Genetic Investigations into the Black Death

Genetic Investigations into the Black Death

By KIRSTEN BOS, B.Sc. Hon, M.A.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2011) Hamilton, Ontario (Anthropology)

TITILE: Genetic Investigations into the Black Death AUTHOR: Kirsten Bos, BSc. Hon., M.A. (McMaster University) SUPERVISOR: Professor H.N. Poinar NUMBER OF PAGES: xiv, 219

Abstract

This dissertation discusses molecular analyses of dental and skeletal material from victims of the Black Death with the goal of both identifying and describing the evolutionary history of the causative agent of the pandemic. Through this work, *Yersinia pestis* DNA was successfully identified in skeletal material from a well-documented Black Death burial ground, the East Smithfield cemetery of London, England (1348 -1350). The thesis presents two major methodological advancements in the field of ancient pathogen research: 1) it describes a protocol to confirm the authenticity of ancient pathogen DNA, thus circumventing tenuous issues relating to modern contaminants, and 2) it demonstrates the applicability of DNA capture methods to isolate ancient pathogen DNA from its complex metagenomic background common to ancient DNA extracts. The dissertation is comprised of three publications. The first, submitted to the journal BMC Systems Biology, describes a computational software program for oligo design that has applications to PCR, and capture techniques such as primer extension capture (PEC) and array-based capture. The second manuscript, published in the Proceedings of the National Academy of Sciences, presents a novel capture technique for retrieval of the *pestis*-specific pPCP (9.6kb) plasmid which can be used as a simple screening tool for the presence of *Y. pestis* DNA in ancient remains, and describes a method for authenticating ancient pathogen DNA. The third paper, published in the journal Nature, presents a draft genome of *Yersinia pestis* isolated from the individuals of the East Smithfield collection, thus presenting the first ancient pathogen genome in published literature. Evolutionary changes as they relate to phylogenetic placement and the evolution of virulence are discussed within an anthropological framework.

Acknowledgements

Beyond the many years of hard work that were required of me to complete this document, it simply would not have been possible without the support of many influential people. I thank Hendrik Poinar for his years of tutelage along with his continued faith in both me and the project, for acquainting me with the adaptive and opportunistic aspects of scientific research, and for teaching me that disappointment is simply a matter of perception. I thank Ann Herring for recognising strength in me when I didn't see it in myself, and for demonstrating that professionalism and friendship can come hand in hand. I'm grateful to all past and present members of the McMaster Ancient DNA Centre who shepherded me through the experimental design features of the project, and provided me with excellent training and scientific advice. I am indebted to Johannes Krause for recognising the potential of the project, for granting me the opportunity to work alongside him at the Max Planck Institute in Leipzig, and for his devotion to seeing the project to completion, even when the powers of Spider-Man were needed. I thank Verena Schuenemann for her impressive and infectious dedication to our shared work. I hold the contributions and comments of all coauthors of my publications in high regard, and I fully appreciate the concerted group effort that made these publications a success. I issue a special 'thank you' to Stephen Forrest for his hours happily spent at a computer terminal implementing design features imposed by me, and his enthusiasm for learning the intricacies of DNA capture methods necessary for writing *PrimerClique* source code appropriately. I thank my committee members Ann Herring, Brian Golding, and external examiner Anne Stone for constructive comments on the dissertation, and I thank Ben Evans and the late Shelley Saunders for their support in earlier phases of my PhD research. Additionally I also thank the late Bill White for granting me access to the material from the East Smithfield collection, Allison Sekuler for expediting my readmission process, and Svante Pääbo for use of his facilities at the Max Planck Institute for Evolutionary Anthropology, especially when timelines were originally optimistic.

In addition to posing intellectual challenges, my years as a PhD student tested my emotional strength time and time again. Overcoming these obstacles would not have been possible without the love and unconditional support of my parents Tony and Ferne Bos, and my loving partner (and recent husband) Stephen Forrest. I thank my parents for their listening ears and for their wisdom evident in their sound advice. I thank Stephen for showing enthusiasm for my work (even on days when I did not share it!),

for his unyielding belief in my abilities, and for his resounding love that could make it seem like the sun was shining on even the darkest of days.

For financial support I thank the Social Sciences and Humanities Research Council of Canada for three years of funding through the Canada Graduate Scholarship programme, the Department of Anthropology of McMaster University for a graduate fellowship, the School of Graduate Studies for travel support, Hendrik Poinar for research, travel, and personal funding, and Stephen Forrest for support throughout the degree, but especially for the post-thesis-submission trip to New York City and the awesome boots.

TABLE OF CONTENTS

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 3

Chapter 4

Declaration of Academic Achievement

Paper 1: "*PrimerClique: Open Source software for design of multiplex primers and array probes for contiguous template coverage"*

KIB and SAF conceived of the program. KIB designed the interface. SAF wrote the source code. KIB wrote the manuscript. HNP provided support.

The computational program was originally conceived in the summer of 2009, and was completed in September 2011.

Paper 2: "*Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death"*

K.B., S.D., B.K.C., J.W.W., D.J.D.E., W.W., J.K., and H.N.P. designed research; V.J.S., K.B., S.D., S.S., J.J., A.M., S.F., J.W.W., and J.K. performed research; B.K.C. and W.W. contributed new reagents/analytic tools; V.J.S., K.B., J.K., and H.N.P. analyzed data; and K.B., J.K., and H.N.P. wrote the paper.

Wet lab procedures and skeletal sampling carried out by K. Bos took place between September 2008 and December 2010. Analysis and manuscript preparation continued until March 2011.

Paper 3: "*A draft genome of* Yersinia pestis *from victims of the Black Death*"

KIB, SND, SS, and JW performed skeletal sampling. KIB, VJS, and JK carried out laboratory work. HAB, KIB, JK, MM, and HNP designed experiments. KIB, GBG, JK, HNP, VJS, and NW analysed the data. BKC, DAH, DJDE, and JBM provided valuable interpretations. PB provided technical support. KIB, JK, and HNP wrote the paper.

Wet lab procedures carried out by K. Bos took place in February and March of 2011. Analysis and manuscript preparation continued until September 2011.

CHAPTER 1: INTRODUCTION

The last several decades have introduced human populations to an historically unprecedented number of emerging or re-emerging infectious diseases, mostly facilitated by anthropogenic factors such as globalization in terms of trade and human travel, changes in local ecology, and antibiotic resistance (Barrett *et al*, 1998). The effectiveness of preventative strategies and treatments will depend on our accurate understanding of the various factors involved in disease emergence (McDonald *et al*, 1989; Relman, 2011). The process of disease emergence, however, is complicated by the complex interplay of the environment, vector dynamics, genetic changes, and overall host susceptibility that will exert influence on pathogen transmission and disease severity. With regard to microbial evolution, pathogens have much shorter generational times than their hosts, and their enhanced ability to share genetic material with their microbial neighbours makes them well-equipped to accumulate genetic changes that may confer advantages in terms of transmissibility and evasion of host defences. Genetic changes tend to accumulate faster in virulence regions (McDonald *et al*, 1989), and the litany of examples demonstrating mutation as enhancing pathogen infiltrations in plant host models suggests that the influence of pathogen evolution may be an important driving force in emerging infections (Joosten *et al*, 1994; reviewed in Woolhouse *et al*, 2002), and thus should be investigated to further our conception of virulence and host adaptation. The paucity of similar examples in a mammalian model,

however, implies a more complicated relationship between animals and communicable diseases.

To better understand factors involved in disease transmission and severity in humans, statistics on health status derived from archaeological populations may be a valuable resource for students of host-pathogen coevolution. Our current epidemiological transition parallels the experience of early sedentary communities, where the change from nomadic to urban lifestyles opened new ecological niches for infectious diseases to occupy. As such, analyses of infectious diseases in archaeological populations can be of great value to elucidate the dynamics of host-pathogen relationships in our new era of emerging infections. Several examples exist in the historical literature of diseases that, upon first introduction to new populations, brought cataclysmic mortality (Baum and Bar-Gal, 2003). Exposure to novel infectious diseases as a result of contact between European and New World populations are thought to have contributed to increased severity of disease on both sides as naive populations met with new communicable diseases from their purportedly better adapted hosts. The initial recorded severity of treponemal infections in 15th century Europe is potentially ascribed to novel exposures from reservoirs in New World populations (discussed in Sherman, 2006), though this pales in comparison to the devastating impact that the wide variety of Old World diseases had on American Aboriginal populations post contact (Dobyns, 1993; Black , 1992).

The Black Death of the mid-14th century provides a seminal example of an emerging disease, bringing about high mortality that later declined in successive waves of infection over the following four centuries in outbreaks that are presumed to have resulted from the same agent. The reasons for initial severity, subsequent decline in disease prevalence, and epidemiological changes in terms of the age structure of susceptible individuals (Cohn, 2002) are at present elusive, since the relative influences of environment, biology, and social factors on transmission of infectious agents and host susceptibility to microbial insults are difficult to evaluate in past populations.

The field of palaeopathology has matured in recent years through the publication of several texts that carefully document and characterise skeletal lesions associated with specific diseases in an attempt to make possible inferences of health status from gross morphology of skeletal remains (Aufderhide and Rodriguez-Martin, 1998; Ortner, 2003). However, even with the impressive literature and the recommendation of standard methods for lesion characterisation (Buikstra and Cook, 1980; Buikstra and Ubelaker, 1994), experts in the field continue to acknowledge the limitations of skeletal assessments and the elusive nature of a true differential diagnosis from gross analysis of archaeological material (Lovell, 2000). Understanding population health is significantly hindered by the fact that not all infectious diseases or other consequences of poor health leave traces on the skeleton. This “osteological paradox”, where morphologically normal skeletons may in fact represent the weaker members of a population (Wood *et al*, 1992), highlights a significant limitation inherent in

morphological assessments of skeletal remains and their suitability for assessing population-level health status in the past.

With the acknowledgement of these limitations, ancient DNA is an attractive tool for anthropologists to extract additional information regarding the health status of past human populations, especially with reference to infectious disease insults. The presence of ancient pathogen DNA in archaeological remains has been used for differential diagnosis (Mays and Taylor, 2002), for identification of infections that leave no gross skeletal indicators (Drancourt *et al*, 1998), and to understand the evolutionary history of some of our most notorious human parasites (Taubenberger *et al*, 2004). Despite these advances, however, accusations of contamination from modern sources (Prentice *et al*, 2004) and inabilities to replicate results independently in other laboratories (Gilbert *et al*, 2004; Barnes and Thomas, 2006) have done much to discredit the approach.

Broadly, the goals of the current work were to establish robust methodological and analytical procedures to isolate and characterise ancient microbial DNA from archaeological remains with adherence to the strictest current requirements for ancient DNA authentication (Pääbo *et al*, 2004). With such standards in place, we will be equipped to begin discussions regarding the relationship between bacterial genetics and population-level disease manifestation, thus establishing the possible contributions of pathogen evolution on the emergence of specific diseases. The Black Death of 1347 –

1351 was chosen as my model infection owing to the availability of well-documented skeletal material from the pandemic, the demonstration of putative ancient *Y. pestis* DNA in skeletal samples (Raoult *et al*, 2000), and pronounced differences in mortality profiles between the Black Death, subsequent resurgences of disease through the Renaissance (Cohn, 2002), and modern *Y. pestis* insults.

1.1 History of the Black Death

The Black Death is a colloquial term used to describe a massive pandemic in the mid 14th century. Medieval European chroniclers were convinced that the disease came to them from the East, though few reliable documents remain to describe the disease's presence east of modern Ukraine, hence preventing knowledge of its original source and the associated human losses prior to its entry into Europe. In his thorough analysis of original documents, Gottfried (1983) surmised that the disease first affected human populations somewhere in the environs of the Gobi Desert in the late 1320s. The disease then followed humans westward along established land-based trade routes, with efficient dissemination offered by nomadic Mongol groups who traveled on horseback (McNeill, 1976). This terrestrial path ultimately ended in Caffa (modern Ukraine), where Genoese merchants contracted the disease and brought it with them via sea voyage to their native Italy. Based on the writings of several chroniclers, the plague first entered Messina, Sicily in October 1347. The most widely-cited account of

the pestilence's introduction to Europe comes from the writings of the Piacenzan chronicler Gabriele de Mussis in his work *Historia de Morbo* (see Horrox, 1994 for a modern translation of the original). Mussis recounts the tale of Genoese traders in Caffa immersed in a battle against the local Muslim and Tartar populations, and just when it appeared the battle had been lost by the Genoese, the Tartar army retreated, overcome by a horrible illness. As a parting gift, however, the Tartars catapulted corpses over the city walls, and according to the Italians, it was this act that acquainted them all too closely with the pestilence. This weaponisation of the Black Death as an implement of medieval bioterrorism has done much to fuel its perception as a symbol of fear for unknown and uncontrollable disease in modern popular culture (Stenseth *et al*, 2008). When the Genoese retreated and sailed back to Italy, most occupants of the vessel had purportedly perished en route, and the remainder docked in Italian ports "as if they had brought evil spirits with them" (Horrox 1994, page 19). It should be noted, however, that Mussis was not present in Messina to greet the arriving Genoese merchants; rather, he received these stories second hand from travelers and returning sailors, hence suggesting that his account may equally represent the traveling of rumours as much as the disease itself (Horrox, 1994; Carmichael, 2008). Moreover, Mussis' accusation of a non-Christian Mongol population as the source of the disease would have been favourably accepted by European populations at the time, so widespread social prejudice may have influenced the faithfulness of actual historical events.

Whatever the context of its entry into Western Europe, dissemination of the disease was rapid. Italy was a commercial centre for the Mediterranean basin, so introduction into its ports in Messina, and soon after in Genoa, allowed for its efficient spread via established sea trade routes. Port cities were always associated with high mortality, and interconnections between trade routes permitted multiple points of entry for the pestilence, as it was then called, to many locales (Gottfried, 1983). Although the general population attributed the calamity to divine punishment, medical faculties sought astrological explanations, or turned to the teachings of Galen and Hippocrates to search for treatments and preventative measures, though to no avail. The pestilence traveled relentlessly through Europe and by 1351 it was retreating back towards the less densely populated Russian Steppe (Benedictow, 2004). Surprisingly, the disease experience was not uniform across Europe. Mortality levels were not always consistent between different communities to the extent that some major urban centres reported casualties on the order 90%, whereas other locations were mysteriously spared of the disease entirely (Gottfried 1983; Cohn, 2002). Furthermore, reported differences in disease symptoms ranging from single cutaneous lesions and stupor to diffuse rashes, hysteria, and extreme thirst are provisionally suggestive of a disease with an oddly heterogeneous presentation (Cohn, 2002).

In all, mortality in Western Europe attributable to the pestilence is popularly estimated to be between 30 and 50%, though Benedictow (2004), in his thorough analysis of demographic records, speculates that 60% might be a more accurate figure.

Such high mortality precipitated massive social and economic reform most evident in the collapse of the feudal system, which had dominated Europe in the previous centuries. Peasant revolts against low pay and taxing work conditions fostered capitalistic sentiments, which ultimately set the stage for major cultural reforms manifest as the European Renaissance of the 14th through 17th centuries (Herlihy, 1997). Although this introductory event ended in the early 1350s, the pestilence returned every 10 to 20 years as local epidemics, though with unique epidemiological characteristics, which suggests pronounced changes in the disease pattern over time (Cohn, 2002). Mysteriously the last of these regular outbreaks occurred in the mid 18th or perhaps the early 19th century (Cohn, 2008), after which time the pestilence retreated from the European experience.

Conventionally the Black Death is assumed to be an outbreak of bubonic plague caused by the bacterium *Yersinia pestis*. Much of this is derived from medieval descriptions of the symptoms, which inevitably describe the presence of large swellings, approximately the size of an egg, most commonly in the groin region, but also in the axilla and neck (Horrox, 1994). This single diagnostic feature is regarded as a hallmark indication that the medieval disease was a widespread outbreak of bubonic plague (Gottfired, 1983; Benedictow, 2004). Much of what we know about *Y. pestis* infections derived from information obtained by the Indian Plague Commission of 1902 issued by the British government (Indian Plague Commission, 1902). Bubonic plague is a zoonotic infection stemming from interactions of immune sylvatic rodents with commensal

rodent populations, such as the infamous black rat *Rattus rattus*. The bacterium is carried in the gut of the rat flea *X. cheopis*. Bacterial proliferation in the flea's gut produces a biofilm that seals the valve connecting the oesophagus to the stomach, thereby preventing the flea from intaking a blood meal. In a desperate attempt to feed, the starving flea vomits successive blood meals, each time depositing live *Y. pestis* bacteria on the dermis of its host. Once in the host interstitium, the bacteria evade immune defences and rapidly travel to the lymph node where bacterial proliferation occurs. The exquisitely painful enlarged lymph nodes, or buboes, are the classic manifestation of bubonic plague. In rare circumstances, bacteria can infiltrate the bloodstream leading to rapid sepsis and colonisation of the lung. This inevitably results in necrosis of lung tissue, where the sufferers clear their lungs by coughing a bloody sputum rich in bacteria, thus making the infection highly communicable.

The Black Death is not the only historical pandemic thought to have been caused by bubonic plague. The Plague of Justinian (541 – 750 AD), arguably implicated in the collapse of the Eastern Roman Empire, supposedly displayed the same suite of symptoms. This has prompted modern medical historians to propose a three wave model of bubonic plague infection, where the Black Death is the second of three successive waves of plague, beginning with the Plague of Justinian, and culminating in the 20th century outbreak that is still responsible for 2000 cases per year (WHO, 2011), and is regarded as re-emerging in certain areas (Chanteau *et al*, 1998). The conundrum of the cataclysmic mortality of the Black Death compared to the much lower mortality

rates from contemporary *Y. pestis* infections, however, calls for explanations to account for epidemiological discrepancies. Herlihy (1997) regards medieval Europe as being ripe for an infectious disease outbreak owing to the decreased health undoubtedly lingering from the Great Famine of the early 14th century, continued food shortages resulting from growing population sizes, and social tensions resulting from the unstable feudal economic system. Others have contended that environment played a large role in plague dissemination: low seasonal temperatures and high rainfalls in the mid 14th century parallel the climatic conditions conducive to large outbreaks of bubonic plague in modern populations (Enscore *et al*, 2002; Parmenter *et al*, 1999; Stenseth *et al*, 2006; Samia *et al*, 2011; Xu *et al*, 2011).

These models, however, seem overly simplistic and factually implausible to some. Graham Twigg (1984) was the first to publish a large document suggesting that the rat-flea model of transmission seemed incompatible with the epidemiological data available for pestilence-associated mortality of the 14th century, and proposed the notion that *Bacillus anthracis*, or anthrax, was the responsible agent. This seminal work has been overshadowed by the more widely cited book by Scott and Duncan *The Biology of Plagues* (2001) where the authors argue that the disease patterns and symptoms are in better accord with a haemorrhagic fever such as an ebola-like filovirus. For Scott and Duncan, the lack of agreement between the symptoms of modern plague and those recorded from Black Death victims prompted a more thorough investigation regarding the aetiology of the medieval disease. Their argument rests most notably

upon the observation that 1) the Black Death disseminated through populations more quickly and with higher mortality than modern bubonic plague; 2) it had high mortality in areas where climatic conditions were not conducive to large rat populations such as Iceland, Norway, and Scotland where human populations were still devastated by the Black Death; and 3) quarantine efforts initiated in Italy in 1377 were successful for a 40-day isolation, but unsuccessful for a 30-day isolation, implying a communicable disease with a 30 – 39 day incubation period (discussed further in Gensini *et al*, 2004). In addition, Stephens *et al* (1998) set the stage for the haemorrhagic fever hypothesis by suggesting that the CCR5-Δ32 mutation, which confers resistance to HIV infiltration of T cells, maps precisely to the geographic areas affected by high Black Death-associated mortality. The coalescence date of 700 years suggested that the mutation was under strong selection, with the Black Death as the obvious candidate for a selective sweep. The confirmation by Mescas *et al* in 2004 that the CCR5 mutation did not confer increased resistance to *Y. pestis* insults in a mouse model was used as further evidence to support Scott and Duncan's initial hypothesis (Duncan *et al*, 2005), even though others urged caution when interpreting analogous laboratory results (Elvin *et al*, 2004), and even demonstrated evidence to the contrary (Styer *et al*, 2007). Ultimately, the CCR5-Δ32 connection with the Black Death has fallen from fashion with the suggestion that initial coalescence dates were far too recent, and that the mutation in question has not experienced recent positive selection (Sabeti *et al*, 2005).

Similar to Scott and Duncan, and using many of the same arguments to support his position, Samuel Cohn (2002; 2008) has also posited that the Black Death must have been caused by a virus rather than a bacterial insult. His premises focus strongly on the fact that the writings of historical chroniclers imply that the disease was highly contagious between individuals thus circumventing the requirement of an insect vector, and that episodes of peak mortality would occur in seasonably unfavourable conditions for the fertility cycle of fleas, such as the extreme heat of summer and the coldest months of winter. The potency of his argument, however, rests on his refusal to assign a known aetiologic agent to the Black Death, claiming rather that the disease was likely caused by a communicable virus that has been lost to the pages of history.

Epidemiological differences as determined from skeletal assemblages are well documented to distinguish the Black Death from modern scourges of bubonic or pneumonic plague (DeWitte and Wood, 2008), thus providing further evidence that the diseases differ in some significant way. Specifically, modern scourges of bubonic plague are considered to be non-selective based on age, sex, and underlying health status, whereas the Black Death provisionally appears to have been selective based on frailty, thus targeting the supposed subset of the population with underlying conditions that adversely affect health status such as non-specific infection and evidence of growth arrest presumably due to nutritional insufficiency. Absent from the discussion, however, is the notion that more affluent individuals and the healthier demographic aged 15 – 24.99 years would have had greater opportunities to flee large urban centres

during the disease, hence introducing biases in disease exposure (Margerieson and Knüsel, 2002; Cohn, 2002). Regardless, these biases went unnoticed by many of the chroniclers who almost unanimously report a non-selective mortality from the disease in the 14th century (discussed in Cohn, 2002).

1.2 Contributions of Ancient DNA and Other Molecular Investigations

With the inaugural study identifying *Y. pestis* DNA in the dental pulp of purported plague victims from Marseilles in 1720, the book on the causative agent of the Black Death was closed in the opinion of some (Drancourt *et al*, 1998). Analogous results obtained in skeletal samples from the 14th century (Raoult *et al*, 2000) provisionally provided further evidence to implicate *Y. pestis* in the pandemic. However, the inability to reproduce such data under conditions that satisfy current requirements for ancient DNA authenticity (Gilbert *et al*, 2004; Pääbo *et al*, 2004) raised scepticism regarding the reliability of initial reports (Wood and DeWitte-Avina, 2003; Prentice *et al*, 2004). The fact that the authors of the original publications obtained their results in a modern microbiology laboratory coupled with the fact that they neglected to use an engineered positive control that could be distinguished from potentially contaminant wildtype sequences, was accepted as evidence that caution should be observed in reliance upon their results (Prentice *et al*, 2004). Regardless, claims that *Y. pestis* DNA has been identified in pestilence-associated catastrophe burials of the Black Death (Drancourt *et*

al, 2007; Tran *et al*, 2011) and the Plague of Justinian (Weichmann and Grupe, 2005; Weichmann *et al*, 2010) continued to be made. In addition, an immunological method that detects the presence of a *Y. pestis* virulence protein has generated positive results in several skeletal collections (Pusch *et al*, 2007; Bianucci *et al*, 2008; Bianucci *et al*, 2009; Kacki *et al*, 2011), however the sensitivity and lack of false positives with this method have yet to be adequately tested and quantified.

The most reliable ancient DNA investigation to date is that of Haensch and colleagues (2010). This study was concerned not only with mere identification of the ancient organism, but also with providing some measure of its phylogenetic placement. This study successfully demonstrated that ancient *Y. pestis* sequences obtained from three different emergency mass burials dated to the 14th century had an ancestral position when compared to modern *Y. pestis* sequences. This ancestral placement adds authenticity to the results, but does little to address the continued objections raised by sceptics. Graham Twigg (2003) contends that confirmation of *Y. pestis* DNA in a few skeletons excavated from mass burials does little to implicate this bacterium as the sole force in a pandemic that claimed the lives of an estimated 30 million people in only five years. This investigation also does little to address the noted epidemiological differences between modern and ancient forms of *Y. pestis* infections (Wood *et al*, 2003), and the ecological arguments raised against transmission via a rat-borne flea such as high mortality in regions where climatic conditions were not conducive to large

domestic rodent populations that would be required for dissemination of plague in the bubonic form (Scott and Duncan, 2001; Cohn, 2002).

Bacterial phenotype is known to influence virulence potential (Joosten *et al*, 1994), hence explanations for discrepancies between ancient and modern *Y. pestis* infections may lie in the ancient genome. PCR investigations are ill-suited for such analyses from ancient templates since they are by necessity limited to characterising short DNA segments. Advancement of molecular capture techniques to isolate target molecules from their complex metagenomic backgrounds (Briggs *et al*, 2009; Maricic *et al*, 2010; Reich *et al*, 2010; Burbano *et al*, 2010) has made full genomic investigations feasible, fast, and for the most part economical. These investigations have set a new standard for ancient DNA research where analyses of long reconstructed sections of ancient genomes are the new academic currency. Thus far, capture methods with ancient pathogens have not been attempted. Only with full genomic data from ancient pathogens can we evaluate the influence of bacterial genetics and phenotype on disease severity and manifestation, and contribute to discussions regarding the influence of microbial adaptation in host-pathogen coevolution models. The future in this area of research is bright, as this dissertation will demonstrate.

1.3 Goals of the Current Work

1. *Confirm or deny the presence of Y. pestis DNA in the remains of Black Death victims*

Proper identification of the aetiologic agent of the Black Death was established as a primary goal. My *a priori* assumption supported the prevailing view that the Black Death was caused by a diffuse outbreak of bubonic plague caused by the Gram-negative bacillus *Y. pestis*. *Y. pestis* is a recent human pathogen, diverging from its soil-dwelling *Y. pseudotuberculosis* progenitor within the last 20,000 years (Acthman *et al*, 1999). Surprisingly few genetic differences distinguish these two bacterial species, though *Y. pestis* evolution is generally characterised by pronounced gene loss (Zhou *et al*, 2004; Pouillot *et al*, 2008), and the acquisition of two virulence-associated plasmids, namely the pMT1 assumed to have been acquired from *Salmonella enterica serovar typhi* (Prentice *et al*, 2001), and the pPCP1 of unknown origin. The uniqueness of the pPCP1 plasmid to *Y. pestis* makes it a good candidate for molecular screening, as the presence of genetic signatures that match this plasmid are unlikely to derive from the ubiquitous microbial contaminants common to ancient DNA extracts (Poinar *et al*, 2006).

To carry out the above analyses, I was fortunate to have access to skeletal material from the East Smithfield burial ground of London, England. Documentary evidence clearly indicates the purchase of this land in late 1348 or early 1349 for the express purpose of depositing victims of the pestilence that had quickly exhausted local parish cemeteries (Hawkins, 1990). Historical sources indicate that the pestilence

entered London in the fall of 1348 (Gottfried, 1983). The fact that only a fraction of the land purchased at East Smithfield was used for human interment suggests the use of the burial ground toward the end of the epidemic, when mortality rates had declined earlier than expected. Regardless, an estimated 2500 individuals were interred in this cemetery between 1348 and 1350 in either individual burials or in one of two main mass burial trenches (Grainger *et al*, 2008). Land developments in the mid-1980s presented the opportunity for archaeologists from the Museum of London to excavate the site, and between the years 1986 and 1988, archaeologists were successful in exhuming the remains of 634 nearly intact skeletons, which now comprise the East Smithfield (or Royal Mint) collection currently curated at the Museum of London. This large collection has been the subject of many skeletal and dental analyses over the years (Waldron, 2001; Roberts and Grauer, 2001; Margerison and Knüsel, 2002; Hillson *et al*, 2005; DeWitte and Wood, 2008). The abundance of material and its unequivocal association with pestilence-associated mortality in London during the Black Death makes it ideally suited for molecular investigations into the causative agent of the Black Death.

To supplement the East Smithfield collection, material from the pre-Black Death cemetery St. Nicholas Shambles (11^{th} – 12^{th} centuries) was collected for use as a negative control.

2. *Investigate the utility of molecular capture techniques for evaluations of bacterial genetics and temporal changes in virulence potential*

The observation that small changes in pathogens can exert significant influence on virulence potential (Joosten *et al*, 1994) highlights the importance of characterising virulence regions in the ancient organism. This allows us to begin discussions of the relative influence of various factors relating to disease emergence and mortality with respect to *Y. pestis* infections. In addition to several virulence factors inherited from its enteric pathogen ancestor *Y. pseudotuberculosis* (Huang *et al*, 2006), *Y. pestis* is equipped with several newly acquired mechanisms of evading host immune responses and making its way to the lymph node, the site where bacterial proliferation occurs (Zhang *et al*, 2008; reviewed in Perry and Fetherston, 1997 and Anisimov *et al*, 2004). The *pestis*-specific pPCP1 plasmid harbours the plasminogen activator gene, which has a dual role both in primary pneumonic infections and in encouraging bacterial dissemination during bubonic episodes (Lathem *et al*, 2007). It is theoretically possible that changes in one or more of these known virulence regions may have been responsible for greater transmissibility of the pathogen in the medieval era. Only ancient DNA data can address that possibility. Furthermore, ancient DNA data may be able to provide insights into adaptive changes in virulence regions acquired since the 14th century, and those required for the organism to change from an avirulent sylvatic pathogen (Song *et al*, 2004) to one capable of causing pandemic human infection on the scale of the Black Death. Together these data may provide an indication of how

selection for virulent traits has operated in this organism, which is important for the establishment of appropriate treatment strategies and for our understanding of the complex web of factors influencing host-pathogen relationships.

3. *Determine the phylogenetic placement of the ancient strain via genomic analysis*

The traditional view of *Y. pestis* phylogeography proposed by Devignat (1954) assumed that the modern geographical distribution of different *Y. pestis* lineages, at that time defined by physiological attributes, paralleled their historical associations with human populations. As such, Devignat proposed a three lineage (biovar) model, whereby the "Antiqua" variant was responsible for the Plague of Justinian, the "Medievalis" variant was associated with the Black Death and subsequent scourges up until the 18th century, and the "Orientalis" type was implicated in modern *pestis* outbreaks of the so-called "third wave". While the biovar model was considered doctrine for several decades, the model did not accord with results coming from genetic studies. Full genomic sequences from all biovars and one non-human *Y. pestis* (Parkhill *et al*, 2001; Deng *et al*, 2002; Song *et al*, 2004) permitted the identification of several phylogenetically important genomic positions. Achtman and colleagues (2004) considered these positions to develop a new scheme of strain placement where Orientalis sequences tended to fall in "branch 1", Medievalis in "branch 2", and *Y. pestis* previously identified as Antiqua were distributed between the two. Sequencing of additional genomes (Chain *et al*, 2006;

Garcia *et al*, 2007; Eppinger 2009; Eppinger, 2010) permitted better resolution of phylogenetic relationships, and coupled with additional SNP-typing for almost 1000 positions, Morelli and colleagues (2010) extrapolated that *Y. pestis* likely emerged as a human pathogen in China sometime in the last 8000 years and radiated to different areas of the world through global trade routes.

The use of modern data to extrapolate positions of ancestral nodes places a heavy reliance upon mutation rates that are merely approximations of actual biological rates of change. Bona fide ancient sequence data from reliable historical contexts is the most valuable resource for proper phylogenetic reconstruction. In the case of *Y. pestis*, the most fruitful methodological approach would be to acquire full ancient genomic data, which would permit precise dating of phylogenetic nodes, hence permitting more accurate calculations for rates of genetic change. This information can then be used to better determine the relationship between the "three waves" of purported *Y. pestis* infection, and to characterise the genetic changes that have accrued in its 600+ years of evolution as a human pathogen. In addition, genomic reconstructions from ancient sequences may permit evaluations of genome architecture, which are known to change and undergo fixation at an alarming rate (Auerbach *et al*, 2007; Liang *et al*, 2010).

1.4 Contents of this dissertation

The dissertation is comprised of three publications written to address the above goals. Each paper presents a necessary methodological advancement to both identify and characterise long reassembled genetic stretches of ancient pathogen DNA obtained via molecular capture techniques. Similar methods have been used in the capture of ancient hominin nuclear genomes (Burbano, 2010; Reich, 2010). The publications in this dissertation present the first time such techniques have been applied to the acquisition of full pathogen genomes isolated from their complex metagenomic background, therefore providing a major methodological advancement to the field of palaeopathology and ancient pathogen research. Full genomes from pathogens derived from archaeological populations will allow us to evaluate how pathogens have adapted over time to their human hosts, thus permitting evaluations of the role that pathogen genetics plays in the landscape of determinants for population health as they related to infectious disease in the past.

Paper 1:

Title: "*PrimerClique: Open Source software for design of multiplex primers and array probes for contiguous template coverage*"

Author list: **Kirsten I. Bos**, Stephen A. Forrest, and Hendrik N. Poinar

Publication status: submitted to BMC Systems Biology, [pending]

Description: Custom designed primer compatibility software that was used to allocate PCR primers into chemically compatible groups for multiplex reactions in Paper 2. Additional features were added to make this a general use program, now called PrimerClique. The program currently supports tasks for unidirectional primer design intended for applications such as primer extension capture (PEC, Briggs *et al*, 2009), PCR primer pair design yielding maximum coverage of a given template in multiplex reactions, and probe design for DNA capture arrays (Hodges *et al*, 2007; Hodges *et al*, 2009).

Paper 2:

Title: "*Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death*".

Author list: **Kirsten Bos***, Verena J. Schueneman*, Sharon DeWitte, Sarah Schmedes, Joslyn Jamieson, Alissa Mittnik,Stephen Forrest, Brian K. Coombes, James W. Wood, David J. D. Earn, William White, Johannes Krause and Hendrik N. Poinar.

Publication status: *Proceedings of the National Academy of Sciences*, 2011, doi/10.1073/pnas.1105107108. Accepted July 22, 2011.

Description: Report on a capture method that is successful for retrieval of ancient pathogen DNA, a method of authenticating pathogen DNA in ancient specimens, and analysis of the full pPCP1 plasmid of ancient *Y. pestis*. This paper demonstrated, for the first time in the published literature, that DNA capture techniques can be successfully applied to isolate pathogen molecules from ancient DNA extracts, and that sufficient amounts of preserved DNA could be accessed from dental pulp to permit reconstruction of full genetic segments of ancient organisms. The genetic homogeneity in the plasmid between ancient and modern forms indicates that this plasmid was not responsible for increased virulence in the ancient strain.

Paper 3:

Title: "*A draft genome sequence of Yersinia pestis from victims of the Black Death*"

Author list: **Kirsten I. Bos***, Verena J. Schuenemann*, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, Sharon N. DeWitte, Matthias Meyer, Sarah Schmedes, James Wood, David J. D. Earn, D. Ann Herring, Peter Bauer, Hendrik N. Poinar & Johannes Krause

Publication status: *Nature*, in press. Accepted September 9, 2011.

Description: Report on full genome capture of ancient *Y. pestis* from individuals securely dated to 1348 – 1350. This paper discusses an array-based capture method, which differs from the one used in paper 2. Coverage, genome architecture via reference-guided assembly, single nucleotide changes, and phylogenetic placement of the ancient sequence are discussed. At our current resolution, the minimal number of genetic differences between ancient and modern strains and the lack of a single unique derived mutation in the medieval organism provisionally suggests that the differences in severity of the Black Death as compared to modern *Y. pestis* infections may not be the result of genetic differences in the bacterium. This invites discussion of alternate

explanations for increased disease severity in the past such as population level susceptibility in terms of underlying health status, host exposure to other co-circulating infections which might exacerbate *Y. pestis* insults, social conditions conducive to disease exposure and infection, and climatic conditions that might favour large vector populations, hence facilitating transmission and disease exposure.

CHAPTER 2:

PrimerClique: Open Source software for design of multiplex primers and array probes for contiguous template coverage

Kirsten I. Bos[1] , Stephen A. Forrest[1] , and Hendrik N. Poinar[1,2]

(1) McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main Street West, Hamilton, ON, L8S 4L8, Canada

(2) Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main Street West, Hamilton, ON, L8S 4L8, Canada

2.1 Author contributions

KIB and SAF conceived of the program. KIB designed the interface. SAF wrote the source code. KIB wrote the manuscript. HNP provided support.

2.2 Abstract

Multiplex design to maximize template coverage and minimize the number of parallel reactions poses computational challenges that can only be accommodated using approximate algorithms. We offer a software package implementing such an algorithm, with the aim of designing oligos that provide maximal tiling while distributing these oligos into the fewest number of chemically compatible groups for multiplex applications. Our program is equally suited to PCR primer pairs and unidirectional oligos for various amplification and molecular capture applications. Array probe design for templates in the mega base pair range is also supported via a convenient graphical interface.

2.3 Background

The multiplex approach, where several primers are combined in parallel reactions, continues to be a useful tool to minimize experimental costs, set-up time, and most importantly DNA sample amounts. This method is commonly used in health diagnostics where screening for several pathogens can be accomplished in a single step [1,2,3], and is equally valuable for the amplification of full genes in an individual organism [4]. In our new genomic era, it persists as a simple and economical tool for contig closure in investigations of genetic rearrangements [5,6] and in full genome closure [7]. Furthermore, multiplex reactions are ideally suited to investigations of degraded molecules such as genomic reconstructions from ancient DNA [8], and those involving small target templates, especially with reference to RT-PCR reactions for genomic investigations of RNA viruses [9]. Within these contexts, multiplex amplifications can increase target template numbers for efficient conversion into DNA libraries for subsequent high-throughput sequencing [10,9]. In addition to amplifications, multiplexes composed of unidirectional primers have proven to be a valuable approach for methods of DNA capture of target molecules from a complex metagenomic background, such as in the primer extension capture (PEC) method [11].

Primer design for the above tasks can be a laborious process since achieving maximal tiling and coverage of a given template is often critical. Additionally, multiplex reactions are inherently subject to tradeoffs between coverage and multiplexing level

[12] because of the opposing objectives of minimizing both the number of reactions necessary to achieve a desired coverage goal and the number of primer-primer interactions within a reaction. An ideal solution to this problem would be maximal tiling of a template with primers partitioned into the minimum number of chemically compatible groups. This problem poses difficult computational challenges since finding a minimal number of such groups is equivalent to the graph-theoretic Clique Cover problem which is known to be **NP**-complete in the general case. While achieving a globally optimal solution may not be computationally feasible, the use of approximation algorithms for Clique Cover can still grant us many of the benefits of high coverage and multiplexing level.

Here we present PrimerClique, oligo design software that makes use of these approximations to produce chemically compatible oligo sets that offer optimal contiguous tiling of a given template while minimizing the number of parallel reactions. This general-use program was originally developed for direct multiplex sequencing [10] in which overlapping primer pairs are designed to cover an entire template; however, it is equally suited for other applications requiring complete amplification of a template via overlapping primer pairs, such as long-range PCR amplification of full genomic regions for subsequent high-throughput sequencing [13] or the manufacture of PCR products for subsequent use as DNA capture probes [14]. PrimerClique also employs a similar approach to generate contiguous unidirectional non-overlapping primers for use in enrichment reactions such as PEC [11]. In addition to primer design, PrimerClique is

also equipped with an array design feature to produce densely tiled array probes in a user-friendly environment. Array capture is popularly used to isolate long contiguous stretches of template molecules permitting reconstruction of large genetic segments, including full genomes [15,16,17]. The program incorporates previously employed stringency criteria to discourage capture of highly repeated genomic elements, and incorporates novel design features to limit design of probes that might bind to non-specific templates. Probes are designed in mere seconds and can be evaluated in a visual interface, a feature currently absent in popular web-based probe design packages.

Similar programs exist to perform primer allocations into chemically compatible groups without tiling, though often with design parameters specific to narrow applications [18,19,20]. By contrast, those that do perform tiling [21] are not concerned with contiguous coverage and they approximate chemical compatibility using BLAST searches as opposed to alignment calculations. To our knowledge PrimerClique is the first open-source primer design software that attempts both dense primer tiling and allocation of primers into a minimum number of groups based on primer compatibility. Primer interactions within a reaction are discouraged by user-defined stringency parameters employed in their design, and primers that may interact with non-target templates such as high-throughput sequencing adaptors are avoided. For certain combinations of input parameters the resulting multiplex allocation is provably optimal; for others it is a best-effort approximation.

2.4 Implementation and Features of PrimerClique

PrimerClique is a publicly-accessible Open-Source program, available via http://primerclique.sourceforge.net/ for Windows, MacOS, and Linux platforms. Primer design features of PrimerClique are based on Primer3 version 2.2.3 [22], which is included in the PrimerClique distribution. A basic flow chart for the functions and steps in PrimerClique is shown in Fig. 1. Oligo design projects can be saved with all settings and intermediate results as PrimerClique's custom format (*.fps) at any point in the design process. As outlined below, PrimerClique can perform 3 tasks:

1) Primer multiplex partitioning by compatibility (without template)

In its most basic function, PrimerClique can distribute pre-designed primers into different suggested multiplex PCR reactions based on user-defined compatibility criteria, including melting temperature (*Tm*) differences between primers and primer compatibility scores computed via local and global alignments, taken as a measure of potential primer pair interactions. In this application, PrimerClique accepts input of either unidirectional primers or primer pairs in FASTA format. Primers entered as a pair are always kept together in suggested multiplex distributions, thus overall compatibility of the reaction is influenced by existing forward and reverse primer interactions. Melting temperature is calculated for all primers using the Santa Lucia nearest neighbour method [23] adapted from Primer3, and the allowable *Tm* range for a multiplex primer set is defined by the user. Compatibility between all primers is

evaluated based on either alignment score or thermodynamic parameters (Figure 2), and assignment into multiplex groups is made based on user-defined alignment or thermodynamic cut-offs.

2) Primer design for multiplex applications (unidirectional or PCR primers)

The multiplex design feature is intended for templates of approximately 25Kb for primer pairs, and 50kb for unidirectional primers. In the first step, PrimerClique uses the Primer3 engine to generate a large list of suitable candidate primers meeting user-defined design criteria. The user must first provide a template in either FASTA or GenBank file format, or enter the base pair sequence directly into the program. Individual base pairs or full stretches can be marked for exclusion from the template if primer binding or amplification is not desired for a given region, such as avoidance of intronic regions in studies of exome capture [24]. To discourage primer binding to non-target templates that are likely contained in the DNA sample, a custom mispriming library can be used to exclude primers that contain sequence similarity detected by alignment score. This is a helpful feature when working with DNA suspected to be under-represented in a DNA sample such as pathogen templates where host genomic DNA is expected to be present in greater quantities.

Multiplex-based amplifications have recently been performed on immortalized DNA products from libraries designed for high throughput sequencing applications [25], hence PrimerClique allows for exclusion of candidate primers that show similarities to

adaptor sequences as measured by global end-anchored alignments. This will discourage the design of primers potentially prone to adaptor binding, an especially relevant design feature for unidirectional primer generation where non-specific binding to adaptors will decrease capture efficiency.

To generate multiplexes, PrimerClique first selects oligos from a candidate list generated by Primer3 to generate a number of dense tilings of the template using a well-known linear-time dynamic programming algorithm for weighted interval scheduling [26] which guarantees that all tilings thus produced have equivalent and optimal coverage of the template. The number of tilings recorded at this stage can be lowered by the user through the "Number of Algorithmic Branches" setting in order to return results more quickly with the possible cost of more multiplexes.

For each optimal tiling thus generated, PrimerClique assigns the oligos in this tiling to different multiplex groups based on compatibility restrictions defined by the user. These can be based on *Tm* difference, maximum multiplexing level, local/global alignment scores (Figure 2), or thermodynamic alignment scores [22]. Alignments are computed using Primer3 in a fashion similar to what Primer3 uses internally for avoidance of primer self-dimerization and primer-pair interactions. The resulting primer set has the maximum possible template coverage for the generated candidate list. When user-specified compatibility criteria are based only on *Tm* difference and maximum multiplexing level, the number of multiplexes returned is the minimum

possible. Alternatively when these criteria include alignments, PrimerClique employs a heuristic approach towards partitioning based on well-known greedy approximation algorithms for Clique Cover and its complementary Graph k-Coloring problem. These include the Welsh-Powell algorithm [27] and Brélaz's *DSATUR* algorithm [28]. The multiplex assignment for the current tiling thus corresponds to the smallest clique cover found using this approach.

At its conclusion, the algorithm returns the tiling with the smallest number of multiplexes. When alignments are used this is still an approximate solution, but has the potential for better results than simply selecting an arbitrary optimal tiling and applying the heuristic multiplex assignment.

3) Design of densely tiled array probes

Probes are normally designed as densely overlapping oligos along the template, with one probe every few base pairs. PrimerClique can generate probes of 50-200bp accommodating user-defined tiling density, where tiling density is defined as the distance between probe 5' ends. Tiling density of the probes is a user-defined numeric value, and will apply uniformly across the template. If a user desires a different tiling density for specific genomic regions, these should be entered as separate templates to incorporate the unique probe design parameters.

Since probe design is based on size and tiling density as opposed to molecular parameters such as *Tm* or GC content, Primer3 is not used as the underlying engine for oligo design in this application. Instead PrimerClique makes use of a custom external library which efficiently manages probe design for a specified probe size and uniform tiling density across the target template. Tiling density is required to be less than or equal to the probe size. To avoid redundancy, duplicate probes are consolidated in the generated output so every resulting probe is unique.

Exclusion of repetitive genomic elements is accomplished by methods previously described [16]. With this approach, an average genome frequency score for the probe is calculated by determining each 15-mer sequence within a putative probe, obtaining the associated frequency count for that 15-mer for the entire genome, then averaging the counts for all such 15-mers in the probe. If the average number of 15-mer repeats in the template is above a pre-set threshold, the probe is excluded. A threshold value of 100 has been used for mammalian genomes [17], and is included as our default value. The genomic frequency table, which records the count of all 15-mers present in the template or its reverse complement, is computed immediately before probe generation. For reasons of file size the table is not stored in the PrimerClique project file along with the generated probes, and is recomputed as necessary. A sample array output is shown in Figure 3.

Although probes designed via PrimerClique can be uploaded into array design programs, arrays may also be designed within the PrimerClique interface. This offers an advantage over other array design programs since PrimerClique permits greater user-defined flexibility of probe allocation into different array groups in terms of numbers of probes per group and molecular compatibility, if desired.

For Agilent microarrays, probes are spatially immobilized on glass slides thus eliminating artifacts of probe-probe interactions. In addition, *Tm* for probe-template hybridization need not be considered in design parameters because established protocols recommend that capture reactions occur at high annealing temperatures (65°C) to make binding as specific as possible [17]. With in-solution-based approaches (e.g. SureSelect, NimbleGen), probes can be distributed between different tubes, thus capture reactions may be performed with probe-specific annealing temperatures if desired. Calculations of probe *Tm* will be most relevant where the probe is equal in length or shorter than the presumed size of the DNA template, such that hybridization would occur along the full length of the probe. If probe grouping based on *Tm* is desired, *Tm* for probes can be calculated via Primer3's special-purpose *Tm* code for long oligos, and probes can be assigned to different groups based on *Tm* compatibility. Since popular in-solution based capture approaches use RNA probes to minimize probe dimerization, PrimerClique does not currently evaluate probes for compatibility based on alignment or thermodynamic properties.

For either of the above capture methods, if complete tiling and template coverage is not critical, probes matching high throughput sequencing adaptors can be removed, thus discouraging co-enrichment of non-target templates. This is accomplished by removing all probes that contain 15-mer sequences present in adaptors. Output for probe design is in FASTA, compressed FASTA (`.fasta.gz`), or CSV (comma-separated values) suitable for import in a spreadsheet application.

2.5 Results and Discussion

Appendix 1 provides a comprehensive list of output that PrimerClique generates for its different tasks, using design criteria appropriate for the intended applications. For testing we use an AMD Phenom II X4 2.80 GHz system with 6.0 GB RAM running Windows 7 (64-bit).

1) Design of unidirectional primers

Implementation of this feature is designed for applications such as PEC for capture of fragmentary DNA templates, where non-overlapping contiguous primers are tiled along the template with as little distance as possible between adjacent primers [11]. In our example, we allowed an overlap of 10bp for primers, though this is fully customizable by the user. On our test system, this task is best used for templates of maximum of 50Kb, since larger templates require a greater number of primers to achieve contiguous

coverage. The default ranges of values for primer design are intentionally relaxed to allow for a large number of candidates, hence allowing greater flexibility in primer set generation. Using these default values with the human mitochondrial genome (16.5Kb) as a template, and using default alignment values for primer set generation would permit 95.12% (629 primers) coverage from a theoretical maximum of 99.41% coverage (from 53,734 candidate primers) in 14 multiplex reactions. Constraining the number of candidate primers to 10,000 lowers the overall coverage to 77.15% (498 primers in 9 multiplexes). Although this tiling is lower, the largest distance between two primers is only 53bp, which is still suitable for capture of templates that are approximately 70bp or greater. To test the effect of adaptor filtering, we excluded the 454 Shotgun adaptors, since these are the longest adaptors and hence would generate the most restrictive results. Using the default alignment criteria, Shotgun adaptor filtering removes nearly 50% of all candidates; however, PrimerClique is still able to generate a primer set that will provide 88.04% coverage in 13 multiplex reactions.

To test the effect of introducing a mispriming library, we used the Sudan Ebola sequence with a human genomic mispriming library to generate theoretically non-promiscuous primers. The mispriming library removed 9% of candidate oligos, and would permit coverage of 80.93% of the template when constraining the number of candidate oligos to 20,000. Filtering for Illumina adaptors removed 36% of candidates, slightly fewer than for the larger Shotgun adaptors. Regardless, filtering for the

mispriming library and the Illumina adaptors would still permit coverage of 72% of the template using default parameters for compatibility based on alignment score.

2) PCR primer generation

The increased number of compatibility checks required for PCR multiplex (Figure 2) introduces greater restrictions on the number of primers that can be included in each reaction. Regardless, PrimerClique produces impressive results for long-range PCR applications. The full human mtDNA genome with a 2000-3000 bp product size can be amplified in 2 reactions yielding 99.53% coverage. This is reduced to 2 multiplexes and 99.02% coverage when filtering for Shotgun adaptors is introduced.

Although smaller product sizes may offer higher coverage since the beginning and end of the template can be better accommodated in primer design, they pose a greater challenge since more primers are required to obtain comparable coverage, thus necessitating a greater number of multiplexes if compatibility criteria are preserved. In our human mtDNA genome example, reducing the product size to 200-300 bp increases the number of primers by an order of magnitude, therefore requiring 13 multiplexes to achieve 99.63% coverage. This is reduced to 11 multiplexes when 454 Shotgun adaptor filtering is used, though this still provides 99.21% coverage. The computational limits imposed by the Clique Cover problem become amplified as product size decreases. Reducing the temperature difference between forward and reverse primers can facilitate placement of primers into compatible groups based on *Tm*, though many

alignment checks required for compatibility between two primer pairs continue to pose difficulties. Implementation of default compatibility criteria for the human mitochondrial genome with a 70-80bp product with 2°C between forward and reverse primers requires 748 primers distributed between 42 multiplex reactions to achieve 96.58% coverage, which is potentially too many reactions for practical use. For such situations, a user may prefer to distribute primers into different groups while considering *Tm* differences only. Implementing a *Tm* difference of at most 5°C per reaction significantly reduced the number of multiplexes required, permitting 96.58% of the template to be covered in only 7 multiplexes with a maximum of 150 primer pairs per reaction, though primer compatibility is undetermined.

3) Array probe generation

Two templates were selected for array generation to demonstrate different design features. The larger 25.6Mb human Y chromosome was chosen because it is the largest template that could be handled with the memory available on our test machine. In addition the small 5.5Kb *E. coli* genome was selected since we could reasonably expect to fit the entire genome on a single one million feature array. The array design feature was demonstrated with 70bp probes for the Y chromosome and *E. coli* genome. Removal of probes that correspond to repetitive regions as identified via 15-mer filtering proved to be influential since it removed 11% (586,418/5,130,530) of probes for the Y chromosome which is known to have many repeated regions, and only 0.07 %

(673/921,397) for the *E. coli* genome, which is not as repetitive. Adaptor filtering differs slightly in this application compared to the unidirectional and PCR primer approaches in that probes are removed only if they share identical 15-mer sequences with the template, which is obviously less restrictive than the alignment compatibility algorithm followed in the other applications. Shotgun adaptors were again used for demonstration because of their longer length. Filtering removed only 88 probes for the Y chromosome and 9 probes for *E. coli* genome, though this number would be increased if longer custom adaptors were chosen, such as those that include unique identifiers [29]. Filtering of duplicate probes removed 2% (over 80,000) and 1% (over 6000) probes for the Y chromosome and *E. coli* genome, respectively. Filtering via these methods does not significantly decrease overall coverage of probes against the template, hence they are highly recommended. In all, the Y chromosome could be captured over 5 one million feature arrays at 5bp tiling, and the *E. coli* genome could be captured on one array with 6bp tiling. For a *Tm*-specific in-solution hybridization, the Y chromosome could be captured in 11 multiplexes with 5°C partitioning, and the *E. coli* genome in 9 with 4°C partitioning.

2.6 Conclusions

We have provided a user-friendly open-source software package for use with multiple oligo design applications where complete tiling and chemical compatibility is

accommodated. In future we hope to extend our features for degenerate probe design and larger input templates. For array probe design, we hope to accommodate different tiling densities across a template more easily by allowing the user to select regions and assign tiling density. Screening out promiscuous probes by building a genomic frequency table for an input mispriming library and removing probes whose weighted score is above a set threshold would also be fruitful.

2.7 Availability and requirements

PrimerClique is supplied as a binary distribution for Windows and MacOS; it requires Windows XP or later or MacOS 10.6+. It is additionally available for other platforms, including Unix/Linux, in source form. All versions of PrimerClique require Java 6.

PrimerClique requires 100MB of space for installation and use and 1GB of RAM; performance of some tasks especially probe design is tied heavily to available RAM.

PrimerClique is distributed under Version 2 of the GNU General Public License (GPL) and includes a redistributed copy of Primer3, which is also distributed under the GPL.

2.8 Acknowledgements

We thank past and current members of the McMaster Ancient DNA Centre for beta testing and helpful comments about design features.

Supporting files:

Table 1: Data obtained from PrimerClique oligo design for various tasks

See appendix 1

2.9 References

1. Mathisen M, Strand TA, Sharma BN, Chandyo RK, Valentiner-Branth P, Basnet S, Adhikari RK, Hvidsten D, Shrestha PS, Sommerfelt H: **RNA viruses in community-acquired childhood pneumonia in semi-urban Nepal; a cross-sectional study**. *BMC Medicine* 2009, **7**(35), [http://dx.doi.org/10.1186/1741-7015-7-35].
2. Schmitt M, de Koning MNC, Eekhof JA, Quint WG, Pawlita M: **Evaluation of a novel multiplex HPV genotyping assay for HPV types in skin warts**. *J. Clin. Microbiol.* 2011, **49**(9):3262-3267, [http://dx.doi.org/10.1128/JCM.00634-11].
3. Roeb E, Arndt M, Jansen B, Schumpelick V, Matern S: **Simultaneous determination of matrix metalloproteinase (MMP)-7, MMP-1, -3, and -13 gene expression by multiplex PCR in colorectal carcinomas**. *Int J Colorectal Dis* 2004, **19**(6):518-524, [http://dx.doi.org/10.1007/s00384-004-0592-6].
4. Zong Z, Lü X: **Characterization of a New SCC*mec* Element in *Staphylococcus cohnii***. *PLoS one* 2010, **5**(11), [http://dx.doi.org/10.1371/journal.pone.0014016].
5. Carraro D, Camargo A, Salim A, Grivet M, Vasconcelos A, Simpson A: **PCR-Assisted Contig Extension: Stepwise Strategy for Bacterial Genome Closure**. *BioTechniques* 2003, **34**(3):626-632.
6. Ticha I, Kleibl Z, Stribrna J, Kotlas J, Zimovjanova M, Mateju M, Zikan M, Pohlreich P: **Screening for genomic rearrangements in *BRCA1* and *BRCA2* genes in Czech high-risk breast/ovarian cancer patients: high proportion of population specific alterations in *BRCA1 gene***. *Breast Cancer Research and Treatment* 2010, **124**(2):337-347, [http://dx.doi.org/10.1007/s10549-010-0745-y].
7. Loman NJ, Snyder LAS, Linton JD, Langdon R, Lawson AJ, Weinstock GM, Wren BW, Pallen MJ: **Genome Sequence of the Emerging Pathogen Helicobacter canadensis**. *Journal of Bacteriology* 2009, **191**(17):5566-5567, [http://dx.doi.org/10.1128/JB.00729-09].
8. Krause J, Dear P, Pollack J, Slatkin M, Spriggs H, Barnes I, Lister A, Ebersberger I, Pääbo S, Hofreiter M: **Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae**. *Nature* 2006, **439**:724-727, [http://dx.doi.org/10.1038/nature04432].
9. Towner JS, Sealy TK, Khristova ML, no CGA, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, Okware S, Lutwama J, Bakamutumaho B, Kayiwa J, Comer JA, Rollin PE, Ksiazek TG, Nichol ST: **Newly Discovered Ebola Virus Associated with Hemorrhagic Fever Outbreak in Uganda**. *PLoS Pathogens* 2008, **4**(11), [http://dx.doi.org/10.1371/journal.ppat.1000212].

10. Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M: **Direct multiplex sequencing (DMPS)-a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA**. *Genome Research* 2009, **19**(10):1843-1848.
11. Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovi\'c D, Zeljko Ku\'can, Gusi\'c I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Pääbo S: **Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes**. *Science* 2009, **325**(5938):318-321, [http://dx.doi.org/10.1126/science.1174462].
12. Rachlin J, Ding C, Cantor C, Kasif S: **Computational tradeoffs in multiplex PCR assay design for SNP genotyping**. *BMC Genomics* 2005, **6**(102), [http://dx.doi.org/10.1186/1471-2164-6-102].
13. Knaus B, Cronn R, Liston A, Pilgrim K, Schwartz MK: **Mitochondrial genome sequences illuminate maternal lineages of conservation concern in a rare carnivore**. *BMC Ecology* 2011, **11**(10), [http://dx.doi.org/10.1186/1472-6785-11-10].
14. Maricic T, Whitten M, Pääbo S: **Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products**. *PloS one* 2011, **5**(11), [http://dx.doi.org/10.1371/journal.pone.0014004].
15. Burbano H, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PLF, Xuan Z, Rooks M, Bhattacharjee A, Brizuela L, Albert FW, de la Rasilla M, Fortea J, Rosas A, Lachmann M, Hannon GJ, Pääbo S: **Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture**. *Science* 2010, **328**(5979):723-725, [http://dx.doi.org/10.1126/science.1188046].
16. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR: **Genome-wide *in situ* exon capture for selective resequencing**. *Nature Genetics* 2007, **39**(12):1522-1527, [http://dx.doi.org/10.1038/ng.2007.42].
17. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ: **Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing**. *Nature Protocols* 2009, **4**(6):960-974, [http://dx.doi.org/10.1038/nprot.2009.68].
18. Rachlin J, Ding C, Cantor C, Kasif S: **MuPlex: multi-objective multiplex PCR assay design**. *Nucleic Acids Research* 2005, **33**:W544-W547, [http://dx.doi.org/10.1093/nar/gki377].
19. Kirsten H, Dienst S, Emmrich F, Ahner P: **CalcDalton: a tool for multiplex genotyping primer design for single-base extension reactions using cleavable primers**. *BioTechniques* 2006, **40**:158-162.

20. Kaplinski L, Andreson R, Puurand T, Remm M: **MultiPLX: automatic grouping and evaluation of PCR primers**. *Bioinformatics* 2005, **21**(8):1701-1702, [http://dx.doi.org/10.1093/bioinformatics/bti219].
21. Gervais AL, Marques M, Gaudreau L: **PCRTiler: automated design of tiled and specific PCR primer pairs**. *Nucleic Acids Research* 2010, **38**:W308-W312, [http://dx.doi.org/10.1093/nar/gkq485].
22. Rozen S, Skaletsky HJ: *Primer3 on the WWW for general users and for biologist programmers*, Humana Press, Totowa, NJ 2000 :365-386.
23. SantaLucia J: **A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics**. *Proc. Natl. Acad. Sci. USA* 1998, **95**:1460-1465, [http://dx.doi.org/10.1073/pnas.95.4.1460].
24. Teer J, Mullikin J: **Exome sequencing: the sweet spot before whole genomes**. *Hum Mol Genet.* 2010, **19**(R2):R145-R151.
25. Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, Mittnik A, Forrest S, Coombes BK, Wood JW, Earn DJD, White W, Krause J, Poinar HN: **Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death**. *Proc. Natl. Acad. Sci. USA* 2011, **108**(38), [http://dx.doi.org/10.1073/pnas.1105107108].
26. Kleinberg J, Tardos E: **Algorithm Design**. Addison Wesley, 2005.
27. Welsh DJA, Powell MB: **An upper bound for the chromatic number of a graph and its application to timetabling problems**. *The Computer Journal* 1967, **10**:85-86, [http://dx.doi.org/1093/comjnl/10.1.85].
28. Brélaz D: **New methods to color the vertices of a graph**. *Communications of the ACM* 1979, **22**(4), [http://dx.doi.org/10.1145/359094.359101].
29. Meyer, M and Kircher M: Illumina **Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing**. *Cold Spring Harb Protoc* 2010 **6**; [http://dx.doi:10.1101/pdb.prot5448]

2.10 Figures

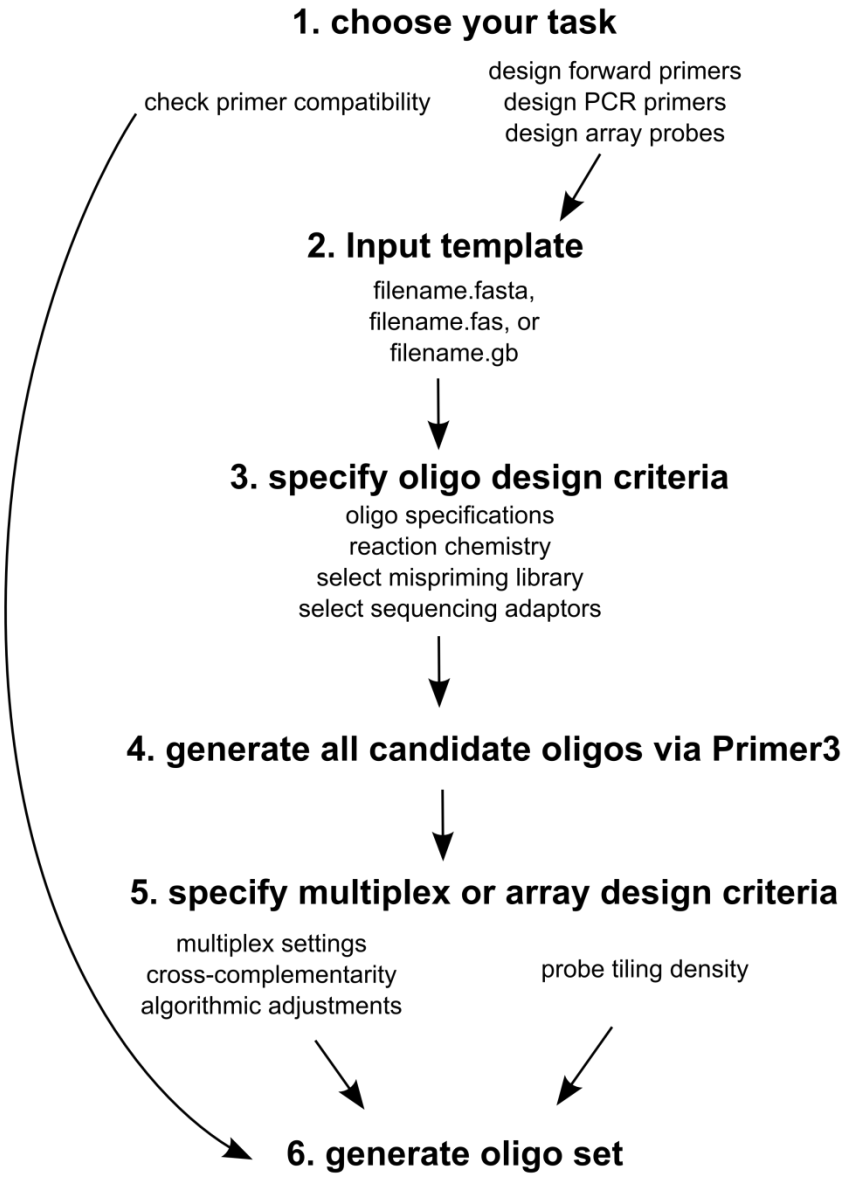


Figure 1 - Flow chart demonstrating the tasks performed by PrimerClique

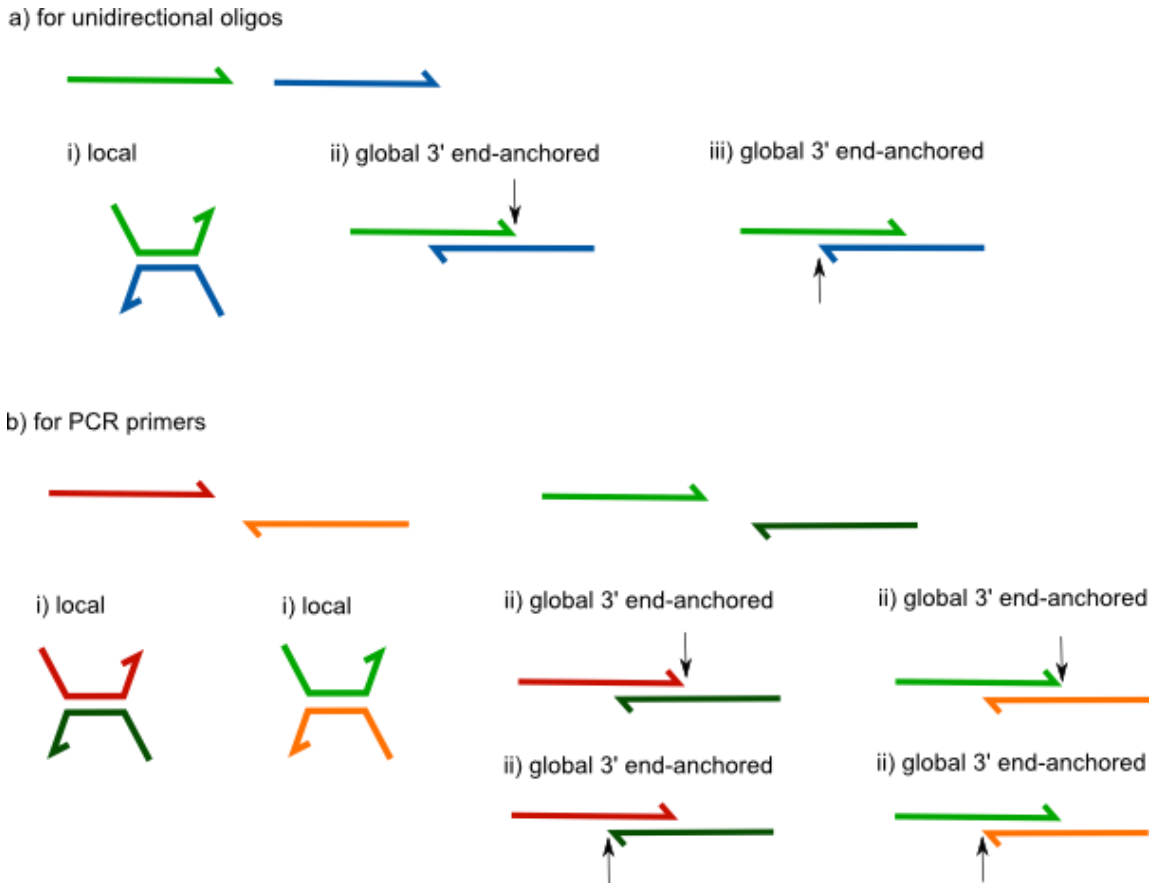


Figure 2 – Primer alignment checks performed for unidirectional and PCR primer cross compatibility.

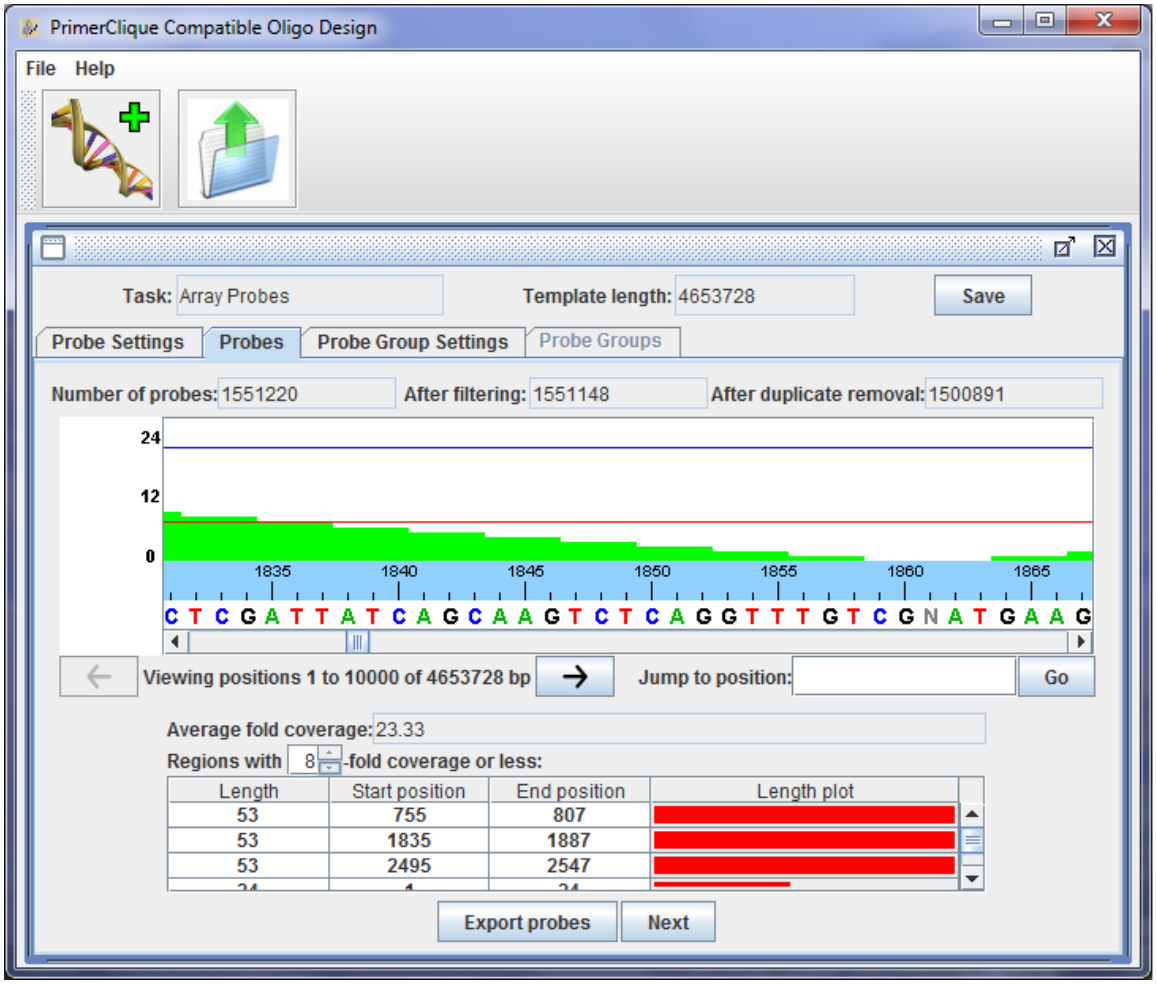


Figure 3 - Sample output display for array probe design. Blue line indicates average coverage, red line indicates coverage displayed at the selected threshold.

CHAPTER 3:

Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death

Bos KI[a]*, Schuenemann VJ[b]*, DeWitte SN[c], Schmedes S[d], Jamieson J, Mittnik A[a], Forrest SA[a], Coombes B[e], Wood JW[f,g], Earn D [h,e], White W[i,†], Krause J[b, j,‡] and Poinar HN[a,e,‡].

[a] McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

[b] Institut für Naturwissenschaftliche Archäologie, University of Tübingen, Rümelinstr 23, 72072 Tübingen, Germany

[c] Department of Anthropology, University at Albany, SUNY, 1400 Washington Avenue, Albany NY, 12222

[d] Institute of Investigative Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd Fort Worth Texas, 76107

[e] Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

[f] Department of Anthropology, Pennsylvania State University, University Park PA, 16802

[g] Population Research Institute, Pennsylvania State University, University Park, PA 16802-6211

[h] Department of Mathematics and Statistics, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

[i] Centre for Human Bioarcheology, Museum of London, 1500 London Wall, London, EC2Y 5HN, United Kingdom

[j] Human Genetics Department, Medical Faculty, University of Tübingen, Calwerstr. 7, 72076 Tübingen, Germany

[†] Deceased.

3.1 Author contributions

K.B., S.D., B.K.C., J.W.W., D.J.D.E., W.W., J.K., and H.N.P. designed research; V.J.S., K.B., S.D., S.S., J.J., A.M., S.F., J.W.W., and J.K. performed research; B.K.C. and W.W. contributed new reagents/analytic tools; V.J.S., K.B., J.K., and H.N.P. analyzed data; and K.B., J.K., and H.N.P. wrote the paper.

Data deposition: The sequence reported in this paper has been deposited in the Genbank database (accession nos. HE576978–HE576987).

[1]V.J.S. and K.B. contributed equally to this work.

[2]Deceased December, 2010.

[3]To whom correspondence may be addressed. E-mail: johannes.krause@uni-tuebingen.de or poinarh@mcmaster.ca.

3.2 Abstract

Although investigations of medieval plague victims have identified *Yersinia pestis* as the putative etiologic agent of the pandemic, methodological limitations have prevented large-scale genomic investigations to evaluate changes in the pathogen's virulence over time. We screened over 100 skeletal remains from Black Death victims of the East Smithfield mass burial site (1348–1350, London, England). Recent methods of DNA enrichment coupled with high-throughput DNA sequencing subsequently permitted reconstruction of ten full human mitochondrial genomes (16kb each) and the full pPCP1 (9.6 kb) virulence-associated plasmid at high coverage. Comparisons of molecular damage profiles between endogenous human and *Y. pestis* DNA confirmed its authenticity as an ancient pathogen, thus representing the longest contiguous genomic sequence for an ancient pathogen to date. Comparison of our reconstructed plasmid against modern *Y. pestis* shows identity with several isolates matching the Medievalis biovar; however, our chromosomal sequences indicate the victims were infected with a *Y. pestis* variant that has not been previously reported. Our data reveal that the Black Death in medieval Europe was caused by a variant of *Y. pestis* that may no longer exist, and genetic data carried on its pPCP1 plasmid were not responsible for the purported epidemiological differences between ancient and modern forms of *Y. pestis* infections.

ancient DNA | paleopathology

3.3 Introduction

The Black Death of 1347–1351 in Europe was one of the most cataclysmic events in history, and it is arguably "one of the most dramatic examples ever of emerging or reemerging disease" (ref. 1, p. 971). The disease is assumed to have been a particularly intense pandemic of bubonic and pneumonic plague caused by the Gram-negative bacillus *Yersinia pestis*. In this model, the medieval plague is considered the second of three pandemic waves, starting with the Plague of Justinian in A.D. 541 and culminating in the 20th century pandemic, which is still responsible for 2,000 cases/y worldwide (2) and is regarded as reemerging (3). Similarities in disease manifestation, mortality rates, and geographical distribution are generally cited as factors relating the three pandemics, though some scholars have argued that the second wave was too distinct in terms of its purported symptoms, epidemiology, and time of year of peak mortality to warrant such a connection. These discussants have argued in favor of other potential microbial associations with the medieval disease, including Bacillus anthracis (4), a filovirus (5), or a pathogen that has since become extinct (6).

Ancient DNA was sought to address the above controversy, although failed attempts to replicate initial work showing the presence of *Y. pestis* in purported victims of the medieval pandemic ignited skepticism regarding the identity of its etiologic agent (7, 8). Recent publications seem to have settled the controversy with the amplification of short segments of *Y. pestis* DNA by PCR-based approaches in several skeletal

collections from time periods associated with medieval plague outbreaks (9, 10), though none of these collections can be conclusively associated with the purported initial disease outbreak in medieval Europe of 1347–1351. A global survey of modern *Y. pestis* variants suggests that the ancient forms possess a unique and ancestral phylogenetic placement (9, 11). Although standard molecular methods can permit identification and limited phylogenetic resolution of ancient microbes, successful characterization of long contiguous stretches of authentic pathogen DNA will facilitate greater insight into the molecular architecture of ancient host–pathogen interactions. For *Y. pestis*, such an approach might be informative in addressing the noted differences between ancient and modern forms of the disease (4, 5, 12, 13).

PCR-based approaches are ill-suited for large-scale genetic investigations of ancient DNA owing to their preferential amplification of less damaged templates that derive from exogenous

contaminants and the highly fragmented nature of endogenous molecules (14, 15). In contrast, targeted enrichment strategies (16) in combination with high-throughput DNA sequencing allow for long stretches of ancient DNA to be reconstructed from a complex metagenomic background, and it is clearly the way forward for ancient pathogen research. The authenticity of endogenous ancient sequences can then be determined by looking for patterns of nucleotide damage typical of ancient DNA (15, 17). To show the suitability of these methods for analyses of

ancient pathogens, we have identified virulence-associated *Y. pestis* DNA fragments by PCR and independently replicated these results through subsequent targeted DNA enrichment (18) and high-throughput sequencing for several skeletal samples from victims securely dated to the initial medieval Black Death pandemic of 1348–1350 in London, England. The data presented

here represent both the oldest and longest assembled authentic sequences from an ancient pathogen, and in turn, they suggest that the Black Death was caused by a *Y. pestis* variant that harbours a pPCP1 plasmid found in some modern isolates.

3.4 Results

3.4.1 Y. pestis *DNA Preservation in Skeletal Remains*. Complete information regarding skeletal screening is available in SI Materials and Methods. In total, we screened DNA extracts from 109 samples — 53 bones and 46 teeth from the East Smithfield (ES) collection as well as 10 samples from St. Nicholas Shambles (SNS) that served as negative controls (Fig. 1 and Table S1). Total DNA content measured fluorometrically revealed less DNA in teeth than in bone, with dental extracts yielding on average 1,065 and 664 pg/μL for supernatant and pellet, respectively, compared with 1,956 and 2,280 pg/μL for bone. Quantitative PCR (qPCR) results showed the presence of amplifiable *Y. pestis pla* DNA in 5.7% of bones (3/53) and 37% of teeth (17/46) (Figs. S1 and S2). The caf1M assay yielded expected products in 5 of 17 teeth that contained

amplifiable pla, although this finding is expected based on *Y. pestis* plasmid numbers where the pPCP1 plasmid outnumbers the pMT1 by an estimated 100 to 1 (19). No *Y. pestis* DNA was

detected in any of the negative control samples. DNA sequences from the qPCR products contained both G to A and C to T transitions, damage patterns typical of ancient DNA (17). Despite their lower whole DNA content, dental samples were consistently a richer source for *Y. pestis* DNA than bone, and higher copy numbers were frequently observed in the EDTA supernatants as opposed to the pellets (Table S1). This finding is consistent with what one might expect from a blood-borne pathogen, where DNA likely resides in the desiccated blood vessels of the pulp chamber.

All amplifiable DNA was extremely low in quantity, with a maximum estimated copy number of 30 copies/µL.

3.4.2 ***Multiplex PCR Data for the*** **pla** ***Gene***. Using a multiplex PCR approach, we were successful in sequencing 78.3% of the *pla* gene and its flanking intergenic spacers at a minimum of four times coverage (Dataset S1). Missing regions were caused by a lack of expected PCR product. Cloned sequence data revealed 63 sites showing DNA damage reflected in the predominance of C to T (53.97%) and G to A (30.16%) transitions, which accounted for 84.13% of all aberrant positions. No transversions were observed. To distinguish damage from potential polymorphisms, all sequences were confirmed by cloned data from two independent amplified libraries. Wherever products from these

two libraries did not match, cloned sequence data from a third amplified library was sought. This process permitted us to resolve all transitions with the exception of three, where single nucleotide differences from published sequences were present in a subset of clones from two independently amplified libraries (Fig. S3). The PCR data did not contain the Microtus-specific C to T (20) at position 1,109 in Dataset S1, and it does not show the T to C transition previously reported in an ancient sequence for this region in the work by Raoult *et al*. (7) (Dataset S1, position 809).

3.4.3 Y. pestis *Chromosomal PCR Assays*. Of 11 primer sets used to define branches 1 and 2 phylogenetic placement (9), only 1 primer set yielded expected amplification products, namely s19: it produced the shortest amplicon of all chromosomal primer sets (80 bp), and it corresponds to a region in the DNA helicase II gene. Sequence data showed the absence of the CTA motif common to extant branch 2 sequences, thus supporting the notion that medieval

plague was not caused by an extant branch 2 variant (9); however, this region also revealed the presence of two synonymous point mutations that, to our knowledge, are not found in any *Y. pestis* sequences either modern or ancient (Fig. 2).

3.4.4 *Enrichment Efficiency*. To evaluate the suitability of targeted enrichment for investigations of virulence in ancient pathogens, we attempted to capture one of the two *pestis*-specific virulence plasmids, namely the 9.6-kb pPCP1 along with complete

human mitochondrial genomes from both Black Death and control teeth to evaluate endogenous DNA preservation. The total number of merged Illumina reads per sample that mapped to the target DNA (human mtDNA and pPCP1 DNA) varied between 0.3% and 49%, suggesting high enrichment efficiencies.

3.4.5 ***Human mtDNA Enrichment and Preservation***. Sequence clustering revealed between 228 and 84,244 unique fragments for the individual human teeth mapping to the reference mtDNA (Table 1), thus revealing high quantities of mtDNA fragments in all 10 human extracts with low quantities in two of the SNS human control extracts. The cave bear control sample yielded 15 mtDNA fragments, 6 of which aligned to the human mtDNA and 9 were identical or highly similar to the cave bear mtDNA genome (21). No human mtDNA fragments were found in the extraction blank. Complete or near complete mitochondrial genomes were constructed from fragments of 58 bp average length for all 10 human samples, with an average coverage of 0.7–280× (Table 1). All 10 mtDNA genomes were different from each other, and two samples (ES 3 and 4) extracted two times yielded the identical mtDNA genome sequence between duplicates. These data suggest that there was no cross-contamination between human samples. Nine of the ten were found to correspond to typical European mtDNA haplogroups based on phylotree (22) (Table 2). Only for SNS1 was the haplogroup not determinable due to the higher contamination level (Table 2). Comparisons of ancient mitochondrial consensus sequences against a worldwide dataset of 311 mtDNAs and

the Cambridge reference sequence revealed between 1 and 17 positions that were either unique to the ancient samples or present at less than 1% frequency in the current modern DNA database. Such positions were used to calculate the extent of polluting human exogenous DNA contamination within our fragments (Table 2 and Fig. S4). To calculate the ratio of endogenous to exogenous human DNA, we selected all fragments that showed the rare variant(s) and considered them to be endogenous, whereas all fragments that showed a common human mtDNA variant were considered to be potential contaminants. In so doing, we found between 21 and 6,724 individual mtDNA fragments overlapping such contamination informative positions in the ES samples. The majority of these DNA fragments were found to be consistent (all unique fragments overlapping the position showed the same substitution), suggesting that the majority of mtDNA fragments from each sample is derived from a single biological source (Fig. S4) (15, 23). Individual mtDNA fragments were compared against the corresponding consensus sequence, and nucleotide substitutions were recorded for each position along the DNA fragment. Based on these estimates, levels of polluting human DNA could be calculated, although these rates were very low for all samples (Table 2). In support of this notion, fragments identified as endogenous DNA contained degradation patterns typical for ancient templates (Fig. S5), again suggesting preservation of endogenous human DNA in all 10 medieval teeth.

3.4.6 Y. pestis *DNA Enrichment and Preservation*. We found between 42 and 36,986 unique fragments mapping to a portion of the *Y. pestis* pPCP1 plasmid reference genome in all libraries, including the pre-Black Death human control, the cave bear sample, and the extraction blank (Table 1). All reads in the non-ES extracts mapped to a region between positions 3,000 and 4,200 of the reference pPCP1 that shows a high similarity to expression vectors that are used for recombinant enzyme production. This region was found to be problematic in a previous *Y. pestis* pPCP1 study using shotgun sequencing (24); hence, it is likely that remnants of the expression vectors are being captured by our enrichment approach. Apart from these conserved motifs found across all samples and controls, the pre-Black Death human control samples (SNS) showed not a single unique fragment homologous to the pPCP1 sequence, whereas the five ES samples contained between 7 and 6,435 unique fragments that mapped to the *Y. pestis* pPCP1 plasmid (Table 1). Because all five individuals from the ES cemetery likely died from the same strain of *Y. pestis*, unique pPCP1 fragments were pooled to reconstruct a consensus sequence at maximum coverage. The pooled consensus has an average coverage of 43 bp, ranging from 0× to 702×, and an average fragment length of 36 bp (Table 1). This finding may be why a previous study using the ES collection that targeted regions in excess of 130 bp (8) was unsuccessful in replicating some original work with plague samples (7, 25). Not surprisingly, the three regions with the poorest coverage had a low GC content (%) (Fig. 3 and Fig. S6). Shotgun sequence data for ancient human samples (23) show similar patterns, suggesting a preservation bias to high GCcontent

and thus likely ruling out enrichment artifacts. Despite lower coverage of AT-rich regions, the consensus sequence covers ~99% of the entire pPCP1 at a minimum of 2× (Table 3).

We tested the authenticity of the pPCP1 DNA by analyzing the substitution patterns along the DNA fragments and found the same nucleotide misincorporation pattern that was observed within the mtDNA fragments from the five medieval samples of the ES site. The predominant C to T substitutions found at the 5' ends and G to A substitutions at the 3' ends are typical of ancient DNA (15, 26, 27). This pattern is not seen in the fragments mapping to the vector region of the pPCP1 plasmid, lending additional support to the notion that the majority of these fragments are indeed derived from contaminating expression vectors of the supplied reagents (Fig. 4 and Fig. S7). Fragment length distribution also indicates longer fragments obtained for this region, which is consistent with their coming from a modern contaminant source (Fig. 5).

Phylogenetic comparison of the 8,299-bp 2× consensus pPCP1 plasmid with modern *Y. pestis* reveals sequence identity with 11 of 14 extant *Y. pestis* strains; it did not match sequences typed as Microtus (AE017046.1), Angola (CP000900.1), and Medievalis (CP001611.1), although it does match the KIM pPCP1 (AF053945.1). It also resolves the three ambiguous positions identified by the multiplex approach of the *pla* gene and replicates the *pla* sequence described above. Considering indel

polymorphisms, the ES strain does not contain the TT insertion common to the Orientalis biovar (AL109969.1 and CA88-4125).

3.5 Discussion

The last several decades have introduced human populations to a historically unprecedented number of emerging or reemerging infectious diseases, mostly facilitated by anthropogenic factors such as globalization of trade and human travel, changes in local ecology, and antibiotic resistance (28). Genetic diversity in pathogens is known to be a major source of phenotypic diversity underlying disease dynamics (29), hence microbial changes cannot be ignored as a potential driving force influencing host–pathogen interactions (30, 31). Genetic investigations of ancient microbes may provide much needed data to elucidate how the virulence of our close microbial companions has evolved over time. Potential factors that influenced the coevolution of humans and their pathogens in the past can be of great value to elucidate the dynamics of host–pathogen relationships in our new era of emerging infections (32, 33). Here we report the presence of short DNA fragments (<60 bp) from two *pestis*-specific plasmids as well as one chromosomal sequence showing synonymous substitutions from individuals from a well-documented Black Death mass burial ground from 1348 to 1350. These results were obtained in a facility that has not been previously exposed to modern sources of *Y. pestis*, and the sequence data contain a damage pattern that is characteristic of ancient DNA (26). Using a targeted DNA capture approach combined

with high-throughput sequencing, we obtained sequences of the pPCP1 plasmid of *Y. pestis* from all five human samples from the ES site that, when used together, reconstruct 99% of the plasmid, excluding a 1,200-bp region because of

modern expression vector contamination. The pPCP1 fragments from the ES cemetery show DNA damage patterns common to ancient DNA (15, 26, 27). Analogous results were obtained for complete human mtDNA sequences from these five victims as well as the control samples, suggesting little postexcavation human DNA contamination and adequate preservation of endogenous DNA. Given the independent replication of the *pla* gene by PCR and enrichment-based strategies in two independent laboratories, the intersample replication of portions of the pPCP1 from five humans each with different mtDNA genomes, and the unambiguous DNA damage patterns, we believe that these data conclusively show the presence of *Y. pestis* DNA in

medieval dental tissues from victims of the first wave of the Black Death, lending additional support to implicate this bacterium as a causative agent in the medieval plague pandemic of 1347–1351. The absence of a single *Y. pestis* DNA fragment in any of the pre- Black Death control teeth (SNS), which consisted of two samples with mtDNA yields comparable with the three best-preserved ES samples, also bolsters this claim and circumvents the concern over soil-dwelling microbial contaminants mimicking pPCP1 sequences in the ES samples.

Two of the authors (SD and JM) have previously argued that the epidemiology, virulence, and population dynamics of the Black Death were too different from those

factors of modern yersinial plague to have been caused by *Y. pestis* (13). Given the growing body of evidence implicating this bacterium as responsible for the pandemic, we believe scientific debates should now shift to addressing the genetic basis of the epidemic's unique characteristics.

Y. pestis is a recently evolved variant of the soil-dwelling bacterium *Y. pseudotuberculosis*, differing mostly in the acquisition of two virulence plasmids, namely the pMT1 (100 kb) and the high copy pPCP1 (9.6 kb), which contains the pesticin (pst) and pesticin immunity genes (pim) involved in bactericidal activity and the virulence-associated plasminogen activator gene, pla. This important gene is associated with enhancing the bubonic form of the disease by facilitating bacterial dissemination at the site of a flea bite and encouraging bacterial proliferation in the respiratory system during pneumonic episodes. The full pPCP1 plasmid is mostly conserved between different modern *Y. pestis* variants; hence, it is not surprising that the plasmid is effectively unchanged in an ancient sample. In all, we feel confident that the pPCP1 plasmid presented here did not contribute to the purported differences between ancient and modern forms of the disease. The ES pPCP1 sequence does not match the Orientalis strain previously defined as branch 1 (11). Although an in-depth discussion of organismal phylogeny falls outside the current focus, the absence of a branch 2-specific marker in the DNA helicase II gene and the presence of two substitutions in the same region that, to our knowledge, are not found in any previously reported modern or ancient sequence suggest that the medieval form of *Y. pestis* may harbor additional

information regarding the organism's evolutionary history as a human pathogen. This information may prove instrumental in identifying factors that influence the different epidemiology of ancient and modern forms of the disease.

3.6 Materials and Methods

3.6.1 ***Experiments Conducted at McMaster University***. Information on the archaeological site, skeletal sampling, and DNA extraction can be found in SI Materials and Methods. We note that the site from which the skeletons were recovered, the ES burial ground in London, is one of just a few excavated sites in the world that can be linked definitively and uniquely to the mid-14th century outbreak of the Black Death by archaeological and documentary evidence (34).

Screening. Total DNA from each extract was measured on a Turner Biosystems DNA fluorometer. PCR inhibition for each DNA extract was determined by an internal positive control (35) using a 63-bp mammoth cytochrome b assay to evaluate the amount of Taq Gold DNA polymerase required to overcome PCR inhibition in the extract. Initial screening of all extracts was done using a 52-bp qPCR assay specific to the plasminogen activator (pla) gene located on the high-copy pPCP1 plasmid of *Y. pestis* (19). All extracts yielding expected pla

products were screened for the presence of an additional *Y. pestis*-specific marker, the chaperone protein for the fraction 1 antigen gene (caf1M) located on the pMT1 plasmid. Sensitivity for both *Y. pestis* assays was determined using synthetically

produced sequences, which contained identical primerbinding sites flanking a *Y. pestis* sequence that differed by a single nucleotide transversion (Figs. S1 and S2). Quantitative assays were performed on an

MX3000 PCR cycler (Stratagene) or a CFX96 Real-Time PCR Detection System (BioRad), and all were sensitive to single copy template molecules. Details on PCR conditions and sensitivity are available in SI Materials and Methods.

3.6.1.1 ***454 library preparation***. Library preparation was performed using the Roche manufacturer's protocol, eliminating the fractionation step (454; Roche). Libraries were quantitated in 20 µL qPCR reactions using methods described elsewhere (36), and amplified products were used as templates for downstream multiplex applications.

3.6.1.2 ***Standard and multiplex PCR reactions***. Forward and reverse primers were designed using Integrated DNA Technologies software (www.idtdna.com) to span the full *pla* gene and its upstream and downstream intergenic spacer regions (37) (Table S2). Limitations in primer design made it necessary to exclude a 123-bp region flanking the transcription start site and an 86-bp region flanking the start of the downstream intergenic region, leaving a 1,191-bp target region. Multiplex reactions were carried out using methods described elsewhere (38). Primer pairs were partitioned into 22 pools through the use of a dynamic programming algorithm for interval scheduling (39), with compatibility assessed by alignment score and melting temperature (Tm). DNA

sequences were confirmed in at least two independent amplified library fractions. For phylogenetic placement, primers for a subset of the branch 1 and 2 diagnostic chromosomal SNPs (40) were used in single PCR reactions using primers and PCR conditions described in ref. 9. Primer sequences, PCR conditions, and the consensus sequence are described in SI Materials and Methods.

3.6.2 ***Experiments Conducted at the Max Planck Institute in Leipzig, Germany***. DNA extraction and enrichment. To independently replicate the PCR-amplified *Y. pestis* DNA from the ES collection and test the suitability of enrichment strategies and subsequent targeted high-throughput sequencing for ancient pathogens, several teeth were analyzed at the Max Planck Institute in Leipzig, Germany. Five dental roots from mature molars from the ES collection were chosen based on *pla* copy numbers and analyzed alongside five dental roots from St. Nicholas Shambles (Table S1), and an ancient cave bear sample was used as a cross-contamination control. An aliquot from each DNA extract was used to produce libraries using a modified Illumina multiplex protocol (41). Endogenous mitochondrial DNA was captured in parallel from all libraries. To enrich for target DNA, we used long-range PCR products as bait for molecular capture through hybridization (18). Additional details on extraction, library preparation, and enrichment are available in SI Materials and Methods.

3.6.2.1 ***Solexa sequencing and analysis (Leipzig)***. The enriched library pool was sequenced on an Illumina Genome Analyzer IIx using the manufacturer's protocols for multiplex sequencing with modifications. The raw reads were aligned to the PhiX 174 reference sequence to obtain a training dataset for the base caller Ibis (42). Raw sequences called by Ibis 1.1.1 were filtered for the individual indexes as previously described (43). The paired-end reads were subjected to a fusion process (including removal of adapter sequences and adaptor dimers) by requiring

at least an 11-nt overlap between the two reads. In the overlapping sequence, quality scores were combined, and the base with the highest base quality score was called. Only sequences merged in this way were used for additional analysis. The sequencing data were analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. Enrichment efficiency was calculated by mapping all 37,502,405 merged reads to the revised Cambridge Reference Sequence (rCRS) for human mtDNA and the *Y. pestis* CO92 pPCP1 plasmid using a custom iterative mapping assembler (23). From this information, the fraction of total reads that mapped to the target DNA was calculated.

3.6.2.2 ***Phylogenetic analysis***. The consensus sequences for the mitochondrial genome fragments were analyzed using a custom Perl script to identify individual haplogroups based on phylotree (22). All pPCP1 fragments from the ES samples were combined into

a single consensus sequence, and they were aligned and compared by eye with 14 modern pPCP1 sequences using the software Muscle (44) (SI Materials and Methods).

3.7 ACKNOWLEDGMENTS.

We thank Tomislav Maricic and Martin Kircher for technical expertise and analysis assistance. We thank Svante Pääbo for access to the clean room and laboratory facilities of the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. We also thank Rebecca Redfern and Jelena Bekvalac for providing access to skeletal samples and facilities for sampling at the Museum of London, Elizabeth Carniel for providing modern *Yersinia pestis* DNA, the entire team at the McMaster Ancient DNA Centre for helpful comments on the design and implementation of laboratory work, and D. Ann Herring and Debi Poinar for helpful comments on earlier versions of the manuscript. Financial support was provided by the Social Sciences and Humanities Research Council of Canada, the Canadian Institute for Health Research, Canada Research Chairs program, the Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, the University at Albany Center for Social and Demographic Analysis, and the Human Genetics Department, Medical Faculty, University of Tübingen, Germany.

3.8 References Cited

1. Wheelis M (2002) Biological warfare at the 1346 siege of Caffa. *Emerg Infect Dis* 8(9):971-975.
2. www.who.int/vaccine_research/diseases/zoonotic/en/index3.html.
3. Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E, et al. (2008) Plague: Past, Present, and Future. *PLoS Med* 5(1): e3. doi:10.1371/journal.pmed.0050003.
4. Twigg G (1984) The Black Death: a biological reappraisal. London: Batsford Academic.
5. Scott S, Duncan CJ (2001) The Biology of Plagues. Cambridge: Cambridge University Press.
6. Cohn SK (2003) The Black Death transformed: disease and culture in early Renaissance Europe. London: Arnoldohn.
7. Raoult D, Aboudharam G, Crubezy E, Larrouy G, Ludes B, Drancourt M (2000) Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of medieval black death. *Proc Natl Acad Sci U S A* 97(23):12800-12803.
8. Gilbert MT, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, Prentice MB (2004) Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology* 150(Pt 2):341-354.
9. Haensch S, Bianucci R, Signoli M, Rajerison M, SchultzM, Kacki S, Vermunt M, Weston DA, Hurst D, Achtman M, et al (2010) Distinct Clones of *Yersinia pestis* caused the Black Death. *PLoS Pathog* 6(10): e1001134. doi:10.1371/journal.ppat.1001134.
10. Weichmann I, Harbeck M, Grupe G (2010) *Yersinia pestis* DNA sequences in Late Medieval Skeletal Finds, Bavaria. *Emerg Infect Dis* 16(11):1806 – 1807.
11. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, et al (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* 42(12):1140 – 1143.
12. Wood, J.W**,** R.J. Ferrell, and S.N. DeWitte-Aviña (2003) The temporal dynamics of the fourteenth-century Black Death: New evidence from ecclesiastical records. *Human Biology* 75:427-448.
13. Wood JW, and DeWitte-Aviña SN. Was the Black Death yersinial plague? *Lancet Infect Dis* 2003; 3:327-328.
14. Pääbo S, Poinar H, Serre D, Jaenicke-Deprés V, Hebler J (2004) Genetic Analyses from ancient DNA. *Annu Rev Genet* 38:645-79.
15. Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianki AP, Pääbo S (2010) A complete mtDNA genome of an Early Modern Human from Kostenki, Siberia. *Curr Biol* 20:231-236.

16. Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PLF, Xuan Z, et al (2010) Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* 328:723 – 725.
17. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, and Paabo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nuc Acids Res* 2001 29(23):4793-4799.
18. Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 5(11): e14004. doi:10.1371/journal.pone.0014004.
19. Parkhill J, Wren BW, Thomson NR, Titball NR, Holden MTG, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, et al (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* 413 Oct; 523 – 527.
20. Song Y, Tong Z, Wang J, Wang L, Guo Z, Han Y, Zhang J, Pei D, Zhou D, Qin H, et al (2004) Complete Genome Sequence of *Yersinia pestis* Strain 91001, an Isolate Avirulent to Humans. *DNA Research* 11:179-197.
21. Krause J, Unger T, Noçon A, Malaspinas A, Kolokotronis S (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evol Biol* 8:220.
22. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondria variation. *Hum Mutat* 30(2): E386-E394.
23. Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al (2008) A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell* 134(3):416-426.
24. Plunkett G, Anderson BD, Burland V, Cabot EL, Glasner JD, et al (2007) *Yersinia pestis* CA88-4125 whole genome shotgun sequencing project, direct submission to NCBI, accession number NZ_ABCD01000008.1.
25. Drancourt M, Aboudharam G, Signoli M, Dutour O, and Raoult D (1998) Detection of 400 year-old *Yersinia pestis* DNA in human dental pulp: An approach to the diagnosis of ancient septicaemia. *Proc Natl Acad Sci U S A* 95:12637-12640.
26. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, and Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A* 104(17):14616-114621.
27. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, and Alan Cooper (2007) Novel high-resolution characterization of ancient DNA reveals C>U-type base modification events as the sole cause of post mortem miscoding lesions. *Nuc Acids Res* 35(17):5717-5728.

28. Barrett R, Kuzawa CW, McDade T, and Armelagos GJ (1998) Emerging and Re-emerging Infectious Diseases: The Third Epidemiologic Transition. *Annu. Rev. Anthropol* 27:247–271.
29. Lawrence JG (2005) Common themes in the genome strategies of pathogens. *Curr Opin Genet Dev* 15(6):584-588.
30. Morse SS (1995) Factors in the Emergence of Infectious Diseases. *Emerg Infect Dis* 1(1):7-15
31. Pybus and Rambaut (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 10:540-550.
32. Baum J and Khalina Bar-Gal G (2003) in Greenblat C and Spigleman M eds. *Emerging Pathogens Archaeology, Ecology, & Evolution of Infectious Diseases* (Oxford University Press, Great Britain) 67-78.
33. Gandon Buckling, E. Decaestecker, and T. Day (2008) Host-parasite coevolution and patterns of adaptation across time and space. *Journal of Evol Biol* 21:1861-1866.
34. Grainger I, and Hawkins D. 1988. Excavations at the Royal Mint site 1986-1988. The London Archaeologist 5:429-436.
35. King CE, Debruyne R, Kuch M, Schwarz C, Poinar HN (2009) A quantitative approach to detect and overcome PCR inhibition in ancient DNA extracts. *BioTechniques* 47(5):941-949.
36. Meyer M, Briggs AW, Maricic T, Höbner B, Höffner B, Krause J, Weihmann A, Pääbo S, Hofreiter M (2007) From micrograms to picograms: quantitative PCR reduces the material demands of high-througput sequencing. *Nucleic Acids Res* 36(1):e5.
37. Kim T, S Chauhan, Motin VL, Goh E, Igo MM, Young GM (2007) Direct Transcriptional Control of the Plasminogen Activator Gene of *Yersinia pestis* by the Cyclic AMP Receptor Protein. *J Bacteriol* 189(24):8890 – 8900.
38. Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, Barnes I, Lister A, Ebersberger I, Pääbo S, HofreiterM (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* 439:724-727.
39. Bos KI, Forrest SA, Poinar HN (2009) Metagenomics and ancient human disease: efficient targeted retrieval of ancient pathogen DNA for high-throughput sequencing applications. Paper presented at the 37th annual meeting for the Canadian Association for Physical Anthropology, Vancouver BC.
40. Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, et al (2004) Microevolution and History of the Plague Bacillus, *Yersinia pestis*. *Proc Natl Acad Sci U S A* 101: 17837-17842.

41. Meyer M and Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc* Issue 6; doi:10.1101/pdb.prot5448.
42. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* 10: R83.
43. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* 35: e97.
44. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* 32 (5): 1792-1797.
45. Rakin A, Boolgakowa E, and Heesemann J (1996) Structural and functional organization of the *Yersinia pestis* bacteriocin pesticin gene cluster. *Microbiology* 142:3415-3424.

3.9 Figures

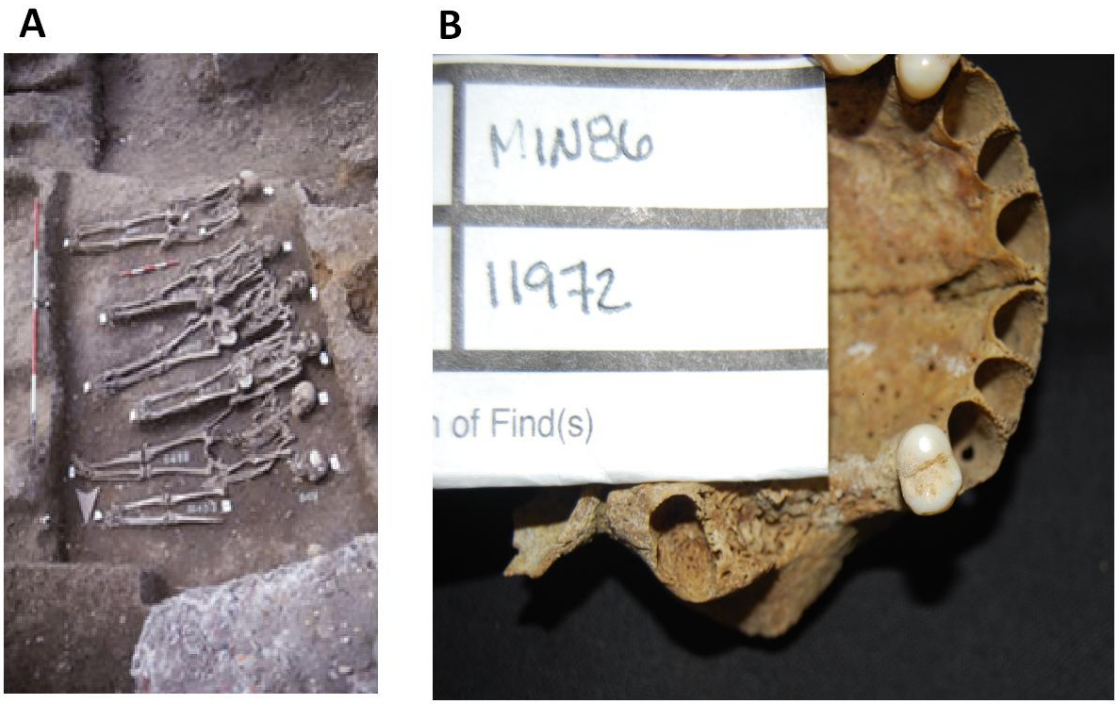


Figure 1 – The excavation at East Smithfield, 1986 – 1988 (A) and sampling at the Museum of London (B). Photos reproduced courtesy of the Museum of London.

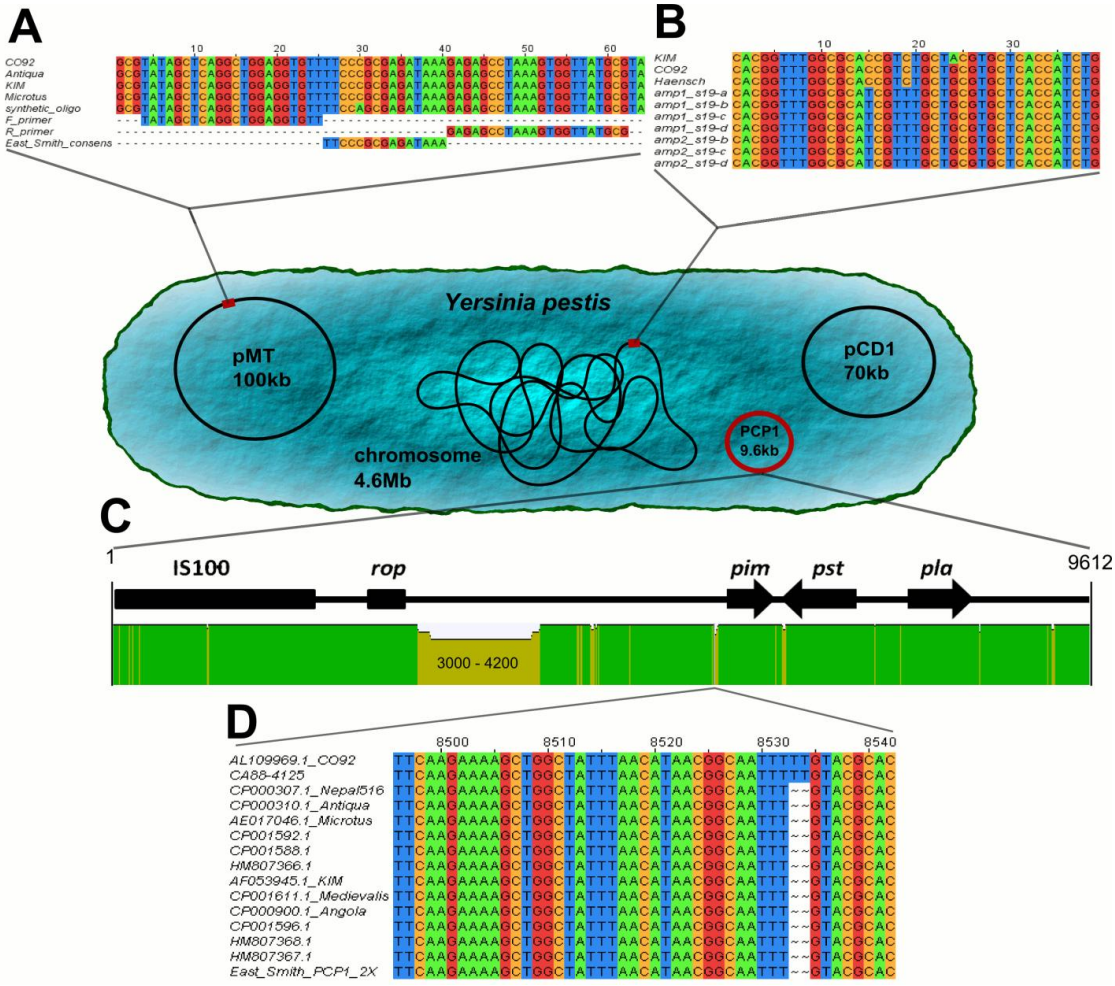


Figure 2 – Positional relationship of the sequences reported for medieval *Y. pestis*. A) Quantitative PCR assay for the *caf1M* locus of the pMT1 plasmid. **B**) Chromosomal sequence data for s19 (9) showing positions of two novel synonymous substitutions in the DNA helicase II gene, and absence of the CTA SNP required for branch 2 designation (boxed nucleotides). **C**) Solexa consensus (minimum of 2-fold coverage) compared against 14 modern *Y. pestis* variants shown with positional information for genes on the PCP1 plasmid (*rop* = plasmid replication regulatory protein; *pim* = pesticin immunity protein; *pst* = pesticin; *pla* = plasminogen activator gene). Transcriptional polarity (45) indicated by arrows. Identity shown in green; region between 3000 and 4200 is not considered (see main text). **D**) Alignment of the medieval *Y. pestis* pPCP1 consensus against 14 modern variants showing presence of an indel common to branch 2 and ancestral isolates.

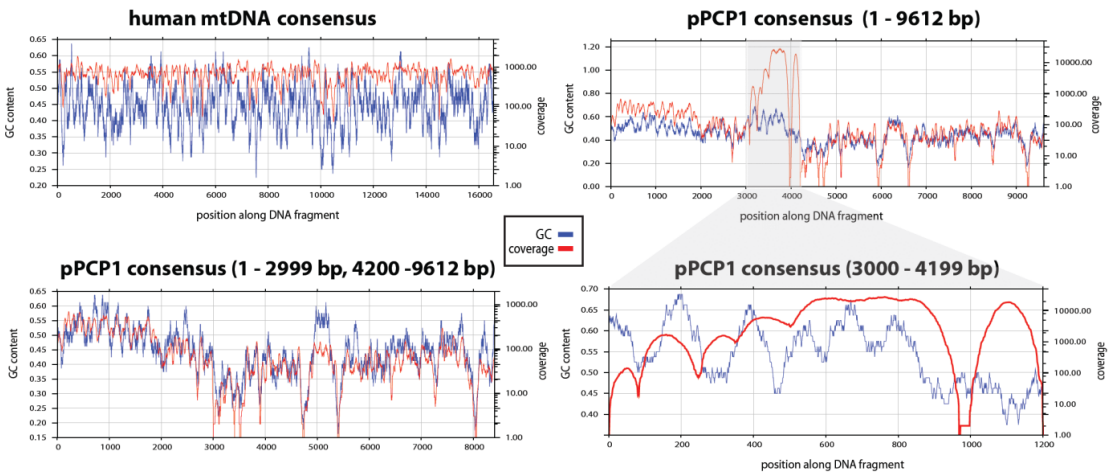


Figure 3 – GC content and DNA coverage for the consensus sequences of mtDNA, *Y.pestis* pPCP1 (1 – 9612 bp), pPCP1 (4200 – 9612 bp, 1 – 3000 bp) and the excluded region pPCP1 (3000 – 4200 bp).

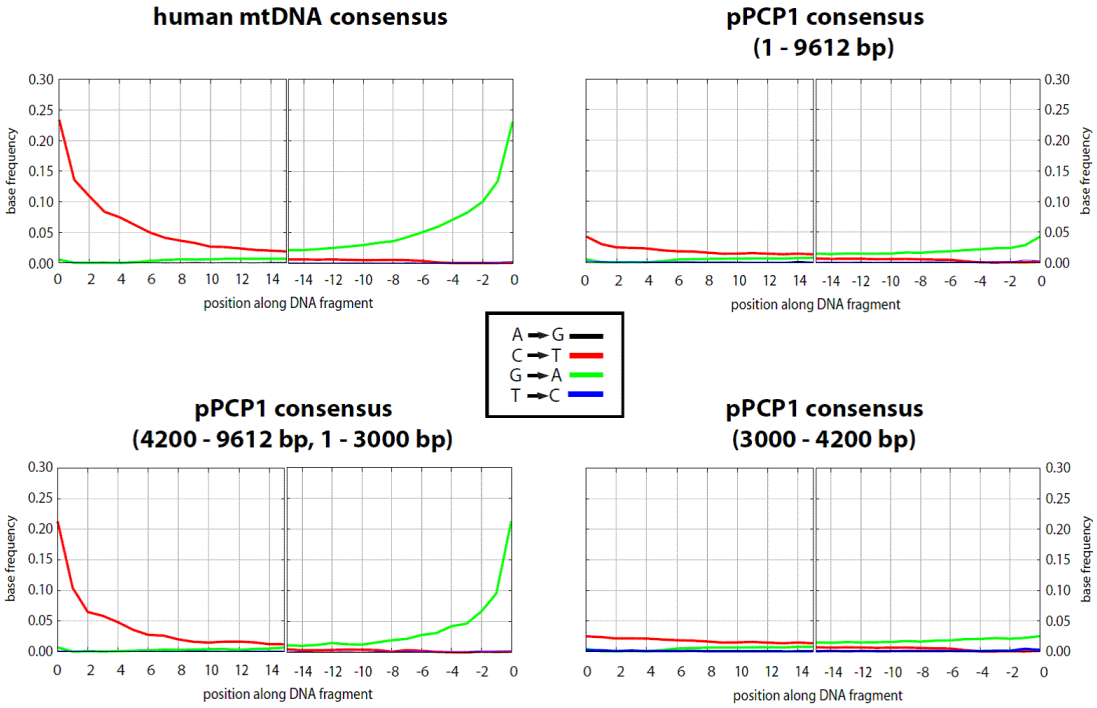


Figure 4 – Nucleotide misincorporation patterns for both the East Smithfield mtDNA consensus sequences and the complete pPCP1 (1 – 9612 bp) including the section showing alignment to contaminating modern vector DNA (3000 – 4200 bp), and the entire pPCP1 excluding the mismapping region (1-3000bp and 4200 – 9612 bp).

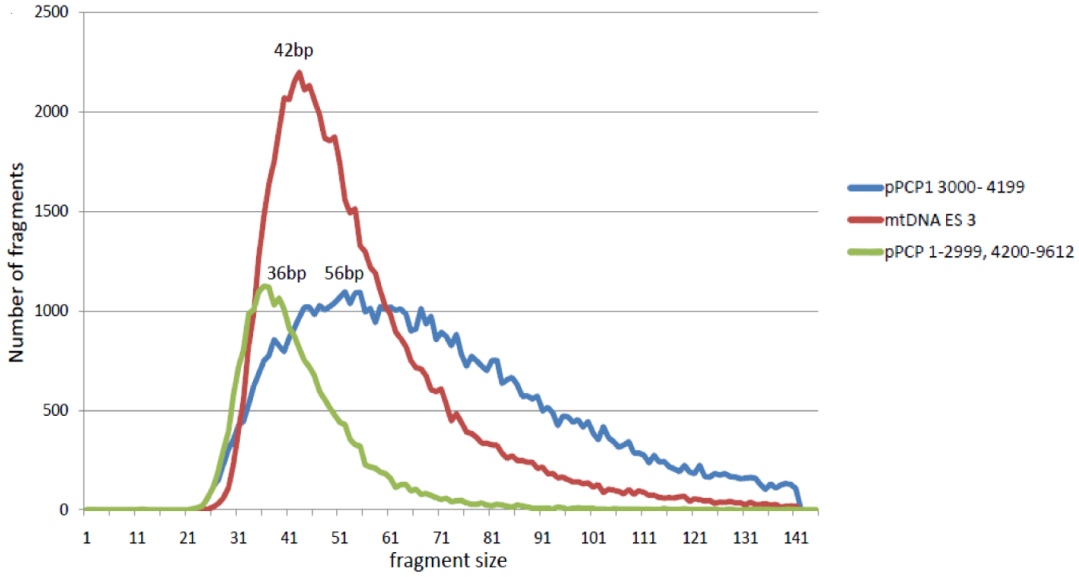


Figure 5 – Graphical depiction of length distribution vs number of fragments for different target regions.

3.10 Tables

Sample		mtDNA				pPCP1_1-2999, 4200-9612			pPCP1_3000-4199		
		fragments (unique)	average coverage	average length	MtDNA Haplogroup	fragments (unique)	Average Coverage	average length	fragments (unique)	average coverage	average length
SNS1	SN 5169	240	0.8	55.6	H2a2	0	0.0	na	42	1.9	54.4
SNS2	SN 5200	31,973	95.8	49.7	K1a1b1	0	0.0	na	16,121	886.1	66.0
SNS3	SN 5194	228	0.7	51.5	H2a	0	0.0	na	110	4.3	46.8
SNS4	SN 5122	77,234	313.1	67.2	T2b	0	0.0	na	7,172	382.4	64.0
SNS5	SN 5053	2,221	9.0	67.2	H	0	0.0	na	8,703	522.0	71.5
ES1	ES6443	19,995	69.7	57.7	HV0b	7	0.0	39.0	17,852	971.0	64.3
ES2	ES8291	6,666	26.8	66.5	H1	4,531	22.8	42.3	15,215	887.1	71.6
ES3.1	ES6330	84,244	280.2	55.1	T2g	6,435	33.1	43.3	36,986	2216.4	74.0
ES3.2	ES6330	47,706	177.6	61.7	T2g	3,559	20.4	48.1	12,829	748.8	72.5
ES4.1	ES11972	5,436	19.5	59.5	H6a1a	2,500	12.0	40.3	13,196	727.6	66.4
ES4.2	ES11972	29,746	102.1	56.8	H6a1a	4,515	22.8	42.5	28,283	1592.2	66.9
ES5	ES8124	7,621	24.0	52.2	U4c1a	1,530	6.9	37.8	28,784	1656.5	68.6
Totals ES 1-5		**202,406**	**703.2**	**57.6**	**na**	**23,112**	**118.2**	**43**	**150,789**	**8735.7**	**69.6**
cave bear		15*	0.1	56.9	na	0	0.0	na	5,397	279.2	62.1
ExtrBlk		0	0	na	na	0	0.0	na	13,463	757.6	67.4
libraryBlk		11	0.03	47.2	na	0	0.0	na	886	38.0	51.3

Table 1 – Average coverage, fragment size and number of fragments mapping to mtDNA, pPCP1 (4200 - 9612 bp, 1 - 3000 bp) and pPCP1 (3000 bp - 4200 bp) for all 6 medieval teeth as well as two controls and the pooled fragments from the East Smithfield site ES1-5. na- not applicable. * sequences are conserved between cave bear mtDNA (9 sequences) and human mtDNA (6 sequences).

			contamination analyis via 311 mtDNA			contamination analysis via rCRS		
Sample		**mtDNA group**	clean fragments	polluting fragments	% contam (95%C.I.)	clean fragments	polluting fragments	% contam (95%C.I.)
SNS 1	SN 5169	na	1	2	**66.7%** (20.8-93.9%)	4	1	**20.0%** (3.6- 62.4%)
SNS 2	SN 5200	K1a1b1	95	0	**0.0%** (0-3.9%)	2362	2	**0.1%** (0.0- 0.3%)
SNS 3	SN 5194	H2a	1	0	**0.0%** (0-79.3%)	15	0	**0.0%** (0.0- 20.4%)
SNS 4	SN 5122	T2b	218	7	**3.1%** (1.5-6.3%)	6365	7	**0.1%** (0.1- 0.2%)
SNS 5	SN 5053	H	3	0	**0.0%** (0-56.2%)	77	0	**0.0%** (0.0- 4.8%)
ES 1	ES 6443	HV0b	69	0	**0.0%** (0.0-5.3%)	696	0	**0.0%** (0.0-0.5%)
ES 2	ES 8291	H1	21	3	**12.5%** (4.3-31.0%)	239	3	**1.2%** (0.4-3.6%)
ES 3.1	ES 6330	T2g	1388	2	**0.1%** (0.0-0.5%)	6724	11	**0.2%** (0.1-0.3%)
ES 3.2	ES 6330	T2g	911	0	**0.0%** (0.0-0.4%)	4067	5	**0.1%** (0.1-0.3%)
ES 4.1	ES 11972	H6a1a	79	1	**1.2%** (0.2-6.7%)	262	1	**0.4%** (0.1-2.1%)
ES 4.2	ES 11972	H6a1a	454	3	**0.7%** (0.2-1.9%)	1315	3	**0.2%** (0.1-0.7%)
ES 5	ES 8124	U4c1a	161	4	**2.4%** (0.9-6.1%)	543	4	**0.7%** (0.-1.9%)
cave bear		na	15	0	**0.0%** (0.0-27.8%)	6	0	**0.0%** (0.0-27.8%)
ExtrBlk		na	0	0	na	0	0	na
LibraryBlk		na	0	0	na	1	0	**0.0%** (0.0-79.3%)

Table 2 – Mitochondrial haplogroup identification based on phylotree (12), as well as mtDNA contamination estimates based on positions where 311 worldwide mtDNAs or the rCRS differ from the consensus mtDNA sequence of a particular sample. Clean fragments were found to be identical to the sample consensus sequence at informative positions. Polluting fragments were found to be in the same state like the 311mtDNAs or the rCRS. The na refers to not analysed.

coverage	positions (total)	positions (covered)	coverage %
1- fold	8412	8371	99.52%
2- fold	8412	8299	98.66%
3- fold	8412	8197	97.45%
4- fold	8412	8109	96.40%
5- fold	8412	8022	95.37%

Table 3 – Coverage for the pPCP1 East Smithfield (1 - 3000 bp, 4200 - 9612 bp) consensus sequence.

3.11 Supporting online material for Shuenemann et al. 2011, **Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death**

3.11.1 METHODS

3.11.1.1 *The Site*. The East Smithfield cemetery is one of two emergency mass burial grounds established in the city of London in late 1348 or early 1349 in response to the mass casualties resulting from the medieval plague which entered London in the autumn of 1348 (1). The burial ground is located northeast of the Tower of London, and lay just on the outskirts of the medieval city. Historical documents clearly confirm the purchase of this land for the express purpose of disposing of plague victims, which had overwhelmed existing parish cemeteries (2). In addition, archeological evidence (coin deposits) confirms that the cemetery dates to the 1340s (2). The site consists of an estimated 2400 burials, and excavations carried out between 1986 and 1988 recovered the remains of 634 individuals, most of which consisted of nearly complete skeletons (Figure 1). This collection is among the largest excavated catastrophe burials in Europe, and the historical evidence linking it to the initial outbreak of the 14th century Black Death in Europe makes it an ideal collection from which to draw tissue samples.

3.11.1.2 *Skeletal sampling*. Skeletal material for the collection is curated at the Museum of London. Sample collection took place at the museum by individuals wearing

protective clothing (latex gloves, tyvek suit, N95 mask). Sections of skeletal elements were removed using a Dremel tool on the lowest setting to reduce heat, which can denature DNA. The Dremel tool was cleaned with bleach (10% sodium hypochlorite) between samples. Care was taken to avoid areas of muscle and ligament attachment, or areas showing pathological changes. Mature in-situ teeth with fused roots and intact enamel surfaces absent of pathological lesions (caries, calculus, excessive enamel ware) were carefully harvested by hand from skulls (Figure 1). In total, 53 skeletal samples and 46 teeth were analysed from individuals of the East Smithfield collection. In addition, 10 teeth were taken from individuals of the St. Nicholas Shambles pre-Black Death attritional burial site of London, England dating to the 11th and 12th centuries (3) for use as a negative control population.

3.11.1.3 DNA extraction and amplification, McMaster University

Manipulations of skeletal and dental tissues took place at the McMaster Ancient DNA Centre in a facility specifically dedicated to the sampling and extraction of ancient human tissues. Prior to extraction, bone samples were pulverised by blunt force with a hammer. To expose the pulp chamber, dental samples were first sectioned at the cementoenamel junction using a Dremel cutting wheel. Crowns were placed in a -40°C freezer, and material from the innermost layer of the pulp chamber was carefully removed using a Dremel tool on its lowest setting fitted with a cold (-40°C) grout cutter

attachment. DNA extraction was performed by a modified phenol-chloroform method (4), using approximately 90mg of bone and 60mg of dental pulp. All manipulations of extracts and downstream products were performed in facilities with no prior exposure to *Y. pestis* DNA. In total, 50 bone samples and 60 teeth samples were extracted. Multiple negative controls were included throughout.

qPCR primers and conditions for sample screening

3.11.1.3.1 Internal positive control (inhibition test)

Forward primer: AGGAGCATGCCTAATTACCCA

Reverse primer: GATGAAAATGCAGTTATTGTGTCA

Mammoth cytochrome b standard (positions 5 – 1031 of the cytochrome b gene):
CCCACATTCGAAAATCTCACCCCTTACTTAAAATCATCAATAAATCCTTCATTGATCTACCTACCC
CATCCAACATCTCAACATGATGAAATTTCGGCTCACTACTAGGAGCATGCCTAATTACCCAAATC
CTAACAGGATTATTCCTAGCCATACATTATACACCCGACACAATAACTGCATTTTCATCTATATCC
CATATTTGCCGAGATGTGAACTACGGCTGAATTATTCGACAACTACACTCAAACGGAGCATCCA
TTTTCTTCCTCTGCCTATACACACACATTGGACGAAACATCTACTATGGGTCCTACCTATACTCGG
AAACTTGAAATACCGGCATTATATTACTACTAATCACCATAGCCACCGCCTTCATAGGATATGTC
CTTCCGTGAGGACAAATATCATTCTGAGGGGCAACCGTAATCACTAACCTTTTCTCAGCAATCCC
TTATATCGGCACAAACCTAGTAGAATGAATCTGAGGAGGCTTTTCAGTAGATAAAGCAACCTTA

AATCGATTTTTCGCCCTCCATTTCATTCTTCCATTTACTATAATTGCACTAGCAGGAGTACACCTA
ACCTTTCTTCACGAAACAGGCTCAAACAATCCACTGGGCCTCACTTCAGACTCAGACAAAATCCC
CTTTCACCCATACTATACCATTAAGGACTTCCTAGGATTACTTATCCTAATTTTACTTCTTCTACTC
CTAGCCCTACTATCTCCTGACATACTAGGAGACCCTGACAACTACATACCAGCCGACCCACTAAA
TACTCCCCTACATATCAAGCCAGAGTGATATTTTCTCTTTGCTTACGCCATCCTACGATCTGTACC
AAACAAACTAGGAGGCGTCCTAGCCCTACTCCTATCAATTCTAATCCTAGGATTAATACCACTTC
TCCATACATCCAAGCACCGAAGCATAATACTCCGACCTCTTAGCCAAGTCCTATTCTGAACTCTA
ACAATAGATTTACTAACACTTACATGAATTGGCAGCCAACCAGTAGA

PCR conditions consisted of 1 unit 10X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM each dNTP, 0.2mg/ml BSA, 300 nM each primer, 0.167 units SYBR, and 0.05 units/µl of *Taq* GOLD DNA polymerase (Applied Biosystems) in 20µl reactions with 5µl of ancient extract and 500 copies of the mammoth cytochrome b standard. Thermal profile included a 7 minute denaturation at 95°C, followed by 45 cycles consisting of a 30 second denaturation at 95°C, a 60 second annealing at 60°C, and a 90 second elongation at 72°C. Amplifications were followed by a dissociation gradient from 55 - 95°C. For extracts that demonstrated a cycle shift greater than 0.5 compared to the standard sans ancient extract, the reaction was again attempted under the same reaction conditions with 0.1 units/µl of *Taq* GOLD.

3.11.1.3.1 Standard generation and design of *pla* and *caf1M* assays:

Standards were generated in Bioedit by aligning all available *Yersinia pestis* sequences in the NCBI database for the region of interest. Primers were designed using IDT software (www.idtdna.com), using the following criteria: product length of 50 – 60 bp, annealing temperature of 46 – 65°C, and 40 – 60% GC content. Selection of primer sets was made based on the specificity to *Yersinia pestis* of the region between primers as determined using the NCBI BLAST feature. For specific regions, a transversion was introduced outside of the primer binding sites. This altered sequence was again BLASTed, and confirmed to fall outside of the variant common to all *Yersinia pestis* sequences in the NCBI database. HPLC-purified synthetic oligos (IDT) we rendered double stranded in PCR reactions under the following conditions: 1 unit 10X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM each dNTP, 300 nM each primer, and 0.05 units/µl of *Taq* GOLD DNA polymerase. Thermal profile consisted of a 5 minute initial denaturation at 95°C, followed by 40 cycles of a 30 second denaturation at 95°C, a 45 second annealing at 61 °C, and a 45 second elongation at 72°C. Products were purified over 10K Pall plates (Pall Corporation) and quantitated via a Turner Biosystems DNA fluorometer and Genequant UV spectrophotometer. Total copy numbers were determined, and 10-fold serial dilutions down to 0.04 copies/µl were made in 0.1x TE.

3.11.1.3.2 Assay for the plasminogen activator sequence (*pla*):

forward primer: ATGCCCTGAAAGACGTGGAGAA

reverse primer: GGGCGCTCATTCTGTTGTTT

synthetic oligonucleotide (synthetic transversion shown in bold): CCATGCCCTGAAAGACGTGGAGAATGTCAA**T**GCAAAACAACAGAATGAGCGCCCCG

PCRs were performed in 20µl reactions consisting of 1 unit 10X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM each dNTP, 0.2mg/ml BSA, 300 nM each primer, 0.167 units SYBR, 0.05 or 0.1 units/µl of *Taq* GOLD DNA polymerase (Applied Biosystems) depending on the level of inhibition for the sample as determined in the IPC test, and 5µl of DNA extract. Thermal profile included a 12 minute denaturation at 95°C, followed by 45 cycles consisting of a 30 second denaturation at 95°C, a 60 second annealing at 61°C, and a 45 second elongation at 72°C. Amplifications were followed by a final elongation at 72°C and a dissociation gradient from 65 - 95°C.

3.11.1.3.3 Assay for the chaperone protein for the fraction 1 antigen (*caf1M*):

forward primer: TATAGCTCAGGCTGGAGGTGTT

reverse primer: CGCATAACCACTTTAGGCTCTC

synthetic oligo (synthetic transversion shown in bold): GCGTATAGCTCAGGCTGGAGGTGTTTTCC**A**GCGAGATAAAGAGAGCCTAAAGTGGTTATGCGT A

PCRs were performed in 20µl reactions consisting of 1 unit 10X PCR Buffer II, 2.0 mM $MgCl_2$, 250 µM each dNTP, 0.2mg/ml BSA, 300 nM each primer, 0.167 units SYBR, 0.05 or 0.1 units/µl of *Taq* GOLD DNA polymerase (Applied Biosystems) depending on the level of inhibition for the sample as determined in the IPC test, and 5µl of DNA extract. Thermal profile included a 5 minute denaturation at 95°C, followed by 45 cycles consisting of a 30 second dissociation at 95°C, a 60 second annealing at 63°C, and a 45 second elongation at 72°C. Amplifications were followed by a final elongation at 72°C and a dissociation gradient from 65 - 95°C.

3.11.1.3.4 Multiplex assays for the *pla* gene and its flanking intergenic spacers.

Primers are displayed in Table S2 below, and were designed with the following criteria: product length of 60 – 90 bp, annealing temperature of 46 – 65°C, and 20 – 80% GC content PCR conditions consisted of 1 unit 10X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM each dNTP, 0.2mg/ml BSA, 200 nM each primer, 0.167 units SYBR, 1 unit of *Taq* GOLD DNA polymerase (Applied Biosystems), and 2µl of a 1:100 dilution of the appropriate amplified 454 library. Thermal profile included a 10 minute denaturation at 95°C, followed by 45 cycles consisting of a 30 second dissociation at 95°C, a 30 second

annealing at the temperature specified by the multiplex, and a 45 second elongation at 72°C. Amplifications were followed by a final elongation at 72°C and a dissociation gradient from 65 - 95°C. All products were cloned with TOPO TA cloning kits (Invitrogen) following the manufacturer's protocols. Sequencing reactions were run on an ABI 3130 using either BD v1.1 or 3.1 chemistry. Sequences were visualised in Bioedit and Geneious 5.1.7.

3.11.1.4 DNA extraction and enrichment at MPI EVA

All pre-amplification steps were carried out in a clean room facility where procedures that minimize contamination from present-day human DNA are rigorously implemented (5). An aliquot from each DNA extract was used to produce libraries using a modified Illumina multiplex protocol (6). Each extract received an individual indexing adapter allowing for easy discrimination between individual samples after pooling and sequencing. After indexing PCR, the libraries were amplified an additional 10 cycles in a 100µl reaction containing 50µl Phusion™ High-Fidelity Master Mix and 500nM of primers sitting at the outer P5 and P7 *Illumina* library grafting sequences. The annealing temperature was 60°C. The amplified product was spin column purified and quantified on an Agilent 2100 Bioanalyzer DNA 1000 chip.

3.11.1.4.1 ***DNA enrichment***. Long-range PCR products spanning the entire human mitochondrial genome were produced following the methods of (7). Long range products for the complete *Y. pestis* pPCP1 were produced from modern *Y. pestis* genomic DNA provided by the Pasteur Institut, Paris. *Yersinia pestis* long range PCR primers are as follows, and were designed using Primer3:

FWD 1 – 23: ATCCACACCCAACGCCTGAA

RVS 1 – 3378: GAGCGAGGAAGCGGAAGAGC

FWD 2 – 3004: ATGCTCTGGATGCCGACGAG

RVS 2 – 6413: TCTTCCCGTTCCTGCCCTTC

FWD 3 3 – 6199: GAAGAACGGACAGCCCGTGA

RVS 3 3 – 69: TTCAGGCGTTGGGTGTGGAT

PCRs were performed using the Roche long range PCR kit in 100ul reactions using 1 unit of 10x PCR Buffer, Mg, 1250 µM each dNTP, DMSO, 0.4 mg/ml SBA, 300 nM each primer, and 0.7U/µl of polymerase. Thermal profile consisted of an initial denaturation at 92C for 2 minutes, followed by 9 cycles consisting of a denaturation for 10s at 92C, 15s annealing at 68C, and a 9 minute elongation at 68C. A further 28 cycles were carried out using the above thermal profile, though with increasing the elongation time by 20 seconds each cycle.

PCR products were purified using Qiagen spin columns and quantified by NanoDrop. The two mitochondrial PCR products were pooled in equimolar amounts to a total amount of 3 µg. The same was done with the pPCP1 products. The pooled PCR products were then sonicated (Bioruptor, Diogenode, Liege, Belgium), producing fragments of 150 to 700 bp as observed on a 2% agarose gel. The products were biotinylated by ligation to a biotin-carrying adapter and immobilized on streptavidin-coated magnetic beads. The amplified libraries were single-stranded by incubating them at 95ºC for 3 min. Each library was divided in an aliquot that was incubated with streptavidin beads coated with the mitochondrial fragmented long range PCR product and a second aliquot incubated with pPCP1 coated beads. The mixture was incubated under rotation at 65ºC in a hybridization oven (SciGene, Model 700, Sunnyvale, CA, USA). After 48 hours, the beads were washed and library molecules were eluted by NaOH melting. The DNA concentration was measured by qPCR (Mx3005P Real Time PCR System, Stratagene, La Jolla, CA), the eluted library was further amplified for 15 cycles using primers complementary to the outer P5 and P7 *Illumina* library grafting sequences. The individual mtDNA as well as pPCP1 libraries were subsequently pooled in a ratio of 20% mtDNA and 80% pPCP1.

3.11.1.4.2 ***Solexa DNA sequencing analysis***. The raw reads were aligned to the PhiX 174 reference sequence to obtain a training data set for the base caller Ibis (8). Raw sequences called by Ibis 1.1.1 were filtered for the individual indexes used as described (9). The paired-end reads were subjected to a fusion process (including removal of adapter sequences and adaptor dimers) by requiring at least an 11nt overlap between the two reads. In the overlapping sequence, quality scores were combined and the base with the highest base quality score was called. Only sequences merged in this way were used for further analysis. The small proportion of molecules longer than 191nt was thus discarded. Since several amplification steps were performed we filtered for uniqueness by grouping sequences with the same direction, start, and end coordinates and considering only fragments that occurred at least three times, since several amplification steps were performed. From each such cluster a consensus sequence was generated by taking, for each position, the base with the highest quality score. That resulted in a total of 1,354,268 unique merged sequences that were used as input of the iterative mapping assembler (10).

3.11.2 Results

3.11.2.1 Figures

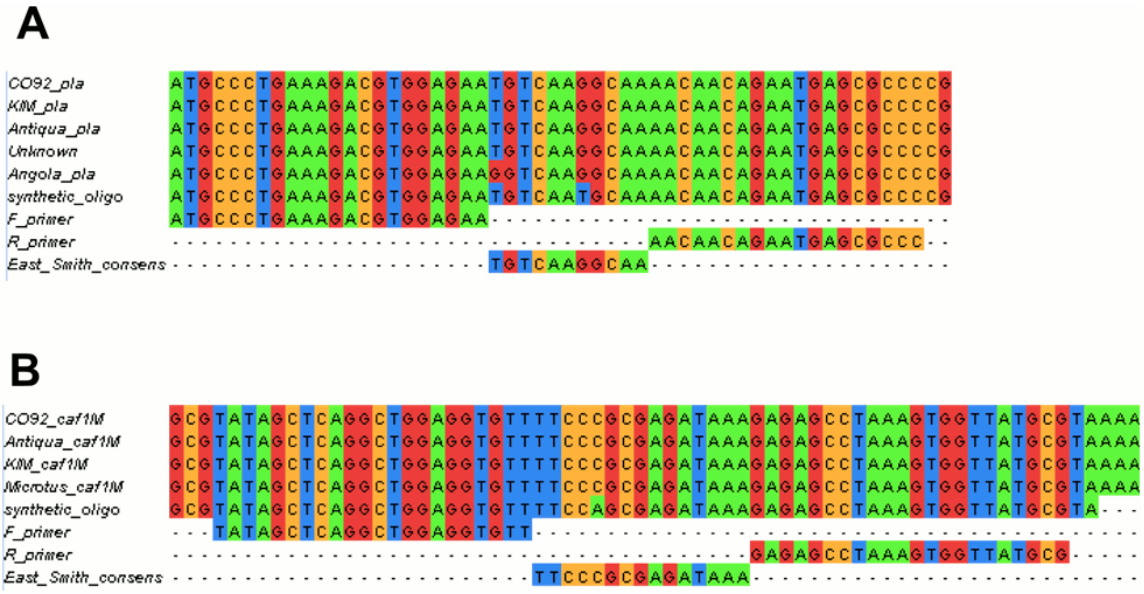


Figure S1 – Sequence alignment showing the positions of qPCR primers and synthetic oligonucleotides for A) plasminogen activator (*pla*) and B) chaperone protein for the fraction 1 antigen (*caf1M*).

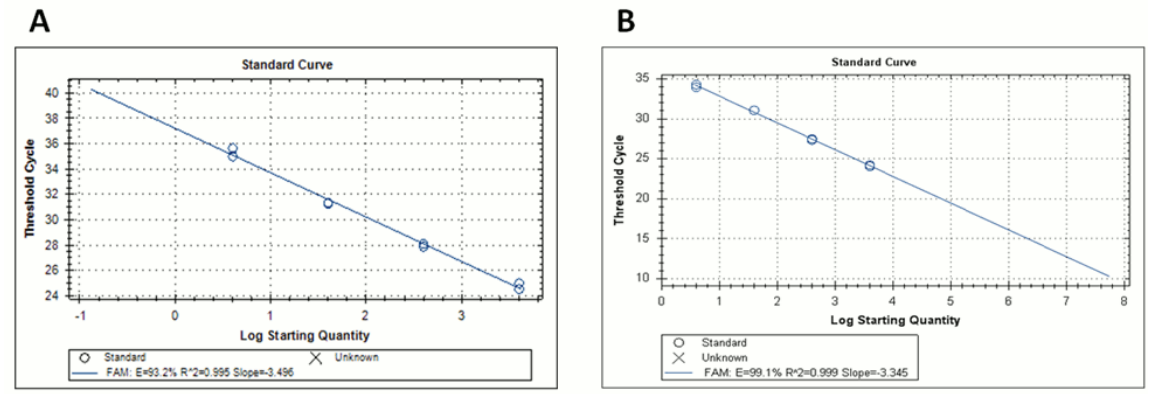


Figure S2 – sensitivity data for the plasminogen activator assay (**A**) and the fraction 1 antigen assay (**B**). Plasminogen activator assay is shown here to be reliably sensitive to 4 copies of initial template, with an R^2 value of 0.995 and an overall reaction efficiency of 93.2%. The fraction 1 antigen assay is shown here to be reliably sensitive to 4 initial template quantities, with an R^2 value of 0.999 and an overall reaction efficiency of 99.1%. Both sensitivity assays were conducted using only synthetically-produced oligonucleotides that differ from all known wild type *Y. pestis* sequences.

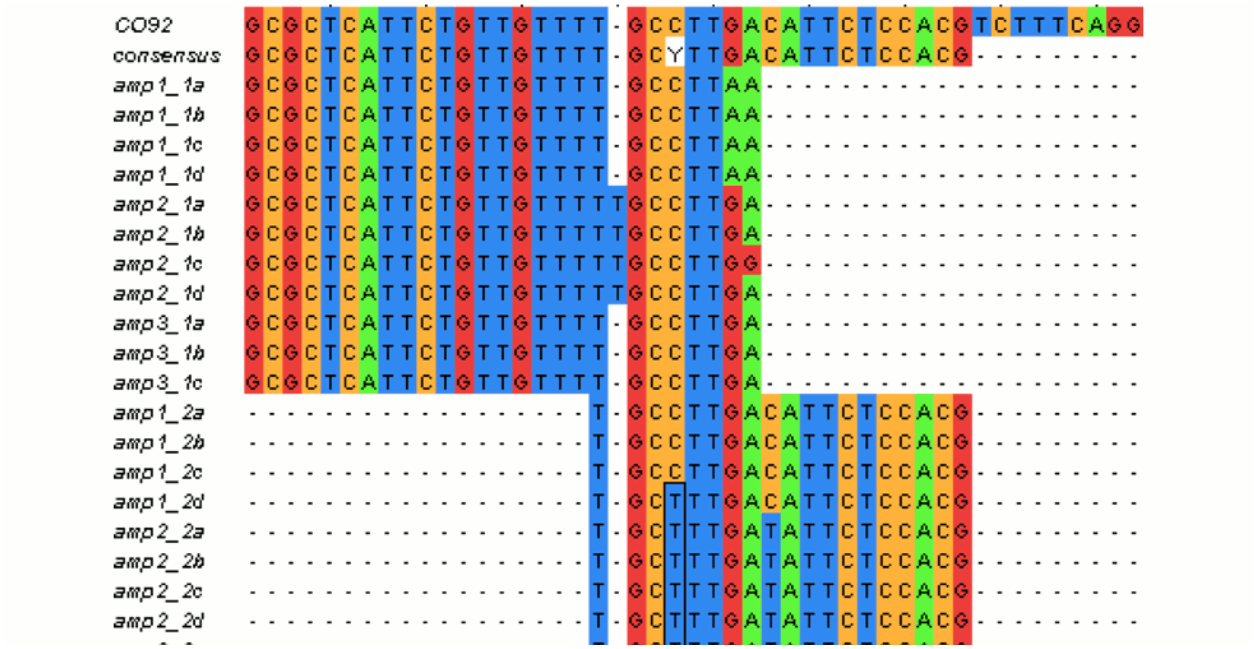


Figure S3 - Sequence data from multiplex PCR method demonstrating the manner in which damaged positions were identified and resolved (amp# = amplified library; # = PCR, lower case letters = individual clones). Boxed nucleotides indicate positions that could not be resolved via this method.

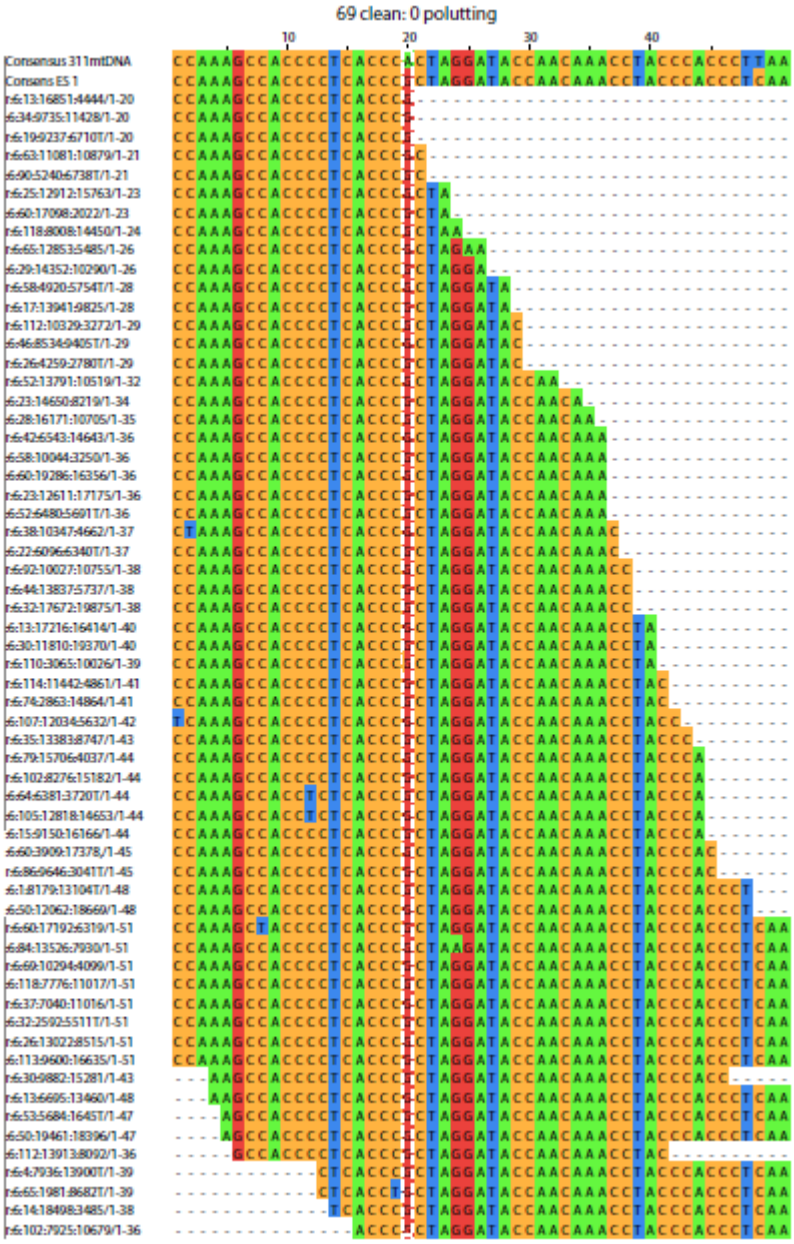


Figure S4 – Depiction of endogenous and contaminant mitochondrial fragments, showing the method used to identify polluting human mtDNA molecules.

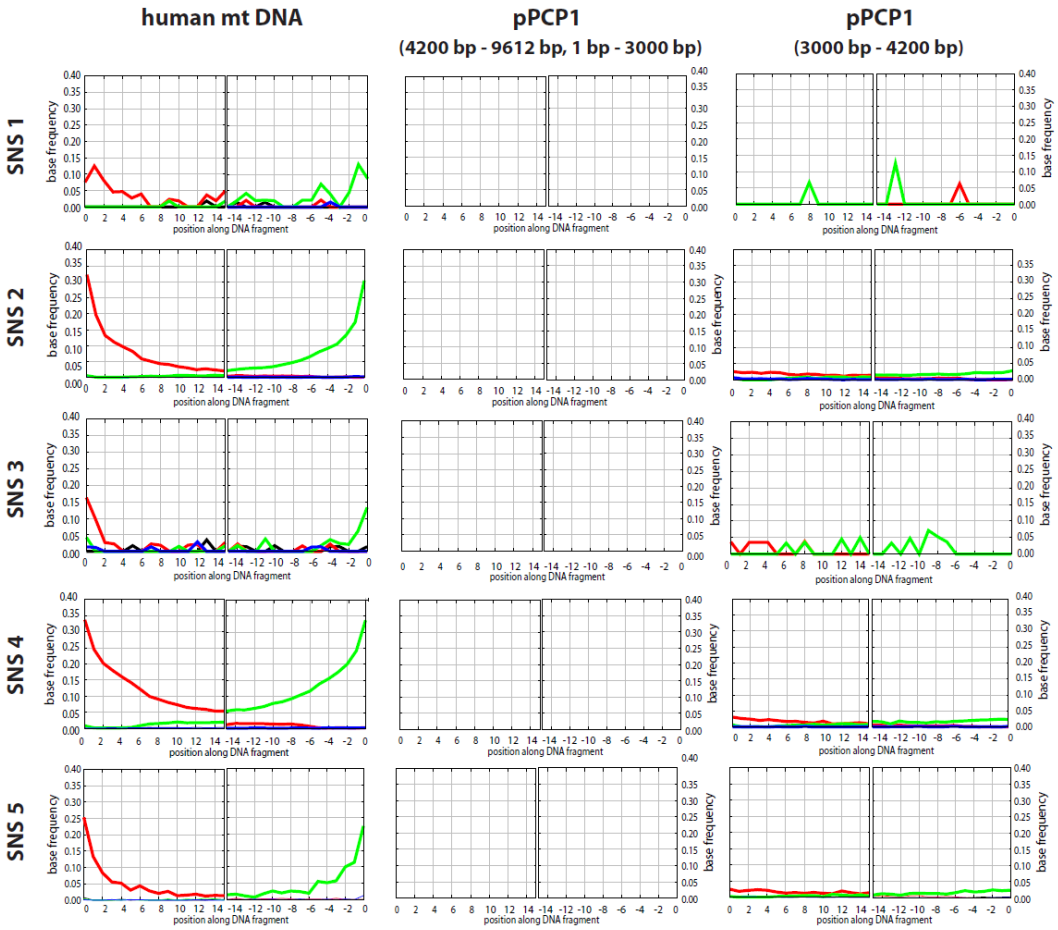


Figure S5 – Damage pattern for mtDNA, pPCP1 (4200 – 9612 bp, 1 – 3000 bp), pPCP1 (3000 – 4200 bp) for the samples SNS1, SNS 2,SNS 3,SNS 4, SNS 5, ES 1, ES 2, ES 3.1, ES 3.2, ES 4.1, ES 4.2, ES 5, cave bear control, extraction blank, and library control.

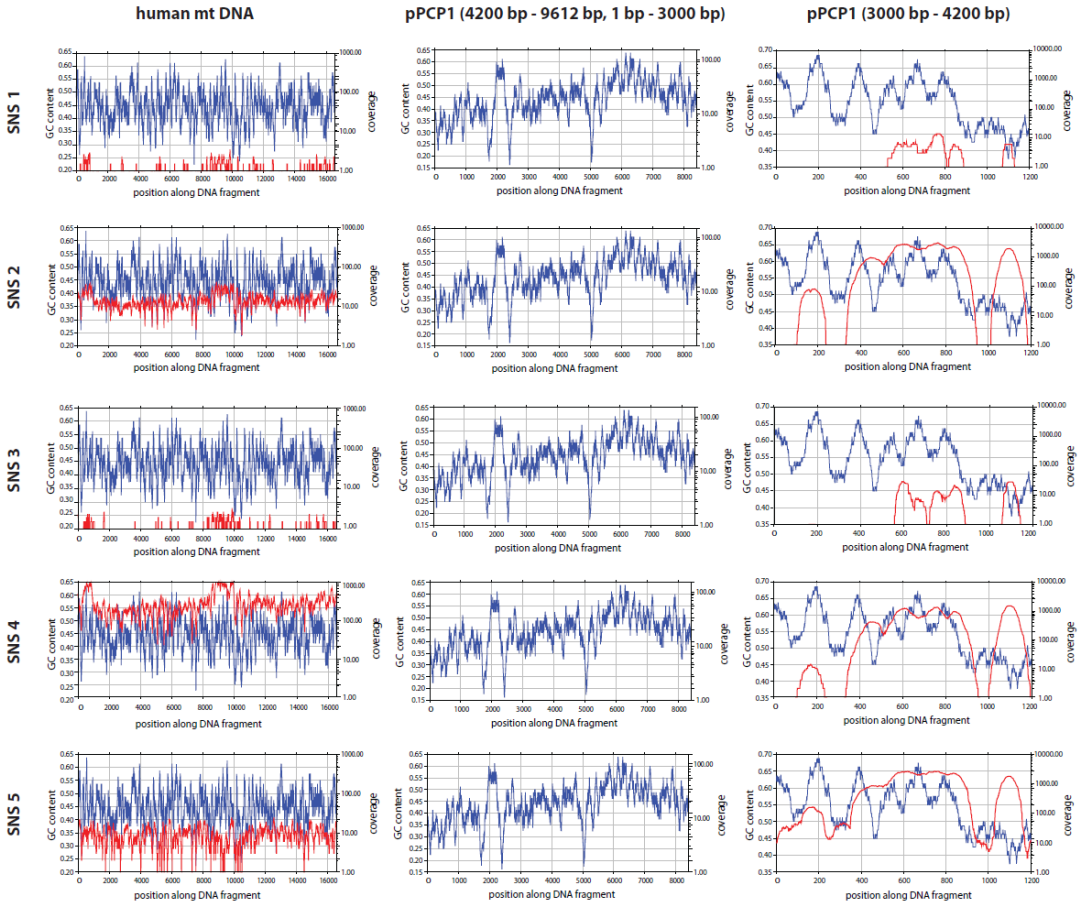


Figure S6 – GC content and coverage for mtDNA, pPCP1 (4200 – 9612 bp, 1 – 3000 bp), pPCP1 (3000 – 4200 bp) for the samples SNS1, SNS 2,SNS 3,SNS 4, SNS 5, ES 1, ES 2, ES 3.1, ES 3.2, ES 4.1, ES 4.2, ES 5, cave bear control, extraction blank, and library control.

3.11.2.2 Tables

Museum of London Identification	Collection	Demographic information	Skeletal Element	Amount used in extraction (mg)	*Taq* GOLD used in qPCR		of *pla*		of *caf1M*	
					SN	P	SN	P	SN	P
5960	ES	adult male	bone	115	0.05	0.10	0	0	-	-
6108	ES	adult female	bone	107	0.10	0.10	0	0	-	-
6327	ES	adult male (senile)	bone	94	0.10	0.10	0	0	-	-
9511	ES	subadult	bone	115	0.10	0.10	0	0	-	-
9741	ES	subadult (male adolescent)	bone	131	0.10	0.10	0	0	-	-
11109	ES	adult male	bone	115	0.05	0.05	0	0	-	-
11117	ES	adult female	bone	94	0.10	0.10	0	0	-	-
11622	ES	adult male	bone	97	0.10	0.10	0	0	-	-
12801	ES	subadult	bone	94	0.05	0.10	0	0	-	-
12850	ES	subadult	bone	92	0.05	0.10	0	0	-	-
6313	ES	adult male	bone	116	0.05	0.10	0	0	-	-
6398	ES	adult female	bone	110	0.10	0.10	0	0	-	-
8305	ES	adult male	bone	112	0.05	0.10	0	0	-	-
9035	ES	adult male	bone	100	0.10	0.10	0	0	-	-
9056	ES	adult	bone	123	0.05	0.10	<4	0	-	-
9576	ES	subadult	bone	110	0.10	0.10	0	0	-	-
11448	ES	adult	bone	125	0.05	0.10	0	0	-	-
11747	ES	subadult	bone	97.4	0.05	0.10	0	0	-	-
12915	ES	subadult	bone	122	0.10	0.10	0	0	-	-
8057	ES	subadult	bone	128	0.05	0.05	0	0	-	-
8345	ES	subadult	bone	88.4	0.10	0.10	0	0	-	-
8424	ES	adult female	bone	97.5	0.05	0.10	0	0	-	-
9525	ES	adult male	bone	113.6	0.05	0.10	0	0	-	-
9782	ES	adult female	bone	96.5	0.05	0.10	0	0	-	-
11449	ES	adult male	bone	87.7	0.05	0.10	<4	0	-	-
11780	ES	adult female	bone	106.7	0.05	0.10	0	0	-	-
11893	ES	adult	bone	105.8	0.05	0.10	0	0	-	-

12809	ES	adult	bone	91.3	0.05	0.10	<4	0	-	-
5209	ES	subadult	bone	85	0.05	0.10	0	0	-	-
5317	ES	subadult	bone	108.2	0.05	0.10	0	0	-	-
5346	ES	subadult	bone	119	0.05	0.10	0	0	-	-
6503	ES	adult	bone	91	0.05	0.10	0	0	-	-
8235	ES	adult male	bone	97	0.05	0.05	0	0	-	-
9575	ES	subadult	bone	104	0.05	0.05	0	0	-	-
11607	ES	adult	bone	94	0.05	0.05	0	0	-	-
11619	ES	subadult	bone	83	0.05	0.05	0	0	-	-
11838	ES	adult female	bone	88.9	0.05	0.10	0	0	-	-
12656	ES	subadult male	bone	98	0.05	0.05	0	0	-	-
5812	ES	adult probable female	bone	81.7	0.05	0.10	0	0	-	-
5916	ES	adult	bone	85.6	0.05	0.10	0	0	-	-
6287	ES	adult female	bone	105.3	0.10	0.10	0	0	-	-
7375	ES	adult male	bone	95	0.10	0.10	0	0	-	-
8400	ES	subadult	bone	99.2	0.10	0.05	0	0	-	-
9710	ES	adult probable male	bone	103	0.10	0.10	0	0	-	-
9848	ES	subadult	bone	102.8	0.10	0.10	0	0	-	-
9856	ES	adult male	bone	88.9	0.05	0.10	0	0	-	-
12503	ES	adult male	bone	114.3	0.10	0.10	0	0	-	-
12795	ES	subadult	bone	110	0.10	0.10	0	0	-	-
5960	ES	adult male	tooth	94.8	0.05	0.05	<4	0	-	-
6409	ES	adult female	tooth	42.1	0.05	0.05	0	0	-	-
8057	ES	subadult	tooth	69.1	0.05	0.05	<4	0	-	-
8124*	ES	adult female	tooth	64.2	0.10	0.05	15	26	4	-
8291*	ES	subadult	tooth	45	0.05	0.05	30	<4	5	<4
8293	ES	adult male	tooth	50.7	0.05	0.05	0	0	-	-
9522	ES	adult male	tooth	62.7	0.05	0.05	0	0	-	-
9731	ES	adult male	tooth	73.3	0.05	0.05	0	0	-	-
11972*	ES	adult female	tooth	44.8	0.05	0.05	26	<4	<4	<4

12522	ES		tooth	55.7	0.05	0.05	<4	0	-	-
12553	ES		tooth	71.6	0.05	0.05	0	0	-	-
12814	ES	adult female	tooth	58.2	0.05	0.05	<4	<4	-	-
5053*	SNS		tooth	90.1	0.05	0.05	0	0	-	-
5086	SNS		tooth	63.8	0.05	0.05	0	0	-	-
5122*	SNS		tooth	92.6	0.05	0.05	0	0	-	-
5168	SNS		tooth	46.6	0.05	0.05	0	0	-	-
5169*	SNS		tooth	39.9	0.05	0.05	0	0	-	-
5191	SNS		tooth	69.9	0.05	0.05	0	0	-	-
5194*	SNS		tooth	41.9	0.05	0.05	0	0	-	-
5200*	SNS		tooth	39.6	0.05	0.05	0	0	-	-
5237	SNS		tooth	54.7	0.05	0.05	0	0	-	-
5243	SNS		tooth	56.8	0.05	0.05	0	0	-	-
5291	ES		tooth	113	0.10	0.05	0	0	-	-
6545	ES	adult male	tooth	83.2	0.05	0.05	0	0	-	-
7055	ES	adult male	tooth	104.7	0.10	0.05	0	0	-	-
8161	ES	adult female (senile)	tooth	99.8	0.10	0.05	<4	<4	-	-
8235	ES	adult male (senile)	tooth	66.5	0.10	0.05	0	0	-	-
8449	ES		tooth	82	0.10	0.05	0	0	-	-
9056	ES	adult	bone	24.8	0.10	0.05	0	0	-	-
9807	ES	adult female	tooth	49.6	0.05	0.05	0	0	-	-
11314	ES		tooth	104.7	0.10	0.05	0	0	-	-
11780	ES	adult female	tooth	49	0.05	0.05	0	0	-	-
11972	ES	adult female	tooth	40.3	0.05	0.05	0	0	-	-
12566	ES		tooth	88.6	0.10	0.05	<4	0	-	-
12636	ES	juvenile	tooth	82.2	0.10	0.05	<4	0	-	-
12822	ES		tooth	53.6	0.05	0.05	0	0	-	-
12843	ES	juvenile	tooth	73.2	0.10	0.05	0	0	-	-
5344	ES		tooth	73.1	0.05	0.05	<4	0	-	-
5812	ES	adult	tooth	84.9	0.05	0.05	6	<4	<4	0

5862	ES	juvenile male	tooth	90.3	0.10	0.05	0	0	-	-
6330*	ES		tooth	98.7	0.05	0.05	8	10	<4	<4
6443*	ES		tooth	73.1	0.05	0.05	15	0	0	0
6518	ES	juvenile	tooth	48.7	0.10	0.05	0	0	-	-
6654	ES		tooth	69.3	0.10	0.05	0	0	-	-
7366	ES		tooth	82.4	0.05	0.05	0	0	-	-
8291	ES	juvenile	tooth	39.8	0.05	0.05	<4	0	-	-
8329	ES		tooth	42.6	0.05	0.05	0	0	-	-
8379	ES		tooth	91	0.10	0.05	0	0	-	-
11109	ES	adult female	tooth	92.6	0.05	0.05	0	0	-	-
11939	ES	juvenile	tooth	51.6	0.05	0.05	<4	0	-	-
12510	ES		tooth	70.4	0.05	0.05	0	0	-	-
12801	ES	juvenile	tooth	54.6	0.05	0.05	0	0	-	-
12839	ES		tooth	112.7	0.10	0.05	0	0	-	-
8075B	ES		tooth	36.9	0.10	0.05	0	0	-	-
8075T	ES		bone	117.1	0.10	0.05	0	0	-	-
8412	ES		tooth	57.2	0.05	0.05	<4	0	-	-
9674	ES		tooth	34	0.05	0.05	0	0	-	-
11449	ES	adult male	bone	85.5	0.10	0.10	0	0	-	-
12522B	ES		bone	115.9	0.10	0.10	0	0	-	-
12522T	ES		tooth	29.8	0.05	0.05	0	0	-	-
12809	ES	adult	bone	79	0.10	0.05	0	0	-	-

Table S1 – List of all skeletal samples obtained from the Museum of London. Demographic information, where available, after (11). Units/µl of *Taq* GOLD used in qPCR determined as per the IPC test. Each *pla* and *caf1M* qPCR was performed one time per sample. ES = East Smithfiled; SNS = St. Nicholas Shambles; SN = supernatant; P = pellet; * = samples sent to the MPI EVA for independent replication.

Primer name	forward primer sequence	Tm (for)	reverse primer sequence	Tm (rev)	Ta	Amp. size (between primers)
M01P01	AAGGGCAGGAACGGGAAGATTTAC	59.2	TTTCAGATCCGACAACGAGCG	57.5		22
M01P02	GCTGAGTGGAAAGTCTCATGAAATGC	58.2	TCTTCCAGTCTAACTGGCTGATCT	57.2		24
M01P03	GAGAATTATAAAGCAGGTATAACAGCAGG	55.3	TGAACCACCTGTAGCTGTCCAA	58		21
M01P04	CGCTGGATATTATGTCACACCTAATGC	57.3	TGAGTACCTCCTTTGCCCTCATCA	59.3	62	33
M02P01	TCCCGTTATCAGTACCATCG	53.7	AAATATCCAGGGAAATATGATCT	49.2		39
M02P02	ACCAACCTTCAGATGTGTG	52.4	AACAACAGAATGAGCGCC	53.5	56	29
M03P01	TCAACGCTCGTTGTCGGATCT	58.7	GACTTTGCTCACATAAGAAACGGTGG	58.1		40
M03P02	GAGAGATTAAGGGTGTCTAATGAAGA	53.8	GCATTAGCACTCCCGGACAGAATA	58.7		28
M03P03	GCCAATGAATATGACCTCAATGTGAAA	56	CCTGATATCCTGCTGTTATACCTGCTT	58		28
M03P04	GAAAGGAGTGCGGGTAATAGGTTAT	56.4	CATTAATGCGATACTGGCCTGC	56.2		33
M03P05	TGCCTTGACATTCTCCACG	55.2	AGGAGAGTAGCGTTCCATGTCT	57.4	60	28
M04P01	AAAGATCATATTTCCCTGGAT	48.8	CTCGTCGTCTGACAGAAT	51.3		36
M04P02	CAGACCATTGATAAGAATAGTGGAG	53	CGCCGTCACAGTATAA	48.3		46
M04P03	TTCACCATAATGACGGG	48.8	CCTGAAAGACGTGGAGAATG	53.2	52	26
M05P01	CGTTCATGCAGAGAGATTAAGGGTG	57.2	GGACAGAATAGTTATAATGGTTGCC	53.3		23
M05P02	CAGGAAGAAAGATCAGCCAGTTAGA	56.3	GGATCCCAGGATATATCACCTTTCAGG	57.9		24
M05P03	CCGGGTCAGGTAATATGGATGACT	58	CAGGATGAGATGAGTGATCTGTCCAC	58.2		28
M05P04	AAAGTCTTTGCGGAATTTACATACAGT	55.1	CAGAATCTCCACTATTCTTATCAATGGTCT	55.9	60	28
M06P01	AGTTCTATTGTGGCAACC	49.4	GGTATTAACTGAGATGATGCTGC	53.3		30
M06P02	ACTACGACTGGATGAATGA	49.8	GGCATGATTAACATTTGTAGCAGG	54.3	56	34
M07P01	AACCATTATAACTATTCTGTCCGGGAG	56.1	AACTGTAAAGCTGTCAGGG	52.1		38
M07P02	GAGTGGACAGATCACTCATCTCA	55.5	GCCTTTCACATTGAGGTCATATTC	53.9		28
M07P03	CACGTTTCAGTTGGACAGCTAC	55.8	TCGGGAAGTTTCCGGTATAAGC	56.5	58	28
M08P01	CAGTTAATACCAAATATATCCC	46.1	GACTTTCCACTCAGCATC	50.4		31
M08P02	ACGTCTCTGGCTTCC	50.5	GTCCACTCAGATTGATTT	46.2	52	38
M09P01	TTTACAGTTGCAGCCTCCAC	55.2	CTTCCTGTTTCTGCGTCA	52.1		37
M09P02	ATGTCGCTATCCTGAAAGGTGA	55.5	CCTGGCATTCAGGGTCAGAAAT	57.1		21
M09P03	ACAGCAGGATATCAGGAAACACG	56.9	GCTCCATTATTATAACTATATGAACCACC	53.4	57	19
M10P01	GCTTTATGACGCAGAAACAGGAAG	56.2	CCTTTCAGGATAGCGACATT	59.1		31
M10P02	AAGACATCCGGCTCACGTTATT	56.5	AGACTTTGGCATTAGGTGTGAC	54.8		29
M10P03	TGAGGGCAAAGGAGGTACTCAG	58.4	GCATCTCCGCCAATAGAGACAGAA	58.4	61	23
M11P01	GGTGATATATCCTGGGATCCATACTCA	56.8	TATTACCTGACCCGGAAGCCAGAGA	60.6		30
M11P02	GGACAGCTACAGGTGGTTCATATAGTTA	57.5	CTATTACCCGCACTCCTTTCGGGAAGTT	61.5		23
M11P03	TGTCACACCTAATGCCAAAGTCT	56.5	TGAGTACCTCCTTTGCCCTCATCA	59.3	63	25
M12P01	TGATATATCCTGGGATCCATACTCATTTCT	56.7	TATTACCTGACCCGGAAGCCAGAGA	60.6		25
M12P02	TTGCAGGCCAGTATCGCATTA	56.9	ATTATCATGTGCCCGAACCCAGTC	58.9		35
M12P03	CACTATATGAGAGATCTTACTTTCCGTGAG	56.1	AGGTGTGACATAATATCCAGCGT	55.9	63	37
M13P01	TTTCTGACCCTGAATGCCAG	54.7	CGTAGTCATCCATATTACCTGACCCG	58.2		21
M13P02	ATGGTACCGTAATTAACGCTGGATATTA	55.9	ACTGTATGTAAATTCCGCAAAGAC	53.3	59	19

M14P01	GGCTGGTTACTCCAGGATGAGAATTA	58	GCTGTCCAACTGAAACGTGTTTCCTGA	60.9		24
M14P02	CTGGGTTCGGGCACATGATAA	57.4	ATAACGTGAGCCGGATGTCTTCTC	58.3		34
M14P03	TATTGGCGGAGATGCTGCCGGTATTT	62.1	TTCAGAAGCGATATTGCAGACCCG	58.9	63	24
M15P01	TTATAATAATGGAGCTTATACCG	47.4	AAAGCGCTGGTTAT	43.2		33
M15P02	TGCCATATATTGGACTTG	46	GTCGCTGAATTTAAATAAT	42.7	52	37
M16P01	ATTATTTAAATTCAGCGAC	42.7	CTTCTCACGGAAAGTAAG	46.8		42
M16P02	TGCCGGTATTTCCA	45.8	GAGAGATATGATCTGTATT	42.2		46
M16P03	TGTGGAAAGGAGGT	45.1	CCGTCATTATGGTG	42	52	43
M17P01	GGGTCTGCAATATCGCTTCTGA	57	TTCCACAGACATCCTCCC	53.8	58	40
M18P01	ACAGATCATATCTCTCTTT	44.4	TGGTCAAACACCAACC	49.4	53	38
M19P01	GGGCGCTCATTCTGTTGTTT	56.3	TGACCTTCTCCATGCCCTGAAAGA	60	62	19
M20P01	GACAGATCACTCATCTCATCCTGC	56.7	AGCCTTTCACATTGAGGTC	52.2	58	28
M21P01	CTGCTACAAATGTTAATCATGCC	52.3	TTCTCATCCTGGAGTAACCAGC	56	58	25
M22P01	TCAGCCAGTTAGACTGGAAGATCA	57.1	GGTCAGAAATGAGTATGGATCCCAGG	58.6	61	30
Yersinia_ plaF	GGAGGTACTCAGACCATTGAT	53.6	TTTTTATTGGAAATACCGGCAG	51.6	55	22
Yersinia_ plaG	AGATGCTGCCGGTATTTC	52.5	AGATATGATCTGTATTTTTCAGAAGC	51.3	55	38
Yersinia_ plaH	CGGCGGGTCTGCAATATC	56.3	ACAGACATCCTCCCCGCT	58.5	59	44
Yersinia_ plaI	TCTCTCTTTTCATCCTCCCC	53.5	GATTTTTCACACATCTGAAGGTT	51.7	55	42
Yersinia_ plaJ	GGAGGTTGGTGTTTGACC	53.5	GCAAAACAACAGAATGAGCG	53	55	47

Table S2 – list of all primers used for multiplex PCR of the *pla* gene and its flanking intergenic spacer regions. Grey shading over primer names indicate pairs that did not yield expected products.

See Appendix 2

Dataset S1 – Consensus sequence for the plasminogen activator gene (*pla*) obtained via multiplex (McMaster University) and targeted enrichment (MPI EVA). For multiplex, each sequence was confirmed in a minimum of two independent amplified DNA libraries, where the number of clones from each library is separated by commas. X's in the multiplex consensus sequence indicate regions not attempted owing to limitations in primer design, and Y's indicate positions that could not be resolved via this method (nucleotides for unresolved positions are shown).

3.11.3 References Cited

1. Gottfried RS. 1983. The Black Death: natural and human disaster in medieval Europe. New York: Free Press.
2. Grainger I, and Hawkins D. 1988. Excavations at the Royal Mint site 1986-1988. The London Archaeologist 5:429-436
3. White WJ. 1988. Skeletal Remains from the Cemetery of St Nicholas Shambles, City of London. London: London and Middlesex Archaeological Society.
4. Schwarz C, Debruyne R, Kuch M, McNally E, Schwarcz H, Aubrey AD, Bada J, and Poinar H. 2009. New insights from old bones: DNA preservation and degradation in permafrost preserved mammoth remains. *Nucleic Acids Res* 37(10):3215-3229.
5. Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, and Hofreiter M. 2004. Genetic analyses from ancient DNA. *Annu Rev Genet* 38:645-679.
6. Meyer M and Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc* Issue 6; doi:10.1101/pdb.prot5448.
7. Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 5(11): e14004. doi:10.1371/journal.pone.0014004.
8. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* 10: R83.
9. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* 35: e97.
10. Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al (2008) A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell* 134(3):416-426.
11. DeWitte SN and Wood JW (2008) Selectivity of Black Death mortality with respect to preexisting health. *Proc Natl Acad Sci U S A* 105(5):1436-1441.
12. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondria variation. *Hum Mutat* 30(2): E386-E394.

CHATPER 4:

A draft genome of *Yersinia pestis* from victims of the Black Death

Bos KI*[1], Schuenemann VJ*[2], Golding GB[3], Burbano HA[4], Waglechner N[5], Coombes BK[5], McPhee JB[5], DeWitte SN[6,7], Meyer M[4], Schmedes S[8], Wood J[9], Earn DJD[5,10], Herring DA[11], Bauer P[12], Poinar HN[1,3,5], and Krause J[2,12].

1. McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main Street West, Hamilton ON, Canada L8S 4L8

2. Institute for Archaeological Sciences, Rümelinstr. 23, University of Tübingen, 72070 Tübingen, Germany

3. Biology Department, McMaster University, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

4. Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

5. Michael DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

6. Department of Anthropology, University of South Carolina, Columbia, SC

7. Department of Biological Sciences, University of South Carolina, Columbia, SC

8. Institute of Applied Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd Fort Worth Texas, 76107

9. Department of Anthropology and Population Research Institute, Pennsylvania State University, University Park, PA

10. Department of Mathematics and Statistics, 1280 Main Street West, Hamilton ON Canada, L8S 4L8

11. Department of Anthropology, McMaster University, 1280 Main Street West, Hamilton ON, Canada L8S 4L8

12. Human Genetics Department, Medical Faculty, University of Tübingen, 72070 Tübingen, Germany

* These authors contributed equally to this work

4.1 Summary paragraph

Technological advances in DNA recovery and sequencing have drastically expanded the scope of genetic analyses from ancient specimens where full genomic investigations are now feasible and are quickly becoming standard[1]. This trend has important implications for infectious disease research, since genomic data from ancient microbes may help to elucidate mechanisms of pathogen evolution and adaptation for emerging and re-emerging infections. Here we report a reconstructed ancient genome of *Yersinia pestis* at 30-fold average coverage from Black Death victims securely dated to episodes of pestilence-associated mortality in London, England, 1348 – 1350. Genetic architecture and phylogenetic analysis indicate that the ancient organism is ancestral to most extant strains and sits very close to the ancestral node of all *Y. pestis* commonly-associated with human infection. Temporal estimates suggest that the Black Death of 1346 – 1351 was the main historical event responsible for the introduction and worldwide dissemination of the ancestor to all currently circulating *Y. pestis* strains pathogenic to humans, and further indicates that contemporary *Y. pestis* epidemics have their origins in the medieval era. Comparisons against modern genomes reveal no unique derived positions in the medieval organism, suggesting that the perceived increased virulence of

the disease during the Black Death may not have been due to bacterial physiology. These findings support the notion that factors other than microbial genetics, such as environment, vector dynamics, and host susceptibility should be at the forefront of discussions regarding emerging *Y. pestis* infections.

4.2 Manuscript

The Black Death of 1347 – 1351, caused by the bacterium *Yersinia pestis*[2,3], provides one of the best historical examples of an emerging infection with rapid dissemination and high mortality, claiming an estimated 50% of the European population in only a five-year period[4]. Discrepancies in epidemiological trends between the medieval disease and modern *Y. pestis* infections have ignited controversy over the pandemic's aetiologic agent[5,6], and although ancient DNA investigations have strongly implicated *Y. pestis*[2,3], genetic changes in the bacterium may be partially responsible for differences in disease manifestation and severity. To understand the organism's evolution it is necessary to characterise the genetic changes involved in its transformation from a sylvatic pathogen to one capable of pandemic human infection on the scale of the Black Death, and to determine its relationship with currently circulating strains. Here we begin this discussion by presenting the first draft genome sequence of the ancient pathogen.

Yersinia pestis is a recently-evolved descendent of the soil-dwelling bacillus *Yersinia pseudotuberculosis*[7], which in the course of its evolution acquired two additional plasmids (pMT and pPCP1) that provide it with specialised mechanisms for infiltrating mammalian hosts. To investigate recent evolutionary changes in of one of these plasmids, we reported on the screening of 46 teeth and 53 bones from the East Smithfield (ES) collection of London, England for presence of the *Y. pestis*-specific pPCP1[3]. Historical data indicates that the ES burial ground was established in late 1348

or early 1349 specifically for interment of Black Death victims[8] (Figure S1 and S2), making the collection well-suited for genetic investigations of ancient *Y. pestis*. DNA sequence data for five teeth obtained via molecular capture of the full *pestis*-specific pPCP1 revealed a C to T damage pattern characteristic of authentic endogenous ancient DNA[9], and assembly of the pooled Illumina reads permitted the reconstruction of 98.68% of the 9.6kb plasmid at a minimum of 2-fold coverage[3].

To evaluate the suitability of capture-based methods for reconstructing the complete ancient genome, multiple DNA extracts from both roots and crowns stemming from four of the five teeth which yielded the highest pPCP1 coverage[3] were used for array-based enrichment (Agilent®) and subsequent high-throughput sequencing on the Illumina GAII® platform[10]. Removal of duplicate molecules and subsequent filtering produced a total of 2,366,647 high quality chromosomal reads (Tables S1a and S1b) with an average fragment length of 55.53bp (Figure S4), which is typical for ancient DNA. Coverage estimates yielded an average of 28.2 reads per site for the chromosome, and 35.2 and 31.2 for the pCD1 and pMT1 plasmids respectively (Figure 1a, c, and d; Tables S1b and S1c). Coverage was predictably low for the pPCP1 (Figure 1e) because probes specific to this plasmid were not included on the arrays. Coverage correlated with GC content (Figure S6), a trend previously observed for high-throughput sequence data[11]. The coverage on each half of the chromosome was uneven due to differences in sequencing depth between the two arrays with 36.46 and 22.41 average reads per site for array 1 and array 2 respectively. Although greater depth contributed to more

average reads per site, it did not increase overall coverage, with both arrays covering 93.48% of the targeted regions at a minimum of 1-fold coverage (Table S1b). This indicates that our capture procedure successfully retrieved template molecules from all genomic regions accessible via this method, and that deeper sequencing would not result in additional data for CO92 template regions not covered in our dataset.

Genome architecture is known to vary widely among extant *Y. pestis* strains[12]. To extrapolate gene order in our ancient genome, we analysed reads mapping to the CO92 reference for all extracts stemming from a single individual who yielded the highest coverage (individual 8291). Despite the short read length of our ancient sequences and the highly repetitive nature of the *Y. pestis* genome, 2221 contigs matching CO92 were extracted, comprising a total of 4,367,867 bp. To identify potential regions of the ancient genome that are architecturally distinct from CO92, all reads not mapping to the CO92 reference were in turn considered for contig construction. A conventional BLAST search queried against the CO92 genome revealed matches for 2105 out of the 2134 resulting contigs. Evidence of altered architecture was identified in 10 contigs (Table S2). An example of such a structural variant is shown in Figure 2, where reference-guided assembly incorporating unmapped reads to span the breakpoint validates its reconstruction. This specific genetic orientation is only found in *Y. pseudotuberculosis* and *Y. pestis* strains Microtus, Angola, Pestoides F, and B42003004 that are ancestral to all *Y. pestis* commonly-associated with human infections (branch 1 and branch 2 strains[13,14]). Furthermore, discrepancies in the

arrangement of this region in branch 1 vs. branch 2 modern *Y. pestis* strains indicate that rearrangements occurred as separate events on the different lineages.

Single nucleotide differences between our ancient genome and the CO92 reference surprisingly consisted of only 97 chromosomal positions, and 2 and 4 positions in the pCD1 and pMT1 plasmids respectively (Table S3), indicating tight genetic conservation in this organism over the last 660 years. Twenty-seven of these positions were unreported in a previous analysis of extant *Y. pestis* diversity[14] (Table S3, S4). Comparison of our ancient genome to its ancestor *Y. pseudotuberculosis* revealed that the medieval sequence contained the ancestral nucleotide for all 97 positions, indicating that it does not possess any derived positions absent in other *Y. pestis* strains. Two previously reported chromosomal differences[3] were not present in our genomic sequence data, suggesting that they likely derived from deaminated cytosines that would have been removed in the current investigation via UDG treatment prior to array capture.

To place our ancient genome in a phylogenetic context, we characterised all 1694 previously identified phylogenetically informative positions[14] (Table S4), and compared those from our ancient organism against aggregate base call data for 17 publically available *Y.pestis* genomes and the ancestral *Y. pseudotuberculosis*. When considered separately, sequences from three of the four victims fall only two substitutions from the root of all extant human pathogenic *Y. pestis* strains (Figure 3a),

and they show a closer relationship to branch 1 *Y. pestis* than to branch 2; however, one of the four victims (individual 6330) was infected with a strain that contained three additional derived positions seen in all other branch 1 genomes[14]. This suggests either the presence of multiple strains in the London 1348 – 1350 pandemic or microevolutionary changes accruing in one strain, which is known to occur in disease outbreaks[15]. Additional support for *Y. pestis* microevolution is indicated by the presence of several variant positions for which sequence data from one individual shows two different nucleotides at comparable frequencies (Table S5). Position 2,896,636, for example, is a known polymorphic position in extant *Y. pestis* populations[14], and this position shows the fixed derived state in one individual (6330) and the polymorphic state in another (individual 8291) at minimum 5-fold coverage (Figure S7). This provides a remarkable example of microevolution captured during an historical pandemic. The remaining variance positions are unchanged in the 18 extant *Yersinia* genomes, thus they may be unique to the ancient organism and are, therefore, of further interest. Additional sampling of ancient genomes will assist in determining the frequency of these mutations in co-circulating *Y. pestis* strains, and will clarify the emergence of branch 2 strains that are as yet unreported in ancient samples.

Consistent tree topologies were produced via several construction methods and all major nodes were supported by posterior probability (pp) values of > 0.96 and bootstrap values >90 (Figure 3b, Figure S8, Figure S9). The trees place the ES sequence close to the ancestral node of all extant human pathogenic *Y. pestis* strains (only 2 of

1694 positions) and at the base of branch 1 (Figure 3b). A secure date for the ES site of 1348 – 1350 allowed us to assign a tip calibration to the ancient sequence and thus date the divergence time of the modern genomes and the ES genome using a Bayesian approach. Temporal estimates indicate that the root of all *Y. pestis* commonly-associated with human infection shared a common ancestor sometime between 660 and 723 years ago (1288 – 1351 AD, 95% highest probability density, HPD), encompassing a much smaller time interval than recently published estimates[14] and further suggesting that all currently circulating branch 1 and branch 2 isolates emerged during the 13th century at the earliest (Figure 3b), potentially stemming from an eastern source as has been previously suggested[14]. This implies that the medieval plague was the main historical event that introduced human populations to the ancestor of all known pathogenic strains of *Y. pestis*. This further questions the aetiology of the 6th through 8th century plague of Justinian, popularly assumed to have resulted from the same pathogen: our temporal estimates imply that the pandemic was either caused by a *Y. pestis* variant that is distinct from all currently circulating strains commonly associated with human infections, or it was another disease altogether.

Although our approach of using an extant *Y. pestis* reference template for bait design precluded our ability to identify genomic regions that may have been present in the ancient organism and were subsequently lost in CO92, genomic comparisons of our ancient sequence against its closest outgroups may yield valuable insights into *Y. pestis* evolution and altered physiology. The Microtus 91001 strain is the closest branch 1 and

branch 2 relative confirmed to be non-pathogenic to humans[16], hence genetic changes may represent contributions to the pathogen's adaptation to a human host. Comparisons against this outgroup revealed 113 changes (Table S6a, and Table S6b), many of which are found in genes affecting virulence-associated functions like biofilm formation (*hmsT*), iron-acquisition (*iucD*) or adaptation to the intracellular environment (*phoP*). Similarly, though its virulence potential in humans has yet to be confirmed to our knowledge, *Y. pestis* B42003004 isolated from a Chinese marmot population[17] has been identified as the strain closest to the ancestral node of all *Y. pestis* most commonly associated with human plague, and thus may provide key information regarding the organism's evolution. Full genome comparison against the ES sequence revealed only 8 single nucleotide differences (Table S6c), six of which result in non-synonymous changes (Table S6d). Although these differences likely do not affect virulence, the influence of gene loss, gene gain, or genetic rearrangements, all of which are well-documented in *Y. pestis*[12,18], is undetermined. In more recent evolutionary terms, single nucleotide differences in several known pathogenicity-associated genes were found between our ancient genome and the CO92 reference sequence (Table S3), which may represent further adaptations to human hosts.

Through enrichment by DNA capture coupled with targeted high throughput DNA sequencing, we have reconstructed a draft genome for what is arguably the most devastating human pathogen in history, and revealed that the medieval plague of the 14th century was likely responsible for its introduction and for its current global

distribution in human populations. This indicates that the disease responsible for the Black Death persists relatively unchanged in the 21st century as both an endemic and emerging biological threat[19]. Introductions of new pathogens to populations are often associated with increased incidence and severity of disease[20] and although the mechanisms governing this process are complex[21], genetic data from ancient infectious diseases will provide invaluable contributions toward our understanding of host-pathogen coevolution. The Black Death is a seminal example of an emerging infection, traveling across Europe and claiming the lives of an estimated 30 million people in only five years, which is much faster than contemporary rates of bubonic or pneumonic plague infection[22] and dissemination[7,8]. Regardless, although no extant *Y. pestis* strain possesses the same genetic profile as our ancient organism, our data suggest that few changes in known virulence-associated genes have accrued in the organism's 660 years of evolution as a human pathogen, further suggesting that its perceived increased virulence in history[23] may not be due to novel fixed point mutations detectable via the analytical approach described here. At our current resolution, we posit that molecular changes in pathogens are but one component of a constellation of factors contributing to changing infectious disease prevalence and severity, where genetics of the host population[24], climate[25], vector dynamics[26], social conditions[27], and synergistic interactions with concurrent diseases[28] should be foremost in discussions of population susceptibility to infectious disease and host-pathogen relationships with reference to *Y. pestis* infections.

4.3 Methods Summary

DNA from dental pulp was extracted and converted into sequencing libraries as previously described[3]. Potential sequencing artefacts resulting from deaminated nucleotides were eliminated by treatment of the DNA extracts with uracil-DNA-glycosylase and endonuclease VIII[29]. DNA extracts were subsequently converted into sequencing libraries and amplified to incorporate unique sequence tags on both ends of the molecule[30]. Two Agilent® DNA capture arrays were designed for capture of the full *Y. pestis* chromosome (4.6Mb), and the pCD1 (70kb) and pMT1 (100kb) plasmids using the modern *Y. pestis* strain CO92 (accession numbers NC_003143, NC_003131, NC_003134, respectively) for bait design with 3 base pair (bp) tiling density. Serial array capture was performed over two copies of each array, using the enriched fraction from the first round of capture as template for a second round. The resulting products were amplified and pooled in equimolar amounts. All templates were sequenced for 76 cycles from both ends on the Illumina GAII® platform, and reads merged into single fragments were included in subsequent analyses only if forward and reverse sequences overlapped by a minimum of 11 bp. Reads were mapped against the CO92 genome using the software BWA, and molecules with the same start and end coordinates were removed with the *rmdup* program in the *samtools* suite[31]. Reference-guided sequence assembly was performed using Velvet version 1.1.03[32], with mapped and unmapped reads supplied in separate channels. Single nucleotide differences were determined at a minimum of 5-fold coverage and base frequency of at least 95% for both a pooled

dataset for all individuals and one in which all individuals were treated separately. A median network was constructed on these base calls using SplitsTree4[33]. Phylogenetic trees were constructed using Parsimony, Neighbhour-Joining (MEGA 4.1), and Bayesian methods[34], and coalescence dates were determined in BEAST using both a strict and a relaxed molecular clock[34] (Fig. S9).

4.4 Acknowledgements: We thank William White (deceased), Jelena Bekvalac, and Rebecca Redfern from the Museum of London Centre for Human Bioarchaeology for access to samples, Martin Kircher and Stephen Forrest for assistance with computational analysis, Gerry Wright for support throughout the project, past and present members of the McMaster Ancient DNA Centre for support throughout the project, and Debi Poinar for constructive comments on earlier versions of the manuscript. We also thank Svante Pääbo and the Max Planck Institute of Evolutionary Anthropology for use of their clean room facilities and molecular biology lab. Funding was provided by the Carl Zeiss Foundation (JK), the Human Genetics department of the Medical faculty in Tübingen (JK), the Canada Research Chair program (HNP, GBG), the Canadian Institute for Health Research (HNP), the Social Science and Humanities Research Council of Canada (HNP), the Michael G, DeGroote Institute for Infectious Disease Research (HNP), and the University at Albany Research Foundation and Center for Social and Demographic Research (SND).

4.5 Author contributions: KIB, SND, SS, and JW performed skeletal sampling. KIB, VJS, and JK carried out laboratory work. HAB, KIB, JK, MM, and HNP designed experiments. KIB, GBG, JK, HNP, VJS, and NW analysed the data. BKC, DAH, DJDE, and JBM provided valuable interpretations. PB provided technical support. KIB, JK, and HNP wrote the paper.

4.6 References Cited

1. Stonkeing, M. and Krause J. Learning about human population history from ancient and modern genomes. *Nat Rev Genet* in press.
2. Haensch, S, *et al*. Distinct Clones of *Yersinia pestis* caused the Black Death. *PLoS Pathog* **6**, e1001134. doi:10.1371/journal.ppat.1001134 (2010).
3. Schuenemann, V.J. *et al*. Fishing for ancient pathogens: *Y. pestis* confirmed in victims of the Black Death via high-throughput sequencing of the pPCP1 plasmid. *Proc Natl Acad Sci USA* in press.
4. Benedictow, O.J. The Black Death 1346-1353: The Complete History. Great Britain: The Boydell Press (2004).
5. Scott, S., Duncan C.J. The Biology of Plagues. Cambridge: Cambridge University Press (2001).
6. Cohn, S.K. The Black Death transformed: disease and culture in early Renaissance Europe. London: Arnoldohn (2002).
7. Achtman, M., *et al*. *Yersinia pestis*, the cause of the plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA* **96**,14043 – 14048 (1999).
8. Cowal, L.., Grainger I., Hawkins D., Mikulski R. 2008. The Black Death Cemetery, East Smithfield. London. Great Britain: Museum of London Archaeological Svc (2008).
9. Briggs A.W., *et al*. Patterns of damage in geniomic DNA sequences from a Neandertal . *Proc Natl Acad Sci USA* **104**, 144616 – 144621 (2007).
10. Hodges, E., *et al*. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols* **4**, 960 – 974 (2009).
11. Green, R.E. *et al*. A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell* **134**, 416-426 (2008).

12. Chain, P.S.G. *et al*. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with Yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA* **11**, 13826 – 13831 (2004).
13. Acthman, M., *et al*. Microevolution and History of the Plague Bacillus, *Yersinia pestis*. *Proc. Natl Acad. Sci. USA* **101**, 17837 – 17842 (2004).
14. Morelli, G., *et al*. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**, 1140 – 1143 (2010).
15. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).
16. Song, Y., *et al*. Complete Genome Sequence of *Yersinia pestis* Strain 91001, an Isolate Avirulent to Humans. *DNA Res* **11**, 179-197 (2004).
17. Eppinger, M., *et al.* Draft Genome Sequences of *Yersinia pestis* Isolates from Natural Foci of Endemic Plague in China. *J. Bacteriol.* 191, 7628 – 7659 (2009).
18. Pouillot, F., Fayolle C, and Carniel E. Characterization of Chromosomal Regions Conserved in *Yersinia pseudotuberculosis* and Lost by *Yersinia pestis*. *Infect. Immun*. **76**, 4592 – 4599 (2008).
19. Stenseth. N.C., *et al*. Plague: Past, Present, and Future. *PLoS Med* **5**, e3 (2008).
20. Baum, J. and Khalina Bar-Gal G. The emergence and coevolution of human pathogens. In Greenblat C and Spigleman M eds. *Emerging Pathogens Archaeology, Ecology, & Evolution of Infectious Diseases* (Oxford University Press, Great Britain) 67-78 (2003).
21. Brown, N.F., *et al*. Crossing the Line: Selection and Evolution of Virulence Traits. *PLoS Pathog*. **2**, e42 (2006).
22. WHO. Interregional meeting on prevention and control of plague. www.who.int/csr/resources/publications/WHO_HSE_EPR_2008_3w.pdf (2008).
23. **Wood, J.W,** Ferrell R.J., and DeWitte-Aviña S.N. The temporal dynamics of the fourteenth-century Black Death: New evidence from ecclesiastical records. *Hum. Biol.* **75**, 427-448 (2003).
24. Joosten M. H. A. J., Cosijnsen T. J., De Wit P. J. G. Host resistance to a fungal tomato pathogen lost by a single base pair change in an avirulence gene. *Nature* **367**, 384 – 386 (1994).
25. Xu, L, *et al*. Nonlinear effect of climate on plague during the third pandemic in China. *Proc Natl Acad Sci USA* doi: 10.1073/pnas.1019486108 (2011).
26. Keeling, M.K. and Gilligan C.A. Metapopulation dynamics of bubonic plague. *Nature* **407**, 903 – 906 (2000).

27. Barrett, R., Kuzawa C.W., McDade T., and Armelagos G.J. Emerging and Re-emerging Infectious Diseases: The Third Epidemiologic Transition. *Annu Rev Anthropol* **27**, 247–-271 (1998).
28. Singer, M. and Clair S. Syndemics and Public Health: Reconceptualizing Disease in Bio-Social Context. *Med Anthropol Q* **17**, 423 – 441 (2003).
29. Briggs, A.W., *et al*. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nuc. Acids Res.***38**, e87 (2009).
30. Kircher, M., Sawyer S., and Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Submitted.
31. Li, H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-60. [PMID: 19451168] (2009).
32. *Zerbino*, D. R., *Birney* E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829 (2008).
33. Huson, D.H. and Bryant D. Application of Phylogentic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23**, 254 – 267 (2006).
34. Drummond, A.J., *et al*. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*. **4**, e88 (2006).

4.7 Figures

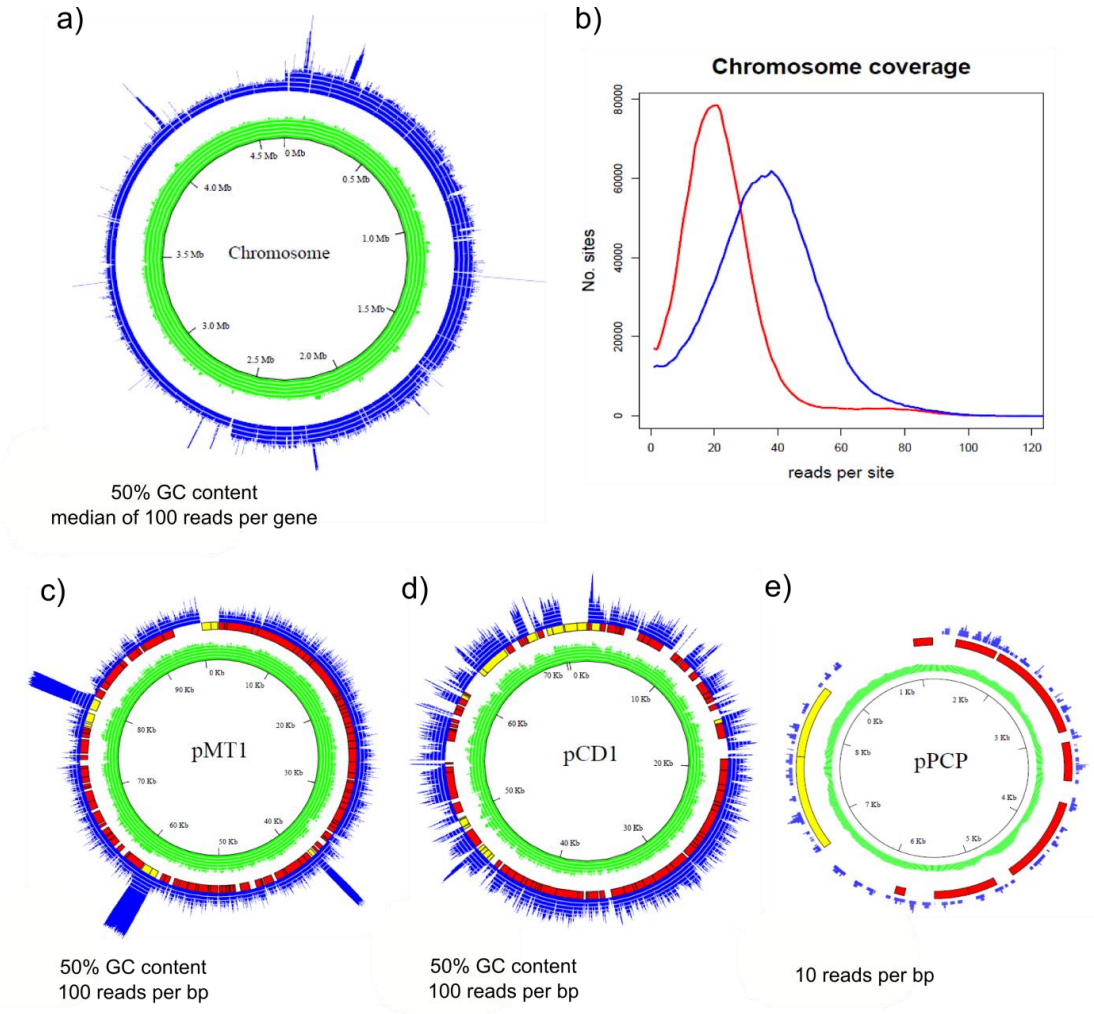


Figure 1 – Coverage plots for a) the chromosome and c), d), and e) the plasmids. Coverage in blue, GC content in green. Scale lines indicate 10, 20, 30, 40, 50 fold coverage and to 0.1, 0.2, 0.3, 0.4, 0.5 GC content. For plasmids, red corresponds to coding regions, yellow to mobile elements. Chromosome shows median gene coverage. Plasmids show each site plotted. Histogram in b) shows chromosomal coverage of array 1 (blue) and array 2 (red) indicating that deeper sequencing increases the number of reads per site, but does not influence overall coverage. Coverage histogram for the plasmids shown in Figure S5.

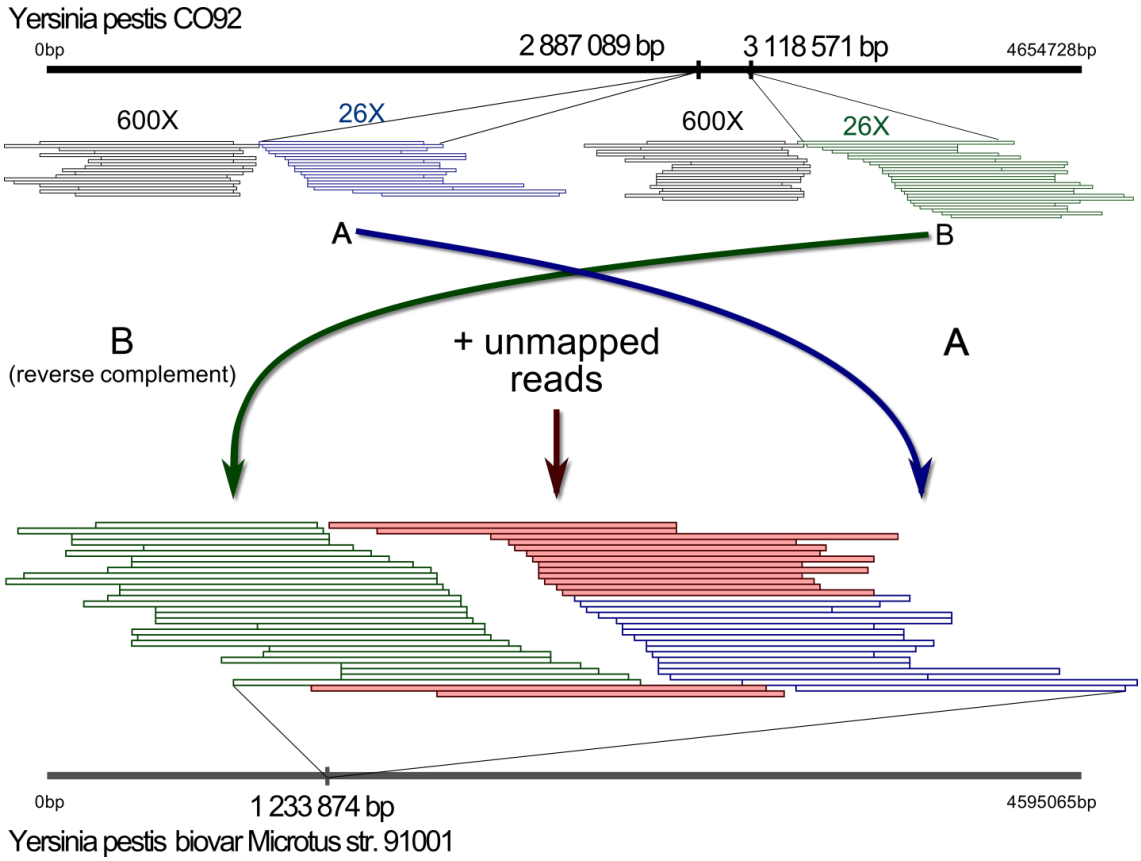


Figure 2 – Alignment of mapped reconstructed contigs against CO92 and Microtus genomes. Reads mapped at positions A (blue) and B (green) are 231 kbp apart in the linearized CO92 genome. Adjacent sequence is high-coverage though only 18x and 20x is shown due to space constraints (black) for A and B, respectively. The structural variant was assembled using reads that did not map to CO92 (red). Its position is shown on the linearized Microtus 91001 chromosome where the 9096 bp contig maps with 100% identity.

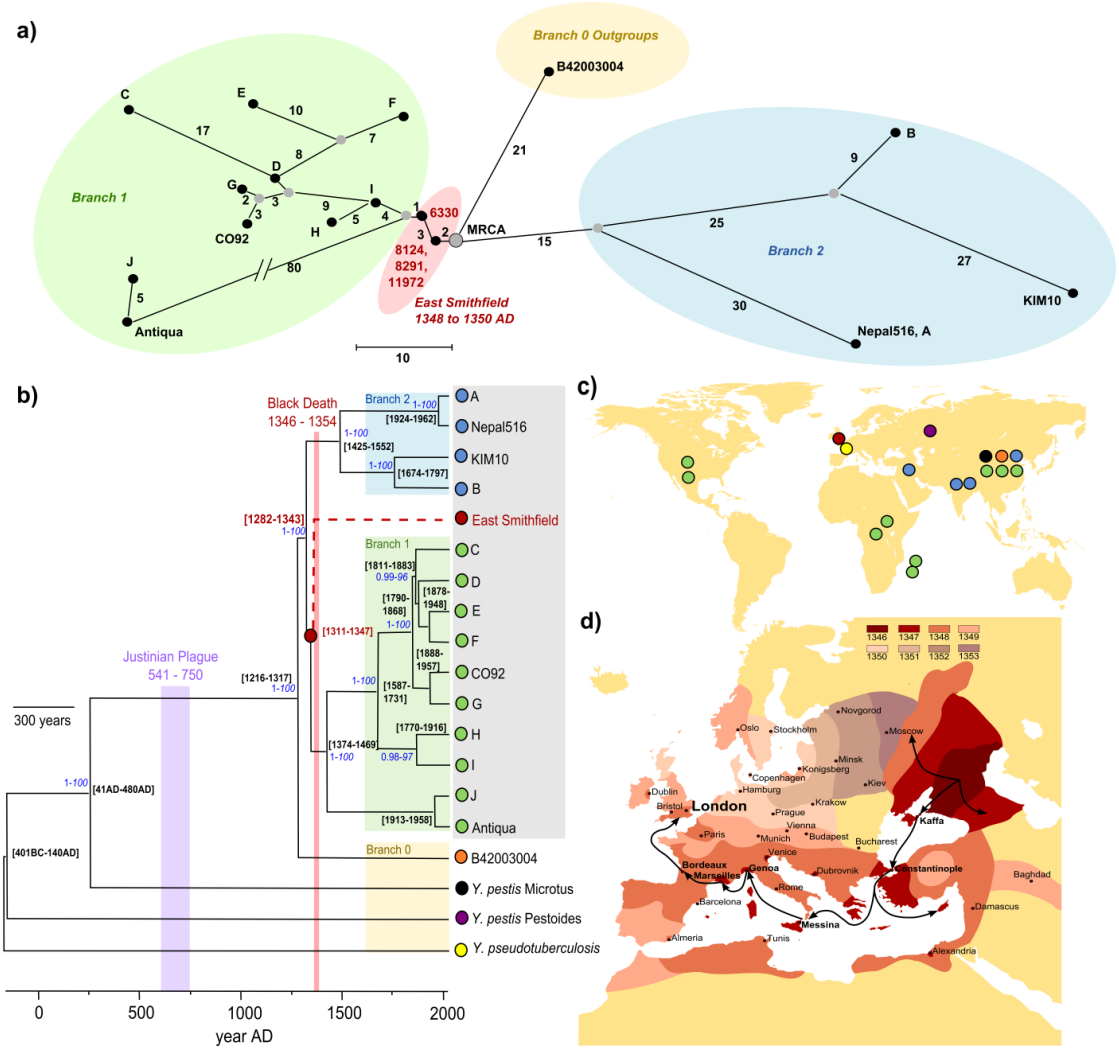


Figure 3 – Phylogenetic placement and historical context for the ES strain. a) Median Network of ancient and modern *Y.pestis* based on 1694 variant positions in modern genomes[14]. Coloured circles represent different clades as defined by Achtman *et al*, 2004[13]. Gray circles represent hypothetical nodes. b) Phylogenetic tree using 1694 variable positions. Divergence time intervals are shown in calendar years, with Neighbour Joining bootstrap support (blue italic) and Bayesian posterior probability (blue). Grey box indicates known human pathogenic strains. c) Geographic origin of genome sequences used in a) and b). d) Geographical spread of the Black Death from infection routes reported in Benedictow, 2004[4]. Abbreviations used in a) and b) for the modern strains: A) NZ ACNQ01000, Nepal516 NC 008149, KIM10 NC 004088, B) NZ AAYT01000, C) NZ ABAT01000, D) NZ ACNS01000, E) NZ AAYS01000, F) NZ AAOS02000, CO92 NC 003143, G) NZ ABCD01000, H) NZ AAYV01000, I) NC 014029, J) NZ AAYR01000, Antiqua NC 008150.

4.8 Supplementary Information (SI) for Bos et al, 2011, "A draft genome of *Yersinia pestis* from victims of the Black Death"

4.8.1 Materials and Methods

4.8.1.1 ***Mortality records***. Figure S1 is based on data collected and published by Cohn, 2002[1] (Fig. 7.33). The City of London did not publish bills of mortality in the 14th century. Cohn counted the numbers of last wills and testaments submitted to the Court of Husting in London each month in 1348 and 1349. While only relatively prosperous individuals would have had last wills and testaments, the temporal pattern of deaths indicated by these monthly counts is likely to be strongly correlated with the full epidemic curve for the population of London as a whole.

4.8.1.2 ***Samples***. Dental material from the four individuals was harvested from the East Smithfield collection at the Museum of London as previously described[2]. The East Smithfield cemetery of London, England (Figure S2) is one of the few excavated medieval catastrophe burials in Europe, and historical documentation clearly indicates purchase of this land in late 1348 or early 1349 specifically for interment of victims of the pestilence[3,4], who had quickly exhausted available plots in local perish cemeteries. Shortly after the Black Death, the Abbey of St. Mary Graces was established on this land in 1350. The plot later housed the nineteenth-century Royal Mint for the city of London,

hence the skeletal collection has been equivocally referred to as the "Royal Mint collection" in published work.

4.8.1.3 ***Extraction and screening***. All manipulations of samples and DNA extracts were performed in facilities specifically dedicated to the extraction of ancient DNA with no prior exposure to sources of *Y. pestis* DNA, and following the most stringent protocols to limit laboratory contamination[5]. Pulverisation of dentin from both roots and molars was performed as previously described[2]. DNA extraction from dental material was performed by a protein kinase mediated lysis, followed by purification via either a modified phenol chloroform method as previously described[2] (McMaster) or a guanidinium-silica method[6] (MPI). A minimum of one negative extraction blank per 8 samples was carried along throughout.

4.8.1.4 ***Library preparation and indexing***. 75µl of each DNA extract, extraction blank control or water library blank control (a minimum of one per every 10 libraries), was used for library preparation following the protocol of Meyer and Kircher (2010)[7] with modifications for ancient DNA[8]. For extracts purified via phenol chloroform, supernatant and pellet fractions were combined, and 75µl of the resultant pool was used in the immortalisation procedure. For all libraries, the blunt-ending step was preceded by treatment with uracil-DNA-glycosylase as well as Endonuclease VIII[9] to

remove deaminated cytosines that contribute to C > T damage motifs in ancient DNA[10]. The chemistry of this procedure was scaled to accommodate a larger volume of extract. Post library preparation, each library was amplified with two 5'-tailed 'indexed' PCR primers, adding sample-specific indexes to both library adapters in this process[8]. Individual PCR reactions contained a maximum of $5x10^8$ library template molecules, and were amplified in 100µl reactions with the following thermal profile: 12 minute initial denaturation at 95°C, 10 cycles consisting of a 30 second denaturation at 95°C, a 30 second annealing at 58°C, and a 45 second elongation at 72°C, ultimately followed by at 10 minute final elongation at 72°C.

4.8.1.5 ***Array design*****.** We used Agilent one million (1M) custom features arrays to capture *Yersinia pestis* genomic DNA libraries. We designed two 1M arrays by tiling 60 bp probes every three bases along the complete 4.6 Mb *Y. pestis* strain CO92 genome (NC_003143), the 70Kb pCD1 (NC_003131) and 96Kb pMT1 (NC_003134) plasmids. We included additional 60 bp probes every 1 base to get extra coverage of 933 single nucleotide polymorphisms (SNPs) that have been used to performed phylogenetic analysis of different *Y. pestis* isolates[11], and eight *Y. pestis* genomic regions that contain the virulence genes *hms* (hemin storage locus, YPO1951- YPO1954), *T3SS* (Type III secretion system, YPO0255 - YPO0273), an enhancing factor (YPO0339), the putative insecticide enhancing factors (YPO3678, YPO3681), and the putative insecticide toxin

(YPO2312, YPO2380, YPO3673, YPO3674). The array design pipeline for mammalian genomes we used removes probes that contain repetitive elements based on 15-mer frequency counts[12]. First a table of frequencies for every 15-mer found in a given genome (sense and antisense) is calculated. Then, for each probe, the genomic frequency values of all of the probe's 15-mers are averaged. If this average genomic frequency of the 15-mers in one probe is above a given threshold, the probe is removed. A threshold of 100 has been used to design arrays based on mammalian genomes[12, 13]. We applied here the same methodology to remove probes that contain repetitive elements. Using 100 as a threshold, no probes were discarded. The highest frequency for a 15-mer in the *Y. pestis* genome was 66. The first array we designed (array A) contained probes covering the first 2.5 Mb of the *Y. pestis* chromosome, the second array (array B) contained the last 2.1 Mb of the chromosome, the plasmid pCD1, the plasmid pMT1, several virulence genes, and all SNPs taken from Morelli et al. 2010[11].

4.8.1.6 ***Array capture***. Capture was performed following methods previously described[17] where all amplification reactions were performed with Phusion Hi-Fidelity DNA polymerase (New England Biosystems) in 100µl reactions with the following chemistry: 50µl Phusion High Fidelity Master Mix, 400 nM each p5 – p7 bridge amplification primer, and 10µl of template. Thermal profile consisted of a 3 minute

initial denaturation at 95°C, between 5 to 10 cycles of a 30 second denaturation at 95°C, 30 second annealing at 60°C, and a 45 second extension time at 72°C, and a final elongation at 72°C for 5 minutes. PCR products were spin column purified over MinElute and quantitated via an Agilent 2100 bioanalyzer following the manufacturer's protocols. The products were captured on two copies of each array A and array B in parallel, four arrays in total. After the first round of capture the 490µl eluate from the two identical arrays were pooled and PCR amplified in 20 independent PCRs in 100µl reactions containing 50µl PhusionTM High-Fidelity Master Mix, 1µM of each p5 – p7 bridge amplification primer and 50 µl library template and a total of 20 cycles. PCR products were spin column purified and quantified via an Agilent 2100 bioanalyzer. A second round of array capture was performed using the same conditions as the first round, only this time each of the two library pools were put on single arrays (array A or array B). The eluted 490µl from each array was subsequently amplified in 10 reactions using the amplification conditions described above.

4.8.1.7 *Sequencing*. The enriched libraries were pooled in equimolar concentrations and sequenced on the *Illumina Genome Analyzer IIx* platform using 2 x 76 + 7 +7 cycles on a single sequencing lane according to the manufacturer's instructions for multiplex sequencing (FC-104-400x v4 sequencing chemistry and PE-203-4001 cluster generation kit v4). The manufacturer's protocol was followed except that an indexed control PhiX

174 library (index 5'- TTGCCGC-3') was spiked into each lane yielding a fraction of 2-3% control reads in all lanes of the run. The sequencing data were analyzed starting from QSEQ sequence files and CIF intensity files from the Illumina Genome Analyzer RTA 1.6 software. The raw reads were aligned to the PhiX 174 reference sequence to obtain a training data set for the base caller Ibis[15]. Raw sequences called by Ibis 1.1.1 were filtered for the 'AATCTTC' index as described[15]. The paired-end reads were subjected to a fusion process (including removal of adapter sequences and adaptor dimers) by requiring at least an 11nt overlap between the two reads. In the overlapping sequence, quality scores were combined and the base with the highest base quality score was called. Only sequences merged in this way were used for further analysis. A small number of larger molecules (longer than 141nt) had lower sequence quality, and were thus discarded. The total number of merged reads for individual samples and the total pool can be found in Table S1b.

4.8.1.8 ***Mapping***. Merged reads were aligned with BWA[16] to the CO92 *Y.pestis* reference genome (NC_003143), the 70Kb pCD1 (NC_003131), the 96Kb pMT1 (NC_003134), and the 9.6Kb pPCP1 (NC_003132.1) CO92 plasmids using default parameters. Mapping alignments were converted to SAM/BAM format[17] with BWA's *samse* command. Accurate mapping into different genomic regions was complicated by the presence of a 1956-bp transposase that is present in one copy in the pCD1 and the

pPCP1, and in multiple copies on the pMT1 and the chromosome. One copy of this transposase was removed from the reference plasmid sequences prior to mapping, though its multiple copies in the chromosome and pMT contributed to inflated coverage estimates for these genetic components prior to additional filtering (Table S1b). The pooled dataset of all enriched extracts generated a total of 20,512,847 merged reads, of which 13,180,582 (64.26%) mapped to the CO92 reference chromosome. Multiple amplifications of the DNA library and enriched fractions performed as part of the capture procedure made it necessary to remove all identical molecules from the merged reads such that each starting library template was considered only once in genome assembly. This poses certain limitations when working with data from a bacterial population, since accuracy in identifying a unique template sequence requires properly distinguishing true genetic variation in the population, which can accrue rapidly even within a single host during pandemic episodes, from artefacts introduced via PCR misincorporations and Illumina sequencing. Duplicate molecules were removed using *rmdup*. This method clusters all reads together that have identical start and end coordinates in the mapping alignment, and uses the read with the highest mapping quality as representative of the cluster. This process, therefore, reduced the error signal resulting from amplification and sequencing, and also reduced the number of reads mapping to the transposase of the chromosome and pMT. The total number of mapping fragments before and after duplicate removal was calculated with *flagstat*[17] (Table S1b). Duplicate removal resulted in a total of 2,812,240 (13.71%) chromosomal reads, and

filtering with Q>30 reduced this number to 2,366,647 (11.53%) high quality chromosomal reads. The average fragment length and length distribution was calculated using the SAM format loaded into *Galaxy*[18, 19]. Using the *pileup* function of *samtools*[17] all BAM files were converted into *pileup* files to calculate average coverage over the genome by dividing the total number of nucleotides mapping to the CO92 genome by its length using a mapping quality of Q>30. Average coverage was estimated at varying fold coverage and with and without mapping quality filtering (Table S1a, S1b, and S1c). To determine if we have sequenced the enriched library to exhaustion we estimated how often we have sequenced each original fragment for array A and array B separately. Average coverage of the target chromosomal region for each array was calculated using a mapping quality filter of Q>30 for the total reads both before and after duplicate removal (*rmdup*). Array A gave an average coverage of 66.09 before and 36.46 after duplicate removal, and array B gave 34.87 before and 22.41 after. Thus for array A, each original fragment was seen on average approximately 1.8 times whereas for array B each fragment was seen on average 1.5 times. This indicates that array A was sequenced to a greater depth than array B. The differences in sequencing depth for the arrays could be potentially explained by differences in enrichment efficiencies or experimental variation. Regardless of different sequencing depths, each array covered a comparable number of sites for the target region at a minimum of 1-fold coverage, indicating that further sequencing would not contribute to additional information on sites not currently covered (Figure 1b, main manuscript).

4.8.1.9 ***Genome architecture and contig assembly***. Using the read dataset for individual 8291 aligned to the CO92 chromosome with BWA[16], all regions with a minimum mapping quality of 25 were extracted. A reference-guided sequence assembly was performed in Velvet version 1.1.03[20] using the extracted reference sequences supplied to the Columbus extension for reference-aided assembly in addition to all mapped and unmapped reads for individual 8291 in separate channels. A hash length of k = 21 was identified as producing an optimal assembly after iterations using several values for k. NCBI BLAST 2.2.21 was used for local similarity searches. The reference-guided assembly produced 130,556 contigs with an N50 of 288 bp. When filtered for contigs 500bp or larger 2134 contigs remained totalling 4,013,009 bp. This consisted of 30,959 unmapped reads. To identify contigs that potentially contained novel architecture distinct from CO92 and stemming from unmapped reads, all contigs were queried against the CO92 chromosome using BLAST[21] with an E-value cutoff of 1e-50. Twenty-nine were contigs that had no hits to CO92: six had 100% identity to human mtDNA, 8 matched mobile elements that were common to several species of bacteria, and 16 had no or low significant similarity to sequences in the NCBI database. A potential variant was considered for further examination if the contig sequence produced multiple hits to the CO92 chromosome. Any such contig would have segments that occur at a distance in the CO92 chromosome that the BLAST algorithm would be unable to extend through. Ideally these segments would be non-overlapping and would occur at a distance of more than one read length (54bp). Each contig was queried against other NCBI *Yersinia*

sequences to check if the sequence proposed by the contig was observed in other species in the genus. Of the 20 such contigs identified (Table S2), half were either eliminated as mapping errors or were removed due to their location within known repetitive regions. The remaining 10 contigs provided indications of altered structure in the ancient organism as compared to the reference CO92 genome. The contig numbers, length, description of the change relative to CO92, and the number of *Yersinia* strains in which each structural variant was observed are recorded in Tables S2a and S2b. An example is provided in Figure 2 in the main manuscript using contig number 6149, which suggests a breakpoint joining sequence in the *pbpC* pseudogene, which is intact in ancestral *Yersinia* strains, but spans a gap of 231 kb in the CO92 genome. The mapping alignment shows both joined ends occurring next to regions of high coverage (insertion element IS285), though a read depth of 20 is depicted in Figure 2 owing to space restrictions and resolution. Only short overlaps of several bases are observed across the breakpoint in the mapping alignment and no read is aligned across its centre at that position. Unmapped reads have spanned this breakpoint producing a natural overlap joining these two ends in reference-guided assembly. This specific orientation is shared only with *Y. pseudotuberculosis*, Microtus, Pestoides, and B42003004, which are ancestral to all currently circulating *Y. pestis* strains commonly associated with human infection. Separate orientations exist for this region in branch 1 (*pestis* Z176003, *pestis* D182038, *pestis* D106004, CO92, Antiqua) and branch 2 (Medievalis str. Harbin 35, Nepal516, KIM) strains, indicating that rearrangements in the *pbpC* pseudogene

occurred as independent events on different lineages. This ancestral character supports the authenticity of our sequence data because these contigs demonstrate marked differences from modern *Y. pestis*. Each proposed breakpoint position was examined in the 8291 mapping alignment as well as a mapping alignment for each contig to determine if 1) the reconstruction was supported by multiple reads, 2) it did not occur in a repetitive region of CO92, and 3) was not produced by misassembly of a collapsed repeat. Additionally, the mapping alignment needed to indicate that the breakpoints occur in relatively poorly mapped regions of CO92 such as that described for contig 6149 above. This process revealed that 10 of the 20 contigs resulted from improper assembly (Table S2b).

4.8.1.10 ***Genomic analysis***. To identify positions that differ from the CO92 *Y.pestis* reference genome, pileup files generated from BAM files were filtered for positions with a mapping score of Q>30, a major base frequency of at least 95% and at least a 5 fold coverage using *Galaxy*[18, 19]. Using these conditions 97 nucleotide positions were found to be different between the CO92 reference genome and the East Smithfield individuals 11972, 8291, 8134, and 94 positions between 6330 and CO92 (Table S3a). All 97 (or 94) nucleotides were found to be in the ancestral state in the East Smithfield individuals when compared to *Y. pseudotuberculosis*, thus not a single unique derived position was found in the ancient bacterial genome. Furthermore, all of these positions are known variants in extant *Y. pestis* strains[11]. The same strategy was used for reference

comparison against the three plasmids. The average coverage of the plasmids was found to be similar for both the pMT1 and the pCD1. The PCP1 plasmid has an estimated copy number of 186 within *Y.pestis*[22] and was not put on the designed capture arrays since it was sequenced at high coverage in a previous work[2]. The estimated low coverage of less than one fold was, therefore, expected. For the pMT1 plasmid 2 differences to CO92 were found, and 4 differences for the pCD1. As with the chromosomal positions, all differences from the reference sequences in the plasmids were ancestral in the East Smithfield individuals. Differences from the reference sequence are recorded in Table S3a and S3b.

Variant positions were identified by comparing base call frequency data from all high-quality reads for each position with individuals treated separately. A position was considered polymorphic if at least two different nucleotides were recorded in reads with a minimum of 5-fold coverage, permitting 0.9675 confidence. All polymorphic positions are shown in Table S5.

Heat plots were generated for the CO92 genome and for the plasmid genomes to show the relationship between read depth and GC content. In each case the number of sites that have a given number of reads per site and a given GC content (averaged over the 30 neighbouring bases) is colour encoded. The heat plots (Figure S6) show that most sites are covered by a comparatively small number of reads and have intermediate

GC content. There is a small skew toward less coverage when the GC content is low. This skew becomes more apparent in the plasmids.

4.8.1.11 ***Phylogenetic analysis***. We used 1694 of the 1748 positions that were previously identified as variant in modern *Y. pestis* genomes[11], since a multiple alignment of 17 modern genomes as well as the outgroup *Y. pseudotuberculosis* using MAUVE[23] revealed only a subset of them to be variant in extant *Y. pestis*. The genome sequence of the Angola strain (NC_010159) was excluded from phylogenetic analysis due to its high divergence compared to the *Y. pseudotuberculosis* outgroup[11]. Strain NZ_AAUB01000 was excluded from further analysis due to poor sequence quality. The 27 previously unreported variant positions discovered in our analysis that are derived in the CO92 reference sequence were not included in the 1694 variant position since they could not be confirmed via HPLC[11] and could potentially represent sequencing artefacts in the publically available reference sequence (NC_003143). All 1694 positions were called from the four East Smithfield individuals (Table S4) using *pileup* files and filtered for a minimum coverage of 5-fold and mapping quality of Q>30 using Galaxy[18, 19]. A 95% majority base filter was not used in this step. Using the software SplitsTree4[24] median networks were calculated for all 1694 positions from the 18 modern genomes as well as a subset excluding the Branch 0 genomes[11], and compared to all four East Smithfield individuals. We found that three of the four East Smithfield individuals clustered

together (8124, 8291, 11972) whereas one individual (6330) fell closer to other Branch 1 strains (Figure 3, main manuscript). Phylogenetic trees using the Maximum Parsimony (MP) and the Neighbour Joining (NJ) algorithm implemented in MEGA 4.1[25] confirmed a distinct position of individual 6330 compared to the others (Figure S8a and S8b). Thus individual 6330 was excluded for further analysis and individuals 8124, 8291, and 11972 were pooled to get a higher number of informative sites (Table S4, pool 3). The resulting MP and NJ trees are well resolved based on 1000 bootstrap replicates with bootstrap values >90 (Figure S8c and S8d).

4.8.1.12 ***Divergence times***. To estimate the divergence times of all *Y. pestis* strains the Bayesian algorithm implemented in BEAST v1.5.3 was used. The pool (8124, 8291, 11972) consensus was analyzed using a relaxed uncorrelated log-normal molecular clock as well as the strict clock option[26]. As a tip calibration point we set the prior distribution age of the East Smithfield sample to 1348 - 1350 AD as well as the isolation years for the modern strains as described elsewhere[11]. For each analysis, we used a model that assumes a constant population size across the phylogeny and ran 50,000,000 generations of the Markov Chain Monte Carlo with the first 5,000,000 generations discarded as burn-in. We chose the HKY sequence evolution model. We performed two BEAST runs and merged both using Logcombiner[26]. The estimated mutation rate was 1.96E-08 (HPD 95% lower 1.7E-08, HPD 95% higher 2.2E-08) mutations per site and per

year over the whole genome, using a strict molecular clock[26] . This estimate is similar to previously determined mutation rates for *Y. pestis*[11]. A consensus tree of all 90,000 trees was inferred using TreeAnnotater V.1.4.8. As observed for the NJ and MP tree all nodes were found to be well resolved in the Bayesian tree with posterior probabilities of >0.96 (Figure S9). The estimated coalescence times for the strict and relaxed molecular clock are shown in Figure S9.

4.8.1.13 ***Comparisons with related genomes***. Descriptions of genomic regions for the nucleotide differences were determined using the current annotation of C092 and its plasmids. The location of the difference is given in the *mpileup* file and this was determined to be either synonymous or non-synonymous given the estimated mapping position and the location within the annotation.

A number of differences in coding regions for proteins involved in various aspects of pathogen virulence were observed between the ES strain and both *Y. pestis* Microtus and *Y. pestis* CO92 (Tables S3 and S6d). Although the specific role of these differences in virulence has not been assessed, the genes in which they appear are well known virulence factors.

4.8.2 Supplementary figures

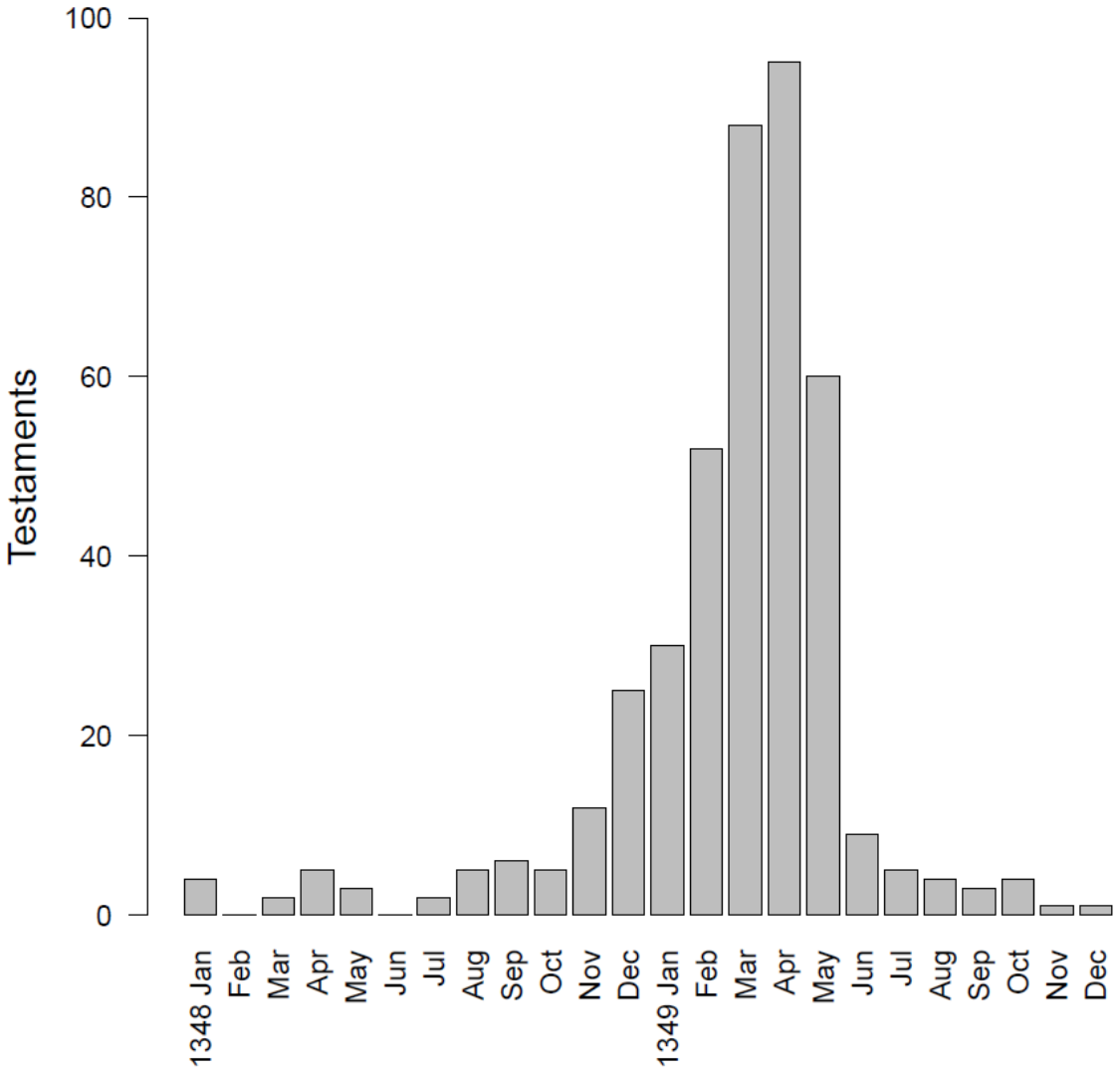


Figure S1 – Monthly numbers of Last Wills and Testaments submitted to the Court of Husting in London in 1348 and 1349, as counted by Cohn, 2002[1]. Deaths reported in this way would have represented a small sample of all deaths in London, but the structure of the epidemic curve is clear nevertheless.

Figure S2 – John Norden's map of London in 1593. East Smithfield is indicated on the far right of the map in red, just north of the river.

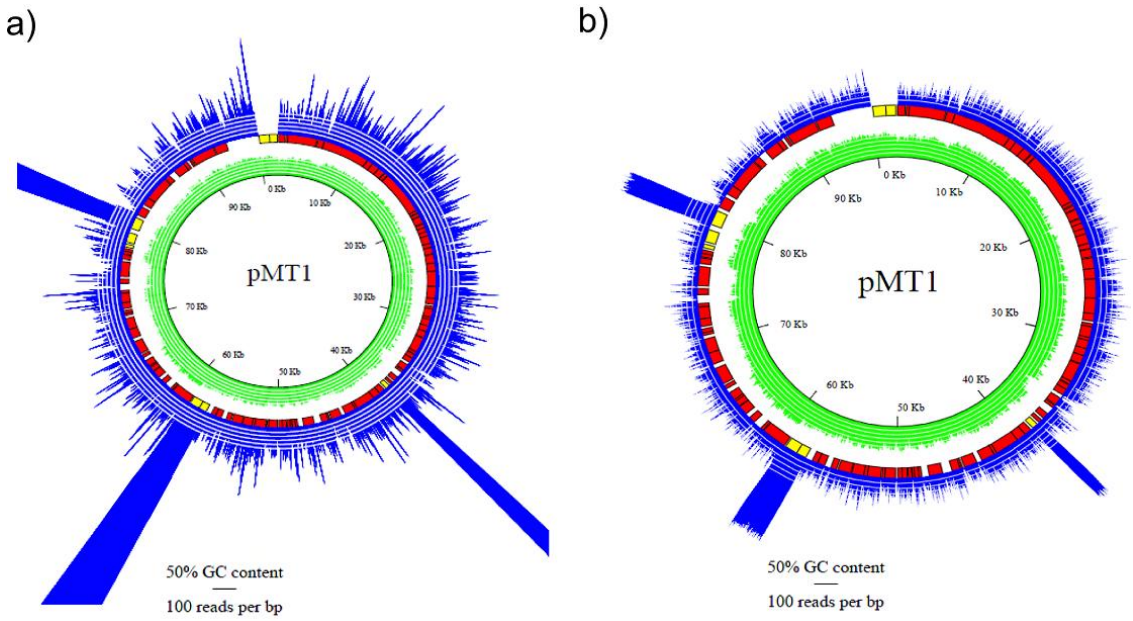


Figure S3 – Effect of duplicate removal via *rmdup*. The coverage plot in a) shows the mapping of all fragments against the pMT1 reference (NC_003134) prior to duplicate removal and before filtering for mapping quality. Coverage is shown in blue, GC frequency for the chromosome in green. Scale lines correspond to 10, 20, 30, 40, 50 fold coverage and to 0.1, 0.2, 0.3, 0.4, 0.5 GC content. Red corresponds to coding regions, yellow to mobile elements. A large tranposase corresponding to positions 1 – 1962 was removed, though three additional copies of this transposase in the pMT have contributed to three areas of dense coverage. The coverage plot in b) shows the effect of duplicate removal, which greatly reduces the number of fragments mapping to the pMT reference, most notable in coverage density in the three remaining transposases.

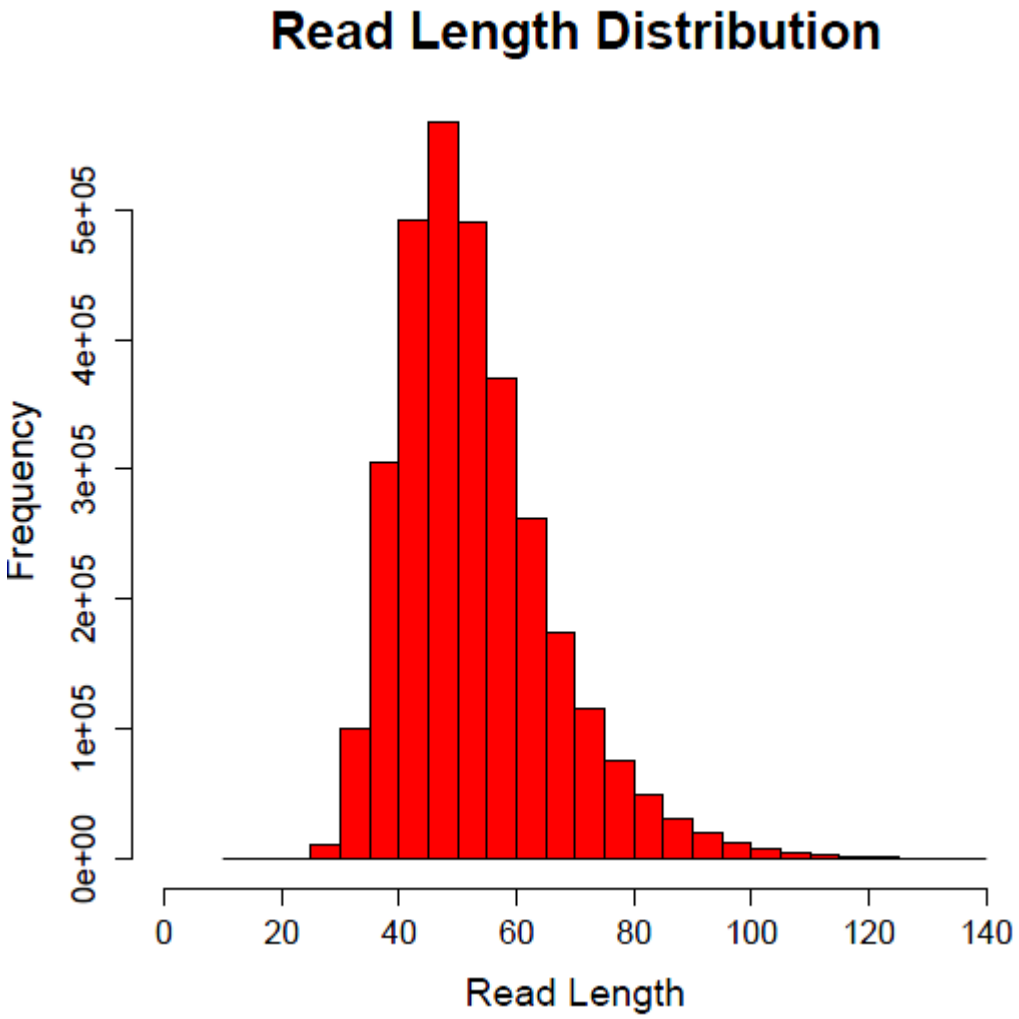


Figure S4 – Fragment size distribution of reads post duplicate removal (*rmdup*) and mapping filtering (Q>30) from a pooled collection of all libraries.

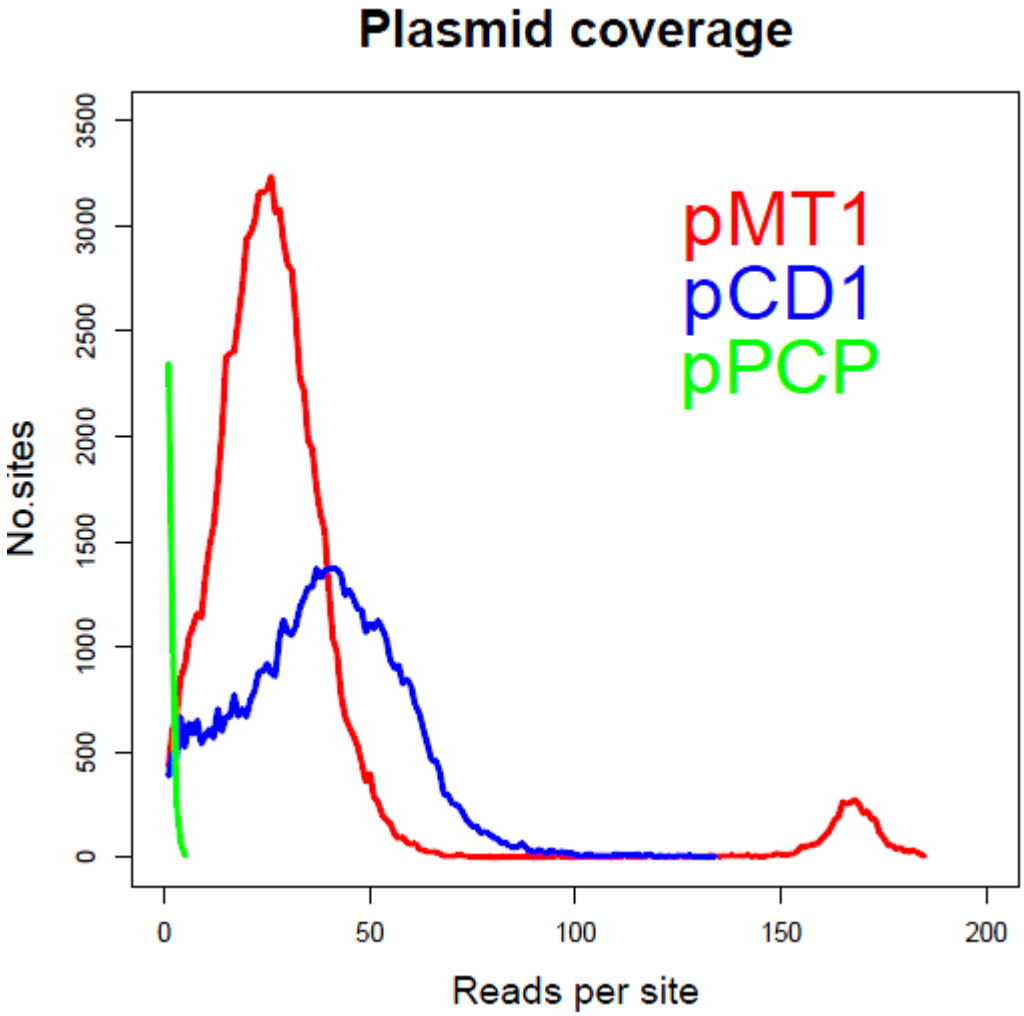


Figure S5 – Coverage plot for East Smithfield *Y. pestis* plasmids

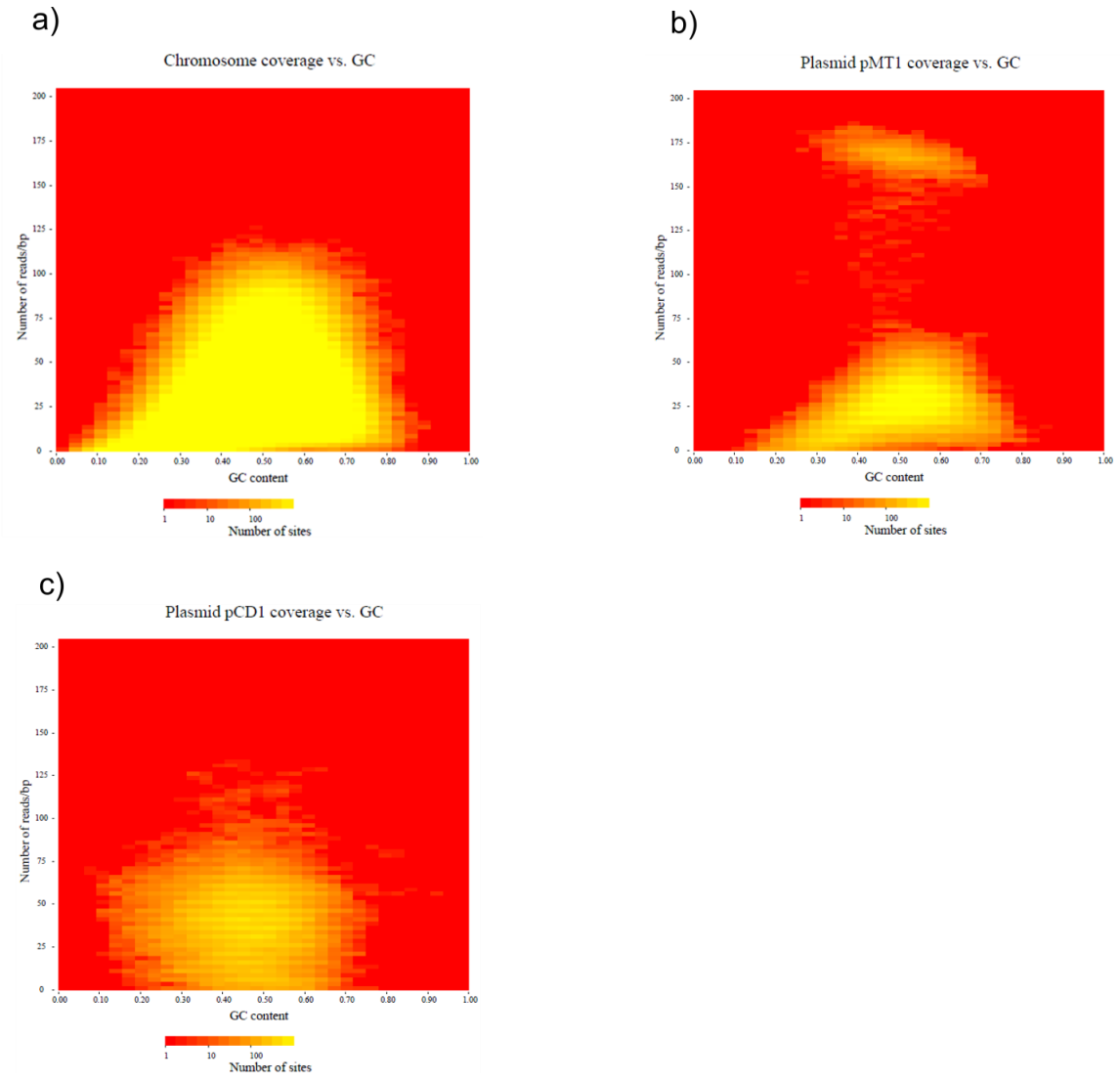


Figure S6 – Heat plots showing relationship of GC content on coverage for a) the chromosome, b) the pMT plasmid, and c) the pCD1 plasmid. The number of sites is shown as a function of the number of reads per site and the GC content averaged over the 30 neighbouring bases. Most sites are covered by fewer than 30 reads and have intermediate GC content. There is a small skew toward less coverage when the GC content is low.

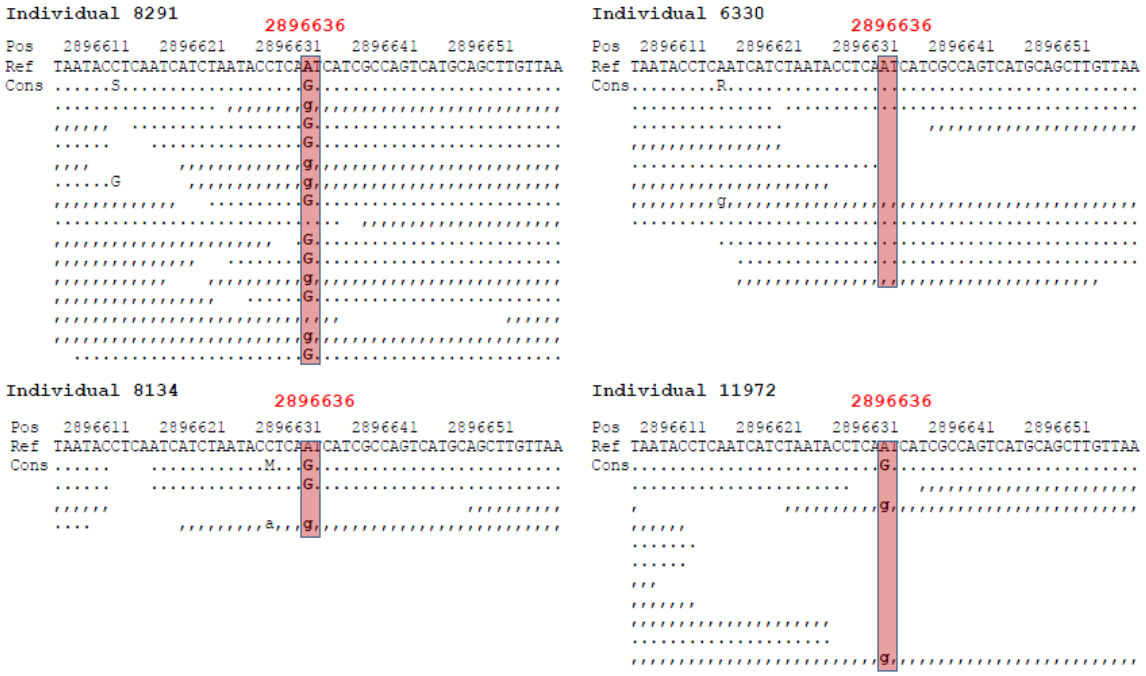


Figure S7 – Alignment showing base frequencies for position 2896636 in all four individuals. The position is polymorphic in individual 8291 and fixed derived in 6330. Individuals 11972 and 8124 show the ancestral state, though coverage is too low to ascertain whether the position is polymorphic with 0.9675 confidence.

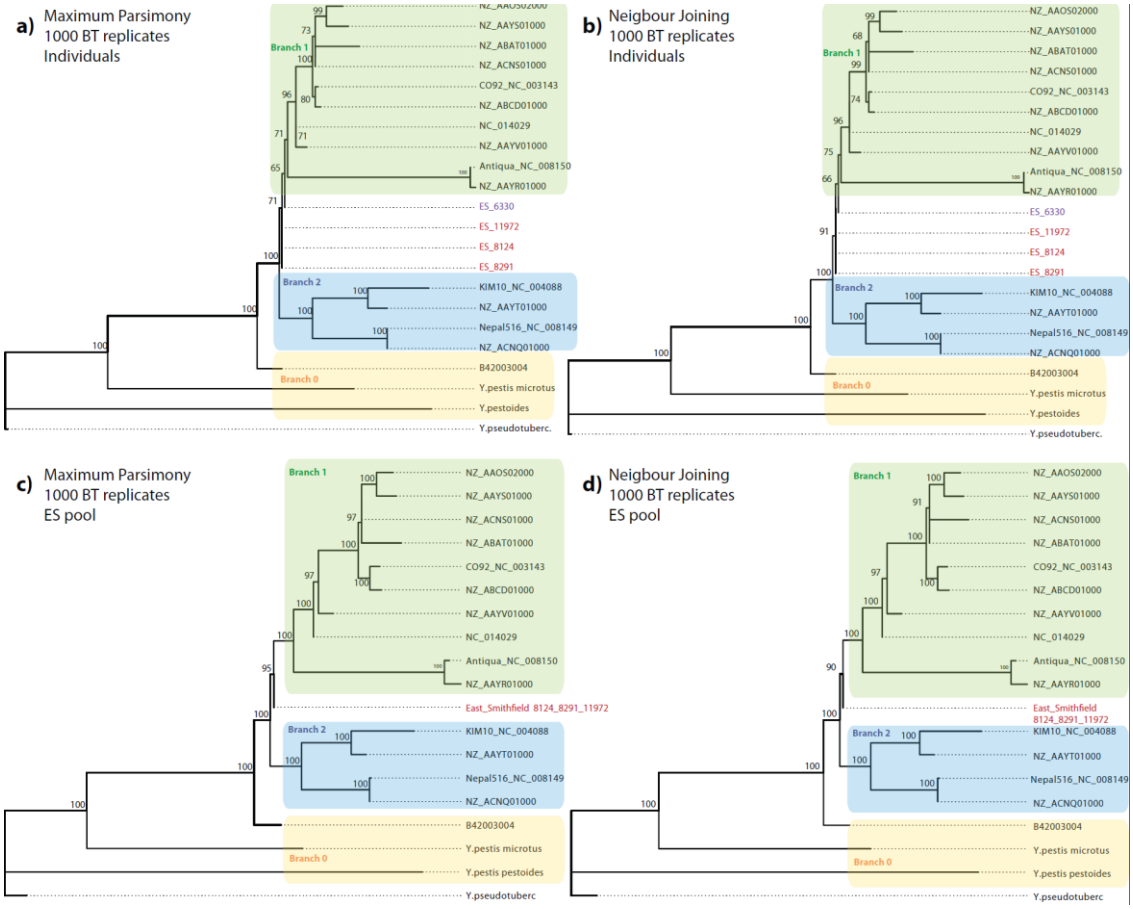


Figure S8 – Phylogenetic trees based on 1694 positions known to be variable within modern *Y. pestis* strains, reconstructed using Maximum Parsimony (a and c) and Neighbour Joining (b and d) algorithms implemented in MEGA 4.1[25]. For a) and b) the East Smithfield individuals are kept separate; for c) and d) individuals 8124, 8291, and 11972 are pooled together.

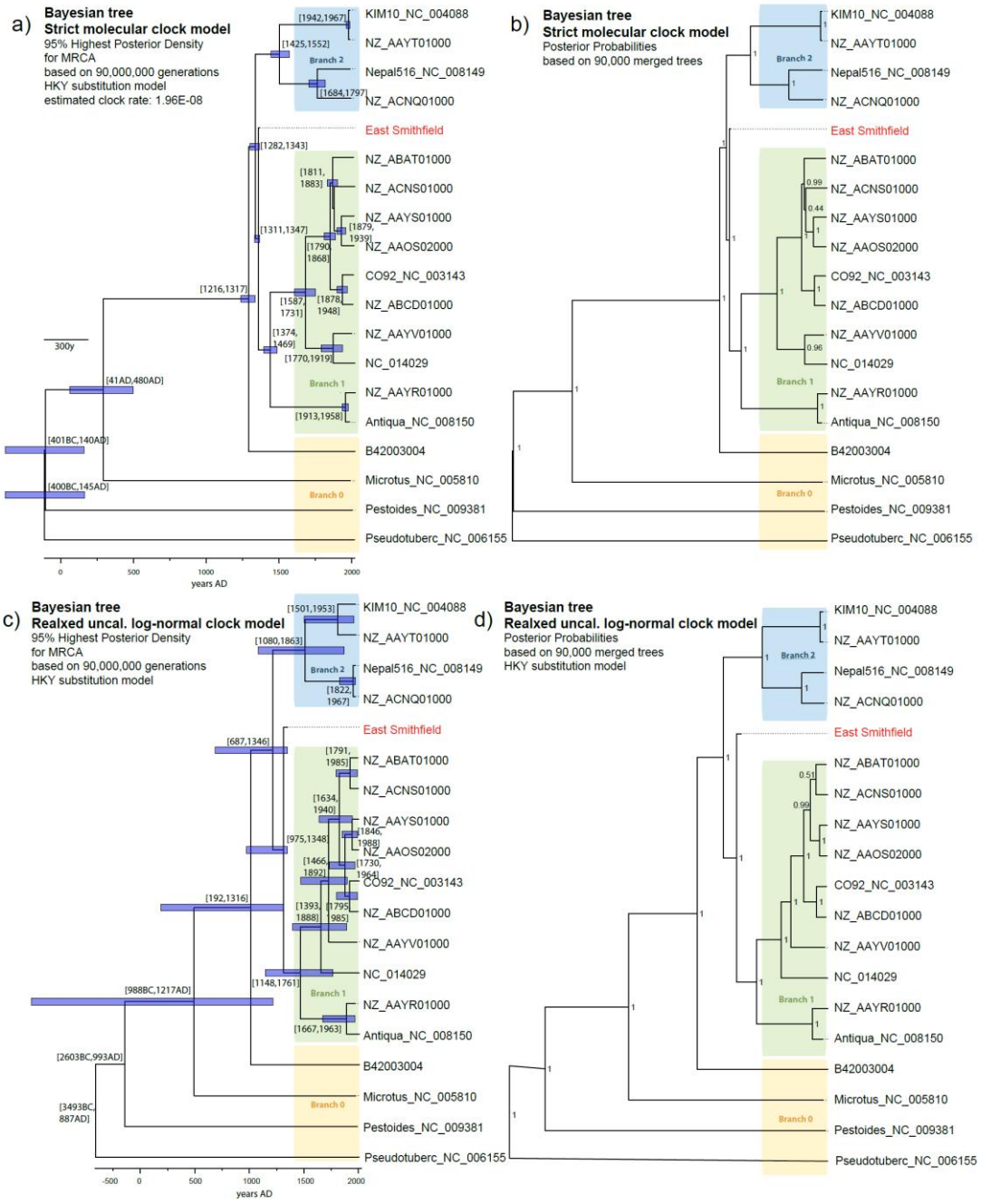


Figure S9 – Phylogenetic trees reconstructed using a Bayesian approach with BEAST and two times 50,000,000 generations. Each tree is derived from a consensus of 90,000 merged trees[26]. Trees a) and b) used a strict molecular clock; trees c) and d) used a relaxed uncorrelated log-normal clock. Trees a) and c) show divergence dates on each node; b) and d) show posterior probabilities for each node. Blue bars show 95% Highest Posterior Density for the divergence date of the different *Y.pestis* strains.

4.8.3 Supplementary tables

Table S1 – Estimates of East Smithfield sequence coverage.

Table S1a – Fold coverage estimates showing effect of duplicate removal (*rmdup*) and mapping filtering.

fold chromosomal coverage	post duplicate removal	post duplicate removal and mapping filtering Q>30
1 fold	99.25%	93.48%
2 fold	98.62%	92.85%
3 fold	97.98%	92.22%
4 fold	97.29%	91.54%
5 fold	96.55%	90.80%
10 fold	91.50%	85.76%
30 fold	48.30%	42.67%

Table S1b – description of all mapped reads

See Appendix 3

S1c – Fold coverage estimates presented as average coverage per site.

individual	Chromosome	pMT1	pCD1	pcp1
ES_11972	6.6	10.1	7.9	0.3
ES_6330	3.5	7.6	5.4	0.1
ES_8124	2.2	7.5	2.7	0
ES_8291	14.8	19.9	19.4	0.2
Pool	28.2	31.2	35.2	0.7
Controls	0	0	0	0

Table S2 – Description of reference-guided contigs showing architectural differences between the East Smithfield genome and the CO92 reference.

Table S2a Reference-guided contigs that differ from CO92

Contig	Length (bp)	Rearrangement	Found in *Y. pestis*	Found in *Y.pseudotuberculosis*
1530	1149	20 bp indel	7	3
208	2062	insertion of IS100 in CO92	10	4
3036	1540	30kb rearrangement	7	4
3293	2519	9 bp indel	7	0
4326	963	insertion of IS100 in CO92	10	0
4746	4923	32 bp indel	9	1
6149*	9096	231 kbp rearragement	3	4
6289	659	insertion of IS285 in CO92	10	4
7016	10311	insertion of IS100 in CO92	9	4
7127	2526	insertion of IS100 in CO92	5	3

*Intact contig found in Pestoides, Microtus, Angola, B42003004 and *Y. pseudotuberculosis*

Table S2b Incorrectly reconstructed reference -guided contigs that differ from CO92

Contig	Length (bp)	Description	Found in *Y. pestis*	Found in *Y.pseudotuberculosis*
2935	1477	in repetitive region in CO92	0	0
3181	1704	near phage protein in CO92	0	0
5765	3992	near repeated tRNA val in CO92	0	0
6107	4825	misassembly indel	0	0
6798	603	near pseudogene in C092	0	0
7052	5267	misassembly indel	0	0
7405	8587	misassembly indel	0	0
7488	4398	misassembly indel	0	0
8057	8059	misassembly indel	0	0
43500	6912	misassembly indel	0	0

Table S3 – nucleotide differences identified between the ancient genome and the CO92 reference genome used for bait design for a) the chromosome, and b) the pCD1 and pMT1 plasmids.

See Appendices 4a and 4b

Table S4 – Known polymorphic positions[11] used for phylogenetic reconstruction.

See Appendix 5

Table S5 – Variant positions recorded in the East Smithfield sequences.

See Appendix 6

Table S6 – Substitutions of note in the two outgroups most closely related to the ES sequence. a) Differences between Microtus 91001 and the other 18 available *Yersinia* genomes. b) Differences between Microtus 91001 and ES with genetic descriptions. c) Differences between B42003004 and the other 18 available *Yersina* genomes. d) Differences between B42003004 and ES with genetic descriptions.

See Appendices 7a, 7b, 7c, and 7d

4.8.4 References Cited:

1. Cohn, S.K. The Black Death transformed: disease and culture in early Renaissance Europe. London: Arnoldohn (2002).
2. Schuenemann, V.J. *et al*. Fishing for ancient pathogens: *Y. pestis* confirmed in victims of the Black Death via high-throughput sequencing of the pPCP1 plasmid. Submitted.
3. Cowal, L.., *et al*. The Black Death Cemetery, East Smithfield. London. Great Britain: Museum of London Archaeological Svc (2008).
4. Hawkins, D. The Black Death and new London cemeteries of 1348. *Antiquity* **64**, 637 – 642 (1990).
5. Pääbo S., *et al*. Genetic analyses from ancient DNA. *Annu Rev Genet* **38**,645-679 (2004).
6. Rohland, N., and Hofreiter M. Ancient DNA extraction from Bones and Teeth. *Nat Protoc* **2**, 1756 – 1762 (2007).
7. Meyer, M. and Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc* **6**; doi:10.1101/pdb.prot5448 (2010).
8. Kircher, M., Sawyer S., and Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Submitted.
9. Briggs, A.W., *et al*. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nuc. Acids Res.***38**, e87 (2009).
10. Hofreiter, M *et al*. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nuc Acids Res* **29**, 4793-4799 (2001).
11. Morelli, G., *et al*. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**, 1140 – 1143 (2010).
12. Hodges, E., *et al*. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* **39**, 1522 – 1527 (2007).
13. Burbano H.A., *et al*. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**, 723 – 725 (2010).
14. Hodges, E., *et al*. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protoc* **4**, 960 – 974 (2009).
15. Kircher, M., Stenzel U, and Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**: R83 (2009).
16. Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**,1754-60 (2009).
17. Li, R., *et al*. SOAP2: and improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966 – 1967 (2009).

18. Goecks, J., *et al*. Galaxy: a comprehensive approack for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86+ (2010).
19. Blankenberg, D., *et al*. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*. Chapter **19**, 1-21 (2010).
20. *Zerbino*, D. *R., Birney* E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821–829 (2008).
21. Altschul, S.F., *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
22. Parkhill J., *et al*. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523 – 527 (2001).
23. Darling A.C., *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).
24. Huson, D.H. and Bryant D. Application of Phylogentic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23**, 254 – 267 (2006).
25. Tamura, K., *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596 – 1599 (2007).
26. Drummond, A.J., *et al*. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*. **4**, e88 (2006).

CHAPTER 5: CONCLUSION

This dissertation has presented the culmination of seven years of work that demonstrates the application of molecular methods for evaluating ancient pathogenic agents. The first two publications provide the necessary methods for retrieval and authentication of ancient pathogen molecules from skeletal remains, and provide the necessary foundation for the final publication, which provides the first ancient pathogen genome, a remarkable methodological achievement that significantly advances the field of pathogen research. Our genome is at present only a draft, owing to limitations in our capture design by using the modern CO92 reference genome as a template for array probes. This has prevented us from determining whether or not ancestral genetic segments such as those present in *Y. pseudotuberculosis*, Microtus 91001, and B42003004 are indeed present in the strains associated with the Black Death. In addition, the use of the CO92 template for reference guided assembly introduced yet another bias: although some contigs from mapped and umapped reads showed evidence of altered genomic rearrangements, full evaluations of genome architecture are not yet possible. True de novo assembly will hopefully soon become a reality as computational power and output continues to grow.

The effect of the single nucleotide differences we recorded on virulence has yet to be tested under controlled laboratory conditions, though at our current resolution, we have not identified the proverbial “smoking gun” for a genetic basis to explain the

differences between ancient and modern forms of *Y. pestis* infections. As is often the case for scientific investigations, our results have generated more questions than answers. What contributed to the high levels of mortality in the 14th century, and how can we avoid a resurgence of such a cataclysmic event in future? Climate and living conditions could have presented an environment conducive to large rodent populations and flea survival, and though this might have supported local surges of *Y. pestis* infections as have been demonstrated with modern populations (Enscore *et al*, 2002; Parmenter *et al*, 1999), it does not seem sufficient to explain dissemination of the disease throughout nearly all of Western Europe in five short years via travel in a rat-flea vector model. To explain the rapid transmission of plague, several scholars have presented the argument that the disease was manifest in the pneumonic form in the densely populated major centres, which allowed for person to person transmission and dissemination through the sparser rural areas en route to adjacent towns and cities (Gottfried, 1983; Benedictow, 2004). Pneumonic episodes are extremely rare in modern plague outbreaks (Gamsa, 2006), and this does little to address the omnipresent mention of bubo-type swellings in historical descriptions of the disease symptoms (Horrox, 1994). Some medical historians have suggested that a change in vector from the rat flea to the human louse may have been responsible for increased transmission of the disease in the bubonic form (reviewed in Cohn, 2008). Laboratory evidence has demonstrated transmissibility of *Y. pestis* via the human louse in a rabbit model (Ayyadurai *et al*, 2011), though data for efficient transmission via this vector has

yet to be demonstrated in modern populations. However, these proposed theories involving biological bases for enhanced transmission speak little to the reasons for variation in the reported symptoms of the disease in different communities, and why certain towns and villages escaped the pandemic entirely (Cohn, 2002; Gottfried, 1983).

Since it is known that host genetics can play a decisive role in susceptibility and resistance to various infectious insults (Anderson and May, 1983; Black 1992; Gagneaux *et al*, 2005), it is perhaps fitting to turn our attention to the affected population as opposed to the vector, especially considering recent investigations that have sought to characterise the precise relationship between the bacterium and human proteins (Yang *et al*, 2011). Infectious disease has been considered a driving force in the evolution of host populations (Haldane, 1949), and such changes would be apparent as genetic signatures demonstrating recent positive selection (Woolhouse *et al*, 2002). Although experimental evidence from non-human models has provisionally confirmed this theoretical concept (Navas *et al*, 2007), translation of this model to human populations is not straight forward owing to the myriad of social determinants on population health and susceptibility to communicable disease (reviewed in Farmer, 2006). This point is acutely relevant to a disease of the medieval era, where poor sanitary conditions, close quotidian association with domestic animals, and malnutrition resulting from frequent famines and pronounced social inequities would make the general population at high risk of infectious insult.

Current trends in infectious disease research are moving away from an orthodox model of “single host, single pathogen, single disease”, which has dominated medical thinking for over a century. A more comprehensive view of disease ecology is surfacing that acknowledges the biological interactions of several pathogens and underlying health status in a system to influence the speed of transmission, infectivity, and severity of disease manifestation. This “epidemiological synergism” (Singer, 2010) posits that interactions between seemingly unrelated pathogens might be a driving force in pathogen evolution and adaptation. The notion that interactions between organisms can result in either “facilitation” or “inhibition” of metabolic systems is a well-established concept in evolutionary theory (Woolhouse *et al*, 2002; Stosor and Wolinsky, 2001; Cheng *et al*, 2009), though the unique influence of ubiquitous social and economic inequalities on the transmission of infectious disease in human populations necessitates a more detailed explanation for why epidemic disease has always been part of the human experience, and why certain groups appear to be more susceptible than others (Herring and Sattenspiel, 2007).

The term “syndemic” has recently emerged in the social sciences literature as “a set of enmeshed and mutually enhancing health problems that, working together in a context of deleterious social and physical conditions that increase vulnerability, significantly affect the overall disease status of a population" (Singer, 2010 p.15). Inherent in this definition is the notion that more vulnerable members of a community tend to be more susceptible to infectious disease. For populations faced with food

insecurity, social and economic marginalisation, and dense living quarters under poor hygienic conditions, persistent epidemics can define the baseline health status. Since such conditions characterise urban centres and many rural communities of medieval Europe, it is conceivable that interactions between co-circulating parasites may have been partially responsible for the heterogeneity of recoded disease symptoms and mortality profiles of the Black Death. Genetic investigations into other endemic infections of medieval Europe may, therefore, bring us one step closer to understanding the biosocial context for the medieval pandemic.

The observation that symbiotic relationships can develop between pathogens to the extent that a single transmission can result in the dissemination of multiple parasites (Rowbotham, 1980) opens the possibility that interactions between pathogens may have been responsible for the rapid spread of the disease through the European population. Facilitation during a massive pandemic such as the Black Death would certainly exert genetic influence on the main causative agent, and the demonstration that the *Y. pestis* pMT1 is derived from an exclusively human pathogen (Prentice *et al*, 2001) indicates its past association with at least one co-circulating disease. Furthermore, our observation of microevolution and the potential of multiple co-circulating strains of *Y. pestis* during the Black Death is worthy of additional comment. There is some evidence in the literature that relates enhanced virulence potential to competition between different strains (Woolhouse *et al*, 2002; Ebert and Bull, 2003),

though additional circulating *Y. pestis* genomes will be required to address population-level differences in bacterial allele frequencies.

While susceptibility and within-host interactions between microbes are important determinants of disease severity, we cannot escape the fact that illness can only occur after exposure. That the Black Death traveled so efficiently through medieval Europe makes investigations into the mode of transmission paramount. Modern *Y. pestis* can survive on a variety of inanimate surfaces for extended periods (Rose *et al*, 2007), and perhaps enhanced survivability of the ancient form would have granted it efficient travel with trade goods, thus circumventing the need for direct transmission from an infected host. Such arguments as they relate to the transmission of human parasites in textiles have been raised previously by Twigg (1984), albeit in support of his view that *Bacillus anthracis*, or anthrax, was primarily responsible for the Black Death.

Central to these discussions, however, is the fact that much has yet to be learned regarding physiological processes in *Y. pestis* with respect to bacterial survival and virulence determinants. *In vitro* models that test the expression of genes under controlled conditions that mimic mammalian infection may help to identify mechanisms of transmission (Chauvaux *et al*, 2007), though models of single genetic changes as contributing to more virulent phenotypes (Joosten *et al*, 1994) are probably rare and overly simplistic since virulence is likely the result of epistatic interactions between

multiple genetic loci. Better characterisations of virulence determinants will help us to understand the complex interplay between host and pathogens with respect to *Y. pestis* infections, and determine whether indeed this infection model matches proposed theoretical concepts such as the "tradeoff model" (Anderson and May, 1982; Ewald, 1983), where the end goal of pathogen evolution is to achieve an optimal level of virulence that balances transmission with host survival, or an "arms race model", where evolutionary forces favour increases in virulence (reviewed in Ebert and Hamilton, 1996 and Brown *et al*, 2006; see Pybus and Rambaut 2009 for an example with human infection). However, the term "virulence" with respect to models of human infection is difficult to define, since genetic factors in host and microbe are not the sole determinants of susceptibility to infection, further exemplifying the need for biosocial evaluations of the landscape of factors that are involved in population level health determinants as they relate to the Black Death. Clearly a comprehensive understanding of the factors influencing infectious disease is necessary for the establishment of modern treatments and preventative strategies, but genetic data from ancient pathogens will provide one necessary element to help us achieve this ultimate goal.

References for Introduction and Conclusion

Achtman M, *et al*. 2004. Microevolution and History of the Plague Bacillus, *Yersinia pestis*. *Proc Natl Acad Sci USA* **101**: 17837-17842

Achtman M, *et al*. 1999. *Yersinia pestis*, the cause of the plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci USA* **96**:14043 – 14048

Anderson RM and May RM. 1982. Coevolution of hosts and parasites. *Parasitology* **85**:411-426

Anisimov AP, Lindler LE, and Pier GB. 2004. Intraspecific Diversity of *Yersinia pestis*. *Clin Micorbiol Rev* **17**(2):434-464

Auerbach RK *et al*. 2007. *Yersinia pestis* Evolution on a Small Timescale: Comparison of Whole Genome Sequences from North America. *PLoS One* **2**(8):e770

Aufderhide AC and Rodríguez-Martin C . 1998. The Cambridge Encyclopedia of Human Paleopathology. United Kingdom: Cambridge University Press

Ayyadurai S, Sebbane F, Raoult, and Drancourt M. 2011. Body Lice, *Yersinia pestis* Orientalis, and Black Death. *Emerg Infect Dis* **16**(5):892 – 893

Barnes I and Thomas MG. 2006. Evaluating bacterial pathogen DNA preservation in museum osteological collections. *Proc R Soc B* **273**:6545-653

Barrett R, Kuzawa CW, McDade T, and Armelagos GJ. 1998. Emerging and Re-emerging Infectious Diseases: The Third Epidemiologic Transition. *Annu Rev Anthropol* **27**:247–271

Baum J and Khalina Bar-Gal G. 2003. In: Greenblat C and Spigleman M eds. Emerging Pathogens Archaeology, Ecology, & Evolution of Infectious Diseases. Oxford University Press: Great Britain: 67-78

Benedictow, OJ. 2004. The Black Death 1346-1353: The Complete History. The Boydell Press: Great Britain

Bianucci R, Rahalison L, Rabino Massa E, Peluso A, Ferroglio E, and Signoli, M. 2008. Technological Note: A Rapid Diagnostic Test Detects Plague in Ancient Human Remains: An Example of the Interaction Between Archaeological and Biological Approaches (Southern France, 16th-18th Centuries). *Am J Phys Anthropol* **136**:361-367

Bianucci R, *et al*. 2009. Plague immunodetection in remains of religious exhumed from burial sites in central France. *J Achaeolo Sci* **36**(3):616-621

Black FL. 1992. Why did they die? *Science* **258**:1739-1741

Briggs AW, *et al*. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**(5938):318-321

Brown NF, *et al*. 2006. Crossing the Line: Selection and Evolution of Virulence Traits. *PLoS Pathog*. **2**, e42

Buikstra JE and Cook DC. 1980. Paleopathology: an American account. *Ann Rev Anthropol* **9**:433-470

Buikstra JE and Ubelaker DH. 1994. Standards for data collection from human skeletal remains. Arkansas Archeological Survey Research Series No. 44. Arkansas Archeological Survey: Fayetteville, Arkansas

Burbano HA, *et al*. 2010. Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328**:723 – 725

Carmichael AG. 2008. Universal and Particular: The Language of Plague. *Med Hist Suppl* **27**:15- 52

Chain PSG *et al*. 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with Yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA* **11**:13826 – 13831

Chanteau S, Ratsifasoamanana L, Rasoamanana B, Rahalison L, Randriambelosoa J, Roux J, and Rabeson D. 1998. Plague, a Reemerging Disease in Madagascar. *Emerg Infec Dis* **4**(1):101-104

Chauvaux S, *et al*. 2007. Transcriptome analysis of *Yersinia pestis* in human plasma: an approach for discovering bacterial genes involved in septicaemic plague. M*icrobiol* **53**:3112-3123

Cheng VCC *et al*. 2009. Fatal co-infection with swine origin influenza virus A/H1N1 and community-acquired methicillin-resistant *Staphylococcus aureus*. *J of Infect* **59**:366-370

Cohn SK. 2002. The Black Death transformed: disease and culture in early Renaissance Europe. London: Arnoldohn

Cohn SK. 2008. Epidemiology of the Black Death and Successive Waves of Plague. *Med Hist Suppl* **27**:74-100

Deng W, *et al*. 2002. Genome Sequence of *Yersinia pestis* KIM. *J Bacteriol* **184**(6):4601-4611

Devignat R. 1954. Biological and biochemical behaviour of *Pasturella pestis* and *Paturella pseudotuberculosis*. *Bull Word Health Organ* **10**(3):463-494

DeWitte SN and Wood JW. 2008. Selectivity of Black Death mortality with respect to pre-existing health. *Proc Natl Acad Sci* **105**(5):1436-1441

Dobyns HF. 1993. Disease transfer at contact. *Ann Rev Anthropol* **22**:273-291

Drancourt M, Aboudharam G, Signoli M, Dutour O, and Raoult D. 1998. Detection of 400 year-old *Yersinia pestis* DNA in human dental pulp: An approach to the diagnosis of ancient septicaemia. *Proc Natl Acad Sci US A* **95**:12637-12640

Drancourt M, Signoli M, Dang LV, Bizot B, Roux V, Tzortis S, and Raoult D. 2007. *Yersinia pestis* Orientalis Remains of Ancient Plague Patients. *Emerg Infect Dis* **13**(2):332-333

Duncan SR, Scott S, and Duncan CJ. 2005. Reappraisal of the historical selective pressures for the CCR5-Δ32 mutation. *J Med Genet* **42**:205-208

Ebert D and Bull JJ. 2003. Challenging the trade-off model for the evolution of virulence: is virulence management feasible? *Trends in Microbiol* **11**(1): 15-20

Ebert D and Hamilton WD. 1996. Sex against virulence: the coevolution of parasitic diseases. *TREE* **11**(2):79-82

Elvin SJ, *et al*. 2004. Ambigious role of CCR5 in *Y. pestis* infection. *Nature* doi:10.1038/nature02822

Enscore RE, *et a*l. 2002. Modeling relationship between climate and the frequency of human plague cased in the southwestern United States, 1960-1997. *Am J Trop Med Hyg* **66**(2):186-196

Eppinger M, *et al*. 2009. Draft Genome Sequences of *Yersinia pestis* Isolates from Natural Foci of Endemic Plague in China. *J. Bacteriol* 191:7628 – 7659

Eppinger M, *et al*. 2010. Genome Sequence of the Deep-Rooted *Yesinia pestis* Strain Angola Reveals New Insights in the Evolution and -Pangenome of the Plague Bacterium. *J Bacteriol* **192**(6):1685-1699

Ewald P. 1983. Host-parasite relations, vectors, and the evolution of disease severity. *Ann Rev Ecol Systemat* **14**:465-485

Farmer P. 2006. Social Inequalities and Emerging Infectious Diseases. *Emerg Infect Dis* **2**(4):259-269

Gagneaux S, *et al.* 2005. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* **130**(8):2869-2873

Gamsa M. 2006. The epidemic of pneumonic plague in Manchuria, 1910-1911. *Past and Present* **190**:147-184

Garcia E *et al*. 2007. Pestoides F, and Atypical *Yersinia pestis* Strain from the Former Soviet Union. *The Genus Yersinia*, in: Advances in Experimental Medicine and Biology **603**(1):17-22

Gensini GF, Yacoub MH, and Conti AA. 2004. The concept of quarantine in history: from plague to SARS. *J of Infection* **24**:257-261

Gilbert MT, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, Prentice MB. 2004. Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. Microbiology **150**(Pt 2):341-354

Gottfried RS. 1983. The Black Death: Natural and Human Disaster in Medieval Europe. Free Press: New York

Grainger I, *et al*. 2008. The Black Death cemetery, East Smithfield, London. Museum of London Archaeology Service: London

Haensch S, *et al*. 2010. Distinct Clones of *Yersinia pestis* caused the Black Death. *PLoS Pathog* **6**(10): e1001134. doi:10.1371/journal.ppat.1001134

Haldane JBS. 1949. Suggestions as to quantitative measurement of rates of evolution. *Evolution* **3**(1):51-56

Hawkins D. 1990. The Black Death and the new London cemeteries of 1348. *Antiquity* **64**:637-642

Herlihy D. 1997. The Black Death and the Transformation of the West. Harvard University Press: United States of America

Herring DA and Sattenspiel L. 2007. Social contexts, syndemics, and infectious disease in Northern Aboriginal Populations. *Am J Hum Biol* **19**:190-202

Hillson S, FitzGerald C, and Flinn H. 2005. Alternative Dental Measurements: Proposals and Relationships With Other Measurements. *Am J Phys Anthropol* **126**:413-426

Hodges E, *et al*. 2007. Genome-wide in situ exon capture for selective resequencing. Nature 39(12):1522-1527

Hodges E, *et al*. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nature Protocols **4**(6):960-974

Horrox R. 1994. The Black Death. Manchester University Press: New York

Huang X-Z, Nikolich MP, Lindler LE. 2006. Current Trends in Plague Research: From Genomics to Virulence. *Clin Med Res* **4**(3):189-199

Indian Plague Commission. 1902. The Report on the Indian Plague Commission. *BMJ* 1218 – 1220

Joosten MHA J, Cosijnsen TJ, De Wit PJG. 1994. Host resistance to a fungal tomato pathogen lost by a single base pair change in an avirulence gene. *Nature* **367**: 384 – 386

Kacki S, Rahalison L, Rajerison M, Ferroglio E, and Biannuci R. 2011. Black Death in the rural cemetery of Saint-Laurent-de-la-Carerisse Aude-Lanuedoc, southern France, 14th century: immunological evidence. *J Archaeol Sci* **38**:581-587

Lathem WW, Price PA, Miller VL, Goldman WE. 2007. A Plasminogen-Activiating Protease Specifically Controls the Development of Primary Pneumonic Plague. *Science* **315**:509-513

Liang Y, *et al*. 2010. Genome Rearrangements of Completely Sequenced Strains of *Yersinia pestis*. *J Clin Microbiol* **48**(5):1619-1623

Lovell NC. 2000. *Paleopathological Description and Diagnosis*. In Katzenberg AM and Saunders SR eds: Biological Anthropolgy of the Human Skeleton. Wiley: United States of America

Margerison BJ and Knüsel CJ. 2002. Paleodemographic Comparision of a Catastrophic and Attritional Death Assemblage. *Am J Phys Anthropol* **119**:134-143

Maricic T, Whitten M, Pääbo S. 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* **5**(11): e14004. doi:10.1371/journal.pone.0014004

May RM and Anderson RM. 1983. Epidemiology and genetics in the coevolution of parasites and hosts. *Proc R Soc Lond B* **219**: 281-313

Mays S and Taylor GM. 2002. Osteological and Biomolecular Study of Two Possbile Cases of Hypertrophic Osteoarthropathy from Medieval England. *J Archaeol Sci* **29**(11):1267-1276

McDonald BA, Goodwin SB, and Allard RW. 1989. The population biology of host-pathogen interactions. *Annu Rev Phytopathol* **27**:77-94

McNeill W. 1976. Plagues and People New York: Doubleday

Mescas J, *et al*. 2004. CCR5 mutation and plague protection. *Nature* **427**:606

Morelli G, *et al*. 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* **42**(12):1140 – 1143

Navas A, *et al*. 2007. Experimental validation of Haldane's hypothesis on the role of infection as an evolutionary force for Metazoans. *Proc Natl Acad Sci* **104**(34):13728-13731

Ortner DJ. 2003. Identification of pathological lesions in human skeletal remains. Academic Press: San Diego

Pääbo S, *et al*. 2004. Genetic Analyses from Ancient DNA. *Annu Rev Genet* **38**:645-679

Parkhill J, *et al*. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**:523 – 527

Parmenter RR *et al*. 1999. Incidence of Plague Associated with Increased Winter-Spring Precipitation in New Mexico. *Am J Trop Med Hyg* **61**(5):814-821

Perry RD and Fetherston JD. 1997. *Yersinia pestis* – Etiologic Agent of Plague. *Clin Microbiol Rev* **10**(1):35-66

Poinar HN, *et al*. 2006. Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science* **311**:392-394

Pouillot, F., Fayolle C, and Carniel E. 2008. Characterization of Chromosomal Regions Conserved in *Yersinia pseudotuberculosis* and Lost by *Yersinia pestis*. *Infect Immun* **76**, 4592 – 4599

Prentice MB, *et al*. 2001. *Yersinia pestis* pFra Shows Biovar-Specific Differences and Recent Common Ancestry with *Salmonella enteric Serovar Typhi* Plasmid. *J Bacteriol* **183**(8):2586-2594

Prentice MB, Gilbert T, and Cooper A. 2004. Was the Black Death caused by *Yersinia pestis*? *Lancet Infect Dis* **4**:72

Pusch CM, Rahalison L, Blin N, Nicholson GJ, Czarmetzki A. 2007. Yersinial F1 antigen and the cause of Black Death. *Lancet Infect Dis* **4**:484-485

Pybus OG and Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* **10**:540-550

Raoult D, Aboudharam G, Crubezy E, Larrouy G, Ludes B, Drancourt M. 2000. Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of medieval black death. *Proc Natl Acad Sci USA* **97**(23):12800-12803

Relman DA. 2011. Microbial Genomics and Infectious Diseases. *N Engl J Med* **365**:347-357

Reich D, *et al*. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**:1053-1060

Roberts CA and Grauer A. 2001. Commentary: Bones, bodies and representivity in the archaeological record. *Int J Epidemiol* **30**(1):109-110

Rose LJ, Donlan R, Banerjee SN, and Arduino MJ. 2007. Survival of *Yersinia pestis* on Environmental Surfaces. *Appl Env Microbiol* **69**(4):2166-2171

Rothschild BM, Calderon F, Coppa A, and Rothschild C. 2000. First European Exposure to Syphilis: The Dominican Republic at the time of Columbian Contact. *Clin Infect Dis* **31**(4):936-941

Rowbotham TJ. 1980. Preliminary report on the pathogenicity of Legionella pneumophila for freshwater and soil amoebae. *J Clin Pathol* **33**:1179-1183

Sabeti PC, *et al*. 2005. The Case for Selection at CCR5-Δ32. *PLoS Biology* **3**(11):e378

Samia NI, *et al*. 2011. Dynamics of the plague-wildlife-human system in Central Asia are controlled by two epidemiological thresholds. *Proc Natl Acad Sci* doi/10.1073/pnas.1015946108

Scott S and Duncan CJ. 2001. The Biology of Plagues. Cambridge: Cambridge University Press

Sherman IW. 2006. The Power of Plagues. ASM Press: Washington, D.C.

Singer M. 2010. Pathogen-pathogen interaction: A syndemic model of complex biosocial processes. *Virulence* **1**(1):10 -18

Singer, M. and Clair S. 2003. Syndemics and Public Health: Reconceptualizing Disease in Bio-Social Context. *Med Anthropol Q* **17**, 423 – 441

Song Y, *et al*. 2004. Complete Genome Sequence of *Yersinia pestis* Strain 91001, an Isolate Avirulent to Humans. *DNA Research* **11**:179-197

Stenseth NC, *et al*. 2006. Plague dynamics are driven by climate variation. *Proc Natl Acad Sci* **103**(35):13110-13115

Stenseth NC, *et al*. 2008. Plague: Past, Present, and Future. *PLoS Med* **5**(1): e3. doi:10.1371/journal.pmed.0050003

Stephens JC, *et al*. 1998. Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes. *Am J Hum Genet* **62**:1507-1515

Stosor V and Wolinsky S. 2001. GB virus C and mortality from HIV infection. *N Engl J Med* **345**(10):761-762

Styer KL, Click EM, Hopkins GW, Frothingham R, and Aballay A. 2007. Study of the role of CCR5 in a mouse model of intranasal challenge with *Yersinia pestis*. *Microb Infect* **9**:1135- 1138

Taubenberger JK, *et al*. 2004. Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**:889-893

Tran T, Raoult D, and Drancourt M. 2011. *Yersinia pestis* DNA Sequences in Late Medieval Skeletal Finds, Bavaria. *Emerg Infect Dis* **17**(5):955-956

Twigg G. 1984. The Black Death: a biological reappraisal. London: Batsford Academic

Twigg G. 2003. The Black Death and DNA. *Lancet Infect Dis* **3**:11

Waldron HA. 2001. Are plague pits particular use to palaeoepidemiologists? *Int J Epidemiol* **30**(1):104-108

Weichmann I and Grupe G. 2005. Detection of *Yersinia pestis* DNA in Two Early Medieval Skeletal Finds from Aschheim (Upper Bavaria, 6th Century A.D.). *Am J Phys Anthropol* **126**:48-55

Weichmann I, Harbeck M, Grupe G. 2010. *Yersinia pestis* DNA sequences in Late Medieval Skeletal Finds, Bavaria. *Emerg Infect Dis* **16**(11):1806 – 1807

WHO. 2008. Interregional meeting on prevention and control of plague. www.who.int/csr/resources/publications/WHO_HSE_EPR_2008_3w.pdf (2008)

Wood J and DeWitte-Aviña S. 2003. Was the Black Death yersinial plague? *Lancet Infect Dis* **3**(6):327-328

Wood JW, Milner GR, Harpending HC, and Weiss KM (1992). The Osteological Paradox: Problems of Inferring Prehistoric Health from Skeletal Samples. *Curr Anthropol* 33(4):343-370

Wood JW, and DeWitte-Aviña SN. Was the Black Death yersinial plague? *Lancet Infect Dis* 2003; 3:327-328

Wood, J.W, R.J. Ferrell, and S.N. DeWitte-Aviña. 2003. The temporal dynamics of the fourteenth-century Black Death: New evidence from ecclesiastical records. *Human Biology* **75**:427-448

Woolhouse MEJ, Webster JP, Domingo E, Charlesworth B, and Levin BR. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet* **32**:569-577

Xu, L, *et al*. 2011. Nonlinear effect of climate on plague during the third pandemic in China. *Proc Natl Acad Sci USA* doi: 10.1073/pnas.1019486108

Yang H, *et al*. 2011. Insight into bacterial virulence mechanisms against host immune response via *Y. pestis*-human protein interaction framework. *Infect Immun* doi:10.1128/IAI.05622-11

Zhang P, *et al*. 2008. Human Dendritic Cell-Specific Intercellular Adhesion Molecule-Grabbing Nonintegrin (CD209) Is a Receptor for *Yersinia pestis* That Promotes Phagocytosis by Dendritic Cells. *Infect Immun* **76**(5): 2070-2079

Zhou D, *et al*. 2004. DNA Microarray Analysis of Genome Dynamics in *Yersinia pestis*: Insights into Bacterial Genome Microevolution and Niche Adaptation. *J Bacteriol* **186**(15):5138- 5146

Appendices

Appendix 1

Table 1 - Results for oligo design features of PrimerClique applying different design criteria.
Program defaults are used for all undefined paramters. Template sizes are as follows:
human mtDNA = 16.5kb, ebola = 18.8Kb, human Y. chromosome = 26Mb, E.coli O157:H7 = 5.5Mb

a) unidirectional primer generation

Template	Primer size (bp)	No. candidates requested	Adaptor filtering?	No. primers filtered out by adpators (% removed)	Mispriming library	No. primers filtered out with mispriming library (% removed)	No. of remaining primers	Maximum potential coverage	Overlap tolerance	Number of multiplexes	Number of primers in best set	Multiplex template coverage (%)	Size of largest excluded region
human mtDNA	18- 30	default	none	0	none	0	53734	99.41	10	14	629	95.12	17
human mtDNA	18- 30	default	Shotgun	24684 (46%)	none	0	29050	95.84	10	13	569	88.04	47
human mtDNA	18- 30	10 000	none	0	none	0	8775	93.11	10	9	498	77.15	53
human mtDNA	18- 30	10 000	Shotgun	3758 (43%)	none	0	5017	81.38	10	8	413	64.13	92
ebola	18- 30	20000	none	0	none	0	16739	96.46	10	14	608	83.7	46
ebola	18- 30	20000	none	0	human	1520 (9%)	15219	93.18	10	16	597	80.93	68
ebola	18- 30	20000	Illumina	5962 (36%)	none	0	10777	91.54	10	13	552	76.25	65
ebola	18- 30	20000	Illumina	5668	human	1520	9551	86.11	10	15	529	72	86

b) PCR primer generation

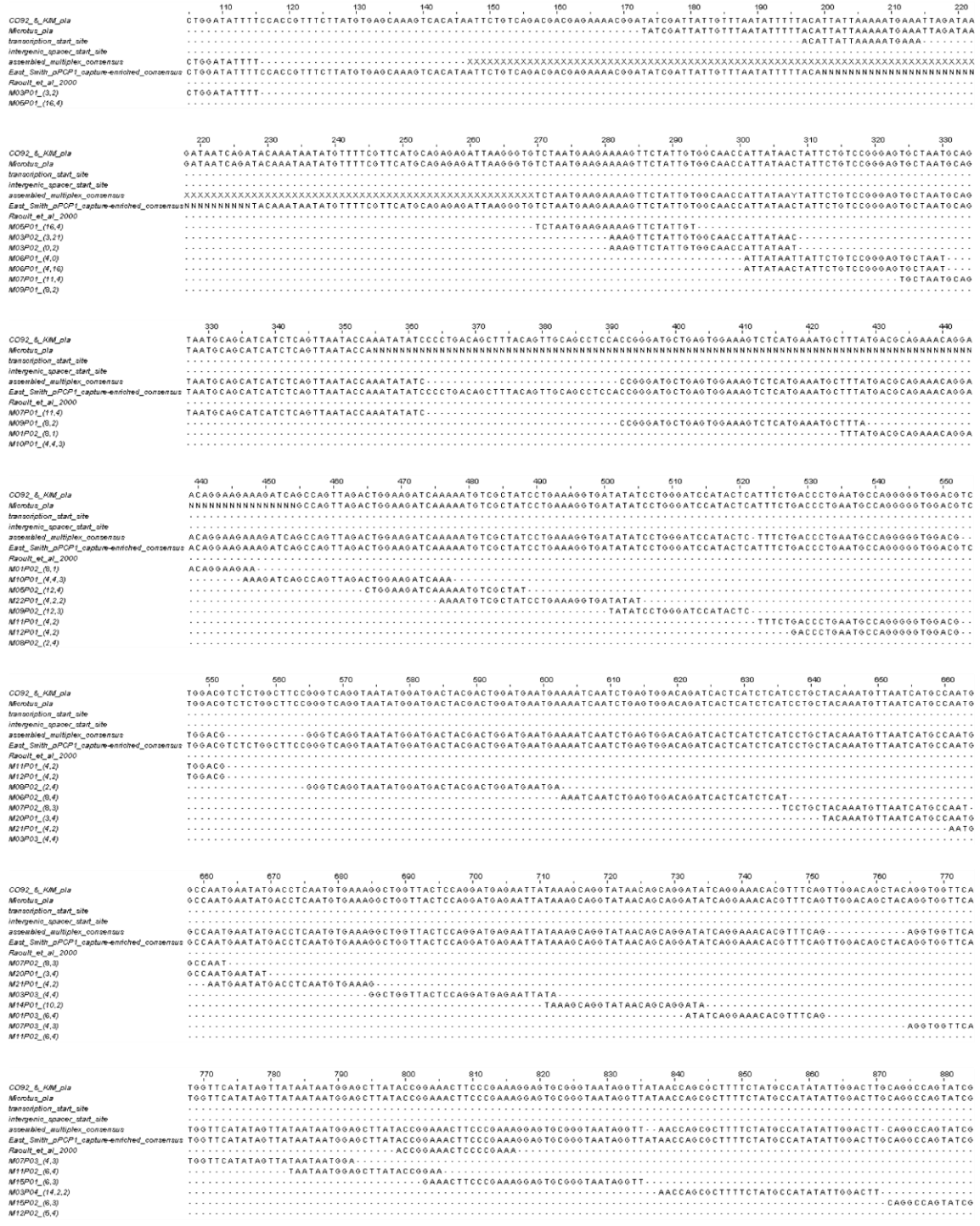
Template	Product length (bp)	No. primer pairs requested	temperature difference between F and R primers	Adaptor filtering?	No. primers filtered out by adaptors	Mispriming library	No. primers filtered out with mispriming library	No. of remaining primers	Maximum potential coverage	Overlap tolerance	Number of multiplexes	Number of primers in best set	Multiplex template coverage (%)	Size of largest excluded region
human mtDNA	70 - 80	default	2	none	0	none	0	11091	97.82	10	42	748	96.58	46
human mtDNA	70 - 80	default	2	none	0	none	0	11091	97.82	10	*7	748	96.58	46
human mtDNA	70 - 80	default	2	Shotgun	7013	none	0	4080	85.67	10	*7	577	82.37	104
human mtDNA	200-300	default	5	none	0	none	0	10543	99.67	20	13	108	99.63	28
human mtDNA	200-300	default	5	Shotgun	6245	none	0	4309	99.45	20	11	99	99.21	62
human mtDNA	2000 - 3000	default	5	none	0	none	0	7783	99.66	100	2	8	99.45	63
human mtDNA	2000 - 3000	default	5	Shotgun	7807	none	0	3045	99.63	100	3	8	99.44	65
ebola	2000 - 3000	default	5	none	0	none	0	10701	99.19	100	3	9	99.19	90
ebola	2000 - 3000	default	5	none	0	human	not determinalble	11465	99.22	100	3	9	98.5	194

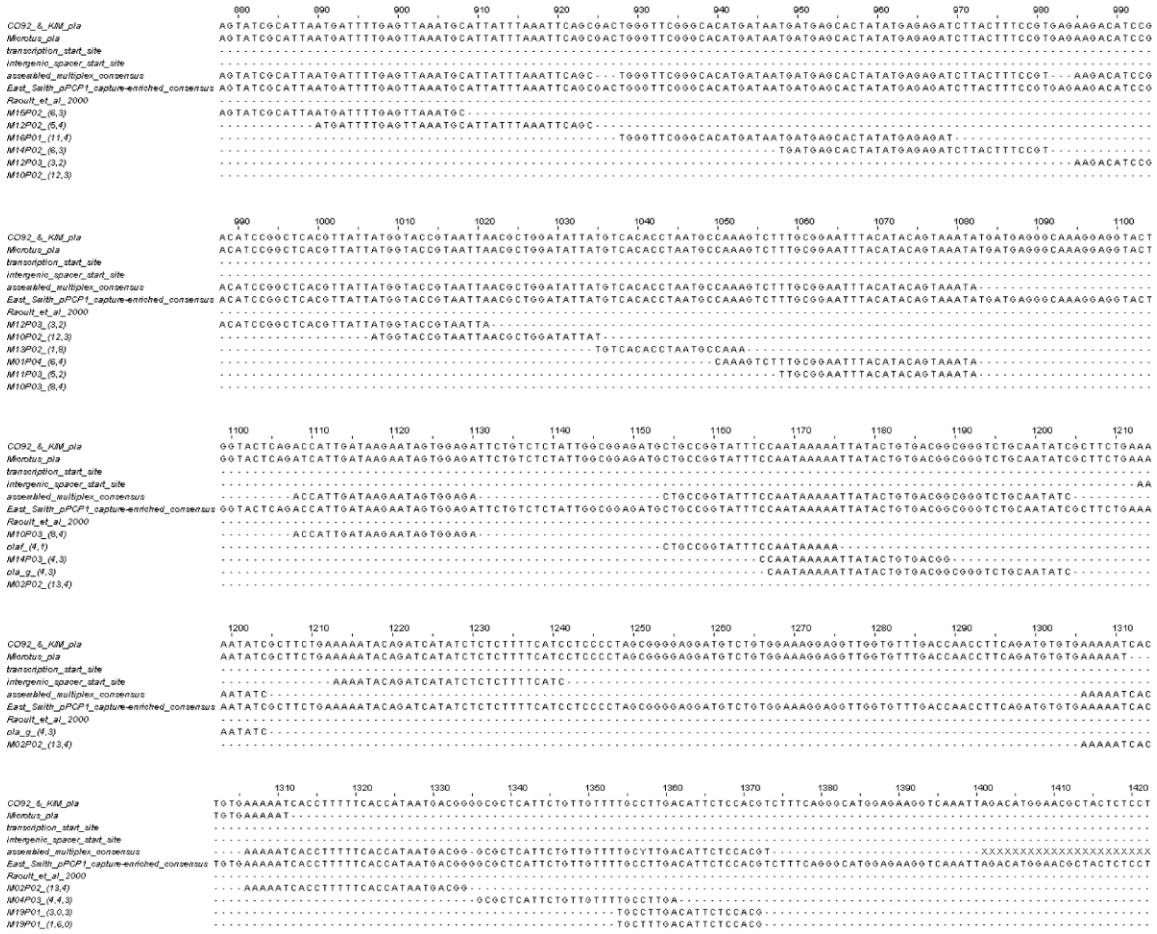
*done without alignment compatibility for multiplex generation

c) array probe generation

Template	probe length	Tiling density (bp)	Number of probes	No. probes after avg. 15-mer repeat filtering (number removed)	Adaptor filtering?	No. probes after adaptor filtering (number removed)	Remove duplicate probes?	Number of unique probes remaining (number removed)	Average fold coverage per template base	probe partitioning?	No. probe groups
large template	70	5	5130530	4544112 (586418)	none	4544112 (0)	N	4544112 (0)	12.4	Y (1 million: 5°C)	11
large template	70	5	5130530	4544112 (586418)	shotgun	4544024 (88)	N	4544024 (0)	12.4	Y (1 million: 5°C)	11
large template	70	5	5130530	4544112 (586418)	none	4544112 (0)	Y	4461907 (82205)	12.4	Y (1 million: 5°C)	11
large template	70	5	5130530	4544112 (586418)	shotgun	4544024 (88)	Y	4461872 (82152)	12.4	Y (1 million: 5°C)	11
E. coli O157:H7	70	6	921397	920724 (673)	none	920724 (0)	Y	914135 (6589)	11.66	Y (1 million: 5°C)	7
E. coli O157:H7	70	6	921397	920724 (673)	shotgun	920715 (9)	Y	914126 (6589)	11.66	Y (1 million: 4°C)	9

Appendix 2 – Dataset S1




CO92_&_KIM_pla
Microtus_pla
transcription_start_site
intergenic_spacer_start_site
assembled_multiplex_consensus
East_Smith_pPCP1_capture-enriched_consensus
Rasulf_et_al_2000
M16P02_(6,3)
M12P02_(5,4)
M16P01_(11,4)
M14P02_(6,3)
M12P03_(3,2)
M10P02_(12,3)
M12P03_(3,2)
M10P02_(12,3)
M13P02_(1,8)
M01P04_(6,4)
M11P03_(5,2)
M10P03_(8,4)
M10P03_(8,4)
clat_(4,1)
M14P03_(4,3)
pla_g_(4,3)
M02P02_(13,4)
M02P02_(13,4)
M04P03_(4,4,3)
M19P01_(3,0,3)
M19P01_(1,6,0)

Appendix 3
Table S1b - mapping results

		STEP 1				STEP 2				STEP 3			
		number of mapping reads (percentage of total reads)				number of mapping reads after duplicate removal (rmdup)				duplicate removal and mapping Q>30			
individual	total fragments	chromosome C092	pMT1	pCD1	pcp1	chromosome C092	pMT1	pCD1	pcp1	chromosome C092	pMT1	pCD1	pcp1
ES_11972	3841602	2521719 (65.64%)	1421426 (37.00%)	24873 (0.65%)	73 (0.00%)	938960 (24.44%)	16363 (0.43%)	13706 (0.36%)	72 (0.00%)	628716	16223	10818	60
ES_6330	2168850	1369390 (63.14%)	767160 (35.37%)	14380 (0.66%)	26 (0.00%)	559162 (27.78%)	11811 (0.55%)	8759 (0.40%)	26 (0.00%)	308916	11713	6934	19
ES_8124	1512672	644143 (42.58%)	352231 (23.29%)	5673 (0.38%)	14 (0.00%)	358582 (23.71%)	9761 (0.65%)	4007 (0.27%)	14 (0.00%)	172096	9667	3055	4
ES_8291	9822862	6331304 (64.45%)	3364246 (34.25%)	59769 (0.61%)	72 (0.00%)	1644894 (16.75%)	31081 (0.32%)	28473 (0.29%)	69 (0.00%)	1253005	30871	23663	44
whole pool	20512847	13180582 (64.26%)	7162613 (34.92%)	128183 (0.62%)	225 (0.00%)	2812240 (13.71%)	50006 (0.24%)	49143 (0.24%)	204 (0.00%)	2366647	49630	42315	137
Controls	684	289 (42.25%)	81 (11.84%)	8 (1.17%)	7 (1.02%)	267 (39.04%)	78 (11.40%)	8 (1.17%)	6 (0.88%)	42	69	0	0

Coverage estimate after Q>30 and duplicate removal				
individual	genome	pMT1	pCD1	pcp1
ES_11972	6.6	10.1	7.9	0.3
ES_6330	3.5	7.6	5.4	0.1
ES_8124	2.2	7.5	2.7	0
ES_8291	14.8	19.9	19.4	0.2
whole pool	28.2	31.2	35.2	0.7
Controls	0	0	0	0

Coverage of CO92 after duplicate removal					
Positions covered	nt mapping	1-fold	2-fold	3-fold	4-fold
4619022	1.62E+08	99.25%	98.62%	97.98%	97.29%

Coverage of CO92 after duplicate removal and filtering at Q>30					
Positions covered	nt mapping	1-fold	2-fold	3-fold	4-fold
4350311	1.31E+08	93.48%	92.85%	92.22%	91.54%

K. Bos; McMaster University

Appendix 4a

Table S3a - differences identified between the ancient chromosome and the CO92 reference chromosome used for bait design.
Numbering follows NC_003143.
*** = not present in individual 6330**
^ = potentially important in virulence

Position	CO92	ES	derived	ancestral	Known polymorphic position[11]?	Codon (ancestral -> derived)	aa change (ancestral -> derived)	syn/non-syn	Gene	Description
130	G	C	G	C	No	Intergenic				
74539	C	T	C	T	Yes	ACC -> GCC	T->A	Nonsyn	YPO0063	hypothetical protein
105187	A	C	A	C	No	Intergenic				
130643	G	A	G	A	Yes	ATC -> ACC	I->T	Nonsyn	glpE YPO0122	thiosulfate sulfurtransferase
150946	C	A	C	A	No	ACT -> ACG	T->T	Synon	pckA YPO0138	phosphoenolpyruvate carboxykinase
155747	A	G	A	G	Yes	AGC -> AGT	S->S	Synon	YPO0142	hypothetical protein
286528	T	A	T	A	Yes	CGT -> CGA	R->R	Synon	YPO0285	hypothetical protein
325836	T	C	T	C	Yes	CCG -> CCA	P->P	Synon	qor YPO0319	quinone oxidoreductase, NADPH-dependent
351821	T	G	T	G	No	CAG -> CAT	Q->H	Nonsyn	YPO0342	putative oxidoreductase Fe-S binding subunit
442439	T	C	T	C	Yes	Intergenic				
545488	T	C	T	C	No	CAA -> TAA	Q->stop	Nonsyn	YPO0508	hypothetical protein
*699494	A	G	A	G	Yes	GGC -> GGT	G->G	Synon	rpoD YPO0643	RNA polymerase sigma factor RpoD
699647	T	C	T	C	Yes	CAG -> CAA	Q->Q	Synon	rpoD YPO0643	RNA polymerase sigma factor RpoD
773110	T	C	T	C	Yes	AAG -> AAA	K->K	Synon	fliM YPO0711	putative flagellar motor switch protein
877258	T	C	T	C	Yes	CTT -> TTT	L->F	Nonsyn	lysR YPO0797	DNA-binding transcriptional regulator LysR

895071	T	C	C	Yes	CGT -> CAT	R->H	Nonsyn	YPO0816	general secretion pathway protein D
917155	A	G	G	No	GGT -> AGT	G->S	Nonsyn	YPO0837	putative PTS permease
1001553	A	G	G	No	Intergenic				
1017647	T	C	C	Yes	GCG -> GCA	A->A	Synon	tktA YPO0926	transketolase
1025278	T	G	G	Yes	GCT -> TCT	A->S	Nonsyn	YPO0932	hypothetical protein
1098675	A	C	C	No	Intergenic				
1178178	T	C	C	No	Intergenic				
1178459	T	C	C	Yes	CTG -> TTG	L->L	Synon	YPO1037	hypothetical protein
1272559	T	C	C	Yes	GCT -> GTT	A->V	Nonsyn	YPO1126	tol-pal system protein YbgF
1306718	T	C	C	Yes	Intergenic				
1385780	T	C	C	Yes	AAG -> AAA	K->K	Synon	ampH YPO1224	beta-lactam binding protein AmpH
1512930	A	G	G	Yes	GCG -> ACG	A->T	Nonsyn	YPO1348	hypothetical protein lies in intergenic region of a T6SS
^1658492	C	A	A	No	Intergenic				lies in intergenic region of a T6SS
^1658495	C	T	T	No	Intergenic				
1735263	A	C	C	Yes	GCC -> GCA	A->A	Synon	YPO1526	putative assembly protein
1749443	T	C	C	Yes	CCG -> CCA	P->P	Synon	YPO1537	putative iron-siderophore receptor
1808946	T	C	C	Yes	Intergenic				
1914093	T	C	C	No	Intergenic				
1939828	T	G	G	No	GCC -> GCA	A->A	Synon	YPO1701	hypothetical protein
1939841	A	G	G	No	CCT -> CTT	P->L	Nonsyn	YPO1701	hypothetical protein
2022335	A	C	C	Yes	Intergenic				
2098628	T	C	C	Yes	AGT -> AAT	S->N	Nonsyn	putA YPO1851	
^2218046	G	T	T	No	GTC -> GGC	V->G	Nonsyn	hmsR YPO1953	N-glycosyltransferase
*2262577	T	G	G	**Yes**	CGG -> CTG	R->L	Nonsyn	YPO1990	hypothetical protein
2273616	G	C	C	No	AGG -> ACG	R->T	Nonsyn	YPO2000	putative two-component system sensor protein
2277583	G	A	A	Yes	Intergenic				
2278317	A	G	G	Yes	GCT -> GTT	A->V	Nonsyn	YPO2005	hypothetical protein

2300659	T	G	T	G	Yes	GCC -> GAC	A->D	Nonsyn	YPO2029	hypothetical protein
2356003	T	A	T	A	Yes	AAA -> AAT	K->N	Nonsyn	fadD YPO2074	long-chain-fatty-acid--CoA ligase
2453454	A	G	A	G	No	GCC -> GCT	A->A	Synon	adhE YPO2180	
2508389	T	C	T	C	Yes	GCA -> ACA	A->T	Nonsyn	cstA YPO2234	putative carbon starvation protein A
2548551	G	T	G	T	Yes	GAT -> GCT	D->A	Nonsyn	YPO2266	hypothetical protein
2575152	G	A	G	A	Yes	TTA -> TTG	L->L	Synon	YPO2291	putative virulence factor
2577686	A	G	A	G	Yes	ACG -> ATG	T->M	Nonsyn	YPO2292	putative lipoprotein
2619611	T	G	T	G	No	GCG -> GAG	A->E	Nonsyn	YPO2328	hypothetical protein
2684793	A	G	A	G	Yes	Intergenic				
2739149	C	A	C	A	Yes	ACA -> ACC	T->T	Synon	yfeA YPO2439	periplasmic-binding protein
2744933	A	G	A	G	Yes	GTT -> ATT	V->I	Nonsyn	YPO2446	2-deoxyglucose-6-phosphatase
2894703	T	C	T	C	Yes	GCT -> ACT	A->T	Nonsyn	YPO2574	hypothetical protein
2903882	T	G	T	G	Yes	GGT -> TGT	G->C	Nonsyn	YPO2582	sugar transport ATP-binding protein
2934972	C	G	C	G	Yes	ACC -> ACG	T->T	Synon	gltL YPO2612	
2936268	G	A	G	A	Yes	TTT -> CTT	F->L	Nonsyn	gltU YPO2614	
2950954	G	A	G	A	Yes	GTG -> GCG	V->A	Nonsyn	nagC YPO2625	putative N-acetylglucosamine regulatory protein
3085079	A	G	A	G	Yes	GCT -> ACT	A->T	Nonsyn	mepA YPO2752	penicillin-insensitive murein endopeptidase
3145523	A	C	A	C	Yes	CGC -> CTC	R->L	Nonsyn	YPO2814	hypothetical protein
3190399	A	G	A	G	Yes	CGG -> CAG	R->Q	Nonsyn	baeR YPO2853	DNA-binding transcriptional regulator BaeR
3244204	A	G	A	G	Yes	GCC -> GTC	A->V	Nonsyn	YPO2901	hypothetical protein
3267118	A	G	A	G	Yes	CCG -> CTG	P->L	Nonsyn	purL YPO2921	phosphoribosylformylglycinam idine synthase
3324959	A	G	A	G	Yes	Intergenic				

3362591	A	G	G	Yes	GGC -> AGC	G->S	Nonsyn	YPO3009	putative two-component response regulator
3397040	A	G	G	Yes	AAG -> AAA	K->K	Synon	YPO3043	aminoglycoside/multidrug efflux system
3421335	A	G	G	Yes	GTG -> ATG	V->M	Nonsyn	bcp YPO3064	thioredoxin-dependent thiol peroxidase
3442617	A	T	T	Yes	AGT -> AGA	S->R	Nonsyn	copA YPO3086	copper exporting ATPase
3564026	C	T	T	Yes	TAT -> TGT	Y->C	Nonsyn	proY YPO3201	putative proline-specific permease
3571531	A	G	G	No	Intergenic				
3608932	T	C	C	No	GGC -> GAC	G->D	Nonsyn	gmhA YPO3243	phosphoheptose isomerase
3616733	A	G	G	Yes	CCA -> CTA	P->L	Nonsyn	hmwA YPO3247	putative adhesin
3645151	C	G	G	Yes	GGC -> GCC	G->A	Nonsyn	yfiQ YPO3272	putative acetyltransferase
3647867	C	T	T	No	GTT -> GCT	V->A	Nonsyn	pssA YPO3273	phosphatidylserine synthase
3655609	T	C	C	No	GAA -> AAA	E->K	Nonsyn	clpB YPO3275	protein disaggregation chaperone
3667806	A	G	G	Yes	GAA -> AAA	E->K	Nonsyn	YPO3287	
3725545	T	C	C	Yes	GAG -> GAA	E->E	Synon	YPO3339	hypothetical protein
3726726	A	G	G	No	Intergenic				
3732919	A	G	G	Yes	Intergenic				
3739401	C	A	A	Yes	GGT -> GGG	G->G	Synon	ydjJ YPO3352	putative Zinc-binding dehydrogenase
3806677	C	T	T	Yes	AAG -> AGG	K->R	Nonsyn	YPO3408	hypoxanthine phosphoribosyltransferase
3886839	T	C	C	No	GAG -> GAA	E->E	Synon	ibeB YPO3481	putative outer membrane efflux lipoprotein
3973746	C	T	T	Yes	ATT -> ACT	I->T	Nonsyn	YPO3559	hypothetical protein
^4026673	T	C	C	No	GCG -> GTG	A->V	Nonsyn	YPO3613	putative Rhs accessory genetic element
4082562	T	C	C	Yes	CAG -> CAA	Q->Q	Synon	YPO3662	putative sulfite oxidase subunit YedY
4083536	A	G	G	Yes	Intergenic				

4194600	G	A	G	A	Yes	ATT -> GTT	I->V	Nonsyn	thiG YPO3742	thiazole synthase
4243823	A	T	A	T	Yes	TAT -> TTT	Y->F	Nonsyn	ubiE YPO3781	
4371886	A	G	A	G	Yes	ACC -> ACT	T->T	Synon	ilvD YPO3897	dihydroxy-acid dehydratase
4421633	T	C	T	C	Yes	GCG -> GTG	A->V	Nonsyn	glpD YPO3937	glycerol-3-phosphate dehydrogenase
4421689	A	G	A	G	Yes	GCA -> ACA	A->T	Nonsyn	glpD YPO3937	glycerol-3-phosphate dehydrogenase
4518401	G	A	G	A	Yes	AAT -> AAC	N->N	Synon	uhpC YPO4007	regulatory protein UhpC
4527483	A	G	A	G	Yes	Intergenic				
4542642	A	G	A	G	No	Intergenic				
4579183	A	G	A	G	Yes	GGC -> AGC	G->S	Nonsyn	fdhD YPO4060	formate dehydrogenase accessory protein
4624135	C	G	C	G	No	CGC -> CCC	R->P	Nonsyn	trmE YPO4103	tRNA modification GTPase TrmE
4634287	A	G	A	G	Yes	CCG -> TCG	P->S	Nonsyn	phoU YPO4113	transcriptional regulator PhoU ST3VE

Appendix 4b

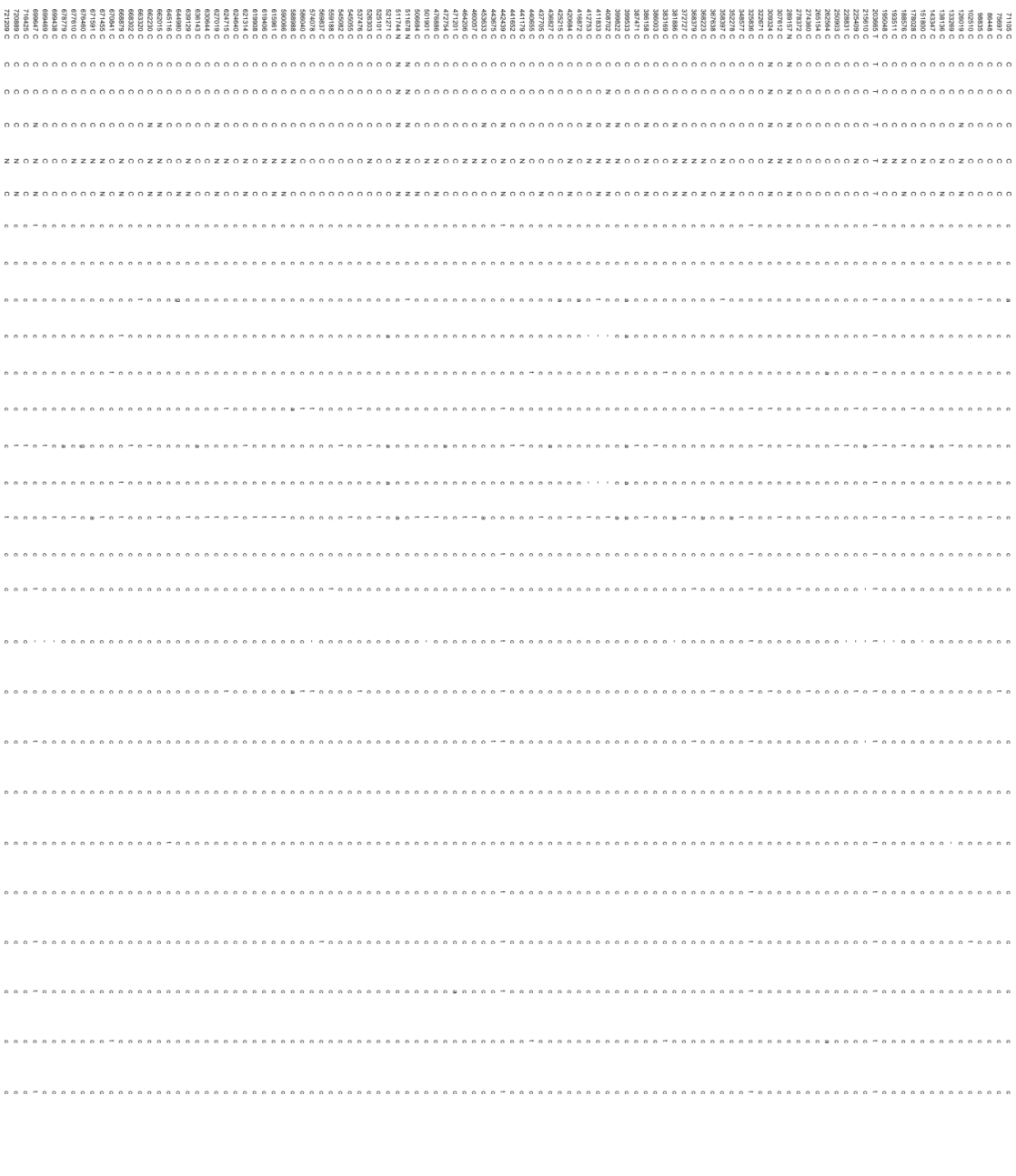
Table S3b - differences identified between the ancient plasmids and the CO92 reference plasmids used for bait design. Numbering follows NC_009595 and NC_009596 for pCD1 and pMT1, respectively.
*** indicates gene on the complement strand**
^ indicates a potential role in virulence

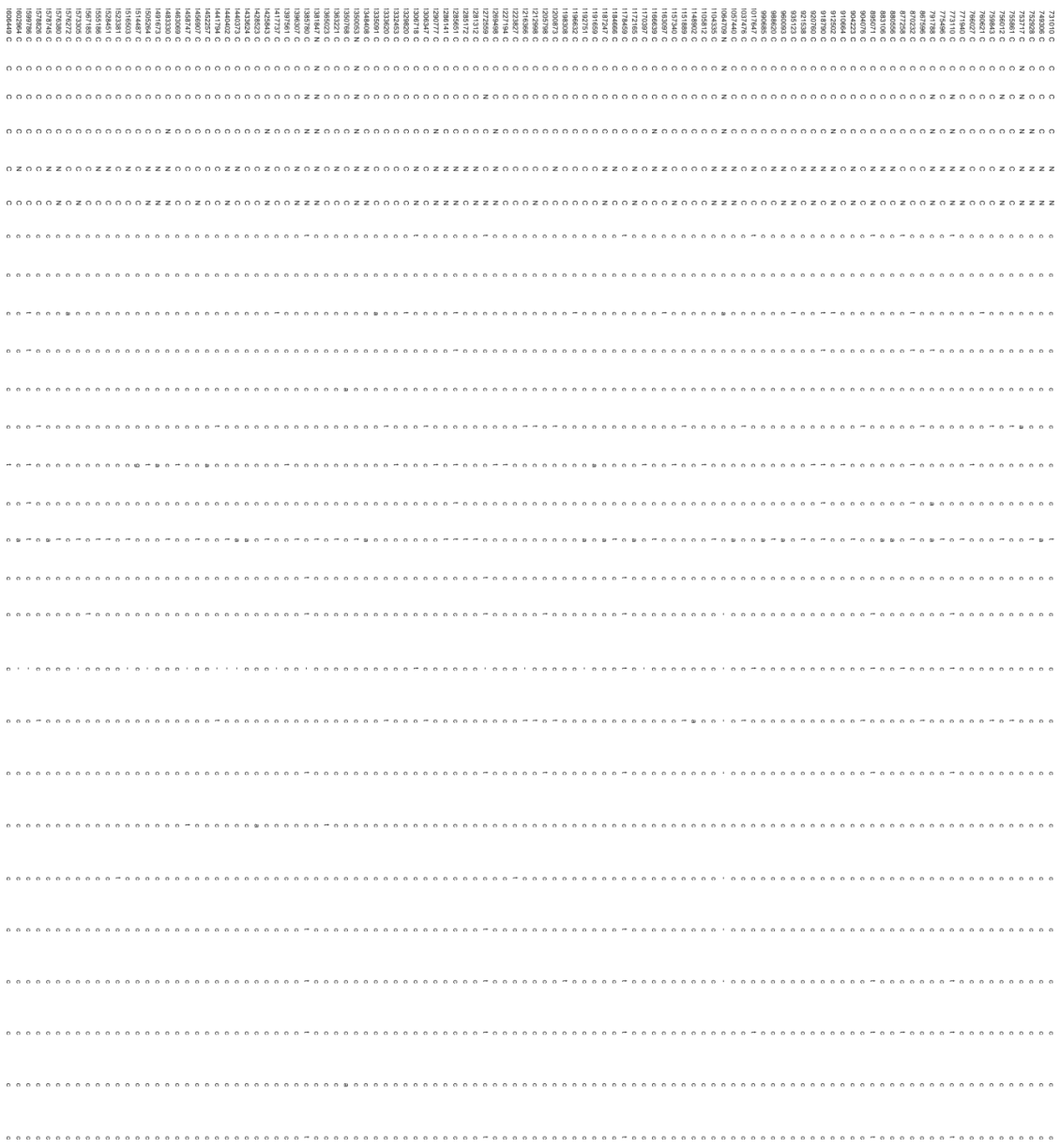
Position	ES	CO92 Ref	Codon	Amino Acid Change	Syn/ non-syn	Gene	Description
pCD1:							
21115	T	C				Intergenic	
^22717	C	A				Intergenic	
^26492	G	A	GAA -> AAA	E -> K	nonsyn	yopJ YPE_4183	
53390	C	A	GCC -> GCA	A -> A	syn	lcr YPE_4221	hypothetical protein
pMT1:							
36403	C	A	ACG -> ACT*	T -> T	synon	YPE_4284	hypothetical protein
84127	T	A				Intergenic	

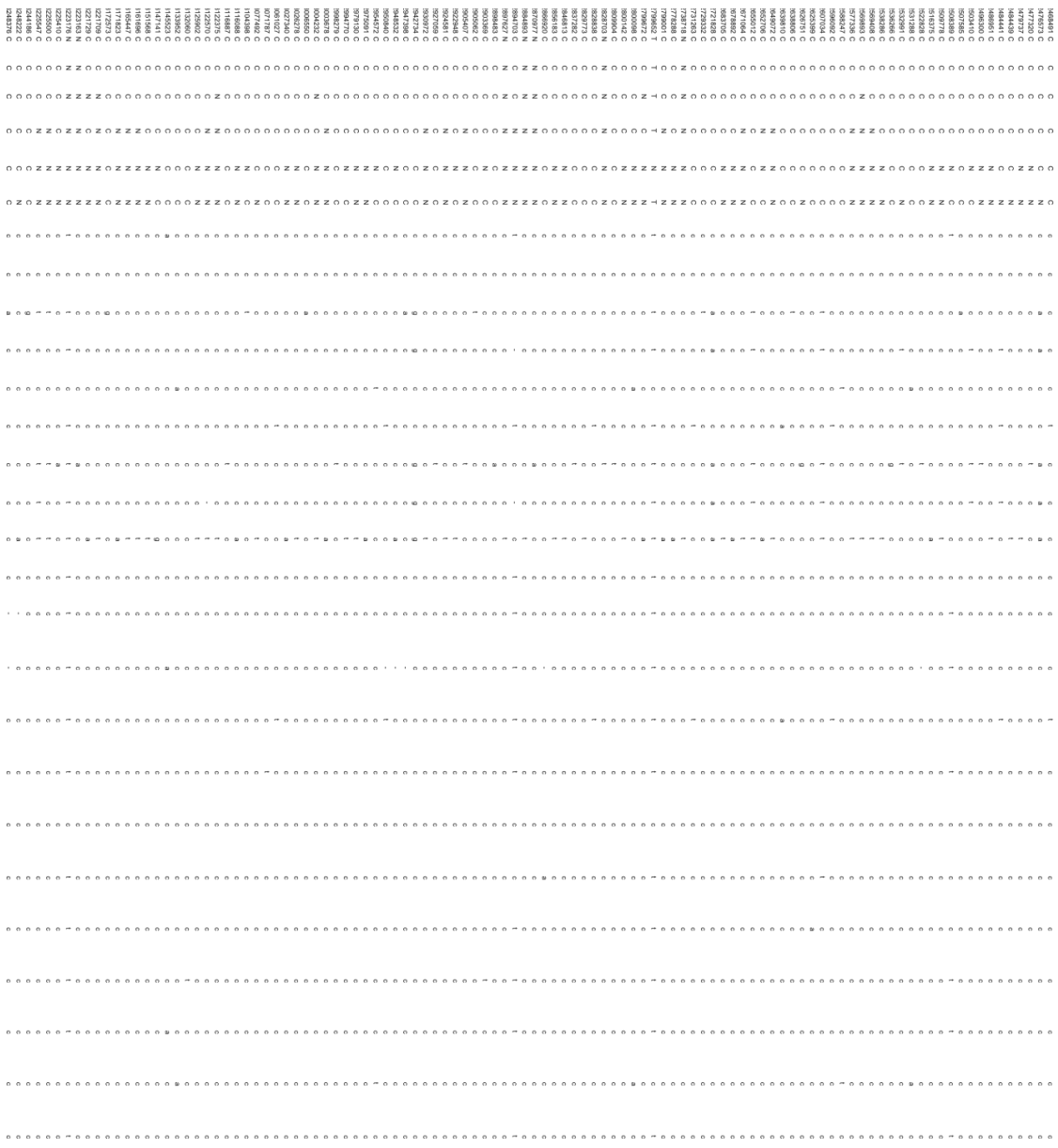
Appendix E
Table S4 - Known variant positions in Y peptis [1], with base calls for 21 Y. pestis genomes

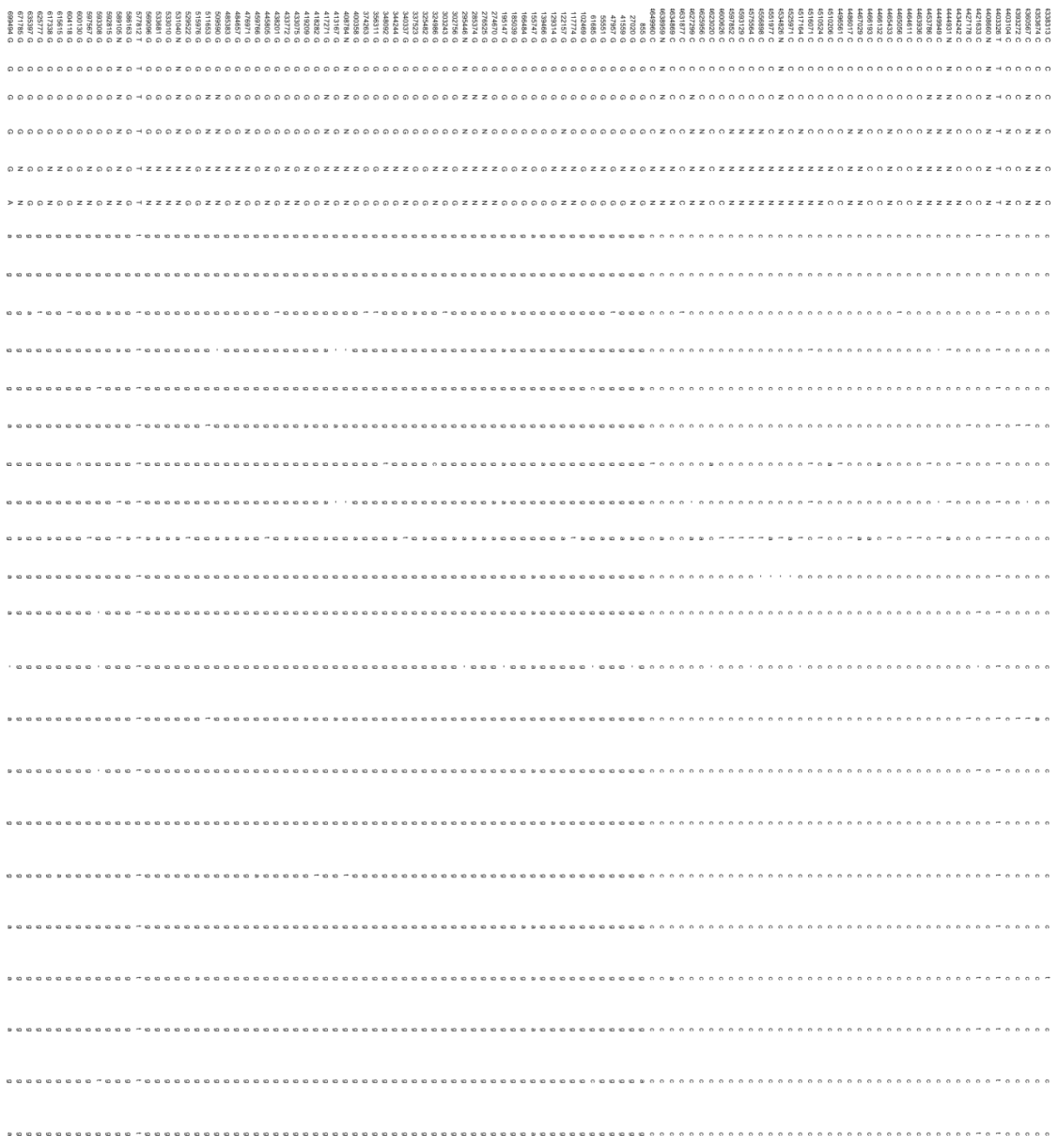
CQ92 Pos	pool.all	pool.3	Ind 8291	Ind 11972	Ind 8124	Ind 8330	NC_003143	NC_004088	NC_005810	NC_005815	NC_008149	NC_009135	NC_009381	NC_009708	NC_010159	NC_014029	NZ_AAOS02000	NZ_AAVB01000	NZ_AAYR01000	NZ_AAYZ01000	NZ_AAYT01000	NZ_AAYU01000	NZ_AAYV01000	NZ_AMJR01000	NZ_ABCD01000	NZ_ACNO01000	NZ_ACNB01000
2256	2256	G	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22687	22687	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
84787	84787	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
90183	90183	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
95031	95031	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
98534	98534	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
126977	126977	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
130543	130543	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
179899	179899	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
212227	212227	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
242986	242986	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
269993	269993	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
282762	282762	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
290302	290302	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
341720	341720	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
411091	411091	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
446699	446699	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
446712	446712	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
447439	447439	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
451790	451790	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
455290	455290	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
568932	568932	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
567790	567790	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
574499	574499	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
596239	596239	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
622580	622580	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
631579	631579	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
663791	663791	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
671548	671548	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
745094	745094	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
745259	745259	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
765084	765084	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
772505	772505	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
783887	783887	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
867647	867647	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
882354	882354	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
935030	935030	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
946159	946159	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
964490	964490	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
967508	967508	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
973328	973328	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
983851	983851	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
984011	984011	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
988389	988389	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1032194	1032194	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1034033	1034033	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1040193	1040193	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1044498	1044498	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1051913	1051913	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1102174	1102174	N	N	N	N	N	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1201887	1201887	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1204131	1204131	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1256501	1256501	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1297062	1297062	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1303193	1303193	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1328350	1328350	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1338535	1338535	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1358986	1358986	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1419496	1419496	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1437064	1437064	G	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1446217	1446217	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1448769	1448769	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1460081	1460081	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1505762	1505762	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1528212	1528212	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1533392	1533392	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1547235	1547235	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1548664	1548664	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1559359	1559359	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1566586	1566586	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1627719	1627719	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1756810	1756810	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1756336	1756336	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1822248	1822248	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1891391	1891391	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

70649 C	65361 C	52924 C	44788 C	44187 C	42303 C	35344 C	22948 C	22702 C	20252 C	7791 C	6103 C	4817 C	4618 C	4089 C	1166 C	650283 G	623917 A	598609 G	562139 N	561458 A	556160 C	548190 A	525085 A	525081 A	518401 A	480365 A	471837 A	469310 A	444189 A	444137 A	433807 A	420550 A	399470 A	365999 A	357515 A	346746 A	330728 A	330584 A	318171 A	271356 A	262988 N	258899 C	256293 A	256238 A	254112 A	210272 A	205098 A	194600 A	185893 A	183377 A	155259 A	084147 A	004885 C	981864 A	976754 A	938510 A	935642 A	933779 A	929620 A	927008 A	916122 A	911329 A	869550 C	860819 A	843152 N	833893 C	824209 N	805148 C	796789 A	789780 A	783583 A	768583 C	750446 A	739401 A	703484 A	680088 A	631152 A	622626 N	622012 A	607609 A	607537 A	602828 A	602775 A	596834 G	595520 A	591858 G	587339 A	566110 A	551475 G	545064 A	543738 A	538123 A	526369 A	524194 A	512606 A	507316 A	501556 A	478224 A	470222 G	465745 A	457559 A	426560 A	421553 A	401645 A	392929 C
C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	G	A	G	N	A	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	N	C	A	A	A	A	A	A	A	A	A	A	C	A	A	A	A	A	A	A	A	A	C	A	N	C	N	C	A	A	A	C	A	A	A	A	A	N	A	A	A	A	A	G	A	G	A	A	G	A	A	A	N	A	A	A	A	A	G	A	N	A	A	A	C
C	C	C	C	C	C	C	C	C	C	C	C	C	N	C	C	G	A	G	N	A	C	A	A	A	A	A	A	A	N	A	A	A	A	A	A	A	A	A	A	A	N	C	N	A	A	A	A	A	A	A	A	A	C	A	A	A	A	A	A	A	A	A	C	A	N	C	N	C	A	A	A	C	A	A	A	A	N	N	A	A	A	A	A	G	A	G	A	A	G	A	A	A	N	A	A	A	A	A	G	A	N	A	A	A	C
C	C	N	N	C	C	C	C	N	C	C	C	C	N	C	C	G	A	G	N	N	C	A	N	N	A	A	A	A	A	A	A	A	A	N	A	A	A	A	A	A	N	C	N	A	A	A	A	A	A	A	A	A	C	A	A	A	A	A	A	N	N	A	C	A	N	C	N	C	A	N	A	C	A	A	A	A	A	N	A	A	A	A	A	G	A	G	A	N	G	A	A	N	N	N	N	N	A	A	G	N	N	A	A	A	C
C	C	N	N	N	N	N	C	N	C	C	C	C	N	N	C	G	A	G	N	N	N	N	N	N	A	A	N	N	N	N	A	A	A	N	A	N	A	A	A	A	N	C	N	N	A	A	N	A	A	A	A	N	C	A	N	N	N	A	A	A	N	A	C	N	N	C	N	N	A	N	N	C	N	A	A	A	N	N	N	A	A	A	N	G	N	N	A	N	G	N	A	N	N	N	N	N	A	N	G	N	N	A	A	N	C
C	C	C	N	C	N	C	C	N	C	C	C	C	N	C	C	G	A	G	N	N	N	N	N	N	A	A	A	A	N	N	A	A	A	N	A	N	A	A	A	A	N	C	N	A	A	A	A	N	A	A	A	N	C	A	N	N	N	A	A	N	N	N	C	N	N	C	N	N	N	N	A	C	N	N	A	A	N	N	N	A	A	A	A	G	N	N	A	N	G	N	A	N	N	N	A	N	A	A	G	N	N	A	A	N	C
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	c	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
c	a	c	c	c	c	c	a	t	t	c	c	c	c	c	c	g	a	g	a	a	c	g	a	a	a	g	a	c	a	c	a	t	g	a	g	a	a	a	g	t	t	c	a	a	a	a	a	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	g	a	g	c	a	a	t	a	g	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	c	g	a	a	g	g	a	c
c	c	c	c	c	c	c	a	t	c	c	c	c	c	c	c	g	a	g	a	a	c	g	a	a	a	a	a	a	c	c	a	a	g	a	g	a	g	a	a	a	t	c	a	a	a	g	a	a	c	a	g	a	c	a	a	a	a	g	g	a	a	a	c	a	a	c	a	c	g	a	a	c	a	a	a	a	g	a	a	a	g	a	a	g	a	g	a	a	g	a	c	a	a	a	a	a	a	c	g	g	a	g	g	c	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	t	g	a	g	a	a	c	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	g	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	g	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	c	a	g	a	a	a	a	a	c
t	c	c	c	c	c	c	a	t	c	c	t	c	c	c	c	g	g	g	a	t	c	g	a	a	a	a	a	a	a	c	a	a	g	a	g	a	g	c	a	a	t	c	a	a	c	g	a	a	c	a	g	a	c	a	a	a	a	g	g	a	a	a	c	g	g	c	g	c	g	a	a	c	a	a	a	c	g	a	a	a	g	g	a	g	a	g	c	a	g	a	c	a	a	a	a	a	a	c	g	g	a	g	g	c	c
c	c	c	c	c	c	c	a	t	c	c	c	c	c	c	c	g	a	g	g	a	c	g	a	a	a	a	a	a	c	c	a	a	g	a	g	a	g	a	a	a	t	c	a	a	a	g	a	a	c	a	g	a	c	a	a	a	a	g	g	a	a	a	c	a	a	c	a	c	g	a	a	c	a	a	a	a	g	a	a	a	g	a	a	g	a	g	a	a	g	a	c	a	a	a	a	a	a	c	g	g	a	g	g	c	c
c	c	a	a	t	c	t	a	t	c	t	c	t	t	c	c	g	a	g	g	a	c	g	a	c	a	a	g	a	a	a	c	a	-	g	g	g	c	a	a	a	t	c	g	a	a	a	g	a	c	g	a	g	-	a	t	t	g	a	g	a	g	g	c	a	a	c	a	c	g	a	a	c	g	a	a	a	g	g	c	a	a	a	a	g	c	g	a	g	g	g	a	-	t	g	t	g	a	c	a	a	t	g	g	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	-	-	-	-	-	-	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	t	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	a	a	a	a	a	a	g	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	-	g	a	a	c	a	a	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	-	-	g	a	-	a	a	c	a	-	a	a	a	a	a	a	a	c	a	a	c	a	c	-	a	a	-	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	t	g	a	g	a	a	c	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	g	a	a	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	g	a	a	g	g	a	g	a	a	g	a	a	a	a	a	a	a	c	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	t	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	a
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	a	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	a	a	a	a	g	a	a	-	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	t	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	t	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	g	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	a	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c
c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	g	a	g	a	a	c	a	a	a	g	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	t	c	a	a	a	a	a	g	a	a	a	a	c	a	a	a	a	a	a	a	a	a	c	a	a	c	a	c	a	a	a	c	a	a	a	a	a	a	a	a	a	a	a	g	a	g	a	a	g	a	a	a	a	a	a	a	a	a	g	a	a	a	a	a	c









701056 G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g
719950 G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
727192 G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
738924 G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
765243 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
765693 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
777016 G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
782297 G	G	G	G	G	G	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	a	g	g
792732 G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
861725 G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
863306 G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	-	g	g	g	g	g	g	g	g	g
872058 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
873426 G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
883504 G	N	N	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
890445 G	G	G	G	G	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
897150 C	C	N	N	N	N	c	g	c	c	c	c	c	c	c	c	c	c	c	c	g	c	c	c	c	c	c
899269 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
903195 G	G	G	G	N	N	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
910389 G	G	G	G	N	N	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g
914103 G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
915269 G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
919892 G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
929487 G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
931248 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
935991 G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
955231 G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
957890 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
958635 G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
965291 G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	a	g	g	a	g	g	g	g	g	g	g
974414 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
981896 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
989455 G	G	G	N	N	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
990954 G	G	G	N	N	N	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1025278 G	G	N	G	N	G	t	g	g	g	g	g	g	g	g	g	t	t	g	t	g	g	g	t	t	g	t
1028008 G	G	G	N	N	N	g	g	g	g	g	a	g	g	g	g	g	-	a	g	g	g	g	g	g	g	g
1044141 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1050995 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1055016 G	G	G	G	G	G	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g
1092473 G	G	G	G	G	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1187931 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1210145 G	G	G	G	G	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1239570 G	G	G	G	N	N	g	g	g	-	g	g	g	-	t	g	g	g	g	g	g	g	g	g	g	g	g
1254127 G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1256749 G	G	G	N	G	G	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1263337 G	G	G	G	G	G	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
1298319 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1301844 G	G	G	G	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1302316 G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1326124 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1348621 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1349263 G	G	G	G	G	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1364621 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1381808 G	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1387745 N	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1392783 G	G	G	G	N	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1398636 G	G	N	G	N	G	g	g	g	a	g	g	a	-	g	g	g	g	g	g	g	g	g	g	g	g	g
1412465 G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1433676 T	T	T	T	T	T	t	g	t	t	t	t	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t
1443460 G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	-	g	g	g	g	a	g	g	g	g
1462604 G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	a	g	g	a	g	g	g	a	g	g	a
1487807 G	G	G	G	N	N	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1489855 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1496194 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1512930 G	G	G	N	N	N	a	g	g	g	g	a	g	g	g	a	a	a	a	a	g	g	a	a	a	g	a
1522213 G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1534159 G	G	G	N	N	N	g	g	g	g	g	g	g	t	t	g	g	g	g	g	g	g	g	g	g	g	g
1535853 G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1537228 G	G	G	G	G	G	g	g	g	a	g	g	a	a	g	g	g	g	g	g	g	g	g	g	g	g	g
1546560 G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1555356 G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1559993 G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g
1569280 G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1575210 G	G	G	G	G	N	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1575831 G	G	G	G	G	G	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g
1576311 G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1583240 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1586298 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1605152 G	G	G	G	N	G	g	g	g	g	g	g	g	g	g	g	a	g	g	a	g	g	g	g	g	g	g
1636065 G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1639211 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1654559 G	G	G	N	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1685196 G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1687397 G	G	G	G	N	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1688895 G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1696254 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1700123 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1700711 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1701019 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1702503 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1702674 G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1702751 G	G	G	N	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1707708 G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1710181 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1710942 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1721070 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1740387 G	G	G	G	G	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1741277 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1743610 G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1748342 G	G	G	G	N	N	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1749278 G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1752134 G	G	G	N	N	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1756399 G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1756795 G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1758561 G	G	G	N	G	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1792125 G	G	G	G	N	N	g	g	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g
1793892 G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g

1801784	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
1802278	G	G	G	N	N	N	g	g	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g
1809133	A	A	A	A	N	N	a	g	a	a	-	-	a	a	a	-	-	-	-	-	-	-	-	-	a	-	a
1813550	G	G	G	G	N	N	g	g	g	g		g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1821484	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	-	g	g	g	g	g	g	g	g	-
1836912	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1846478	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1867878	G	G	G	G	N	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1871363	N	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1893755	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1923870	G	G	G	G	N	N	g	g	-	g	g	t	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g
1926334	G	G	G	N	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1933069	G	G	G	G	G	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
1950598	G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1958106	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1961366	N	N	N	N	N	N	g	g	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g
1965511	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1965682	G	G	G	G	G	G	g	g	g	t	g	g	t	t	t	g	g	g	g	g	g	g	g	g	g	g	g
1966917	G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1969011	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1969561	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
1977401	G	G	G	G	G	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
1984202	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2000216	N	N	N	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2014097	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2014796	G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2016824	G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2018367	G	G	G	G	N	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
2025192	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g
2063793	T	T	T	N	T	N	t	g	t	t	g	t	-	t	-	t	t	t	t	t	g	t	t	t	t	g	t
2067815	G	G	G	G	N	N	g	g	g	g	g	g	-	g	-	g	g	g	a	g	g	g	g	g	g	g	g
2112974	G	G	G	G	N	N	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
2113701	G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2131090	G	G	G	G	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2250633	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2262577	G	G	G	G	N	T	t	g	g	g	g	t	g	g	g	t	t	t	t	t	g	g	t	t	t	g	t
2263648	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2266535	G	G	G	G	G	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2278013	G	G	G	G	N	N	g	g	-	-	g	g	g	-	a	g	g	g	g	g	g	g	g	g	g	g	g
2278317	G	G	G	N	G	N	a	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g	g	g	g	g
2278777	G	G	G	G	N	G	g	g	a	-	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2279352	G	G	G	G	N	G	g	g	g	-	g	g	a	-	g	g	g	g	g	g	g	g	g	g	g	g	g
2299160	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
2300659	G	G	G	G	G	G	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2321103	G	G	G	N	N	N	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g
2332600	G	G	G	G	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2343925	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2432007	G	G	G	G	G	G	g	g	g	t	g	g	t	t	g	g	g	g	g	g	g	g	g	g	g	g	g
2445134	G	G	G	G	G	G	g	g	t	g	g	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g
2466459	N	N	N	N	N	N	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2468339	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2480200	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2480736	G	G	G	G	N	N	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g
2481985	G	G	G	G	G	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2482482	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2519192	N	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2528093	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2537024	G	G	G	G	N	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2541447	G	G	G	G	N	G	g	g	g	-	g	g	g	-	a	g	g	g	g	g	g	g	g	g	g	g	g
2548214	G	G	G	N	N	N	g	g	g	g	g	g	g	g	-	g	g	g	g	g	g	a	g	g	g	g	g
2569996	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2574796	G	G	G	G	N	G	g	g	g	g	g	g	g	a	t	g	g	g	g	g	g	g	g	g	g	g	g
2576590	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2576788	C	C	C	N	N	N	c	g	c	c	c	c	c	c	c	c	c	c	c	c	g	c	c	c	c	c	c
2577686	G	G	G	G	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
2592207	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2598796	G	G	N	N	N	N	g	g	g	g	-	-	g	g	t	-	g	g	-	g	g	g	g	g	g	-	g
2607374	G	G	G	G	N	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2642504	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2651733	G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2663285	G	G	G	G	G	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2677598	G	G	G	G	N	G	g	g	g	a	g	g	a	a	g	g	g	g	g	g	g	g	g	g	g	g	g
2678365	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2679056	G	G	G	G	N	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2684462	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2684662	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2684793	G	G	G	G	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
2685390	G	G	G	G	G	G	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2685591	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2697434	G	G	N	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2721618	G	G	G	N	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2722524	G	G	G	G	G	G	g	g	g	a	g	g	a	a	g	g	g	g	g	g	g	g	g	g	g	g	g
2728710	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2744933	G	G	G	G	G	G	a	g	g	g	g	g	g	g	g	g	a	-	g	a	g	g	g	a	a	g	a
2748467	G	G	G	N	N	N	g	g	g	g	g	g	t	g	g	g	g	g	-	g	g	-	g	g	g	g	g
2749358	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2751086	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g
2751771	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
2755510	N	N	N	N	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2766007	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2767376	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
2768594	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g
2768892	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2779028	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2797762	G	G	G	N	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2808649	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2812384	G	G	G	G	G	G	g	g	t	t	g	g	t	t	t	g	g	g	g	g	g	t	g	g	g	g	g
2824130	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2828810	G	G	G	G	G	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2828968	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2830100	G	G	G	N	N	N	g	g	g	a	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2830646	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2835409	G	G	N	N	N	N	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
2836077	G	G	G	G	G	G	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
2839659	G	G	G	N	N	N	g	g	g	g	g	g	g	a	g	g	g	g	a	g	g	g	g	g	g	g	g
2858983	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g

2865641	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
2868665	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2871320	G	G	G	N	N	N	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g
2873293	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2889154	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2896636	G	G	G	N	N	A	a	g	g	g	g	a	g	g	g	a	a	a	a	a	g	g	a	a	a	g	a
2898374	G	G	G	G	G	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2899068	G	G	G	G	N	G	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g
2902270	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2903882	G	G	G	G	G	N	t	g	g	g	g	g	g	g	g	g	t	t	g	t	g	g	g	t	t	g	t
2920611	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2924187	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
2933957	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2934653	G	G	G	G	G	G	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2934972	G	G	G	N	N	N	c	g	g	g	g	c	g	g	g	c	c	c	c	c	g	g	c	c	c	g	c
2937916	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2938178	G	G	N	N	N	N	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g
2953232	G	G	N	G	N	N	g	g	g	g	g	g	g	g	t	g	g	-	g	g	g	g	g	g	g	g	g
2958797	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2967943	G	G	G	G	G	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2974551	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2977732	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
2979662	G	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
2989125	G	G	G	N	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3002442	G	G	G	G	G	G	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3007357	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3011797	G	G	G	G	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3019946	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3024294	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3026456	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3033049	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3045663	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3060587	N	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3075828	G	G	G	N	N	N	g	g	g	g	g	g	g	g	c	g	g	-	g	g	g	g	g	g	g	g	g
3076716	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3084869	G	G	G	N	N	N	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3085079	G	N	N	N	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3089821	T	T	T	T	N	N	t	g	t	t	t	t	t	t	t	t	t	-	t	t	t	t	t	t	t	t	t
3090589	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3096319	G	G	G	N	N	N	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3096883	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3099378	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
3102875	G	G	G	G	G	G	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
3122352	G	G	N	N	N	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3122477	G	G	G	N	N	N	g	g	g	g	g	g	a	-	g	g	g	g	g	g	g	g	g	g	g	g	g
3128007	G	G	G	N	N	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3137262	T	T	T	T	N	T	t	g	t	t	t	t	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t
3140212	G	G	G	G	G	G	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
3150735	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3159744	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3166303	G	G	G	G	N	G	g	g	g	g	g	g	g	a	g	g	g	g	a	g	g	g	g	g	g	g	g
3166562	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
3173575	G	G	G	G	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3190399	G	G	G	G	G	N	a	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g
3194193	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3209579	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3214425	A	A	A	N	N	N	a	g	a	a	a	a	a	a	a	a	a	-	a	a	a	a	a	a	a	a	a
3217391	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3225479	G	G	G	G	N	N	g	g	t	g	g	g	t	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3225497	G	G	G	N	N	N	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3225501	G	G	G	N	N	N	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3225506	G	G	G	N	N	N	g	g	c	c	g	g	c	t	c	g	g	g	g	g	g	g	g	g	g	g	g
3225509	G	G	G	N	N	N	g	g	c	c	g	g	c	c	c	g	g	g	g	g	g	g	g	g	g	g	g
3225513	G	G	G	N	N	N	g	g	c	c	g	g	c	c	c	g	g	g	g	g	g	g	g	g	g	g	g
3225521	G	G	N	N	N	N	g	g	c	c	g	g	c	c	c	g	g	g	g	g	g	g	g	g	g	g	g
3225522	G	G	G	N	N	N	g	g	a	a	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3225528	G	G	G	N	N	N	g	g	t	t	g	g	t	t	t	g	g	g	g	g	g	g	g	g	g	g	g
3225554	G	G	G	N	N	N	g	g	a	g	g	g	a	a	a	g	g	g	g	g	g	g	g	g	g	g	g
3228082	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3244204	G	G	G	N	G	G	a	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g
3248389	G	G	G	G	G	G	g	g	g	g	g	g	g	g	t	g	-	-	g	g	g	g	g	g	g	g	g
3250836	G	G	N	N	N	N	g	g	g	g	a	g	g	g	g	g	g	-	g	g	g	g	g	g	g	a	g
3267118	G	G	G	N	N	N	a	g	g	g	g	a	g	g	g	a	a	a	a	a	g	g	a	a	a	g	a
3291060	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3291254	G	G	G	G	N	G	g	g	g	g	g	g	g	g	g	g	t	g	g	t	g	g	g	t	g	g	t
3295047	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3296657	G	G	G	G	N	N	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	t	g
3297934	G	G	G	G	G	G	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3299204	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3301818	G	G	G	G	G	G	g	g	g	g	g	g	c	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3324959	G	G	G	G	N	G	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3327979	C	C	C	C	C	C	c	g	c	c	c	c	c	c	c	c	c	c	c	c	g	c	c	c	c	c	c
3341360	G	G	G	G	G	G	g	g	t	t	g	g	t	t	t	g	g	-	g	g	g	g	g	g	g	g	g
3357379	C	C	C	C	C	N	c	g	c	c	g	c	c	c	c	c	c	c	c	c	g	c	c	c	c	g	c
3360868	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3361311	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3362591	G	G	G	G	G	G	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3363364	G	G	G	G	G	G	g	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g	t	g	g	g
3363946	G	G	G	G	G	G	g	g	g	g	g	g	t	-	g	g	g	g	g	g	g	g	g	g	g	g	g
3366169	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3369097	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g
3370950	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3374754	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3376387	A	A	A	A	N	N	a	g	a	a	g	a	a	a	a	a	a	-	a	a	g	a	a	a	a	g	a
3388304	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3395095	G	G	G	G	N	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3395268	G	G	G	G	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3397040	G	G	G	N	N	N	a	g	g	g	g	g	g	g	g	a	a	a	g	a	g	g	a	a	a	g	a
3398946	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3409415	G	G	G	N	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3409454	G	G	G	G	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3414433	G	G	G	G	G	N	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g
3421335	G	G	G	G	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3431905	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3446841	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3454765	T	T	T	T	N	T	t	g	t	t	g	t	t	t	t	t	t	t	t	t	g	t	t	t	t	g	t

3484993	G	G	G	G	G	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3497425	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3502610	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3519841	G	G	G	N	N	G	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3532033	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3535454	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3579994	G	G	G	G	G	G	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
3586778	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g
3589425	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3594885	G	N	N	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3602127	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3605437	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3610750	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3616733	G	G	G	N	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3620064	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3630645	G	G	G	N	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3635257	G	G	G	N	N	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3635663	G	G	G	G	G	G	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3645151	G	G	G	N	N	N	c	g	g	g	g	g	g	g	g	c	c	c	g	c	g	g	c	c	c	g	c
3655814	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3656171	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3659175	G	G	G	G	N	G	g	g	t	t	g	g	t	t	t	g	g	g	g	g	g	g	g	g	g	g	g
3660617	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3667806	G	N	N	N	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	a	a	a	g	a
3668322	N	N	N	N	N	N	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g
3673226	G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3698745	G	G	G	G	N	N	g	g	g	g	g	g	a	g	a	g	g	g	a	g	g	g	g	g	g	g	g
3699234	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3705489	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3715815	T	T	T	T	N	T	t	g	t	t	t	t	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t
3717147	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3717777	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3728411	G	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3732919	G	G	G	G	N	N	a	g	g	g	g	g	g	g	g	g	a	a	g	a	g	g	g	a	a	g	a
3739257	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3756048	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3762390	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3769752	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3778284	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	-	a	g	g	g	g	g	g	g	g
3789603	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3789691	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3791153	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	-	g	g	g	g	g	g	g	g	g
3794907	G	G	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3802012	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3830055	G	G	G	G	G	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3844040	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	a	-	g	a	g	g	g	g	g	g	g
3854440	G	G	G	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
3870393	G	G	G	N	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3887319	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3897676	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	-	g	a	g	g	g	g	g	g	g
3898451	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g
3900188	G	G	G	G	G	G	g	g	g	g	a	g	g	g	g	g	g	-	g	g	g	g	g	g	g	a	g
3913056	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	-	g	g	g	g	g	g	g	g	g
3936625	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3940629	T	T	T	T	T	T	t	g	t	t	t	t	t	t	t	t	t	-	t	t	g	t	t	t	t	t	t
3946037	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	-	g	g	g	g	g	g	g	g	g
3955281	N	N	N	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3957761	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3958669	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3962272	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3964074	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3966465	G	G	G	G	G	G	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
3972375	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3977650	G	G	G	G	G	G	g	g	g	g	g	c	g	g	g	g	g	-	c	g	g	g	g	g	g	g	g
3991709	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
3993496	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
3994985	G	G	G	G	G	N	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4037937	G	G	N	N	N	N	g	g	t	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g	g	g	g
4053446	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4060082	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
4083536	G	G	G	N	G	N	a	g	g	g	g	g	g	g	g	a	a	a	g	a	g	g	a	a	a	g	a
4120633	G	G	G	G	N	N	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g
4123916	G	G	G	N	N	N	g	g	g	g	g	t	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g
4147421	G	G	G	G	G	G	g	g	g	g	g	t	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g
4148980	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4159123	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4159411	G	G	N	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4164195	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4171391	G	G	G	N	N	G	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g
4183938	G	G	G	G	N	N	g	g	g	g	g	g	c	g	g	g	g	-	g	g	g	g	g	g	g	g	g
4185559	G	G	G	G	G	G	g	g	g	g	g	g	a	a	g	g	g	g	g	g	g	g	g	g	g	g	g
4190717	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4208049	G	G	G	G	N	G	g	g	g	g	g	g	g	g	c	g	g	-	g	g	g	g	g	g	g	g	g
4210181	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	-	g	g	g	g	g	g	g	g	g
4239328	G	G	G	G	G	G	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g
4268134	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
4269446	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4288438	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4293068	G	G	G	G	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
4305301	G	G	G	G	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4306829	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4309661	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4311175	G	G	G	G	N	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4312299	G	G	G	G	N	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4319081	G	G	G	N	N	N	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4320553	G	G	G	G	N	G	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
4332632	G	G	G	G	G	G	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4343118	G	G	G	G	G	G	g	g	g	g	g	a	g	g	g	g	g	-	a	g	g	g	g	g	g	g	g
4343766	G	G	G	G	G	G	g	g	g	g	g	g	t	g	g	g	g	-	g	g	g	g	g	g	g	g	g
4352178	N	N	N	N	N	N	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g
4371886	G	G	G	N	N	G	a	g	g	g	g	g	g	g	g	a	a	a	g	a	g	g	a	a	a	g	a
4393265	G	G	G	G	G	G	g	g	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g
4400732	G	G	G	G	G	N	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	a	g	g	g	g	g
4416198	G	G	G	G	G	G	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4421689	G	G	G	G	G	G	a	g	g	g	g	g	g	g	t	a	a	-	g	a	g	g	a	a	a	g	a
4426771	G	G	G	G	N	N	g	g	g	g	g	g	a	g	g	g	g	g	g	g	g	g	g	g	g	g	g

1359942 T	T	T	T	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1374526 T	T	T	T	N	N	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1387768 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1388315 T	T	T	T	T	T	t	t	a	a	t	t	a	a	a	t	t	t	t	t	t	t	t	t	t	t	t
1389352 T	T	T	T	N	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1411175 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	t	-	t	t	t	t	a	t	t	t	t
1425412 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1426775 T	T	T	T	N	T	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1437533 T	T	N	N	N	N	t	t	t	a	t	t	a	a	t	t	t	t	t	t	t	t	t	t	t	t	t
1440842 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	-	t	t	t	t	t	t	t	t	t
1450348 T	T	T	T	N	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1452142 T	T	T	T	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1453663 T	T	T	T	T	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1455700 T	T	T	N	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1470653 T	T	T	N	N	T	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1487372 G	G	G	G	N	G	g	t	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g
1487493 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1505789 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	t	-	t	t	t	t	t	t	c	t	t
1523715 T	T	T	T	T	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1541862 T	T	T	T	N	T	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1545694 T	T	T	T	N	T	t	t	t	t	t	t	t	t	a	t	t	-	t	t	t	t	t	t	t	t	t
1546885 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1566978 T	T	T	T	N	T	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1568321 T	T	T	T	T	N	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1569438 T	T	T	T	T	T	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1573651 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
1586141 C	C	C	C	C	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
1629155 T	T	N	N	N	N	t	t	t	t	t	t	t	t	c	t	t	-	t	t	t	t	t	t	t	t	t
1632446 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t
1632565 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1637512 T	T	T	T	T	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1673569 G	G	G	G	G	G	g	t	g	t	g	g	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g
1714258 T	T	T	T	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1729368 T	T	T	T	N	N	t	t	t	t	t	t	t	t	g	t	t	-	t	t	t	t	t	t	t	t	t
1748616 T	T	T	T	N	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
1755324 T	T	T	T	T	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1760302 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	g	t	t	t	t	t
1771880 T	T	T	N	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1773435 T	T	T	T	N	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1777473 T	T	T	T	T	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1783007 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t
1791460 T	N	N	N	N	N	t	t	c	c	t	t	t	c	t	t	t	-	t	t	t	t	t	t	t	t	t
1796044 T	T	T	T	N	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1832390 T	T	T	T	N	N	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
1837282 T	T	T	T	T	T	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
1859946 T	T	T	T	N	N	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1861680 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
1875101 T	T	T	T	T	N	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1877285 T	T	T	T	T	T	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
1911655 T	T	T	T	N	T	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1917295 T	T	T	T	N	T	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1923495 T	T	T	N	N	N	t	t	-	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1942932 T	T	T	N	T	N	t	t	g	t	t	t	t	t	t	t	t	-	t	t	t	t	t	t	t	t	-
1946881 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
1956130 T	T	T	T	N	N	t	t	c	c	t	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t
1963493 T	T	N	N	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
1987554 T	T	T	N	N	N	t	t	t	t	c	t	t	t	t	t	t	-	t	t	t	t	t	t	t	c	t
2002713 T	T	N	N	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2010107 T	T	T	T	T	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2012524 T	T	T	T	N	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
2025677 G	G	G	G	G	G	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2077107 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t
2117591 N	N	N	N	N	N	c	t	c	-	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
2132450 T	T	T	T	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2242931 C	C	C	N	N	N	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
2253174 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2255614 T	T	T	T	N	N	t	t	-	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2261325 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t
2267519 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2269099 G	G	G	N	N	N	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
2277197 T	T	N	T	N	N	t	t	t	t	t	t	t	-	g	t	t	t	t	t	t	t	t	t	t	t	t
2293322 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	-	t	t	-	t	c	t	-	t	t	t
2309711 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2310319 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2310626 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
2341346 T	T	T	T	N	T	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	g	t	t	t	t	t
2355850 T	T	T	T	N	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
2410185 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	-	t	t	t	t	t	t	t	t	t
2414771 T	T	T	T	N	T	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2416435 T	T	T	T	T	T	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2431250 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
2437858 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2439992 N	N	N	N	N	N	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2459541 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	-	t	t	t	t	t	t	t	t	t
2467734 T	T	T	T	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2485389 T	T	T	T	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2493895 C	N	N	N	N	N	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
2548551 T	T	T	T	N	N	g	t	t	t	t	t	t	t	-	g	g	g	t	g	t	t	g	g	g	t	g
2551008 G	G	G	N	N	N	g	t	g	g	g	g	g	g	-	g	g	g	g	g	t	g	g	g	g	g	g
2577752 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2594754 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2597523 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2597543 T	T	T	N	N	N	t	t	t	-	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2605170 T	T	T	N	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
2640081 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2656129 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	g	t	t	t	t	t
2667041 T	T	T	T	T	T	t	t	t	t	t	g	t	t	t	t	t	-	g	t	t	t	t	t	t	t	t
2670390 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	-	t	t	t	t	t	t	t	t	t
2681696 T	T	T	T	T	T	t	t	t	t	t	c	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t
2691491 T	T	N	N	N	T	t	t	t	t	t	t	t	t	c	t	-	-	t	-	t	t	t	t	t	t	t
2724357 T	T	T	T	N	N	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2728232 T	T	T	T	T	T	t	t	t	g	t	t	g	g	t	t	t	-	t	t	t	t	t	t	t	t	t
2732999 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	-	t	t	t	t	t	t	t	t	t
2738714 N	N	N	N	N	N	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2741382 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
2754024 T	T	N	N	N	N	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t

2757750 T	T	T	T	N	T	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
2764996 T	T	T	T	T	T	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t	g	t
2799903 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	c	t	t	t
2800065 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2803529 T	T	T	T	T	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2805612 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
2823630 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	c	t	t	t	t	t
2840617 N	N	N	N	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2843592 T	T	T	T	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
2847692 C	C	C	C	C	C	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
2858689 T	T	T	T	N	N	t	t	t	t	t	t	t	t	t	t	t	-	t	t	t	t	t	c	t	t	t
2876328 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2882454 T	T	T	T	T	T	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t	c	t
2902125 C	C	C	C	C	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
2913038 T	T	T	T	T	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
2923498 T	T	T	N	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
2934864 C	C	C	C	C	C	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
2943162 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
2945027 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
2959217 T	T	T	T	N	N	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
2968017 C	C	C	C	C	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
2973614 T	T	T	T	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
2976049 T	T	T	N	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
3023974 T	T	T	T	T	T	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3024220 T	T	T	T	T	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3025936 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
3026489 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3030163 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
3052609 T	T	T	T	T	T	t	t	a	a	t	t	a	a	a	t	t	t	t	t	t	t	t	t	t	t	t
3071710 T	T	T	T	T	T	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3074598 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
3077177 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	a	t	t	t
3085186 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3101838 T	T	T	T	T	T	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t	g	t
3105510 T	T	T	N	N	N	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	a	t	t	t
3108649 T	T	T	N	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
3111131 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3113533 T	T	N	N	N	N	t	t	t	t	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t
3124232 T	T	T	T	T	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3144072 T	T	T	T	T	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3215984 T	T	T	T	N	N	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3239590 T	T	T	T	T	T	t	t	t	t	t	t	a	t	t	t	t	-	t	t	t	t	t	t	t	t	t
3248217 T	T	T	T	T	N	t	t	t	t	t	t	t	t	c	t	-	t	t	t	t	t	t	t	t	t	t
3314013 G	G	G	G	G	G	g	t	g	g	t	g	g	g	g	g	g	g	g	g	t	g	g	g	g	t	g
3342989 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	-	c	c	c	c	c	c	c	c	c
3343029 T	T	T	T	T	T	t	t	c	t	t	t	t	t	t	t	t	-	t	t	t	t	t	t	t	t	t
3371630 T	T	T	T	T	T	t	t	t	t	t	c	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t
3393976 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
3403449 N	N	N	N	N	N	t	t	-	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3405483 T	T	T	T	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3407398 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
3415081 T	T	T	T	T	T	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3442617 T	T	T	T	T	N	a	t	t	t	t	t	t	t	t	t	a	a	t	a	t	t	a	a	a	t	a
3442988 T	T	T	T	N	N	t	t	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t
3493530 C	C	C	C	N	C	c	t	c	c	c	c	c	c	c	c	c	-	c	c	c	c	c	c	c	c	c
3539709 C	C	C	C	C	C	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
3564026 T	T	T	T	T	T	c	t	t	t	t	t	t	t	t	t	c	c	t	c	t	t	t	c	c	t	c
3566853 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3596495 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3634102 T	T	T	N	N	N	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	a	t	t	t
3636012 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
3639873 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
3642221 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
3658233 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
3688024 T	T	T	T	T	T	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
3701692 N	N	N	N	N	N	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
3712455 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
3728410 T	N	N	N	N	N	t	t	t	t	t	t	c	c	t	t	t	-	t	t	t	t	t	t	t	t	t
3771852 T	T	T	N	N	N	t	t	t	t	t	a	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t
3784042 T	T	T	T	T	T	t	t	t	t	t	a	t	t	t	t	t	t	a	t	t	t	t	t	t	t	t
3806677 T	T	T	N	N	N	c	t	t	t	t	c	t	t	t	c	c	c	c	c	t	t	c	c	c	t	c
3949188 T	T	T	T	T	T	t	t	t	t	t	t	c	t	t	t	t	-	t	t	t	t	t	t	t	t	t
3973746 T	T	T	N	N	N	c	t	t	t	t	c	t	t	t	c	c	c	c	c	t	t	c	c	c	t	c
3986268 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
4045281 T	T	T	T	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
4053034 N	N	N	N	N	N	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
4067118 N	N	N	N	N	N	c	t	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
4080579 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4081612 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4087224 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4103232 T	T	T	T	T	N	t	t	t	t	t	-	t	t	c	t	-	-	-	-	t	-	-	-	t	t	t
4133113 T	T	T	T	N	T	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
4152231 T	T	T	T	T	T	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t	t	t
4154108 T	N	N	N	N	N	t	t	t	t	t	t	t	t	g	t	t	t	t	t	t	t	t	t	t	t	t
4160733 T	T	T	T	N	N	t	t	t	t	t	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4166491 T	T	T	T	T	T	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	c	t	t	t
4201619 T	T	T	T	T	T	t	t	t	t	t	c	t	t	t	t	t	-	c	t	t	t	t	t	t	t	t
4243823 T	T	T	N	N	N	a	t	t	t	t	a	t	t	t	a	a	a	a	a	t	t	a	a	a	t	a
4253544 T	T	T	T	T	T	t	t	t	c	t	t	c	c	t	t	t	t	t	t	t	t	t	t	t	t	t
4277985 T	T	T	T	T	T	t	t	c	t	t	t	t	t	t	t	t	-	t	t	t	t	t	t	t	t	t
4287044 G	G	G	G	N	G	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4311918 C	C	C	C	C	C	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
4318701 T	T	T	T	T	T	t	t	t	t	t	c	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t
4321135 T	T	T	T	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
4339366 T	T	T	T	T	T	t	t	g	g	t	t	g	g	g	t	t	t	t	t	t	t	t	t	t	t	t
4339924 T	T	T	N	N	N	t	t	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t
4352091 T	T	T	N	N	N	t	t	t	t	t	t	t	t	t	t	c	t	t	c	t	t	t	t	t	t	t
4377190 T	T	T	T	T	T	t	t	t	t	a	t	t	t	t	t	t	t	t	t	t	t	t	t	t	a	t
4408510 C	C	C	N	N	N	c	t	c	c	c	c	c	c	c	c	c	c	c	c	t	c	c	c	c	c	c
4433790 G	G	G	G	G	G	g	t	g	g	g	g	g	g	g	g	g	g	g	g	t	g	g	g	g	g	g
4562052 N	N	N	N	N	N	t	t	t	t	t	t	t	t	c	-	t	t	t	t	t	t	t	t	t	t	t
4585561 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4589706 T	T	T	T	T	T	t	t	c	c	t	t	c	c	c	t	t	t	t	t	t	t	t	t	t	t	t
4595001 C	C	C	C	C	C	c	t	c	c	t	c	c	c	c	c	c	c	c	c	t	c	c	c	c	t	c
4630765 G	G	G	G	G	G	g	t	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
4633814 N	N	N	N	N	N	t	t	t	t	t	t	c	t	t	t	t	t	t	t	t	t	t	t	t	t	t

Appendix 6

Table S5 - variant positions in the East Smithfield sequences compared against CO92 (NC_03143.1)

* = position that is fixed in all individuals, though base call for one individual differs from the others

^ = position that is fixed in at least one individual and polymorphic in at least one other

Genes in **bold** indicate potential role in virulence

Position	Known polymorphic site[11]?	CO92	8291			6330			8124			11972			gene
		Base	Base	Coverage	Freq. Variant	Base	Coverage	Freq. Variant	Base	Coverage	Freq. Variant	Base	Coverage	Freq. Variant	
41481	N	C	M	18	0.44							M	14	0.50	YPO0032 conserved membrane protein
308840	N	G	D	5	0.20										Intergenic
314599	N	A	R	9	0.22							R	5	0.40	YPO0309 putative exported protein
350035	N	G	K	7	0.29							K	5	0.80	YPO0342 putative oxidoreductase
386545	N	G	K	10	0.20										YPO0371 DNA mismatch repair protein "mutL", "b4170"
477380	N	C	M	21	0.38							M	6	0.17	YPO0452 putative soluble lytic murein transglycosylase "slt", "sltY"
677870	N	C	M	25	0.56							M	8	0.50	YPO0616 putative glycosyl hydrolase
*699494	Y	A	G	33	1.00	A	8	0.00	G	6	1.00	G	17	1.00	YPO0643 RNA polymerase sigma factor "rpoD"
768365	N	G	K	14	0.36							K	14	0.50	Intergenic
827175	N	A	R	60	0.23				R	11	0.10	R	35	0.20	YPO0765 putative adhesin (pseudogene)
827554	N	G	K	14	0.29							K	6	0.33	YPO0765 putative adhesin (pseudogene)
914282	N	G	K	20	0.35				K	5	0.60	K	11	0.09	YPO0833 putative phosphosugar isomerase
1333333	N	G	K	23	0.43							K	14	0.50	YPO1184 putative ABC transport membrane permease
1398797	N	C	M	7	0.43										Intergenic
1458918	N	G	K	15	0.60	K	9	0.44	K	5	0.40	K	10	0.70	YPO1299 1-phosphofructokinase "fruK"
1543875	N	G	K	8	0.38							K	6	0.33	Intergenic
2024938	N	G	K	18	0.67	K	7	0.50	K	5	0.40	K	12	0.42	YPO1779 putative exported protein
2056058	N	G	K	12	0.50							K	6	0.17	YPO1810 putative transcriptional regulatory protein
2077729	N	C	Y	5	0.80										Intergenic
2077732	N	G	R	6	0.83	Y	23	0.68							Intergenic
2125410	N	C	Y	74	0.58				Y	13	0.92	Y	45	0.58	YPO1884 transposase (pseudogene) "y1062", "ypmt1.80C"
2231022	N	C	M	33	0.21							M	8	0.63	YPO1964 succinylglutamic semialdehyde dehydrogenase "astD"
*2262577	Y	T	G	29	1.00	T	6	0.00				T	9	1.00	YPO1990 hypothetical protein
2290222	N	G	K	21	0.33							K	9	0.11	YPO2017 peptide chain release factor 1 "prfA", "sueB", "uar"

2655987	N	C	Y	7	0.29						5	0.20	YPO2361 putative proton dependent peptide transporter
2797369	N	C	M	6	0.33								**YPO2490 putative haemolysin**
2855269	N	C	M	5	0.60								YPO2543 NADH dehydrogenase I chain N "nuoN"
^2896636	Y	A	R	14	0.86	A	6	0.00					Intergenic
3063732	N	G	K	12	0.25						K		YPO2731 putative membrane protein
3155659	N	C	M	8	0.25						9	0.22	YPO2828 putative phosphoribosylaminoimidazole synthetase "purT", "purM"
^3349970	N	G	G	7	0.00	K	6	0.50	K				**YPO2999 putative HlyD family secretion protein**
3622057	N	G	K	10	0.20				K		5	0.40	K	7	0.43	YPO3248 putative surface-exposed protein (hxuB)
3665978	N	C	M	8	0.50				M		6	0.33				YPO3285) protein (includes: chorismate mutase and prephenate
3855653	N	C	M	11	0.55						5	0.20				dehydrogenase) "tyrA"
3979405	N	G	K	15	0.20											YPO3448 putative extracellular solute-binding protein (pseudogene)
4110561	N	G	K	59	0.39	K	14	0.79	K		6	0.33	K	6	0.17	YPO3566 protease "degQ", "hhoA"
4151776	N	C	M	6	0.33									20	0.30	**YPO3679 insecticidal toxin complex protein TcaB (pseudogene)** **"tcaB"**
4192480	N	C	M	5	0.60											Intergenic
4208402	N	G	K	5	0.40											thiC YPO3739 thiamine biosynthesis protein ThiC
4368681	N	C	M	5	0.60											YPO3751 50S ribosomal protein L11 "relC", "rplK"
4441063	N	A	R	16	0.25											YPO3914 soluble pyridine nucleotide transhydrogenase "sth", "sthA", "udhA"
4504259	N	G	R	5	0.20											**YPO3944 putative invasin**
4511375	N	G	K	7	0.29				K		5	0.20				YPO3996 putative membrane protein "yhjK"
4517080	N	C	M	7	0.29											YPO4002 dipeptide transport system permease protein "dppB"
																YPO4006 putative exported protein

Appendix 7a

Table S6a - Chromosomal differences between Microtus (NC_005810) and the East Smithfield (ES) genome

Microtus Pos	Microtus	Ancestral	ES	Derived	Co92	KIM10	Nepal 516	Microtus	IP32953	Antiqua	NC_009381	NC_009708
									Y. pseudo		Pestoides	
22497	t	T	C	c	c	c	c	t	t	c	t	t
211794	c	C	T	t	t	t	t	c	c	t	c	c
211939	t	T	A	a	a	a	a	t	t	a	t	t
266734	t	A	G	c	c	c	c	t	t	c	t	t
294434	g	C	T	a	a	a	a	g	g	a	g	g
323567	t	A	G	c	c	c	c	t	t	c	t	t
378080	g	C	A	t	t	t	t	g	g	t	g	g
437054	t	T	C	c	c	c	c	t	t	c	c	t
522726	g	G	A	a	a	a	a	g	g	a	g	g
539902	c	C	A	a	a	a	a	c	c	a	a	c
553965	c	C	T	t	t	t	t	c	c	t	c	c
579849	a	A	C	c	c	c	c	a	a	c	a	a
699134	t	A	C	g	g	g	g	t	t	g	t	t
700076	g	C	A	t	t	t	t	g	g	t	g	g
718436	g	C	A	t	t	t	t	g	g	t	g	g
797929	c	G	A	t	t	t	t	c	c	t	c	c
875377	c	G	T	a	a	a	a	c	c	a	c	c
975069	a	T	A	t	t	t	t	a	a	t	a	a
988148	c	G	A	t	t	t	t	c	c	t	c	c
1029862	g	C	T	a	a	a	a	g	g	a	g	g
1099213	t	A	G	c	c	c	c	t	t	c	t	t
1119592	a	T	C	g	g	g	g	a	a	g	a	a
1131883	c	G	A	t	t	t	t	c	c	t	c	c
1160296	c	G	A	t	t	t	t	c	c	t	c	c
1175799	c	G	A	t	t	t	t	c	c	t	c	c
1176227	g	C	G	c	c	c	c	g	g	c	g	g
1290920	g	C	T	a	a	a	a	g	g	a	g	g
1290935	g	C	A	t	t	t	t	g	g	t	g	g
1327376	g	C	T	a	a	a	a	g	g	a	g	g
1329145	c	G	A	t	t	t	t	c	c	t	c	c
1335276	g	C	T	a	a	a	a	g	g	a	g	g
1386605	c	G	A	t	t	t	t	c	c	t	c	c
1423760	c	G	A	t	t	t	t	c	c	t	c	c
1427890	g	C	T	a	a	a	a	g	g	a	g	g
1433228	c	G	A	t	t	t	t	c	c	t	c	c
1433576	c	G	T	a	a	a	a	c	c	a	c	c
1534714	c	C	A	a	a	a	a	c	c	a	c	c
1616932	c	C	T	t	t	t	t	c	c	t	t	c
1621516	c	C	T	t	t	t	t	c	c	t	c	c
1630030	t	T	C	c	c	c	c	t	t	c	t	t
1639545	c	C	A	a	a	a	a	c	c	a	c	c

Microtus Pos	Microtus	Ancestral	ES	Derived	Co92	KIM10	Nepal 516	Microtus	IP32953	Antiqua	NC_009381	NC_009708
									Y. pseudo		Pestoides	
1924644	c	C	T	t	t	t	t	c	c	t	t	c
1956188	c	C	T	t	t	t	t	c	c	t	c	c
1957922	g	G	T	t	t	t	t	g	g	t	g	g
1971540	t	T	C	c	c	c	c	t	t	c	t	t
1973997	t	T	C	-	-	c	c	t	t	-	t	c
2044313	c	C	A	a	a	a	a	c	c	a	c	c
2180376	g	G	T	t	t	t	t	g	g	t	g	g
2222448	a	A	C	c	c	c	c	a	a	c	a	a
2345507	t	T	C	c	c	c	c	t	t	c	t	t
2393482	t	T	C	c	c	c	c	t	t	c	t	t
2394599	g	G	T	t	t	t	t	g	g	t	g	g
2458357	a	A	C	c	c	c	c	a	a	c	a	a
2469528	g	G	T	t	t	t	t	g	g	t	g	g
2494404	g	C	T	a	a	a	a	g	g	a	g	g
2543275	g	G	A	a	a	a	a	g	g	a	g	g
2563610	g	G	A	a	a	a	a	g	g	a	g	g
2564041	g	G	A	a	a	a	a	g	g	a	g	g
2578438	t	T	G	g	g	g	g	t	t	g	t	t
2593928	g	G	A	a	a	a	a	g	g	a	g	g
2600172	a	A	G	g	g	g	g	a	a	g	a	a
2670085	a	T	C	g	g	g	g	a	a	g	a	a
2714773	a	T	A	t	t	t	t	a	a	t	a	a
2743354	g	G	A	a	a	a	a	g	g	a	g	g
2750020	a	A	G	g	g	g	g	a	a	g	a	a
2773531	c	C	T	t	t	t	t	c	c	t	c	c
2819888	c	G	T	a	a	a	a	c	c	a	c	c
2825317	c	G	T	a	a	a	a	c	c	a	c	c
2845969	g	G	C	c	c	c	c	g	g	c	g	g
2866305	t	T	C	c	c	c	c	t	t	c	c	t
2877871	t	T	C	c	c	c	c	t	t	c	t	t
2882536	g	G	A	a	a	a	a	g	g	a	g	g
2899888	t	T	G	g	g	g	g	t	t	g	t	t
2929688	g	G	A	a	a	a	a	g	g	a	g	g
2951043	t	T	C	c	c	c	c	t	t	c	t	t
2952498	g	G	T	t	t	t	t	g	g	t	g	g
2963444	c	C	T	t	t	t	t	c	c	t	c	c
2977599	g	G	A	a	a	a	a	g	g	a	g	g
2982606	g	G	A	a	a	a	a	g	g	a	g	g
3168368	c	G	T	a	a	a	a	c	c	a	c	c
3180490	t	A	R	c	c	c	c	t	t	c	t	t
3234553	t	T	A	a	a	a	a	t	t	a	t	t
3247603	a	A	C	c	c	c	-	a	a	c	a	a
3346229	c	C	T	t	t	t	t	c	c	t	c	c
3391266	c	C	A	a	a	-	-	c	c	a	-	c

Microtus Pos	Microtus	Ancestral	ES	Derived	Co92	KIM10	Nepal 516	Microtus	IP32953	Antiqua	NC_009381	NC_009708
									Y. pseudo		Pestoides	
3400361	t	T	C	c	c	c	c	t	t	c	t	t
3446576	c	C	A	a	a	a	a	c	c	a	c	c
3491428	g	C	T	a	a	a	a	g	g	a	g	g
3533348	g	C	T	a	a	a	a	g	g	a	g	g
3550138	g	C	A	t	t	t	t	g	g	t	g	g
3702947	c	C	A	a	a	a	a	c	c	a	c	c
3807000	g	G	A	a	a	a	a	g	g	a	g	g
3826835	c	C	T	t	t	t	t	c	c	t	-	c
3917068	g	C	T	a	a	a	a	g	g	a	g	g
3967354	g	C	T	a	a	a	a	g	g	a	g	g
4020111	a	A	T	t	t	t	t	a	a	t	a	a
4023576	t	T	C	c	c	c	c	t	t	c	t	t
4072466	g	G	A	a	a	a	a	g	g	a	g	g
4141711	a	A	G	g	g	g	g	a	a	g	g	g
4178619	g	C	A	t	t	t	t	g	g	t	g	g
4236091	c	G	A	t	t	t	t	c	c	t	c	c
4241745	c	G	A	t	t	t	t	c	c	t	c	c
4279391	c	G	A	t	t	t	t	c	c	t	c	c
4279634	c	G	A	t	t	t	t	c	c	t	c	c
4290808	g	C	A	t	t	t	t	g	g	t	g	g
4417815	c	G	A	t	t	t	t	c	c	t	c	c
4423427	c	G	A	t	t	t	t	c	c	t	c	c
4424460	c	G	A	t	t	t	t	c	c	t	c	c
4526645	c	C	T	t	t	t	t	c	c	t	c	c
4530790	c	C	T	t	t	t	t	c	c	t	c	c

Microtus Pos	NC_010159 Angola	NC_014029	NZ_AAOS02000	NZ_AAUB01000	NZ_AAYR01000	NZ_AAYS01000	NZ_AAYT01000
22497	t	c	c	c	c	c	c
211794	c	t	-	-	t	-	t
211939	t	a	-	-	a	-	a
266734	t	c	c	c	c	c	c
294434	g	a	a	-	a	a	a
323567	t	c	c	-	c	c	c
378080	g	t	t	t	t	t	t
437054	c	c	c	c	c	c	c
522726	g	a	a	a	a	a	a
539902	a	a	a	a	a	a	a
553965	c	t	t	t	t	t	t
579849	a	c	c	c	c	c	c
699134	t	g	g	g	g	g	g
700076	g	t	t	t	t	t	t
718436	g	t	t	t	t	t	t
797929	c	t	t	t	t	t	t
875377	c	a	a	a	a	a	a
975069	a	t	t	t	t	t	t
988148	c	t	t	t	t	t	t
1029862	g	a	a	a	a	a	a
1099213	t	c	c	c	c	c	c
1119592	a	g	g	g	g	g	g
1131883	c	t	t	-	t	t	t
1160296	c	t	t	t	t	t	t
1175799	c	t	t	t	t	t	t
1176227	g	c	c	c	c	c	c
1290920	g	a	a	a	a	a	a
1290935	g	t	t	t	t	t	t
1327376	g	a	a	a	a	a	a
1329145	c	t	t	t	t	t	t
1335276	g	a	a	a	a	a	a
1386605	c	t	t	t	t	t	t
1423760	c	t	t	t	t	t	t
1427890	g	a	a	a	a	a	a
1433228	c	t	t	-	t	t	t
1433576	-	a	a	-	a	a	a
1534714	c	a	a	a	a	a	a
1616932	t	t	t	-	t	t	t
1621516	c	t	t	t	t	t	t
1630030	t	c	c	c	c	c	c
1639545	c	a	a	a	a	a	a

Microtus Pos	NC_010159 Angola	NC_014029	NZ_AAOS02000	NZ_AAUB01000	NZ_AAYR01000	NZ_AAYS01000	NZ_AAYT01000
1924644	c	t	t	t	t	t	t
1956188	c	t	t	t	t	t	t
1957922	g	t	t	t	t	t	t
1971540	t	c	c	c	c	c	c
1973997	t	c	-	-	-	-	c
2044313	c	a	a	a	a	a	a
2180376	g	t	t	t	t	t	t
2222448	a	c	c	c	c	c	c
2345507	t	c	c	c	c	c	c
2393482	t	c	c	c	c	c	c
2394599	g	t	t	t	t	t	t
2458357	a	c	c	c	c	c	c
2469528	g	t	t	-	t	t	t
2494404	g	a	a	-	a	a	a
2543275	g	a	a	a	a	a	a
2563610	g	a	a	a	a	a	a
2564041	g	a	a	a	a	a	a
2578438	t	g	g	g	g	g	g
2593928	g	a	a	a	a	a	a
2600172	a	g	g	g	g	g	g
2670085	a	g	g	g	g	g	g
2714773	a	t	t	t	t	t	t
2743354	g	a	a	a	a	a	a
2750020	a	g	g	g	g	g	g
2773531	c	t	t	t	t	t	t
2819888	c	a	a	a	a	a	a
2825317	c	a	-	a	a	a	a
2845969	g	c	c	c	c	c	c
2866305	c	c	c	c	c	c	c
2877871	t	c	c	c	c	c	c
2882536	g	a	a	-	a	a	a
2899888	t	g	g	-	g	g	g
2929688	g	a	a	a	a	a	a
2951043	t	c	c	c	c	c	c
2952498	g	t	t	t	t	t	t
2963444	c	t	t	t	t	t	t
2977599	g	a	a	a	a	a	a
2982606	g	a	a	a	a	a	a
3168368	c	a	a	a	a	a	a
3180490	t	c	c	c	c	c	c
3234553	t	a	a	a	a	a	a
3247603	a	c	c	c	c	c	-
3346229	c	t	t	t	t	t	t
3391266	c	a	a	a	a	a	-

Microtus Pos	NC_010159 Angola	NC_014029	NZ_AAOS02000	NZ_AAUB01000	NZ_AAYR01000	NZ_AAYS01000	NZ_AAYT01000
3400361	t	c	c	c	c	c	c
3446576	c	a	a	a	a	a	a
3491428	g	a	a	a	a	a	a
3533348	g	a	a	a	a	a	a
3550138	g	t	t	t	t	t	t
3702947	c	a	a	a	a	a	a
3807000	g	-	a	a	a	a	a
3826835	-	t	t	t	-	t	t
3917068	g	a	a	a	a	a	a
3967354	g	a	a	a	a	a	a
4020111	a	t	t	t	t	t	t
4023576	t	c	c	c	c	c	c
4072466	g	a	a	a	a	a	a
4141711	g	g	g	g	g	g	g
4178619	g	t	t	t	t	t	t
4236091	c	t	t	t	t	t	t
4241745	c	t	t	g	t	t	t
4279391	c	t	t	t	t	t	t
4279634	c	t	t	t	t	t	t
4290808	g	t	t	t	t	t	t
4417815	c	t	t	t	t	t	t
4423427	c	t	t	t	t	t	t
4424460	c	t	t	t	t	t	t
4526645	c	t	t	t	t	t	t
4530790	c	t	t	t	t	t	t

Microtus Pos	B42	NZ_AAYU01000	NZ_AAYV01000	NZ_ABAT01000	NZ_ABCD01000	NZ_ACNQ01000	NZ_ACNS01000
22497	c	c	c	c	c	c	
211794	t	t	t	t	t	t	
211939	a	a	a	a	a	a	
266734	c	c	c	c	c	c	
294434	a	a	a	a	a	a	
323567	c	c	c	c	c	c	
378080	t	t	t	t	t	t	
437054	c	c	c	c	c	c	
522726	a	a	a	a	a	a	
539902	a	a	a	a	a	a	
553965	t	t	t	t	t	t	
579849	c	c	c	c	c	c	
699134	g	g	g	g	g	g	
700076	t	t	t	t	t	t	
718436	t	t	t	t	t	t	
797929	t	t	t	t	t	t	
875377	a	a	a	a	a	a	
975069	t	t	t	t	t	t	
988148	t	t	t	t	t	t	
1029862	a	a	a	a	a	a	
1099213	c	c	c	c	c	c	
1119592	g	g	g	g	g	g	
1131883	c	t	t	t	t	t	
1160296	t	t	t	t	t	t	
1175799	t	t	t	t	t	t	
1176227	c	c	c	c	c	c	
1290920	a	a	a	a	a	a	
1290935	t	t	t	t	t	t	
1327376	a	a	a	a	a	a	
1329145	t	t	t	t	t	t	
1335276	g	a	a	a	a	a	
1386605	t	t	t	t	t	t	
1423760	t	t	t	t	t	t	
1427890	a	a	a	a	a	a	
1433228	t	t	t	t	t	t	
1433576	c	a	a	a	a	a	
1534714	c	a	a	a	a	a	
1616932	t	t	t	t	t	t	
1621516	t	t	t	t	t	t	
1630030	t	c	c	c	c	c	
1639545	c	a	a	a	a	a	

Microtus Pos	B42	NZ_AAYU01000	NZ_AAYV01000	NZ_ABAT01000	NZ_ABCD01000	NZ_ACNQ01000	NZ_ACNS01000
1924644	t	t	t	t	t	t	
1956188	t	t	t	t	t	t	
1957922	t	t	t	t	t	t	
1971540	c	c	c	c	c	c	
1973997	c	c	-	-	c	-	
2044313	a	a	a	a	a	a	
2180376	t	t	t	t	t	t	
2222448	c	c	c	c	c	c	
2345507	t	c	c	c	c	c	
2393482	c	c	c	c	c	c	
2394599	g	t	t	t	t	t	
2458357	c	c	c	c	c	c	
2469528	t	t	t	t	t	t	
2494404	a	a	a	a	a	a	
2543275	a	a	a	a	a	a	
2563610	a	a	a	a	a	a	
2564041	a	a	a	a	a	a	
2578438	t	g	g	g	g	g	
2593928	a	a	a	a	a	a	
2600172	g	g	g	g	g	g	
2670085	g	g	g	g	g	g	
2714773	t	t	t	t	t	t	
2743354	a	a	a	a	a	a	
2750020	g	g	g	g	g	g	
2773531	t	t	t	t	t	t	
2819888	a	a	a	a	a	a	
2825317	a	a	a	a	a	a	
2845969	c	c	c	c	c	c	
2866305	c	c	c	c	c	c	
2877871	c	c	c	c	c	c	
2882536	a	a	a	a	a	a	
2899888	g	g	g	g	g	g	
2929688	a	a	a	a	a	a	
2951043	c	c	c	c	c	c	
2952498	t	t	t	t	t	t	
2963444	t	t	t	t	t	t	
2977599	a	a	a	a	a	a	
2982606	a	a	a	a	a	a	
3168368	a	a	a	a	a	a	
3180490	c	c	c	c	c	c	
3234553	a	a	a	a	a	a	
3247603	c	c	c	c	-	c	
3346229	t	t	t	t	t	t	
3391266	a	a	a	a	-	a	

Microtus Pos	B42	NZ_AAYU01000	NZ_AAYV01000	NZ_ABAT01000	NZ_ABCD01000	NZ_ACNQ01000	NZ_ACNS01000
3400361	c	c	c	c	c	c	c
3446576	a	a	a	a	a	a	a
3491428	a	a	a	a	a	a	a
3533348	a	a	a	a	a	a	a
3550138	t	t	t	t	t	t	t
3702947	a	a	a	a	a	a	a
3807000	a	a	a	a	a	a	a
3826835	t	t	t	t	t	t	t
3917068	g	a	a	a	a	a	a
3967354	a	a	a	a	a	a	a
4020111	t	t	t	t	t	t	t
4023576	c	c	c	c	c	c	c
4072466	a	a	a	a	a	a	a
4141711	g	g	g	g	g	g	g
4178619	t	t	t	t	t	t	t
4236091	t	t	t	t	t	t	t
4241745	t	t	t	t	t	t	t
4279391	t	t	t	t	t	t	t
4279634	t	t	t	t	t	t	t
4290808	t	t	t	t	t	t	t
4417815	t	t	t	t	t	t	t
4423427	t	t	t	t	t	t	t
4424460	t	t	t	t	t	t	t
4526645	t	t	t	t	t	t	t
4530790	t	t	t	t	t	t	t

Appendix 7b

Table S6b

Positions differing between East Smithfield (ES) and Microtus 91001 (NC_005810.1)

Numbering as in Microtus (NC_005810.1)

* denotes gene found on complement strand

^ indicates a potential role in virulence

Position	Substitution (Microtus -> ES)	Codon	aa change	Syn/ non-syn	Gene
22497	T -> C				Intergenic
211794	C -> T				Intergenic
211939	T -> A				Intergenic
266734	A -> G				Intergenic
*294434	C -> T	GAG -> AAG	E -> K	nonsyn	glnS1 YP_0287 glutamyl-Q tRNA(Asp) synthetase
323567	A -> G				Intergenic
*378080	C -> A	GTG -> TTG	V -> L	nonsyn	glpR2 YP_0359 DeoR family transcriptional regulator
*437054	T -> C	AAA -> AAG	K -> K	synon	baeS1 YP_0409 two-component sensor/regulator
522726	G -> A	CGC -> CAC	R -> H	nonsyn	rhaS YP_0487 transcriptional activator RhaS
*539902	C -> A	GTG -> TTG	V -> L	nonsyn	aspA YP_0502 aspartate ammonia-lyase
*553965	C -> T	CGT -> CAT	R -> H	nonsyn	YP_0518 hypothetical protein
579849	A -> C	TTA -> TTC	L -> F	nonsyn	aidB YP_0538 isovaleryl CoA dehydrogenase
699134	A -> C				Intergenic
700076	C -> A	AGC -> GTA	S -> V	nonsyn	rluD YP_0654 23S rRNA pseudouridine synthase D
*718436	C -> A	GGG -> GGT	G -> G	synon	emrB YP_0663 multidrug resistance protein B
797929	G -> A				Intergenic
875377	G -> T	CCG -> GTT	P -> V	nonsyn	recR YP_0809 recombination protein RecR
*975069	T -> A	CAG -> CTG	Q -> L	nonsyn	YP_0912 hypothetical protein
*988148	G -> A	TCA -> TTA	S -> L	nonsyn	rcsC YP_0920
*1029862	C -> T	AAG -> AAA	K -> K	synon	YP_0952 putative ABC transport membrane permease
^1099213	A -> G	GAC -> GGC	D -> G	nonsyn	galE YP_1020 UDP-galactose-4-epimerase
1119592	T -> C				Intergenic
*1131883	G -> A	ACC -> ATC	T -> I	nonsyn	grpE YP_1049 heat shock protein GrpE
*1160296	G -> A	CCA -> TCA	P -> S	nonsyn	YP_1083 hypothetical protein
1175799	G -> A	GAC -> AAC	D -> N	nonsyn	miaB YP_1093

Position	Substitution (Microtus -> ES)	Codon	aa change	Syn/ non-syn	Gene
1176227	C -> G			Intergenic	
1290920	C -> T			Intergenic	
1290935	C -> A			Intergenic	
*1327376	C -> T	GCT -> ACT	A -> T	nonsyn	ftsK YP_1217 putative cell division protein
*1329145	G -> A	GCG -> GTG	A -> V	nonsyn	ftsK YP_1217 putative cell division protein
1335276	C -> T	TAC -> GTT	Y -> V	nonsyn	cydC YP_1221
1386605	G -> A			Intergenic	
*1423760	G -> A	GGC -> GGT	G -> G	synon	xylB1 YP_1300 putative carbohydrate kinase
1427890	C -> T	GCG -> GTG	A -> V	nonsyn	serA1 YP_1303
1433228	G -> A			Intergenic	
1433576	G -> T	GAC -> TAC	D -> Y	nonsyn	mtlD1 YP_1309 putative D-mannonate oxidoreductase
*1534714	C -> A	GGG -> GTG	G -> V	nonsyn	adhC YP_1392 alcohol dehydrogenase
1616932	C -> T			Intergenic	
1621516	C -> T	GAC -> GTT	D -> V	nonsyn	YP_1462 hypothetical protein
*1630030	T -> C	GAA -> GAG	E -> E	synon	galA YP_1469 alpha-galactosidase
*1639545	C -> A	CGA -> CTA	R -> L	nonsyn	araC2 YP_1478 AraC family transcription regulator
*1776537	C -> A	GTG -> TTG	V -> L	nonsyn	hpaI YP_1626 2, 4-dihydroxyhept-2-ene-1
*1901080	G -> C	CCC -> CGC	P -> R	nonsyn	YP_1720 hypothetical protein
1924644	C -> T			Intergenic	
^*1956188	C -> T	GGT -> AGT	G -> S	nonsyn	phoP YP_1764 DNA-binding transcriptional regulator PhoP
*1957922	G -> T	ACC -> AAC	T -> N	nonsyn	purB YP_1766 adenylosuccinate lyase
1971540	T -> C			Intergenic	
1973997	T -> C			Intergenic	
*2044313	C -> A	GGC -> GTC	G -> V	nonsyn	YP_1836 hypothetical protein
*2180376	G -> T	CCT -> CAT	P -> H	nonsyn	sppA YP_1962 protease 4
2222448	A -> C	GAC -> GCC	D -> A	nonsyn	ompW YP_1999 outer membrane protein W
2345507	T -> C	CTT -> CCT	L -> P	nonsyn	YP_2106 putative pepetidase
2393482	T -> C			Intergenic	
2394599	G -> T	CCG -> GTT	P -> V	nonsyn	tppB YP_2150 putative tripeptide transporter permease
*2458357	A -> C	GGT -> GGG	G -> G	synon	wcaG9 YP_2207
2469528	G -> T			Intergenic	
2494404	C -> T			Intergenic	

Position	Substitution (Microtus -> ES)	Codon	aa change	Syn/ non-syn	Gene
*2543275	G -> A	GCC -> GTC	A -> V	nonsyn	YP_2291 hypothetical protein
*2563610	G -> A	AGC -> AGT	S -> S	synon	fhaB1 YP_2305 putative hemolysin
*2564041	G -> A	CGC -> CGT	R -> R	synon	fhaC2 YP_2306 putative hemolysin activator protein
*2578438	T -> G	CAG -> CCG	Q -> P	nonsyn	gutB YP_2317 putative zinc-binding dehydrogenase
*2593928	G -> A	CGC -> TGC	R -> C	nonsyn	YP_2331 putative SAM-dependent methyltransferase
*2600172	A -> G	GTG -> GCG	V -> A	nonsyn	menD YP_2338
*2670085	T -> C	GAA -> GGA	E -> G	nonsyn	araC5 YP_2401 AraC family transcriptional regulator
*2714773	T -> A	GTA -> GTT	V -> V	synon	icmF2 YP_2442 hypothetical protein
2743354	G -> A				Intergenic
*2750020	A -> G	GGT -> GGC	G -> G	synon	ccrB YP_2477 camphor resistance protein CrcB
2773531	C -> T	CAT -> TAT	H -> Y	nonsyn	YP_2501 putative hydrolase-oxidase
2819888	G -> T	CAG -> GTT	Q -> V	nonsyn	glnB YP_2545 nitrogen regulatory protein P-II 1
^2825317	G -> T	GTT -> TTT	V -> F	nonsyn	ail2 YP_2550 attachment invasion locus protein
2845969	G -> C				Intergenic
2866305	T -> C	GGT -> GTC	G -> V	nonsyn	sfuB YP_2585 iron(III)-transport system permease
2877871	T -> C				Intergenic
*2882536	G -> A	CGC -> TGC	R -> C	nonsyn	avtA1 YP_2600 aminotransferase
2899888	T -> G	ATC -> AGC	I -> S	nonsyn	cysZ YP_2616 putative sulfate transport protein CysZ
*2929688	G -> A	CCC -> CTC	P -> L	nonsyn	nanT YP_2640 putative sialic acid transporter
*2951043	T -> C	ACC -> GCC	T -> A	nonsyn	napA YP_2661 nitrate reductase catalytic subunit
*2952498	G -> T	CAC -> AAC	H -> N	nonsyn	napF3 YP_2663 ferredoxin-type protein NapF
2963444	C -> T				Intergenic
2977599	G -> A				Intergenic
2982606	G -> A	CTG -> GTA	L -> V	nonsyn	arsC2 YP_2692 putative oxidoreductase
3168368	G -> T	GCT -> TCT	A -> S	nonsyn	cheD2 YP_2855 methyl-accepting chemotaxis protein
3180490	A -> R	GAC -> GRC	D -> stop	nonsyn	aas YP_2864
*3234553	T -> A	ACC -> TCC	T -> S	nonsyn	sPS1 YP_2912 putative kinase protein
^3247603	A -> C	CAG -> CCG	Q -> P	nonsyn	fhaB2 YP_2919 putative adhesin
3346229	C -> T				Intergenic
3391266	C -> A	CGC -> GTA	R -> V	nonsyn	fliA2 YP_3054 flagellar biosynthesis sigma factor
*3396441	G -> A	GCA -> GTA	A -> V	nonsyn	YP_3062 hypothetical protein
*3400361	T -> C	ATC -> GTC	I -> V	nonsyn	YP_3067 hypothetical protein

Position	Substitution (Microtus -> ES)	Codon	aa change	Syn/ non-syn	Gene
*3446576	C -> A	ATG -> ATT	M -> I	nonsyn	YP_3095 hypothetical protein
*3491428	C -> T	GGC -> AGC	G -> S	nonsyn	lpd YP_3131 dihydrolipoamide dehydrogenase
*3533348	C -> T	GGG -> GGA	G -> G	synon	YP_3166 hypothetical protein
^3550138	C -> A	CGT -> AGT	R -> S	nonsyn	wzzE YP_3180 lipopolysaccharide biosynthesis protein WzzE
^3702947	C -> A				Intergenic
3807000	G -> A				Intergenic
^3826835	C -> T				Intergenic
^*3917068	C -> T	GCA -> ACA	A -> T	nonsyn	iucD YP_3447 putative siderophore biosynthesis protein IucD
*3967354	C -> T	GGG -> AGG	G -> R	nonsyn	YP_3488 hypothetical protein
4020111	A -> T	CCA -> GTT	P -> V	nonsyn	YP_3531 putative PTS permease
4023576	T -> C	CTA -> CCA	L -> P	nonsyn	kduD2 YP_3535 2-deoxy-D-gluconate 3-dehydrogenase
4072466	G -> A				Intergenic
*4141711	A -> G	GAT -> GAC	D -> D	synon	YP_3642 hypothetical protein
*4178619	C -> A	CCG -> CCT	P -> P	synon	YP_3671 hypothetical protein
4236091	G -> A	GCG -> GTA	A -> V	nonsyn	YP_3720 hypothetical protein
4241745	G -> A				Intergenic
^4279391	G -> A	AGC -> AAC	S -> N	nonsyn	hmsT YP_3756 HmsT protein
4279634	G -> A				Intergenic
^*4290808	C -> A	GGC -> TGC	G -> C	nonsyn	xylB3 YP_3766 autoinducer-2 (AI-2) kinase
4417815	G -> A				Intergenic
4423427	G -> A				Intergenic
4424460	G -> A	CGC -> CAC	R -> H	nonsyn	accB YP_3887

Appendix 7c

Table S6c - differences between NZ_AAYU01000000 (*Y. pestis* B42003004) and the East Smithfield (ES) chromosome

B42003004 Position	in B42003004	B42	ES	Co92	KIM10	Microtus	*Y. pseudotuberculosis*	IP32953, *Y. pseudotuberculosis*	Nepal516	Antiqua	Pestoides	Angola	NC_014029	NZ_AAOS02000
923462	ancestral	T	G	g	g	t	t	t	g	t	t	t	g	g
3360238	ancestral	C	T	t	t	c	c	c	t	t	c	t	t	t
3569901	ancestral	C	A	a	a	c	c	c	a	c	-	a	a	a
3668184	ancestral	G	A	a	a	g	g	g	a	a	g	g	a	a
3833674	ancestral	C	A	a	a	c	c	c	a	a	c	a	a	a
3930477	ancestral	T	C	c	c	t	t	t	c	c	t	t	c	c
3943308	ancestral	C	A	a	a	c	c	c	a	a	c	a	a	a
4555931	ancestral	T	C	c	c	t	t	t	c	a	t	t	c	c
4605038	ancestral	G	T	t	t	g	g	g	t	t	g	g	t	t

B42003004 Position	NZ_AAUB01000	NZ_AAYR01000	NZ_AAYS01000	NZ_AAYT01000	NZ_AAYV01000	NZ_ABAT01000	NZ_ABCD01000	NZ_ACNQ01000	NZ_ACNS01000
923462	g	g	g	g	t	g	g	g	g
3360238	-	t	t	t	c	t	t	t	t
3569901	-	a	a	a	c	a	a	a	a
3668184	a	a	a	a	g	a	a	a	a
3833674	a	a	a	a	c	a	a	a	a
3930477	c	c	c	c	t	c	c	c	c
3943308	a	a	a	a	c	a	a	a	a
4555931	c	c	c	c	t	c	c	c	c
4605038	t	t	t	t	g	t	t	t	t

Appendix 7d

Table S6d - differnces between B42003004 and ES with genetic descriptions

B42003004 Position	in B42003004	B42003004	ES	CO92	Gene Locus Tag	Description	Syn/ Non-syn?	Codon Change	Amino Acid change
923462	ancestral	T	G	G	YpB42003004_0036		Nonsyn	CAG -> CCG	Q -> P
3360238	ancestral	C	T	T	grpE YpB42003004_2082	co-chaperone GrpE	Nonsyn	ACC -> ATC	T -> I
3569901	ancestral	C	A	A	YpB42003004_2300	mannitol dehydrogenase family protein	Nonsyn	GAC -> TAC	D -> Y
3668184	ancestral	G	A	A	YpB42003004_2405	cydC	Syn	TAC -> TAT	Y -> Y
3833674	ancestral	C	A	A	YpB42003004_2543		Nonsyn	GGG -> GTG	G -> V
3930477	ancestral	T	C	C	YpB42003004_2632	agaN alpha-galactosidase AgaN	Syn	GAA -> GAG	E -> E
3943308	ancestral	C	A	A	YpB42003004_1108	transcriptional regulator, AraC family	Nonsyn	CGA -> CTA	R -> L
4555931	ancestral	T	C	C	YpB42003004_1717	putative protease	Nonsyn	CTT -> CCT	L -> P