tips. This can be due to a lack of recombination in the terminal branches, or because of the difficulty to detect homoplasies that occur in the recent past. Since little recombination is observed in the terminal branches, we assume that their relationship is correct. With this assumption, we coloured the tree based on location (i.e. continent) in order to detect a phylogeographic structure (Figure 4A). We do not observe a clear phylogeographic signal.

To further assess the phylogeographic distribution of *S. mutans*, we also performed a PCA based on all SNP calls (Figure 5). Similar to what we observed in the phylogenetic tree, there is no obvious clustering of the samples based on geography, as already observed in Cornejo et al., (2013). We suspect that the lack of phylogeographic signal is due to the modern dataset. A phylogeographic pattern for *S. mutans* is expected due to its transmission being predominantly vertical (da Silva Bastos et al., 2015; Lapirattanakul and Nakano, 2014; Li and Caufield, 1995). This is also the case for other pathobionts associated with humans, such as *Helicobacter pylori*, where a strong phylogeographic signal is observed (Mégraud et al., 2016). Notably, a different data collection strategy was applied to acquire the modern reference datasets for these two bacterial species. The modern dataset of *H. pylori* has been collected taking in consideration two criteria: known ancestry of the donor and a comprehensive world-wide collection strategy. In the case of *S. mutans*, the dataset has been assembled in a more opportunistic manner, whereby strains sequenced across multiple studies, mainly for the purpose of studying virulence, and with little associated metadata were simply compiled together. For that reason, the modern dataset has an uneven world-wide distribution, with 65.4% of the strains sampled from the Americas, and Brazil alone representing the 46.9% of the dataset. Furthermore, since little to no information about the donor was collected for the modern strains, we cannot take into account the ancestry of the donor, which is an important variable when seeking to understand the phylogeographic structure of recombining bacteria such as *S. mutans*. Recent globalization together with past events, such as colonialism, have created numerous opportunities for *S. mutans* strains to recombine, thus dampening the phylogeographic signal.

Capture design and evaluation of its performance

The capture was based on 191 *S. mutans* genomes indicated in Supplementary Table 1. We designed 2 probe sets with a 3bp offset between them, and after filtering of low complexity regions and duplicated probes the final probe sets consisted of 926,935 and 926,903 unique probes respectively (See methods for more details). In order to check for the specificity of the designed probes, we performed MALT analysis with the full-nt dataset as described in the methods. The majority of the probes (90,99%) are assigned to the *S. mutans* species node or beyond such as strain level (Figure 6A). 63,578 reads (4%) are classified as *S. troglodytae*, however this is not unexpected since this species is the closest relative to *S. mutans*. Overall, these results show that the capture is specific to *S. mutans*. In order to test the efficiency of this new capture, we captured the FAR004 sample and compared it with the shotgun genome. We calculated a capture efficiency of 199.24-fold, increasing the endogenous DNA from ~0.18% in shotgun to 35.836% after capture. We retrieved a ~34X genome from ~8,793,624 sequenced reads in comparison to ~1,340,046,653 reads needed from shotgun data to recover a similarly covered genome which showcases the designed capture as a cost-effective technique to recover ancient *S. mutans* DNA. To assess any capture bias, we compared the coverage across the reference to detect regions in which coverage differs between shotgun and capture data using bedtools genomecov and intersect. When performing side-by-side visualisation of the coverage across the reference (Figure 6B), we observe similar trends in both the capture and the shotgun. Even though we detected differences in coverage with bedtools intersect (9429 intervals, 9,386 not covered in the

capture and 43 not covered in the shotgun genome), the majority are less than 10 bp long intervals (9,038) with only 1 interval being 107 bp (Figure 6C).

Additionally, we did not observe any significant difference in the coverage for the *S. mutans* virulence factors (Figure 2). When comparing the overall SNP calls, we observe that shotgun and capture genomes only differ with missing data at a total of 8,568 positions affected (7,145 positions called in the shotgun and 1,423 called in the capture). The missing data is either due to the position not being covered with a minimum of 5 reads (threshold set to call SNPs) or with the position having an allele frequency lower than 90% (we referred to this as heterozygous call). These results are summarised in Table 2. We checked the specific SNP calls for the shotgun and capture data, and found a difference in the number of specific positions called (1501 vs 1400). As with the overall calls, we did not observe a difference other than missing data and resulting from the SNPs not meeting the thresholds used for SNP calling as explained above (Table 2). We compared both the positioning on the tree and in the PCA of the shotgun data versus the capture data. In the tree, both genomes are identical with no extra positions, since it was computed based on a complete deletion alignment (Figure 4B). In the PCA space, the shotgun and capture data fall closer together, however they do not overlap (Figure 5B). We hypothesise that this is due to the missing data described above. To test this hypothesis, we performed the PCA including sites with 50% missing data, 90% missing data or only with sites without missing data (equivalent to the complete deletion performed for the phylogeny building). We show that in the case of complete deletion both the capture and the shotgun overlap completely, however, once we start including missing data the shotgun and the capture genomes start differing in their position (Figure 7). This indicates that our hypothesis was correct and that the reason why the shotgun and capture genomes differ in their position in our initial PCA (Figure 5) is due to missing data.
Overall, these results show that the capture efficiently targets and enriches for *S. mutans* DNA and there is no observed bias in the capture data, when performing standard analysis done in ancient pathogenomics work when using an alignment containing no missing data.

The reconstruction of a large number of ancient *S. mutans* genomes from shotgun data requires an incredible economic investment that most labs cannot afford. We present the designed capture as an alternative to shotgun genome generation that is cost-effective thus allowing for the gathering of an ancient *S. mutans* dataset with a world-wide distribution, as well as, sampled through time. Even considering the economic advantage of the capture, though, one should keep in mind its limitations. *S. mutans* has been shown to have an open pan-genome. Meng et al. (2017) showed in their pan-genome analysis based on 183 strains that the size of the pangenome increased steadily without reaching a plateau and estimated that each genome added will contribute with 5 new genes. The designed capture will represent the pangenome of the 191 strains used for its design, thus limiting our analysis of the pangenome to this representation.

Discussion

Here we have recovered the first ancient *S. mutans* genome, a bacterium highly involved in dental disease, from a 2000-year-old archaeological individual with carious lesions. This genome presents similar patterns to modern *S. mutans* genomes in terms of the presence and absence of known virulence factors. This indicates that a similar pathogenicity potential was already present in strains already 2000 years ago, and highlights the long association of this bacterium with its human host. Being able to use *S. mutans* phylogeographic signal as a proxy to infer human past migrations would be of interest in cases where the study of human DNA is not possible due to ethical concerns, such as social implications of the conducting and outcome of the research. Due to its vertical inheritance, it has been proposed that *S. mutans* should hold information on the dispersal of its host, *Homo sapiens*. We did not find evidence of phylogeography; reproducing the results of Cornejo et al., (2013). This could be due to a more complex inheritance pattern in the bacterium since horizontal transmission of *S. mutans* has been previously observed (see for example, Alves et al., 2009; Baca et al., 2012; Douglass et al., 2008; Spolidorio et al., 2003). However, the current modern dataset has been assembled in an opportunistic manner, and recent human admixture events could have led to the admixture of highly divergent *S. mutans* strains, and thus dilutes phylogeographic signals. In order to properly test for phylogeography, the dataset must be compiled using a strategic sampling to obtain a world-wide diversity of *S. mutans* and to avoid signal loss due to recent human admixture events. Ancient DNA is the ideal tool for generating such a dataset, since it allows for the recovery of strains prior to admixture events due to colonialism but also allows for the exploitation of changes in the bacterium through time. Furthermore, the fact that we can now obtain time transects in interesting periods where we have archaeological evidence of changes in human behaviour, such as introduction of agriculture or dentistry practises, opens the possibility to study the co-evolution in real time of this pathobiont with its human host. This will have implications for understanding health and disease of long-term cultural change in modern society. Furthermore , we can also learn about the demographic history of *S. mutans* by utilising methods that account for a temporal signal in the data such as the Bayesian framework implemented in BEAST (Drummond et al., 2012), providing more robust estimations for the demographic parameters, such as population size of the bacterium that can inform in the spreading of the bacterium in human populations and provide a resource for epidemiological modelling in modern contexts.

Despite the benefits from ancient DNA, it is also common to observe a low endogenous DNA for the species of interest (Bos et al., 2019), which poses a challenge to reconstruct genomes from shotgun data without a considerable financial investment. Capture techniques have been employed in other ancient DNA studies to provide a cost-effective alternative (See for example: Bos et al., 2011; Fu et al., 2013; Haak et al., 2015; Maixner et al., 2016; Vågene et al., 2018). To aid the recovery of the large sample sizes required for this type of study, we designed a probe set to target and enrich S. *mutans* reads from metagenomic samples. We have shown here that it is possible to obtain *S. mutans* from archaeological material and the high efficiency of the capture in recovering this bacterium from the metagenomic context of an ancient DNA sample. The generation of a large ancient *S. mutans* dataset has to be combined with the collection of extensive metadata such as radiocarbon dates, location, subsistence strategy of the population, known cultural practises such as tooth picking and brushing. This will allow for hypothesis testing

such as how *S. mutans* adapted to the introduction of agriculture in a finer scale than it is now possible.

Summary

Here we recovered the first ancient genome of *S. mutans* from a hunter-gathered individual from South Africa that lived between 2017-1748 cal years BP. We show that this strain has a genetic repertoire of virulence factors and indels that fit within the modern diversity of *S. mutans*. This suggests that the modern diversity in terms of presence/absence of virulence factors and indels is rather old; however, we cannot determine if certain virulent factors/indels became more prevalent after humans changed their subsistence strategy since we currently only have one data point from the past. We also failed to detect a phylogeographic signal in the *S. mutans* data, and question that this could be detected with the current available modern dataset. Globalisation and colonialism which has contributed to diverse human populations across the world, thus diluting the phylogeographic signal in *S. mutans*. Ancient DNA can be used as a tool to explore the diversity and phylogeography of *S. mutans* prior to these events. It also allows for temporal sampling during key changes in human behaviour, such as subsistence and cultural practises. This opens the possibility to study the co-evolution of this pathobiont and its human host. This knowledge could inform modern strategies to treat carious lesions associated with *S. mutans*. The large sample size required to answer these questions would require a considerable financial investment given the common problem in ancient DNA, where the species of interest usually represents a small fraction of the DNA recovered. To facilitate the recovery of *S. mutans* from archaeological material in a cost-effective manner, we designed a *S. mutans* capture to target for this pathogen DNA. The designed capture efficiently recovers ancient *S. mutans* DNA and we observe no bias in the data recovered. We hope to apply this designed capture to more individuals from the past. Applying the analytical methods established in this paper to a large dataset will shed some light into the population history of *S. mutans*; how its repertoire of virulence genes came to be by testing whether there is an association between virulence genes prevalence and subsistence or behavioural changes in human populations.

References

Aas, J.A., Griffen, A.L., Dardis, S.R., Lee, A.M., Olsen, I., Dewhirst, F.E., Leys, E.J., Paster, B.J., 2008. Bacteria of Dental Caries in Primary and Permanent Teeth in Children and Young Adults. J. Clin. Microbiol. 46, 1407–1417. https://doi.org/10.1128/JCM.01410-07

Adler, C.J., Dobney, K., Weyrich, L.S., Kaidonis, J., Walker, A.W., Haak, W., Bradshaw, C.J.A., Townsend, G., Sołtysiak, A., Alt, K.W., Parkhill, J., Cooper, A., 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat. Genet. 45, ng.2536. https://doi.org/10.1038/ng.2536

Ajdić, D., McShan, W.M., McLaughlin, R.E., Savić, G., Chang, J., Carson, M.B., Primeaux, C., Tian, R., Kenton, S., Jia, H., Lin, S., Qian, Y., Li, S., Zhu, H., Najar, F., Lai, H., White, J., Roe, B.A., Ferretti, J.J., 2002. Genome sequence of *Streptococcus mutans* UA159, a

cariogenic dental pathogen. Proc. Natl. Acad. Sci. 99, 14434–14439. https://doi.org/10.1073/pnas.172501299

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. Curr. Biol. 27, 3683-3691.e8. https://doi.org/10.1016/j.cub.2017.10.025

Argimón, S., Caufield, P.W., 2011. Distribution of Putative Virulence Genes in *Streptococcus mutans* Strains Does Not Correlate with Caries Experience. J. Clin. Microbiol. 49, 984–992. https://doi.org/10.1128/JCM.01993-10

Argimón, S., Konganti, K., Chen, H., Alekseyenko, A.V., Brown, S., Caufield, P.W., 2014. Comparative genomics of oral isolates of *Streptococcus mutans* by in silico genome subtraction does not reveal accessory DNA associated with severe early childhood caries. Infect. Genet. Evol. 21, 269–278. https://doi.org/10.1016/j.meegid.2013.11.003

Bender, G.R., Thibodeau, E.A., Marquis, R.E., 2016. Reduction of Acidurance of Streptococcal Growth and Glycolysis by Fluoride and Gramicidin: J. Dent. Res. https://doi.org/10.1177/00220345850640021701

Biswas, I., Drake, L., Erkina, D., Biswas, S., 2008. Involvement of Sensor Kinases in the Stress Tolerance Response of *Streptococcus mutans*. J. Bacteriol. 190, 68–77. https://doi.org/10.1128/JB.00990-07

Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., Campbell, T.J., Majander, K., Wilbur, A.K., Guichon, R.A., Wolfe Steadman, D.L., Cook, D.C., Niemann, S., Behr, M.A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J.E., Gagneux, S., Stone, A.C., Krause, J., 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514, 494–497. https://doi.org/10.1038/nature13591

Bos, K.I., Kühnert, D., Herbig, A., Esquivel-Gomez, L.R., Valtueña, A.A., Barquera, R., Giffin, K., Kumar Lankapalli, A., Nelson, E.A., Sabin, S., Spyrou, M.A., Krause, J., 2019. Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens. Annu. Rev. Microbiol. 73, 639–66. https://doi.org/10.1146/annurev-micro-090817-062436.

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. https://doi.org/10.1038/nature10549

Briggs, AdrianW., Heyn, P., 2012. Preparation of Next-Generation Sequencing Libraries from Damaged DNA, in: Shapiro, B., Hofreiter, M. (Eds.), Ancient DNA, Methods in Molecular Biology. Humana Press, pp. 143–154. https://doi.org/10.1007/978-1-61779-516-9_18

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., Pääbo, S., 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. 104, 14616–14621. https://doi.org/10.1073/pnas.0704665104

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Caufield, P.W., Saxena, D., Fitch, D., Li, Y., 2007. Population Structure of Plasmid-Containing Strains of *Streptococcus mutans*, a Member of the Human Indigenous Biota. J. Bacteriol. 189, 1238–1243. https://doi.org/10.1128/JB.01183-06

Chen, P.-M., Chen, H.-C., Ho, C.-T., Jung, C.-J., Lien, H.-T., Chen, J.-Y., Chia, J.-S., 2008. The two-component system ScnRK of *Streptococcus mutans* affects hydrogen peroxide resistance and murine macrophage killing. Microbes Infect. 10, 293–301. https://doi.org/10.1016/j.micinf.2007.12.006

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6, 80–92. https://doi.org/10.4161/fly.19695

Clarke, J.K., 1924. On the Bacterial Factor in the aetiology of Dental Caries. Br. J. Exp. Pathol. 5, 141–147.

Cohen, M.N., Armelagos, G.J., 1984. Paleopathology and the origins of agriculture. Academic Press, Orlando (FL).

Cornejo, O.E., Lefébure, T., Bitar, P.D.P., Lang, P., Richards, V.P., Eilertson, K., Do, T., Beighton, D., Zeng, L., Ahn, S.-J., Burne, R.A., Siepel, A., Bustamante, C.D., Stanhope, M.J., 2013. Evolutionary and Population Genomics of the Cavity Causing Bacteria *Streptococcus mutans*. Mol. Biol. Evol. 30, 881–893. https://doi.org/10.1093/molbev/mss278

da Silva Bastos, V. de A., Freitas-Fernandes, L.B., Fidalgo, T.K. da S., Martins, C., Mattos, C.T., de Souza, I.P.R., Maia, L.C., 2015. Mother-to-child transmission of *Streptococcus mutans*: A systematic review and meta-analysis. J. Dent. 43, 181–191. https://doi.org/10.1016/j.jdent.2014.12.001

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., Meyer, M., 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. 110, 15758–15763. https://doi.org/10.1073/pnas.1314445110

De La Fuente, C., Flores, S., Moraga, M., 2013. DNA from human ancient bacteria: a novel source of genetic evidence from archaeological dental calculus. Archaeometry 55, 767–778. https://doi.org/10.1111/j.1475-4754.2012.00707.x

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973. https://doi.org/10.1093/molbev/mss075

Fellows Yates, J.A., Drucker, D.G., Reiter, E., Heumos, S., Welker, F., Münzel, S.C., Wojtal, P., Lázničková-Galetová, M., Conard, N.J., Herbig, A., Bocherens, H., Krause, J., 2017. Central European Woolly Mammoth Population Dynamics: Insights from Late Pleistocene Mitochondrial Genomes. Sci. Rep. 7, 17714. https://doi.org/10.1038/s41598-017-17723-1

Formicola, V., 1987. Neolithic transition and dental changes: the case of an Italian site. J. Hum. Evol. 16, 231–239. https://doi.org/10.1016/0047-2484(87)90078-9

Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., Pääbo, S., 2013. DNA analysis of an early modern human from Tianyuan Cave, China. Proc. Natl. Acad. Sci. 110, 2223–2227. https://doi.org/10.1073/pnas.1221359110

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., Conesa, A., 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28, 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Gibbons, R.J., Berman, K.S., Knoettner, P., Kapsimalis, B., 1966. Dental caries and alveolar bone loss in gnotobiotic rats infected with capsule forming streptococci of human origin. Arch. Oral Biol. 11, 549-IN4. https://doi.org/10.1016/0003-9969(66)90220-2

González-Ittig, R.E., Carletto-Körber, F.P.M., Vera, N.S., Jiménez, M.G., Cornejo, L.S., 2016. Population genetic structure and demographic history of *Streptococcus mutans* (Bacteria: Streptococcaceae). Biol. J. Linn. Soc. n/a-n/a. https://doi.org/10.1111/bij.12904

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R.G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S.L., Risch, R., Rojo Guerra, M.A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K.W., Reich, D., 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211. https://doi.org/10.1038/nature14317

Humphrey, L.T., Groote, I.D., Morales, J., Barton, N., Collcutt, S., Ramsey, C.B., Bouzouggar, A., 2014. Earliest evidence for caries and exploitation of starchy plant foods in Pleistocene hunter-gatherers from Morocco. Proc. Natl. Acad. Sci. 111, 954–959. https://doi.org/10.1073/pnas.1318176111

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R., 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput. Biol. 12, e1004957. https://doi.org/10.1371/journal.pcbi.1004957

Huson, D.H., Bryan, D., 2006. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23, 254-267. https://doi.org/10.1093/molbev/msj030

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. https://doi.org/10.1093/bioinformatics/btt193

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

Lapirattanakul, J., Nakano, K., 2014. Mother-to-child transmission of mutans streptococci. Future Microbiol. 9, 807–823. https://doi.org/10.2217/fmb.14.37

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence

Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Caufield, P.W., 1995. The fidelity of initial acquisition of mutans streptococci by infants from their mothers. J. Dent. Res. 74, 681–685. https://doi.org/10.1177/00220345950740020901

Loesche, W.J., 1986. Role of *Streptococcus mutans* in human dental decay. Microbiol. Rev. 50, 353–380.

Lukacs, J.R., 1992. Dental paleopathology and agricultural intensification in South Asia: New evidence from Bronze Age Harappa. Am. J. Phys. Anthropol. 87, 133–150. https://doi.org/10.1002/ajpa.1330870202

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. https://doi.org/10.1126/science.aad2545

Manhire, A., 1993. A report on the excavations at Faraoskop Rock Shelter in the Graafwater district of the South-Western Cape. South. Afr. Field Archaeol. 2, 3–23.

Maruyama, F., Kobata, M., Kurokawa, K., Nishida, K., Sakurai, A., Nakano, K., Nomura, R., Kawabata, S., Ooshima, T., Nakai, K., Hattori, M., Hamada, S., Nakagawa, I., 2009. Comparative genomic analyses of *Streptococcus mutans* provide insights into chromosomal shuffling and species-specific content. BMC Genomics 10, 358. https://doi.org/10.1186/1471-2164-10-358

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E.R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J.L., de Castro, J.M.B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M.A.R., Roodenberg, J., Vergès, J.M., Krause, J., Cooper, A., Alt, K.W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., Reich, D., 2015. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528, 499–503. https://doi.org/10.1038/nature16152

Mégraud, F., Lehours, P., Vale, F.F., 2016. The history of *Helicobacter pylori*: from phylogeography to paleomicrobiology. Clin. Microbiol. Infect. 22, 922–927. https://doi.org/10.1016/j.cmi.2016.07.013

Meng, P., Lu, C., Zhang, Q., Lin, J., Chen, F., 2017. Exploring the Genomic Diversity and Cariogenic Differences of *Streptococcus mutans* Strains Through Pan-Genome and Comparative Genome Analysis. Curr. Microbiol. 74, 1200–1209. https://doi.org/10.1007/s00284-017-1305-z

Meyer, M., Kircher, M., 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. 2010, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Moynihan, P., Petersen, P.E., 2004. Diet, nutrition and the prevention of dental diseases. Public Health Nutr. 7, 201–226. https://doi.org/10.1079/phn2003589

Nicklisch, N., Ganslmeier, R., Siebert, A., Friederich, S., Meller, H., Alt, K.W., 2016. Holes in teeth – Dental caries in Neolithic and Early Bronze Age populations in Central Germany. Ann.

Anat. - Anat. Anz., SI: Dental Morphology Research - Past meets Present 203, 90–99. https://doi.org/10.1016/j.aanat.2015.02.001

Okamoto, M., Naito, M., Miyanohara, M., Imai, S., Nomura, Y., Saito, W., Momoi, Y., Takada, K., Miyabe-Nishiwaki, T., Tomonaga, M., Hanada, N., 2016. Complete genome sequence of *Streptococcus troglodytae* TKU31 isolated from the oral cavity of a chimpanzee (Pan troglodytes). Microbiol. Immunol. 60, 811–816. https://doi.org/10.1111/1348-0421.12453

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

Peres, M.A., Sheiham, A., Liu, P., Demarco, F.F., Silva, A.E.R., Assunção, M.C., Menezes, A.M., Barros, F.C., Peres, K.G., 2016. Sugar Consumption and Changes in Dental Caries from Childhood to Adolescence. Journal of Dental Research 95, 388–394. https://doi.org/10.1177/0022034515625907

Picard Tools - By Broad Institute [WWW Document], n.d. URL http://broadinstitute.github.io/picard/index.html (accessed 3.12.20).

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sälzer, S., Alkilzy, M., Slot, D.E., Dörfer, C.E., Schmoeckel, J., Splieth, C.H., 2017. Socio-behavioural aspects in the prevention and control of dental caries and periodontal diseases at an individual and population level. J. Clin. Periodontol. 44, S106–S115. https://doi.org/10.1111/jcpe.12673

Sealy, J.C., Patrick, M.K., Morris, A.G., Alder, D., 1992. Diet and dental caries among later stone age inhabitants of the Cape Province, South Africa. Am. J. Phys. Anthropol. 88, 123–134. https://doi.org/10.1002/ajpa.1330880202

Simón, M., Montiel, R., Smerling, A., Solórzano, E., Díaz, N., Álvarez-Sandoval, B.A., Jiménez-Marín, A.R., Malgosa, A., 2014. Molecular analysis of ancient caries. Proc. R. Soc. B Biol. Sci. 281, 20140586. https://doi.org/10.1098/rspb.2014.0586

Skoglund, P., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J.I., Chiumia, C., Crowther, A., Gomani-Chindebvu, E., Gidna, A.O., Grillo, K.M., Helenius, I.T., Hellenthal, G., Helm, R., Horton, M., López, S., Mabulla, A.Z.P., Parkington, J., Shipton, C., Thomas, M.G., Tibesasa, R., Welling, M., Hayes, V.M., Kennett, D.J., Ramesar, R., Meyer, M., Pääbo, S., Patterson, N., Morris, A.G., Boivin, N., Pinhasi, R., Krause, J., Reich, D., 2017. Reconstructing Prehistoric African Population Structure. Cell 171, 59-71.e21. https://doi.org/10.1016/j.cell.2017.08.049

Song, L., Wang, W., Conrads, G., Rheinberg, A., Sztajer, H., Reck, M., Wagner-Döbler, I., Zeng, A.-P., 2013. Genetic variability of mutans streptococci revealed by wide whole-genome sequencing. BMC Genomics 14, 430. https://doi.org/10.1186/1471-2164-14-430

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

Stoppelaar, J.D. de, Houte, J. van, Dirks, O.B., 1970. The Effect of Carbohydrate Restriction on the Presence of *Streptococcus mutans, Streptococcus sanguis* and Iodophilic Polysaccharide-Producing Bacteria in Human Dental Plaque. Caries Res. 4, 114–123. https://doi.org/10.1159/000259633

Vågene, Å.J., Herbig, A., Campana, M.G., García, N.M.R., 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. Nat. Ecol. Evol. 2, 520. https://doi.org/10.1038/s41559-017-0446-6

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, in: Current Protocols in Bioinformatics. p. 43:11.10.1–11.10.33.

Vos, T., Abajobir, A.A., Abate, K.H., Abbafati, C., Abbas, K.M., Abd-Allah, F., Abdulkader, R.S., Abdulle, A.M., Abebo, T.A., Abera, S.F., Aboyans, V., Abu-Raddad, L.J., Ackerman, I.N., Adamu, A.A., Adetokunboh, O., Afarideh, M., Afshin, A., Agarwal, S.K., Aggarwal, R., Agrawal, A., Agrawal, S., Ahmadieh, H., Ahmed, M.B., Aichour, M.T.E., Aichour, A.N., Aichour, I., Aiyar, S., Akinyemi, R.O., Akseer, N., Lami, F.H.A., Alahdab, F., Al-Aly, Z., Alam, K., Alam, N., Alam, T., Alasfoor, D., Alene, K.A., Ali, R., Alizadeh-Navaei, R., Alkerwi, A., Alla, F., Allebeck, P., Allen, C., Al-Maskari, F., Al-Raddadi, R., Alsharif, U., Alsowaidi, S., Altirkawi, K.A., Amare, A.T., Amini, E., Ammar, W., Amoako, Y.A., Andersen, H.H., Antonio, C.A.T., Anwari, P., Ärnlöv, J., Artaman, A., Aryal, K.K., Asayesh, H., Asgedom, S.W., Assadi, R., Atey, T.M., Atnafu, N.T., Atre, S.R., Avila-Burgos, L., Avokphako, E.F.G.A., Awasthi, A., Bacha, U., Badawi, A., Balakrishnan, K., Banerjee, A., Bannick, M.S., Barac, A., Barber, R.M., Barker-Collo, S.L., Bärnighausen, T., Barquera, S., Barregard, L., Barrero, L.H., Basu, S., Battista, B., Battle, K.E., Baune, B.T., Bazargan-Hejazi, S., Beardsley, J., Bedi, N., Beghi, E., Béjot, Y., Bekele, B.B., Bell, M.L., Bennett, D.A., Bensenor, I.M., Benson, J., Berhane, A., Berhe, D.F., Bernabé, E., Betsu, B.D., Beuran, M., Beyene, A.S., Bhala, N., Bhansali, A., Bhatt, S., Bhutta, Z.A., Biadgilign, S., Bicer, B.K., Bienhoff, K., Bikbov, B., Birungi, C., Biryukov, S., Bisanzio, D., Bizuayehu, H.M., Boneya, D.J., Boufous, S., Bourne, R.R.A., Brazinova, A., Brugha, T.S., Buchbinder, R., Bulto, L.N.B., Bumgarner, B.R., Butt, Z.A., Cahuana-Hurtado, L., Cameron, E., Car, M., Carabin, H., Carapetis, J.R., Cárdenas, R., Carpenter, D.O., Carrero, J.J., Carter, A., Carvalho, F., Casey, D.C., Caso, V., Castañeda-Orjuela, C.A., Castle, C.D., Catalá-López, F., Chang, H.-Y., Chang, J.-C., Charlson, F.J., Chen, H., Chibalabala, M., Chibueze, C.E., Chisumpa, V.H., Chitheer, A.A., Christopher, D.J., Ciobanu, L.G., Cirillo, M., Colombara, D., Cooper, C., Cortesi, P.A., Criqui, M.H., Crump, J.A., Dadi, A.F., Dalal, K., Dandona, L., Dandona, R., Neves, J. das, Davitoiu, D.V., Courten, B. de, Leo, D.D.D., Defo, B.K., Degenhardt, L., Deiparine, S., Dellavalle, R.P., Deribe, K., Jarlais, D.C.D., Dey, S., Dharmaratne, S.D., Dhillon, P.K., Dicker, D., Ding, E.L., Djalalinia, S., Do, H.P., Dorsey, E.R., Santos, K.P.B. dos, Douwes-Schultz, D., Doyle, K.E., Driscoll, T.R., Dubey, M., Duncan, B.B., El-Khatib, Z.Z., Ellerstrand, J., Enayati, A., Endries, A.Y., Ermakov, S.P., Erskine, H.E., Eshrati, B., Eskandarieh, S., Esteghamati, A., Estep, K., Fanuel, F.B.B., Farinha, C.S.E.S., Faro, A., Farzadfar, F., Fazeli, M.S., Feigin, V.L., Fereshtehnejad, S.-

M., Fernandes, J.C., Ferrari, A.J., Feyissa, T.R., Filip, I., Fischer, F., Fitzmaurice, C., Flaxman, A.D., Flor, L.S., Foigt, N., Foreman, K.J., Franklin, R.C., Fullman, N., Fürst, T., Furtado, J.M., Futran, N.D., Gakidou, E., Ganji, M., Garcia-Basteiro, A.L., Gebre, T., Gebrehiwot, T.T., Geleto, A., Gemechu, B.L., Gesesew, H.A., Gething, P.W., Ghajar, A., Gibney, K.B., Gill, P.S., Gillum, R.F., Ginawi, I.A.M., Giref, A.Z., Gishu, M.D., Giussani, G., Godwin, W.W., Gold, A.L., Goldberg, E.M., Gona, P.N., Goodridge, A., Gopalani, S.V., Goto, A., Goulart, A.C., Griswold, M., Gugnani, H.C., Gupta, Rahul, Gupta, Rajeev, Gupta, T., Gupta, V., Hafezi-Nejad, N., Hailu, G.B., Hailu, A.D., Hamadeh, R.R., Hamidi, S., Handal, A.J., Hankey, G.J., Hanson, S.W., Hao, Y., Harb, H.L., Hareri, H.A., Haro, J.M., Harvey, J., Hassanvand, M.S., Havmoeller, R., Hawley, C., Hay, S.I., Hay, R.J., Henry, N.J., Heredia-Pi, I.B., Hernandez, J.M., Heydarpour, P., Hoek, H.W., Hoffman, H.J., Horita, N., Hosgood, H.D., Hostiuc, S., Hotez, P.J., Hoy, D.G., Htet, A.S., Hu, G., Huang, H., Huynh, C., Iburg, K.M., Igumbor, E.U., Ikeda, C., Irvine, C.M.S., Jacobsen, K.H., Jahanmehr, N., Jakovljevic, M.B., Jassal, S.K., Javanbakht, M., Jayaraman, S.P., Jeemon, P., Jensen, P.N., Jha, V., Jiang, G., John, D., Johnson, S.C., Johnson, C.O., Jonas, J.B., Jürisson, M., Kabir, Z., Kadel, R., Kahsay, A., Kamal, R., Kan, H., Karam, N.E., Karch, A., Karema, C.K., Kasaeian, A., Kassa, G.M., Kassaw, N.A., Kassebaum, N.J., Kastor, A., Katikireddi, S.V., Kaul, A., Kawakami, N., Keiyoro, P.N., Kengne, A.P., Keren, A., Khader, Y.S., Khalil, I.A., Khan, E.A., Khang, Y.-H., Khosravi, A., Khubchandani, J., Kiadaliri, A.A., Kieling, C., Kim, Y.J., Kim, D., Kim, P., Kimokoti, R.W., Kinfu, Y., Kisa, A., Kissimova-Skarbek, K.A., Kivimaki, M., Knudsen, A.K., Kokubo, Y., Kolte, D., Kopec, J.A., Kosen, S., Koul, P.A., Koyanagi, A., Kravchenko, M., Krishnaswami, S., Krohn, K.J., Kumar, G.A., Kumar, P., Kumar, S., Kyu, H.H., Lal, D.K., Lalloo, R., Lambert, N., Lan, Q., Larsson, A., Lavados, P.M., Leasher, J.L., Lee, P.H., Lee, J.-T., Leigh, J., Leshargie, C.T., Leung, J., Leung, R., Levi, M., Li, Yichong, Li, Yongmei, Kappe, D.L., Liang, X., Liben, M.L., Lim, S.S., Linn, S., Liu, P.Y., Liu, A., Liu, S., Liu, Y., Lodha, R., Logroscino, G., London, S.J., Looker, K.J., Lopez, A.D., Lorkowski, S., Lotufo, P.A., Low, N., Lozano, R., Lucas, T.C.D., Macarayan, E.R.K., Razek, H.M.A.E., Razek, M.M.A.E., Mahdavi, M., Majdan, M., Majdzadeh, R., Majeed, A., Malekzadeh, R., Malhotra, R., Malta, D.C., Mamun, A.A., Manguerra, H., Manhertz, T., Mantilla, A., Mantovani, L.G., Mapoma, C.C., Marczak, L.B., Martinez-Raga, J., Martins-Melo, F.R., Martopullo, I., März, W., Mathur, M.R., Mazidi, M., McAlinden, C., McGaughey, M., McGrath, J.J., McKee, M., McNellan, C., Mehata, S., Mehndiratta, M.M., Mekonnen, T.C., Memiah, P., Memish, Z.A., Mendoza, W., Mengistie, M.A., Mengistu, D.T., Mensah, G.A., Meretoja, T.J., Meretoja, A., Mezgebe, H.B., Micha, R., Millear, A., Miller, T.R., Mills, E.J., Mirarefin, M., Mirrakhimov, E.M., Misganaw, A., Mishra, S.R., Mitchell, P.B., Mohammad, K.A., Mohammadi, A., Mohammed, K.E., Mohammed, S., Mohanty, S.K., Mokdad, A.H., Mollenkopf, S.K., Monasta, L., Montico, M., Moradi-Lakeh, M., Moraga, P., Mori, R., Morozoff, C., Morrison, S.D., Moses, M., Mountjoy-Venning, C., Mruts, K.B., Mueller, U.O., Muller, K., Murdoch, M.E., Murthy, G.V.S., Musa, K.I., Nachega, J.B., Nagel, G., Naghavi, M., Naheed, A., Naidoo, K.S., Naldi, L., Nangia, V., Natarajan, G., Negasa, D.E., Negoi, R.I., Negoi, I., Newton, C.R., Ngunjiri, J.W., Nguyen, T.H., Nguyen, Q.L., Nguyen, C.T., Nguyen, G., Nguyen, M., Nichols, E., Ningrum, D.N.A., Nolte, S., Nong, V.M., Norrving, B., Noubiap, J.J.N., O'Donnell, M.J., Ogbo, F.A., Oh, I.-H., Okoro, A., Oladimeji, O.,

Olagunju, T.O., Olagunju, A.T., Olsen, H.E., Olusanya, B.O., Olusanya, J.O., Ong, K., Opio, J.N., Oren, E., Ortiz, A., Osgood-Zimmerman, A., Osman, M., Owolabi, M.O., PA, M., Pacella, R.E., Pana, A., Panda, B.K., Papachristou, C., Park, E.-K., Parry, C.D., Parsaeian, M., Patten, S.B., Patton, G.C., Paulson, K., Pearce, N., Pereira, D.M., Perico, N., Pesudovs, K., Peterson, C.B., Petzold, M., Phillips, M.R., Pigott, D.M., Pillay, J.D., Pinho, C., Plass, D., Pletcher, M.A., Popova, S., Poulton, R.G., Pourmalek, F., Prabhakaran, D., Prasad, N.M., Prasad, N., Purcell, C., Qorbani, M., Quansah, R., Quintanilla, B.P.A., Rabiee, R.H.S., Radfar, A., Rafay, A., Rahimi, K., Rahimi-Movaghar, A., Rahimi-Movaghar, V., Rahman, M.H.U., Rahman, M., Rai, R.K., Rajsic, S., Ram, U., Ranabhat, C.L., Rankin, Z., Rao, P.C., Rao, P.V., Rawaf, S., Ray, S.E., Reiner, R.C., Reinig, N., Reitsma, M.B., Remuzzi, G., Renzaho, A.M.N., Resnikoff, S., Rezaei, S., Ribeiro, A.L., Ronfani, L., Roshandel, G., Roth, G.A., Roy, A., Rubagotti, E., Ruhago, G.M., Saadat, S., Sadat, N., Safdarian, M., Safi, S., Safiri, S., Sagar, R., Sahathevan, R., Salama, J., Saleem, H.O.B., Salomon, J.A., Salvi, S.S., Samy, A.M., Sanabria, J.R., Santomauro, D., Santos, I.S., Santos, J.V., Milicevic, M.M.S., Sartorius, B., Satpathy, M., Sawhney, M., Saxena, S., Schmidt, M.I., Schneider, I.J.C., Schöttker, B., Schwebel, D.C., Schwendicke, F., Seedat, S., Sepanlou, S.G., Servan-Mori, E.E., Setegn, T., Shackelford, K.A., Shaheen, A., Shaikh, M.A., Shamsipour, M., Islam, S.M.S., Sharma, J., Sharma, R., She, J., Shi, P., Shields, C., Shifa, G.T., Shigematsu, M., Shinohara, Y., Shiri, R., Shirkoohi, R., Shirude, S., Shishani, K., Shrime, M.G., Sibai, A.M., Sigfusdottir, I.D., Silva, D.A.S., Silva, J.P., Silveira, D.G.A., Singh, J.A., Singh, N.P., Sinha, D.N., Skiadaresi, E., Skirbekk, V., Slepak, E.L., Sligar, A., Smith, D.L., Smith, M., Sobaih, B.H.A., Sobngwi, E., Sorensen, R.J.D., Sousa, T.C.M., Sposato, L.A., Sreeramareddy, C.T., Srinivasan, V., Stanaway, J.D., Stathopoulou, V., Steel, N., Stein, M.B., Stein, D.J., Steiner, T.J., Steiner, C., Steinke, S., Stokes, M.A., Stovner, L.J., Strub, B., Subart, M., Sufiyan, M.B., Sunguya, B.F., Sur, P.J., Swaminathan, S., Sykes, B.L., Sylte, D.O., Tabarés-Seisdedos, R., Taffere, G.R., Takala, J.S., Tandon, N., Tavakkoli, M., Taveira, N., Taylor, H.R., Tehrani-Banihashemi, A., Tekelab, T., Terkawi, A.S., Tesfaye, D.J., Tesssema, B., Thamsuwan, O., Thomas, K.E., Thrift, A.G., Tiruye, T.Y., Tobe-Gai, R., Tollanes, M.C., Tonelli, M., Topor-Madry, R., Tortajada, M., Touvier, M., Tran, B.X., Tripathi, S., Troeger, C., Truelsen, T., Tsoi, D., Tuem, K.B., Tuzcu, E.M., Tyrovolas, S., Ukwaja, K.N., Undurraga, E.A., Uneke, C.J., Updike, R., Uthman, O.A., Uzochukwu, B.S.C., Boven, J.F.M. van, Varughese, S., Vasankari, T., Venkatesh, S., Venketasubramanian, N., Vidavalur, R., Violante, F.S., Vladimirov, S.K., Vlassov, V.V., Vollset, S.E., Wadilo, F., Wakayo, T., Wang, Y.-P., Weaver, M., Weichenthal, S., Weiderpass, E., Weintraub, R.G., Werdecker, A., Westerman, R., Whiteford, H.A., Wijeratne, T., Wiysonge, C.S., Wolfe, C.D.A., Woodbrook, R., Woolf, A.D., Workicho, A., Xavier, D., Xu, G., Yadgir, S., Yaghoubi, M., Yakob, B., Yan, L.L., Yano, Y., Ye, P., Yimam, H.H., Yip, P., Yonemoto, N., Yoon, S.-J., Yotebieng, M., Younis, M.Z., Zaidi, Z., Zaki, M.E.S., Zegeye, E.A., Zenebe, Z.M., Zhang, X., Zhou, M., Zipkin, B., Zodpey, S., Zuhlke, L.J., Murray, C.J.L., 2017. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet 390, 1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2

Warinner, C., 2016. Dental Calculus and the Evolution of the Human Oral Microbiome. J. Calif. Dent. Assoc. 44, 411–420.

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. Annu. Rev. Genomics Hum. Genet. 18, 321–356. https://doi.org/10.1146/annurev-genom-091416-035526

Warinner, C., Matias Rodrigues, J.F., 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., Samaniego Castruita, J.A., Barkow-Oesterreicher, S., Teoh, K.Y., Kelstrup, C., Olsen, J.V., Nanni, P., Kawai, T., Willerslev, E., von Mering, C., 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. Nat. Genet. 46, 336–344. https://doi.org/10.1038/ng.2906

Willmann, C., Mata, X., Hanghoej, K., Tonasso, L., Tisseyre, L., Jeziorski, C., Cabot, E., Chevet, P., Crubézy, E., Orlando, L., Esclassan, R., Thèves, C., 2018. Oral health status in historic population: Macroscopic and metagenomic evidence. PLoS ONE 13. https://doi.org/10.1371/journal.pone.0196482

Figures and Tables

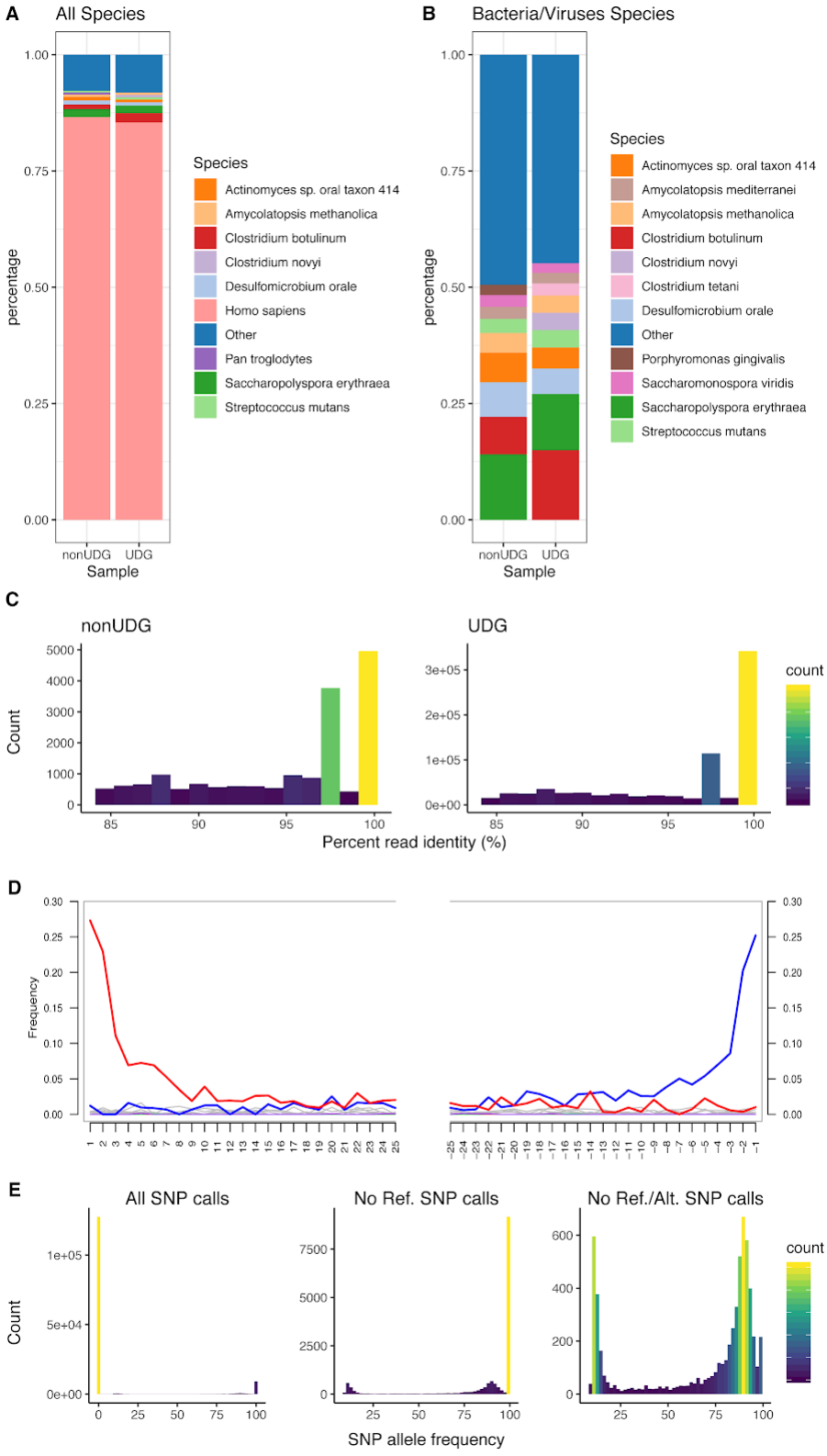


Figure 1: Assessment of metagenomic reads and authenticity of the *S. mutans* signal in FAR004. Metagenomic composition of the FAR004 sample including (A) all species in the full nt-database; (B) only bacterial hits. (C) Distribution of the percent identity of the reads classified as *S. mutans* by MALT. (D) Deamination patterns of FAR004 produced by MapDamage2. Red displays the C->T substitutions, blue the G->A substitutions and in grey all the other possible substitutions (E) Distribution of the allele frequencies for the SNP calls of FAR004. Reference calls have an allele frequency of 0, while alternative calls have a frequency between 90-100% (determined by our threshold to call a SNP, see methods) and heterozygous calls have a frequency between 10-90%. This plot was generated using ggplot and ggarrange in R.

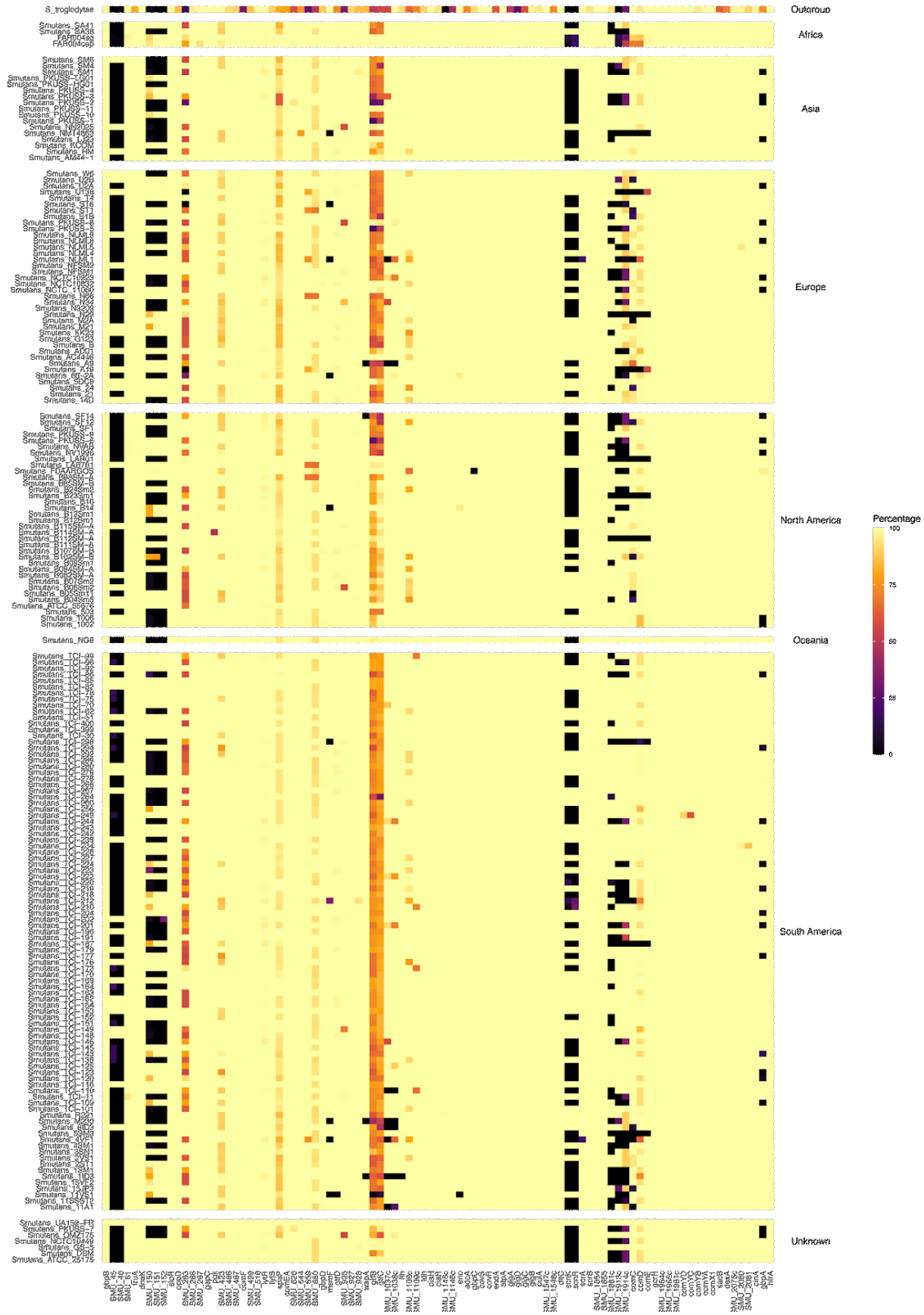
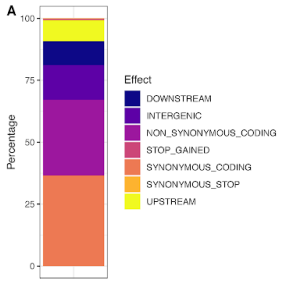

Outgroup
Africa
Asia
Europe
North America
Oceania
South America
Unknown
Percentage
100
75
50
25
0

Figure 2: Presence/absence profile of 93 described virulence factors of modern and ancient (FAR004) *S. mutans* strains. The gradient indicates the percent of the gene covered at least 1X being black 0% and yellow 100% covered. The capture and shotgun reconstructed genomes for the FAR004 strains are referred to as FAR004cap and FAR004sg respectively. This plot was generated using ggplot in R.

A



B

Ref. Position	Ref. SNP	FAR004 SNP	Gene	Product	Protein ID	AA Ref./ AA FAR004	Codon Ref./ Codon FAR004	Codon number	CDS Size
171960	A	C	SMU_174c	hypothetical protein	NP_720643.1	L/*	tTa/tGa	195	791
172042	G	A	SMU_174c	hypothetical protein	NP_720643.1	Q/*	Caa/Taa	168	791
324739	C	T	SMU_343	hypothetical protein	NP_720796.1	Q/*	Caa/Taa	166	521
396663	G	T	SMU_423	hypothetical protein	NP_720871.1	G/*	Gga/Tga	45	230
480347	G	A	SMU_512c	hypothetical protein	NP_720950.1	Q/*	Cag/Tag	132	455
596649	C	T	SMU_637c	hypothetical protein	NP_721065.1	W/*	tgG/tgA	56	503
926691	C	T	bglB1	BglB fragment	NP_721376.1	Q/*	Caa/Taa	53	887
944287	C	T	SMU_997	inorganic ion ABC transporter ATP-binding protein	NP_721392.1	Q/*	Caa/Taa	256	806
1191984	C	T	SMU_1259	restriction endonuclease	NP_721635.1	R/*	Cga/Tga	104	326
1397861	C	A	SMU_1471c	hypothetical protein	NP_721821.1	E/*	Gaa/Taa	75	899

Figure 3: Summary of the effects found by SnpEff from the 1,501 unique SNPs from the ancient strains FAR004. (A) Percentage distribution of the effects detected by SnpEff. (B) Table containing predicted Stop gain effects containing the SNP position and the gene affected by a stop.


A
B
S_troglodytae
S_troglodytae

Figure 4: Maximum Likelihood (ML) tree inferred from 193 modern *S. mutans* strains, the ancient *S. mutans* FAR004 strain and *S. troglodytae* as an outgroup. The strains have been coloured based on the continent where they were isolated. (A) Only with the FAR004 shotgun (FAR004sg) reconstructed genome, (B) including both shotgun (FAR004sg) and capture (FAR004cap) genomes. The ancient genomes are inside a red-box. The bootstraps are indicated with the bubbles in the nodes, coloured with a gradient with red being 100% support and blue is 0%, and the size of the bubbles also indicating support the bigger the bubble the closer to 100% support.

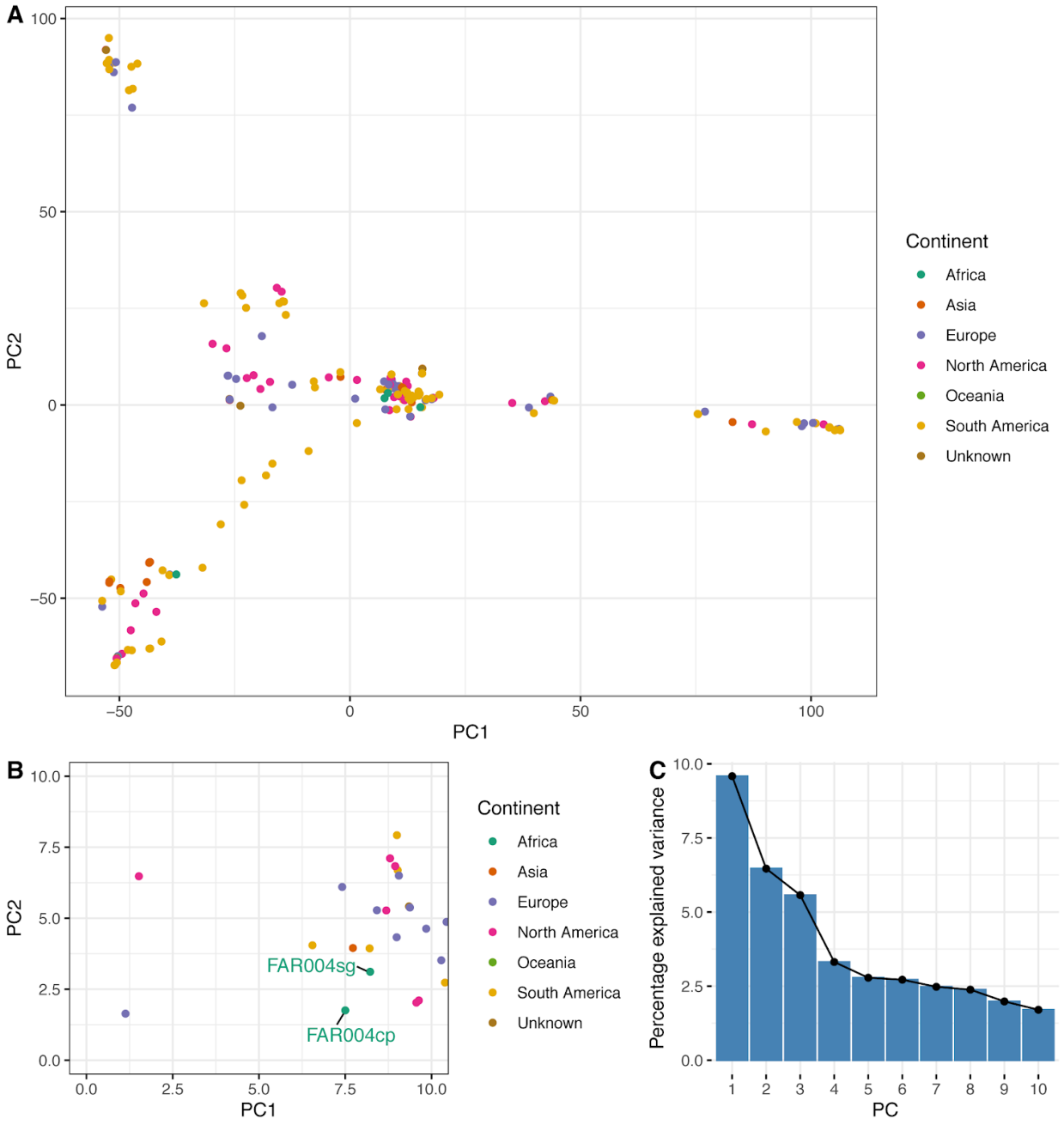


Figure 5: Principal component analysis (PCA) calculated using all variable sites (120,746 positions) including 193 modern *S. mutans* strains and including both the capture (FAR004cp) and the shotgun (FAR004sg) genomes reconstructed for the FAR004 ancient *S. mutans* strain. (A) PCA plot with the strains coloured based on the isolation continent. (B) Zoom in of plot A to display the positioning of the shotgun

and capture FAR004 reconstructed genomes. (C) Percentage of variance explained by the different principal components (PC).

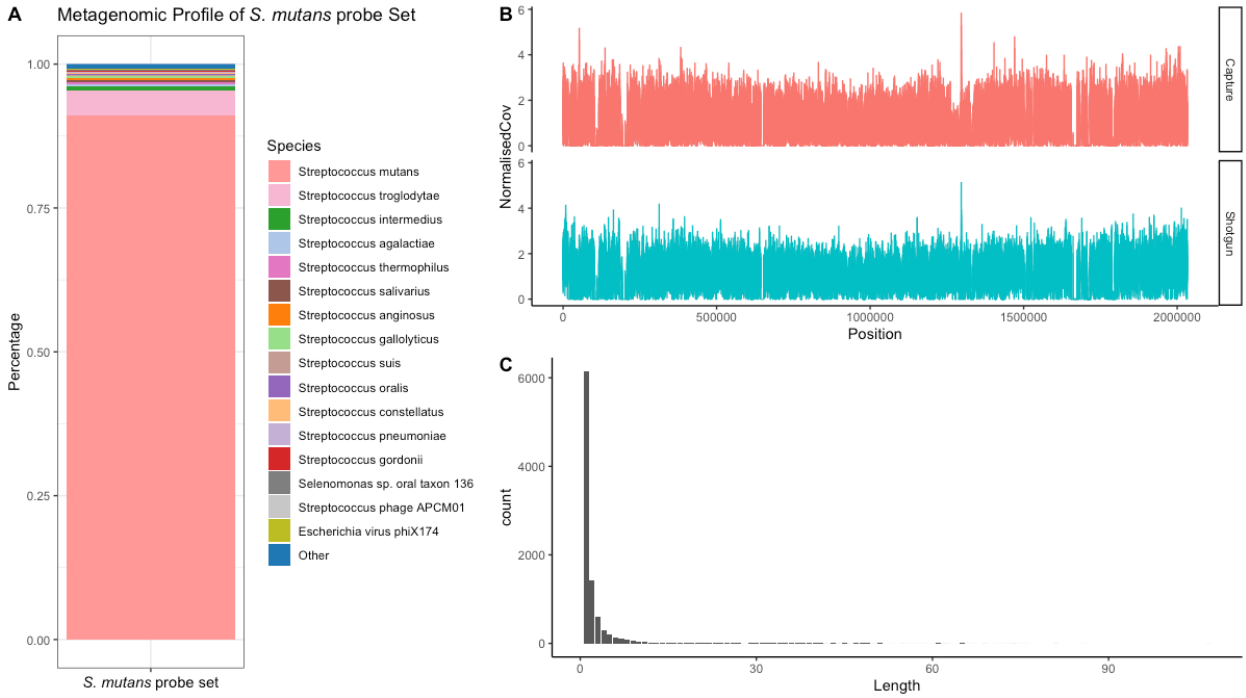


Figure 6: Evaluation of the probe set for the designed *S. mutans* capture and its performance compared to the shotgun genome. (A) Metagenomic profile of the probe set using the full nt-database in MALT. (B) Normalised coverage across the reference genome (*S. mutans* UA159) by the capture and shotgun genomes. (C) Histogram of the size of non-covered windows between the shotgun and the capture reconstructed genomes for the FAR004 strain.

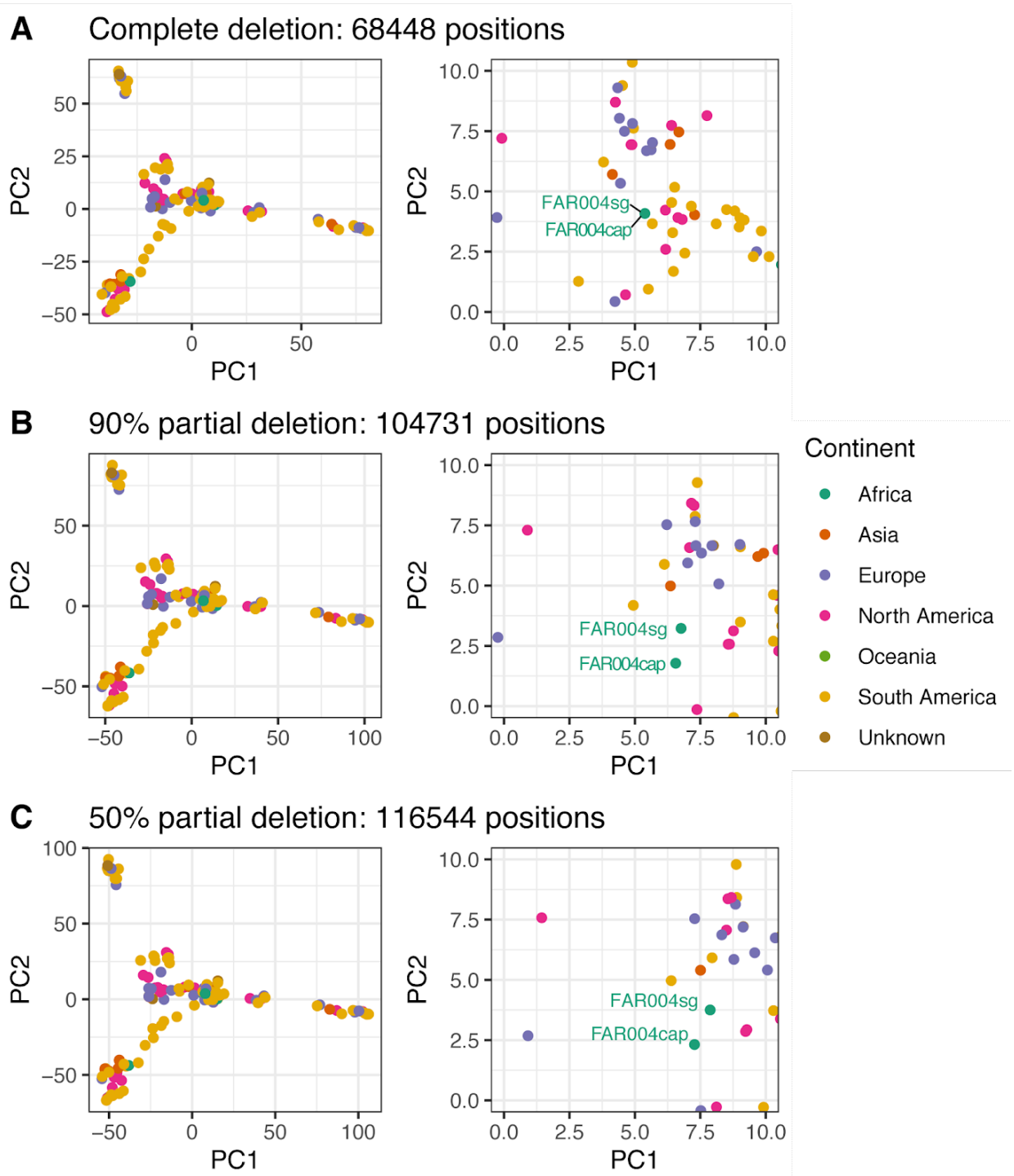


Figure 7: PCA computed with (A) no missing data, (B) allowing 10% missing data or (C) allowing up to 50% missing data. The PCA includes 193 modern *S. mutans* strains and the ancient FAR004 strain represented by the capture (FAR004cap) and the shotgun (FAR004sg) reconstructed genomes.

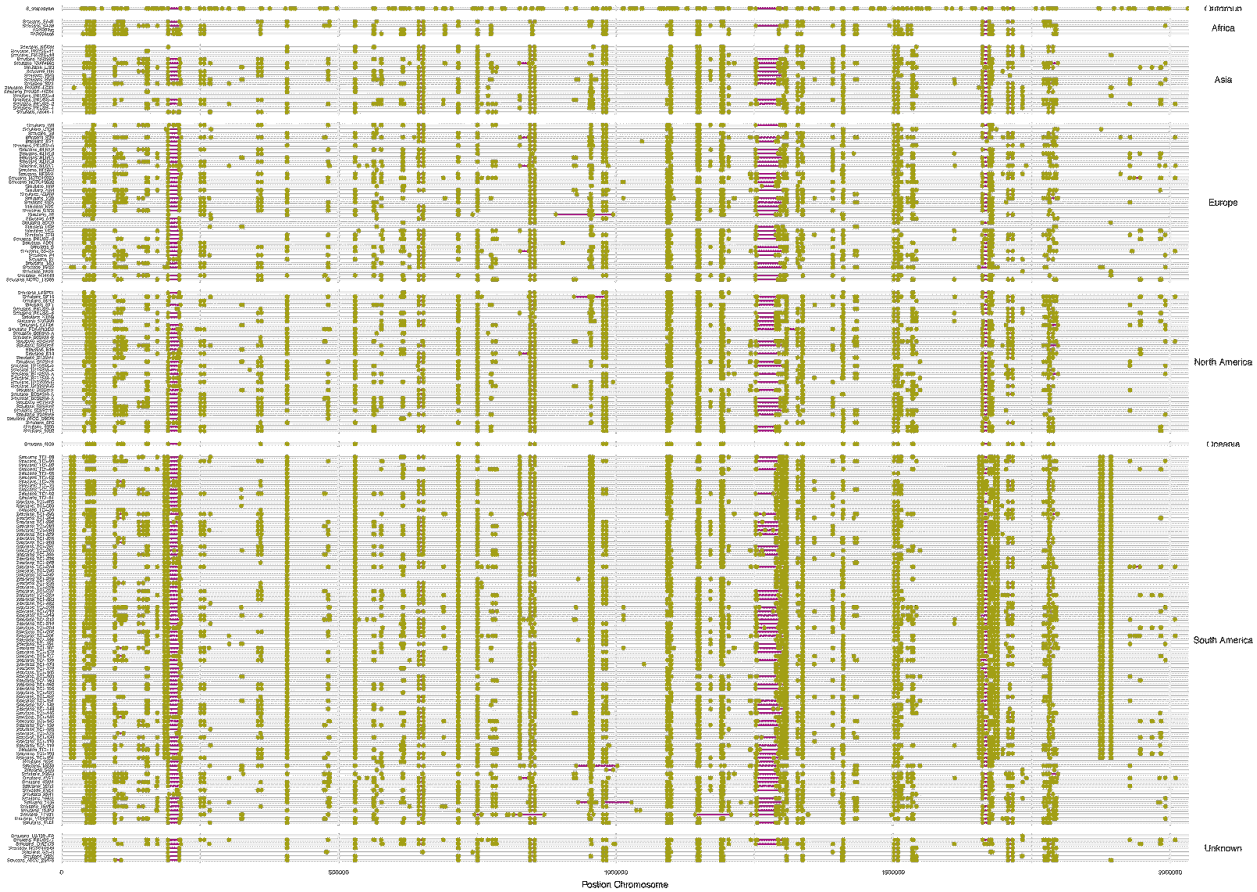
Table 1: Summary of the FAR004 sample including the archaeological information, C14 dating and genomic mapping statistics. calBP=calibrated Before Present (1950); sg=shotgun; cp= capture; Endo. =Endogenous; Cov. =Coverage

Individual	Tissue Sampled	Site	Radiocarbon date (14δC)	2σ Interval [calBP]	Data type	Clipped, Merged and Quality-Filtered Reads before mapping	Unique Reads Mapping to *S. mutans* UA159	Endo. DNA (%)	Mean Cov.	Cov. ≥1X (%)	Cov. ≥5X (%)	Publication
FAR004-nonUDG	Tooth	Faraoskop (South Africa)	2000 ± 50 (Pta-5283)	2017-1748 calBP	sg	1,503,374	1,680	0.17	0.04	4.02	0	Skoglund et al. (2017)
FAR004-UDG					sg	6,849,788,085	6,535,378	0.18	175.51	90.94	90.02	
FAR004-Capture					cp	8,793,624	1,218,659	35.83	34.66	89.72	86.25	

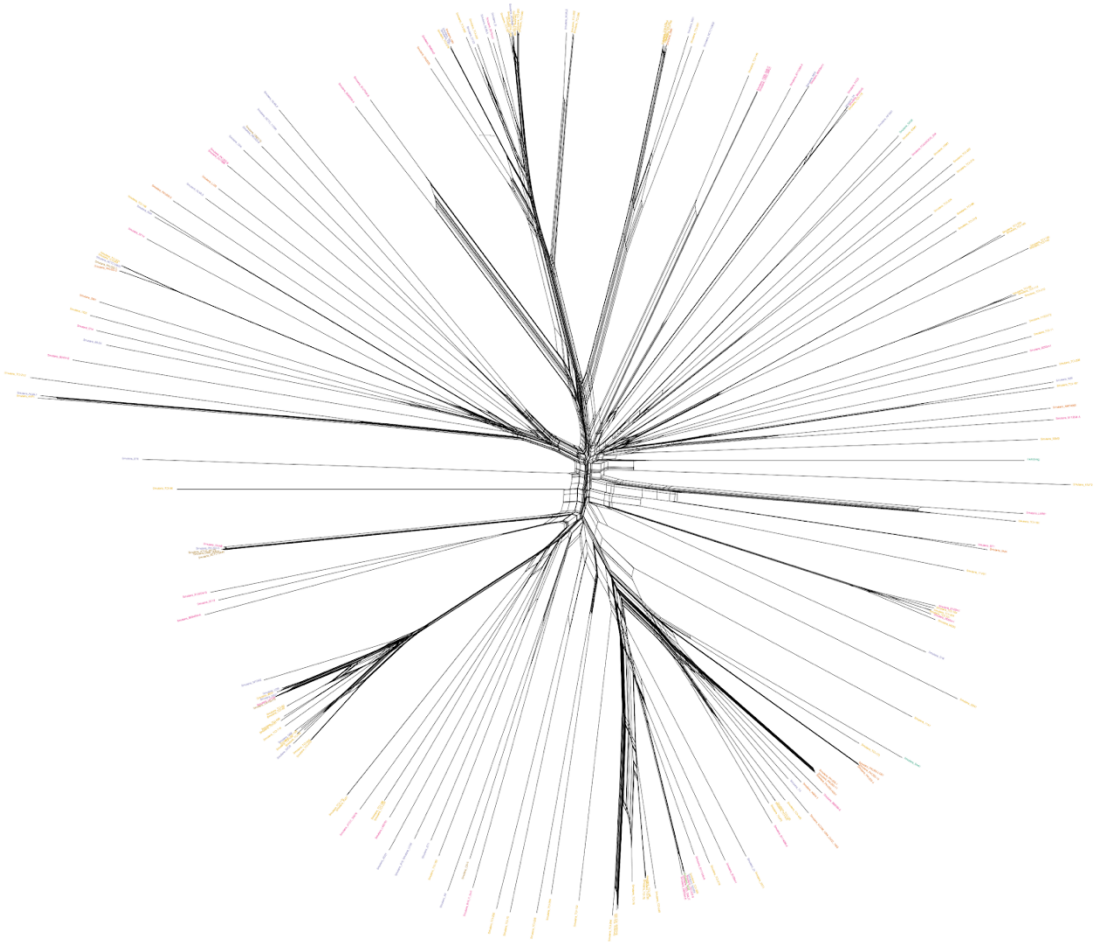
Table 2: Summary of the not called SNPs differing between the shotgun (FAR004-UDG) and the capture (FAR004-Capture) reconstructed genomes for the FAR004 ancient *S. mutans* strain. This was calculated for all the SNP calls (8,568 differing positions) and also for the unique SNPs of the FAR004 strain (799 differing positions).

No Calls	Genome	Coverage less than 5X	Heterozygous call
All no calls	FAR004-Capture	6150	995
	FAR004-UDG	15	1408
Unique SNP no calls	FAR004-Capture	251	199
	FAR004-UDG	2	347

Supplementary Information



Supplementary figure 1: Dumbbell plot showing the missing regions in comparison with the reference (*S. mutans* UA159). The green dots indicate either the start or the end of the missing region while the purple lines indicate the missing regions. This plot was generated using ggplot in R.



Supplementary figure 2: NeighborNet network of 195 *S. mutans* genomes including the ancient FAR004 strain. The tree has been coloured based on the geographic isolation of the strain.

Supplementary Table 1: Metadata available for the 194 modern *S. mutans* and 1 modern *S. troglodytae* strains used in this study

Discussion

Where to look for ancient pathogen DNA?

Research into ancient pathogen DNA started by looking for signs in either the bones of past human populations, historical accounts of diseases, or palaeoepidemiological signs of a catastrophic event such as mass graves. This context-based targeted approach led to the recovery of *Mycobacterium tuberculosis* from tuberculosis cases from pre-contact America (Bos et al. 2014), the understanding of leprae in medieval Europe (Schuenemann et al. 2013), and confirmation that *Y. pestis* was indeed the responsible agent for the first and second pandemic (Bos et al. 2011; Schuenemann et al. 2013; Wagner et al. 2014). Untreated oral disease can also result in lesions on the teeth, e.g. caries, or bone, such as bone loss due to periodontal disease. Ancient oral disease has been previously studied by utilising dental calculus (Warinner et al. 2014). Dental calculus has been shown to be exceptionally well preserved and to be a good representation of the oral microbiome present in modern dental calculus (Weyrich et al. 2017; Velsko et al. 2019; Warinner et al. 2014). This material is the result of the calcification of biofilms that form on the tooth surface, composed of a complex community of microorganisms encompassing over 600 taxa (Dewhirst et al. 2010). However, the presence of the rich microbial background can make it challenging to recover single genomes from dental calculus, since there is the risk of cross-mapping with either close relatives or different strains of the pathogen/pathobiont of interest. Furthermore, since dental calculus is not part of the human host itself, it can only inform us on the species present in the oral cavity and provide *indirect* evidence of potential pathobionts involved in the disease's manifestation. On the other hand, sampling of host tissues affected by oral diseases can provide direct evidence of the strains involved in the lesion formation. In Manuscript E, we have used this approach to analyse a carious tooth of an individual from South Africa who lived around 2000 years ago in order to gain insights into the bacterial background that lead to this lesion. Manuscript E represents the first study to reconstruct a complete genome of a taxa found in an ancient caries. This research has demonstrated the involvement of *S. mutans* in the formation of caries since ancient times, thus pointing to a long co-evolution of this pathogen with the human host. It further shows the potential of caries as a source for the study of oral disease and ancient microbial genomes. The fact that caries is diagnostically specific, easily identified and ubiquitous in the archaeological record, together with the successful results from this study suggests that this material type could allow for production of large genomic datasets in which to study the different microorganisms involved in this disease,

as well as species-specific dynamics at fine resolution. Additionally, in contrast to dental calculus, carious lesions are formed by the overgrowth of specific taxa, thus a lower microbial background is expected, which, in combination with the designed capture, could allow for the easier and cleaner reconstruction of specific pathobionts, further facilitating large scale studies.

Physical evidence on skeletons, archaeological or textual/pictorial evidence of disease allows for the detection of time periods and samples that could be relevant to the study of past epidemics. However, the fact that the disease needs to be visible to us by the described means limits the research to periods/samples with evidence, such are the fields of medical history or palaeopathology. Historical records are scarce for some periods and regions and they only represent a small part of human history. Additionally, most diseases do not result in osteological evidence due to a rapid progression of the disease, e.g. plague, or lack of bone involvement, such as chronic infection with the hepatitis B virus. In order to overcome this, there is a need to analyse individuals from the past where there is not clear evidence of disease. Finding pathogen DNA is similar to finding a needle in a haystack, requiring the screening of large datasets to find the infected individual where pathogen DNA has preserved. In the past years, the advancement of NGS techniques have allowed for the compilation and publication of extensive human DNA datasets from ancient prehistoric contexts (see for example Allentoft et al. 2015; Damgaard et al. 2018; Haak et al. 2015). Use of NGS has opened the possibility to use a by-product of these studies - the off-target non-human reads - to explore past disease in human populations. In order to look for pathogenic DNA, which usually represents a small amount of the total DNA retrieved from a sample, there was a need for the development of highly sensitive screening techniques for the ancient pathogen DNA field to account for these large sample sizes. Tools such as MALT (Vågene et al. 2018), which allows ultra-fast alignment of shotgun reads to large databases such as the NCBI Nucleotide (Nt) database, have since been used to detect these small amounts of pathogenic DNA. For example, the application of this tool led to the discovery of *Salmonella enterica* Paratyphi C – a bacterium that causes paratyphoid fever – in victims of the *cocoliztli* epidemic (see introduction), thus shedding light on the pathogen potentially responsible for this epidemic. Despite the taxonomic profiling functionality of MALT and subsequent analysis in MEGAN (Huson et al. 2007; 2016), this approach still required extensive manual work, which involves loading each dataset in MEGAN and checking the taxonomic profile to detect potential pathogens and evaluation of every potential positive. Manual work is prone to human error, such as missing a potential hit amongst the metagenomic background, and would be impossible to accomplish in larger datasets.

The need to screen, and more importantly, to evaluate hundreds of ancient human DNA libraries led to the implementation of a fast, targeted and automated screening method for the detection of *Y. pestis* reads. In Manuscript A (Andrades Valtueña et al. 2017), I show the successful implementation of a fast-screening method that allowed for the screening of 563 ancient human individuals dating to the Late Neolithic and Bronze Age (LNBA) period. This method allows to control for potential environmental contaminants that can lead to the false identification (false positive hit) of *Y. pestis* in a sample by employing a competitive mapping that include other *Yersinia* species. The application of this method, in parallel to the work by Rasmussen et al. (2015), led to the unexpected discovery of *Y. pestis* signals in individuals prior to the first pandemic. Specifically, we reconstructed the first 6 genomes dating to the LNBA period from Europe and demonstrated that these strains form a basal clade in the *Y. pestis* phylogeny with the previously sequenced strains in the Altai region (Rasmussen et al. 2015). This transformed our knowledge of plague by showing a long history of plague association with humans and questioned the previously hypothesised Chinese origin of *Y. pestis*.

However, this type of targeted screening method is limited to a single bacterial species, which restricts our results in terms of discovery of other pathogens in the dataset. Since the publication of Manuscript A, new tools have been developed to automate the detection of multiple possible human pathogens at once, such as the HOPS pipeline (Hübler et al. 2019) . HOPS utilizes MALT for the initial alignment of reads to a database and to extract information on a list of taxa of interest, which the user provides as an input. It not only allows for the fast screening of large datasets of ancient DNA libraries for a collection of pathogens, but also importantly collects information that allows for the palaeogenetic authentication of potential positive hits. This is essential in the context of ancient pathogenomics, where closely related environmental bacteria can generate false positives, and lead to wrong conclusions about the presence of possible pathogens. Typically, the authentication criteria that one would like to evaluate are: close similarity between the aligned reads and to the pathogen genomes in the database, which is evaluated with the edit distance (number of mismatches between a read and the reference it is aligned, see introduction), indicating a correct taxonomic assignment of the reads; the presence of cytosine deamination patterns, typically observed in true aDNA due to accumulation of hydrolytic damage at the end of the DNA molecules over time; and short fragment lengths. This fast screening technique was successfully used in Manuscript B where we detected 15 *Y. pestis* positives samples from 227 individuals using the HOPS screening, which allowed for the recovery of 14 genomes from the LNBA branch as well as the discovery of a new ancient *Y. pestis* lineage. Furthermore, the

possibility to screen for multiple pathogens at the same time has opened the door to the detection of multiple infections in single ancient individuals, such as in the case of a medieval individual co-infected by treponematosis as well as plague (Giffin et al. 2020). This is particularly interesting since it may indicate a synergy between pathogens, such as the ones seen between childhood illnesses and tuberculosis (Whittaker et al. 2019). Long lasting diseases, such as treponematosis, could weaken the immune system of the host, thus facilitating the infection of the host by other pathogens. This could also explain the rapid dispersal of plague during the second pandemic since the general health of the European population could have been weakened by other diseases and famines, such as the Great Famine of 1315-1317.

After the detection of potential pathogens in a DNA library, one should recover the whole genome of the organism to further ensure that the signal is not due to environmental bacteria. This can be achieved by deep sequencing of the DNA library to retrieve sufficient coverage of the genome, however, this often requires the sequencing of millions or billions of DNA molecules since pathogenic DNA represents normally a small fraction of the library (see for example Rasmussen et al. (2015) or Manuscript A (Andrades Valtueña et al. 2017)). This can constrain this research to well established groups with extensive funding. Techniques for cost-effective recovery of pathogenic DNA have also been established, such as capture-based techniques that allow for the enrichment of the DNA of interest in the library, thus reducing sequencing cost. In Manuscript A and Manuscript E, we have designed capture probes for the retrieval of *Y. pestis* and *S. mutans*, respectively. This will help - and in the case of *Y. pestis* is already helping (Spyrou, Keller, et al. 2019; Keller et al. 2019; Giffin et al. 2020; Spyrou et al. 2018) - the recovery of those pathogens to increase sample sizes of ancient strains in future studies to numbers that will be essential for more sophisticated analysis that require fine-scale resolution, for example to model the rate of spread and evolution that can inform long-term planning for modern outbreaks. These analyses include for example molecular dating, where ancient DNA provides a molecular fossil of past strains thus informing when those strains were present, functional evolution (as is applied in Manuscript A, B and E), or epidemiological modelling. The developed probe-sets in Manuscript A and E will assist the field by democratising access to smaller groups with less funding to also perform research on ancient *Y. pestis* and *S. mutans*. Furthermore, these captures, as well as the screening methods, are not limited to ancient material and could be applied to modern clinical-settings for the early detection of these bacteria in compromised samples due to bad preservation or containing small amounts of pathogen DNA.

What have we learned from phylogenies?

Through a variety of approaches to generate sufficient numbers of genomes, the study of ancient pathogens to date has been focused on the phylogenetic placement of the ancient strains in relation to their modern counterparts as well as to other ancient strains. Phylogenetics can also help the understanding of disease dynamics and dispersals in ancient times, which can be used to inform epidemiological modelling of modern epidemics. For example, ancient plague studies have shown that the Black Death outbreak in Europe was caused by a single clone that expanded quickly to the whole continent, which was ancestral to the strains that gave rise to the third pandemic, as well as to strains that established local reservoirs responsible for recurrent outbreaks in Europe until the 18th century (Spyrou et al. 2016; Namouchi et al. 2018; Spyrou, Keller, et al. 2019; Guellil et al. 2020; Morozova et al. 2020), which later disappeared from Europe. Phylogenies can also inform us of the past diversity of strains responsible for epidemics. The study of medieval leprosy has led to the discovery of a great diversity of strains of *Mycobacterium leprae* responsible for the disease in Europe, that either became extinct or are not present in Europe anymore (Schuenemann, Avanzi, et al. 2018). Why leprosy disappeared from Europe after the 16th century despite it being highly diverse and widespread is still an open question. Additionally, phylogenies including ancient genomes can be used for the inference of the animal reservoirs responsible for the zoonotic outbreaks in past human populations. This approach led to the astonishing discovery that ancient tuberculosis cases from the Americas were caused by strains that are now present in modern pinnipeds (Bos et al. 2014). The particular advantage of adding ancient DNA to the phylogenetic analysis is that it provides a deep-time calibration point, which, if combined with historical context, can inform on when the pathogen emerged, and pinpoint past events, or anthropological or environmental changes that can explain the long-term emergence of the pathogens.

In particular phylogenetic analysis has contributed greatly to the understanding of the early evolution of *Y. pestis*. The phylogenetic placement of the 25 recovered ancient genomes from the LNBA period has shown the presence of at least 4 ancient plague lineages (Rasmussen et al. 2015; Manuscript A (Andrades Valtueña et al. 2017); Spyrou et al. 2018; Rascovan et al. 2019, Manuscript B), thus highlighting that plague diversified quickly after emergence, which was not previously known. These ancient strains have also allowed for more refined dating of the ancestor of all known *Y. pestis*, with the most recent estimate being 5700 years Before Present (yBP) with a 95% Highest Probability Density (HPD) comprising 5250–6364 yBP. This time period is characterised by increase mobility of human populations (Allentoft et al. 2015; Haak et al. 2015),

as well as intensification of agro-pastoralism (Zvelebil 2001). Given the importance of animals as intermediate host in plague transmission to humans, this is a highly interesting correlation between intensification of human-animal interaction and the rapid diversification of *Y. pestis* during this period.

Despite the advantages of phylogenetic analysis, these can only be applied to species that evolve in a tree-like manner or clonally. Bacteria are renowned for their capacity to exchange genetic material through recombination (Didelot and Maiden 2010), as well as to incorporate exogenous DNA in their genome via Horizontal Gene Transfer (HGT; Sun, 2018). Events such as recombination or HGT led to parts of the genome evolving in a non-tree like-pattern that can lead to wrong phylogenetic inference. *S. mutans* is a highly recombining bacterium that also incorporates DNA via its competence pathway (Shanker and Federle, 2017). In Manuscript E, we performed a phylogenetic analysis of *S. mutans*, which resulted in a phylogeny with extremely low bootstrap support (a common technique to evaluate the robustness of the phylogenetic topology). Further exploration using network phylogenetic analysis, which allows for connections between different branches of the phylogenetic representation (reticulations) to account for HGT, showed an extremely reticulated network (See Supplementary Figure 2, in Manuscript E) particularly close to the root of the tree. These reticulations are indicative of the presence of non-tree-like-evolving regions in the genome and questioned the phylogeny reconstructed with the Maximum Likelihood algorithm. This clearly demonstrated the presence of extensive recombination contributing to the evolution in *S. mutans*, also shown via a different method in Cornejo et al. (2013). A typical approach to reconstruct phylogenies from recombining organisms is to remove recombinant regions from the alignment with tools such as ClonalFrameML (Didelot and Wilson, 2015), however I question the adequacy of this approach for *S. mutans* since recombination seems to be widespread across its genome. Since the field of ancient pathogenomics is diversifying, there will be a need to find new analytical techniques to deal with recombination. The study of gene content or population genetic methods to detect admixture could be better suited to the study of recombinant bacteria, such as *S. mutans*, however there is only one study to date that has applied them in an ancient context to study a 5300 year old *Helicobacter pylori* strain recovered from the mummy of Ötzi (Maixner et al. 2016). The results of Manuscript E will help to emphasise to the field the current problem of the lack of recombinant aware methods for ancient pathogenomics, and has identified it as an important area of analytical development that must be carried out to open the possibility to study other bacterial species where phylogenetic analyses are inadequate.

New insights from known genes and deletions

Phylogenies can inform us on which strains caused past outbreaks, however they do not provide any information on how the pathogen functioned at the time. In modern genomics, this is achieved by genome-scale comparative analysis, where strains with differing phenotypes such as virulence are compared in their genomic content to see which genes differentiates them and can explain their different phenotypes. Comparative genomics of *Y. pseudotuberculosis* and *Y. pestis* have led to the discovery of several important virulence factors that play a role in the disease outcome and transmission in *Y. pestis*. Functional understanding of past strains is more relevant for modern clinical context, where this knowledge is used for combating the disease, however this has been rarely the focus of ancient pathogenomics studies.

In order to gain insights on this, in Manuscript A I established a workflow to check for the presence and absence of a set of genes in a genome. This workflow was applied to check for known virulence factors of *Y. pestis* in ancient strains. This workflow led to the discovery of the presence of two forms of *Y. pestis* in Eurasia during LNBA period (5000-2400 yBP) that differ in their transmission and disease potential (Spyrou et al. 2018, Manuscript B). What differentiate these lineages is a series of genomic adaptations required for an efficient blocked-flea transmission. One of the genes absent in the LNBA lineage and the earlier lineage represented by the Gok2 genome (Rascovan et al. 2019) is the *ymt* gene. This gene is required for the survival of *Y. pestis* inside of the gut of the flea vector, which enhances its transmission potential. Furthermore, the genes related to the biofilm regulation and formation were still active in those earlier lineages. The loss of function of these genes allows for the formation of a biofilm in the entrance of the stomach of the flea (proventriculus), which prevents the meal blood to reach the stomach of the flea, resulting in an increase of bite frequency by the hungry flea, therefore increasing the changes of *Y. pestis* being transmitted to the host. Finally, the earlier strains have still an active *ureD*, which has been silenced in *Y. pestis* strains since its expression causes toxicity and morbidity to the flea vector, thus reducing its transmission capability. The LNBA and Gok2 genomes potentially represent a pre-flea adaptation form of plague, although we cannot discard that fleas played a role in the transmission of this earlier plague. The study of animal remains can add valuable information on the ecology of the disease and the cycle of transmission to human populations. This cycle may have been rather different to the one we observe for modern plague. The only attempt to detect ancient *Y. pestis* from animals is in the context of the second pandemic, where signs of this pathogen were found in a rat, thus suggesting rats played a role in the dispersal of plague during the second pandemic (Morozova et al. 2020). The analysis of known virulence

factors coupled with the phylogenetic analysis also showed that the specific plasmids pMT1 and pPCP1 were already present in the early stages of the evolution of plague, as expected since those plasmids differentiate *Y. pestis* from its ancestor. Surprisingly, however, the *ymt* which is encoded in the pMT1 plasmid in modern strains was absent, indicating the uptake by the bacterium of that gene occurred at a later stage. Based on modern strains, it was hypothesized that the *pla* gene, encoded in the pPCP1 plasmid, was acquired after the adaptation to the flea vector (Sun et al. 2014).The application of this workflow showed that the earlier forms of plague had already acquired the *pla* gene prior to the adaptation to the flea vector, thus rejecting the inference based on modern genomes. This workflow provided insights into the long-term genetic changes of *Y. pestis* that were crucial for its emergence and transmission adaptations, therefore showing that new ways of analysing ancient genomes can inform on the long-term evolution of pathogens. Furthermore, it has provided new perspectives on the past genetic diversity *Y. pestis*, and new questions for future research such as did the efficient blocked-flea transmitted and earlier non-blocked-flea adapted forms occupy the same ecological niches and why did these ancient strains become extinct.

By modifying the workflow developed in Manuscript A, I adapted this type of analysis to detect presence and absence across the entire genetic content of the chromosome, not only to the known virulence factors. In Manuscript B, we looked for regions of the reference chromosome (*Y. pestis* CO92) that were deleted in ancient strains. We have shown an increase in genomic degradation along the LNBA lineage, which can be indicative of selection or adaptation in bacteria (Hottes et al. 2013; Koskiniemi et al. 2012). The deleted regions contain genes that are related to the flagellar machinery as well as membrane proteins which may play a function in immune recognition. Through this type of analysis, it has since been shown that this pattern of deletions also occurs during the first and second pandemics, in which we observe the parallel loss of the same genomic region (Spyrou, Keller, et al. 2019; Keller et al. 2019). Interestingly, the region loss during the first and second pandemic also contained genes of the same categories than those lost in later strains of the LNBA lineage. Those strains all represent past diversity of plague that we do not observe nowadays, thus indicating that the genes contained in the loss regions are possibly of importance for the survival of modern *Y. pestis* strains since they are present in all modern strains. Missing genes in past strains could be potential targets for the development of new antibiotics for the treatment of plague, which may become more relevant in future times, since the presence of antibiotic-resistant *Y. pestis* strains has already been reported (Cabanel et al. 2018; Galimand et al. 1997; Galimand, Carniel, and Courvalin 2006).

The workflows for analysing virulence and gene insertion and deletion (indels) are not limited to the study of *Y. pestis*, as they only require ancient reads mapped to a modern reference and a list of genes of interest. In Manuscript E, we applied the same workflow to the study of a 2000-year-old genome of the pathobiont *S. mutans*. By analysing known virulence factors, we observed that the ancient strain cannot be distinguished from modern strains based on the presence/absence of those genes. However, inferences on how the prevalence of genes in the *S. mutans* population changed cannot be made from a single genome. More ancient *S. mutans* genomes will provide insights on how the presence/absence patterns may have changed with the introduction of more processed and carbohydrate-rich diets after the introduction of agriculture or after the introduction of dietary sugar during the industrial revolution. Genes that change in prevalence in the *S. mutans* population are likely to play an important role in the adaptation of this pathobiont to changes in behaviour of the human host, and they could provide information for the prevention of caries, such as dietary recommendations depending on the strains present in the patient.

This workflow opens the possibility to explore the gene content of other pathogens, pathobionts or microorganisms of interest to gain insights into the potential of virulence, adaptation, or disease ecology of those. Additionally, the fact that the workflow is based on percent of the gene covered, it accounts for low-coverage genomes that are common in aDNA, thus contributing to the field a method to gain information on past microorganism even when the data available is scarce and insufficient for phylogenetic analysis.

Moving on from a single reference genome: *de novo* assembly and pangenomics

While manuscripts A, B and E successfully used single-reference based approaches to infer changes in the gene content of *Y. pestis* and *S. mutans*, relying on single genomes restricts the analysis to the single genome used as reference, therefore limiting the scope of discovery of other important genes for the evolution of the microorganism. To overcome this limitation, Manuscript C and D propose the modification of techniques used for modern comparative analysis, specifically *de novo* assembly and pangenomics, to allow for the inclusion of ancient data.

De novo assembly allows for the reconstruction of a genome in a reference-free manner. The application of *de novo* assembly tools to ancient DNA has remained elusive due to the intrinsic characteristics of aDNA (see introduction). *De bruijn* algorithms employed for *de novo* assembly of genomes rely on the overlap of many sequencing reads to link reads together to form ‘contigs’

(a set of overlapping reads that represent a part of the genome). Contigs are then ordered into scaffold, thus representing longer parts of a genome. aDNA is highly fragmented in nature and in combination with low coverage, results in short contigs due to the absence of a sufficient overlap-paths linking the reads. Furthermore, the absence of long reads in aDNA datasets obstructs the reconstruction of repetitive regions, such as insertion elements or transposons usually associated with genomic rearrangements, which could result in wrongly and/or short assembled contigs. For those reasons, *de novo* assembly has only occasionally be performed (see for example Schuenemann et al. 2013; Devault et al. 2017; Krause-Kyora et al. 2018; Manuscript D). *De novo* assemblies of ancient microorganisms could allow for the detection of new genetic components that are absent in the diversity of modern strains. It also opens the door to explore the genomic organisation of past strains. This is important in the case of plague as specific rearrangements have been linked to different foci and could be used to infer and distinguish between different reservoirs.

Prior *Y. pestis* assemblies were performed on highly fragmented aDNA, which resulted on assemblies containing multiple contigs (Bos et al. 2011; Luhmann, Doerr, and Chauve 2017). In Manuscript D, we successfully assembled an exceptionally well-preserved genome (mean read length of 102bp) of *Y. pestis* from the 17th century. The resulting assembly consisted of 2 scaffolds and represents an unprecedented highly contiguous ancient plague genome. The workflow proposed in Manuscript D did not only allowed for the assembly but also determined the necessary coverage to obtain the best assembly, and showed that for the case of *Y. pestis* a minimum of 200X coverage is required for a good assembly from aDNA. This will allow other researchers of the field to plan the sequencing experiments to obtain the best assembly based on the organism of interest. On contrast with the previous studies, we used the SPAdes assembler in combination with the Ragout scaffolding tool to order the contigs on the basis of different genomic architectures present in *Y. pestis*. Additionally, in contrast to previous attempts, we annotated the genome with Prokka, providing the community with the first annotated ancient plague genome. We also explored the gene content of this strain in comparison to other *Y. pestis* strains. We showed that there was no acquisition of new genetic material during the second pandemic. Furthermore, we confirmed the presence of a 49kb deletion in this strain as described in Spyrou, Keller, et al. (2019), which contains important genes for the virulence of *Y. pestis*. This workflow could be used to perform *de novo* assembly of other well-preserved specimens, which are likely to become available since the screening for pathogenic DNA is becoming routine practice. However, there is also room for development of assembly tools specifically for ancient DNA, which could open the door to reconstruct *de novo* assembled genomes even when

preservation is not optimal. Steps in that direction have started, for example the two layer assembly tool MADAM (Seitz and Nieselt 2017) or MADMAN (https://github.com/maxibor/madman), a newly developed pipeline that utilises damage to authenticate the reconstructed contigs, thus ensuring the ancient origin of the reconstructed genomes.

As an alternative approach to *de novo* assembly – which Manuscript D shows that high coverage is required, and in many cases will still be unfeasible for aDNA samples - one could instead employ pangenomics to overcome the biases of a single genome reference, to get a wider view of the gene-content potential in a particular sample. Pangenomics is an emerging field that allows for the study of all the genes present in a given taxon (see introduction). The pangenome has been proposed as a concept to study not only gene content, but also as a method for taxonomic classification and understanding bacterial evolution in general (Rouli et al. 2015). Most of the available tools to compute a pangenome, such as panX (Ding, Baumdicker, and Neher 2018) or Roary (Page et al. 2015), rely on well annotated genomes as input. Since the assembly of ancient genomes is limited to well-preserved samples as described above, the already existing pangenome tools would not allow the inclusion of ancient genomes, which are normally available as raw sequencing reads. Incorporating ancient genomes in the pangenome can provide information on which genes have played a role in the development of a species, as well as how the pangenome was formed through time.

In order to overcome this, I have proposed a new workflow in Manuscript C. This workflow generates a pangenome based on the well annotated genomes of the species of interest. Additionally, and most importantly for the ancient pathogenomics field, it includes a mapping step that allows for the incorporation of ancient genomes as well as modern genomes that lack annotations or where only raw sequencing data is available. Furthermore, by utilising sensitive mapping parameters, we can also incorporate low-coverage aDNA data, thus opening the possibility to gain insights on the gene content of this ancient genomes that are normally excluded from analysis such as phylogenies. Furthermore, this new workflow is not limited to the pangenome reconstruction, we also incorporate in the workflow the inference of presence and absence patterns across the pangenome as well as the detection of pseudogenes, genes that no longer function due to frameshift or premature stop codons mutations. This furthers the analytical capabilities of the field to understand the evolution of pathogens based on the total gene content as well as loss of function through pseudogenisation, which has been shown to be of importance in host specialisation in for example *S. enterica* (Zhou et al. 2018). We demonstrated the

successful application of the workflow by generating the first pangenome of *Yersinia pseudotuberculosis* complex that includes ancient data. The gain and loss of functions has been pivotal to the emergence of *Y. pestis* (Hinnebusch, Chouikha, and Sun 2016), however only few studies have been focus on the pangenome of *Y. pseudotuberculosis* and *Y. pestis* and based on gene content (Califf et al. 2015). Manuscript C represents then the first study to provide insights into the evolution of *Y. pestis* by analysing both the gene content as well as the pseudogenisation events that played a role in the emergence of *Y. pestis*. Additionally, this approach could help to provide a holistic view of the functional development of *Y. pestis* since its emergence from *Y. pseudotuberculosis*. The Manuscript C workflow is not restricted in its application to the *Y. pseudotuberculosis* complex and could be used to gain new insights in the evolution of other pathogens, as well as, microbiome associated microorganisms. Its application to *S. mutans* could help to understand the formation of its open pangenome and link it to its co-evolution with the human host. Despite the representation being limited by the well annotated genomes, Manuscript C proposes a method to avoid single reference bias and allows for the exploration of all the gene content in the species in ancient genomes.

What is knowledge without context: future outlooks for the field of ancient pathogenomics

Ancient pathogen studies have been highly focused in the phylogenetic evolution of the ancient pathogen, however the application of those studies to modern context are limited. In this thesis, I have provided the field with new approaches to understand not only the relationships with their modern counterparts, but also starting to uncover how those ancient pathogens (*Y. pestis* and *S. mutans*) differ in their functions with respect to modern strains. The results from these studies could guide modern functional studies on these pathogens to check the genes and variants present in ancient strains that could be important for the survival, transmission and virulence of the pathogen. By studying the genomic evolution of past pathogens, one can start to understand potential mechanisms employed by microorganisms that lead to pathogen emergence. However, the emergence and evolution of pathogens does not happen in an empty void, and aDNA pathogen studies have rarely discussed in-depth the context in which those pathogens are found. In order to provide an ecological context to the emergence and evolution in the aDNA pathogenomics field, there is a need to start to incorporate other disciplines into the research.

Studying changes in human behaviour is essential to understand past and modern zoonotic diseases. Humans have been modifying their environments to suit their lifestyle for millennia. One

of such modifications was the introduction of agriculture and domestication of animals, which could have increased the chances of humans getting into contact with animals as well as disease vectors (Cohen and Armelagos 1984; Ronald Barrett et al. 1998; Key et al. 2020). Human mobility and behaviour are also of high relevance to understanding the introduction and spreading of a disease, as we have seen with the COVID-19 pandemic in 2020, where bushmeat, one of the proposed sources of the initial outbreak (Shereen et al. 2020), together with highly interconnected human networks led to quick spread of the disease worldwide. We can learn about human mobility from archaeology and ancient DNA studies. Ancient human DNA studies have suggested an increase in human mobility around 5,000 years ago (Allentoft et al. 2015; Haak et al. 2015). This is also the time where we find the first evidence of the LNBA *Y. pestis* lineage in Europe. Based on this observation, we linked the increase in human mobility with the appearance of *Y. pestis* in Europe in Manuscript A. We hypothesised that the disease was introduced into Europe in a process associated with the dispersal of human groups from the Eurasian steppes into Europe. However, to understand the ecology of the earlier plague we need to start exploring its presence in other parts of the world. Most ancient human DNA studies have been focused on Eurasia and more specifically Europe (Slatkin and Racimo 2016), and even these datasets are currently limited by comprising only petrous bones, which are not ideal to detect pathogens (Margaryan et al. 2018). We have now detected two closely related flea-adapted *Y. pestis* lineages that are separated by thousands of kilometres: one represented by a genome in Spain, identified in Manuscript B, and the other represented by RT5 found in the Samara region (Spyrou et al. 2018). Analysing human remains from other regions in Europe such as France or Italy could help understand how those flea-adapted *Y. pestis* strains spread during the LNBA period. Even though other parts of the world have been explored by the means of ancient human DNA, the available data from these studies consist of human reads only, thus making it impossible to re-use this data to conduct pathogen screening. Since ancient remains represent a limited resource, we should maximise the scientific yield that we can obtain to minimise destructive sampling. There is then a need to establish a conversation between the field of ancient human genomics and pathogenomics to set up collaborations to ensure that off-target reads are being analysed by pathogenomics experts. In addition to geographical sampling, we need to extend the sampling outside of human individuals and start looking into other species. This is particularly relevant for the study of ancient plague. We know from today that neither *Y. pestis* nor its ancestor *Y. pseudotuberculosis* are human specific pathogens. *Y. pestis* is a rodent disease, while *Y. pseudotuberculosis* can be found in the environment and has a less restricted host range, which include rodents as well as birds, wild boars and pigs among others (Fukushima and Gomyoda

1991; Childs-Sanford et al. 2009; Reinhardt, Hammerl, and Hertwig 2018; Chakraborty et al. 2015).

In order to understand what was the host range of the earlier plague, we need to start sampling animal remains. In Manuscript B, we have observed that the LNBA *Y. pestis* branch is widely distributed across Eurasia, which evolved in a clock-like manner from a single population. This was formally demonstrated with a novel analysis in which we look for the correlation between genetic and geographical distance versus genetic and time distance, which has only been possible due to the recovery of multiple genomes in the LNBA branch – 22 of which were contributed by Manuscript A and B. The fact that we show that geography does not play a role in the formation of the LNBA branch and a strong correlation between time and genetic distance demonstrates that the genomes in this branch form a single lineage that does not experience diversification into parallelly evolving lineages. This observation is an indication of the presence of a single reservoir of this branch, and point towards the reservoir or the intermediate host of human infection being highly mobile. Based on this analysis and the wide-spread distribution of the lineage, we have proposed that the increase in mobility and pastoralism practise by the LNBA human populations could have increased the chances of humans coming into contact with infected wild-animals, either directly or through domesticates. Screening of animal remains present in archaeological sites could help to illuminate which other species were susceptible to plague and could have acted as intermediate hosts for human infection. Understanding the intermediate hosts and potential vectors is not only relevant for the earlier plague. Despite phylogenetic analysis of ancient *Y. pestis* strains elucidating the dynamics and routes of spread of plague during the second and first pandemic, which species acted as vectors during those is still up for debate. Although black rats and their fleas have been proposed as the spreaders of the disease, there is no description of rats' mortality from the second pandemic as observed in modern plague epidemics (e.g. Walløe, 2008; Ell, 1979). Rats should not be disregarded as potential intermediate hosts as a recent study has detected *Y. pestis* signals in a medieval rat (Morozova et al. 2020). Therefore, exploring of further faunal remains as well as blood-borne insects in search of *Y. pestis* aDNA is needed to advance the field. Human ectoparasites have been proposed as an alternative to the rat-flea model to explain the rapid spreading of the disease during the second pandemic (Dean et al. 2018; Drancourt, Houhamdi, and Raoult 2006). Ancient DNA from ectoparasites found in clothing of the time could provide some insights on which of them were involved in the transmission of the disease.

Another important factor to have in account when studying past disease is climate. Environmental changes, such as climatic change, as well as species extinction have been shown to increase the likelihood of emergence of new pathogens (Brooks and Boeger, 2019). In the case of plague, climatic variables such as precipitation and temperature have been suggested to be good predictors for the incidence of plague in human populations (Enscore et al. 2002; Stenseth et al. 2006). This line of evidence has been used for example to argue for the multiple introduction of plague in Europe during the second pandemic (Schmid et al. 2015). The study of strains from the second pandemic have shown the opposite, where the establishment of local reservoirs have been argued for (Spyrou, Keller, et al. 2019; Namouchi et al. 2018). Integrating long-term climatic data from throughout, with the current growing molecular knowledge of plague, as provided by Manuscript A and B, may illuminate the environmental conditions that led to the emergence of plague.

Context is not only relevant for virulently lethal disease. Past diet is of particular relevance for the study of more long-term and systematic oral and gut microbiota associated diseases such as caries. Since caries is a multi-causal disease in which diet plays an important role, we need to couple the microbial research with dietary information. The study of plant and animal remains together with isotopic studies can help to reconstruct past diets.

While in these studies I have begun to push the ancient pathogenomics field towards a more functional focus, the detection of presence of genes does not mean they were being expressed. Ancient protein is a relatively unexplored avenue for the detection of ancient pathogens. The preservation of the dental proteome in calculus (Warinner et al. 2014) as well as from a *Homo erectus* individual from Dmanisi dated to 1.77 Ma (Welker et al. 2020), shows the potential of proteins to explore disease even in deeper evolutionary times. Furthermore, ancient protein research can provide us a new line of evidence for which proteins were expressed during the course of disease, which is not possible to do with DNA alone. This could open the possibility to compare the expression profile of modern and ancient strains, which could increase our understanding on how they differ in terms of disease course.

Finally, the ancient pathogenomics field needs to more extensively reach out to the modern pathogen community. The modern community can provide insights into the biology of a specific pathogen which can help to contextualise the findings of ancient pathogenomics studies. This is particularly relevant for ancient pathogenomics where researchers conducting the studies may have been trained in other field other than microbiology, such as archaeology or computer

science. In turn, ancient DNA can bring new insights in the long-term evolution of a pathogen that can provide the timing and order of key adaptations of pathogenic species. The genomic-centric approaches presented and the data generated in this thesis can provide targets for empirically functional testing. The expression in controlled laboratory settings of identified genes that differentiate ancient strains from modern ones could open the possibility to reconstruct the behaviour of past strains, as well as to reveal genes important for the function of the pathogen that can be new targets for the development of treatments or vaccines for modern diseases.

The multidisciplinary approach to the study of past diseases would allow to create an ecological and epidemiological framework to study the emergence and long-term evolution of diseases, not only infectious but also microbiome-associated ones, which could inform models for the prevention, management and prediction of modern outbreaks and pandemics.

Conclusion

In this thesis I have presented new analytical approaches to move away the focus in the ancient pathogenomics field from phylogenetic-centric analysis of when and why events occurred, to start answering questions about how and why pathogens evolved. The implemented analytical approaches in this thesis can answer questions about the virulence evolution of past pathogens by analysing their gene content. Furthermore, patterns of presence and absence of genes, as well as pseudogenisations patterns of past ancient strains can inform us about disease ecology. I apply these approaches to study the evolution of two important bacteria: *Y. pestis and S. mutans*.

For *Y. pestis,* I have explored previously unknown lineages, composed by early strains of plague dating to the LNBA period. In manuscript A, we recovered the first strains from Europe, which together with the previous recovered ancient genomes from the Altai formed an undocumented branch of *Y. pestis*, named LNBA branch. This manuscript showed that the LNBA branch was present in a wider territory than predicted before. Additionally, we proposed that the entrance of plague in Europe was related to the expansion of human populations from the Eurasian steppe, since the phylogeny of the pathogen mirrors the human movements. Furthermore, we show that the genetic background of those strains is not compatible with the model of transmission via a blocked flea. We extend the genomic *Y. pestis* dataset from the LNBA in Manuscript B, representing one of the largest ancient pathogen genomic datasets from prehistoric times to date. We describe a novel lineage represented by a genome recovered from an individual in Spain. This lineage together with another ancient genome recovered from the Samara region contain all

the necessary genes for the efficient transmission via the blocked flea, thus demonstrating the presence of at least two forms of plague that differ in their disease and transmission potential. By utilising a novel analysis, we show that the genomes in the LNBA branch form a lineage that does not experience parallel diversification. This has implications for the disease ecology suggesting a restricted reservoir from which human infection occurs.

Rather than focusing on phylogenies, in order to understand the genetic changes that contributed to the virulence and functional evolution of *Y. pestis*, I have developed a novel workflow that successfully reconstructed a pangenome for 912 genomes, including the available ancient genomes, of the *Y. pseudotuberculosis* complex. Both the large number of genomes as well as the inclusion of ancient data in the pangenome reconstruction is unprecedented in the study of *Y. pestis* emergence and evolution. This represents one of the first studies to explore the pangenome of ancient strains, both in terms of presence and absence of genes and the presence of pseudogenes. In the future, this data will allow for the exploration of the whole gene content of the complex and to detect specific genes or pseudogenes that are exclusive to either *Y. pseudotuberculosis* or *Y. pestis*, thus providing new insights into their differences. Finally, in Manuscript D I explore the potential of *de novo* assembly from ancient pathogens. This led to the *de novo* assembly of an exceptionally well preserve *Y. pestis* genome from the 17th century, which represent the most contiguous ancient assembly of this species to date. Additionally, we provide the community with the first annotated ancient *Y. pestis* genome.

For *S. mutans*, I successfully reconstruct the first genome of this species from an ancient context (Manuscript E). I show that the ancient strain presents similar patterns in terms of presence and absence of genes that its modern counterparts, therefore suggesting a rather old diversity of the virulence factors for this species. I demonstrate that carious tooth is an ideal source to study ancient oral disease which could allow for large comparative analysis of the genetic background responsible of caries formation. Understanding caries formation and how it has changed through time, potentially influenced by major dietary and habits changes, such as the introduction of agriculture or increased access of refined sugar during industrialisation, is of special interest since caries still represents one of the most common disease affecting a large proportion of the human population. The past could provide new information on which microorganisms and conditions have been influencing caries formation, knowledge that could inform disease management in modern times.

Summary

Reconstructing genomes of ancient pathogens and microbes can provide information on how infectious diseases were affecting past human populations as well as the composition of the human microbiota. The field of ancient pathogenomics has been mostly focused on phylogenetic analysis to answer questions on how the diversity observed today came to be, by reconstructing the relationship of past strains in comparison with each other as well as with modern strains. Phylogenetic analysis of ancient strains can help us model the long-term dispersal of past diseases in terms of geography, as well as, speed. However, this type of analysis cannot provide information on the mechanisms that led to the dispersal, adaptation and virulence of past pathogens neither on their past ecology. In this thesis, I present new approaches that allow us to explore the functional evolution of past pathogens. I have applied those approaches to understanding the early evolution of *Yersinia pestis*, which is the bacterium responsible for plague. *Y. pestis* represents one of the best studied pathogens in historical times, and until recently it was thought that the first time it affected human populations was during the Plague of Justinian in the 6th century AD. In Manuscripts A and B of this thesis, I have shown that plague was affecting human populations since the Late Neolithic and Bronze Age (LNBA) long before historical times. By applying functional analysis to those early strains, we have identified the presence of two forms of plague during the LNBA period that differ in their transmission and disease potential. To extend the analysis from a reference-based approach to a pangenomic one, in which *all* the gene content of a given species is analysed, I propose a new workflow in Manuscript C, which allows the computation of the pangenome, as well as the inclusion of ancient genomes in the pangenome representation, which is unprecedented. By applying this workflow to the *Yersinia pseudotuberculosis* complex, which includes *Y. pestis*, its ancestor *Yersinia pseudotuberculosis* and *Yersinia similis*, we gain insights into the functional evolution of this complex in order to understand the emergence of plague. Additionally, reference-based approaches do not allow to explore novel insertions/deletions nor the structural organisation of ancient genomes. In Manuscript D, I present a workflow for the *de novo* assembly of ancient *Y. pestis* and provide the community with the first annotated *Y. pestis* genome reconstructed from a victim of plague from the 17th century. These workflows are not limited to their application to *Y. pestis*. In Manuscript E, I reconstruct the first ancient genome of *Streptococcus mutans*, a pathobiont involved in the formation of caries, from an individual who lived around 2,000 years ago. I show that phylogenetic analysis of this species may not be appropriate for the study of its

evolution and show the potential to gain insights into the co-evolution with the human host by applying the functional workflows developed in the course of this thesis.

Overall, this thesis showcases how the application of functional analytical approaches to ancient pathogen genomes opens the possibility to study their emergence and long-term evolution of their virulence.

Zusammenfassung

Die Rekonstruktion von Genomen alter Krankheitserreger und anderer Mikroben kann Informationen darüber liefern wie Infektionskrankheiten menschliche Populationen in der Vergangenheit beeinflusst haben und wie damalige Mikrobiome zusammengesetzt waren. Das Forschungsfeld der Palaeopathogenetik hat sich zumeist auf die vergleichende phylogenetische Analyse alter und moderner Genome konzentriert, um die Frage zu beantworten wie die heute zu beobachtende genetische Vielfalt von Mikroorganismen zustande gekommen ist. Derartige Analysen können uns helfen die geographische Verbreitung von Krankheiten in der Vergangenheit sowie deren Verbreitungsgeschwindigkeit zu modellieren. Allerdings können diese Methoden keinerlei Informationen darüber liefern durch welche Mechanismen diese Verbreitung erfolgte oder wie sich die Erreger mit der Zeit anpassten bzw. wie sich deren Virulenz und Ökologie entwickelte. In dieser Dissertation präsentiere ich neue Ansätze zur Untersuchung der funktionalen Evolution alter Krankheitserreger. Ich habe diese angewandt um die frühe Evolution von *Yersinia pestis* zu studieren, dem Bakterium, das die Pest auslöst. Es handelt sich um eines der eingehendst untersuchten historischen Erreger und bisher ging man davon aus, dass dieser im Rahmen der Justianischen Pest im sechsten Jahrhundert zum ersten Mal die Menschen befiel. In den Manuskripten A und B dieser Dissertation konnte ich zeigen, dass Menschen bereits im späten Neolithikum und der frühen Bronzezeit - also schon zu prähistorischen Zeiten - mit der Pest infiziert wurden. Durch funktionale Analysen konnten wir zwei verschiedene Pest-Stämme aus dieser Zeitperiode identifizieren, die sich hinsichtlich ihres Infektionspotenzials unterschieden. Um die Analyse von einem rein referenzbasierten Ansatz zu einem pangenomischen zu erweitern, in dem sämtliche Gene einer Spezies berücksichtigt werden, schlage ich in Manuskript C eine neue Methodik vor, die es erstmals erlaubt das Pangenom auch unter Einbeziehung alter Genome zu erstellen. Die Anwendung dieser Methodik auf den *Yersinia pseudotuberculosis*-Komplex, der *Y. pestis* sowie dessen Vorfahren *Y. pseudotuberculosis* und auch *Y. similis* enthält, erlaubte uns einen Einblick in die funktionale Evolution dieses Komplexes und dadurch in die Entstehung der Pest. Zudem erlauben referenzbasierte Ansätze weder die Untersuchung von bislang unentdeckten Insertionen und Deletionen noch die Aufdeckung der Genomstruktur alter Mikroben. In Manuskript D präsentiere ich eine Methodik für die *de novo*-Rekonstruktion alter *Y. pestis*-Genome und präsentiere das erste annotierte Pestgenom, das für ein Pestopfer aus dem siebzehnten Jahrhundert rekonstruiert werden konnte. Diese Methoden sind nicht auf die Anwendung bei *Y. pestis* begrenzt. In Manuskript E rekonstruiere ich das erste alte Genom des pathobiotischen Kariesbakteriums

Streptococcus mutans auf Basis eines Individuums, das vor ca. 2000 Jahren gelebt hat. Ich demonstriere, dass phylogenetische Analysen bei dieser Spezies möglicherweise nicht für das Studium ihrer Evolution geeignet sind und zeige durch die Anwendung funktionaler Analysen, die im Rahmen dieser Dissertation entwickelt wurden, das Potenzial auf, Einblicke in die Coevolution dieses Bakteriums mit seinem menschlichen Wirt zu erlangen.

Somit demonstriert diese Dissertation wie die Anwendung funktionaler analytischer Ansätze auf Genome alter Krankheitserreger neue Möglichkeiten zum Studium ihrer Entstehung sowie der Evolution ihrer Virulenz ermöglicht.

References

Aas, Jørn A., Ann L. Griffen, Sara R. Dardis, Alice M. Lee, Ingar Olsen, Floyd E. Dewhirst, Eugene J. Leys, and Bruce J. Paster. 2008. 'Bacteria of Dental Caries in Primary and Permanent Teeth in Children and Young Adults'. *Journal of Clinical Microbiology* 46 (4): 1407–17. https://doi.org/10.1128/JCM.01410-07.

Achtman, Mark, Kerstin Zurth, Giovanna Morelli, Gabriela Torrea, Annie Guiyoule, and Elisabeth Carniel. 1999. '*Yersinia Pestis*, the Cause of Plague, Is a Recently Emerged Clone of *Yersinia Pseudotuberculosis*'. *Proceedings of the National Academy of Sciences* 96 (24): 14043–48. https://doi.org/10.1073/pnas.96.24.14043.

Adler, Christina J., Keith Dobney, Laura S. Weyrich, John Kaidonis, Alan W. Walker, Wolfgang Haak, Corey J. A. Bradshaw, et al. 2013. 'Sequencing Ancient Calcified Dental Plaque Shows Changes in Oral Microbiota with Dietary Shifts of the Neolithic and Industrial Revolutions'. *Nature Genetics* 45 (4): ng.2536. https://doi.org/10.1038/ng.2536.

Ajdić, Dragana, William M. McShan, Robert E. McLaughlin, Gorana Savić, Jin Chang, Matthew B. Carson, Charles Primeaux, et al. 2002. 'Genome Sequence of *Streptococcus Mutans* UA159, a Cariogenic Dental Pathogen'. *Proceedings of the National Academy of Sciences* 99 (22): 14434–39. https://doi.org/10.1073/pnas.172501299.

Allam, Adel H. 2009. 'Computed Tomographic Assessment of Atherosclerosis in Ancient Egyptian Mummies'. *JAMA* 302 (19): 2091. https://doi.org/10.1001/jama.2009.1641.

Allam, Adel H., Randall C. Thompson, L. Samuel Wann, Michael I. Miyamoto, Abd el-Halim Nur el-Din, Gomaa Abd el-Maksoud, Muhammad Al-Tohamy Soliman, et al. 2011. 'Atherosclerosis in Ancient Egyptian Mummies: The Horus Study'. *JACC: Cardiovascular Imaging* 4 (4): 315–27. https://doi.org/10.1016/j.jcmg.2011.02.002.

Allentoft, Morten E., Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, et al. 2015. 'Population Genomics of Bronze Age Eurasia'. *Nature* 522 (7555): 167–72. https://doi.org/10.1038/nature14507.

Alves, Alessandra C., Ruchele D. Nogueira, Rafael N. Stipp, Flávia Pampolini, Antonio B. A. Moraes, Reginaldo B. Gonçalves, José F. Höfling, Yihong Li, and Renata O. Mattos-Graner. 2009. 'Prospective Study of Potential Sources of *Streptococcus Mutans* Transmission in Nursery School Children'. *Journal of Medical Microbiology* 58 (4): 476–81. https://doi.org/10.1099/jmm.0.005777-0.

Andrades Valtueña, Aida, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, et al. 2017. 'The Stone Age Plague and Its Persistence in

Eurasia'. *Current Biology* 27 (23): 3683-3691.e8. https://doi.org/10.1016/j.cub.2017.10.025.

Arbaji, A., S. Kharabsheh, S. Al-Azab, M. Al-Kayed, Z. S. Amr, M. Abu Baker, and M. C. Chu. 2005. 'A 12-Case Outbreak of Pharyngeal Plague Following the Consumption of Camel Meat, in North–Eastern Jordan'. *Annals of Tropical Medicine & Parasitology* 99 (8): 789–93. https://doi.org/10.1179/136485905X65161.

Arriaza, B. T., W. Salo, A. C. Aufderheide, and T. A. Holcomb. 1995. 'Pre-Columbian Tuberculosis in Northern Chile: Molecular and Skeletal Evidence'. *American Journal of Physical Anthropology* 98 (1): 37–45. https://doi.org/10.1002/ajpa.1330980104.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing'. *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Barberis, I., N.L. Bragazzi, L. Galluzzo, and M. Martini. 2017. 'The History of Tuberculosis: From the First Historical Records to the Isolation of Koch's Bacillus'. *Journal of Preventive Medicine and Hygiene* 58 (1): E9–12.

Barquera, Rodrigo, Thiseas C. Lamnidis, Aditya Kumar Lankapalli, Arthur Kocher, Diana I. Hernández-Zaragoza, Elizabeth A. Nelson, Adriana C. Zamora-Herrera, et al. 2020. 'Origin and Health Status of First-Generation Africans from Early Colonial Mexico'. *Current Biology* 30 (11): 2078-2091.e11. https://doi.org/10.1016/j.cub.2020.04.002.

Bedoya-Correa, Claudia María, Ramiro Javier Rincón Rodríguez, and Monica Tatiana Parada-Sanchez. 2018. 'Genomic and Phenotypic Diversity of *Streptococcus Mutans*'. *Journal of Oral Biosciences*, November. https://doi.org/10.1016/j.job.2018.11.001.

Begier, Elizabeth M., Gershim Asiki, Zaccheus Anywaine, Brook Yockey, Martin E. Schriefer, Philliam Aleti, Asaph Ogden-Odoi, et al. 2006. 'Pneumonic Plague Cluster, Uganda, 2004'. *Emerging Infectious Diseases* 12 (3): 460–67. https://doi.org/10.3201/eid1203.051051.

Belizário, José E., and Mauro Napolitano. 2015. 'Human Microbiomes and Their Roles in Dysbiosis, Common Diseases, and Novel Therapeutic Approaches'. *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.01050.

Bender, G. R., E. A. Thibodeau, and R. E. Marquis. 2016. 'Reduction of Acidurance of Streptococcal Growth and Glycolysis by Fluoride and Gramicidin': *Journal of Dental Research*, November. https://doi.org/10.1177/00220345850640021701.

Benedictow, Ole Jørgen. 2004. *The Black Death, 1346-1353: The Complete History*. Boydell Press.

Berkowitz, R. J., and P. Jones. 1985. 'Mouth-to-Mouth Transmission of the Bacterium *Streptococcus Mutans* between Mother and Child'. *Archives of Oral Biology* 30 (4): 377–79. https://doi.org/10.1016/0003-9969(85)90014-7.

Bertherat, Eric, Philippe Thullier, Jean Christophe Shako, Kathleen England, Mamadou Lamine Koné, Lorraine Arntzen, Herbert Tomaso, et al. 2011. 'Lessons Learned about Pneumonic Plague Diagnosis from Two Outbreaks, Democratic Republic of the Congo'. *Emerging Infectious Diseases* 17 (5): 778–84. https://doi.org/10.3201/eid1705.100029.

Binnewies, Tim T., Yair Motro, Peter F. Hallin, Ole Lund, David Dunn, Tom La, David J. Hampson, Matthew Bellgard, Trudy M. Wassenaar, and David W. Ussery. 2006. 'Ten Years of Bacterial Genome Sequencing: Comparative-Genomics-Based Discoveries'. *Functional & Integrative Genomics* 6 (3): 165–85. https://doi.org/10.1007/s10142-006-0027-2.

Bos, Kirsten I., Kelly M. Harkins, Alexander Herbig, Mireia Coscolla, Nico Weber, Iñaki Comas, Stephen A. Forrest, et al. 2014. 'Pre-Columbian Mycobacterial Genomes Reveal Seals as a Source of New World Human Tuberculosis'. *Nature* 514 (7523): 494–97. https://doi.org/10.1038/nature13591.

Bos, Kirsten I., Alexander Herbig, Jason Sahl, Nicholas Waglechner, Mathieu Fourment, Stephen A. Forrest, Jennifer Klunk, et al. 2016. 'Eighteenth Century *Yersinia Pestis* Genomes Reveal the Long-Term Persistence of an Historical Plague Focus'. *ELife* 5 (January): e12994. https://doi.org/10.7554/eLife.12994.

Bos, Kirsten I., Denise Kühnert, Alexander Herbig, Luis Roger Esquivel-Gomez, Aida Andrades Valtueña, Rodrigo Barquera, Karen Giffin, et al. 2019. 'Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens'. *Annual Review of Microbiology* 73 (1): 639–66. https://doi.org/10.1146/annurev-micro-090817-062436.

Bos, Kirsten I., Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, et al. 2011. 'A Draft Genome of *Yersinia Pestis* from Victims of the Black Death'. *Nature* 478 (7370): 506–10. https://doi.org/10.1038/nature10549.

Briggs, Adrian W., Udo Stenzel, Philip L. F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, et al. 2007. 'Patterns of Damage in Genomic DNA Sequences from a Neandertal'. *Proceedings of the National Academy of Sciences* 104 (37): 14616–21. https://doi.org/10.1073/pnas.0704665104.

Brooks, Daniel R., and Walter A. Boeger. 2019. 'Climate Change and Emerging Infectious Diseases: Evolutionary Complexity in Action'. *Current Opinion in Systems Biology*, •

Systems biology of model organisms • Systems ecology and evolution, 13 (February): 75–81. https://doi.org/10.1016/j.coisb.2018.11.001.

Bubeck, Sarah S., Angelene M. Cantwell, and Peter H. Dube. 2007. 'Delayed Inflammatory Response to Primary Pneumonic Plague Occurs in Both Outbred and Inbred Mice'. *Infection and Immunity* 75 (2): 697–705. https://doi.org/10.1128/IAI.00403-06.

Butler, T. 2014. 'Plague History: Yersin's Discovery of the Causative Bacterium in 1894 Enabled, in the Subsequent Century, Scientific Progress in Understanding the Disease and the Development of Treatments and Vaccines'. *Clinical Microbiology and Infection* 20 (3): 202–9. https://doi.org/10.1111/1469-0691.12540.

Cabanel, Nicolas, Christiane Bouchier, Minoarisoa Rajerison, and Elisabeth Carniel. 2018. 'Plasmid-Mediated Doxycycline Resistance in a *Yersinia Pestis* Strain Isolated from a Rat'. *International Journal of Antimicrobial Agents* 51 (2): 249–54. https://doi.org/10.1016/j.ijantimicag.2017.09.015.

Califf, Katy J., Paul S. Keim, David M. Wagner, and Jason W. Sahl. 2015. 'Redefining the Differences in Gene Content between *Yersinia Pestis* and *Yersinia Pseudotuberculosis* Using Large-Scale Comparative Genomics'. *Microbial Genomics* 1 (2). https://doi.org/10.1099/mgen.0.000028.

Caufield, P.W., G.R. Cutter, and A.P. Dasanayake. 1993. 'Initial Acquisition of Mutans Streptococci by Infants: Evidence for a Discrete Window of Infectivity'. *Journal of Dental Research* 72 (1): 37–45. https://doi.org/10.1177/00220345930720010501.

Chakraborty, Apurba, Kenneth Komatsu, Matthew Roberts, Jim Collins, Jennifer Beggs, George Turabelidze, Tom Safranek, et al. 2015. 'The Descriptive Epidemiology of Yersiniosis: A Multistate Study, 2005-2011'. *Public Health Reports (Washington, D.C.: 1974)* 130 (3): 269–77. https://doi.org/10.1177/003335491513000314.

Childs-Sanford, Sara E., George V. Kollias, Noha Abou-Madi, Patrick L. McDonough, Michael M. Garner, and Hussni O. Mohammed. 2009. '*Yersinia pseudotuberculosis* in a Closed Colony of Egyptian Fruit Bats (Rousettus Aegyptiacus)'. *Journal of Zoo and Wildlife Medicine: Official Publication of the American Association of Zoo Veterinarians* 40 (1): 8–14. https://doi.org/10.1638/2007-0033.1.

Chouikha, Iman, and B Joseph Hinnebusch. 2012. 'Yersinia–Flea Interactions and the Evolution of the Arthropod-Borne Transmission Route of Plague'. *Current Opinion in Microbiology*, Ecology and industrial microbiology/Special section: Microbial proteomics, 15 (3): 239–46. https://doi.org/10.1016/j.mib.2012.02.003.

Chouikha, Iman, and B. Joseph Hinnebusch. 2014. 'Silencing Urease: A Key Evolutionary Step That Facilitated the Adaptation of *Yersinia Pestis* to the Flea-Borne Transmission Route'. *Proceedings of the National Academy of Sciences* 111 (52): 18709–14. https://doi.org/10.1073/pnas.1413209111.

Christie, A. B., T. H. Chen, and Sanford S. Elberg. 1980. 'Plague in Camels and Goats: Their Role in Human Epidemics'. *The Journal of Infectious Diseases* 141 (6): 724–26.

Clarke, J. Kilian. 1924. 'On the Bacterial Factor in the aetiology of Dental Caries'. *British Journal of Experimental Pathology* 5 (3): 141–47.

Cohen, Mark Nathan, and George J. Armelagos. 1984. *Paleopathology and the Origins of Agriculture.* Orlando (FL): Academic Press.

Cohn JR, Samuel K. 2008. '4 Epidemiology of the Black Death and Successive Waves of Plague'. *Medical History. Supplement*, no. 27: 74–100.

Cooper, Alan, and Hendrik N. Poinar. 2000. 'Ancient DNA: Do It Right or Not at All'. *Science* 289 (5482): 1139–1139. https://doi.org/10.1126/science.289.5482.1139b.

Cornejo, Omar E., Tristan Lefébure, Paulina D. Pavinski Bitar, Ping Lang, Vincent P. Richards, Kirsten Eilertson, Thuy Do, et al. 2013. 'Evolutionary and Population Genomics of the Cavity Causing Bacteria *Streptococcus mutans*'. *Molecular Biology and Evolution* 30 (4): 881–93. https://doi.org/10.1093/molbev/mss278.

Cornelis, G. R., and H. Wolf-Watz. 1997. 'The Yersinia Yop Virulon: A Bacterial System for Subverting Eukaryotic Cells'. *Molecular Microbiology* 23 (5): 861–67. https://doi.org/10.1046/j.1365-2958.1997.2731623.x.

Cornelis, Guy R. 2002. 'The Yersinia Ysc-Yop "type III" Weaponry'. *Nature Reviews. Molecular Cell Biology* 3 (10): 742–52. https://doi.org/10.1038/nrm932.

Cowal, Lynne, Ian Grainger, Duncan Hawkins, and Richard Mikulski. 2008. *The Black Death Cemetery, East Smithfield, London*. London: Museum of London Archaeology Service.

Damgaard, Peter de Barros, Nina Marchi, Simon Rasmussen, Michaël Peyrot, Gabriel Renaud, Thorfinn Korneliussen, J. Víctor Moreno-Mayar, et al. 2018. '137 Ancient Human Genomes from across the Eurasian Steppes'. *Nature* 557 (7705): 369. https://doi.org/10.1038/s41586-018-0094-2.

De La Fuente, C., S. Flores, and M. Moraga. 2013. 'DNA From Human Ancient Bacteria: A Novel Source Of Genetic Evidence From Archaeological Dental Calculus'. *Archaeometry* 55 (4): 767–78. https://doi.org/10.1111/j.1475-4754.2012.00707.x.

Deacon, A. G., A. Hay, and J. Duncan. 2003. 'Septicemia Due to *Yersinia pseudotuberculosis*—a Case Report'. *Clinical Microbiology and Infection* 9 (11): 1118–19. https://doi.org/10.1046/j.1469-0691.2003.00746.x.

Dean, Katharine R., Fabienne Krauer, Lars Walløe, Ole Christian Lingjærde, Barbara Bramanti, Nils Chr. Stenseth, and Boris V. Schmid. 2018. 'Human Ectoparasites and the Spread of Plague in Europe during the Second Pandemic'. *Proceedings of the National Academy of Sciences* 115 (6): 1304–9. https://doi.org/10.1073/pnas.1715640115.

Devault, Alison M, Tatum D Mortimer, Andrew Kitchen, Henrike Kiesewetter, Jacob M Enk, G Brian Golding, John Southon, et al. 2017. 'A Molecular Portrait of Maternal Sepsis from Byzantine Troy'. Edited by George H Perry. *ELife* 6 (January): e20983. https://doi.org/10.7554/eLife.20983.

Dewhirst, Floyd E., Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. 2010. 'The Human Oral Microbiome'. *Journal of Bacteriology* 192 (19): 5002–17. https://doi.org/10.1128/JB.00542-10.

Didelot, Xavier, and Martin C. J. Maiden. 2010. 'Impact of Recombination on Bacterial Evolution'. *Trends in Microbiology* 18 (7): 315–22. https://doi.org/10.1016/j.tim.2010.04.002.

Ding, Wei, Franz Baumdicker, and Richard A Neher. 2018. 'PanX: Pan-Genome Analysis and Exploration'. *Nucleic Acids Research* 46 (1): e5. https://doi.org/10.1093/nar/gkx977.

Dorp, Lucy van, Pere Gelabert, Adrien Rieux, Marc de Manuel, Toni de-Dios, Shyam Gopalakrishnan, Christian Carøe, et al. 2020. '*Plasmodium vivax* Malaria Viewed through the Lens of an Eradicated European Strain'. *Molecular Biology and Evolution* 37 (3): 773–85. https://doi.org/10.1093/molbev/msz264.

Douglass, Joanna M., Yihong Li, and Norman Tinanoff. 2008. 'Association of Mutans Streptococci Between Caregivers and Their Children'. *Pediatric Dentistry* 30 (5): 375–87.

Drancourt, Michel, Gérard Aboudharam, Michel Signoli, Olivier Dutour, and Didier Raoult. 1998. 'Detection of 400-Year-Old *Yersinia pestis* DNA in Human Dental Pulp: An Approach to the Diagnosis of Ancient Septicemia'. *Proceedings of the National Academy of Sciences* 95 (21): 12637–40. https://doi.org/10.1073/pnas.95.21.12637.

Drancourt, Michel, Linda Houhamdi, and Didier Raoult. 2006. '*Yersinia pestis* as a Telluric, Human Ectoparasite-Borne Organism'. *The Lancet Infectious Diseases* 6 (4): 234–41. https://doi.org/10.1016/S1473-3099(06)70438-8.

Eisen, Rebecca J., Scott W. Bearden, Aryn P. Wilder, John A. Montenieri, Michael F. Antolin, and Kenneth L. Gage. 2006. 'Early-Phase Transmission of *Yersinia pestis* by Unblocked Fleas as a Mechanism Explaining Rapidly Spreading Plague Epizootics'. *Proceedings of the*

National Academy of Sciences 103 (42): 15380–85. https://doi.org/10.1073/pnas.0606831103.

Eisen, Rebecca J., and Kenneth L. Gage. 2009. 'Adaptive Strategies of *Yersinia pestis* to Persist during Inter-Epizootic and Epizootic Periods'. *Veterinary Research* 40 (2): 1. https://doi.org/10.1051/vetres:2008039.

Eisen, Rebecca J., Aryn P. Wilder, Scott W. Bearden, John A. Montenieri, and Kenneth L. Gage. 2007. 'Early-Phase Transmission of *Yersinia pestis* by Unblocked *Xenopsylla cheopis* (Siphonaptera: Pulicidae) Is as Efficient as Transmission by Blocked Fleas'. *Journal of Medical Entomology* 44 (4): 678–82. https://doi.org/10.1093/jmedent/44.4.678.

Ell, Stephen R. 1979. 'Some Evidence for Interhuman Transmission of Medieval Plague'. *Reviews of Infectious Diseases* 1 (3): 563–66. https://doi.org/10.1093/clinids/1.3.563.

Enscore, Russell E., Brad J. Biggerstaff, Ted L. Brown, Ralph E. Fulgham, Pamela J. Reynolds, David M. Engelthaler, Craig E. Levy, et al. 2002. 'Modeling Relationships between Climate and the Frequency of Human Plague Cases in the Southwestern United States, 1960-1997.' *The American Journal of Tropical Medicine and Hygiene* 66 (2): 186–96.

Featherstone, J. D. B., and B. E. Rodgers. 1981. 'Effect of Acetic, Lactic and Other Organic Acids on the Formation of Artificial Carious Lesions'. *Caries Research* 15 (5): 377–85. https://doi.org/10.1159/000260541.

Feldman, Michal, Michaela Harbeck, Marcel Keller, Maria A. Spyrou, Andreas Rott, Bernd Trautmann, Holger C. Scholz, et al. 2016. 'A High-Coverage *Yersinia pestis* Genome from a 6th-Century Justinianic Plague Victim'. *Molecular Biology and Evolution*, August, msw170. https://doi.org/10.1093/molbev/msw170.

Ferrari, Giada, Judith Neukamm, Helle T. Baalsrud, Abagail M. Breidenstein, Mark Ravinet, Carina Phillips, Frank Rühli, Abigail Bouwman, and Verena J. Schuenemann. 2020. 'Variola Virus Genome Sequenced from an Eighteenth-Century Museum Specimen Supports the Recent Origin of Smallpox'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1812): 20190572. https://doi.org/10.1098/rstb.2019.0572.

Formicola, Vincenzo. 1987. 'Neolithic Transition and Dental Changes: The Case of an Italian Site'. *Journal of Human Evolution* 16 (2): 231–39. https://doi.org/10.1016/0047-2484(87)90078-9.

Fukushima, H., and M. Gomyoda. 1991. 'Intestinal Carriage of *Yersinia pseudotuberculosis* by Wild Birds and Mammals in Japan'. *Applied and Environmental Microbiology* 57 (4): 1152–55. https://doi.org/10.1128/AEM.57.4.1152-1155.1991.

Galimand, Marc, Elisabeth Carniel, and Patrice Courvalin. 2006. 'Resistance of *Yersinia pestis* to Antimicrobial Agents'. *Antimicrobial Agents and Chemotherapy* 50 (10): 3233–36. https://doi.org/10.1128/AAC.00306-06.

Galimand, Marc, Annie Guiyoule, Guy Gerbaud, Bruno Rasoamanana, Suzanne Chanteau, Elisabeth Carniel, and Patrice Courvalin. 1997. 'Multidrug Resistance in *Yersinia pestis* Mediated by a Transferable Plasmid'. *New England Journal of Medicine* 337 (10): 677–81. https://doi.org/10.1056/NEJM199709043371004.

Gelabert, Pere, Marcela Sandoval-Velasco, Iñigo Olalde, Rosa Fregel, Adrien Rieux, Raül Escosa, Carles Aranda, et al. 2016. 'Mitochondrial DNA from the Eradicated European *Plasmodium vivax* and *P. falciparum* from 70-Year-Old Slides from the Ebro Delta in Spain'. *Proceedings of the National Academy of Sciences* 113 (41): 11495–500. https://doi.org/10.1073/pnas.1611017113.

Gepts, Paul. 2010. 'Crop Domestication as a Long-Term Selection Experiment'. In *Plant Breeding Reviews*, 1–44. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470650288.ch1.

Giffin, Karen, Aditya Kumar Lankapalli, Susanna Sabin, Maria A. Spyrou, Cosimo Posth, Justina Kozakaitė, Ronny Friedrich, et al. 2020. 'A Treponemal Genome from an Historic Plague Victim Supports a Recent Emergence of Yaws and Its Presence in 15 Th Century Europe'. *Scientific Reports* 10 (1): 9499. https://doi.org/10.1038/s41598-020-66012-x.

Gilbert, M. Thomas P., Jon Cuccui, William White, Niels Lynnerup, Richard W. Titball, Alan Cooper, and Michael B. Prentice. 2004. 'Absence of *Yersinia pestis*-Specific DNA in Human Teeth from Five European Excavations of Putative Plague Victims'. *Microbiology,* 150 (2): 341–54. https://doi.org/10.1099/mic.0.26594-0.

González-Ittig, Raúl E., Fabiana P. M. Carletto-Körber, Noelia S. Vera, María G. Jiménez, and Lila S. Cornejo. 2016. 'Population Genetic Structure and Demographic History of *Streptococcus mutans* (Bacteria: Streptococcaceae)'. *Biological Journal of the Linnean Society*, n/a-n/a. https://doi.org/10.1111/bij.12904.

Guellil, Meriam, Oliver Kersten, Amine Namouchi, Egil L. Bauer, Michael Derrick, Anne Ø Jensen, Nils C. Stenseth, and Barbara Bramanti. 2018. 'Genomic Blueprint of a Relapsing Fever Pathogen in 15th Century Scandinavia'. *Proceedings of the National Academy of Sciences* 115 (41): 10422–27. https://doi.org/10.1073/pnas.1807266115.

Guellil, Meriam, Oliver Kersten, Amine Namouchi, Stefania Luciani, Isolina Marota, Caroline A. Arcini, Elisabeth Iregren, et al. 2020. 'A Genomic and Historical Synthesis of Plague in 18th Century Eurasia'. *Proceedings of the National Academy of Sciences* 117 (45): 28328–35. https://doi.org/10.1073/pnas.2009677117.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. 'Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe'. *Nature* 522 (7555): 207–11. https://doi.org/10.1038/nature14317.

Haas, Christian J., Albert Zink, György Pálfi, Ulrike Szeimies, and Andreas G. Nerlich. 2000. 'Detection of Leprosy in Ancient Human Skeletal Remains by Molecular Identification of Mycobacterium Leprae'. *American Journal of Clinical Pathology* 114 (3): 428–36. https://doi.org/10.1093/ajcp/114.3.428.

Hagan, Richard W., Courtney A. Hofman, Alexander Hübner, Karl Reinhard, Stephanie Schnorr, Cecil M. Lewis, Krithivasan Sankaranarayanan, and Christina G. Warinner. 2020. 'Comparison of Extraction Methods for Recovering Ancient Microbial DNA from Paleofeces'. *American Journal of Physical Anthropology* 171 (2): 275–84. https://doi.org/10.1002/ajpa.23978.

Hagen, Stephen J., and Minjun Son. 2017. 'Origins of Heterogeneity in Streptococcus Mutans Competence: Interpreting an Environment-Sensitive Signaling Pathway'. *Physical Biology* 14 (1): 015001. https://doi.org/10.1088/1478-3975/aa546c.

Harbeck, Michaela, Lisa Seifert, Stephanie Hänsch, David M. Wagner, Dawn Birdsell, Katy L. Parise, Ingrid Wiechmann, et al. 2013. '*Yersinia pestis* DNA from Skeletal Remains from the 6th Century AD Reveals Insights into Justinianic Plague'. *PLOS Pathogens* 9 (5): e1003349. https://doi.org/10.1371/journal.ppat.1003349.

Harper, Kristin N., Molly K. Zuckerman, Megan L. Harper, John D. Kingston, and George J. Armelagos. 2011. 'The Origin and Antiquity of Syphilis Revisited: An Appraisal of Old World Pre-Columbian Evidence for Treponemal Infection'. *American Journal of Physical Anthropology* 146 (S53): 99–133. https://doi.org/10.1002/ajpa.21613.

Hinnebusch, B. Joseph. 2012. 'Biofilm-Dependent and Biofilm-Independent Mechanisms of Transmission of *Yersinia pestis* by Fleas'. In *Advances in Yersinia Research*, 237–43. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3561-7_30.

Hinnebusch, B. Joseph, Iman Chouikha, and Yi-Cheng Sun. 2016. 'Ecological Opportunity, Evolution, and the Emergence of Flea-Borne Plague'. *Infection and Immunity*, May, IAI.00188-16. https://doi.org/10.1128/IAI.00188-16.

Hinnebusch, B. Joseph, Elizabeth R. Fischer, and Tom G. Schwan. 1998. 'Evaluation of the Role of the *Yersinia pestis* Plasminogen Activator and Other Plasmid-Encoded Factors in Temperature-Dependent Blockage of the Flea'. *Journal of Infectious Diseases* 178 (5): 1406–15. https://doi.org/10.1086/314456.

Hinnebusch, B. Joseph, Amy E. Rudolph, Peter Cherepanov, Jack E. Dixon, Tom G. Schwan, and Åke Forsberg. 2002. 'Role of *Yersinia* Murine Toxin in Survival of *Yersinia pestis* in the Midgut of the Flea Vector'. *Science* 296 (5568): 733–35. https://doi.org/10.1126/science.1069972.

Höss, Matthias, Pawel Jaruga, Tomasz H. Zastawny, Miral Dizdaroglu, and Svante Paabo. 1996. 'DNA Damage and DNA Sequence Retrieval from Ancient Tissues'. *Nucleic Acids Research* 24 (7): 1304–7. https://doi.org/10.1093/nar/24.7.1304.

Hottes, Alison K., Peter L. Freddolino, Anupama Khare, Zachary N. Donnell, Julia C. Liu, and Saeed Tavazoie. 2013. 'Bacterial Adaptation through Loss of Function'. *PLOS Genetics* 9 (7): e1003617. https://doi.org/10.1371/journal.pgen.1003617.

Hübler, Ron, Felix M. Key, Christina Warinner, Kirsten I. Bos, Johannes Krause, and Alexander Herbig. 2019. 'HOPS: Automated Detection and Authentication of Pathogen DNA in Archaeological Remains'. *Genome Biology* 20 (1): 280. https://doi.org/10.1186/s13059-019-1903-0.

Humphrey, Louise T., Isabelle De Groote, Jacob Morales, Nick Barton, Simon Collcutt, Christopher Bronk Ramsey, and Abdeljalil Bouzouggar. 2014. 'Earliest Evidence for Caries and Exploitation of Starchy Plant Foods in Pleistocene Hunter-Gatherers from Morocco'. *Proceedings of the National Academy of Sciences* 111 (3): 954–59. https://doi.org/10.1073/pnas.1318176111.

Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. 'MEGAN Analysis of Metagenomic Data'. *Genome Research* 17 (3): 377–86. https://doi.org/10.1101/gr.5969107.

Huson, Daniel H., Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. 2016. 'MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data'. *PLOS Computational Biology* 12 (6): e1004957. https://doi.org/10.1371/journal.pcbi.1004957.

Kehrmann, Jan, Walter Popp, Battumur Delgermaa, Damdin Otgonbayar, Tsagaan Gantumur, Jan Buer, and Nyamdorj Tsogbadrakh. 2020. 'Two Fatal Cases of Plague after Consumption of Raw Marmot Organs'. *Emerging Microbes & Infections* 9 (1): 1878–80. https://doi.org/10.1080/22221751.2020.1807412.

Keller, Marcel, Maria A. Spyrou, Christiana L. Scheib, Gunnar U. Neumann, Andreas Kröpelin, Brigitte Haas-Gebhard, Bernd Päffgen, et al. 2019. 'Ancient *Yersinia pestis* Genomes from across Western Europe Reveal Early Diversification during the First Pandemic (541–750)'.

Proceedings of the National Academy of Sciences 116 (25): 12363–72. https://doi.org/10.1073/pnas.1820447116.

Key, Felix M., Cosimo Posth, Luis R. Esquivel-Gomez, Ron Hübler, Maria A. Spyrou, Gunnar U. Neumann, Anja Furtwängler, et al. 2020. 'Emergence of Human-Adapted *Salmonella enterica* Is Linked to the Neolithization Process'. *Nature Ecology & Evolution* 4 (3): 324–33. https://doi.org/10.1038/s41559-020-1106-9.

Khan, Mohd Danish, Hong Ha Thi Vu, Quang Tuan Lai, and Ji Whan Ahn. 2019. 'Aggravation of Human Diseases and Climate Change Nexus'. *International Journal of Environmental Research and Public Health* 16 (15): 2799. https://doi.org/10.3390/ijerph16152799.

Kleinberg, I. 2002. 'A Mixed-Bacteria Ecological Approach to Understanding the Role of the Oral Bacteria in Dental Caries Causation: An Alternative to *Streptococcus mutans* and the Specific-Plaque Hypothesis'. *Critical Reviews in Oral Biology & Medicine* 13 (2): 108–25. https://doi.org/10.1177/154411130201300202.

Kolman, Connie J., Arturo Centurion-Lara, Sheila A. Lukehart, Douglas W. Owsley, and Noreen Tuross. 1999. 'Identification of *Treponema pallidum* Subspecies *pallidum* in a 200-Year-Old Skeletal Specimen'. *Journal of Infectious Diseases* 180 (6): 2060–63. https://doi.org/10.1086/315151.

Koskiniemi, Sanna, Song Sun, Otto G. Berg, and Dan I. Andersson. 2012. 'Selection-Driven Gene Loss in Bacteria'. *PLOS Genetics* 8 (6): e1002787. https://doi.org/10.1371/journal.pgen.1002787.

Koster, Frederick, David S. Perlin, Steven Park, Trevor Brasel, Andrew Gigliotti, Edward Barr, Leslie Myers, Robert C. Layton, Robert Sherwood, and C. R. Lyons. 2010. 'Milestones in Progression of Primary Pneumonic Plague in Cynomolgus Macaques'. *Infection and Immunity* 78 (7): 2946–55. https://doi.org/10.1128/IAI.01296-09.

Krause-Kyora, Ben, Julian Susat, Felix M Key, Denise Kühnert, Esther Bosse, Alexander Immel, Christoph Rinne, et al. 2018. 'Neolithic and Medieval Virus Genomes Reveal Complex Evolution of Hepatitis B'. Edited by Stephen Locarnini. *ELife* 7 (May): e36666. https://doi.org/10.7554/eLife.36666.

Krzyściak, W., A. Jurczak, D. Kościelniak, B. Bystrowska, and A. Skalniak. 2014. 'The Virulence of *Streptococcus mutans* and the Ability to Form Biofilms'. *European Journal of Clinical Microbiology & Infectious Diseases* 33 (4): 499–515. https://doi.org/10.1007/s10096-013-1993-7.

Lapirattanakul, Jinthana, and Kazuhiko Nakano. 2014. 'Mother-to-Child Transmission of Mutans Streptococci'. *Future Microbiology* 9 (6): 807–23. https://doi.org/10.2217/fmb.14.37.

Lathem, Wyndham W., Seth D. Crosby, Virginia L. Miller, and William E. Goldman. 2005. 'Progression of Primary Pneumonic Plague: A Mouse Model of Infection, Pathology, and Bacterial Transcriptional Activity'. *Proceedings of the National Academy of Sciences of the United States of America* 102 (49): 17786–91. https://doi.org/10.1073/pnas.0506840102.

Li, Y., and P. W. Caufield. 1995. 'The Fidelity of Initial Acquisition of Mutans Streptococci by Infants from Their Mothers'. *Journal of Dental Research* 74 (2): 681–85. https://doi.org/10.1177/00220345950740020901.

Lien-teh, Wu, J. W. H. Chun, R. Pollitzer, and C. Y. Wu. 1936. 'Plague : A Manual for Medical and Public Health Workers.' *Plague : A Manual for Medical and Public Health Workers.* https://www.cabdirect.org/cabdirect/abstract/19362901460.

Lindahl, Tomas. 1993. 'Instability and Decay of the Primary Structure of DNA'. *Nature* 362 (6422): 709–15. https://doi.org/10.1038/362709a0.

Llamas, Bastien, Guido Valverde, Lars Fehren-Schmitz, Laura S. Weyrich, Alan Cooper, and Wolfgang Haak. 2017. 'From the Field to the Laboratory: Controlling DNA Contamination in Human Ancient DNA Research in the High-Throughput Sequencing Era'. *STAR: Science & Technology of Archaeological Research* 3 (1): 1–14. https://doi.org/10.1080/20548923.2016.1258824.

Loesche, W J. 1986. 'Role of *Streptococcus mutans* in Human Dental Decay.' *Microbiological Reviews* 50 (4): 353–80.

Luhmann, Nina, Daniel Doerr, and Cedric Chauve. 2017. 'Comparative Scaffolding and Gap Filling of Ancient Bacterial Genomes Applied to Two Ancient *Yersinia pestis* Genomes'. *Microbial Genomics* 3 (9). https://doi.org/10.1099/mgen.0.000123.

Lukacs, John R. 1992. 'Dental Paleopathology and Agricultural Intensification in South Asia: New Evidence from Bronze Age Harappa'. *American Journal of Physical Anthropology* 87 (2): 133–50. https://doi.org/10.1002/ajpa.1330870202.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. 'SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler'. *GigaScience* 1: 18. https://doi.org/10.1186/2047-217X-1-18.

Maixner, Frank, Ben Krause-Kyora, Dmitrij Turaev, Alexander Herbig, Michael R. Hoopmann, Janice L. Hallows, Ulrike Kusebauch, et al. 2016. 'The 5300-Year-Old *Helicobacter pylori* Genome of the Iceman'. *Science* 351 (6269): 162–65. https://doi.org/10.1126/science.aad2545.

Majander, Kerttu, Saskia Pfrengle, Arthur Kocher, Judith Neukamm, Louis du Plessis, Marta Pla-Díaz, Natasha Arora, et al. 2020. 'Ancient Bacterial Genomes Reveal a High Diversity of *Treponema pallidum* Strains in Early Modern Europe'. *Current Biology* 30 (19): 3788-3803.e10. https://doi.org/10.1016/j.cub.2020.07.058.

Mann, Allison E., Susanna Sabin, Kirsten Ziesemer, Åshild J. Vågene, Hannes Schroeder, Andrew T. Ozga, Krithivasan Sankaranarayanan, et al. 2018. 'Differential Preservation of Endogenous Human and Microbial DNA in Dental Calculus and Dentin'. *Scientific Reports* 8 (1): 9822. https://doi.org/10.1038/s41598-018-28091-9.

Margaryan, Ashot, Henrik B. Hansen, Simon Rasmussen, Martin Sikora, Vyacheslav Moiseyev, Alexandr Khoklov, Andrey Epimakhov, et al. 2018. 'Ancient Pathogen DNA in Human Teeth and Petrous Bones'. *Ecology and Evolution* 8 (6): 3534–42. https://doi.org/10.1002/ece3.3924.

Marsh, P. D. 1994. 'Microbial Ecology of Dental Plaque and Its Significance in Health and Disease'. *Advances in Dental Research* 8 (2): 263–71. https://doi.org/10.1177/08959374940080022001.

Maruvada, Padma, Vanessa Leone, Lee M. Kaplan, and Eugene B. Chang. 2017. 'The Human Microbiome and Obesity: Moving beyond Associations'. *Cell Host & Microbe* 22 (5): 589–99. https://doi.org/10.1016/j.chom.2017.10.005.

McHugh, T D, L E Newport, and S H Gillespie. 1997. 'IS6110 Homologs Are Present in Multiple Copies in Mycobacteria Other than Tuberculosis-Causing Mycobacteria.' *Journal of Clinical Microbiology* 35 (7): 1769–71.

Mégraud, F., P. Lehours, and F. F. Vale. 2016. 'The History of *Helicobacter pylori*: From Phylogeography to Paleomicrobiology'. *Clinical Microbiology and Infection* 22 (11): 922–27. https://doi.org/10.1016/j.cmi.2016.07.013.

Meng, Peiqi, Chang Lu, Qian Zhang, Jiuxiang Lin, and Feng Chen. 2017. 'Exploring the Genomic Diversity and Cariogenic Differences of *Streptococcus mutans* Strains Through Pan-Genome and Comparative Genome Analysis'. *Current Microbiology* 74 (10): 1200–1209. https://doi.org/10.1007/s00284-017-1305-z.

Morozova, Irina, Artem Kasianov, Sergey Bruskin, Judith Neukamm, Martyna Molak, Elena Batieva, Aleksandra Pudło, Frank J. Rühli, and Verena J. Schuenemann. 2020. 'New Ancient Eastern European *Yersinia pestis* Genomes Illuminate the Dispersal of Plague in Europe'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1812): 20190569. https://doi.org/10.1098/rstb.2019.0569.

Moynihan, Paula, and Poul Erik Petersen. 2004. 'Diet, Nutrition and the Prevention of Dental Diseases'. *Public Health Nutrition* 7 (1A): 201–26. https://doi.org/10.1079/phn2003589.

Mühlemann, Barbara, Terry C. Jones, Peter de Barros Damgaard, Morten E. Allentoft, Irina Shevnina, Andrey Logvin, Emma Usmanova, et al. 2018. 'Ancient Hepatitis B Viruses from the Bronze Age to the Medieval Period'. *Nature* 557 (7705): 418. https://doi.org/10.1038/s41586-018-0097-z.

Mühlemann, Barbara, Lasse Vinner, Ashot Margaryan, Helene Wilhelmson, Constanza de la Fuente Castro, Morten E. Allentoft, Peter de Barros Damgaard, et al. 2020. 'Diverse Variola Virus (Smallpox) Strains Were Widespread in Northern Europe in the Viking Age'. *Science* 369 (6502). https://doi.org/10.1126/science.aaw8977.

Nagata, E., H. Okayama, H.-O. Ito, Y. Yamashita, M. Inoue, and T. Oho. 2006. 'Serotype-Specific Polysaccharide of *Streptococcus mutans* Contributes to Infectivity in Endocarditis'. *Oral Microbiology and Immunology* 21 (6): 420–23. https://doi.org/10.1111/j.1399-302X.2006.00317.x.

Nakano, Kazuhiko, Ryota Nomura, Hirotoshi Nemoto, Takao Mukai, Hideo Yoshioka, Yasuhiro Shudo, Hiroki Hata, et al. 2007. 'Detection of Novel Serotype k *Streptococcus mutans* in Infective Endocarditis Patients'. *Journal of Medical Microbiology* 56 (Pt 10): 1413–15. https://doi.org/10.1099/jmm.0.47335-0.

Nakano, Kazuhiko, and Takashi Ooshima. 2009. 'Serotype Classification of *Streptococcus mutans* and Its Detection Outside the Oral Cavity'. *Future Microbiology* 4 (7): 891–902. https://doi.org/10.2217/fmb.09.64.

Namouchi, Amine, Meriam Guellil, Oliver Kersten, Stephanie Hänsch, Claudio Ottoni, Boris V. Schmid, Elsa Pacciani, et al. 2018. 'Integrative Approach Using *Yersinia pestis* Genomes to Revisit the Historical Landscape of Plague during the Medieval Period'. *Proceedings of the National Academy of Sciences* 115 (50): E11790. https://doi.org/10.1073/pnas.1812865115.

Nerlich, Andreas, and Raffaella Bianucci. 2020. 'Paleo-Oncology and Mummies'. In . https://doi.org/10.1007/978-981-15-1614-6_381.

Neukamm, Judith, Saskia Pfrengle, Martyna Molak, Alexander Seitz, Michael Francken, Partick Eppenberger, Charlotte Avanzi, et al. 2020. '2000-Year-Old Pathogen Genomes Reconstructed from Metagenomic Analysis of Egyptian Mummified Individuals'. *BMC Biology* 18 (1): 108. https://doi.org/10.1186/s12915-020-00839-8.

Nicklisch, Nicole, Robert Ganslmeier, Angelina Siebert, Susanne Friederich, Harald Meller, and Kurt W. Alt. 2016. 'Holes in Teeth – Dental Caries in Neolithic and Early Bronze Age

Populations in Central Germany'. *Annals of Anatomy - Anatomischer Anzeiger*, SI: Dental Morphology Research - Past meets Present, 203 (January): 90–99. https://doi.org/10.1016/j.aanat.2015.02.001.

Nomura, Ryota, Kazuhiko Nakano, Hirotoshi Nemoto, Kazuyo Fujita, Satoko Inagaki, Toshiki Takahashi, Kazuhiro Taniguchi, et al. 2006. 'Isolation and Characterization of *Streptococcus mutans* in Heart Valve and Dental Plaque Specimens from a Patient with Infective Endocarditis'. *Journal of Medical Microbiology,* 55 (8): 1135–40. https://doi.org/10.1099/jmm.0.46609-0.

Ortner, Donald J. 2003. *Identification of Pathological Conditions in Human Skeletal Remains*. 2nd ed. Cambridge: Academic Press.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. 'Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis'. *Bioinformatics* 31 (22): 3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Papagrigorakis, Manolis J., Christos Yapijakis, Philippos N. Synodinos, and Effie Baziotopoulou-Valavani. 2006. 'DNA Examination of Ancient Dental Pulp Incriminates Typhoid Fever as a Probable Cause of the Plague of Athens'. *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases* 10 (3): 206–14. https://doi.org/10.1016/j.ijid.2005.09.001.

Peres, M.A., A. Sheiham, P. Liu, F. F. Demarco, A.E.R. Silva, M.C. Assunção, A.M. Menezes, F.C. Barros, and K.G. Peres. 2016. 'Sugar Consumption and Changes in Dental Caries from Childhood to Adolescence' 95 (4): 388–94. https://doi.org/10.1177/0022034515625907.

Pezo Lanfranco, Luis, and Sabrine Eggers. 2012. 'Caries Through Time: An Anthropological Overview by Luis Nicanor Pezo Lanfranco, S. Eggers · 10.5772/38059 [PDF]'. In *Contemporary Approach to Dental Caries*, edited by Dr. Ming-Yu Li, 3–34. InTech.

Pflughoeft, Kathryn J., and James Versalovic. 2012. 'Human Microbiome in Health and Disease'. *Annual Review of Pathology: Mechanisms of Disease* 7 (1): 99–122. https://doi.org/10.1146/annurev-pathol-011811-132421.

Pha, Khavong, and Lorena Navarro. 2016. '*Yersinia* Type III Effectors Perturb Host Innate Immune Responses'. *World Journal of Biological Chemistry* 7 (1): 1–13. https://doi.org/10.4331/wjbc.v7.i1.1.

Raoult, Didier, Gérard Aboudharam, Eric Crubézy, Georges Larrouy, Bertrand Ludes, and Michel Drancourt. 2000. 'Molecular Identification by "Suicide PCR" of *Yersinia pestis* as the Agent

of Medieval Black Death'. *Proceedings of the National Academy of Sciences* 97 (23): 12800–803. https://doi.org/10.1073/pnas.220225197.

Rascovan, Nicolás, Karl-Göran Sjögren, Kristian Kristiansen, Rasmus Nielsen, Eske Willerslev, Christelle Desnues, and Simon Rasmussen. 2019. 'Emergence and Spread of Basal Lineages of *Yersinia pestis* during the Neolithic Decline'. *Cell* 176 (1): 295-305.e10. https://doi.org/10.1016/j.cell.2018.11.005.

Rasmussen, Simon, Morten Erik Allentoft, Kasper Nielsen, Ludovic Orlando, Martin Sikora, Karl-Göran Sjögren, Anders Gorm Pedersen, et al. 2015. 'Early Divergent Strains of *Yersinia pestis* in Eurasia 5,000 Years Ago'. *Cell* 163 (3): 571–82. https://doi.org/10.1016/j.cell.2015.10.009.

Ratsitorahina, M., S. Chanteau, L. Rahalison, L. Ratsifasoamanana, and P. Boisier. 2000. 'Epidemiological and Diagnostic Aspects of the Outbreak of Pneumonic Plague in Madagascar'. *Lancet (London, England)* 355 (9198): 111–13. https://doi.org/10.1016/S0140-6736(99)05163-6.

Reinhardt, Marie, Jens A. Hammerl, and Stefan Hertwig. 2018. 'Complete Genome Sequences of 10 *Yersinia pseudotuberculosis* Isolates Recovered from Wild Boars in Germany'. *Genome Announcements* 6 (19). https://doi.org/10.1128/genomeA.00266-18.

Richard, Vincent, Julia M. Riehm, Perlinot Herindrainy, Rahelinirina Soanandrasana, Maherisoa Ratsitoharina, Fanjasoa Rakotomanana, Samuel Andrianalimanana, Holger C. Scholz, and Minoarisoa Rajerison. 2015. 'Pneumonic Plague Outbreak, Northern Madagascar, 2011'. *Emerging Infectious Diseases* 21 (1): 8–15. https://doi.org/10.3201/eid2101.131828.

Rifkin, Riaan F., Surendra Vikram, Jean-Baptiste Ramond, Alba Rey-Iglesia, Tina B. Brand, Guillaume Porraz, Aurore Val, et al. 2020. 'Multi-Proxy Analyses of a Mid-15th Century Middle Iron Age Bantu-Speaker Palaeo-Faecal Specimen Elucidates the Configuration of the "Ancestral" Sub-Saharan African Intestinal Microbiome'. *Microbiome* 8 (1): 62. https://doi.org/10.1186/s40168-020-00832-x.

Roberts, Charlotte. 2018. 'The Bioarchaeology of Leprosy: Learning from the Past'. In *International Textbook of Leprosy*, edited by David M. Scollard and Tom P. Gillis. Vol. Chapter 11.1. www.internationaltextbookofleprosy.org.

Ronald Barrett, Christopher W. Kuzawa, Thomas McDade, and and George J. Armelagos. 1998. 'Emerging And Re-Emerging Infectious Diseases: The Third Epidemiologic Transition'. *Annual Review of Anthropology* 27 (1): 247–71. https://doi.org/10.1146/annurev.anthro.27.1.247.

Rouli, L., V. Merhej, P. -E. Fournier, and D. Raoult. 2015. 'The Bacterial Pangenome as a New Tool for Analysing Pathogenic Bacteria'. *New Microbes and New Infections* 7 (September): 72–85. https://doi.org/10.1016/j.nmni.2015.06.005.

Russell, Josiah C. 1968. 'That Earlier Plague'. *Demography* 5 (1): 174–84. https://doi.org/10.1007/BF03208570.

Sallares, Robert. 2006. 'Ecology, Evolution, and Epidemiology of Plague'. In *Plague and the End of Antiquity: The Pandemic of 541–750*, edited by Lester K. Little, 231–89. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511812934.014.

Sälzer, Sonja, Mohammad Alkilzy, Dagmar E. Slot, Christof E. Dörfer, Julian Schmoeckel, and Christian H. Splieth. 2017. 'Socio-Behavioural Aspects in the Prevention and Control of Dental Caries and Periodontal Diseases at an Individual and Population Level'. *Journal of Clinical Periodontology* 44 (S18): S106–15. https://doi.org/10.1111/jcpe.12673.

Schmid, Boris V., Ulf Büntgen, W. Ryan Easterday, Christian Ginzler, Lars Walløe, Barbara Bramanti, and Nils Chr Stenseth. 2015. 'Climate-Driven Introduction of the Black Death and Successive Plague Reintroductions into Europe'. *Proceedings of the National Academy of Sciences* 112 (10): 3020–25. https://doi.org/10.1073/pnas.1412887112.

Schuenemann, Verena J., Charlotte Avanzi, Ben Krause-Kyora, Alexander Seitz, Alexander Herbig, Sarah Inskip, Marion Bonazzi, et al. 2018. 'Ancient Genomes Reveal a High Diversity of *Mycobacterium leprae* in Medieval Europe'. *PLOS Pathogens* 14 (5): e1006997. https://doi.org/10.1371/journal.ppat.1006997.

Schuenemann, Verena J., Kirsten Bos, Sharon DeWitte, Sarah Schmedes, Joslyn Jamieson, Alissa Mittnik, Stephen Forrest, et al. 2011. 'Targeted Enrichment of Ancient Pathogens Yielding the pPCP1 Plasmid of *Yersinia pestis* from Victims of the Black Death'. *Proceedings of the National Academy of Sciences* 108 (38): E746–52. https://doi.org/10.1073/pnas.1105107108.

Schuenemann, Verena J., Aditya Kumar Lankapalli, Rodrigo Barquera, Elizabeth A. Nelson, Diana Iraíz Hernández, Víctor Acuña Alonzo, Kirsten I. Bos, Lourdes Márquez Morfín, Alexander Herbig, and Johannes Krause. 2018. 'Historic Treponema Pallidum Genomes from Colonial Mexico Retrieved from Archaeological Remains'. *PLOS Neglected Tropical Diseases* 12 (6): e0006447. https://doi.org/10.1371/journal.pntd.0006447.

Schuenemann, Verena J., Pushpendra Singh, Thomas A. Mendum, Ben Krause-Kyora, Günter Jäger, Kirsten I. Bos, Alexander Herbig, et al. 2013. 'Genome-Wide Comparison of Medieval and Modern *Mycobacterium leprae*'. *Science* 341 (6142): 179–83. https://doi.org/10.1126/science.1238286.

Sebbane, Florent, Donald Gardner, Daniel Long, Brian B. Gowen, and B. Joseph Hinnebusch. 2005. 'Kinetics of Disease Progression and Host Response in a Rat Model of Bubonic Plague'. *The American Journal of Pathology* 166 (5): 1427–39. https://doi.org/10.1016/S0002-9440(10)62360-7.

Sebbane, Florent, Clayton O. Jarrett, Donald Gardner, Daniel Long, and B. Joseph Hinnebusch. 2006. 'Role of the *Yersinia pestis* Plasminogen Activator in the Incidence of Distinct Septicemic and Bubonic Forms of Flea-Borne Plague'. *Proceedings of the National Academy of Sciences of the United States of America* 103 (14): 5526–30. https://doi.org/10.1073/pnas.0509544103.

Seitz, Alexander, and Kay Nieselt. 2017. 'Improving Ancient DNA Genome Assembly'. *PeerJ* 5: e3126. https://doi.org/10.7717/peerj.3126.

Shanker, Erin, and Michael J. Federle. 2017. 'Quorum Sensing Regulation of Competence and Bacteriocins in *Streptococcus pneumoniae* and *mutans'*. *Genes* 8 (1). https://doi.org/10.3390/genes8010015.

Shapiro, Beth, Andrew Rambaut, and M. Thomas P. Gilbert. 2006. 'No Proof That Typhoid Caused the Plague of Athens (a Reply to Papagrigorakis et Al.)'. *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases* 10 (4): 334–35; author reply 335-336. https://doi.org/10.1016/j.ijid.2006.02.006.

Shereen, Muhammad Adnan, Suliman Khan, Abeer Kazmi, Nadia Bashir, and Rabeea Siddique. 2020. 'COVID-19 Infection: Origin, Transmission, and Characteristics of Human Coronaviruses'. *Journal of Advanced Research* 24 (July): 91–98. https://doi.org/10.1016/j.jare.2020.03.005.

Silva Bastos, Valeria de Abreu da, Liana Bastos Freitas-Fernandes, Tatiana Kelly da Silva Fidalgo, Carla Martins, Cláudia Trindade Mattos, Ivete Pomarico Ribeiro de Souza, and Lucianne Cople Maia. 2015. 'Mother-to-Child Transmission of Streptococcus Mutans: A Systematic Review and Meta-Analysis'. *Journal of Dentistry* 43 (2): 181–91. https://doi.org/10.1016/j.jdent.2014.12.001.

Simón, Marc, Rafael Montiel, Andrea Smerling, Eduvigis Solórzano, Nancy Díaz, Brenda A. Álvarez-Sandoval, Andrea R. Jiménez-Marín, and Assumpció Malgosa. 2014. 'Molecular Analysis of Ancient Caries'. *Proceedings of the Royal Society B: Biological Sciences* 281 (1790): 20140586. https://doi.org/10.1098/rspb.2014.0586.

Slatkin, Montgomery, and Fernando Racimo. 2016. 'Ancient DNA and Human History'. *Proceedings of the National Academy of Sciences* 113 (23): 6380–87. https://doi.org/10.1073/pnas.1524306113.

Sohn, Jang-il, and Jin-Wu Nam. 2018. 'The Present and Future of de Novo Whole-Genome Assembly'. *Briefings in Bioinformatics* 19 (1): 23–40. https://doi.org/10.1093/bib/bbw096.

Spyrou, Maria A., Kirsten I. Bos, Alexander Herbig, and Johannes Krause. 2019. 'Ancient Pathogen Genomics as an Emerging Tool for Infectious Disease Research'. *Nature Reviews Genetics* 20 (6): 323–40. https://doi.org/10.1038/s41576-019-0119-1.

Spyrou, Maria A., Marcel Keller, Rezeda I. Tukhbatova, Christiana L. Scheib, Elizabeth A. Nelson, Aida Andrades Valtueña, Gunnar U. Neumann, et al. 2019. 'Phylogeography of the Second Plague Pandemic Revealed through Analysis of Historical *Yersinia pestis* Genomes'. *Nature Communications* 10 (1): 1–13. https://doi.org/10.1038/s41467-019-12154-0.

Spyrou, Maria A., Rezeda I. Tukhbatova, Michal Feldman, Joanna Drath, Sacha Kacki, Julia Beltrán de Heredia, Susanne Arnold, et al. 2016. 'Historical *Y. pestis* Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics'. *Cell Host & Microbe* 19 (6): 874–81. https://doi.org/10.1016/j.chom.2016.05.012.

Spyrou, Maria A., Rezeda I. Tukhbatova, Chuan-Chao Wang, Aida Andrades Valtueña, Aditya K. Lankapalli, Vitaly V. Kondrashin, Victor A. Tsybin, et al. 2018. 'Analysis of 3800-Year-Old *Yersinia pestis* Genomes Suggests Bronze Age Origin for Bubonic Plague'. *Nature Communications* 9 (1): 2234. https://doi.org/10.1038/s41467-018-04550-9.

Stenseth, Nils Chr, Bakyt B. Atshabar, Mike Begon, Steven R. Belmain, Eric Bertherat, Elisabeth Carniel, Kenneth L. Gage, Herwig Leirs, and Lila Rahalison. 2008. 'Plague: Past, Present, and Future'. *PLOS Medicine* 5 (1): e3. https://doi.org/10.1371/journal.pmed.0050003.

Stenseth, Nils Chr, Noelle I. Samia, Hildegunn Viljugrein, Kyrre Linné Kausrud, Mike Begon, Stephen Davis, Herwig Leirs, et al. 2006. 'Plague Dynamics Are Driven by Climate Variation'. *Proceedings of the National Academy of Sciences* 103 (35): 13110–15. https://doi.org/10.1073/pnas.0602447103.

Sun, Dongchang. 2018. 'Pull in and Push Out: Mechanisms of Horizontal Gene Transfer in Bacteria'. *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.02154.

Sun, Yi-Cheng, B. Joseph Hinnebusch, and Creg Darby. 2008. 'Experimental Evidence for Negative Selection in the Evolution of a *Yersinia pestis* Pseudogene'. *Proceedings of the National Academy of Sciences* 105 (23): 8097–8101. https://doi.org/10.1073/pnas.0803525105.

Sun, Yi-Cheng, Clayton O. Jarrett, Christopher F. Bosio, and B. Joseph Hinnebusch. 2014. 'Retracing the Evolutionary Path That Led to Flea-Borne Transmission of *Yersinia pestis*'. *Cell Host & Microbe* 15 (5): 578–86. https://doi.org/10.1016/j.chom.2014.04.003.

Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019. 'The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations'. *Cell Host & Microbe* 26 (5): 666-679.e7. https://doi.org/10.1016/j.chom.2019.08.018.

Theilade, E. 1986. 'The Non-Specific Theory in Microbial Etiology of Inflammatory Periodontal Diseases'. *Journal of Clinical Periodontology* 13 (10): 905–11. https://doi.org/10.1111/j.1600-051x.1986.tb01425.x.

Tognotti, Eugenia. 2013. 'Lessons from the History of Quarantine, from Plague to Influenza A'. *Emerging Infectious Diseases* 19 (2): 254–59. https://doi.org/10.3201/eid1902.120312.

Vågene, Åshild J., Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, et al. 2018. '*Salmonella enterica* Genomes from Victims of a Major Sixteenth-Century Epidemic in Mexico'. *Nature Ecology & Evolution* 2 (3): 520. https://doi.org/10.1038/s41559-017-0446-6.

Velsko, Irina M., James A. Fellows Yates, Franziska Aron, Richard W. Hagan, Laurent A. F. Frantz, Louise Loe, Juan Bautista Rodriguez Martinez, et al. 2019. 'Microbial Differences between Dental Plaque and Historic Dental Calculus Are Related to Oral Biofilm Maturation Stage'. *Microbiome* 7 (1): 102. https://doi.org/10.1186/s40168-019-0717-3.

Vos, Theo, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M. Abbas, Foad Abd-Allah, Rizwan Suliankatchi Abdulkader, et al. 2017. 'Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 328 Diseases and Injuries for 195 Countries, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016'. *The Lancet* 390 (10100): 1211–59. https://doi.org/10.1016/S0140-6736(17)32154-2.

Wade, William G. 2013. 'The Oral Microbiome in Health and Disease'. *Pharmacological Research*, SI:Human microbiome and health, 69 (1): 137–43. https://doi.org/10.1016/j.phrs.2012.11.006.

Wagner, David M, Jennifer Klunk, Michaela Harbeck, Alison Devault, Nicholas Waglechner, Jason W Sahl, Jacob Enk, et al. 2014. '*Yersinia pestis* and the Plague of Justinian 541–543 AD: A Genomic Analysis'. *The Lancet Infectious Diseases* 14 (4): 319–26. https://doi.org/10.1016/S1473-3099(13)70323-2.

Walløe, Lars. 2008. 'Medieval and Modern Bubonic Plague: Some Clinical Continuities'. *Medical History* 52 (S27): 59–73. https://doi.org/10.1017/S0025727300072094.

Warinner, Christina. 2016. 'Dental Calculus and the Evolution of the Human Oral Microbiome'. *Journal of the California Dental Association* 44 (July): 411–20.

Warinner, Christina, Nelly Robles García, Ronald Spores, and Noreen Tuross. 2012. 'Disease, Demography, And Diet In Early Colonial New Spain: Investigation Of A Sixteenth-Century Mixtec Cemetery At Teposcolula Yucundaa'. *Latin American Antiquity* 23 (4): 467–89.

Warinner, Christina, Alexander Herbig, Allison Mann, James A. Fellows Yates, Clemens L. Weiß, Hernán A. Burbano, Ludovic Orlando, and Johannes Krause. 2017. 'A Robust Framework for Microbial Archaeology'. *Annual Review of Genomics and Human Genetics* 18 (1): 321–56. https://doi.org/10.1146/annurev-genom-091416-035526.

Warinner, Christina, João F. Matias Rodrigues, Rounak Vyas, Christian Trachsel, Natallia Shved, Jonas Grossmann, Anita Radini, et al. 2014. 'Pathogens and Host Immunity in the Ancient Human Oral Cavity'. *Nature Genetics* 46 (4): 336–44. https://doi.org/10.1038/ng.2906.

Welker, Frido, Jazmín Ramos-Madrigal, Petra Gutenbrunner, Meaghan Mackie, Shivani Tiwary, Rosa Rakownikow Jersie-Christensen, Cristina Chiva, et al. 2020. 'The Dental Proteome of *Homo antecessor*'. *Nature* 580 (7802): 235–38. https://doi.org/10.1038/s41586-020-2153-8.

Weyrich, Laura S., Sebastian Duchene, Julien Soubrier, Luis Arriola, Bastien Llamas, James Breen, Alan G. Morris, et al. 2017. 'Neanderthal Behaviour, Diet, and Disease Inferred from Ancient DNA in Dental Calculus'. *Nature* 544 (7650): 357–61. https://doi.org/10.1038/nature21674.

Whittaker, Elizabeth, Elisa López-Varela, Claire Broderick, and James A. Seddon. 2019. 'Examining the Complex Relationship Between Tuberculosis and Other Infectious Diseases in Children'. *Frontiers in Pediatrics* 7. https://doi.org/10.3389/fped.2019.00233.

Wilbur, Alicia K., Abigail S. Bouwman, Anne C. Stone, Charlotte A. Roberts, Luz-Andrea Pfister, Jane E. Buikstra, and Terence A. Brown. 2009. 'Deficiencies and Challenges in the Study of Ancient Tuberculosis DNA'. *Journal of Archaeological Science* 36 (9): 1990–97. https://doi.org/10.1016/j.jas.2009.05.020.

Wilkinson, David A., Jonathan C. Marshall, Nigel P. French, and David T. S. Hayman. 2018. 'Habitat Fragmentation, Biodiversity Loss and the Risk of Novel Infectious Disease Emergence'. *Journal of The Royal Society Interface* 15 (149): 20180403. https://doi.org/10.1098/rsif.2018.0403.

Willerslev, Eske, and Alan Cooper. 2005. 'Review Paper. Ancient DNA'. *Proceedings of the Royal Society of London B: Biological Sciences* 272 (1558): 3–16. https://doi.org/10.1098/rspb.2004.2813.

Wong, David, Margaret A. Wild, Matthew A. Walburger, Charles L. Higgins, Michael Callahan, Lawrence A. Czarnecki, Elisabeth W. Lawaczeck, et al. 2009. 'Primary Pneumonic Plague

Contracted from a Mountain Lion Carcass'. *Clinical Infectious Diseases* 49 (3): e33–38. https://doi.org/10.1086/600818.

World Health Organisation. 2020. 'WHO | Plague – Democratic Republic of the Congo'. WHO. World Health Organization. 2020. http://www.who.int/csr/don/23-july-2020-plague-drc/en/.

Zauli, Danielle Alves Gomes. 2019. 'PCR and Infectious Diseases'. *Synthetic Biology - New Interdisciplinary Science*, April. https://doi.org/10.5772/intechopen.85630.

Zerbino, Daniel R., and Ewan Birney. 2008. 'Velvet: Algorithms for *de novo* Short Read Assembly Using de Bruijn Graphs'. *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhou, Zhemin, Inge Lundstrøm, Alicia Tran-Dien, Sebastián Duchêne, Nabil-Fareed Alikhan, Martin J. Sergeant, Gemma Langridge, et al. 2018. 'Pan-Genome Analysis of Ancient and Modern *Salmonella enterica* Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia'. *Current Biology* 28 (15): 2420-2428.e10. https://doi.org/10.1016/j.cub.2018.05.058.

Zvelebil, Marek. 2001. 'The Agricultural Transition and the Origins of Neolithic Society in Europe'. *Documenta Praehistorica* 28 (December): 1–26. https://doi.org/10.4312/dp.28.1.

Ehrenwörtliche Erklärung

Entsprechend §5 Nr. 4 der Promotionsordnung der Fakultät für Biowissenschaften der Friedrich-Schiller-Universität Jena erkläre ich hiermit,

(a) dass mir die geltende Promotionsordnung bekannt ist,

(b) dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in der Arbeit angegeben habe,

(c) dass ich alle Personen habe, die mir bei der Auswahl und Auswertung sowie bei der Herstellung des Manuskriptes unterstützt haben, in der Autorenliste der Manuskripte und den entsprechenden Danksagungen namentlich erwähnt

(d) dass ich die Hilfe einer kommerziellen Promotionsvermittlung nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,

(e) dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe,

(f) dass ich weder die gleiche, eine in wesentlichen Teilen ähnliche noch eine andere Abhandlung bei einer anderen Hochschule oder anderen Fakultät als Dissertation eingereicht habe.

Jena, den 16.12.2020

_____

Aida Andrades Valtueña

Curriculum Vitae

Personal information

Name: Aida Andrades Valtueña
Address: Max-Steenbeck-Str. 5, 07745 Jena, Germany
E-mail: aida.andrades@gmail.com; andrades@shh.mpg.de
Date of Birth: 03/05/1991
Affiliation: Max Planck Institute for the Science of Human History, Department of Archaeogenetics, Kahlaische Strasse 10, 07745 Jena, Germany

Education

Oct 2015 - Present **PhD Student**
Max Planck Institute for the Science of Human History, Department of Archaeogenetics
Accepted as PhD student at Friedrich-Schiller-Universität Jena in 22.11.2018
Thesis title: "Beyond phylogenies: advancing analytical approaches for the field of ancient pathogenomics"

Oct 2013 - Oct 2015 **MSc. Archaeological Sciences (specialisation Palaeogenetics)**
Eberhard Karls Universität Tübingen, Tübingen, Germany
Master Thesis: "Reconstruction of ancient *Clostridium botulinum* from metagenomic data"

Sep 2009 - Jul 2013 **BSc Genetics**
Universitat Autònoma de Barcelona, Cerdanyola del Valles, Spain
Exchange program at University at Buffalo, SUNY, Buffalo, NY, USA (2012-2013)

Publications

First author

Aida Andrades Valtueña, Alissa Mittnik, Felix M. Key, Wolfgang Haak, Raili Allmäe, Andrej Belinskij, Mantas Daubaras, Michal Feldman, Rimantas Jankauskas, Ivor Janković, Ken Massy, Mario Novak, Saskia Pfrengle, Sabine Reinhold, Mario Šlaus, Maria A. Spyrou, Anna Szécsényi-Nagy, Mari Tõrv, Svend Hansen, Kirsten I. Bos, Philipp W. Stockhammer, Alexander Herbig, Johannes Krause: *The Stone Age Plague and Its Persistence in Eurasia*. Current Biology 11/2017; 27(23)., DOI:10.1016/j.cub.2017.10.025

Contributions

James A. Fellows Yates, Aida Andrades Valtueña, Åshild J. Vågene, Becky Cribdon, Irina M. Velsko, Maxime Borry, Miriam J. Bravo-López, Antonio Fernandez-Guerra, Eleanor J.

Green, Shreya L. Ramachandran, Peter D. Heintzman, Maria A. Spyrou, Alexander Hübner, Abigail S. Gancz, Jessica Hider, Aurora F. Allshouse, Christina Warinner: *Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir*, bioRxiv, 09/2020, DOI: 10.1101/2020.09.02.279570

James A Fellows Yates, Thiseas Christos Lamnidis, Maxime Borry, Aida Andrades Valtueña, Zandra Fagneräs, Stephen Clayton, Maxime U Garcia, Judith Neukamm, Alexander Peltzer: *Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager* bioRxiv, 06/2020, DOI: 10.1101/2020.06.11.145615

Kirsten I. Bos, Denise Kühnert, Alexander Herbig, Luis Roger Esquivel-Gomez, Aida Andrades Valtueña, Rodrigo Barquera, Karen Giffin, Aditya Kumar Lankapalli, Elizabeth A. Nelson, Susanna Sabin, Maria A. Spyrou, Johannes Krause: *Paleomicrobiology: Diagnosis and Evolution of Ancient Pathogens* Annual Review of Microbiology 09/2019,73, DOI: 10.1146/annurev-micro-090817-062436

Maria A. Spyrou, Rezeda I. Tukhbatova, Chuan-Chao Wang, Aida Andrades Valtueña, Aditya K. Lankapalli, Vitaly V. Kondrashin, Victor A. Tsybin, Aleksandr Khokhlov, Denise Kühnert, Alexander Herbig, Kirsten I. Bos, Johannes Krause: *Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague*. Nature Communications 12/2018; 9(1)., DOI:10.1038/s41467-018-04550-9

Alissa Mittnik, Chuan-Chao Wang, Saskia Pfrengle, Mantas Daubaras, Gunita Zariņa, Fredrik Hallgren, Raili Allmäe, Valery Khartanovich, Vyacheslav Moiseyev, Mari Tõrv, Anja Furtwängler, Aida Andrades Valtueña, Michal Feldman, Christos Economou, Markku Oinonen, Andrejs Vasks, Elena Balanovska, David Reich, Rimantas Jankauskas, Wolfgang Haak, Stephan Schiffels, Johannes Krause: *The genetic prehistory of the Baltic Sea region*. Nature Communications 01/2018; 9(1)., DOI:10.1038/s41467-018-02825-9

Åshild J. Vågene, Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, Aida Andrades Valtueña, Daniel Huson, Noreen Tuross, Kirsten I. Bos, Johannes Krause: *Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico*. Nature Ecology & Evolution 01/2018; 2(3)., DOI:10.1038/s41559-017-0446-6

Gary J Iacobucci, Noura Abdel Rahman, Aida Andrades Valtueña, Tapan Kumar Nayak, Shermali Gunawardena: *Spatial and Temporal Characteristics of Normal and Perturbed Vesicle Transport*. PLoS ONE 05/2014; 9(5):e97237., DOI:10.1371/journal.pone.0097237

Scientific outreach

Aida Andrades Valtueña, Ken Massy, Marcel Keller: *Die Steinzeitpest im Lechtal. Bayerische Archäologie* 2020/3, 16–19.

Conferences

Oral presentation

Andrades Valtueña, A. et al. (2017) "The Stone Age Plague: 1000 years of Persistence in Eurasia" Conference talk, 7th European Conference on Prokaryotic and Fungal Genomics (ProkaGenomics 2017), Göttingen, Germany

Andrades Valtueña, A. et al. (2017) "The Stone Age Plague: 1000 years of Persistence in Eurasia" Conference talk, Annual Meeting of the Society for Molecular Biology and Evolution 2017 (SMBE 2017), Texas, USA

Andrades Valtueña, A. et al. (2017) "The Stone Age Plague: 1000 years of Persistence in Eurasia" Workshop presentation, One Past Health: Workshop on zoonotic diseases and ancient DNA, Max Planck Institute for Evolutionary Biology, Plön, Germany

Poster presentation

Andrades Valtueña, A. et al. (2019) “Temporal and spatial insights into the genomic evolution of *Yersinia pestis* through comparative analysis”, Congress of the European Society for Evolutionary Biology (ESEB), Turku, Finland

Andrades Valtueña, A. et al. (2019) “Comparative genomic analysis of *Yersinia pestis* and *Yersinia pseudotuberculosis* and a de novo assembly of a second pandemic plague genome provide insights into the evolution of Yersinia pestis” Poster, Annual Meeting of the Society for Molecular Biology and Evolution 2019 (SMBE 2019), Manchester, UK

Andrades Valtueña, A. et al. (2018) "De novo assembly of a Second Pandemic Plague genome and the genomic evolution of *Yersinia pestis*" Poster, 8th International Symposium on Biomolecular Archaeology (ISBA2018), Jena, Germany

Attended

Conference: Infection Diseases in the 21st Century (2018), Jena, Germany
Workshop: Linguistics, Archaeology and Genetics, SHH-MPI, Jena, Germany
Annual conference of the Association for General and Applied Microbiology (VAAM2016), Jena,Germany
7th International Symposium on Biomolecular Archaeology (ISBA2016), Oxford, UK

Invited lectures

2019 Title: La Peste en la Edad de Piedra: perspectivas arqueológicas y genéticas sobre enfermedades (pre-)históricas
Facultat de Farmacia, Universidad del Pais Vasco, Spain

Awards/Scholarships

2008 **VI Convocatoria Ajuts Botet i Sisó**,
Granted by: Universitat de Girona (2008)
Description: scholarship to cover cost for the student and the education center to carry out a research project mandatory to graduate from “Batxillerat”. Small

course in the Universitat de Girona and additional supervision by the Universitat de Girona.

2009-2013 **Beca salari Ítaca-Banco Santander**
Granted by: Universitat Autònoma de Barcelona and banco Sabadell
Description: Scholarship covering the tuition, rent expenses and a monthly salary (500E)

2012-2013 **Beca Modalitat C**
Granted by: Fundació Joan Riera i Gubau
Description: Scholarship covering tuition and living expenses for studies in the USA

Service

2016 Reviewer for Evolutionary Bioinformatics Journal

Work Experience

Oct 2015 - Present **PhD Student**
Max Planck Institute for the Science of Human History, Department of Archaeogenetics, Jena, Germany

May 2015 - Sep 2015 **HiWi**
Max Planck Institute for the Science of Human History, Department of Archaeogenetics, Jena, Germany

Nov 2014 - May 2015 **Palaeogenetics HiWi**
Eberhard Karls Universität Tubingen, Department of Geosciences, Tübingen, Germany

Sep 2011 - May 2012 **Laboratory Assistant**
University at Buffalo, Department of Biological Sciences, SUNY, Buffalo NY, USA

Acknowledgments

I would like to extend my gratitude to the following people:

To Johannes Krause, for giving me the opportunity to join the Archaeogenetics department and introducing me to the field of ancient DNA.

To Alexander Herbig who has guided me through the journey into ancient pathogenomics. Thank you for all your supervision and support, which has allowed me to become an independent researcher.

To Maria Spyrou who has inspired and supported me during my doctorate. Thank you for being a great co-worker and for all the invaluable discussion about plague, science and life.

To the LNBA team, Gunnar U. Neumann, Maria A. Spyrou, Lyazzat Musralina, Wolfgang Haak and Alexander Herbig, for being a great team and for all the discussions on the LNBA plague ecology.

To Megan Michel for being my companion in the *Streptococcus mutans* journey. Thank you for all the insightful conversations.

To the pathogen group and Archaeogenetics group for providing a great work place, all your feedback and great discussions.

To all my co-authors for allowing me to work on such precious samples, for providing insights into the archaeological context and all the creative ideas about plague transmission.

To the lab team for generating the data needed for all this work, without which none of this work could have been possible.

To the IMPRS coordinators for all the help given to navigate the university system.

I was supported in the production of this document by suggestions from Alexander Herbig and Christina Warriner. Alexander Herbig assisted me with the translation of the summary to German. James Fellows Yates helped me with the proof-reading.

On a more personal note, I would also like to thank:

My friends, Ke, Thiseas, Zandra, Rodrigo, Susanna, Kathrin, Marieke, Aditya for all the support, the great memories and trips.

My Anyi and Bet, for all those years of friendship and crazy adventures. Thank you for always being there for me and for making me smile.

Raphaela Stahl and Elizabeth Nelson for your friendship, all of your advice and support and keeping me sane during the past few months.

My family, my parents, Mari Luz and Joan, and my brother Sergi, for always believing in and supporting me.

James Fellows Yates, my partner in crime, for your unconditional support and love; for all the discussions on science, politics and strange things; for being my personal proof-reader and life coach; for being the best dad and partner.

And finally, Maia, my daughter, for reminding me what is really important and giving me the best cuddles.