Recovery of pathogen genomes from ancient human samples: individual cases, disease and epidemics

Toni de Dios Martínez

TESI DOCTORAL UPF / 2021

DIRECTOR DE LA TESI

Dr. Carles Lalueza-Fox

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

A mi madre,

¡Ésta va por ti!

Acknowledgments

No se'm donen gaire bé aquest tipus de coses, però voldria aprofitar aquesta secció de la tesi per agrair de tot cor, a tota la gent que m'ha acompanyat durant aquesta etapa de la meva vida, m'han fet costat, m'han vist riure, plorar i fer el boig, i al cap i a la fi, m'han fet ser qui soc ara.

El primer és donar-li les gràcies a en Carles, per donar-me l'oportunitat de poder treballar en aquest grup. No tinc paraules per agrair-te el bon tracte que he rebut i per vetllar per que la cosa tirés endavant. De veritat Carles, moltes gràcies per tota la confiança dipositada en mi.

A en Tomàs Marquès per aquelles pràctiques d'estiu en les que vaig aprendre el més bàsic de la bioinformàtica, i sense les quals aniria perdudíssim en la vida.

M'agradaria també dedicar unes paraules a la Montse Colilles, la meva professora de biologia durant ESO i Batxillerat, gràcies per despertar-me la curiositat en la biologia i convèncer-me per que escollis la carrera de Biologia Humana. També a en Bernat Torres i la Teresa Ros, per la vostra guia i consells durant la meva etapa a l'institut.

A tots els membres l'Escola de Grallers de Sitges. Una gran família amb la que he compartit molts bon moments, feliços 50 anys! Una forta abraçada a en Blai pare, en Blai fill, Marius, Jose, Virgili, Amelia, Perez, Xavi Alaman, Gerard Lopez, Xavi Alcazar, Joel, Quintana, Sergi, Duran, Ot, Ferran, Gari, Pere, Paco, Vicent, Bayot, Suri, i als demès (si m'haig de posar a escriure un per un el nom de tots, m'acabareu pagant vatrus la impressió).

A la Blanc Subur, Llorenç, Toni i Pau, la millor colla de Sitges o si més no ho intentem. Sou uns fills de puta, però us estimo. Per molts anys més de bolus, festes majors i farres. pEipVng.

Al Bambú i a la Yami, on els últims mesos m'he passat més temps del que voldria reconèixer, i on segurament me deixat més d'un sou (i de dos i de tres). Agraïments també a l'Anna Lardies pel suport aquests últims dos anys.

Agraïments al Ali Baba Döner Kebab i a les pizzes de Casa Tarradellas, pel seu menjar nutritiu que m'ha mantingut motivat i en bona forma física.

Especials agraïments per a *Monster Energy*, en Jan Laporta, el *WoW*, les Eleccions als Estats Units, Vengamonjas, per mantenir-me despert i amb alguna motivació en la vida quan tot era fosc. Gràcies també al meu antic llogater, ell ja sap qui és, si estàs llegint això, torna'ns la fiança.

Als usuaris del clúster, *ameseguer, boliva, sguirao mbogaert, ebianco i asantini*, per col·lapsar-lo constantment quan l'havia de fer servir. Se qui sou.

Als meus amics de Ribes, en Pau, l'Albert, en Karim, la Raquel, el Lucas, l'Adri, la Daniela i en Nil. Per tots els moments viscuts junts. En els últims anys em anat prenent camins lleugerament diferents, però tot i això sempre traiem algun moment per retrobar-nos i tornar a fer gresca.

Als amics que he fet durant la carrera i màster. Neus per evitar que caigués en la bogeria aquests últims anys i per sempre estar allà quan se't necessita. Als titus Ramon, Xevi i Carlos, que sou com germans i amb els que he passat una de les millors etapes de la meva al pis de Joan Güell. A l'Anna, en Marc, la Monica, l'Hugo i en Fidel, per totes les hores que em passat plegats passant les estones mortes a la Universitat, i totes les birres, festes i caps de setmana junts.

Als companys del grup de Paleogenòmica. A en Pere Gelabert per haver fet l'esforç d'ensenyar-me com funcionava tot quan vaig arribar, al Manu per estar sempre animat fins i tot quan tot era fosc, a l'Iñigo per ser la font de coneixement que és i il·luminar-nos, al Pere Renom per l'entusiasme i la set de coneixement, i a la Laia per tota la seva feina al laboratori, i sense la qual molts dels projectes serien impossibles.

Als meus companys de riures i penúries al laboratori, en Pablo Villegas, la Sandra Walsh, la Isa, en Luis, l'Aitor, l'Andrea, en Julen, en Pablo Carrión i la Tania. Un any francament estrany i ple d'entrebancs. No se com ho em fet, però em aconseguit treure el millor de la situació i gaudir com mai. Un especial agraïment a la Sandra Acosta i a la Gabriela, per posar el seny que falta.

A tota la gent del IBE; l'Esther, en Marc, en David, en Manolo, la Laura, la Nerea, l'André, la Claudia, en Juan, la Paula, la Marina, en Lukas, en Martin, en Marco, en Fabio, en Txema, en Xavi, la Carla, la Laura, la Jessica, la Begoña, la Judit, la Mònica, i si m'he deixat a algú si us plau, no m'ho tingueu en compte.

Thanks to François Balloux, Lucy van Dorp and all the people from the Computational Biology Group at the UCL for bringing me the opportunity to do an *E-visit* to your lab although the circumstances with the pandemic were not the best, but nevertheless, from which I have learned a lot.

Gràcies a la Txiki, el meu gos, que tants moments de goig i felicitat m'has aportat des de ben petit. Mentre escrivia aquesta tesi t'has apagat; aquí i al meu record podràs viure, si més no, uns anys més.

Un fuerte abrazo a mis tíos y tías, Angustias, Isabel, Hilario y Carme, Merin y Maruja, y a mis primos, Bea, Carolina, Alfredo, José Luis, Santi y María del Mar.

A mi familia de Almería, Maravillas, Paca, José, Merceditas, Caniche, Chon, Pedro, Rosa, Luci, Pedro Antonio, Ginés, Ascen, Irene, Gonzalo y Jose Luis. A mi padrino y a mi madrina, i a también a mis primos, Rosica, Jose y Carolina, mis otros hermanos. Os quiero mucho, que la distancia no sea un impedimento para seguir disfrutando de vosotros, especialmente a ti madrina, que pese a todas las adversidades que estas afrontando, te mantienes siempre con entereza y ánimos, mucha fuerza.

A mi abuela, mi segunda madre, que me ha criado y tenido cuidado de mi desde que era un mocoso, que se ha desvivido siempre por mí y por toda la familia y nos ha colmado de amor y cariño. A mi hermano Luis y a Rocío, aunque a veces no soy la persona más animada, con mejor humor y me olvido de fechas señaladas, siempre están ahí para hacer la cosa más llevadera. A mi padre; ya sé que no siempre coincidimos en las cosas, a pesar de todo ello sé que nunca me has deseado mal y has hecho todo pensando en mi bienestar, aunque no sea el mejor hijo, siempre serás mi padre y te querré por ello.

Finalmente, durante la redacción de este manuscrito, se cumplieron 5 años de tu partida. La herida va cerrando poco a

poco, pero siempre me acordaré de ti. A ti mama, que te has dejado la piel para darnos lo mejor, para educarnos y para que fuéramos ante todo personas. Se que no te hizo mucha gracia que escogiera biología en vez de tirar para médico, pese a eso me dijiste *pa'lante* si era lo que quería, y nunca dejaste de apoyarme. Sin ti nada de esto hubiera sido possible. Te quiero.

Abstract

Infectious diseases have affected humanity since its apparition in Africa 300,000 years ago. Demographic changes associated to the Neolithic transition, and ensuing population movements, have facilitated the emergence and expansion of those diseases around the world. The usage of ancient DNA has allowed us to have a snapshot of ancient pathogens' genomes. In this thesis I present the genomes of different ancient pathogens associated to a global disease, to an historical individual case and to past epidemics. In the first case, we retrieve the partial genome of a European *Plasmodium falciparum* strain, which hints the arrival of the parasite to Europe during antiquity. We also take a look at French revolutionary Jean-Paul Marat's condition, in order to shed light to his mysterious condition. Finally, we analyse an ancient *Salmonella enterica* Paratyphi C strain, which suggests that, although infections by this particular serovar are fairly scarce in the present, in the past could have been the responsible agent of epidemics around the globe.

Resum

Les malalties infeccioses han afectat a l'ésser humà des de la seva aparició a Àfrica fa 300,000 anys. Canvis demogràfics associats a la transició neolítica, i posteriors moviments poblacionals, han afavorit l'aparició i dispersió d'aquestes malalties al voltant del mon. L'ús de ADN antic permet obtenir una finestra temporal des d'on observar com eren aquests patògens en el passat. En aquesta tesi presento els genomes d'una sèrie de patògens antics associats a malalties, casos individuals històrics i epidèmies. En el cas de la malaltia, recuperem un genoma parcial d'una soca Europea erradicada de *Plasmodium falciparum,* la qual dona indicis de l'arribada del paràsit a Europa durant l'antiguitat. Fem una ullada també al cas del revolucionari francès Jean Paul Marat amb la intenció d'esbrinar l'origen de la seva condició. Finalment analitzem una soca de *Salmonella enterica* Paratifoide C que podria suggerir que aquest patogen, actualment escàs, era el responsable d'epidèmies al voltant del mon.

Preface

In the last 15 years, the development of NGS techniques in conjunction with aDNA has allowed the retrieval of ancient pathogens' genomic data. This data allows to have a window to the past, from where we can gasp of how an organism genome was. This is particularly interest in the study of pathogens, since ancient samples allow to understand how those pathogens gain the genomic traits necessary to infect humans or acquire virulence. From a phylogenetic point of view, they also provide data to infer their divergence time with extant strains, thus, enabling to get an approximate date and geographical place of origin. Finally, and since the appearance of antibiotics in the last 100 years, and consequent emergence of resistances, ancient pathogens' genomes give the unique opportunity to take a look at a naïve genome which has not been affected by selective pressures.

Given the possibilities that aDNA offers when combined with the study of pathogens, is important to mention the case of Plague. This disease, caused by the enterobacteria *Yersinia pestis,* is the paradigm in the field with multiple samples recovered of several historical and prehistorical strains. However, the status and reputation of the disease has made other relevant infectious diseases to be overlooked or go unnoticed. In this thesis I present the genomes of several other ancient pathogens with current and past clinical importance,

from the perspective of different manifestations of disease, a common disease, an individual case and an epidemic.

Table of contents

Abbreviations

1KGP: 1000 genomes project

A: Adenine

aDNA: ancient DNA

BAM: Binary Alignment Map

BCE: Before Common Era

C: Cytosine

CE: Common Era

DNA: Deoxyribonucleic Acid

dNTP: Deoxyribonucleotide triphosphate

EDTA: Ethylenediaminetetraacetic Acid

FGS: First Generation Sequencing

G: Guanine

HBV: Hepatitis B Virus

HIV: Human Immunodeficiency Virus

HLA: Human Leukocyte Antigen

Kb: Kilo bases

LCA: Lowest Common Ancestor

LD: Linkage Disequilibrium

LNBA: Late Neolithic Bronze Age

MAF: Minor Allele Frequency

mDNA: Mitochondrial DNA

MeV: Measles Virus

mgtB: Magnesium transporter B

mgtC: Magnesium transporter C

ML: Maximum Likelihood

MRCA: Most Recent Common Ancestor

MSA: Multiple Sequence Alignment

NGS: Next Generation Sequencing

NRAMP1: Natural Resistance Associated Macrophage Protein 1

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

pde2: phosphodiesterase 2

pde3: phosphodiesterase 3

pfama1: *P. falciparum* Apical Membrane Antigen 1

pfcrt: *P. falciparum* chloroquine resistance transporter

pfdhps: *P. falciparum* Dihydropteroate Synthase

pla: Plasminogen activator

pfmdr1: *P. falciparum* Multidrug Resistance 1 gene

pfmrp1: *P. falciparum* Multidrug Resistance Protein 1

pvcrt: *P. vivax* chloroquine resistance transporter

pvdhfr-ts: *Plasmodium vivax* dihydrofolate reductase-thymidylate synthase

pvdhps: *P. vivax* Dihydropteroate Synthase

pvmdr1: *P. vivax* Multidrug Resistance 1 gene

SAM: Sequence Alignment Map

SBS: Sequence by Synthesis

SGS: Second Generation Sequencing

SNP: Single Nucleotide Polymorphism

T: Thymine

TGS: Third Generation Sequencing

U: Uracil

VARV: Variola Virus

VCF: Variant Calling Format

ymt: *Yersinia* Mouse Toxine

YpfΦ: *Yersinia pestis* Filamentous phage

1 Introduction

The main goal of this segment of the thesis is to provide a basic understanding of the field of genetics, and to explain which are the characteristics of ancient DNA (aDNA). Furthermore, it will explore the uses and feats that the discipline in the last years, and its use in genomic studies to understand diseases and pandemics, and their emergence.

1.1 History of aDNA

1.1.1 About DNA, Genetics and Genomics

For most of its history, humanity has perceived the natural world as a sort of static and immutable snapshot where the human species was the pinnacle. This vision was extended from astronomy to geology, and life forms were not exempted from it. One of the first authors that questioned the invariability of life forms and the status of human as the apex of the life's pyramid was Charles Darwin. In 1859, and after travelling for 5 years around the world in the *HMS Beagle* (voyage which inspired his works), published the book "*On the Origin of* Species"[1]. In this work he proposed the revolutionary idea of the evolution of life organisms. The basis is that, given an initial variability among individuals and a competence for the natural resources, the concept of natural selection arises. In this

scenario only the fittest survive and reproduce, and their descendants would inherit those advantageous traits. The means from which those traits were able to be inherited by descendent individuals was not known. Around the same time that Darwin's works were published, the Augustinian monk Gregor Mendel published the article *Experiments on Plant Hybrids ("Versuche über Pflanzen-Hybriden")*[2]. In this works, Mendel described the laws of traits inherited in different plants crossing. Nevertheless, it was still unknown which mechanism drives the transmission of those traits.

It was not until the start of the 20th century, when Thomas Morgan and colleagues published the book *The Mechanism of Mendelian Heredity*[3]*.* The book was based in their work with heredity in the fly *Drosophila melanogaster,* in which they describe how the genes present in the chromosomes are responsible of the inheritance of phenotypical traits. Note that despite chromosomes were already known structures, it was a mystery what was the gene's nature or what allows it to transmit the characters from generation to generation[4,5]. In 1944, Avery, MacLeod and McCarty discovered that using a molecule known as Deoxyribonucleic Acid (DNA) from a virulent strain dead pneumococcus, they were able to convert non-virulent pneumococcus in virulent ones[6]. Years later Watson and Crick, thanks to the previous crystallography works of Rosalyn Franklin, described the structure of DNA[7]. The molecule consists in a double helix chain with a phosphor-

deoxyribose backbone and 4 nucleobases (Adenine, Cytosine, Guanine and Thymine; A,C,G,T) which are “in contact” through hydrogen bonds. Finally, Jacob and Monod decode how DNA store the information necessary to synthetise proteins, in which groups of 3 nucleotides known as codons encode for a specific amino acid[8].

Since those times, genetics has been rapidly advancing. Some of the benchmarks on the field I would like to highlight are the following. The creation of the first sequencing techniques (FGS) by Sanger and colleagues[9], and Maxam and Gilbert[10], which allowed to retrieve and study the first gene sequence. Related to this, the discovery of the Polymerase Chain Reaction (PCR) by Kary Mullis, allowing the replication of DNA sequences in the laboratory[11]. Those breakthrough discoveries allowed for one of the greatest achievements in science history, the publication of the human genome sequence by the Human Genome Project consortium[12]. The titanic effort costed an estimated 3 billion dollars and lasted for 13 years[13]. This feat, besides providing unvaluable information for science, paved the way for the development of more efficient sequencing techniques (Second Generation Sequencing or sometimes Next Generation Sequencing; SGS - NGS)[14–16], and ultimately cheaper genome sequencing and the possibility of systematically sequencing individuals[17][Figure 1]. Since then, projects such as the 1000 genomes project (1KGP) have retrieved hundreds of human genomes[18], or from other

species[19]. The development in the recent years of even more efficient sequencing techniques (Third Generation Sequencing)[20,21] could mean a faster development in fields such as population genetics, medicine, personalised medicine, conservation biology, among others.

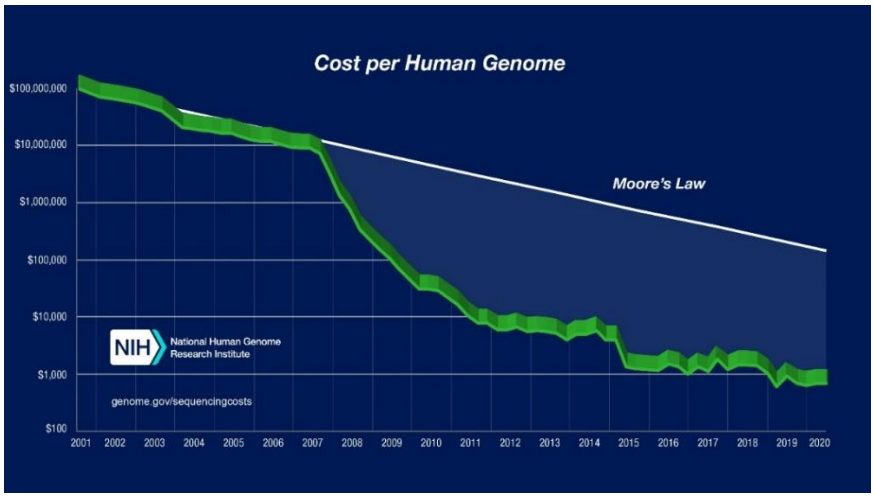


Figure 1. Evolution of genome sequencing cost (2000 - 2020). The price of genome sequencing has decreased faster than expected by Moore's Law during the last 14 years. As for 2020, the price of sequencing a genome is about 1000 USD. From the National Human Genome Research Institute[22].

1.1.2 About aDNA and Paleogenomics

Ancient DNA offers us the unique opportunity to have a window to the past. It allows us to get a glimpse of the genome of extinct animal species and know which factors influenced their extinction [23–25], understand human population movements and events (both historical and prehistorical)[26–28], comprehend past pandemics and how they have shaped humans today[29,30], and has implication in genetic studies of historical figures[31–33].

The history of aDNA is tightly related to the development of First Generation Techniques. It all started in 1984 with the retrieval of DNA fragments extracted from the remains of a *quagga,* an extinct zebra subspecies[34]. As commented in the previous section, the discovery of PCR was crucial for the genetics and genomics field, and aDNA was no alien to it. The field flooded with optimism, new reports of new aDNA sequences, recovered from diverse sources such human remains[35,36], plant remains[37] and extinct animals[38–40]. It seems that aDNA was the panacea, that it would be possible to retrieved it from every imaginable source. Then a series of articles claiming to extract DNA from million years old samples, such as dinosaur bone[41] and amber preserved insects[42], were published. The articles were promptly debunked as being caused by modern contamination, but the credibility of the field remain damaged for a time[43–45].

Slowly but steadily, and thanks to more dedicated quality controls to ensure that contaminants were no longer identified as authentic, aDNA field credibility recovered. Prove of this is the publish of the first sequences of an ancient hominid, the Homo neanderthalensis[46,47]. Also, the first ancient pathogen strains were started to be discovered[48–50] .Years later, after the release of the Human Reference Genome, the development of NGS started. Those technics have a great impact in aDNA studies. This was reflected in the first genome wide DNA retrieval of an extinct species, the mammoth[51,52]. Soon followed the first ancient anatomically modern humans[53,54], the whole Neanderthal genome[24], the genome of a 700,000 year horse (oldest full genome recovered to the date)[55], among others. Other advances include the retrieval of whole bacterial pathogen genomes, such as the case of medieval strains of Yersinia pestis[56]. But among the most relevant discoveries, is necessary to remark the description of a new extinct hominid from genomic data[57,58]. The Denisova hominid is a special case since until recently no major body remains were found[59,60], and phenotypical traits had been inferred using genetic data[61].

As we have seen, the later years are characterised by an increment of the generation of wide genome data. This has also been the case for aDNA, in each new study hundreds of new ancient human samples alone are published[62–65], making the number of ancient genomes of the order of thousands. This

would allow to do large scale population analysis in the following years. As for future perspectives for the aDNA field, paleoproteomics is worth of mentioning. Proteins are much more stable than DNA, being able to survive for millions of years, past the window that we have using aDNA [55,66,67], effectively extending the window of study by millions of years. This is fast advance has manifested in the last 4 years with the recovery of the first ancient hominid peptides[68,69], which could help to get calibrated phylogenies for species whose their DNA is not possible to retrieve.

1.2 aDNA characteristics

1.2.1 aDNA conservation

In comparison with modern DNA, aDNA has a set of particularities which characterised it and condition the way that analyses are performed. Conservation is a key element since is the responsible of the time window which aDNA offers to explore. As other complex molecules, DNA has an estimated lifetime[45]. When an organism is alive this process is not evident, since different molecular machinery exists to prevent DNA damage and decay. After the death of an organisms, the DNA is degraded by the enzymes present in the cell[70], and with the time, microorganisms also start degrading it[45]. Is for this reason that aDNA fragments are usually short (25 – 80 bp), usually but not always, correlated with its age[71–74].

The location where a sample is found it also critical for the aDNA conservation. Factors such as humidity (water causes hydrolytic damage to the DNA)[45], oxidative chemical compounds found in this water or which can change the pH [75], and temperature affect DNA preservation[71,72]. This is one of the main reasons that ancient genomes are typically recovered from temperate and cold latitudes, in contrast to more humid and hot regions [Figure2][76]. For example, the remains of a 700,000 prehistoric horse were found in permafrost. Those optimal conditions are close to the higher bound of life

expectancy of DNA[55]. Other optimal conditions can be found in caves, where temperatures are usually low the whole year without dramatic oscillations[67,77]. The maximum age estimated for a DNA molecule is in the order of hundreds of thousands of years to a million years [45,72].

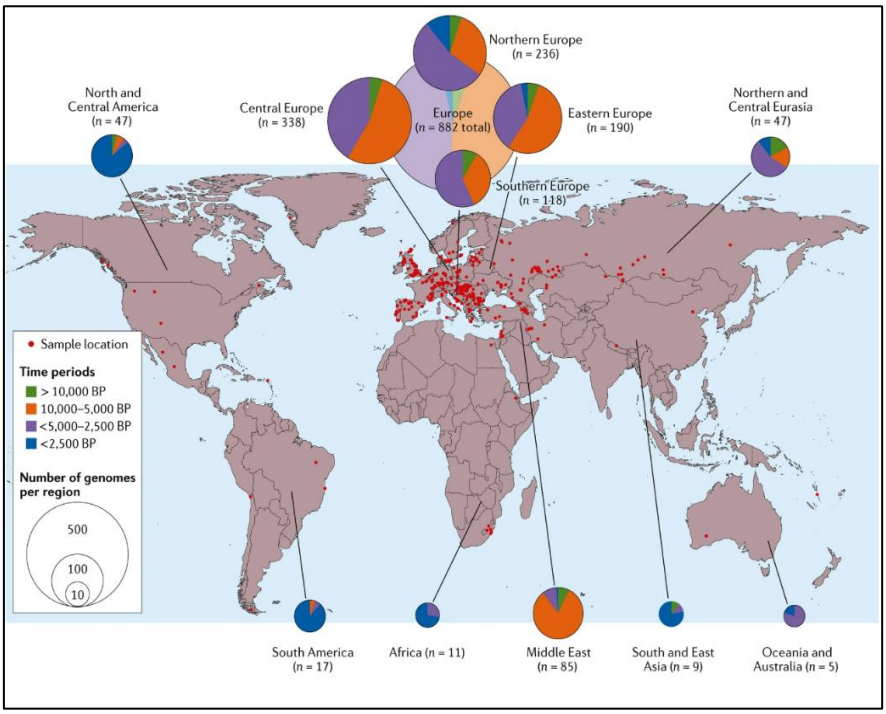


Figure 2. Geographical location of all ancient hominid genomes recovered. Samples location are marked with red dots. The number of genomes recovered per region is represented with a pie chart. The size is proportional to the number of genomes. Note that most of the genomes are recovered from cold latitudes. From Marciniak and Perry 2017[76].

1.2.2 aDNA content and recovery

The explained factors that affect aDNA preservation are responsible for the quantity and quality of aDNA. Here we have to introduce the concept of endogenous DNA, which is the proportion of DNA in a sample which is attributable to the individual. Due to the aforementioned factors which affects to the DNA lifetime, but also to the presence of environmental microbial community DNA found in the sample, and the possible presence of other modern sources of contamination, the proportion of endogenous DNA is usually low (~1%). Nevertheless, in exceptional conditions in preservation, the endogenous DNA can be dramatically increased[78,79].

In general, the most abundant samples are bone remains. Before extracting the aDNA from a sample is of importance to know that different types of bones yield different proportions of endogenous or possible pathogen DNA. Due to its density and isolation[80], petrous part of the temporal bone[81], especially the otic capsule[82] and auditory ossicles[83], are regarded as the optimal sources of endogenous DNA. Another excellent source of endogenous DNA is the thin cementum layer covering the dental root[84,85]. Nevertheless, other bones can be used to retrieve endogenous sequences[46,58,77]. Endogenous DNA content may differ from extractions within the same individual[24].

Despite its high abundance of endogenous DNA, pathogen DNA content is low in the petrous bone[86], which is hypothesized to be caused by the high density and poor irrigation of this particular bone[80]. In contrast, bacterial DNA is readily found in teeth, specifically in the dentine, cementum and pulp[86,87]. This can also be explained by differences of bone turnover rate and blood irrigation when compared with temporal bone[86]. Despite different species of pathogens can be sampled from teeth, due to their lifecycle particularities, other species have to be extracted from specific bones or other tissues. Some examples of those pathogens are *Mycobacterium leprae*[88], *Mycobacterium tuberculosis*[89], *HIV*[90], *Plasmodium falciparum* and *P. vivax*[91], *Helicobacter pylori*[92], *Variola virus*[93], *Vibrio cholerae*[94].

Finally, in the recent years, the improvement in metagenomics sampling techniques has allowed the recovery of aDNA from a diverse array of samples such as parchments[95], ancient latrine deposits[96], cave soil[97,98] and ancient “chewing gum”[99].

1.2.3 Deamination in aDNA damage

As mentioned in the previous sections, aDNA is characterised for being affected by post-mortem chemical reactions. The most iconic damage pattern present in aDNA is the hydrolytic deamination of Cytosine to Uracil (U)[100,101] [Figure 3a]. This change is manifested in sequenced data as a C to T

substitution in the 5' end (due to complementarity of U and A after replication) and as a G to A substitution in the 3' end of the complementary chain[102]. Age and preservation conditions affect to the proportion of deamination at reads end (increasing it), but also the present of modern contaminants can artificially reduce it[103,104]. Cytosine deamination occurs mainly in the ends as a result of the presence of single strand breaks, as they are more chemically exposed when compared tom double-stranded DNA[45]. The rate of deamination is estimated to be about 100 times faster in the end of a DNA molecule than in the centre[45,103]. The presence of deamination is not trivial since it can bias analyses, so enzymatic treatments using *Uracil-DNA-glycosylase* have been developed to revert this damage[105].

1.2.4 Other types of aDNA damage

Although Cytosine deamination is the main damage present in aDNA, other types of lesions can occur. Another remarkable type of aDNA damage is depurination. Depurination causes the break of the *N-glycosyl* between an A or G and the ribose which conforms the DNA backbone. This results in an abasic site, and the subsequent DNA strand-break through β elimination [Figure 3b][45,106,107]. At the end this results in the fragmentation of DNA molecules responsible of the short length of aDNA molecules[108]. Furthermore, G is describe to be more susceptible to depurination than A[103]. In addition to

depurination, it has been suggested the existence of blocking lesions caused by pyrimidine oxidation[71] and Mallard reaction[109].

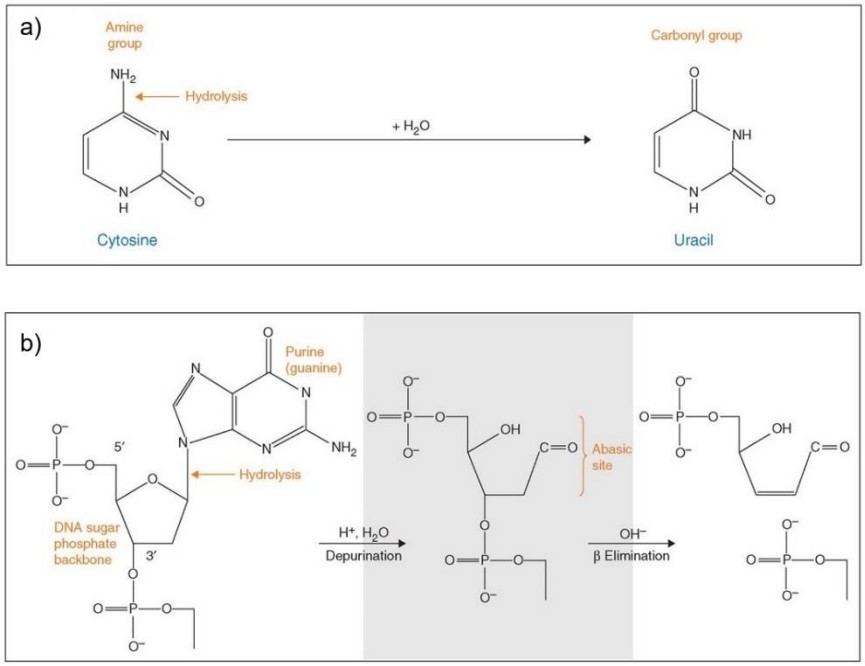


Figure 3. The 2 of the most common DNA lesions occurred after the death of an organism are hydrolytic Cytosine deamination into Uracil (a), and Adenosine or Guanine depurination and subsequent strand break (b). From Dabney, Meyer and Pääbo 2013[110].

1.3 Paleomicrobiology

1.3.1 Disease, Human Prehistory and Neolithic Transition

Pathogens have accompanied humankind since the apparition of first humans in Africa 300,000 years ago[111,112]. Infectious diseases are believed to be one of the major mortality factors in early humans based in comparison with modern hunter-gatherer populations[113]. Archaeological evidence of infections is difficult to assess[114], but molecular studies evidenced the coexistence of ancient humans and pathogens for tens of thousands of years[112,115]. Malaria is arguably one of the most relevant pathogens which has affected humans during millennia and has exert a great evolutionary pressure[116,117], to the point of reshaping human populations genomes[118–121].

Some of the first infectious and parasitic diseases affecting Palaeolithic hunter gatherers were tapeworms, pinworms, lice, malaria parasites, *Bordetella spp.*, *Staphylococcus spp., Mycobacterium lepra* and *tuberculosis*, *Salmonella typhus*, *Helicobacter pylori*[112,117,122–128]. Most of those diseases have been hypothesized to have coexist for long time with human or be acquired by sporadic zoonosis[124,129]. Is important to note that most of the infectious diseases which are common nowadays appeared during the Neolithic transition, caused are

either caused by zoonotic jumps, or by the adoption of sedentarism and the subsequent changes in lifestyle[124,129].

During the Neolithic revolution, human populations experienced a major population boom. The discovery of agricultural and husbandry techniques allowed to maintain larger populations, but at the expense of establishing permanent settlements in order to protect the crops and animals[130,131]. This resulted in an increase of population density, reduced sanitary conditions, close contact with animals, and changes in diet and lifestyle, propitiating the spread of diseases[132,133]. Pathogens emerging or expanding during the Neolithic are *Salmonella enterica* strains, *Yersinia* species, *Vibrio cholera*, *Mycobacterium lepra*, *Mycobacterium tuberculosis*, and *HBV*[29,134–137]. There are pathogens such as *Plasmodium falciparum* that despite having a suggested Palaeolithic origin[117], can associate its expansion to the Neolithic revolution[138]. Others such as *MeV* and *VARV* have a probable younger origin due to its virulence and Neolithic populations not being large enough to maintain the rate of spread of the disease[139–141].

As a final remark, the directionality of the disease jump seems to not be exclusively restricted from animal to human but is apparent that multiple jumps in both directions have happened in different pathogens. The study of ancient *S. enterica* strains demonstrate a generalist origin incompatible with a

transmission of the bacteria into humans by pigs[134]. Other diseases such as *Mycobacterium* species have been suggested to coexist with human for thousands of years but are also present in animals[89,115,127,142]. This data suggests that, despite a considerable number of diseases could sporadically infect ancient humans prior to the Neolithic revolution, most of human common pathogens and epidemics could attribute their global expansion to it.

1.3.2 Pandemics: The Plague from a Paleogenomic Perspective

Now that we have taken a look at how most infectious diseases appeared or expanded during Neolithic is time to analyse one of the most paradigmatic cases of pandemic, the Plague. Is also one of the most studied pathogens from and aDNA perspective, data from 3 historical pandemics have been recovered, and additionally, 2 previously unknown prehistoric epidemics have been discovered.

The Plague is an infectious disease caused by the enterobacteria *Yersinia pestis*. It has 3 different type of clinical manifestations, the bubonic plague, the pneumonic plague and the scepticemic plague[143]. Rodents are the main reservoir of the disease, while fleas carry the disease as vectors[144]. Its main route of transmission is by contact or ingestion contaminated sources, by air or by a vector. It has a mortality

ratio of 30%, up to 100% if is not treated[144]. *Y. pestis* has a 4.5 Kb genome, and 3 virulence plasmids *pCD1*, *pMT1* and *pPCP1*[145–147]. The virulence of the bacteria is linked to a determinate set of genes and mutations such as the *ymt* gene (*pMT1* gene required for the bacteria survivability in the flea's gut[148]); loss of function mutations in the genes *pde2, pde3* and *rcsA* (also linked to flea transmission[149]); *pla* gene (*pCPC1* gene associated to bacterial dissemination in mammals and progression to pneumonic plague[150]); *YpfΦ* (confers higher fitness to the bacteria during infection[151,152]); and *ureD* (associated with the loss of ureolytic activity[153]). The main biovars are *Antiqua* (ANT), *Medievalis* (MED), *Orientalis* (Ori), *Intermediate* (IN) and *Pestoides* (PE)[154] [Figure4]. Despite its past global presence, nowadays is restricted to regions of Central Asia, Africa, and the Americas[155,156].

The Plague its mainly known for its historically recorded epidemics during the late Antiquity (the Justinian Plague, 6th century CE), the Medieval times (the Black Death and its subsequent epidemics, 14th - 18th century) and the 19th century China epidemics. Other historical have been labelled as *Plague* such as the Athens plague (5th century BCE) and the Antonine and Cyprian Plagues (2nd – 3rd century CE) but were not caused by *Y. pestis*. In the case of the Athens Plague the causal agent seems to be a *Salmonella* related *enterobacteria*[157,158], while the causes of the Antonine and Cyprian Plague remain unknown and open to speculation[159].

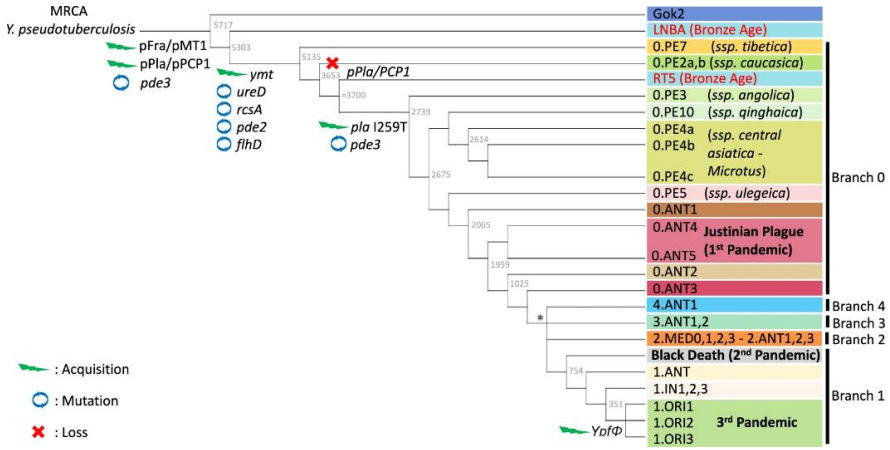


Figure 4. Current and historical diversity of *Yersinia pestis* lineage. Ancient strains are Gok2, LNBA, RTS, the Justinian Plague, the Black Death and the third pandemic. Gain, Mutation or Loss of genomic traits are marked at the bottom of each clade. The *Big Bang* polytomy is marked with an asterisk. From Demeure *et al.* 2019[160].

Yersinia pestis originated from *Yersinia pseudotuberculosis*, a zoonotic pathogen which is capable of infecting humans causing a mild disease known as the Far East scarlet-like fever[145]. Recent aDNA studies placed the divergence of Y. pestis from *Y. pseudotuberculosis* around the year 3,700 BCE[161]. This first strain isolated from a Neolithic Scandinavian individual is basal to all know *Y. pestis* diversity, forming its own independent clade. The strain lacks the gene *ymt* in the *pMT1* plasmid. This specific clade of Y. pestis is linked to the late Neolithic decline and predates the arrival of the step ancestry to Western Eurasia.

Following the Neolithic decline, a new independent clade of *Y. pestis appears*, this time associated to the arrival of the

Eurasian-steppe component into Europe during the Bronze Age (3000 – 800 BCE)[29,162]. This clade (LNBA) is basal to all the modern *Y. pestis* diversity. All those strains but the youngest (800 BCE) also lack the *ymt,* and all Bronze Age strains lack the necessary mutations in *pde2, pde3* and *rcsA*; indicating that those strains were not transmitted by fly. The LNBA reached as far East as Eastern Siberia[163]. In parallel and contemporary to those Bronze Age strains, another Central Asian different lineage dating to 2800 BCE has been discovered[164]. These strains have the *ymt* gene and other virulence factors associated to flea transmission. The strains (RT5) are reported to fall within the diversity of the 0.PE clade, separated from the clade 0.ANT (includes most of the modern diversity, and the historical Justinian, Black Death and China Pandemics).

The third ancient Plague pandemic, and the first historically recorded one, is the Justinian Plague (541 – 543 CE), known with this name for affecting the roman emperor Justinian I [165]. The first record of the pandemic is from the eastern ports of Egypt. From it spread to other provinces of the Eastern Roman Empire and to other regions of Europe. It reportedly caused the death of between 35 and 55% Roman Empire Population[165]. The reported strains of Justinian Plague fall within the 0.ANT diversity and has a probable origin in Central Asia but is debated in how was introduced to Europe[166–169].

The second historical pandemic is the Black Death (from 1346 to 1353) with subsequent epidemic until the 18th century. It started in the Genoese possessions in Crimea, extending via trade routes to the Mediterranean basin and Northern Europe in the following years[170]. It killed between 1 and 2 thirds of European populations (estimates between 70 to 200 million deaths) and started a series of cultural and demographic changes associated to the end of the Medieval Epoch[171–173]. Archaeologically, this elevated mortality is represented in mass burials, also known as *Plague pits* [Figure 5][174]. All the Black Death recovered strains fall inside the same extinct clade originated from a diversification event known as the "*Big Bang*" polytomy[175]. The clade is found in the Branch 1 of *Y. pestis* diversity, being basal to the modern existent strains and the 3rd pandemic strains[56,176–182]. It has been suggested, due to a lack of diversity, an unique introduction from central Asia have occurred[182], and together with processes of persistence, originated the late European epidemics[178,181,182]. Most of those later Europeans epidemics lacked the *pla* virulence gene[182]. Some of them also reported the deletion of the *mgtB* and *mgtC* genes with unknown effects in virulence, but are hypothesised to help them to replicate in macrophages [182–184]. Finally, those same strains seem to have spread eastwards, contributing to the origin of the 3rd plague pandemic and modern diversity[178,181].

The 3rd historical pandemic originated in China in the mid-19th century. From here it spread into America, Africa and across Asia, expanding in to the local rodent population, and originating the 3 main *Orientalis* linages present nowadays[156,185–188].

Figure 5. *Plague pit* associated to the Black Death. Mass burials dating back to the second Plague pandemic are common through all Europe. This one in particular is from the Medieval site of *L'Esquerda,* Barcelona, Spain. The site was inhabited since the Bronze Age, but was depopulated after the pandemic[189].

We can observe a gradual adaptation of *Yersinia pestis* from its zoonotic origins to a more human specialised pathogen with the acquisition of different virulence factors which favour its transmissibility through vectors (*ymt,* in *pde2, pde3* and *rcsA*), favour the transition to a more infective stages of the disease (*pla*), increase the fitness in mammalian hosts (*pla*, *YpfΦ* and *ureD*), and evasion of the host immune system (*mgtB* and *mgtC*).

To finish this section, and as we will see at in the results section, the historical importance of plague has resulted in it being the most studied ancient pathogen but has also contributed to other potentially relevant pathogens which have had a role to pandemics being somewhat overlooked.

1.3.3 The New World, spreading of infectious diseases during colonial times

The Americas were known as the New World during the Age of Discovery and subsequent European Colonial time period. The continent is situated in the Western Hemisphere, being comprised of two main landmasses (North and South America) and a major archipelago (the Antilles). Despite its relatively recent discovery by Europeans powers, briefly settled by the Vikings in the 10th century and its rediscovered by Cristopher Columbus in 1492, the continent was already populated by diverse ethnic groups.

Taking the human out Africa 50,000 years ago as a starting point, the Americas were the last continent on being populated[190]. Is believed that the first ancestors of current Native American (Ancient North Eurasian) populations lived 24,000 years ago and were related to the ancient inhabitants of eastern Siberia and modern day Western Eurasians [191]. During the last glaciation, the Bering strait dried out, emerging a landmass known as Beringia, where those Ancient North Eurasian population settled[163]. Approximately between 22,000 and 18,000 years ago, and after a period of isolation, those Beringian populations split[192], and the ancestors of modern-day Native American settled in what is now Alaska and Yukon. Here between 17,000 and 14,000 years ago[193], the Ancestral Native American diverged in tow main North Ancestral and South Ancestral populations, which in different migratory weaves, populated the whole continent [194–198] [Figure 6].

In 1492, Christopher Columbus, in name of the Castilian crown, set foot in the in the Antilles. The first Spanish settlements were placed around the Caribbean, but 30 years late, and after the conquest of the Aztecs and Incan empire, the territories spanned from modern day Chile up to California and Florida, incorporating those territories to the crown[199]. In parallel the Portuguese started colonising Eastern South America, parts of Coastal Africa and India. By 1600, and after a temporal dynastic union between Portugal and Spain, the Iberian Union controlled most of South America and Western

North America, the Macaronesia, the Philippines, parts of Europe, and numerous outposts along the Asian and African Coast. Additionally, the French and English had also set colonies in parts of South America and the North American Eastern Coast.

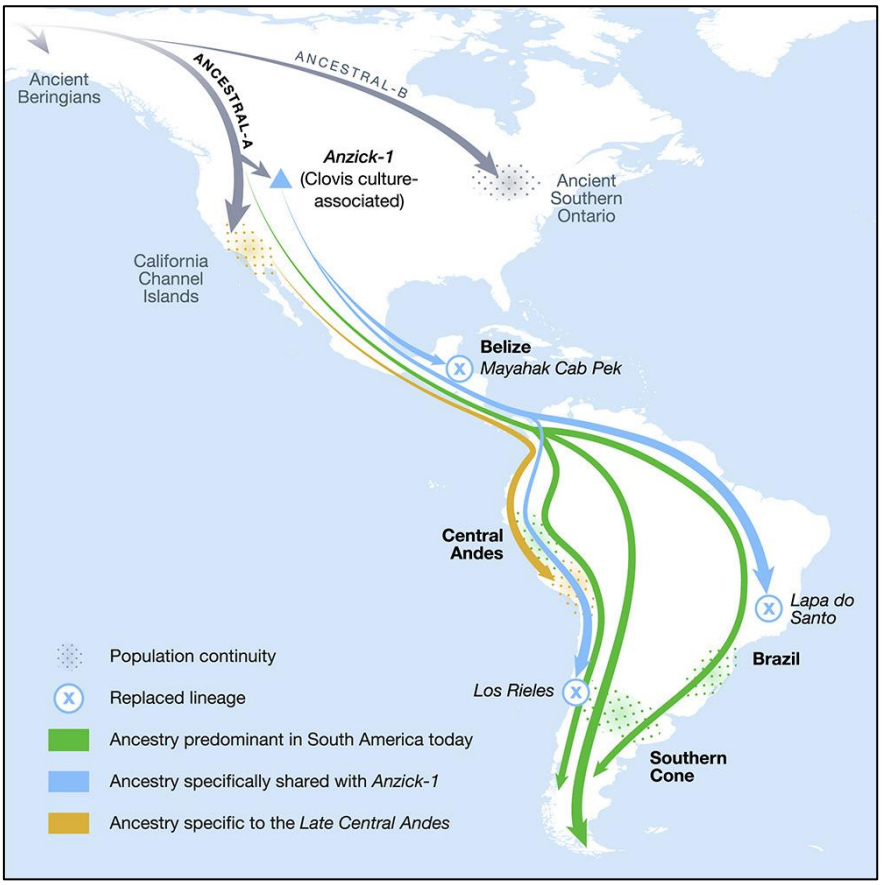


Figure 6. Routes of colonisation used by Ancient Native American populations. Ancestral B populations (North Ancestral populations) from which present day Inuit and Northern Native Americans are descendants, and Ancestral A populations (South Ancestral populations, colorised in the map) from which present day Central and South Native Americans are descendants. From Posth *et al*. 2018[194].

The rise of those first global empires effectively connected distant human population around the world. Millions of people were brought from Western Africa to work as slaves in different American Colonies. Associated the arrival of colonists and slaves, several epidemics occurred. If we compare the time of European arrival and founding of first colonies to date of the first recorded epidemic events, the later befall strikingly early. The first testimony refers to a smallpox outbreak in La Hispaniola in only 1518, extending to Central America the next year [200]. Also, unknown diseases had probably reduced the native population of the island already, since the population decayed from 3,700,000 in 1493 to 15,000 prior to the smallpox epidemic[200]. Although smallpox is the best known pathogen to ravage the Americas, other epidemics also devastated the indigenous population, including typhus, influenza, measles, variola, yellow fever, typhoid fever, malaria, and other of unknown origin as the *Cocolitzli*[200–205].

It has been argued that, from an evolutive perspective, European and African population which settled in the Americas had been selected for those diseases, and despite that, smallpox reached a mortality of 30% among to already exposed populations[206]. During millennia Old World populations have been subjected to conditions of high population density, lack of hygiene and contact with domesticated animals which propitiate the emergence of a variety of infectious diseases, and therefore, their immune

systems have been eventually selected for those selective pressures[207]. In contrast, Native populations did not reach the populations densities typical of the Old World and did not domesticated animals (only the lama), fact which could had hampered the jump of zoonotic borne diseases[207,208]. At the end, is estimated that about 90% of the original Native Americans population succumbed to different infectious diseases[200].

From a Paleogenomic perspective, there are pathogens of European or African origin which have been associated to post-Columbian contact epidemics. In contrast, the evidence of infectious agents imported from the Americas into the Old World is limited and debated. Recent discoveries for example suggest the pre-Columbian presence of tuberculosis or syphilis[89,209]. As we will explore in the results, the demographic changes resultant of colonialism has reshaped both human and pathogens ancestry compositions.

1.3.3.1 *Salmonella enterica* and the *Cocolitzli*

The *Cocolitzli* (Nahuatl word for disease) were a series of epidemics which affected Central America shortly after the arrival of the Spanish[201,202]. Several sources reported the disease to be some kind of haemorrhagic fever, which caused bleedings from the eyes, nose, mouth, haemorrhagic diarrhoea, and a net-like rush[205,210–212] [Figure 7]. Spanish

missionaries described one of these epidemics as *tabardillo*, medieval Spanish name for typhus[210,213]. Additionally, colonist differentiated the *Cocolitzli* ethology from the one of previous smallpox outbreaks[214]. Several infectious entities have been suggested as a possible agent of the disease such as typhus, enteric fever, bartonella, measles, yellow fever, an unknown haemorrhagic virus, among others[204,215–218].

Figure 7. Different Native representations of the 1545 *Cocolitzli* epidemic. The documents are the *Codex en Cruz* (a), *Codex Mexicanus* (b) and *Codex Aubin* (c). The images depict the effects of the epidemic, rashes and a red fluid (possibly blood). The *Codex Aubin* reads in Nahuatl "*The year 1545. At this time an epidemic spread so that everyone's nose bled. It had prevailed for a year when the market at San Hipólito opened"*. From Vågene *et al.* 2018[30].

The recent discovery of Colonial cemetery in Mexico dating back to the 1545-1550 shed some light regarding the origin of the mysterious *Cocolitzli*. The analysed samples show the unequivocal presence of *Salmonella enterica* Paratyphi C[30]. The found Mexican strains were basal to current day Paratyphi C diversity[134], with only a Norwegian medieval strain being below them in the phylogeny[219]. This particular serovar of *Salmonella* is sparsely reported nowadays when compared to other serovars such as Typhi or Paratyphi A[220,221]. Despite this, it causes an enteric fever similar to typhoid fever or typhus common in developing countries, which if left untreated, can reach mortality rates of 20%[222,223]. The disease is spread by contaminated water or food, and infected individuals can became asymptomatic chronical carriers[224].

The evidence points towards *S. enterica* Paratyphi C as the main causal agent of the *Cocolitzli*. As described, it has been previously suggested the epidemics were caused by either typhus or enteric fever. The symptoms described also are plausible with the disease, and the colonists recognised the disease as typhus. Finally, the capacity of the S. enterica to create asymptomatic carriers could explain how it could had been unnoticed during the 2 months of transatlantic voyage[225]. Nevertheless, and as described above, there are other diseases which could match the symptoms exhibit by the *Cocolitzli*, and other studies contradict those evidence[226]. Further discoveries have to be made to either confirm the

presence of Paratyphi C in more *Cocolitzli* associated burials or demonstrate the presence of other pathogens.

1.3.3.2 *Plasmodium spp.*

The *Plasmodium* parasites are the causal agents of Malaria. Species which infect humans are *P.falciparum, P vivax, P malariae, P. ovale* and *P. knowlesi*. Their natural reservoir are simians[138,227,228]. The malaria parasites life cycle includes several stages, being the merozoite the forms which infect erythrocytes causing their lysis the responsible of the clinical manifestations of the disease: anaemia, fevers and in some cases, neurological disorders [Figure 8][229].

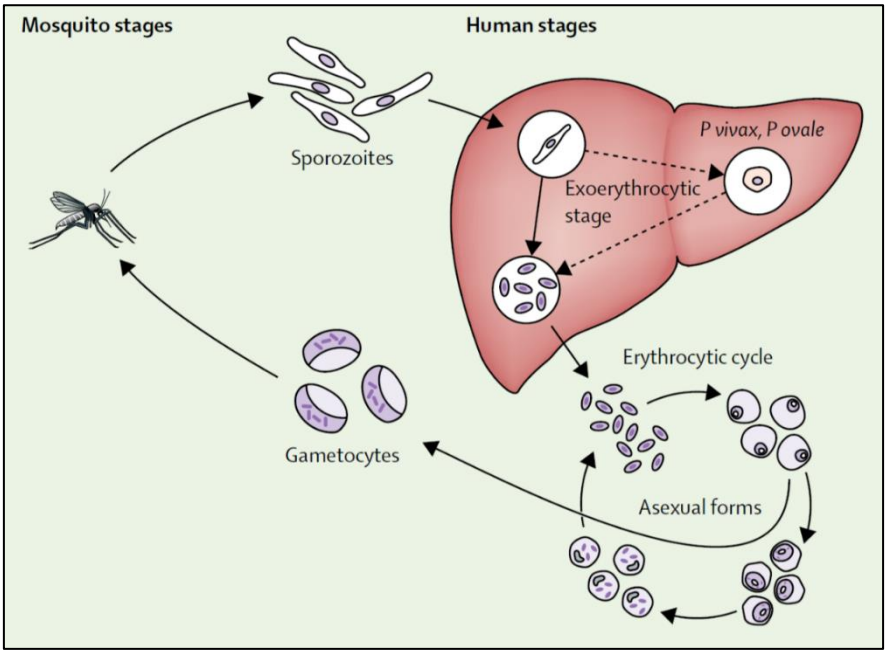


Figure 8. *Plasmodium spp.* life cycle. From Ashley *et al.* 2018[229].

As for 2019, is estimated that there are 229 million cases annually and 409,000 deaths caused by malaria[230]. Most of the infections worldwide are caused by either *P. falciparum* or *P. vivax*. Both species are present in the Americas[230]. Environmental conditions in different American locations are suitable habitat for *Anopheles* mosquitoes, which acts as the disease vectors[231].

a) Plasmodium falciparum

P. falciparum causes a severe form of malaria, which can produce episodes of cerebral malaria, with high index of mortality[232]. Cerebral malaria is associated with changes in cytokine expression and sequestration of *Plasmodium* infected erythrocytes in the encephalic endothelium[233]. In recent years, this parasite has become a major health concern[230]. Since the first usage of earliest antimalarial compounds in the 17th century till today, most of the first line antimalarial drugs have resistant strains[234]. The presence of determined SNPs in several *P. falciparum* genes has been associated to drug resistance. Some of those genes and the drugs which affect include: *pfmdr1* (quinine, chloroquine, amodiaquine)[235–240], *pfcrt* (chloroquine, piperaquine, artemisinin, quinine)[241–248], *pfmrp1* (sulfadoxine-pyrimethamine, chloroquine, quinine, pyronaridine)[245,249–251], *kelch13* (artemisinin)[243,252–256], *pfdhps* (sulfadoxine-pyrimethamine)[257], *pfama1* (spiroindolones,

pyrazole-amides)[258,259], and several other genes and pseudo-genes[243,245,257,259–261].

Most of present-day *P. falciparum* diversity is found primarily in South East Asia, despite *Plasmodium* parasites are thought to be originated from Africa[138]. The increase in diversity in South East Asian countries is attributed a recent bottlenecks and founder effects caused by the introduction of antimalarial policies, and population movements[243,262]. Additionally, recent aDNA studies have also demonstrated the presence of eradicated *P. falciparum* strains in Europe[91,263]. American *P. falciparum* has a very high genetic affinity to African populations[264]. The presence of P. falciparum in the Americas is linked to multiple independent introductions due to the Atlantic Slave trade[265,266] [Figure 9].

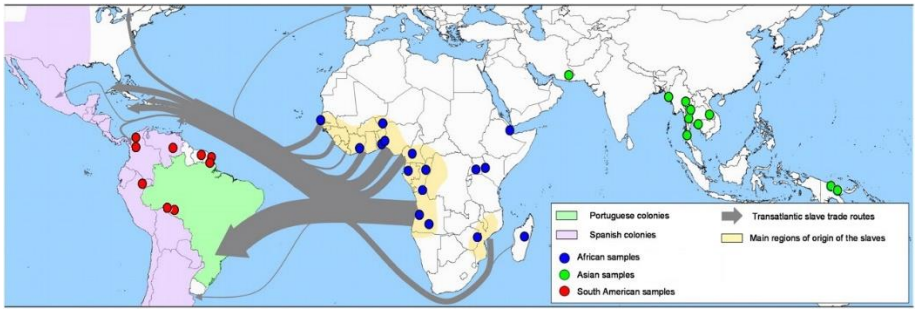


Figure 9. Introduction of *P. falciparum* in the Americas with the Atlantic Slave trade main route (grey arrows). African strains of *P. falciparum* are the primary source population of American *P. falciparum*, although Asian and Southern European strains could have also contributed. From Yalcindag *et al.* 2012[265].

b) *Plasmodium vivax*

P. vivax, in contrast to *P. falciparum*, is characterised by a milder infection, but has the capacity of becoming latent and producing relapses months after[267,268]. This property is conferred by a special stage in the parasite's life cycle known as *hypnozoite*, stage in which the parasite become dormant in the liver[268]. The world distribution of *P. vivax* is restricted to the Americas, South Asia, South East Asia, and Melanesia, with only a few holds in Africa[230]. P. vivax is the most common agent of malaria out of Africa[230]. As in the case of *P. falciparum,* recent events have favoured the emergence of drug resistances in *P. vivax* populations[269]. The presence of polymorphisms on 4 specific genes has been directly associated with drug resistance; *pvdhps* (sulfadoxine)[270–272] , *pvdhfr-ts* (pyrimethamine, cycloguanil, chloroquine) [273–277], *pvcrt* (chloroquine)[278]*, pvmdr1* (chloroquine,) [279–281]*.*

Although *P. vivax* is originally from Africa, the arise of *Duffy* negative allele in Sub-Saharan Africa made the parasite absent from this region[227,282]. The study of current and past populations of *P. vivax* has allowed to determine the affinity of African, European and Indian samples to present-day American parasites[91,266]. Furthermore, the molecular dating of an eradicated 1942 Spanish strain determined a compatible scenario between the discovery of America and the introduction of *P. vivax* in the continent[283] [Figure 10]. This is

in accordance with historical recordings of the disease[284]. Finally, that same strain has been used to create models which can predict the time of emergence of mutation in genes associated to drug resistance[283]. This has been able to be done thanks to the fact that the European strain predates the introduction of most antimalarial treatments, and hence, its genome has not been selected by this evolutionary force.

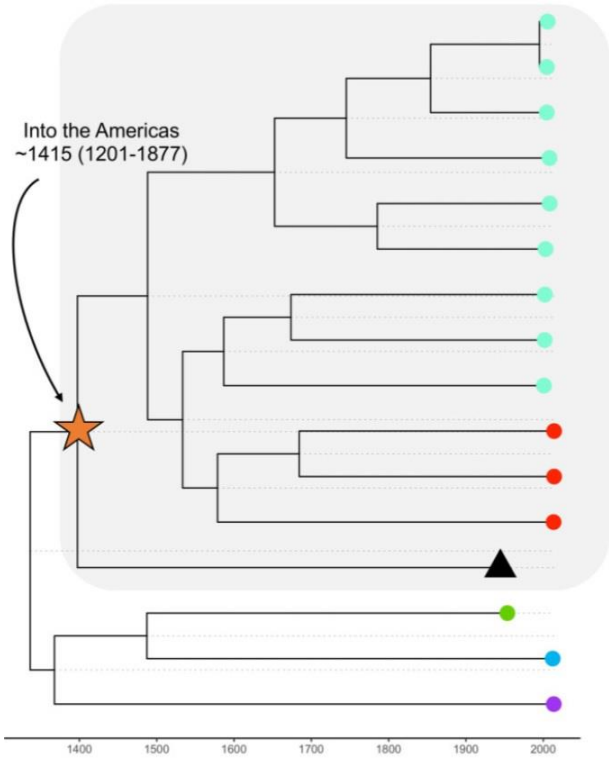


Figure 10. Time-Calibrated phylogenetic tree of *P. vivax* strains from America (cyan and red), East Asia (green), South East Asia (blue), South Asia (purple) and a historical European strain (black). The split between European and American strains supports a post-contact introduction of the parasite. From van Dorp et al. 2020[283].

1.3.3.3 Treponema pallidum

Infections by *treponema pallidum* subspecies englobe different clinical entities such as syphilis, yaws, pinta and bejel[285,286]. Syphilis, a sexually transmitted disease, cause damage to soft tissues and skin lesion in the early stages, and un the most advanced stages induce neurological damage, cardiovascular disease, and necrotic lesions[287]. Endemical treponemal diseases (yaws, pinta and bejel) are transmitted by contaminated fomites or zoonotic jumps, and they primarily attack skin and soft tissue, causing lesions and granuloma[288]. Treponemal diseases can be cured, the main line antibiotic used is penicillin and derivates[287,288]. Only certain strains have reported resistance to other used drugs, in this case, macrolides[289]. Is estimated that venereal syphilis causes more than 5,600,000 new cases each year, while more than 90,000,000 worldwide are at risk of suffering endemic *T. pallidum* infections, especially in developing countries[290,291]. Although of slow progression, mortality of venereal syphilis can reach 60% if the disease is not treated[292], and it is a major causing agent of death and disability worldwide[293].

Syphilis has been regarded as a post-Columbian importation. It has been argued the presence of human American remains predating 1492 with signs of syphilis[294], and the record of outbreaks of an unknown disease after Columbus return to Spain as syphilis[294,295]. In support of a pre-Columbian origin in

Europe, numerous authors suggests the possible presence of ancient and medieval bodies with sings of Treponema pallidum infection, but those have not been tested using molecular methods[296–298].

Genetic studies do not shed much light either, aDNA from an infected individual from Colonial Mexico is dated between 1600-1800[299], while the analysis of global diversity fixes the MRCA less than 500 years ago[300]. New estimates using new ancient North European *T. pallidum* lineages places this MRCA in the year 1000, suggesting a high diversity of *T. pallidum* in Europe prior to Columbus voyage[300]. Those linages found are similar to both venereal syphilis and yaws linages, but different at the same time[300]. The fact the remains where they were found were dated to a post-Contact date, adds a new piece to the puzzle. Different authors suggest the possibility that treponema species where globally distributed with early humans, and that in Europe a yaw like linage existed prior to 1492[294,300]. The importation of American *T. pallidum* could have allowed the recombination between existing lineages and the emergence of Syphilis as we know it[300].

1.3.3.4 *Mycobacterium tuberculosis*

Tuberculosis is a life-threatening infectious respiratory disease. *M. tuberculosis* primarily infects the lungs, causing granulomas, but it can also infect other parts of the

organisms[301,302]. The via of dissemination of the disease is by air droplets exhaled by infected individuals which can directly or indirectly infect other individuals; immunosuppress or malnourished people are especially susceptible[303]. Tuberculosis can produce latent infections[301]. In the later years, the disease has become a major threat due to the emergence of multi-drug resistant *M. tuberculosis* strains[304,305]. Is estimated that a quarter of the world population is infected with the disease, with a reported 10,000,000 new cases annually and causing 1,400,000 deaths[306].

The Old-World origin of *M. tuberculosis* is supported from archaeological studies and genetic studies[307]. From genomic perspective, humans and *M. tuberculosis* has coexisted for millennia, as the selective pressure derived from the host-pathogen relationship between them has left its marks in both human and bacterial genomes[115,308]. *T. tuberculosis* genomic diversity also indicate an African origin of the bacteria, emerging and expanding worldwide during the Neolithic[309–311]. This is in concordance with aDNA data of a 1000 CE Peruvian remains infected by tuberculosis strains which are in the same lineage as *M. pinnipedii*, which infects seals[89]. It has been proposed that, since they share a common ancestor with *M. tuberculosis* 2,500 years ago, marine fauna has imported the bacteria into the Americas, and natives became infected upon seal consumption[89,309]. The diverse population movements and global interconnection of the last 4 centuries has produced

a re-dissemination of the pathogen around the world[309,310,312,313].

1.3.3.5 *Mycobacterium leprae*

Leprosy, also known as Hansen's disease, is an infectious disease caused by the bacillus *M. leprae.* The disease affects mainly to the skin, bones, peripheral nerves, mucosa and testes[314]. The mechanism of transmission of *M. lepreae* is still unknown, but is likely by skin contact or via respiratory route[315]. Human susceptibility to the disease has been linked to certain HLA alleles and the gene NRAMP1 [316–319]. Leprosy is successfully treated with a combination of rifampicin, clofazimine, and dapsone[320], with reported cases of rifampicin resistant strains[321]. Today leprosy is restricted to regions of South Asia, Africa, Melanesia and South America. Due to the slow progress of the disease, the death is not immediate. A study on untreated populations estimated that individuals suffering from leprosy have a mortality rate 4 times higher than the general population[322].

Mentioned in ancient Egyptian, Hindu and Biblical texts[323–325], the earliest evidence of leprosy in history are the remains of an individual from India dating back to the 2nd millennium BCE, where is thought to be originated[137]. Other examples of past infections have been found in Egyptian mummies[326]. Until recent years, it was thought that leprosy arrived at Europe from

India, carried by the Macedonian armies during the Hellenistic period. In contrast to that, the latest genomic studies have placed the origin of *M. leprae* in East Africa, and following population movements, spread worldwide[127]. The arrival of the bacillus to the Americas is linked to 2 main sources, the arrival of European colonists and the Atlantic slave trade from West Africa[88,327]. In recent times has appeared evidence of an inverse zoonotic jump from humans to armadillos, which may act as vector in future infections[142,328].

1.3.3.6 Viral Infections

As we have seen in previous sections, viral infectious rapidly spread after the arrival of first Europeans to the Americas. It has been documented that the first 3 viral infections introduced by Spanish into the Americas were in that order, influenza, yellow fever and smallpox[284]. The paradigmatic case of viral agent which wreaked havoc in Native American population after contact was smallpox[329][Figure 11]. Other virus which have been associated to outbreaks during the 16th century are parvovirus and *Hepatitis B virus*[226].

The causal agent of smallpox is the variola virus. The virus is relatively young, appearing only in the late antiquity – early middle ages[141,330]. The clinical stages of the disease are characterised by around 12 days of incubation period, then appear the first symptoms, fever, malaise, headache, limb pain

and vomiting[331]. Between 3 and 5 after the first symptoms have appear, the skin eruptions develop, starting to mature 2 weeks later and falling down 3 weeks after the symptoms' onset. Variola has a mortality rate between 15 and 45%[331]. Smallpox completely wiped out Native American populations after the disease arrived at the Americas with the Spanish colonist, reaching and ravaging communities even before Spanish could physically contact them[284]. Chronicles tell that the natives, either died because of the disease or by faming, because there was nobody who could take care of them[329]. In Mexico alone, the first smallpox epidemic killed 8 from the original 22 million of people, and the following epidemics of *Cocolitzli* the population dropped to 2 million by 1576[332].

Figure 11. Depiction of Native American suffering from an outbreak during the mid-16th century. Spanish colonist described the epidemic as smallpox. From the *Codex Florentino*[205].

2 Methods

In this section I will explain which experimental procedures I followed to study the different samples I have had at my disposal. Note that despite my PhD has been entirely focused on bioinformatics, I have also briefly explained the experimental methodology used to retrieve the raw data used in the computational analysis.

2.1 Laboratory procedures

2.1.1 Samples

The samples used in the projects are the following:

Four microscopical slides dated between 1942 and 1944 of patients from the Sant Jaume d'Enveja's anti-malarial hospital (Spain)[91]. Dr Ildefonso Canicio was in charge of the hospital from 1925 to 1961. The hospital's patients were local people who worked in the Ebro rice fields, location known for the abundance of stagnant water sources. The patients did not have known record of travelling. The slides were ceded by Dr Canicio's family, from his personal collection. The slides were labelled and stained with Giemsa. The parasites are still visible when looked under the microscope.

Two forensic swabs obtained from one of the newspapers (*l'Ami du Peuple*) which Marat was annotating at the time of his assassination. Due to his health condition, Marat spend most of his last years inside a medicinal bathtub. When he was assassinated by the Girondist's supporter Charlotte Corday (13-7-1793), his sister Charlotte-Albertine kept two issues of the newspaper which were stained by Marat's blood. Albertine gave the issues to the collector François-Nicolas Maurin in 1837, and after his death, were given to the baron Carl De Vinck. In 1906 the newspapers were donated to the Département des Estampes, Bibliothèque National de France, in Paris. One swab was obtained from the blood stain located in the newspaper and the other was obtained from an unstained part of the same document.

The teeth of two putative Spanish soldiers recovered from la Sagrera archaeological site in Barcelona, Spain. The site is attributed to the siege of Barcelona, occurred between 1651 and 1652, in the context of the Thirty Year's war[333]. All the individuals found in the site were males, with ages comprised between 16-40 years, signs of have undergo intense physical activities, and none of them did display signs of a violent death. In chronicles is described that both besiegers and defenders of the city suffered from a Plague epidemic[334]. An upper canine of each individual was extracted for their analysis.

2.1.2 DNA Extraction and Library Preparation

Once the sample is available, the extraction protocol usually starts with the liberation of the DNA from the bone tissue or teeth. First the bone has to be pulverised. Then an extraction buffer composed by 2 main reagents is used to digest the bone or teeth tissue[335,336]. The buffer reagents are proteinase K and Ethylenediaminetetraacetic acid (EDTA). The proteinase K degrades collagen fibres, while the EDTA degrades the hydroxyapatite found in calcified tissue. After the DNA is released from the tissue it has to be captured. A silica column is used to separate the extracted DNA from the buffer[77]. During all then process, strict aseptic conditions have to be maintained in order to minimise the contamination of the sample with modern DNA[337]. The whole protocol has to be performed in a dedicated facility. All surfaces, consumables, tools and instruments have to be cleaned with bleach and ethanol; and irradiated with UV light before and after their use.

To prepare the extracted DNA for its sequencing using NGS techniques it has to undergo a series of chemical process. Since all the projects in this thesis have used double stranded libraries, I am going to focus on them. Is worth to mention that an alternative method known as single stranded library creation exist, which allows to recover short and highly damaged reads[338]. Going back to the double stranded library preparation, the first step is to break up the DNA sequences in

short fragments (this step can be omitted since aDNA sequences are already short). Then the ends are repaired by degrading the overhanging 3' sequences and completing the overhanging 5' ends, creating molecules of the same length. The resultant sequences are known as blunt-end molecules. Finally, the sequencing adapters will be ligated at then ends[16] [Figure 12a]. Is usually after this final step that the library sequences are amplified using PCR.

2.1.3 Sequencing

Multiple commercial sequencing platforms exists. In this section I will explain the basic functioning of the Illumina platform. All the projects of thesis have been sequenced using this technology. The amplified library sequences are loaded onto a flow cell. The flow cell contains oligonucleotide sequences which are complementary to those found in the adapter's sequences. Library sequences are denaturalised in order to hybridise with the complementary sequences found in the flow cell. Then, in a process known as cluster generation, each now single stranded DNA molecule bonded to its primer is amplified creating cluster of up to 1000 clonal copies. The clusters are generated through bridge amplification[Figure 12b]. Once the copies in the cluster are generated, the sequence by synthesis (SBS) can start. The SBS is based in fluorescence, there are fluorescent counterparts for each dNTP (A,C,G,T) with a unique emission spectre. A polymerase

will incorporate those dNTPs to the bound template strand sequence, and the Illumina machine will record which base has been added to the growing strand in each clonal cluster. After the incorporation, the reaction is blocked by the same added fluorophore, the unincorporated dNTPs are washed, and the cycle is started again[339][Figure 12c].

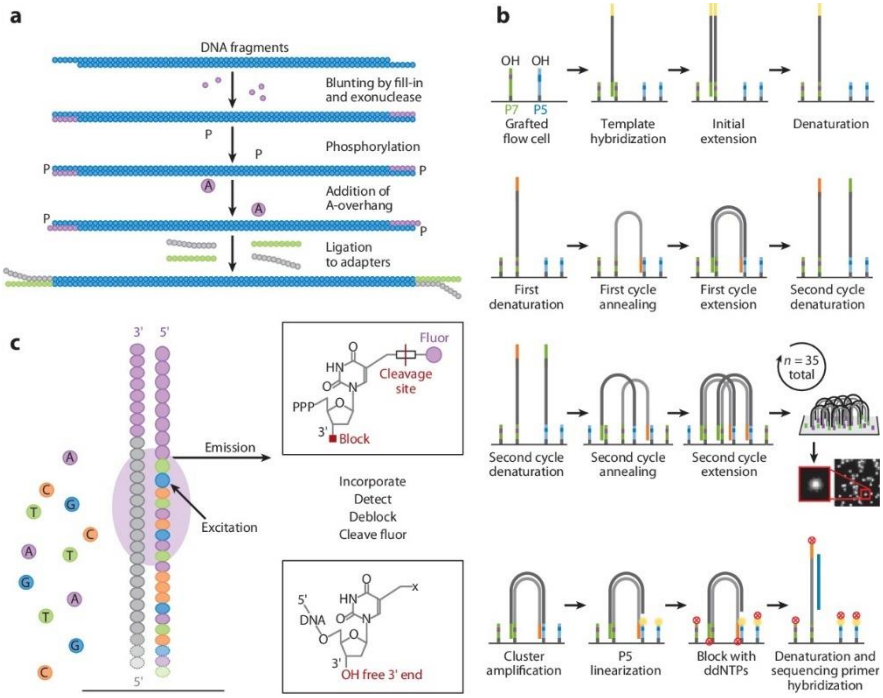


Figure 12. Illumina sequencing process. Library preparation (a), Flow cell ligation and cluster generation (b) and fluorophore light emission (c). From Mardis 2013[340].

2.2 Informatic processing

2.2.1 Adapters trimming

2.2.1.1 The FastQ format and Initial Sequences Quality Control

After sequencing, hundreds of millions of unprocessed reads have been generated. There are a set of steps that those raw sequences have to follow in order to be properly analysed. Due to the high amount of data generated, we are in the need of using computational tools in order to be able to process it.

Raw reads are stored in a specific file format known as FastQ. The FastQ is similar to the FASTA format but incorporate data regarding base quality. The format is composed by blocks of 4 lines. The first line in the always starts with the character '@' and contains the read ID. The second line is the raw read sequence. The third line contains the character '+' and optionally the read ID. The fourth line contains as many characters as in the second line, those characters represent the quality values of the raw read[341].

Before starting to process the raw reads, its recommendable to take a look in to them in order to check that the data is all right. To do this we use FastQC, a software which analyse the raw reads and creates a report with the number of sequenced

reads, their sequence and base quality, base content, sequence length distribution, sequence duplication levels, presence of overrepresented sequences and presence of sequencing adapters[342]. This step is critical since it can point possible errors occurred during the sequencing, or to get a glimpse of the sample's quality before starting the mapping pipeline.

2.2.1.2 Adapter Removal

For the purpose of sequence adapter trimming, FastQ quality trimming and paired end merge I have mainly used AdapterRemoval2[343]. This software is an update of the original AdapterRemoval, which use the same approximation as the later but being more computationally efficient (albeit it also introduces some new features).

The main capacities of AdapterRemoval are adapter trimming from single end and pair end FastQ reads (both one or more adapter sequences simultaneously), demultiplexing single or double indexed reads, reconstruction of pair-end reads' adapter sequence in absence of the original sequence, merge of overlapping pair end reads, reading and writing of interleaved FastQ files, reading and writing compressed files in formats *ggzip* and *bgzip2,* and capability of multi-threading all process for an increased performance.

AdapterRemoval uses a variation of the Needleman-Wunsch algorithm, an algorithm used to compare and align amino acid and DNA sequences[344]. Briefly describing, this algorithm uses a matrix which will compare both DNA sequences and a scoring system (value of events of match, mismatch and gap). The best alignment/s will be decided based on the final score obtained from the matrix after it has been filled. The variant used in AdapterRemoval performs ungapped semiglobal alignments in order to maths the read's 3' end and the adapter sequence 5' end. Additionally, and since the read's 5' end is often not fully sequenced, the alignments are slightly extended allowing for shift towards the read's 3' end, avoiding spuriously aligned adapters and thus, adapter contamination in the resulting sequence. As stated before, AdapterRemoval is able of merging pair end reads. The resultant read have a combined quality based in the combination of the position qualities in the overlapped regions.

As reads tend to have less quality and higher error rates at 5' and 3' ends, AdapterRemoval is able to trim N from both ends (N as a character representing ambiguity) or trim nucleotides based on its qualities.

Due to the particularities of aDNA reads, I have used specific parameters for AdapterRemoval. First, minimum base quality is set to 3, if not the base will be trimmed. Second, N will be automatically trimmed. Third, if reads are pair ended, there will

be merged using the default minimum overlapping length. Only the reads with a final length equal or above 30 bases will be kept for the subsequent analysis. This final length threshold is necessary to avoid ambiguous mappings of the sequences.

Finally, once the raw reads are processed and cleaned from adapter sequences, recommendable to check them again using FastQC. Is important to validate that adapters have been correctly trimmed, and that still are enough sequences after trimming by quality and removing short reads. It is only then when we can jump to the following step in the pipeline, mapping the reads.

2.2.2 Mapping

2.2.2.1 BWA

For mapping most of the raw sequences during the project associated to this thesis, I have used Burrows Wheeler Aligner (BWA)[345]. BWA is a fast mapping software created by Li and Durbin which uses the Burrows-Wheeler transform algorithm to match a large number of short reads against a reference genome, while being able to deal with indels[345,346]. When considering its performance, BWA is more efficient than other available mappers, being capable of processing a million of 70 bp reads in 26 minutes, with a confidence of 90% and error rate of 0.12%[345].

The characteristics of aDNA make not possible to use the standard mapping pipelines. As described in passages above, aDNA tends to be short and damaged. For this reason, BWA is used with slightly modified parameters[347,348]. The first think to consider is the usage of BWA-backtrack instead of BWA-mem. The reasoning behind this is that mem is designed to work with > 70 bp sequences, in which it performs local alignments of the read. This feature is advantageous when dealing with long reads which could include structural variants[349]. We although have to take in mind that aDNA molecules are short, and usually the endogenous content is low, so local alignments could lead to a sort of erroneous mapping against possible contaminants[347].

In the other hand backtrack performs as efficient as mem with short reads, but its capable to perform a global alignment of the read against the reference. To do this we have to disable the parameter seeding (flag -l)[347]. The seed is the length of a read's subset which the mapper will use to carry out the first steps of the alignment, and from which will expand the alignment. By assigning a value larger than the length of the read, we are forcing the software to use the whole read to find the best match to start the alignment from, thus correcting for erroneous matches induced by contamination. The default value is 32. Using an arbitrarily long length it is enough to disable it given reads are short (for example -l 1024).

Another parameter to take in to account it is the edit distance (flag -n). The edit distance is a metric which reflects the degree of similarity between to sequences. In the context of aDNA mapping, the misincorporation of A and T nucleotides at the reads' ends due to post-mortem damage, increases the edit distance between the sequenced reads and the reference. If not relaxed, sequences will not properly map. We can assign two different types of values to this parameter. If we use an integer, it will represent the maximum edit distance between the sequence (i.e. 3). If instead we use a float, it will represent the fraction of missing alignment given a un uniform base error ratio of 2%. The ventage of using the float is that it will change the maximum edit distance depending on the read's length. The default used by BWA is 0.04. The lower this value is, the higher both the maximum edit distance and the number of aligned sequences will be. In the other hand, contaminant sequences could spuriously map when the value is too low[347]. For most of the analysis carried through the thesis I have used an edit distance value of 0.01.

The last parameter which is worth to mention when facing aDNA sequences is the gap open penalty (flag -o). The default value is 3. Lowering the gap open penalty allow for more recovered sequences. The value used for most of the analysis the value used was 2.

Using all this parameter with their modified values, we conclude the first part of BWA's backtrack algorithm, aln. From here we obtain genomic coordinates of the reads, which we will properly align using *samse* or *sampe* (for single or pared end read respectively). In this last step we will use the BWA default values. This will result in a mapped genome in SAM format.

As a brief resume of this segment, the characteristics properties of aDNA make it necessary for modified pipelines in order to be able to properly recover its reads. aDNA damage is the issue we have to deal with, since it increases the distance between the sample and the reference genome. Other particularities of aDNA such as contamination are not usually dealt with when mapping, but downstream the pipeline. Finally, it is of importance to mention that despite reads are mapped, we still have to filter the mapped reads in order to retrieve only high-quality unique sequences.

2.2.2.2 Sequence Alignment Map format

Sequence alignment map (SAM) is the format in which sequences aligned against a reference genome are stored. SAM files are TAB-delimited text files composed of 2 different parts[350]. In this section I am going briefly explain its structure and the software used to edit it.

The first part of a SAM file is the header and its optional. Each of its lines start with the symbol '@' followed by two-character tag and code (i.e., @HD:VN). The main required elements in the file's header are:

- [@HD:VN] The version of SAM.
- [@SQ:SN] The name of the reference genome file used.
- [@SQ:LN] The length of the reference genome used.
- [@RG:ID] The read group ID. Can be more than one. It will appear in each read record.
- [@PG:ID] The programme group ID. Can be more than one. It will appear in each read record.

The second part of the SAM file is the alignment section. The lines are TAB-delimited and contain at least 11 obligatory fields. These fields are in the order the appear:

- QNAME: Is the query template name. It is a string which are sort of an "ID" of the read we are processing. Depending on the software used to create the reads, or the source of the repository from they proceed, they could present different structures.

- FLAG: It is a numerical value composed of bitwise FLAGs. It can describe multiple characteristics of the read (i.e., if it is an optical duplicate or if tis mapped).
- RNAME: It is the name of the alignment's reference sequence. An example of this could be the name of a chromosome find in the human reference genome ; chrX.
- POS: It is a 1-based number. This number is the reference genome position in which the left-most (first base) of the read is mapped.
- MAPQ: It is the mapping quality. Is logarithmic value which indicates the probability of an erroneous mapping position.
- CIGAR: The CIGAR string is a composite of numeric positions and characters. Each character represents possible events that can happen when the read is aligned to the reference genome, for example deletions (D) or mismatches (X).
- RNEXT: Is the RNAME where the mate read is aligned. (Only relevant when mapping paired end reads).
- PNEXT: Is the POS where the mate read is aligned. (Only relevant when mapping paired end reads).
- TLEN: Is the distance between paired end reads. It can be calculated as [end of the first mate] – [start of the second mate] + 1.
- SEQ: It is the nucleotide sequence of the read.

- QUAL: Is the same ASCII base quality found in the FastQ plus 33.

Since SAM files are heavy due to all the data they contain, they are usually compress in a *bgzip* file known as BAM. Those BAM files must be accompanied by an index file (BAI) in order to speed up the process of specific coordinate/region data retrieval.

To finalise is important to remark that SAM/BAM files can contain multiple other fields or tags in both header and alignment parts. Additionally, there are restrictions in characters, specifications in certain flag values and other characteristics which are not explain above. Again, since SAM/BAM files are computationally speaking heavy (up to hundreds of Gigabytes), is nor feasible nor advisable to edit the files manually (or using custom made scripts). Is not common then to deal with most of those not mentioned flags and restrictions. Instead, we use specialised software to filter BAM files and edit them.

2.2.2.3 BAM processing and filtering

As described in the last segment, is not advisable to work by ourselves with SAM/BAM files. In order to filter the mapped reads, we use mainly two software platforms, SAMtools and picard[350,351].

SAMtools is a set of utilities used to handle SAM or BAM files. Picard is another set of tools to manage SAM/BAM files, but it can also use to manipulate other type of files such as VCF. In the basic aDNA mapping pipeline we use both. SAMtools is used to compress data and filter it by mapping quality. Picard is used to replace ID groups and remove PCR or optic duplicated reads.

With SAMtools we compress the SAM files outputted by BWA into a BAM file using the tool view (view -Sbh, indicating that the input is a SAM and the output BAM). The ventage is that once the BAM is created, is not necessary to decompress it. When the BAM is created, we then use picard in order to replace the @RG:ID and @PG:ID, a step necessary because different ID can alter the way BAM files are merged and ultimately, the final results. After this if there is more than unfiltered BAM file for sample/individual we merge them. The final process would be to extract map reads with SAMtools, remove duplicated reads using picard and extracting q30 reads again with SAMtools.

It is important to note that by default, BWA output both mapped and unmapped reads. At the end, only mapped reads will be present in the final BAMs, but unmapped reads can also be stored and used for other type of analysis.

With this, we finalised the standard mapping pipeline I have used during my thesis. The following segments will explain the different set of tools used to determine the quality, aDNA damage, contamination, and variant calling of a final BAM file.

2.2.3 Post Mapping processing

2.2.3.1 Coverage

A useful statistic which sometimes can hint us the overall quality of a sample is its average coverage (a.k.a. coverage, depth, depth of coverage or average depth of coverage). By definition, the coverage is the number of times a position of the reference is overlapped by one or more mapped reads. By averaging each position's coverage in the genome, we obtain the average depth of coverage. Is necessary to remark that the average coverage not always is a good indicator of the sample's quality nor authenticity. For example, contaminant sequences can be susceptible to be mapped against conserved regions (in a case like this we should expect a small number of regions with exceedingly high coverages, while the majority of the reference genome is not covered). A value which can complement the coverage is the reference recovered fraction, which is the percentage of the reference genome positions which are covered by at least one read.
Given this, from a 1 average coverage sample (1X) we could expect to have 100% of the reference recovered but is not

typical. In principle there are 2 factors which can alter the reference recovered fraction. The first is the mappability of the reference genome and the second is the reads we are mapping by themselves.

The mappability is the fraction of the genome which can be mapped against using reads of a specific length. It is almost never a 100% because of the presence of low complexity regions or the presence of copy number elements. In those regions, reads will not properly map, hence reducing the mappability.

The other factor which affects to the coverage is the quality of the sample's reads. For example, a highly degraded or old sample will present less reads in comparison to a better and younger sample. The genetic distance between the reference genome and the sample; despite the sample could have a high depth if mapped against a genome of its specie, when mapped against a closely related specie, only conserved regions will map but at a high depth. The last scenario that can be faced is the presence of copy number variant in our sample; the deletion of a region could lead to less coverage than expected. To ascertain if the reference recovered fraction of a sample adjusted the expected, we assume a random distribution matching the actual coverage if the sample is authentic, given the presence of r reads of length l using the following formula:

$$c = 1 - \prod_{i=1}^{N}(1 - \frac{l_i}{g})^{r_i}$$

in which N are the different l_1 to l_N, read lengths with counts r_1 to r_N. This value must be corrected by multiplying c for the mappability of the reference genome, otherwise the actual coverage could not match the expected [29].

To obtain the average depth of coverage and the recovered fraction of the genome I have used Qualimap2[352]. Qualimap2 is a computationally efficient set of tools used for the quality control of high-throughput sequencing data. For the quality control of the filtered BAMs obtained during the mapping pipeline, we use *bamqc* option. This will output a file with different graphs values, including the depth and fraction recovered, but also GC content, base read composition, coverage distribution, presence of duplicates, among others. Other software such as GATK, SAMtools or bedtools can be used to calculate the average depth of coverage[353,354].

2.2.3.2 Post-Mortem Damage Detection

As explained in a previous segment, DNA damage is characteristic, specially misincorporations of C to T in 5' end (whit the complementary substitution G to A in 3'). The detection of aDNA damage is crucial to assess the authenticity of a sample, but also to value the approximation to deal with

degree of damage. During the fulfilment of this thesis, I have mainly used two software to characterise aDAN damage patterns, PMDtools and mapDamage2.0.

2.2.3.2.1 PMDtools

Starting with PMDtools, is a set of different utilities for aDNA published by Skoglund and colleagues for both aDNA damage characterisation and read filtering[355]. The basis can be described as following:

PMDtools considers a scenario with 3 non mutually exclusive events which explain a substitution of C to T. First an actual biological polymorphism occurring at a rate π (assumed to be 0.001); second a sequencing error occurring at a rate ε (which depends on the read quality, $\varepsilon = 1/3 \times 10^{-Q/10}$); and third event, which is the degradation of DNA occurring at a rate D_z (this value is obtained from empirical observations). The combination of the different possibilities under a model of PMD (M_{PMD}) results in the following formula:

$$P(Match|z) = (1 - \pi) \cdot (1 - \varepsilon) \cdot (1 - D_z) + (1 - \pi) \cdot \varepsilon \cdot D_z + \pi \cdot \varepsilon \cdot (1 - D_z)$$

While under a model of no PMD (M_{NULL}, D_z= 0 for all distances z), then the probability of a mismatch C to T for any particular site is :

$$P(Mismatch|z) = 1 - P(Match|z)$$

An approximation to the value of D_z is:

$$D_z = (1-p)^{z-1} \cdot P + C$$

Where *z* is the distance to the read end terminus; and P is assumed to be 0.3 and C 0.01 for several thousand-year-old samples. For older, more damaged samples the value of P is increased, and the following approximation is used (in which X is 0 when S_i is a Mismatch or X is 1 when S_i is a Match):

$$(L(M_{PMD}|S_i) = X \cdot P(S_i = Match|M_{PMD}) + (1-X) \cdot P(S_i = Mismatch|M_{PMD})$$

To finally assess if the read alignment originated under a PMD model, the natural logarithm is used:

$$PMDS = \log\left(\frac{\prod_i^l (M_{PMD}|S_i)}{\prod_i^l (M_{NULL}|S_i)}\right)$$

Were *l* being the positions in the read S. Higher values of the PMD score (PMD) indicates a more damaged, and putatively, older read.

Given this information, PMDtools is capable of calculate the PMD score of all reads present in a BAM file. PMD score distribution is characteristic of the age of the sample and its level of contamination. PMDtools can be used to filter contaminants based in a threshold PMD score. Additionally, PMDtools can display the frequency of missincorporated bases at the read's ends [.

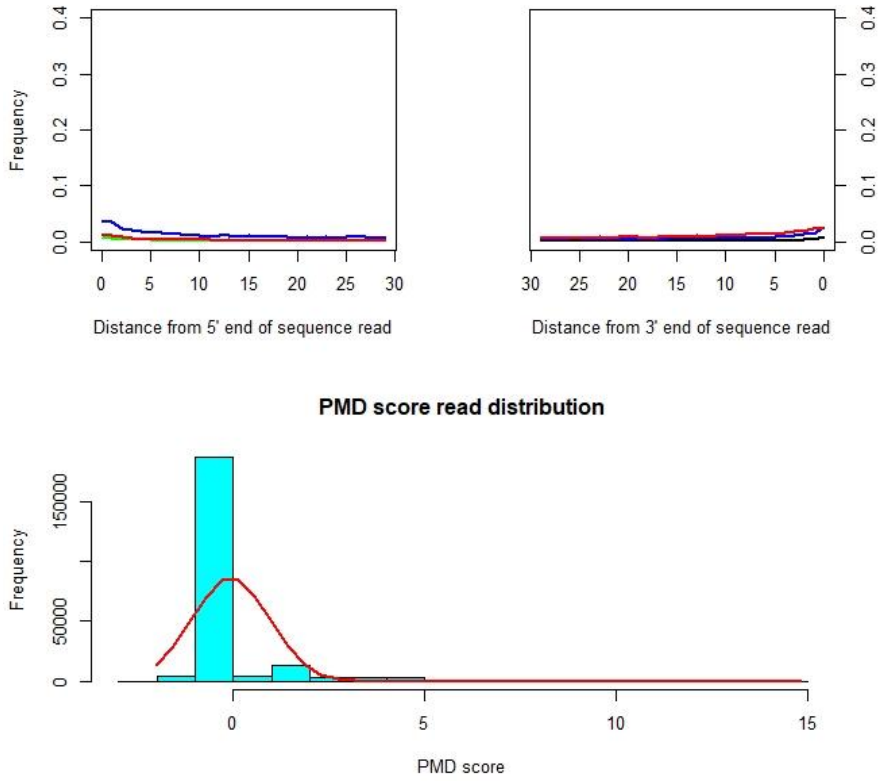


Figure 13. Example of the damage patterns of a historical *P. falciparum* strain estimated using PMDtools. On the top are the frequency of base misincorporation at the ends of the reads (5' end left and 3' end right), x axis describes the position in relation to the end. On the bottom there is the reads' PMD score distribution. The data was visualised using R base[356].

2.2.3.2.2 mapDamage2.0

mapDamage2.0 is a tool published by Jónsson and colleagues[357]. The software is an improvement of a previous version, mapDamage, adding a Bayesian framework[358]. It relies in the aDNA damage model described by Briggs and colleagues and assume that actual biological mutations and aDNA damage mismatches are independent. Given this assumption, aDNA damage occurrence depends only on the distance to the reads' ends.

mapDamage2.0 mutates bases using a Hasegawa, Kishino and Yano (HKY) transition matrix and then, independently, it adds aDNA damage on top of those mutated bases. The authors used a series multinomial distribution to describe the position specific substitution for any given base ($S_{A,i,}$, $S_{C,i,}$, $S_{G,i}$ and $S_{T,i}$).

$$S_{A,i} \sim Mul(D_a, (1,0,0,0) \cdot \Theta(\mu, p) \cdot P_{dam}(\delta_d, \delta_s, \lambda, v, i))$$

Θ is the HKY matrix, and P_{dam} is the DNA damage transition matrix, assuming that cytosine deamination is the main cause of nucleotide miss incorporation in aDNA post-mortem damage. P_{dam} can then be described as:

$$P_{dam} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & (1-p_{ct}) & 0 & p_{ct} \\ p_{ga} & 0 & (1-p_{ga}) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

And the specific damage probabilities C to T and G to A are:

$$p_{ct}(\delta_d, \delta_s, \lambda, v, i) = v_i(\lambda_i \delta_s + \delta_d(1-\lambda_i))$$
$$p_{ga}(\delta_d, \delta_s, \lambda, v, i) = (1-v_i)(\lambda_i \delta_s + \delta_d(1-\lambda_i))$$

The substitutions C to T and G to A are assumed to be caused by actual biological mutations or by aDNA damage derived misincorporations. The probability that one of those changes to be caused by aDNA damage misincorporation at position *i* in the read is:

$$P_{dam}(i) = \frac{\Theta_{c,c} \cdot p_{ct}(i)}{\Theta_{c,c} \cdot p_{ct}(i) \cdot \Theta_{c,t}}$$

Finally, a correction based on base quality scores can be added. Base quality scores are extracted from the BAM files for each position *i* in each read *r.*

$$P'_{err}(i,r) = 1 - \left(1 - P_{perr}(i,r)\right) \cdot \left(1 - P_{dam}(i)\right)$$

mapDamage2.0 needs the reference genome used to map the sample, and the BAM file. Contrary to PMDtools, mapDamage generate a series of plain text output and plot files, providing of

a graphical interpretation of the results, and specifically of the frequency of C to T and G to A misincorporation at reads' ends [Figure 14].

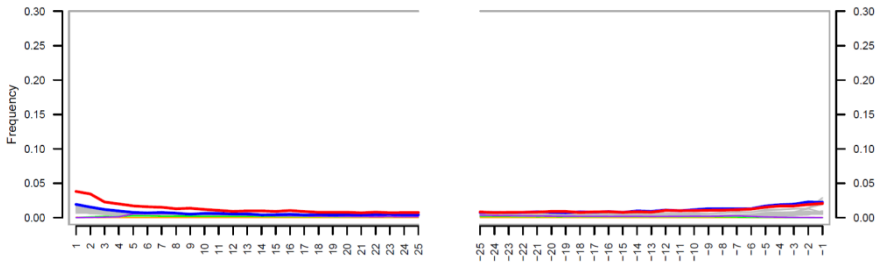


Figure 14. Example of the damage patterns of the same historical *P. falciparum* estimated using MapDamage2. The frequency of misincorporations is represented for the 5' end (left) and the 3' end (right). The x axis describes the position starting from the read's end, so 1 would be the last base. This is the default plot generated by MapDamage2.

2.2.3.3 Genetic Sex Determination

When working with human remains, one of the first things to look at after mapping is the chromosomal sex. Identifying the genetic sex can be useful when the anthropometric measurements are not concluding, and additionally, can evidence the presence of contaminant sequences.

During my thesis, I have used the approximation published by Skoglund and colleagues[359]. In it, is assumed that an individual is female when the presence of the Y chromosome cannot be determined, and male when the opposite is true.

Following this approach, sex assignment is performed by calculating the proportion of reads mapped against the Y chromosome (n_y) against the total reads mapped against both sex chromosomes ($n_x + n_y$). This can be expressed as:

$$R_y = \frac{n_y}{(n_{x+}n_y)}$$

and adding a 95% confidence interval:

$$R_y \pm \left(1.96 \cdot R_y \cdot \frac{(1 - R_y)}{(n_{x+}n_y)}\right)$$

Individuals are considerate female when its R_y upper CI boundary value is below 0.016. In the contrary, the sample is considered male when its R_y lower CI boundary value is above 0.075. Those values were obtained from empirical observations.

2.2.3.4 Contamination

As explained in aDNA characteristics paragraph, modern DNA contamination is serious problem with working with aDNA. The problem gets exacerbated when dealing with human samples, since modern contaminants can alter. Is for this reason that a series of tools and strategies have been developed in order to detect, and to certain degree overcome, the presence of

modern contaminant sequences. In this regard, PMDtools, although not specifically created to detect contamination, can be used to purge modern DNA sequences based on their PMD score.

2.2.3.4.1 Mitochondrial Contamination

Mitochondrial DNA is a valuable resource in population genetics, since is inherited via maternal line. It is also remarkable its abundance, multiple copies of the mitochondrial genome are present in a single cell. Having this in mind, mDNA is an informative and relatively abundant source of genetic material in ancient samples. In the other hand, those same characteristics make it very easy for an investigator to contaminate the sample with their own mitochondrial DNA.

A very simple procedure to get an approximate range of contamination in a sample is to observe the ratio of discordant positions associated to a mitochondrial haplogroup. A mitochondrial haplogroup is a combination of determinate variants in specific positions. Those variants can be shared by two different haplogroups, but usually they are exclusive to them. By counting the number of reads that support a determinate allele in haplogroup defining position, for each position found in the haplogroup, we can get an estimate of contaminant mitochondrial haplogroups. This approximation is far from perfect. aDNA post-mortem damage can be

interpreted as real variations, especially when the depth of coverage of the sample is not high. In those extreme situations, interpretation of the genomic context of the variant using a genome browser could be needed.

A more sophisticated approach I have used during this thesis is the one applied by schmutzi[360]. This software relies in the identification of contaminant sequences by the presence of deamination patterns, read length distribution and SNPs. Schmutzi first creates an estimate of contamination based on the proportion of reads which present deamination in the n las bases of the reads (this parameter can be modified). This alone can provide us of vital information regarding the amount of contaminant sequence. From here we can take a step further by the iterative assembler feature of schmutzi. The programme will then compare the sequences against a human mitochondrial database, which contains both modern and archaic variability. With this, it will create consensus sequences of the endogenous genome and the contaminants. The disadvantages of this software are that since it relies in finding deamination along the read's ends, if the analysed sample is fairly recent and/or well preserved, the detection of a modern contaminant could be hindered. Furthermore, for an optimal detection of contaminants it needs a quantity of mapped reads not feasible for aDNA studies.

2.2.3.4.2 Nuclear Contamination

Although the data provided by the mitochondrial genome could be useful when performing certain ancestry analysis, the fact that it is a uniparental decrease its relevance in other types of population genetic analysis. Added to this, since the number of mitochondrial genome copies is not correlated to the nuclear genome, the presence of contaminants in mDNA could not be representative of the whole sample. As mentioned before, if there is the suspicion of contamination, we can simply use PMD tools to get rid of the lowest PMD score reads.

In the other hand, there are other methods to deal with contaminant sequences in the nuclear genome. The method I use for most of the analysis (when possible) is ANGSD[361]. ANGSD calculates the heterozygosity levels at the X chromosome to test the presence of contaminant sequence. The X chromosome is only haploid in males, so we can only use ANGSD in those individuals when their genetic sex is XY. To do this we first perform a recount of the allele frequencies in a X chromosome 100kb window using ANGSD. Then we compare those calling against the HapMap database, which contains global allele frequency estimates. If the positions' frequencies deviate from those global frequencies, it is an indication of contaminant presence, and the estimates will be calculated. Again, as it happens with schmutzi, ANGSD could not perform well if there is not enough data available.

2.2.4 Variant Calling

Variant calling is the process in which variant positions are identified in mapped genomes. They involved some sort of software which compares a mapped genome in BAM format against the reference genome. When working with aDNA, we are usually limited to call SNP, with exceptions when the samples' quality is high. Here I am going to describe the different approximation I have followed depending on the sample characteristics.

2.2.4.1 Pseudo-haploid calls

The process of pseudo-haploid is usually used when a sample coverage is low. In this case SNPs are called by randomly choosing a read for the interrogated position. The selected read will be the representative of a haploid genotype, hence the name of pseudo-haploid calling. This approximation is effective to correct for the biased induced by aDNA damage or reference bias. To obtain the list of positions to call we can use a SNP reference panel. For humans, I have mostly used the Human Origins panel since it contains both informative SNPs and modern and ancient human genotypes[28,362]. When working with other species I create a custom curated reference panel using published high-quality genomes called using GATK (see the following section).

The process to generate the pseudo-haploid genotypes is rather simple. First, we generate a pileup from BAM. A pileup is a tab-delimited format which contains a line for every position found in the reference genome[350]. The file is comprised by at least 6 columns; the first column corresponds to the chromosome, the second the position, the third the reference base, the fourth the number of reads covering the position, the fifth what bases have the reads for the particular position, the sixth these bases qualities. It can contain additional fields. To create this pileup, we have to ignore Read Group tags in the BAM file (-R), disable the alignment quality computation (-B), supply a set of positions to call (-l followed by the file) and supply the reference genome used to map the sample (-f followed by the file). Once the pileup is created, we use the software pileupCaller to generate the genotypes[363]. It is important to use the random calling mode (-m RandomCalling) and to use the reference SNP panel, so the software can distinguish which alleles are ancestral or derivate (-f followed by the file).

With this process a genotype file for each individual is created. The resultant genotypes only have to be merged with the desired dataset and are ready to be used in population genetic analyses.

2.2.4.2 GATK

This approach is the one I take when the sample has enough coverage. The genome analysis toolkit (GATK) is a set of different utilities to identify SNP and indel in genomic data[353,364]. GATK has numerous utilities, but I will mainly focus on the variant calling functionality. UnifiedGenotyper is a deprecated calling algorithm used by GATK, which call genotypes using a per-base method[365]. In contrast, the recommended HaplotypeCaller (is the only one available from the 2 in GATK version 4 or higher) call variants by creating a local de-novo assembly of the region of interest[366]. There is a problem, HaplotypeCaller is designed for high-quality datasets, and from personal experience, it can generate spurious calls when trying to realign aDNA substitutions.

Taking now a look in how UnifiedGenotyper works, it uses a Bayesian algorithm which can both genotype and calculate allele probabilities. Allele probabilities (AA,AB,BB) are estimated for the A reference allele and the B alternative allele in N individuals. It takes in account the aligned bases at a specific position (D_i) in an individual (*i*), where the genotype likelihood (GT_i) of observing the D_i bases for the genotypes AA,AB and BB given this equation:

$$\Pr\{D_i|GT_i\} = \prod_j Pr\left\{D_{i,j}\middle|GT_i\right\}$$

$$\Pr\{D_i|GT_i = AB\} = \frac{\left(\Pr\{D_{i,j}|\, A\} + \Pr\{D_{i,j}|\, B\}\right)}{2}$$

$$Pr\{D_{i,j} \mid B\} = \begin{Bmatrix} 1 - \varepsilon_{i,j} \\ \varepsilon_{i,j} \cdot \Pr\{B\ is\ true \mid D_{i,j}\ is\ miscalled\} \end{Bmatrix} D_{i,j} = B, otherwise$$

where $\Pr\{D_i|D_{i,j}\}$ is the probability of observing a base ($D_{i,j}$) under a hypothetical genotype (GT_i), $Pr\{D_{i,j} \mid B\}$ and $Pr\{D_{i,j} \mid A\}$ are the probabilities of observing a base ($D_{i,j}$) given that the actual present base is A or B (respectively), $\varepsilon_{i,j}$ is the probability of a miscalled base given the quality score of the base ($D_{i,j}$) and $\Pr\{B\ is\ true \mid D_{i,j}\ is\ miscalled\}$ is the probability of B $_{\text{is true}}$ being the true base given that b is a miscall.

To generate a call considering different individuals at the same time other assumptions have to be made. Being $q_i = \{0,1,2\}$ the number of alternate B alleles which are carried by the individual *i*. Then, $q = \sum_i^N q_i$ is the number of chromosomes carrying the B allele among all individuals N. Given this:

$$\Pr\{q = X\} = \frac{(\Pr\{q = X|\}\Pr\{D|\, q = X\})}{\sum_\gamma \Pr\{D|\, q = \gamma\}}$$

$$\Pr\{q = X\} = \begin{Bmatrix} \theta/X \\ 1 - \theta \sum_{i=1}^{2N} 1/i \end{Bmatrix} X > 0 \; otherwise$$

$$\Pr\{\mathrm{D}|q = X\} = \sum_{GT \in \Gamma} \Pr\{D_i|GT_i\}$$

$$\Gamma = \left\{ GT \text{ where } \sum_i q_i = X \right\}$$

Here Γ is the set of all genotype's callings for all N individuals which have q = X B alleles, $\Pr\{q = X\}$ is the expectation to observe X alternative alleles in all individuals' chromosomes(2N) given a determined heterozygosity (θ), GT_i is the genotype for a determined individual i and D_i is the analysed reads for that same individual.

Finally, to compute the probability of a variant segregating at a site at some frequency it used the following formula:

$$QUAL = -10 \cdot log_{10}[Pr\{q = 0|D\}]$$

The final output of all this process is a Variant Calling Format (VCF) file containing variant positions, which carry information of the callings for each position in the analysed data (see the following section). To use GATK UnifiedGenotyper is necessary a BAM file as input and the reference genome used to create this BAM file. There are a multitude of optional

parameters, of which worth mentioning are to emit all confident sites (independently that the position has the reference or derivate allele), the option to use a reference panel for the callings, or the option to only output a list of selected positions (this can save a considerable amount of time when the data is intended to be merged with an already created dataset).

2.2.4.3 The Variant Calling Format

The VCF file is a tab-delimited that, as stated by its documentation, is composed by 3 differentiated sections[367]. The metainformation, followed by the header, and after that, the lines with information for a position in the genome (genotype being included in it). As I have stated before in this thesis, genomic data files are usually impracticably large, and as previously mentioned, is not advisable to edit the files manually. Is for this reason that is important to know what is in each part.

The first part of a VCF file is comprised by the meta-information lines (preceded by the characters '##'). Just to describe them (since they are not mandatory), the different possible types of lines are the file format, information field format, filter field format, individual field format, alternative allele field format, assembly field format, Contig field format, Sample field format and Pedigree field format.

The second part, this time mandatory, is the header. The header is preceded by a single '#' character. It contains 8 fields plus as additional fields as different samples are present in the VCF. The fields are:

- #CHROM: Name of the chromosome.
- POS: Position relative to the chromosome.
- ID: ID of the variant found. If exists, is recommended to use the dbSNP *rs* number for the variant.
- REF: Allele found in the reference. Must be A,C,G,T or N. Multiple bases are allowed, separated by a coma.
- ALT: Alternative allele. Must be A,C,G,T,N or *. . Multiple bases are allowed, separated by a coma. * indicates an upstream deletion. In case that the alternative allele is the same as the reference, the missing data character must be used.
- QUAL: Phred score of the alternative calling (see GATK section).
- FILTER: Filter status of the position. If all determined filters are passed, the PASS tag will appear. In case that no filters are applied, the missing data character must be used.
- INFO: List of additional information tags. Certain subfields might be reserved. Subfields are separated by a colon.

Last, and technically not mandatory, is the Genotype field. There is an additional Genotype field for each individual present in the VCF. The main subfields (present in the INFO field, separated by colon) are:

- GT: The genotype, which encode for the allele values found at this position (0 for reference allele, 1 for alternative allele, and 2 for a second alternative allele if present). The values are separated by the characters '/' (for diploid calling) and '|' (phased date). For example, 0/0 will be a homozygote for the reference allele, 0/1 a heterozygote, and 1/1 a homozygote for the alternative allele. If the call is not available, the missing data character must be used.
- AD: Allele depth, which is the number of reads which support each one of the alleles. Values are separated by a coma.
- DP: Depth of coverage. The number of positions which support the position.
- GQ: Genotype quality, present only in variant sites. Is the phred quality score indicating the probability of an erroneous call.
- PL: Phred-scaled genotype likelihoods rounded to the closest integer.

After the header, there are as many lines as genomic positions in the VCF. The data lines contain the information for each of

the aforementioned fields. Missing data in one of the fields is represented with the character '.').

To finalise, I usually convert VCF files into PLINK format genotype files. In the process information contained in the QUAL, FILTER and INFO field is lost. From the genotype field, only the GT subfield information is conserved. This format is the default input format for certain population genetic analysis software.

2.2.4.4 The PLINK genotype formats

PLINK is a command line-based toolset used for genome association analysis. It can perform a set analysis, from filtering variants or introducing thresholds to more complex analyses such as imputation[368]. PLINK is able to use VCF as input, or generate them as output, but as described previously, information is lost in the process. The two PLINK basic file formats are the plain genotype files (ped and map) and the plink binary files (bed, bim and fam).

The plain genotype files are ped and map. Ped files are tab delimed files containing the genotypes of one or more individuals. The first 6 fields are in that order the family ID, the Individual ID, the paternal ID, the maternal ID, the genomic sex (1 male, 2 female and any other value unknown) and the phenotype. The following fields correspond to the alleles for

each variant, there are 2 fields for each one (the total number of fields is 6 + 2n, where n is the number of variants). Alleles can be encoded in bases (A,C,G,T) or in numbers (1,2,3,4). The corresponding map file is also tab delimed, it contains the variant marker information. The fields are, in that order, the chromosome ID, the rs or SNP ID, the genetic distance in Morgans, and the genomic position in base pairs.

The binary files are the bed (not to be confused with .bed coordinate files), bim and fam. They are all tab delimited files, except for bed which is binary. The fam contains the first 6 column found in a normal ped file. The bim contains the same 4 fields of a map file plus 2 extra fields, the first the reference allele for that variant and last the alternative allele. Managing binary files is noticeable faster that plain files (binary files are processed in seconds). Furthermore, some of the options are only allowed to use binary format.

2.2.4.5 VCF and PLINK files manipulation

Similar to the case of BAM files, is not advisable to manually edit VCF files. VCF manipulation is common, for example for removing low depth positions, selecting a list of positions found in a dataset, merging different VCFs, remove indels, filter by MAF or MAC (Minor Allele Frequency/Count), performing statistics, among others. There are a set of tools to edit VCFs, of which I have mainly used vcftools and bcftools[367,369,370]. The two software are similar, but bcftools is noticeable faster. Both

accept VCF and its bgzip compressed variants. Is recommendable to index bgzip VCF files using tabix[371]. In a similar manner, PLINK can be used to filter variants, in both PLINK file format or VCF.

2.2.4.6 Imputation

As we have been seeing, lack of high-quality data is a constant when working with aDNA. In the context of variant-calling and genotyping, this translates in a high proportion of positions not covered or called. A process which can compensate for the lack of data is imputation. Imputation is a process of statistical inference from which, with the usage of a phased genotype reference panel, missing data in a sample is deduced. The basis behind imputation is that SNPs are associated in haplotypes, so if all but one of the haplotype's SNPs are present in the sample, that missing SNP is likely to be also present. Linkage disequilibrium (LD) also has to be considered when generating the genotype likelihoods.

To fulfil this, I have used the software Beagle[372]. The programme, created by Browning and Browning, is designed to infer haplotypes, impute genotype data and perform genetic association studies. The input for Beagle is the VCF file we want to be imputed. It also needs a phased reference panel (in example, for humans the 1000 Genomes dataset or the Human Origins Dataset[18,28,362]) and a genomic recombination map.

This will generate a new VCF with the genotype likelihoods for the analysed sample.

It is important to bear in mind that imputation is a computationally costly process, more if a large (both in SNP number and/or samples number) reference panel is used. Additionally, imputation has it limits when a sample has too much missing data. In the case human samples used for population genetic analysis, it has been determined that below 0.8X depth of coverage, the imputed genotypes have no better resolution that the ones generated using pseudo-haploid callings[373].

2.2.5 Uniparental markers analysis

For the most part, I have worked only with human mitochondrial DNA. Human mitochondrial haplogroups are characteristic of specific geographical regions, and thus, are valuable to know the origin or affinities of a sample. As seen in the contamination section, they also allow to detect possible sources of modern DNA contamination.

To infer those haplogroups first we have to perform a variant calling using GATK UnifiedGenotyper[365]. The subsequent VCF is then uploaded the HaploGrep2 web service[374]. Haplogrep2 classify the haplogroups using pre-calculated phylogenetic weights for each variant. Those values are based in the

occurrence per position in the Phylotree database[375], and it also takes in to account the mutational stability of the variant. All the process is performed remotely in the web server and outputted in mere seconds. All of this will generate a report in which are specified (for each sample) the top ranked haplogroups, the quality score for the assignation, the haplogroup's polymorphism found in the sample, the polymorphism not found, and the presence of not described variants. Another approximation to generate a valid input for HaploGrep2 is to use the FASTA files generated by schmutzi (section 2.2.3.4.1), being able to analyse both endogenous and contaminant DNA.

Finally, in certain extreme cases where there is low amount of data, I have used a genome browser and the Phylotree database to manually annotate the mitochondrial haplogroup. This is advisable in those cases since it allows to get a view in the context of each SNP found in the sample, and to discard SNPs possibly caused due to aDNA damage.

2.2.6 Population Genetics analysis

2.2.6.1 Principal Component Analysis

One of the standard analysis to detect genetic structure in a population is the Principal Component Analysis (PCA)[376,377].

The PCA is a statistical method which is used to reduce the dimensions of large variant datasets to a few main principal components, which explain most of the diversity in the sample. For example, performing a PCA of European individuals with a SNP panel, and then visualising the 2 first principal components, we get a distribution which resembles the geographical origin of the samples, and thus, Europe's map. This procedure can also hint individual's ancestry. An individual for which its main ancestry components are European and African (50/50), it will be placed at the middle ground between the European and African clusters. Nevertheless, to fully understand the genetic relationship between different populations, other methods are needed.

A ventage of the PCA is its robustness, since it can well tolerate the inclusion or exclusion of SNPs in large datasets. A SNP dataset can be considerable purged and still maintain a similar structure and resolution. This can be used in our advantage, filtering by MAF or missingness, since the process of very large datasets is computationally slow and expensive[376].

To perform a PCA I have used EIGENSOFT, a package designed by Reich's lab. The tool, smartPCA, is also capable of handling high missingness aDNA data. The option lsqPROJECT uses a set of individuals to generate the main component matrix (in this case would be high quality modern samples). Then, the ancient samples where be projected in this

matrix. EIGENSOFT has its own genotype data format (EIGENSTRAT) but is also capable to handle PLINK files.

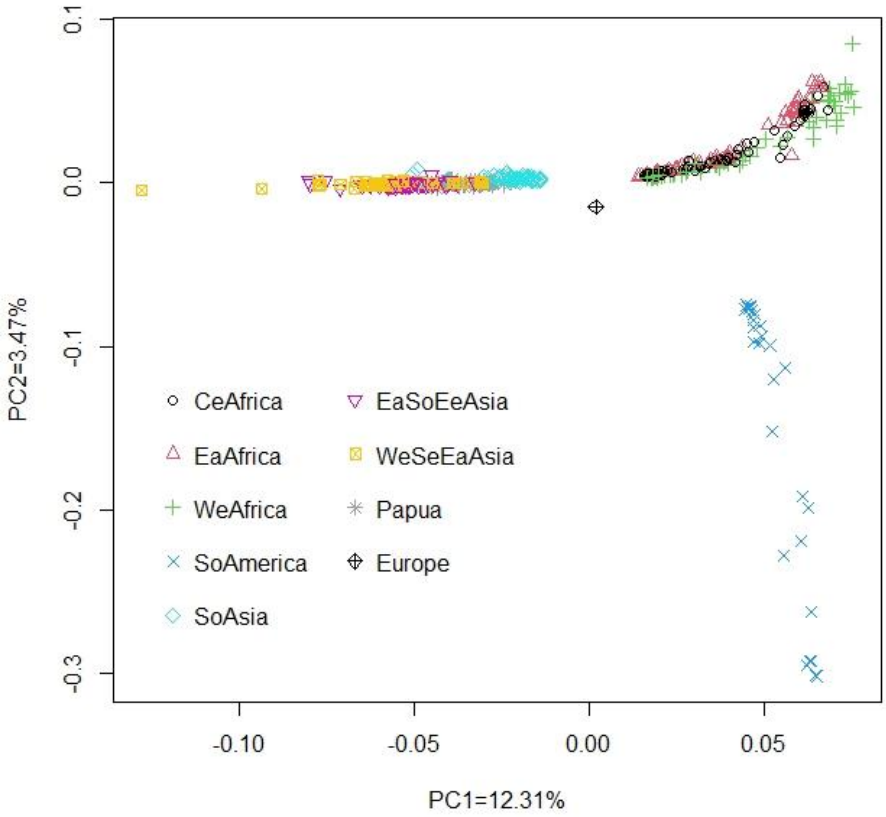


Figure 15. Example of a Principal Component Analysis computed using smartPCA. The same historical strain of European *P. falciparum* is projected into current day *P. falciparum* diversity. The axis displays the percentage of diversity which each Principal Component (PC) displays. This plot was generated using R base.

2.2.6.2 ADMIXTURE

Another approximation for the detection of genetic population stratification is the one used by ADMIXTURE[378]. The software only takes into consideration the global ancestry, estimating the ancestry fraction of a certain number of contributing populations. What differences ADMIXTURE from a PCA, is that the former uses a Model-based approach, based in ancestry proportions and population allele frequencies.

To use ADMIXTURE, we have to supply the number of contributing populations (K) to the ancestry. This number can be originated from knowledge of the sample's populational history or can be inferred by comparison of models. Generally, we delimit a maximum K value and run a number of replicates for each K. The best K will be decided in basis of the value of its cross-validation error (CV). The CV will be calculated as the mean CV for K in each group of replicates [Figure 16].

ADMIXTURE does not consider LD. Is recommendable to prune the dataset for LD, for both removing the SNPs in LD and to reduce the number of SNPs (the larger the dataset, the longer the software it will take to run, and the more computational resources will use). It has been estimated that at least 10,000 SNPs are necessary to detect structure in inter-continental human populations, and up to 100,000 for populations inside the same continent[378,379].

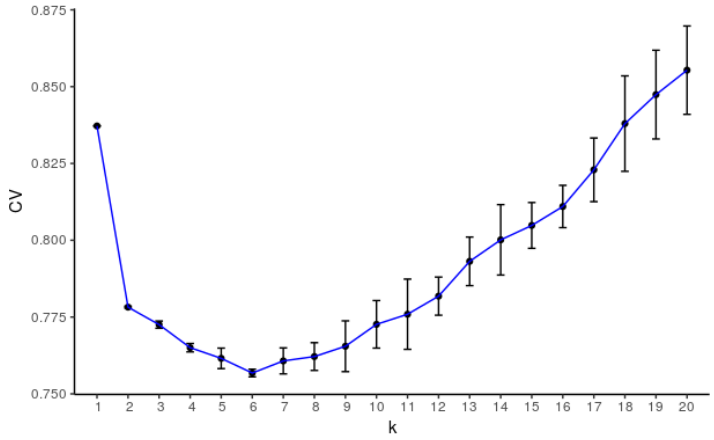


Figure 16. Example of the CV mean values for each different K with 10 replicates of a *P. falciparum* unsupervised ADMIXTURE. The error bars show the CV error standard deviation. In this case, K 6 seems to have the lowest CV error of all the computed K. The plot was visualised using R package *ggplot2*[380].

The programme will generate 2 matrices for each K replicate, one with the ancestry proportions (Q) and one with the allele frequencies for each population (P). Due to the fact that each replicate is not equal, combining and interpreting the resulting Q matrices is not trivial. To do this, we use pong, which both combine the resultant Q matrices for each K and output a graphical visualisation of them [Figure 17][381]. ADMIXTURE can use PLINK (both standard and binary) or EIGENSTRAT format.

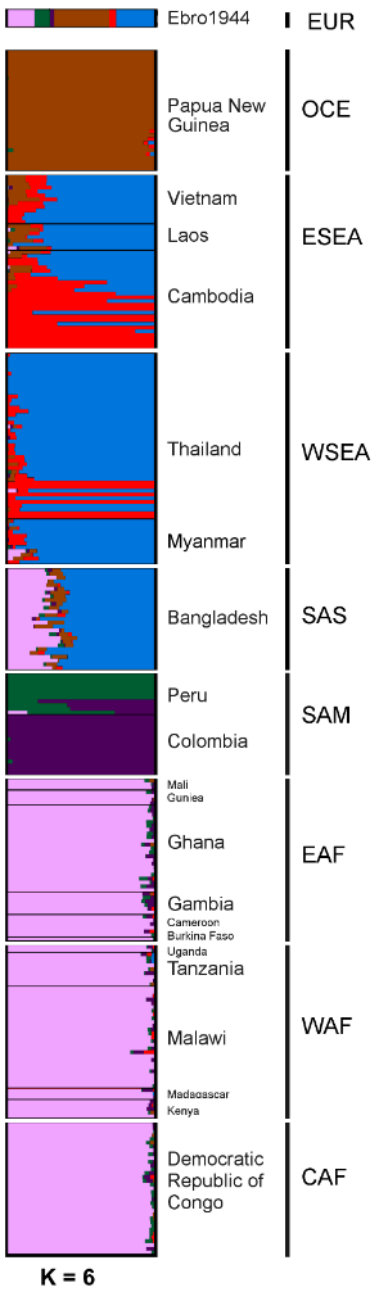


Figure 17. K6 of the previous ADMIXTURE. The optimal K was chosen to be visualised. This plot is a combination of the all the Q matrices resulting from the 10 iterations. Pong was used to combine the matrices and create the visual output.

2.2.6.3 Fixation index and F-statistics

A statistic used to quantify the divergence between 2 populations is the F_{st} or Fixation index proposed by Wright[382]. This index measures the genetic distance between populations using values which range between 0 and 1. A value of 0 implies a scenario of panmixia, while a value 1 implies a total stratification of the 2 populations and thus no shared diversity. F_{st} between two populations can be estimated as:

$$F_{st} = \frac{H_T - H_s}{H_T}$$

where H_T is the total expected heterozygosity found in the populations, and H_S the within subpopulation expected heterozygosity[383]. F_{st} can be calculated using PLINK or vcftools.

Considering a tree model, where each analysed population is located at the tip of a branch, we can interpret the length of the different branches connecting the populations as the genetic drift[384]. The different F correspond to the branch lengths (shared genetic drift) between models of 2, 3 and 4 populations [Figure 18]. Besides its use in population phylogenetics models, F-statistics can be used to model population admixture events[24,384,385], demographic history and events[195,362,386,387], and to find closest relative to extinct populations[388].

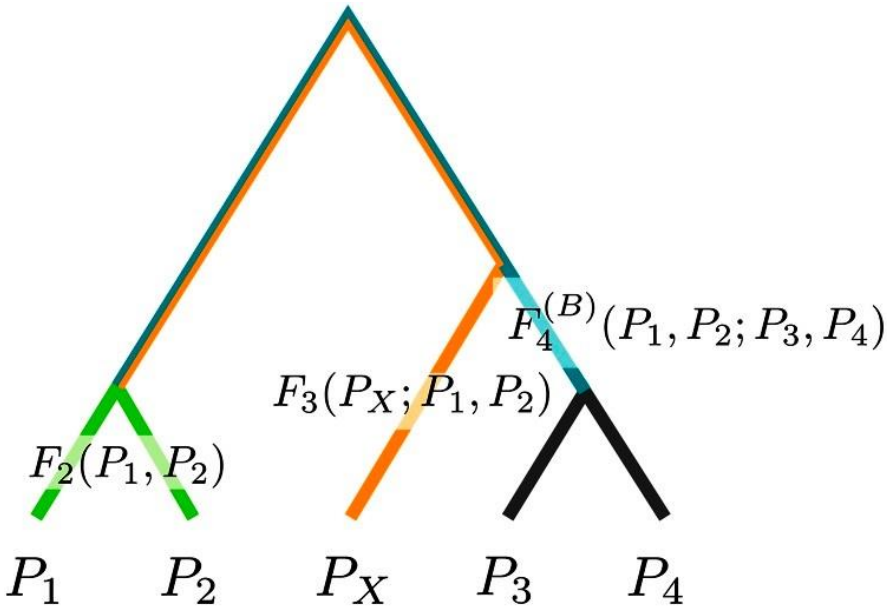


Figure 18. Genetic distances calculated by F-statistics represented in a tree fashion. Here are represented F2 (green), F3 (orange) and F4 (blue). From Peter 2016 [389].

As brief description of each F model regarding a phylogenetic tree [Figure 18]:

- F_2 can be interpreted as the genetic distance (the branch length) between populations P1 and P2. This statistic is similar to F_{st}.
- F_3 can be interpreted as the shared genetic drift between populations P1 and P2, and its genetic distance to an outgroup P3.
- F_4 can be interpreted as the distance between the clade formed by P1 and P2, and the clade formed by P3 and P4.

F-statistics and other related tests are implemented in AdmixTools[387]. Statistical significance is assessed with a jacknife resampling test. Only Z-score values over 3 would be considered as significant. The file system used is the same as in EIGENSOFT.

2.2.6.4 Haplotype Based methods

In the methods describe previously, independent polymorphisms (in our case, SNPs) are used to infer genetic relationships and genetic structure among populations. In contrast to that, Chromopainter uses haplotype to find genetic structure[390]. What the software does is to find the shared chromosomal blocks (the haplotypes) between a set of donor samples and the targets samples. At the end, the target individual's chromosome is defined as a mix of the different donor individual's chromosome. An important remark is Chromopainter needs phased SNP data for both target and donor data. Our sample's coverage can represent a limiting factor in the case is diploid, since phasing is sensible to coverage.

2.2.2.7 Phylogenetics

2.2.7.1 Phylogenetic trees

A representation of the genetic and evolutionary relationships between different individuals is the phylogenetic tree. This tree has a set of differentiated parts: the branches which represent the genetic distance between the different taxa, the leaves which represent the taxa, the nodes which represents the most recent common ancestor (MRCA) between pair of leaves, and the root which is a special case of node which represents the MRCA of all the leaves present in the tree. One of the methods used to build a phylogenetic tree is the Maximum Likelihood (ML) framework[391]. This method assigns the probability in which a given tree has evolved the observed data. Is important to remind that the more probable tree does not have to be the correct one.

To create ML trees, I have used RAxML[392]. This programme needs of a multiple sequence alignment (MSA), the selection of substitution model, and optionally, an outgroup. The MSA can be created by either aligning sequences using and aligner (MAFT or Muscle i.e.[393,394]) or by creating one from VCF. Note that if multiple VCF are used, they have to be called against the same reference genome or they would not be actually aligned. Missing positions have to be replaced with an ‘N’ character. For models, GTRCAT and GTRGAMMA are the

ones recommended for this type of data. GTRGAMMA is regarded as the best model to use, but as is not computationally efficient, when working with large datasets is preferable to use GTRCAT (50 or more taxa)[395].

Once run, the programme will generate a resume file text with all the processes performed, a reduced MSA file, a file with all bootstraps (if the option was selected) and a file with the best tree [Figure 19].

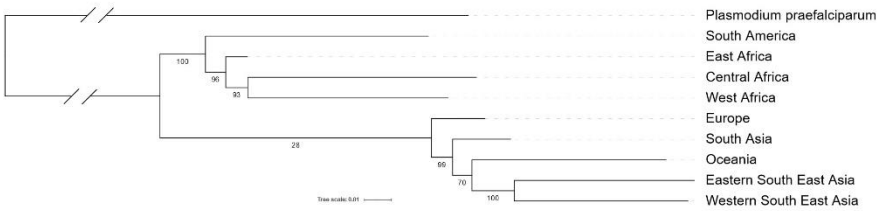


Figure 19. ML tree of *P. falciparum* strains created using RAxML. The tree is rooted to *P. praefalciparum.* As the number of taxa is small, GTRGAMMA model was used. 100 bootstraps were performed, the best tree is the one displayed. Support for each node is displayed at the bottom of the node. Branch length is proportional to the genetic distance between the taxa. Note that the distance between the root and the other taxa is not the real, otherwise, the tree could not be properly visualised. The plot was created using *iTOL*[396]

2.2.7.2 Recombination

DNA recombination is a process in which genetic material is exchanged between one or more chromosomes, generally in regions with high homology. Recombination can occur within a single individual (for example, during meiosis in humans) or among different individuals (such as in bacterial recombination).

Recombination events can alter the phylogenetic tree topology[397], so a dedicated set of tools have been developed in order to remove them from phylogenetic datasets. One of these tools is ClonalFrameML[398], an update of the ClonalFrame software[399], designed specifically to locate recombined regions along the phylogeny of bacterial species.

ClonalFrameML uses as input data a phylogenetic tree, the alignment used to build the tree and the rate of transitions/transversions for the aforementioned alignment. The software will output a new alignment file without the recombinant regions, the list of recombination events and their location, and graphical visualisation of the recombination events and homoplasic regions in respect to the phylogeny.

2.2.7.3 Homoplasy

Homoplasies are substitution events which have arisen independently in different separate linages, but they are not shared with a common ancestor. Homoplasies are determined by their consistency index value[400]. The index's values range between 0 and 1, being the lower values the ones that denote the presence of a homoplasy. As in the case of recombination events, homoplasies can also alter the topology of a phylogenetic tree. In the case of aDNA, they can also be artefactual. The spurious calling of damaged bases in low coverage samples can originate a homoplasy.

ClonalFrameML can be used to remove both recombination and homoplasy events[398]. Another tool which I have used is HomoplasyFinder[401]. This software detects positions with a consistency index value below 1 and removes them from the alignment. Again, as with ClonalFrameML, HomoplasyFinder needs an already build phylogenetic tree and its correspondent alignment. The software then will generate a new alignment without homoplasies, the list of all positions in the original alignment and their consistency index values, and an annotated tree with the present homoplasic events.

2.2.7.4 Time-calibrated phylogenies

Given a constant evolutionary rate, it can be assumed that the branch length is proportional to its age. By having data about the age (date of collection) of each of the leaves found in a phylogenetic tree, we can infer the time at which branches have split, and thus, obtain an approximate date when the MRCA existed [402].

In order to create time-calibrated phylogenies, I have used BactDating, a software focused on bacterial phylogenies[403]. This tool uses a Bayesian framework, but the phylogeny that is being dated has already been created. That allows to a faster computational performance when compared with other software available such as BEAST[404].

BactDating needs of a robust phylogenetic tree (hence the removal of recombination and homoplasies). Additionally, sampling dates for each of sample present in the phylogeny are necessary. The software will then perform a linear regression of the branch root-to-tip distance versus its collection date, which will estimate the mutation rate in means of substitutions/base/year. If there is a strong enough temporal signal, the software can be used to date of each tree node. Ancient samples are useful in this regard since they help to obtain a stronger temporal signal in the phylogeny.

2.2.8 Metagenomics

2.2.8.1 BLAST

BLAST (acronym of Basic Local Alignment Search Tool) is a tool used to find homologous sequences of our query by aligning them against a database[405]. The usage of a heuristic method allow BLAST to find short matches of the query against the database. From this seed matches, the programme will try to extend the match by performing a local sequence alignment. The resultant alignments are not necessarily the optimal ones[406].

All different tools implemented in BLAST can be access at BLAST NCBI web service but can be run remotely using BLAST+[407]. The input accepted are the sequences in Fasta or FastQ format. The resultant output contains an entry for each hit found. Those entries have a brief description of the best match, the scientific name of the specie, the score, the percentage of query sequence covered, the E value and percentage of identity between the sequences.

From experience, and if is not strictly necessary, I do not recommend running BLAST as exploratory method for NGS data because is not as computationally efficient as other available programmes. Despite this, BLAST is still among the most accurate alignment software and its database is

extensive. The software can be still used when the number of sequence reads is reduced or to test that already mapped data is mapped against a proper reference genome. Finally, is worth to mention that there are other tools designed to analyse and visualise the taxonomies outputted by BLAST, being MEGAN one of them[408].

2.2.8.2 Kraken

Another software used for taxonomic labelling is Kraken. Kraken is characterised for classifying reads using exact *k-mer* aligning against a database[409]. The database contains a series is composed of *k-mer* and the respective Lowest Common Ancestor (LCA) of the organisms which contain that same k-mer. Regarding its performance, the latest version of the software Kraken2 [410], can align 90 million of reads in 30 seconds, orders of magnitude faster than BLAST.

Due to its fast runtime, Kraken can be used as an exploratory method before mapping the reads. The databases are customisable, but in the project associated to this thesis, I have used the default Kraken database, with genomes of the taxonomic groups Archaea, Bacteria, Fungi, Protozoa, Virus and Human. The output of Kraken consist in a file with each read's taxonomic assignation and a file with the recount of reads for each taxonomic group. The taxonomic composition of sample can be visualised using Krona[411].

3 Objectives

The objectives of this thesis and its associated studies is to characterise the presence of ancient pathogen DNA through 3 different perspectives, the analysis of a life-threatening disease, the study of an historically documented clinical case, and the outbreak of an epidemic.

The specifics objectives are:

1) A global disease (malaria)

- The retrieval of *Plasmodium falciparum* genome-wide data from antique medical slides.
- The characterisation of an eradicated European *P. falciparum* strain and its phylogenetic relationship with present day strains.
- The analysis of the temporal emergence of drug resistance associated mutations.

2) A historical, individual case of disease

- The first recovery of human genome-wide data from a 18th century blood stain found in a manuscript.
- To demonstrate the authenticity of the retrieved human DNA and explore his ancestry background.

- The characterisation of the microbial community found in the blood and its association with a documented clinical case.

3) An ancient pandemic (paratyphoid epidemics)

- The recovery of human and pathogen DNA from human remains.
- The characterisation of the ancestry composition of 17th century soldiers.
- The characterisation of a European strain of ancient *Salmonella enterica*.
- To date the split between this historical strain from other modern and ancient *S. enterica* strains.

4 Results

4.1 Genetic affinities of an eradicated European *Plasmodium falciparum* strain.

Toni de-Dios, Lucy van Dorp, Pere Gelabert, Christian Carøe, Marcela Sandoval-Velasco, Rosa Fregel, Raül Escosa, Carles Aranda, Silvie Huijben, François Balloux, M Thomas P Gilbert, Carles Lalueza-Fox

Microb Genom. 2019 Sep;5(9):e000289.doi: 10.1099/mgen.0.000289. Epub 2019 Aug 20.

MICROBIAL GENOMICS

SHORT COMMUNICATION

de-Dios *et al.*, *Microbial Genomics* 2019;5

DOI 10.1099/mgen.0.000289

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Genetic affinities of an eradicated European *Plasmodium falciparum* strain

Toni de-Dios[1]†, Lucy van Dorp[2]†, Pere Gelabert[1], Christian Carøe[3], Marcela Sandoval-Velasco[3], Rosa Fregel[4,5], Raül Escosa[6], Carles Aranda[7], Silvie Huijben[8], François Balloux[2], M. Thomas P. Gilbert[3,9] and Carles Lalueza-Fox[1,*]

Abstract

Malaria was present in most of Europe until the second half of the 20th century, when it was eradicated through a combination of increased surveillance and mosquito control strategies, together with cross-border and political collaboration. Despite the severe burden of malaria on human populations, it remains contentious how the disease arrived and spread in Europe. Here, we report a partial *Plasmodium falciparum* nuclear genome derived from a set of antique medical slides stained with the blood of malaria-infected patients from Spain's Ebro Delta, dating to the 1940s. Our analyses of the genome of this now eradicated European *P. falciparum* strain confirms stronger phylogeographical affinity to present-day strains in circulation in central south Asia, rather than to those in Africa. This points to a longitudinal, rather than a latitudinal, spread of malaria into Europe. In addition, this genome displays two derived alleles in the *pfmrp1* gene that have been associated with drug resistance. Whilst this could represent standing variation in the ancestral *P. falciparum* population, these mutations may also have arisen due to the selective pressure of quinine treatment, which was an anti-malarial drug already in use by the time the sample we sequenced was mounted on a slide.

DATA SUMMARY

Plasmodium falciparum reads from Ebro-1944 have been deposited in the European Nucleotide Archive under accession number ERP114811. All modern *P. falciparum* samples used for the population genomics analyses were reported by the MalariaGEN *Plasmodium falciparum* Community Project in 2016 [1] and are provided in Table S1 (available in the online version of this article). All software used in the bioinformatic analyses are publicly available. Positions screened for anti-malarial drug resistance are available in Table S2.

INTRODUCTION

Classical Greek accounts in the fourth and fifth centuries BCE describe people with intermittent fevers and infectious symptoms characteristic of malaria [2]. The Roman author Celsius was able to accurately differentiate the clinical symptoms of *Plasmodium vivax* versus *Plasmodium malariae* infections [3]. In contrast, it is unclear whether *Plasmodium falciparum*, the deadliest form of the pathogen, was already present in classical times. While some authors argue that *P. falciparum* only spread to southern Europe with the dawn of the Roman Empire, historical accounts suggest it may have affected western and central Italy as early as 400–100 BCE [4], before reaching the Po Delta region of northern Italy around 1000 years later [3]. The spread of this pathogen through Italy in historical times is supported by the discovery of a large infant cemetery at Lugnano (Teverina, Umbria, Italy), dating to approximately 450 CE, and the report of *P. falciparum* ribosomal DNA sequences obtained from one of these skeletons

Received 15 April 2019; Accepted 31 July 2019; Published 27 August 2019

Author affiliations: [1]Institute of Evolutionary Biology (CSIC-UPF), 08003 Barcelona, Spain; [2]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK; [3]Section for Evolutionary Genomics, Department of Biology, University of Copenhagen, 1353 Copenhagen, Denmark; [4]Department of Genetics, Stanford University, Stanford, CA, USA; [5]Department of Biochemistry, Microbiology, Cell Biology and Genetics, Universidad of La Laguna, 38206 La Laguna, Spain; [6]Consorci de Polítiques Ambientals de les Terres de l'Ebre (COPATE), 43580 Deltebre, Spain; [7]Servei de Control de Mosquits, Consell Comarcal del Baix Llobregat, 08980 Sant Feliu de Llobregat, Spain; [8]Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA; [9]Norwegian University of Science and Technology (NTNU) University Museum, N-7491 Trondheim, Norway.

***Correspondence:** Carles Lalueza-Fox, carles.lalueza@upf.edu

Keywords: malaria; *Plasmodium falciparum*; ancient genomics; drug resistance.

Abbreviations: aDNAancient DNA; mtDNA, mitochondrial DNA; PCA, principal component analysis; SNP, single nucleotide position.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary material is available with the online version of this article.

by means of traditional PCR [5]. The recent retrieval of larger amounts of *P. falciparum* genetic data, by means of capture baits and second-generation sequencing technologies, from teeth sampled in Velia and Vagnari cemeteries (Italy), directly places this parasite in southern Italy by the beginning of the Roman Imperial period [6].

Genetic analyses of DNA recovered from a unique set of antique microscopy slides (1942–1944), stained with the blood of malaria-infected patients in the Ebro Delta of Spain for immediate diagnostic purposes, allowed us to report the complete mitochondrial DNA (mtDNA) genomes of historical *P. falciparum* and *P. vivax* [7]. In this study, we analyse data generated through merging the sequence information derived from four different slides, allowing us to reconstruct the partial nuclear genome of this eradicated European *P. falciparum* strain.

Impact Statement

Malaria is a serious infectious disease affecting over 200 million people annually. The disease is caused by species of parasitic protozoans from the genus *Plasmodium*, which are transmitted by several species of mosquitoes from the genus *Anopheles*. Today, *Plasmodium* is restricted to tropical and subtropical latitudes. However, malaria was historically present in most of Europe, spanning from southern Britain and the Mediterranean to as far north as Finland. Spain represented one of its last footholds, where it persisted until the 1960s. Here, we report a substantial fraction of the genome of a 20th century European *Plasmodium falciparum* strain isolated from slides stained with the blood of malaria-infected patients in the 1940s. We analyse this genome in the context of worldwide modern strains to trace the historical dispersal of *P. falciparum* in Europe. We find evidence supporting a longitudinal spread from Asia into Europe, over a latitudinal spread from Africa, as well as variants in anti-malarial resistance genes predating the use of most common anti-malarial drugs. Our work highlights the potential of collections of antique medical slides to open new possibilities in the study of ancient microbial genomics, including malaria.

METHODS

Sample collection

Four microscope slides, dated between 1942 and 1944, were selected for this study. All were obtained from Dr Ildefonso Canicio's family collection. Dr Canicio was in charge of the anti-malarial hospital at Sant Jaume d'Enveja (Ebro Delta, Spain), inaugurated in 1925, until his death in 1961. Patients were predominantly local people who worked in the Ebro rice fields and had no history of international travel. The samples consist of four labelled microscopy slides stained with Giemsa – probably not previously fixed – in which parasites were still visible under the microscope. DNA extraction was performed in dedicated ancient DNA (aDNA) laboratories at the Institute of Evolutionary Biology in Barcelona (Spain) and the Centre for GeoGenetics in Copenhagen (Denmark) in 2015 and 2017, respectively, as described in the Supplementary Material.

Ancient sample mapping and assembly

We first analysed the sequenced reads obtained from our eradicated European *P. falciparum* samples using *FastQC* (v0.11.7) [8] in order to determine their quality before and after trimming of the adapter sequences. We removed sequencing adapters and reads shorter than 30 bp using *cutadapt* (v1.3) [9]. We then mapped our reads against the *P. falciparum* 3D7 and *P. vivax* Sal1 reference genomes using BWA (v0.7.17) [10] aln, specifying no seeding, a gap open penalty of 2, an edit distance of 0.01 and no trimming; parameters shown to optimize mapping of ancient microbial samples [11]. We removed duplicated reads using *Picard* (v2.18.6) 'MarkDuplicates' and retained all mapped reads with a map quality of at least 30 in *SAMtools* (v1.6) [12]. As the blood samples had known co-infection with *P. vivax*, we extracted the sequencing reads mapping more confidently to *P. falciparum* by selecting those reads that had a lesser edit distance between *P. falciparum* 3D7 than with *P. vivax* Sal1 (Fig. S1). The obtained G+C content of the reads matched the expected and extremely low G+C content characteristic of the *P. falciparum* genome (Fig. S2).

Post-mortem aDNA damage patterns were determined using *MapDamage* (v2.0.8) [13] (Fig. S3). Most of the *P. falciparum* reads (0.6294× mean coverage, representing 88% of the total genome) come from a single slide (labelled CA) sequenced in 2015. The remaining reads come from three other slides with mean genome-wide coverage of 0.0213×, 0.0285× and 0.0314×. The reduced coverage obtained from these slides relates to the overall endogenous sequence content and quality, which may also vary depending on the stage of the parasite's life cycle when the slides were prepared. To increase the overall coverage, we therefore merged data from all four slides, always calling the dominant allele. We call our resultant composite genome Ebro-1944.

We also generated a reference panel of modern *P. falciparum* samples from publicly available sequence data [1]. In all cases, we mapped reads using *BWA* (v0.7.17) [10] *mem*, before removing duplicated sequence reads and filtering by mapping quality using thresholds as described above. The same procedure was followed for mapping the genome of *Plasmodium praefalciparum*, which was used as an outgroup. *Qualimap* (v2.2) [14] was used to generate the mapping metrics for all samples.

Variant calling and dataset creation

To compare Ebro-1944 to current strains in global circulation, we selected modern strains with a mean depth of coverage equal or above 50× and with at least 90% of the reference genome covered. This filtering strategy resulted in a dataset

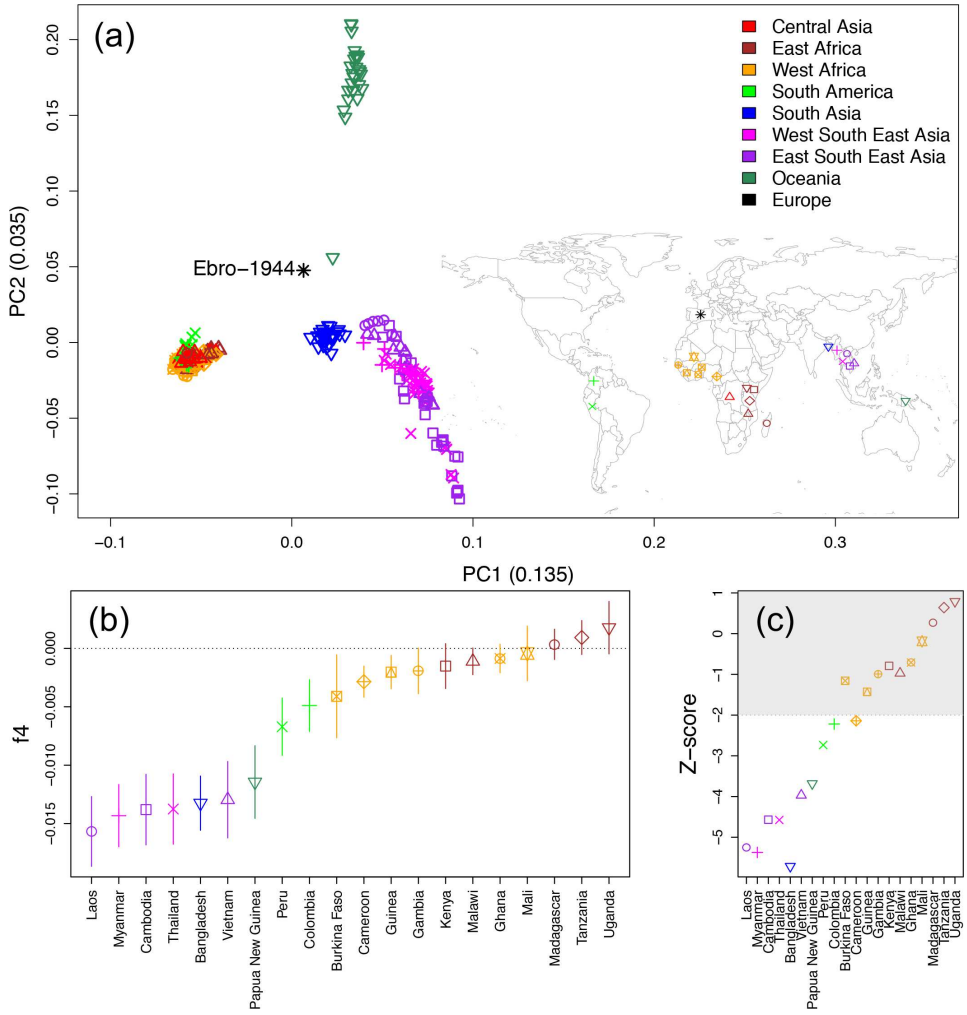


Fig. 1. (a) PCA with worldwide sample locations (inset map). (b) *f*4-statistics under the test relationship *f*4(*P. praefalciparum*, Ebro-1944, X, Democratic Republic of Congo), where X iterates through the geographical sampling locations of our included modern *P. falciparum* strains. A more negative *f*4 value indicates a closer relationship of Ebro-1944 to X relative to strains sampled from the Democratic Republic of Congo. (c) Z-scores under the test relationship *f*4(*P. praefalciparum*, Ebro-1944, X, Colombia) assessed through block jackknife resampling. All possible *f*4 topologies were tested (Fig. S4), with Ebro-1944 showing a consistent closer affinity to strains sampled from central south Asia.

comprising 206 global samples. We used *GATK* (v3.7) [15], algorithm *UnifiedGenotyper*, to call variants, specifying *EMIT_ALL_CONFIDENT_SITES* and a standard confidence threshold over 50. We subsequently used *VCFtools* (0.1.14) [16] to filter out positions with less than 40× coverage, heterozygous positions, multi-allelic positions, indels and recombinant sub-telomeric regions (see the Supplementary Material). These samples were later merged using *bcftools* (v1.3.1) [12] and all variants that were not present in at least three samples were removed leaving 681 486 single nucleotide positions (SNPs). We used these positions to call additional modern samples, which mapped with lower coverage (<50×), applying the same filters. The resultant calls were merged using *bcftools*. SNPs were called for *P. praefalciparum* using the same parameters.

For Ebro-1944, we used a different approach to overcome the post-mortem damage associated with aDNA samples. We generated pseudo-haploid calls at the positions identified in the population genetics dataset using *SAMtools* to

call a random base drawn from all possible reads at each site [17]. We then merged all the samples (including Ebro-1944) using *bcftools*. The complete dataset was filtered for a minimal minor allele frequency of 0.01 and multi-allelic SNPs were removed using *VCFtools*. All SNPs that were not present in at least 75% of the samples considered were also removed. This procedure resulted in a final dataset comprising 435 samples and 14346 bi-allelic SNPs, with 50.85% of these covered in Ebro-1944.

Population genetics analyses

This dataset was compiled by filtering all modern samples for sites called in Ebro-1944 and retaining all samples that had at least 50% of these sites covered. This resulted in a dataset of 306 samples and 6 755 SNPs. A principal component analysis (PCA) was performed on this dataset using SmartPCA within the *EIG* (v.6.0.1) suite of tools [18].

To formally test the relative affinity of Ebro-1944 to South Asian, Oceanian and African strains, we used the full 14346 biallelic SNP dataset to generate *f*4 statistics of the form *f*4(*P. praefalciparum*, Ebro-1944; X, Y), where we use *P. praefalciparum* as an outgroup and iterate through all combinations of geographical samples (X and Y) included in our global dataset. This statistic is designed to quantify the covariance in allele frequency differences between *P. praefalciparum* and Ebro-1944, and X and Y. If *P. praefalciparum* and Ebro-1944 form a clade with respect to X and Y, then their allele frequency differences should be uncorrelated and *f*4 will have a value of 0. Deviations from 0, thus, provide the relative affinity of Ebro-1944 to X and Y; positive values indicating a closer relationship to Y relative to X and negative values indicating a closer relationship to X relative to Y. *f*4 statistics were generated in qpDstat of *AdmixTools* (v.5.0) [19] and statistical significance was assessed through Z-scores following block jack-knife resampling (Fig. S4).

We additionally sub-sampled the global population genetics dataset more stringently, retaining only the strains with no missingness across the entire alignment, which led to a dataset of 30 strains over 8195 sites (Table S1). We generated patterns of pairwise haplotype sharing across this dataset using Chromopainter v2 [20], specifying a uniform recombination rate of 9.6 kb cM^{-1} [21] (Fig. S5) and using default estimates of mutation and switch parameters.

We also considered the mitochondrial relationships by placing our Ebro-1944 mtDNA genome (16.1×) in a minimum spanning network together with 435 global strains (Fig. S6). The mtDNA genome was also screened for specific variants restricted to South Asian strains.

Resistance variants screening

A set of variants previously reported in the literature as associated with resistance to different anti-malarial drugs (Table S2) was screened in Ebro-1944 [22–46]. The position of the derived allele was determined using *Jvarkit* (v.2018.04.05) [47].

RESULTS

Our sequencing of the four antique microscopy slide samples yielded 218 952 *P. falciparum* reads producing a nuclear genome of 0.67× mean depth and covering 40.99% of the *P. falciparum* reference genome 3D7 (Fig. S1). In addition, 1398 reads mapped to the mtDNA genome, generating a mean coverage of 16.1× over 99.43% of the reference genome. Sequence reads showed damage patterns typical of aDNA (C to T and G to A substitutions at the 5′ and 3′ ends) in ratios of 3.8 and 2.2%, respectively (Fig. S3). The characteristic aDNA damage of miscoding lesions indicates that our *P. falciparum* reads are authentically old, deriving from our historic specimen rather than from modern contamination where we would expect no such damage profile.

PCA (Fig. 1a), *f*4 allele sharing statistics (Figs 1b, c, S4) and haplotype sharing analyses (Fig. S5) indicate that the closest genetic affinity of our European strain is to contemporary samples from central south Asia, including those currently in circulation in Laos, Myanmar and Vietnam. We also detect a shared genetic component with samples from Papua New Guinea, in particular PN0008-C (Fig. 1a). Whilst Ebro-1944 is more closely related to current strains from central south Asia than those from Africa, we found a higher proportion of African haplotypes in Ebro-1944 than in any central south Asian samples (two sample *t*-test $P<2.2\times10^{-16}$). Such a pattern would be consistent with a common origin of the European and central south Asian *P. falciparum* populations, followed by secondary introgression of African strains into the European population after the split between the European and central south Asian lineages (Fig. S5).

The close genetic affinity of the nuclear genome to strains in circulation in central south Asia is further confirmed at the mtDNA level (Fig. S6). Ebro-1944 carries two Indian-specific mtDNA substitutions at positions 276 and 2763 [48]. One additional mutation (position 725) is also present in our sample and is shared with nine contemporary Indian samples and two African samples included in our modern reference dataset (Table S1). We observe mutations at these three positions across all four of our sampled slides, suggesting the strains combined in Ebro-1944 are phylogenetically similar (Table S3). Unfortunately, these three mutations are not covered in a previously published partial strain retrieved from Roman cemetery sites [6]. Additionally, none of our reads overlapped with the ribosomal DNA fragment of European *P. falciparum* retrieved in a previous work [5].

We screened the Ebro-1944 nuclear genome for the presence of 117 variants that have been previously associated with anti-malarial drug resistance (Table S2). We achieved sequence coverage of at least one read at 62 positions. Of these, only two mutations (H191Y and I876V) displayed the resistance-associated allele (Table S2). Both alleles are located in the multidrug resistance protein 1 encoding gene (*pfmrp1)*, which encodes an ABC family transporter, and has been associated to alterations in quinine, chloroquine, artemisinin, piperaquine and primaquine sensitivity [49]. Although one of these mutations, H191Y, is a C to T substitution that is

characteristic of post-mortem DNA deamination, and might, thus, be expected in ancient samples, we believe this to be authentic as the transition is present in two overlapping reads, and is not located at the end of the reads where damage tends to accumulate [50]. While the second variant, I876V, is only covered by a single read, it is an A to G substitution and is, thus, not a common DNA damage motif. To provide additional support for the presence of these derived alleles, we imputed these positions in our partial genome by using a reference panel of modern strains carrying the H191Y and I876V variants (see the Supplementary Material). Interestingly, while the latter variant is distributed worldwide, the former is currently restricted to Asia and Oceania, consistent with an Asian dispersal of *P. falciparum* into Europe.

DISCUSSION

We showed that an eradicated European *P. falciparum* strain from the 20th century is most closely related to extant strains from central south Asia, such as those currently in circulation in Laos, Myanmar and Vietnam. Although we detect some evidence for secondary introgression of African *P. falciparum* into the extinct European population, the significantly stronger affinity to Asian strains argues against a direct African origin [4] of European *P. falciparum*, and points instead to a migration event between Europe and Asia.

We cannot infer the directionality of the migration between Europe and Asia from the genetic evidence alone. A recent expansion of *P. falciparum* from Europe to Asia and Oceania, coinciding with European colonial expansion, might be conceivable given the well-characterized role of colonialism in the widespread dissemination of other major infectious diseases, such as *Mycobacterium tuberculosis* lineage 4, the globally distributed agent of tuberculosis [51].

However, a migration of *P. falciparum* from Europe to Asia does not sit well with historical evidence of the arrival of the parasite in Europe during antiquity [2–6]. Given that our extinct European strain shares significantly more alleles with extant Asian rather than African strains, the arrival of *P. falciparum* malaria parasites into Europe likely took place from the Asian sub-continent rather than spreading from Africa via the Mediterranean during the Roman Empire.

Plausible historical migrations responsible for a spread into Europe from the East include the extensive commercial exchanges and movement between people of various ethnicities in the Achaemenid Empire (550-330 BCE). Alternatively, *P. falciparum* might have reached Europe during the subsequent Hellenistic period, connecting India with the Mediterranean, following the conquest of the Achaemenid Empire by Alexander the Great.

The availability of a strain pre-dating most of the currently used anti-malarial drugs allowed us to look for the presence of resistance variants that may inform on the spread of such resistances in the future. The two resistance variants we observed in the *pfmrp1* gene could be explained by standing variation for drug-resistance mutations in *P. falciparum* or may have arisen following the use of quinine for over three centuries; chloroquines were not introduced in Spain until 1948 and initially only in the African colonies.

Our results provide novel insights into the evolution and past demography of one of the world's deadliest pathogens, which could not have been reached by studying the genomes of extant strains alone. Additional genomic evidence from both medical collections and ancient remains will be needed to reconstruct more precise timings and routes for the spread of malaria into Europe, and could also help in determining the emergence and drivers of resistance to anti-malarial drugs.

Funding information
C. L. -F. is supported by Obra Social 'La Caixa', Secretaria d'Universitats i Recerca Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880) and by FEDER-MINECO (PGC2018-095931-B-100). L. v. D. and F. B. acknowledge financial support from the MRC Newton Fund UK–China NSFC initiative (grants MR/P007597/1 and 81661138006).

Acknowledgements
We are grateful to the descendants of Dr Canicio, Miquel and Ildefons Oliveras, for sharing their slides with us.

Author contributions
F. B., M. T. P. G. and C. L.-F. conceived the research; C. A. and R. E. collected the samples and associated information; C. C., M. S.-V. and R. F. performed experimental procedures; T. d.-D., L. v. D. and P. G. performed computational analysis; S. H. provided information on resistance variants; T. d.-D., L. v. D., F. B., M. T. P. G. and C. L.-F wrote the paper.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement
Experimental protocols have been approved by the Clinical Research Ethics Committee of Institut Hospital del Mar d'Investigacions Mèdiques (CEIC-PSMAR).

Data bibliography
1. The *Plasmodium falciparum* reference genome used was *P. falciparum* 3D7, assembly ASM276v2.

2. The *Plasmodium vivax* reference genome used was *P. vivax* Sal-1, assembly ASM241v2.

3. Details of the *Plasmodium falciparum* samples used in the population genetics analyses can be found under the study accession number PRJEB2136. Accession numbers for each sample are described in Table S1.

4. The sequence of *Plasmodium praefalciparum* can be found under accession number SAMEA2464702.

References
1. MalariaGEN Plasmodium falciparum Community Project. 2016. Genomic epidemiology of artemisinin resistant malaria. https://elifesciences.org/articles/08714
2. Jones WHS. *Malaria, a Neglected Factor in the History of Greece and Rome*. Cambridge: Cambridge University Press; 1907. pp. 97–102.
3. Sallares R, Bouwman A, Anderung C. The spread of malaria to southern Europe in antiquity: new approaches to old problems. *Med Hist* 2004;48:311–328.
4. De Zulueta J. Malaria and Mediterranean history. *Parassitologia* 1973;15:1–15.
5. Sallares R, Gomzi S. Biomolecular archaeology of malaria. *Ancient Biomolecules* 2001;3:195–213.

6. Marciniak S, Prowse TL, Herring DA, Klunk J, Kuch M *et al*. *Plasmodium falciparum* malaria in 1st–2nd century CE southern Italy. *Current Biology* 2016;26:R1220–R1222.
7. Gelabert P, Sandoval-Velasco M, Olalde I, Fregel R, Rieux A *et al*. Mitochondrial DNA from the eradicated European *Plasmodium vivax* and *P. falciparum* from 70-year-old slides from the Ebro Delta in Spain. *Proc Natl Acad Sci USA* 2016;113:11495–11500.
8. Andrews S. FastQC, a quality control tool for high throughput sequence data; 2010. http://www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/
9. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.. *EMBnet.Journal* 2011;17:10–12.
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
11. Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining metagenomic data sets for ancient DNA: recommended protocols for authentication. *Trends Genet* 2017;33:508–520.
12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
13. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L *et al*. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 2013;29:1682–1684.
14. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S *et al*. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012;28:2678–2679.
15. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G *et al*. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;43:11.10.1–11.10.33.
16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E *et al*. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–2158.
17. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U *et al*. A draft sequence of the Neandertal genome. *Science* 2010;328:710–722.
18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909.
19. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N *et al*. Ancient admixture in human history. *Genetics* 2012;192:1065–1093.
20. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet* 2012;8:e1002453.
21. Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S *et al*. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biol* 2011;12:R33.
22. Foote SJ, Kyle DE, Martin RK, Oduola AM, Forsyth K *et al*. Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in *Plasmodium falciparum*. *Nature* 1990;345:255–258.
23. Price RN, Cassar C, Brockman A, Duraisingh M, van Vugt M *et al*. The *pfmdr1* gene is associated with a multidrug-resistant phenotype in *Plasmodium falciparum* from the Western border of Thailand. *Antimicrob Agents Chemother* 1999;43:2943–2949.
24. Dahlström S, Ferreira PE, Veiga MI, Sedighi N, Wiklund L *et al*. *Plasmodium falciparum* multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *J Infect Dis* 2009;200:1456–1464.
25. Rottmann M, McNamara C, Yeung BKS, Lee MCS, Zou B *et al*. Spiroindolones, a potent compound class for the treatment of malaria. *Science* 2010;329:1175–1180.
26. Setthaudom C, Tan-ariya P, Sitthichot N, Khositnithikul R, Suwandittakul N *et al*. Role of *Plasmodium falciparum* chloroquine resistance transporter and multidrug resistance 1 genes on in vitro chloroquine resistance in isolates of *Plasmodium falciparum* from Thailand. *Am J Trop Med Hyg* 2011;85:606–611.
27. Veiga MI, Ferreira PE, Jörnhagen L, Malmberg M, Kone A *et al*. Novel polymorphisms in *Plasmodium falciparum* ABC transporter genes are associated with major ACT antimalarial drug resistance. *PLoS One* 2011;6:e20212.
28. Takala-Harrison S, Clark TG, Jacob CG, Cummings MP, Miotto O *et al*. Genetic loci associated with delayed clearance of *Plasmodium falciparum* following artemisinin treatment in Southeast Asia. *Proc Natl Acad Sci USA*2013;110:240–245.
29. Gupta B, Xu S, Wang Z, Sun L, Miao J *et al*. *Plasmodium falciparum* multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east Myanmar. *J Antimicrob Chemother* 2014;69:2110–2117.
30. Vaidya AB, Morrisey JM, Zhang Z, Das S, Daly TM *et al*. Pyrazoleamide compounds are potent antimalarials that target Na+ homeostasis in intraerythrocytic *Plasmodium falciparum*. *Nat Commun* 2014;5:5521.
31. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J *et al*. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat Genet* 2015;47:226–234.
32. Pelleau S, Moss EL, Dhingra SK, Volney B, Casteras J *et al*. Adaptive evolution of malaria parasites in French Guiana: reversal of chloroquine resistance by acquisition of a mutation in *pfcrt*. *Proc Natl Acad Sci USA* 2015;112:11672–11677.
33. Callaghan PS, Siriwardana A, Hassett MR, Roepe PD. *Plasmodium falciparum* chloroquine resistance transporter (PfCRT) isoforms PH1 and PH2 perturb vacuolar physiology. *Malar J* 2016;15:186.
34. Reed MB, Saliba KJ, Caruana SR, Kirk K, Cowman AF. Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum*. *Nature* 2000;403:906–909.
35. Mishra N, Bharti RS, Mallick P, Singh OP, Srivastava B *et al*. Emerging polymorphisms in falciparum Kelch 13 gene in Northeastern region of India. *Malar J* 2016;15:4–9.
36. Wang Z, Cabrera M, Yang J, Yuan L, Gupta B *et al*. Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in *Plasmodium falciparum* from China-Myanmar border. *Sci Rep* 2016;6:33891.
37. Ye R, Hu D, Zhang Y, Huang Y, Sun X *et al*. Distinctive origin of artemisinin-resistant *Plasmodium falciparum* on the China-Myanmar border. *Sci Rep* 2016;6:20100.
38. Kobasa T, Talundzic E, Sug-aram R, Boondat P, Goldman IF *et al*. Emergence and spread of kelch13 mutations associated with artemisinin resistance in *Plasmodium falciparum* parasites in 12 Thai provinces from 2007 to 2016. *Antimicrob Agents Chemother* 2018;62:e02141-17.
39. Ross LS, Dhingra SK, Mok S, Yeo T, Wicht KJ *et al*. Emerging Southeast Asian PfCRT mutations confer *Plasmodium falciparum* resistance to the first-line antimalarial piperaquine. *Nat Commun* 2018;9:3314.
40. Mu J, Ferdig MT, Feng X, Joy DA, Duan J *et al*. Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol Microbiol* 2003;49:977–989.
41. Pickard AL, Wongsrichanalai C, Purfield A, Kamwendo D, Emery K *et al*. Resistance to antimalarials in Southeast Asia and genetic polymorphisms in PfMDR1. *Antimicrob Agents Chemother* 2003;47:2418–2423.
42. Durrand V, Berry A, Sem R, Glaziou P, Beaudou J *et al*. Variations in the sequence and expression of the *Plasmodium falciparum* chloroquine resistance transporter (Pfcrt) and their relationship to chloroquine resistance in vitro. *Mol Biochem Parasitol* 2004;136:273–285.
43. Happi CT, Gbotosho GO, Folarin OA, Akinboye DO, Yusuf BO *et al*. Polymorphisms in *Plasmodium falciparum* dhfr and dhps genes and age related in vivo sulfadoxine-pyrimethamine resistance in malaria-infected patients from Nigeria. *Acta Trop* 2005;95:183–193.
44. Sidhu ABS, Valderramos SG, Fidock DA. Pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in *Plasmodium falciparum*. *Mol Microbiol* 2005;57:913–926.
45. Echeverry DF, Holmgren G, Murillo C, Higuita JC, Björkman A *et al*. Polymorphisms in the *pfcrt* and *pfmdr1* genes of *Plasmodium*

falciparum and in vitro susceptibility to amodiaquine and desethylamodiaquine. *Am J Trop Med Hyg* 2007;77:1034–1038.

46. Dahlström S, Veiga MI, Mårtensson A, Björkman A, Gil JP. Polymorphism in PfMRP1 (*Plasmodium falciparum* multidrug resistance protein 1) amino acid 1466 associated with resistance to sulfadoxine-pyrimethamine treatment. *Antimicrob Agents Chemother* 2009;53:2553–2556.
47. Pierre L. JVarkit: java-based utilities for bioinformatics 2015.
48. Tyagi S, Pande V, Das A. New insights into the evolutionary history of *Plasmodium falciparum* from mitochondrial genome sequence analyses of Indian isolates. *Mol Ecol* 2014;23:2975–2987.
49. Raj DK, Mu J, Jiang H, Kabat J, Singh S *et al*. Disruption of a *Plasmodium falciparum* multidrug resistance-associated protein (PfMRP) alters its fitness and transport of antimalarial drugs and glutathione. *J Biol Chem* 2009;284:7687–7696.
50. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J *et al*. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* 2007;104:14616–14621.
51. Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J *et al*. Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci Adv* 2018;4:eaat5869.

Supplementary Materials

DNA extraction

The slides were immersed in 10 –15 mL of lysis buffer (consisting of 0,5% SDS, 0,25 mg/mL proteinase K, 10 mM Tris and 0.5 M EDTA pH 8.0) and left in 50 mL Falcon tubes at 37ºC overnight. In order to protect the written labels (if present), one of the ends remained unsubmerged. Afterwards, the supernatant was concentrated using a silica column-based method, as described by a protocol used to recover short and highly degraded DNA fragments from very ancient samples (1).

Library preparation and amplification

Double stranded libraries were created using NEBNext DNA Sample Prep Master Mix Set 2 (E6070; New England Biolabs) following the manufacturer's instructions with Illumina adapters as described in Dabney *et al.* 2013 (1). We determined the optimal number of required cycles necessary to amplify the samples and thus obtain a suitable amount of DNA (100-500 ng) using quantitative (q)PCR.

Capture depletion

As the expected quantity of *Plasmodium* DNA present in the slides is minimal in comparison to the more abundant human DNA from the host's cells, we used the following procedures: we first tried to reduce the human DNA content through whole genome capture with human baits and shotgun-sequenced the waste product. Additionally, we carried out a capture enrichment approach using whole genome baits synthesized for *P. falciparum* genomic (g)DNA, as described in March *et al.* 2013 (2). Genomic DNA obtained from the *P. falciparum* African strain 3D7 in *in vitro* culture (MRA-102G, MR4; ATCC) was fragmented and built into different libraries with a T7 adapter incorporated. These *Plasmodium* T7 libraries were subsequently used to generate biotinylated RNA baits by *in vitro* transcription. The capture-depletion assay

using whole-genome human baits was done following Mybait Human Whole genome to manual version 3.01 (from www.microarray.com/pdf/Mybaits-manual-v3.pdf). After hybridization of the ancient (a)DNA libraries with the human baits for 24 hours, we let it bind to streptavidin magnetic beads for 30 minutes at 65 ºC. Finally, we collected the supernatant (fraction that did not bind to the beads) and cleaned it using QiaQuick PCR Purification Kit (Qiagen) following the manufacturer's instructions. The samples were eluted in 30 µl of Elution Buffer (EB, Qiagen) after 10-minutes of incubation at 37 ºC.

Amplification of capture-depleted products

After we had estimated the optimal number of cycles with qPCR, capture-depletion products were amplified for five cycles and *P. falciparum* DNA reads captured for 22 cycles using 2x KAPA HotStart ReadyMix and re-amplification primers IS5 and IS6 (3). The samples were then quantified on an Agilent 2100 Bioanalyzer (Agilent technologies) and pooled in equimolar amounts. The pool was sequenced in one lane of an Illumina HiSEQ 4500 run in 80 SR mode. A library blank and an extraction blank control were included and showed no evidence of contamination with exogenous *P. falciparum* DNA. This extraction and amplification process resulted in us using all available material and slides.

Reference dataset

In order to represent the global *P. falciparum* diversity, we selected 434 worldwide *P. falciparum* samples from Amato et al 2016 (4). The final population genetics dataset consisted of 58 samples from Central Africa, 62 from East Africa, 62 from West Africa, 27 from South America, 48 from South Asia, 70 from West South East Asia, 64 from East South East Asia and 43 from Oceania. The full list of countries, samples and identifiers represented in each group can be found in Amato et al 2016 (4). We also selected a *P. praefalciparum* genome (accession code ERS437570) as an outgroup species (5). We detail the dataset in Supplementary Table S1.

Highly recombinant genes

In order to reduce the noise produced by highly variable recombinant regions, we removed a set of sub-telomeric genes at positions described in the literature (6–20).

Mitochondrial analysis

We extracted only the mtDNA alignment of the population genetics dataset, which comprised 5967 bp and 251 SNPs across 426 samples, including the *P. praefalciparum* outgroup. A pairwise similarity matrix was constructed based on the raw SNP differences and used to generate a minimum spanning network using the spantree() function from the R package Vegan(21).

***P. falciparum* 18s rDNA sequence**

We compared the reads mapped against the 18s rDNA gene with 2 previously published sequences of the same gene (22). Using BLASTn, both published sequences showed 100% identity with the 18s rDNA of *P. falciparum* (23). Unfortunately, our Ebro-1944 sequence did not overlap with this specific genetic region, but showed a high degree of identity with *P. falciparum* sequences (>95%).

Imputation of H191Y and I876V genetic variants at the *pfmrp1* gene

Given the low coverage of the Ebro-1944 nuclear genome, we conducted additional analyses to validate the presence and absence of variants involved in anti-malarial drug resistance. We performed imputation over a 100kb window of the genome containing the *pfmrp1* gene, and used GATK *UnifiedGenotyper* to call variants in this region (24). We then selected all samples with more than 80% of the positions called at a depth of 20x. We filtered out all the positions with indels and with a minor allele count below 2; and retained only the biallelic SNPs. This resulted in a dataset of 183 samples and 478 SNPs with no missing genotypes. This dataset was used as a reference for the imputation of Ebro-1944 using Beagle v.4.1 (25). The imputed results confirm the

presence of the derived allele at the positions chr_1:465296 (H191Y) and chr_1:467351 (I876V).

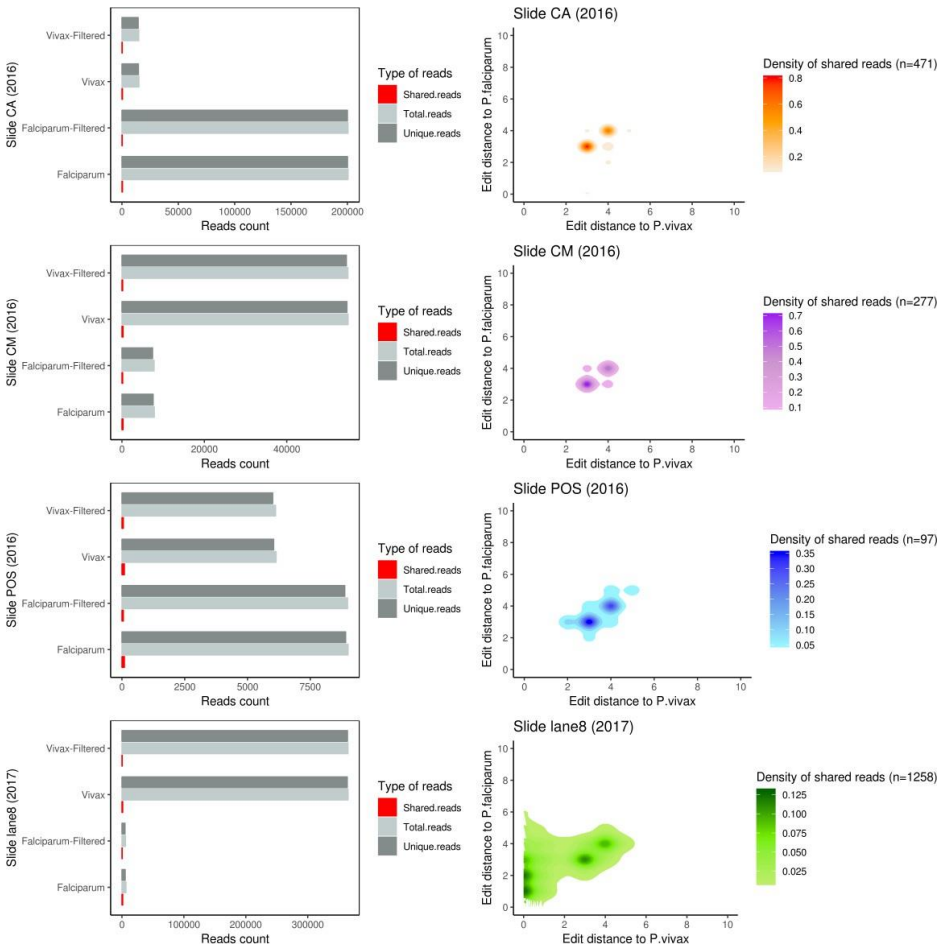
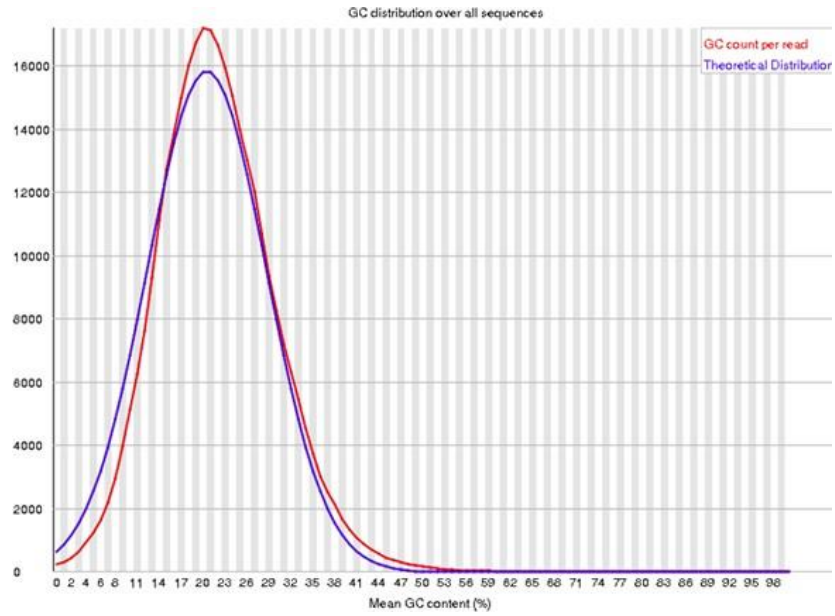
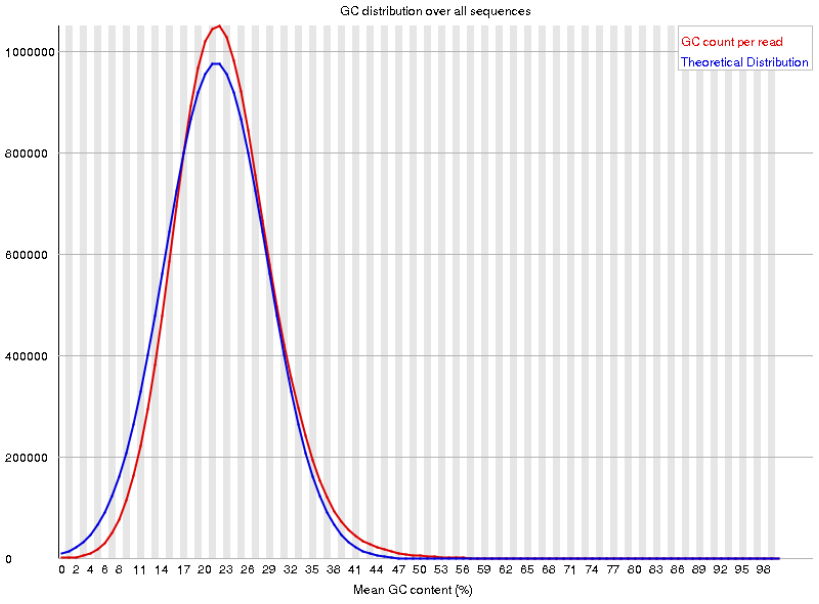


Figure S1: Comparison of read edit distance. Total amount (count) of reads mapped to *P. falciparum* and *P. vivax* in each slide (left), and comparison of the edit distance between shared reads mapped against the *P. falciparum* and *P. vivax* reference genomes (right). The majority of reads were obtained from Slide CA (top row).

A


GC distribution over all sequences
GC count per read
Theoretical Distributior
16000
14000
12000
10000
8000
6000
4000
2000
0
0 2 4 6 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83 86 89 92 95 98
Mean GC content (%)

B


GC distribution over all sequences
GC count per read
Theoretical Distribution
1000000
800000
600000
400000
200000
0
0 2 4 6 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83 86 89 92 95 98
Mean GC content (%)

C

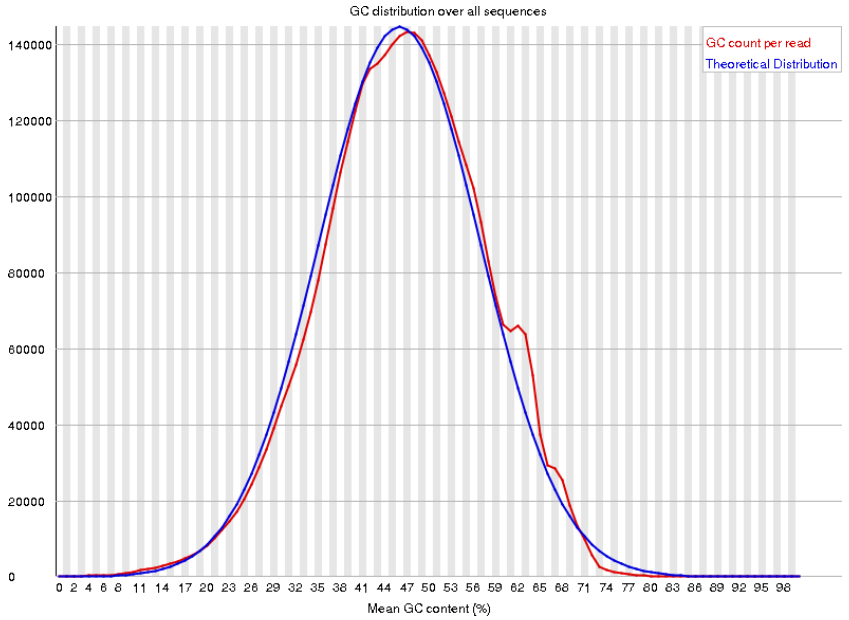


Figure S2: GC content. GC content of Ebro-1944 mapped against *3D7* (A). The y-axis provides the number of reads with an observed GC content (x-axis) in Ebro-1944 (red) and their theoretical distribution (blue). Figures S2 B and C provide the same profile for an example of modern *P. falciparum* (PR0124) and an example of modern *P. vivax* (PNG 030 sample), respectively (4,26).

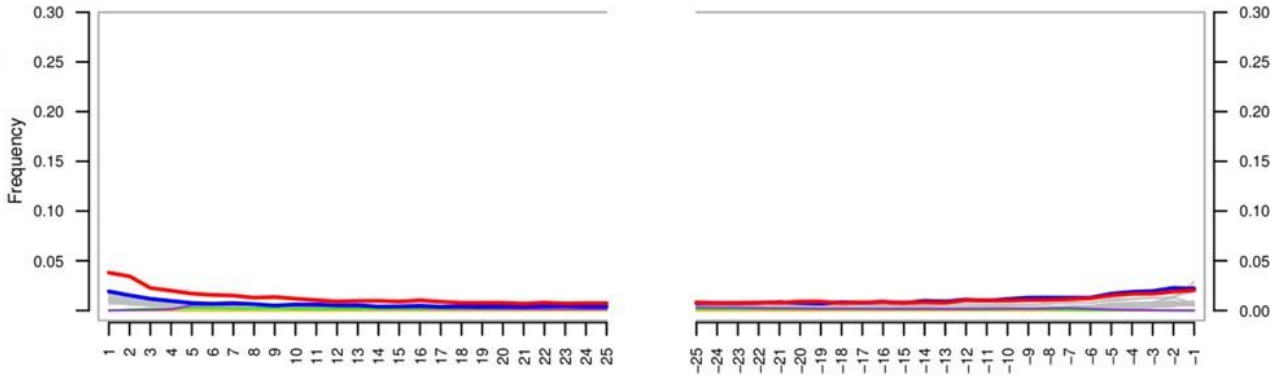
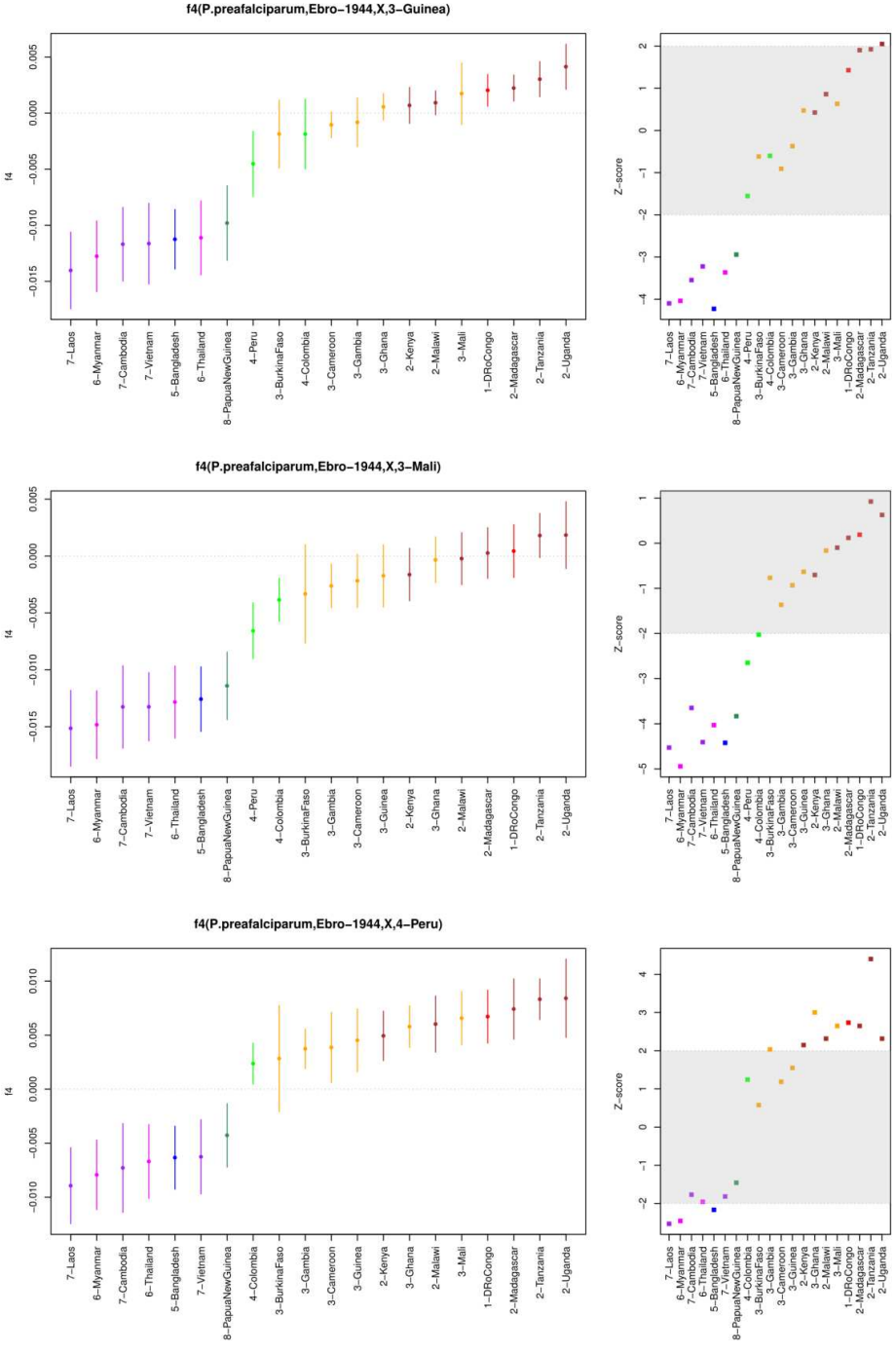
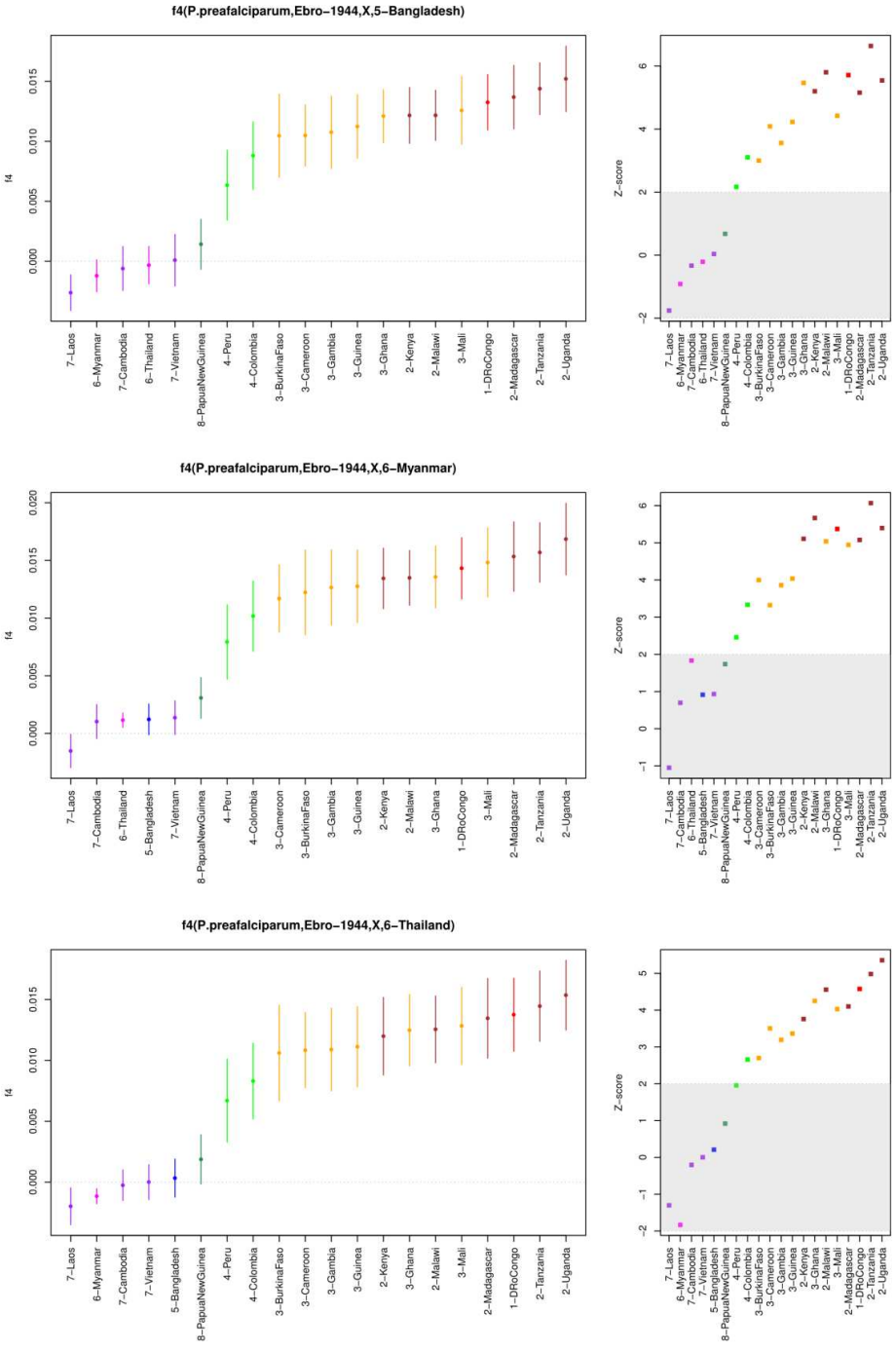
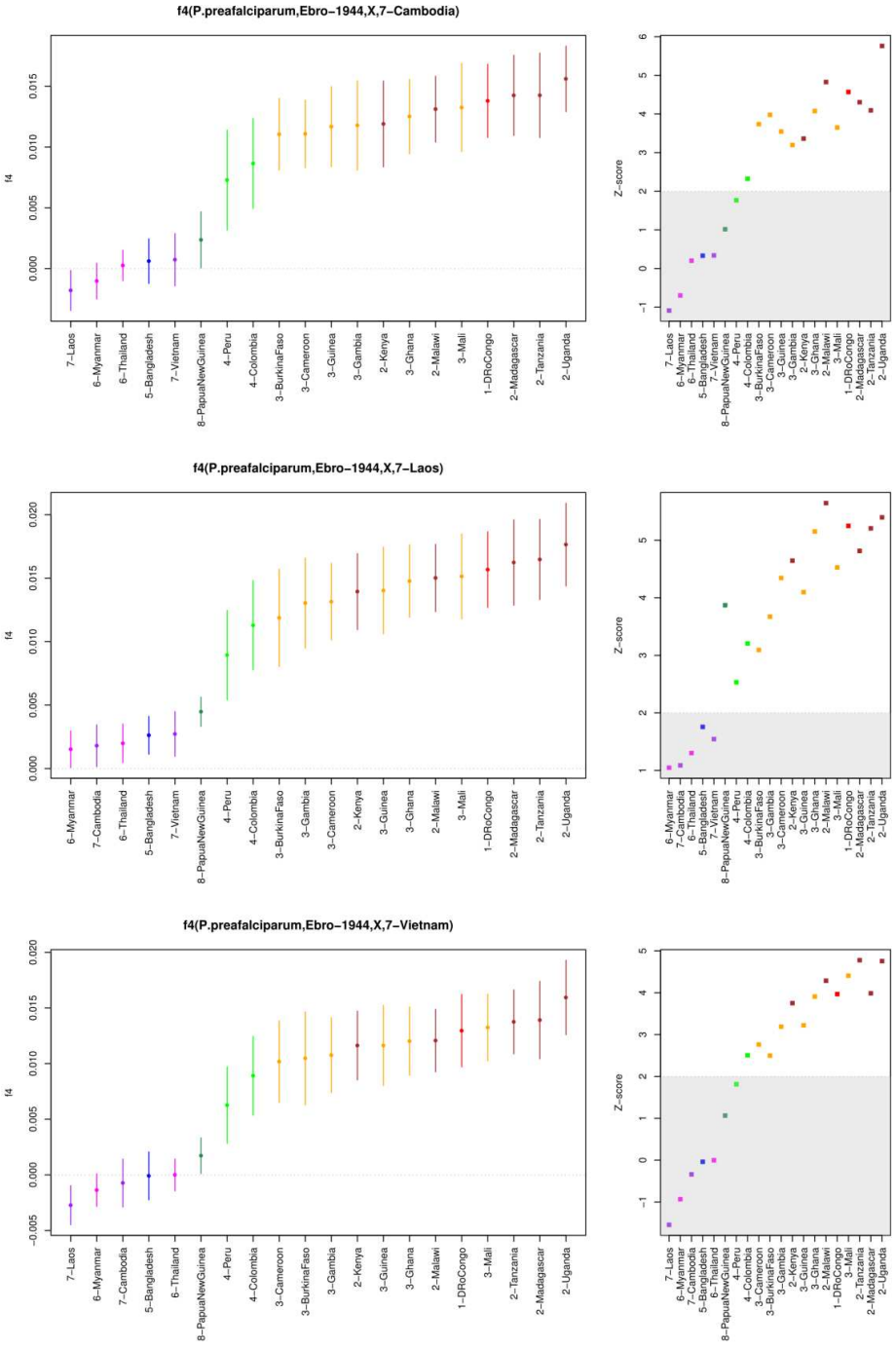
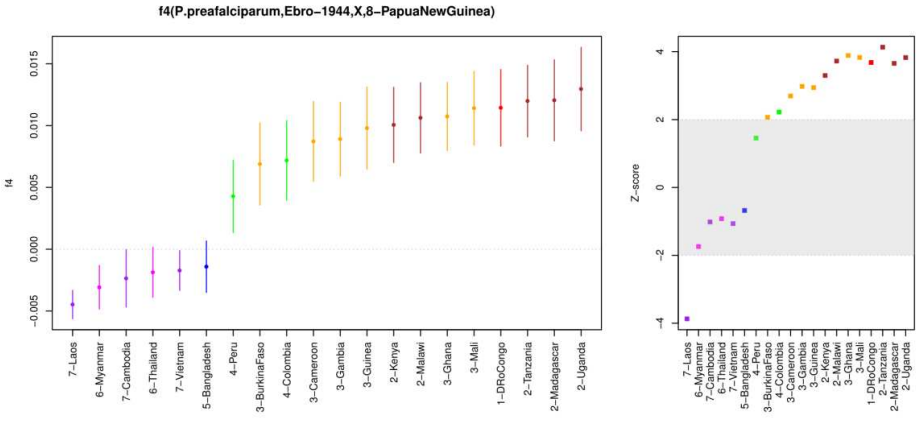


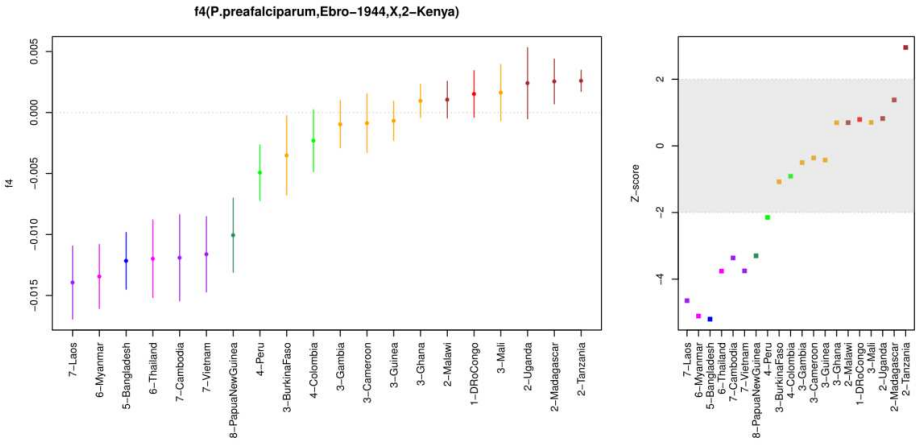
Figure S3: Post-mortem damage profiles of Ebro-1944 reads. The specific nucleotide positions (x-axis) at which a substitution is present at the 5' end (left) and 3' end (right) of the mapped reads is provided. In red, the C to T substitution frequency; in blue, the G to A substitution frequency; in grey, the frequency of all other substitutions. The elevation in C to T substitutions at the 5' end and G to A substitutions at the 3' end suggest DNA damage in Ebro-1944 is consistent with the degradation expected in post-mortem historical samples rather than modern contamination.

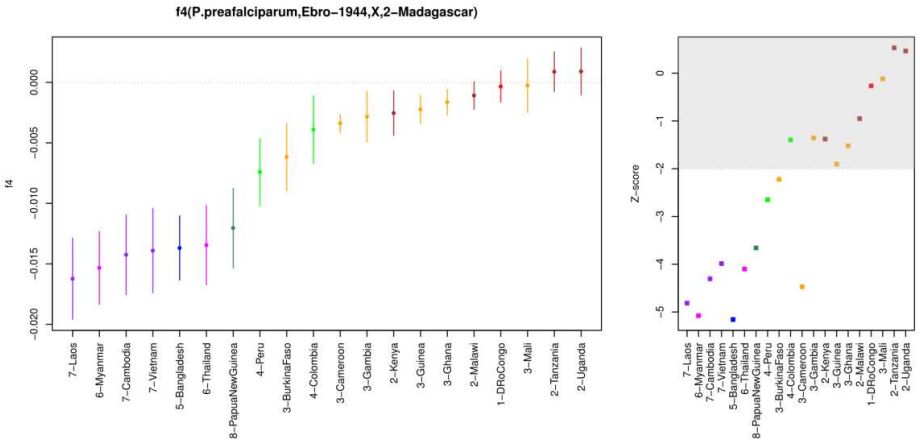

f4(P.preafalciparum,Ebro−1944,X,3−Guinea)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
7−Vietnam
5−Bangladesh
6−Thailand
8−PapuaNewGuinea
4−Peru
3−BurkinaFaso
4−Colombia
3−Cameroon
3−Gambia
3−Ghana
2−Kenya
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,3−Mali)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
7−Vietnam
6−Thailand
5−Bangladesh
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Gambia
3−Cameroon
3−Guinea
2−Kenya
3−Ghana
2−Malawi
2−Madagascar
1−DRoCongo
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,4−Peru)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Colombia
3−BurkinaFaso
3−Gambia
3−Cameroon
3−Guinea
2−Kenya
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda

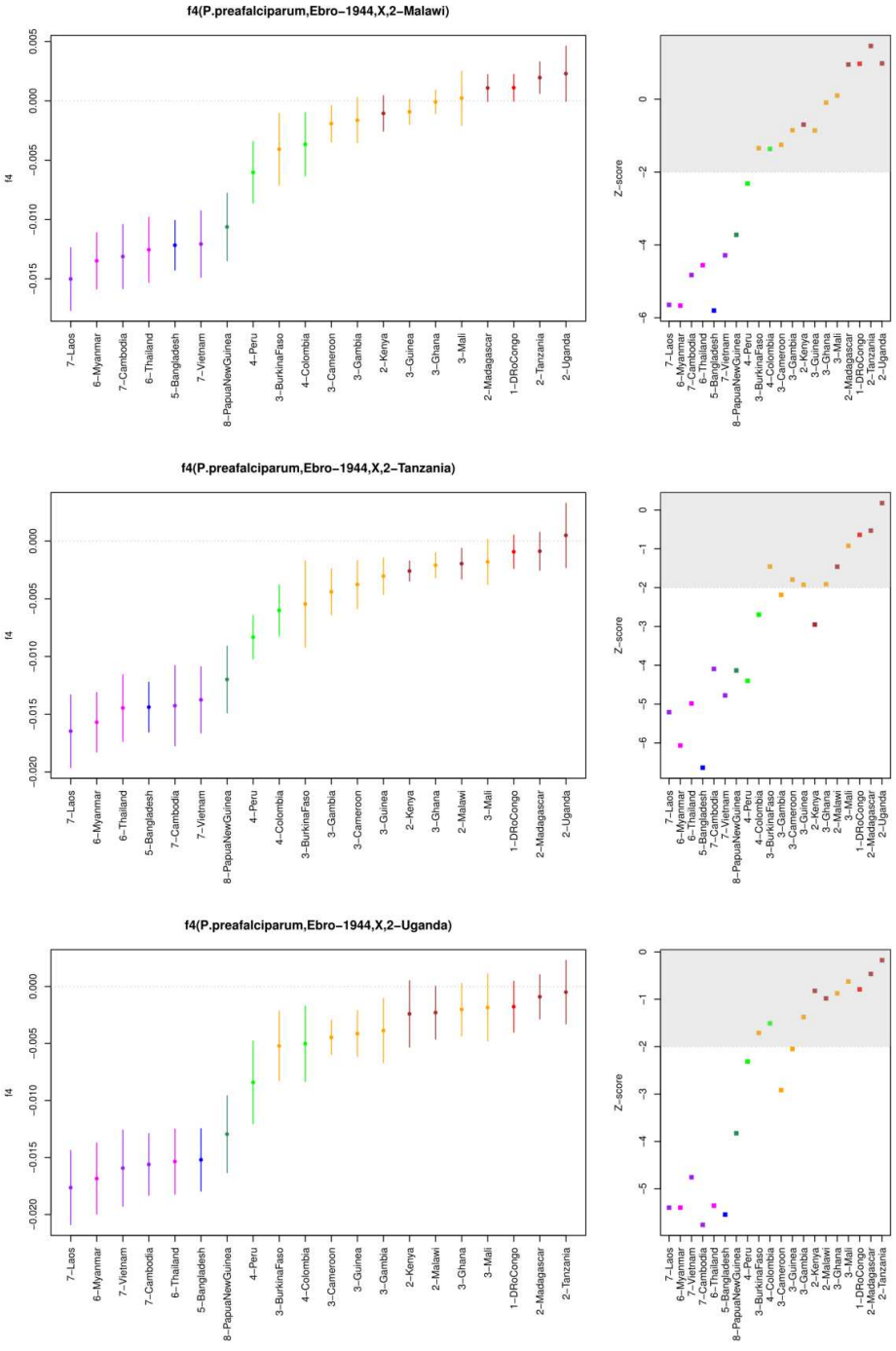

f4(P.preafalciparum,Ebro−1944,X,5−Bangladesh)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
7−Vietnam
5−Bangladesh
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Cameroon
3−Gambia
3−Guinea
3−Ghana
2−Kenya
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,6−Myanmar)
f4
Z−score
7−Laos
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−Cameroon
3−BurkinaFaso
3−Gambia
3−Guinea
2−Kenya
2−Malawi
3−Ghana
1−DRoCongo
3−Mali
2−Madagascar
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,6−Thailand)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
7−Vietnam
5−Bangladesh
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Cameroon
3−Gambia
3−Guinea
2−Kenya
3−Ghana
2−Malawi
3−Mali
2−Madagascar
1−DRoCongo
2−Tanzania
2−Uganda

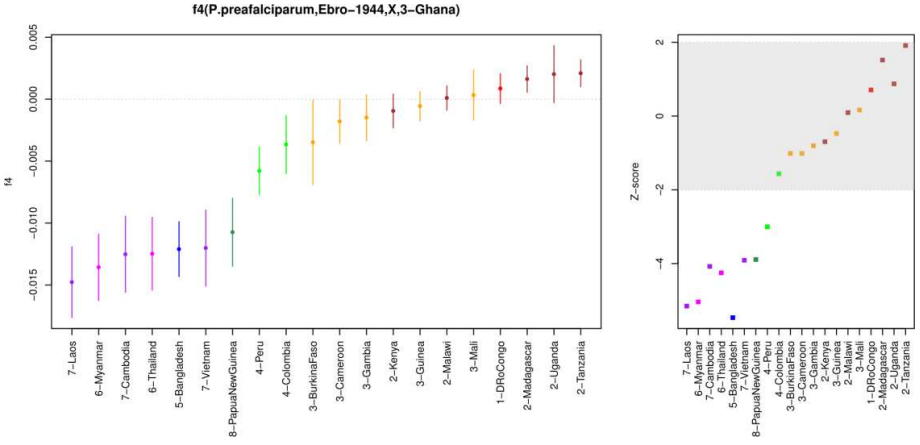

f4(P.preafalciparum,Ebro−1944,X,7−Cambodia)
f4
Z−score
7−Laos
6−Myanmar
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Cameroon
3−Guinea
3−Gambia
2−Kenya
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,7−Laos)
f4
Z−score
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Gambia
3−Cameroon
2−Kenya
3−Guinea
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,7−Vietnam)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
5−Bangladesh
6−Thailand
8−PapuaNewGuinea
4−Peru
4−Colombia
3−Cameroon
3−BurkinaFaso
3−Gambia
2−Kenya
3−Guinea
3−Ghana
2−Malawi
1−DRoCongo
3−Mali
2−Tanzania
2−Madagascar
2−Uganda

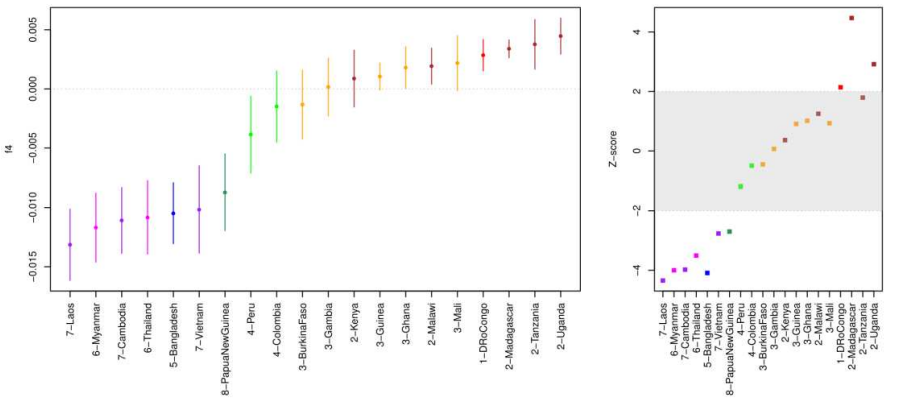

f4(P.preafalciparum,Ebro−1944,X,8−PapuaNewGuinea)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
7−Vietnam
5−Bangladesh
4−Peru
3−BurkinaFaso
4−Colombia
3−Cameroon
3−Gambia
3−Guinea
2−Kenya
2−Malawi
3−Ghana
3−Mali
1−DRoCongo
2−Tanzania
2−Madagascar
2−Uganda

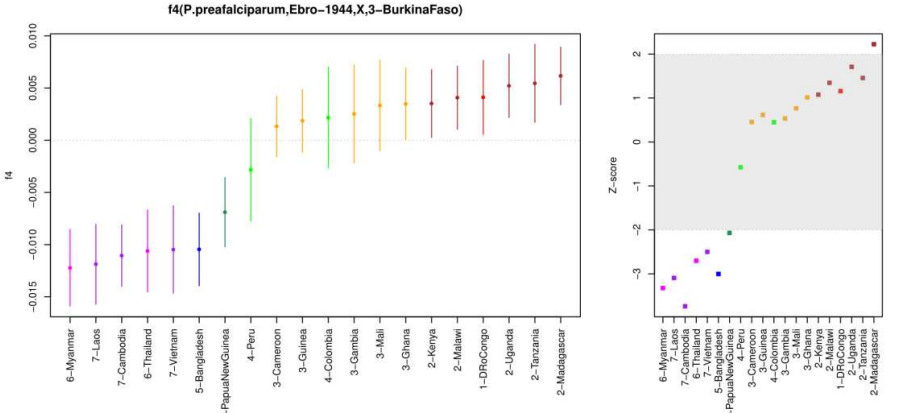

f4(P.preafalciparum,Ebro−1944,X,2−Kenya)
f4
Z−score
7−Laos
6−Myanmar
5−Bangladesh
6−Thailand
7−Cambodia
7−Vietnam
8−PapuaNewGuinea
4−Peru
3−BurkinaFaso
4−Colombia
3−Gambia
3−Cameroon
3−Guinea
3−Ghana
2−Malawi
1−DRoCongo
3−Mali
2−Uganda
2−Madagascar
2−Tanzania


f4(P.preafalciparum,Ebro−1944,X,2−Madagascar)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
7−Vietnam
5−Bangladesh
6−Thailand
8−PapuaNewGuinea
4−Peru
3−BurkinaFaso
4−Colombia
3−Cameroon
3−Gambia
2−Kenya
3−Guinea
3−Ghana
2−Malawi
1−DRoCongo
3−Mali
2−Tanzania
2−Uganda


f4(P.preafalciparum,Ebro−1944,X,2−Malawi)
f4
Z−score
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
3−BurkinaFaso
4−Colombia
3−Cameroon
3−Gambia
2−Kenya
3−Guinea
3−Ghana
3−Mali
2−Madagascar
1−DRoCongo
2−Tanzania
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,2−Tanzania)
f4
Z−score
7−Laos
6−Myanmar
6−Thailand
5−Bangladesh
7−Cambodia
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Gambia
3−Cameroon
3−Guinea
2−Kenya
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Uganda
f4(P.preafalciparum,Ebro−1944,X,2−Uganda)
f4
Z−score
7−Laos
6−Myanmar
7−Vietnam
7−Cambodia
6−Thailand
5−Bangladesh
8−PapuaNewGuinea
4−Peru
3−BurkinaFaso
4−Colombia
3−Cameroon
3−Guinea
3−Gambia
2−Kenya
2−Malawi
3−Ghana
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania


f4(P.preafalciparum,Ebro−1944,X,3−Ghana)
f4
0.005
0.000
−0.005
−0.010
−0.015
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Cameroon
3−Gambia
2−Kenya
3−Guinea
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Uganda
2−Tanzania
Z−score
2
0
−2
−4
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Cameroon
3−Gambia
2−Kenya
3−Guinea
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Uganda
2−Tanzania


f4(P.preafalciparum,Ebro−1944,X,3−Cameroon)
f4
0.005
0.000
−0.005
−0.010
−0.015
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Gambia
2−Kenya
3−Guinea
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda
Z−score
4
2
0
−2
−4
7−Laos
6−Myanmar
7−Cambodia
6−Thailand
5−Bangladesh
7−Vietnam
8−PapuaNewGuinea
4−Peru
4−Colombia
3−BurkinaFaso
3−Gambia
2−Kenya
3−Guinea
3−Ghana
2−Malawi
3−Mali
1−DRoCongo
2−Madagascar
2−Tanzania
2−Uganda


f4(P.preafalciparum,Ebro−1944,X,3−BurkinaFaso)
f4
0.010
0.005
0.000
−0.005
−0.010
−0.015
6−Myanmar
7−Laos
7−Cambodia
6−Thailand
7−Vietnam
5−Bangladesh
8−PapuaNewGuinea
4−Peru
3−Cameroon
3−Guinea
4−Colombia
3−Gambia
3−Mali
3−Ghana
2−Kenya
2−Malawi
1−DRoCongo
2−Uganda
2−Tanzania
2−Madagascar
Z−score
2
1
0
−1
−2
−3
6−Myanmar
7−Laos
7−Cambodia
6−Thailand
7−Vietnam
5−Bangladesh
8−PapuaNewGuinea
4−Peru
3−Cameroon
3−Guinea
4−Colombia
3−Gambia
3−Mali
3−Ghana
2−Kenya
2−Malawi
1−DRoCongo
2−Uganda
2−Tanzania
2−Madagascar

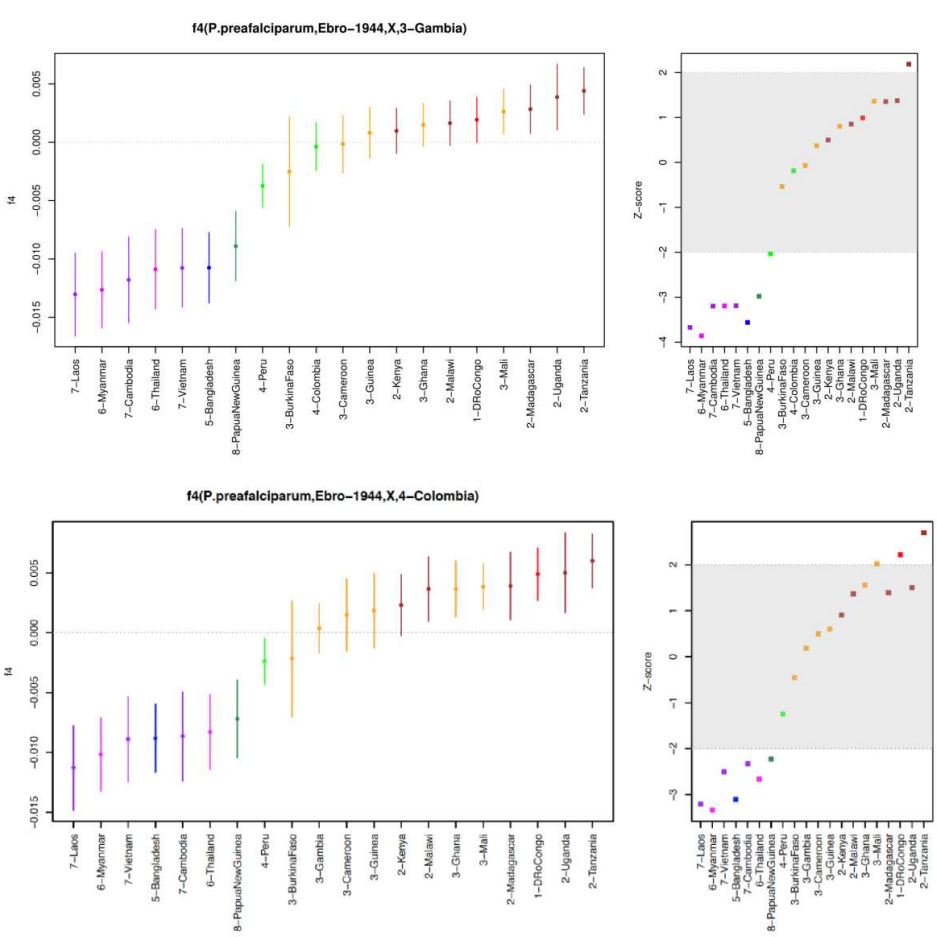


Figure S4: *f*4-statistics. (left-panel) *f*4 statistics and 5-95% confidence intervals (y-axis) testing the relationship *f*4(*P. preafalciparum*, Ebro-1944; X, Y), where X and Y iterate through all groups included in our global dataset, as given in the header of each plot. (right-panel) Z-scores following jackknife resampling where an absolute Z-score greater than 2 is considered significant. Regional groupings are coloured as in Fig. 1 of the main text. A negative *f*4 statistic indicates that Ebro-1944 has a greater affinity to X (x-axis label) over Y.

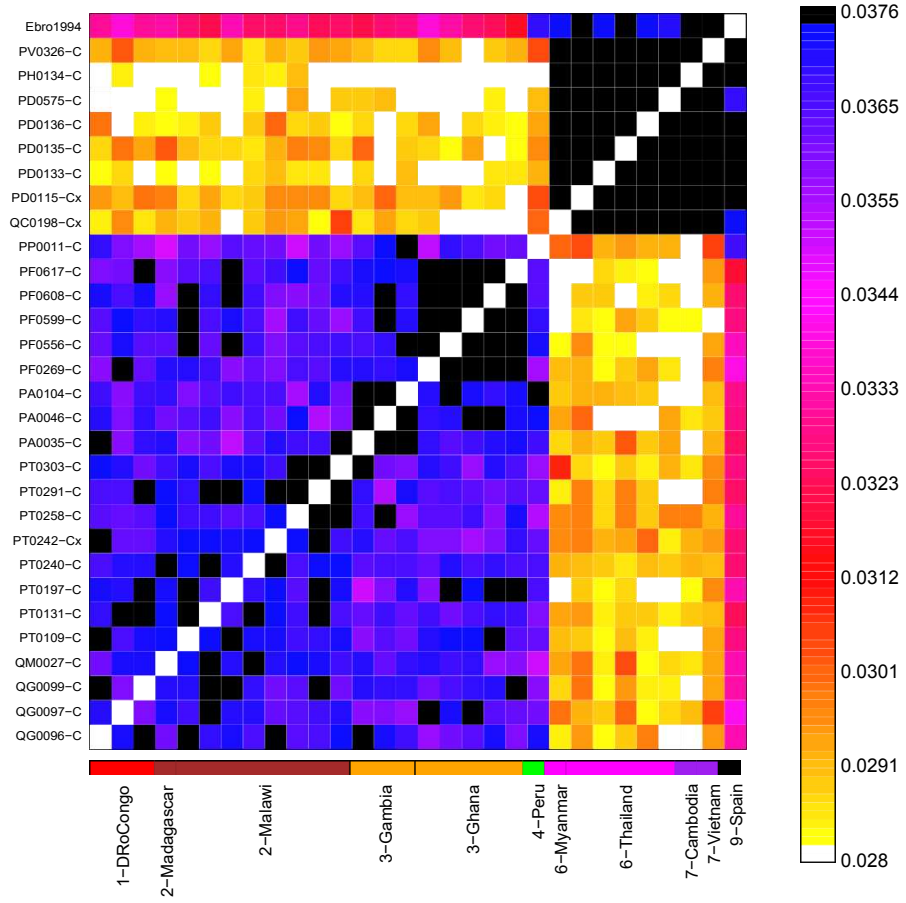


Figure S5: Chromopainter's inferred proportion of haplotypes shared (colour scale) between 30 strains of *P. falciparum* with 100% SNP overlap across 8195 variant sites. The x-axis colour provides the continental region where strains were collected. Ebro-1944 shares more haplotypes with strains from central south Asia relative to Africa.

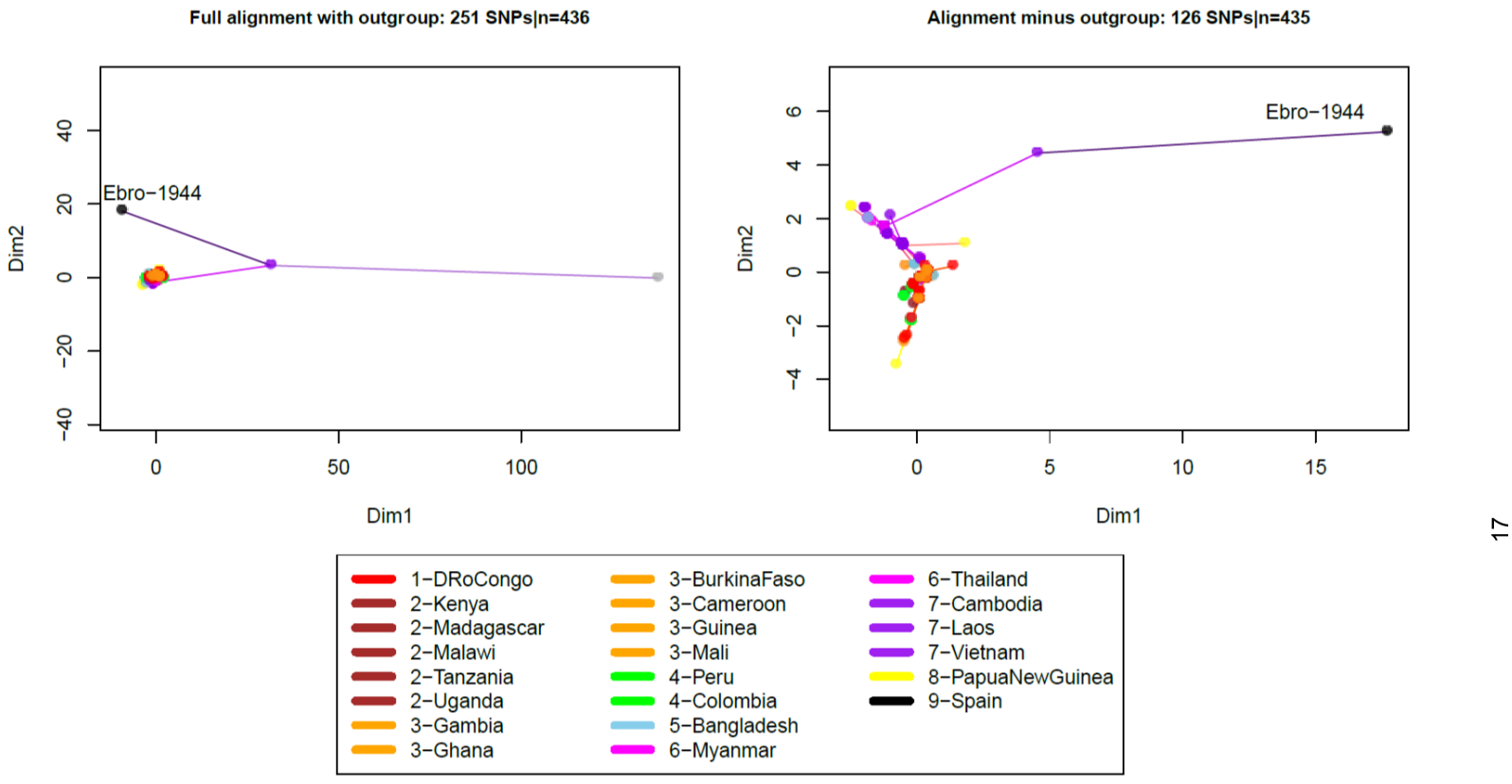


Figure S6: Minimum spanning network of the Ebro1944 mitochondrial genome.

Table S1: *Plasmodium falciparum* strains used in the population genetics analyses. See supplementary excel file.

Table S2: *Plasmodium falciparum* drug resistance variants described in the literature and screened in the Ebro-1944 strain. See supplementary excel file.

Table S3: mtDNA mutations overlapping in different slides. Distribution of the three geographically diagnostic mtDNA mutations across the four analysed slides. None of the ancestral variants are present in any slide, which suggests that the four slides contain a very similar *P. falciparum* strain. The nt276 and nt2763 positions are only found in Indian strains; the nt725 position is found in Indian and East African strains.

mtDNA nt position	Slide CA	Slide CM	Slide POS	Slide Lane8
276	Derived (N=12)	Derived (N=3)	Derived (N=8)	Derived (N=4)
725	Derived (N=8)	not covered	Derived (N=1)	Derived (N=3)
2763	Derived (N=23)	Derived (N=2)	not covered	Derived (N=5)

.

Supplementary References

1. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A. 2013;110(39):15758–63.
2. March S, Ng S, Velmurugan S, Galstian A, Shan J, Logan DJ, et al. A microscale human liver platform that supports the hepatic stages of *Plasmodium falciparum* and *vivax*. Cell Host Microbe. 2013;14(1):104–15.
3. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;5(6):pdb.prot5448.
4. Amato R, Miotto O, Woodrow CJ, Almagro-Garcia J, Sinha I, Campino S, et al. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5:e08714.
5. Otto TD, Gilabert A, Crellen T, Böhme U, Arnathau C, Sanders M, et al. Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. Nat Microbiol. 2018;3(6):687–97.
6. Helmby H, Cavelier L, Pettersson U, Wahlgren M. Rosetting Plasmodium falciparum-infected erythrocytes express unique strain-specific antigens on their surface. Infect Immun. 1993;61(1):284–8.
7. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, et al. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell. 1995;82(1):77–87.
8. Oberli A, Slater LM, Cutts E, Brand F, Mundwiler-Pachlatko E, Rusch S, et al. A *Plasmodium falciparum* PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface. FASEB J. 2014;28(10):4420–33.

9. Nunes MC, Okada M, Scheidig-Benatar C, Cooke BM, Scherf A. *Plasmodium falciparum* fikk kinase members target distinct components of the erythrocyte membrane. PLoS One. 2010;5(7):e11747.

10. Goel S, Palmkvist M, Moll K, Joannin N, Lara P, Akhouri RR, et al. RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. Nat Med. 2015;21:314.

11. McRobert L, Preiser P, Sharp S, Jarra W, Kaviratne M, Taylor MC, et al. Distinct trafficking and localization of STEVOR proteins in three stages of the *Plasmodium falciparum* life cycle. Infect Immun. 2004;72(11):6597–602.

12. Lavazec C, Sanyal S, Templeton TJ. Hypervariability within the Rifin, Stevor and Pfmc-2TM superfamilies in P*lasmodium falciparum*. Nucleic Acids Res. 2006; 34: 6696–6707.

13. Su X zhuan, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell. 1995;82(1):89–100.

14. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, et al. stevor and rif are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol. 1998;97(1–2):161–76.

15. Fernandez V. Small, Clonally Variant Antigens Expressed on the Surface of the *Plasmodium falciparum*-infected Erythrocyte Are Encoded by the rif Gene Family and Are the Target of Human Immune Responses. J Exp Med. 2002;190(10):1393–404.

16. Kyes SA, Rowe JA, Kriek N, Newbold CI. Rifins: A second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2002;96(16):9333–8.

17. Kaviratne M, Khan SM, Jarra W, Preiser PR. Small Variant STEVOR Antigen Is Uniquely Located within Maurer's Clefts in *Plasmodium falciparum* -Infected Red

Blood Cells. Eukaryot Cell. 2002;1(6):926–35.

18. Niang M, Bei AK, Madnani KG, Pelly S, Dankwa S, Kanjee U, et al. STEVOR is a *Plasmodium falciparum* erythrocyte binding protein that mediates merozoite invasion and rosetting. Cell Host Microbe. 2014;16(1):81–93.
19. Spielmann T, Fergusen DJP, Beck H-P. etramps, a new *Plasmodium falciparum* gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface. Mol Biol Cell. 2003;14(4):1529–44.
20. Winter G, Kawai S, Haeggström M, Kaneko O, von Euler A, Kawazu S, et al. SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. J Exp Med. 2005;201(11):1853–63.
21. Oksanen J, Blanchet FG, Friendly M, Roeland Kindt P, Legendre DM, Minchin PR, et al. Vegan: Community Ecology Package. R Package Version 2.2-0. 2019. Available at: http://CRAN.Rproject.org/package=vegan.
22. Sallares R, Gomzi S. Biomolecular archaeology of malaria. Ancient Biomolecules 2001; 3: 195–213.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):254–60.
25. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98(1):116–26.
26. Broad Institute. *Plasmodium vivax* Hybrid Selection initiative [Internet].

4.2 Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743–1793).

Toni de-Dios, Lucy van Dorp, Philippe Charlier, Sofia Morfopoulou, Esther Lizano, Celine Bon, Corinne Le Bitouzé, Marina Alvarez-Estape, Tomas Marquès-Bonet, François Balloux, Carles Lalueza-Fox

Infect Genet Evol. 2020 Jun;80:104209.doi: 10.1016/j.meegid.2020.104209. Epub 2020 Jan 29.

Infection, Genetics and Evolution 80 (2020) 104209

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743–1793)

Toni de-Dios[a,1], Lucy van Dorp[b,1,*], Philippe Charlier[c,d,1], Sofia Morfopoulou[b,e], Esther Lizano[a], Celine Bon[f], Corinne Le Bitouzé[g], Marina Alvarez-Estape[a], Tomas Marquès-Bonet[a,h,i,j], François Balloux[b,*], Carles Lalueza-Fox[a,*]

[a] Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[b] UCL Genetics Institute, University College London, London WC1E 6BT, UK
[c] Département de la Recherche et de l'Enseignement, Musée du Quai Branly - Jacques Chirac, 75007 Paris, France
[d] Université Paris-Saclay (UVSQ), Laboratory Anthropology, Archaeology, Biology (LAAB), 78180 Montigny-le-bretonneux, France
[e] Division of Infection and Immunity, University College London, London WC1E 6BT, UK
[f] Département Hommes, Natures, Sociétés, Muséum National d'Histoire Naturelle, 75116 Paris, France
[g] Archives Nationales, 75004 Paris, France
[h] Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain
[i] CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), 08036 Barcelona, Spain
[j] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain

ARTICLE INFO

Keywords:
Ancient DNA
Metagenomics
Infection

ABSTRACT

The French revolutionary Jean-Paul Marat (1743–1793) was assassinated in 1793 in his bathtub, where he was trying to find relief from the debilitating skin disease he was suffering from. At the time of his death, Marat was annotating newspapers, which got stained with his blood and were subsequently preserved by his sister. We extracted and sequenced DNA from the blood stain and also from another section of the newspaper, which we used for comparison. Results from the human DNA sequence analyses were compatible with a heterogeneous ancestry of Marat, with his mother being of French origin and his father born in Sardinia. Metagenomic analyses of the non-human reads uncovered the presence of fungal, bacterial and low levels of viral DNA. Relying on the presence/absence of microbial species in the samples, we could cast doubt on several putative infectious agents that have been previously hypothesised as the cause of his condition but for which we detect not a single sequencing read. Conversely, some of the species we detect are uncommon as environmental contaminants and may represent plausible infective agents. Based on all the available evidence, we hypothesize that Marat may have suffered from a fungal infection (seborrheic dermatitis), possibly superinfected with bacterial opportunistic pathogens.

1. Introduction

Jean-Paul Marat (1743–1793) was a famous French physician, scientist and journalist, best known for his role as Jacobin leader during the French Revolution. Marat's parents were Giovanni Mara, born in Cagliari, Sardinia, who later added a "t" to his family name to give it a French feel and Louise Cabrol, a French Huguenot from Castres. Marat was stabbed to death in his bathtub by the Girondist' supporter Charlotte Corday on July 13th, 1793 (Fig. 1a). Upon his death, his sister Charlotte Albertine kept two issues of Marat's newspaper *l'Ami du Peuple* (n°506 and n°678, published on June 30th, 1791 and August 13th, 1792, respectively), which he was annotating the day of his assassination and that got stained with his blood (Fig. 1b). Albertine gave the issues to the collector François-Nicolas Maurin (1765–1848) in 1837. After his death, as explained by a handwritten note by writer Anatole France dated from October 10th, 1864, the two issues ended up in the possession of baron Carl De Vinck who in 1906 donated them to the Département des Estampes, Bibliothèque National de France, in Paris (see notice in the Catalogue Général: https://catalogue.bnf.fr/ark:/12148/cb40261215w).

Marat's health during the last years before his assassination is shrouded in mystery. He suffered from a severe itching skin disease

* Corresponding authors.
E-mail address: lucy.dorp.12@ucl.ac.uk (L. van Dorp).
[1] These authors equally contributed to this work.

https://doi.org/10.1016/j.meegid.2020.104209
Received 8 November 2019; Received in revised form 24 January 2020; Accepted 26 January 2020
Available online 29 January 2020

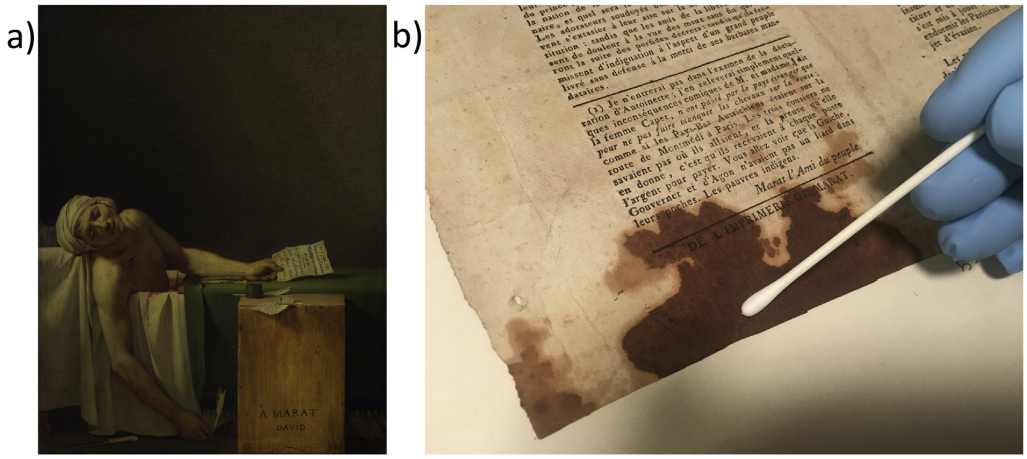


Fig. 1. a) "La mort de Marat"; portrait of Jean-Paul Marat after his assassination, by Jacques-Louis David (1793). Preserved at Musées Royaux des Beaux-Arts de Belgique, Brussels. b) Sampling the page of l'*Ami du Peuple* stained with Marat's blood that has been analysed.

from which he found some relief by spending most of his time in a medicinal bathtub over which he placed a board to use as a writing desk. His condition, which he attributed to his stay in the sewers of Paris while hiding from his political enemies, has been the subject of numerous medical debates and has been alternatively attributed to scabies, syphilis, atopic dermatitis, seborrheic dermatitis or dermatitis herpetiformis (Bayon, 1945; Cohen and Cohen, 1958; Dotz, 1979; Jelinek, 1979; Dale, 1952), the latter as a potential manifestation of celiac disease (Coto-Segura et al., 2011). It has been suggested that his condition affected his character and turned it more violent (Bayon, 1945).

With the intention of shedding light on these issues, we retrieved two samples from one of the newspapers stained with Marat's blood, one sample from the blood stain and a second one from a non-stained area in the upper corner of the paper, to be used as a comparison. A principal concern was to use a non-destructive approach to explore Marat's genomic footprint; therefore, the samples were taken with forensic swabs. The DNA extracted from both samples was used to build genomic libraries that were subjected to second-generation sequencing using the Illumina platform. DNA reads where subsequently classified, separating the human reads – most likely deriving from Marat's blood – from those assigned to microbial species. The analysis of both sets of DNA sequences allowed characterisation of Marat's ancestry as well as identification of the potential pathogens responsible for his debilitating skin condition.

2. Material and methods

2.1. DNA extraction and sequencing

Forensic swabs were obtained from one of the newspapers Marat was annotating at the time of his assassination (Fig. 1). One swab was taken from the blood stain and another from an area of the newspaper without visual evidence of blood. The blood swab was extracted with a buffer composed of 10 mM TrisHCl, 10 mM EDTA, 2 mM SDS, 50 mM DTT; proteinase K was added after one hour incubation. The extract was subsequently concentrated and purified using a Qiagen column kit. DNA extraction from both swabs was performed together with extraction blanks (no sample). A total of 35 µl of each sample was used for library preparation following the BEST protocol (Carøe et al., 2018). Libraries were quantified using BioAnalyzer and sequenced by HiSeq 4000 (Illumina). Library blanks were also performed for each library batch. We generated 568,623,176 DNA reads from the blood stain, of which 74,244,610 reads mapped to the human reference genome (Table S1).

2.2. Mapping and variant calling

Raw sequences adapters were removed using *Cutadapt (Martin, 2011)*. Reads were then aligned against the Human Reference genome (GRCh37/hg19) and the mitochondrial reference genome (rCRS) as well as for a set of microbial candidates using *BWA* v.07.3 (Li and Durbin, 2009) and Bowtie2 (Langmead and Salzberg, 2012). We employed two aligners as mapping sensitivity of different aligners can vary between different samples when working with aDNA (Taron et al., 2018). Duplicate reads were discarded using *Picard* tools (Broad Institute, 2020). Unique mapped reads were filtered for a mapping quality equal of above 30 (Table S1). All mapped sequences (human nuclear, human mitochondrial and microbial) were assessed for post-mortem damage patterns at the ends of reads using *MapDamage* v.2 (Jónsson et al., 2013), which can be used as a sign of historic authenticity over modern contamination (Fig. S1). Post-mortem damage signals were also obtained for each read using *pmdtools (Skoglund et al., 2014)* (Fig. S2). Mapping statistics including the depth of coverage were recorded using *Qualimap (García-Alcalde et al., 2012)*. Due to the low coverage of the human sample, we performed a pseudo-haploid calling approach, common to the processing of aDNA, using the *SAMtools Pileup* tool (Li et al., 2009). This data was then merged with the Human Origins dataset for its use in population genetics analyses (Lazaridis et al., 2014; Lazaridis et al., 2016).

2.3. Modern DNA contamination

Schmutzi was used to estimate the amount of modern DNA contamination in the mitochondrial (mtDNA) genome (Renaud et al., 2015) likely deriving from the DNA of those who have handled the newspaper in the years following Marat's death. We identified mitochondrial contamination based on the inferred deamination patterns as 52.5% +/− 4.5% with the full haplogroup profiles provided in Table S2. This allowed the modern DNA sequences to be delineated from the ancient DNA sequences using *Jvarkit* and a custom script (Pierre, 2015) by selecting the human reads with mismatches in their first or last three nucleotides. This reduced the amount of modern mtDNA contamination to 0–0.1%. The depth of coverage was recorded

using *Qualimap*.

We also independently estimated the amount of contamination based on the heterozygous sites in the X chromosome using *angsd v0.925–21* (Korneliussen et al., 2014; Skoglund et al., 2013). We obtained an estimate of 3.2% modern contamination. As with the mtDNA genome we filtered reads with mismatches in the first or last three nucleotides, taking forward only those reads for additional population genetics analyses. After applying both filters, the resultant mean depth of coverage for the mitochondrial genome was 4.038× and 0.029× for the nuclear genome.

2.4. Uniparental markers and sex determination analyses

The mtDNA haplogroup was determined using *SAMtools* pileup tool calling the positions defined in the *Phylotree* database (van Oven, 2015). We used a genome browser (*IGV*.v2.4.14) to study the genomic context of each possible SNP (Robinson et al., 2011). Only those SNPs that were present in two or more reads, and those which were not located at the ends of the reads, were considered. The contamination was estimated by calculating the ratio of discordant reads at haplogroup-diagnostic positions. Molecular sex was assigned with Ry_compute (Skoglund et al., 2013), a script designed for the sex identification of low coverage individuals (Fig. S3).

2.5. Population genetics analysis

Principal Component Analysis (PCA) was performed using *SmartPCA* in *EIG* v6.0.1 with a subset of modern individuals from the Human Origins dataset (Patterson et al., 2006). This subset contained 434 present-day Europeans and 616,938 autosomal SNPs, plus our sample (Fig. 2). The Marat sample was projected using the option *lsqproject*. We also considered the Marat sample projected into an expanded dataset of West Eurasian populations (Fig. S4). As projected individuals' components tend to 0, we also carried out a control analysis using Han Chinese, French and Marat (Fig. S5). The results were visualised using the R package *GGplot2 (*Wickham, *2016)*. This dataset confirmed that Marat is not artefactually placed at the centre of the plot.

To formally test the relationship of the Marat sample to relevant geographic regions we calculated *f4* statistics of the form *f4*(Mbuti, Marat; X. Y) where X and Y are tested for combinations of possible ancestral sources: Sardinian, French, English, Italian_North, Basque, Spanish. *f4* values were calculated in qpDstat of AdmixTools v.5.0 (Patterson et al., 2012) with statistical significance assessed through *Z*-scores following jack-knife resampling (Table S3). This statistic tests the covariance in allele frequency differences between an African outgroup (Mbuti) and Marat relative to the clade formed by X and Y. Positive values of *f4* indicate a closer affinity of Marat to Y relative to X, with negative values indicating a closer relationship of Marat to X relative to Y.

We additionally ran an unsupervised clustering analysis using *ADMIXTURE* v1.3 and another subset of the Human Origins dataset (Alexander et al., 2009). This subset included 881 individuals from Europe, West Asia and North Africa typed over 616,938 shared autosomal SNPs. We filtered the dataset by removing SNPs in high linkage disequilibrium using PLINK.v1.9 (Purcell et al., 2007), removing all SNPs with a r^2 threshold of 0.4 within a 200 SNP sliding window, advancing by 50 SNPs each time. We performed the clustering analysis using K values ranging from 1 to 10, with 10 replicates for each value of K. We selected K according to the lowest cross-validation error value (K = 4). The *ADMIXTURE* results at K = 4 were visualised using Pong (Behr et al., 2016) (Fig. 2).

2.6. Metagenomic analysis

We first removed adapters and merged the paired-end reads into longer single-end sequences using AdapterRemoval v2 (Schubert et al., 2016). We removed PCR duplicates with exact sequence identity using dedupe from the BBMap suite of tools (https://sourceforge.net/projects/bbmap/). We subsequently used the default preprocessing pipeline designed for metaMix which consists of removing human and rRNA sequences using bowtie2 followed by megaBLAST, as well as low quality and low complexity reads using prinseq (Schmieder and Edwards, 2011) (−lc_method dust -lc_threshold 7 -min_qual_mean 15). The number of reads filtered at each step are provided in Table S4. We screened the remaining high quality DNA reads for the presence of possible pathogens using both KrakenUniq (Breitwieser et al., 2018) against the Kraken database compiled in Lassalle et al. 2018 (Lassalle et al., 2018) and metaMix (Morfopoulou and Plagnol, 2015) using megaBLAST and a local custom database consisting of the RefSeq sequences of bacteria, viruses, parasites, fungi and human, as of July 2019. KrakenUniq was run with default parameters. The metaMix-nucleotide mode was run with the default read support parameter of 10 reads was used (Table S5) and the default number of 12 MCMC chains. The number of the MCMC iterations is automatically calculated by metaMix based on the number of species to explore for each dataset, resulting in 10,000 iterations for the blood sample and 3230 iterations for the paper swab.

The relative proportion of reads assigned to different species by KrakenUniq and metaMix was highly correlated; R^2 = 0.94 and R^2 = 0.82, for the blood stain and the unstained paper, respectively (Fig. S6). However, metaMix tended to assign a higher number of reads to individual species, closer to the number found by mapping directly to the microbial genomes and we observed important discrepancies for the number of reads assigned to some of the species (Table S6). Additionally, metaMix results for both the blood stain and the unstained paper consisted of fewer species compared to KrakenUniq, even when the same read support threshold was applied to KrakenUniq, indicating increased specificity due to the MCMC exploration of the species space, that comes at an increased computational cost.

In order to compare the accuracy of the two assignment tools, we further explored the presence of clinically relevant species by mapping the quality-filtered subset of reads (Table S2) used in metagenomic assignment against the reference genomes of different candidate genera of fungi and bacteria using *bowtie2* (Langmead and Salzberg, 2012) and *BWA* v.07.3 (Fig. S7-S13). For all reads mapping to individual reference genomes, mapDamage v2 (Jónsson et al., 2013) was also run to assess evidence of nucleotide mis-incorporation characteristic of post-mortem damage. These mapping results were systematically supporting the metaMix assignments over those obtained with KrakenUniq (Table S6). This led us to rely on metaMix for all metagenomic assignments presented in the paper.

Besides testing for the presence and absence of species, we tested whether some microorganisms were overrepresented in the blood stain compared to the unstained section of the paper using a one-sided binomial test and a significance threshold of 0.95 (Table S5).

As an additional control, we also conducted metagenomic analysis of two publicly available ancient metagenomes obtained from parchment of comparable age to the Marat newspaper (Teasdale et al., 2015a). We followed the same pre-processing pipeline described for the Marat samples, first removing adapters and PCR duplicates before employing the default metaMix pre-processing pipeline, this time removing reads that mapped to either the human or sheep, cow and goat reference genomes. As before, metaMix-nucleotide mode was run with a read support parameter of 10 reads and with 12 MCMC chains x 2325 and 6130 iterations respectively for ERR466100 and ERR466101. We provide the breakdown of read filtering steps in Table S7 and our raw metaMix results in Table S8.

2.7. Phylogenetic analysis

In the case of *Malassezia*, a phylogenetic analysis of the

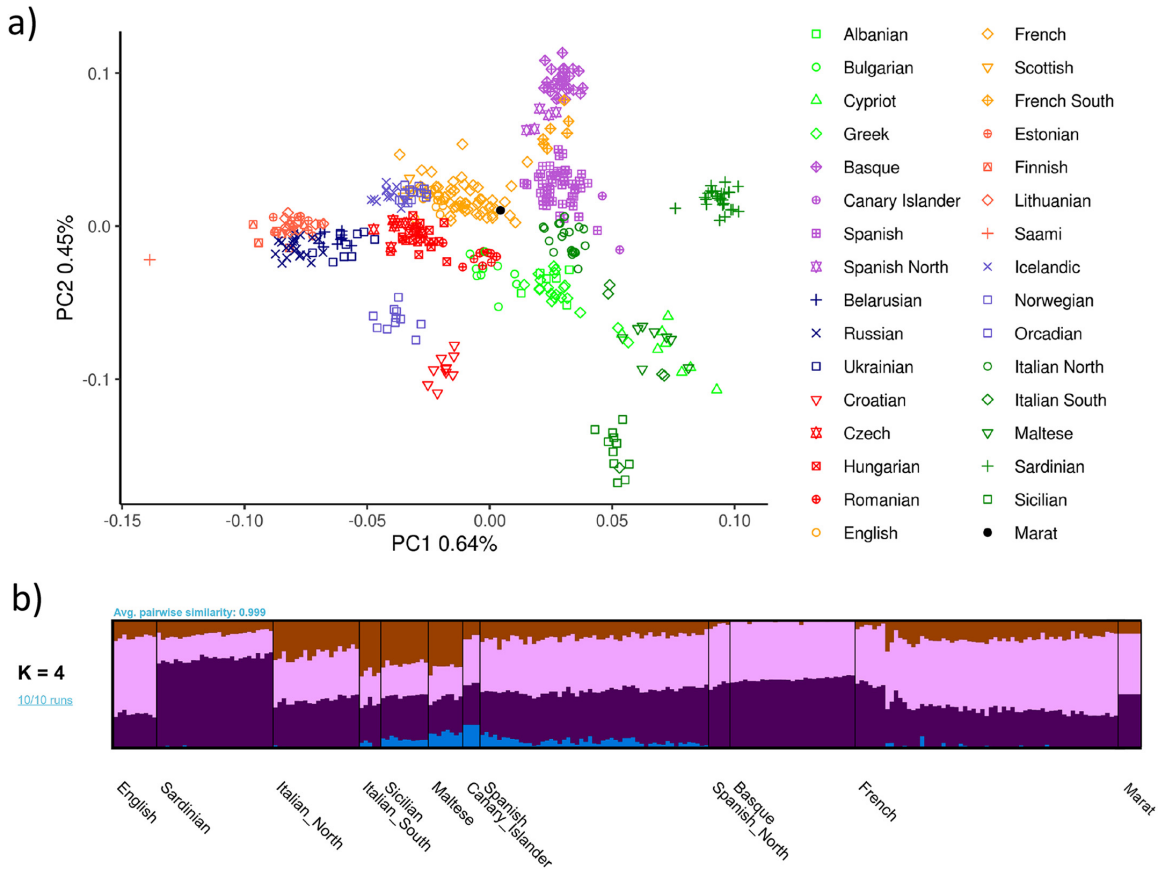


Fig. 2. a) Principal Component Analysis (PCA) of modern human European populations with Marat's ancient DNA reads projected. Symbols provide the country and region, where provided, as given in the legend at right. b) Admixture analysis with modern European samples and Marat. Both analyses are coherent with Marat's suggested French and Italian combined ancestry.

mitochondrial DNA genome with available modern strains on the Short Read Archive (SRA) was performed. We called variant positions using *GATK UnifiedGenotyper (*McKenna et al., *2010)* and generated a Maximum Likelihood tree using *RAxML-NG* specifying a GTR substitution model and 100 bootstrap resamples (Kozlov et al., 2019). The tree was rooted with *M. globosa* (Fig. S10).

We also conducted a phylogenetic analysis for *C. acnes*, combining our historical strain with all *C. acnes* genomes deposited in the SRA covering the reference at an average depth $> 10\times$, and with *C. namnetense* as an outgroup (SRR9222443). The only *C. acnes* genomes sequenced at medium to high depth are those reported by Gomes et al. 2017 (Gomes et al., 2018). A Maximum Likelihood tree was generated over the 21,751 SNP alignment using RAxML-NG (Fig. S13) and clonal complexes and phylotypes were assigned based on the PubMLST *C. acnes* definitions database (https://pubmlst.org/bigsdb?db=pubmlst_pacnes_seqdef).

3. Results

3.1. Human ancestry analysis

We generated 568,623,176 DNA reads from the blood stain, of which 74,244,610 reads mapped to the human reference genome (Table S1). From these, we retrieved a complete human mitochondrial (mtDNA) genome at a mean depth of coverage of $4.038\times$ and the nuclear genome at $0.029\times$ (Table S1). The predominant mtDNA haplotype was H2a2a1f, although we found evidence of some additional mtDNA sequences, notably a K1a15 haplotype. The ratio of sexual chromosome to autosomal DNA reads indicated that the sample donor was male (Fig. S3).

The human DNA reads showed evidence of post-mortem deamination occurring in 1% of the ends of sequencing reads, indicating authentic ancient DNA damage (Fig. S1-S2). This is similar to the degree of damage that has been observed in aDNA obtained from other human specimens of a similar age (Rasmussen et al., 2011). For further analyses we selected only those reads that displayed C to T or G to A substitutions at the 5′ or 3′ end, respectively. After this procedure, the degree of mitochondrial contamination was reduced to 0–0.01%.

To explore the ancestry of Marat in the context of modern European populations, we performed Principal Component Analysis (PCA) (Fig. 2a and Fig. S4–5) and unsupervised clustering in ADMIXTURE (Fig. 2b). Our sample projected among modern French individuals sampled from France in the population genetic analyses. This result is broadly compatible with proposed hypotheses relating to the ancestry of Marat (Cohen and Cohen, 1958). *f4* statistics suggest a closer affinity of Marat to modern Italian, English, Sardinian, Basque and French populations relative to those from Spain (Table S3). However, these trends are subtle and we note that mixed ancestries are difficult to

discern, especially when only limited genetic data is available.

3.2. Metagenomic analysis

We conducted metagenomic species assignments with the 9,788,947 deduplicated, quality controlled and low complexity filtered DNA reads (combined merged and non-merged) that did not map to the human genome (see Methods and Table S4). We used metaMix (Morfopoulou and Plagnol, 2015), a Bayesian mixture model framework developed to resolve complex metagenomic mixtures, which classified ~9% of the non-human reads into 1328 microbial species (Table S5). The species assignments were replicated with KrakenUniq (Breitwieser et al., 2018), which led to largely consistent, if less accurate, results (~7% classified into 3213 species, Fig. S6, Table S6). Thus, we relied on the metaMix species assignments throughout the paper, unless stated otherwise.

We detected the presence of a wide range of microorganisms, including some expected to develop on decaying cellulose and/or dried blood, but also others recognized as opportunistic human pathogens from the following bacterial genera: *Acidovorax, Acinetobacter, Burkholderia, Chryseobacterium, Corynebacterium, Cutibacterium, Micrococcus, Moraxella, Paraburkholderia, Paracoccus, Pseudomonas, Rothia, Staphylococcus, Streptococcus* and the fungal genera *Aspergillus, Penicillium, Talaromyces* and *Malassezia* as well as HPV (type 179 and type 5) and HHV6B viruses, albeit the latter supported by a very low numbers of reads (Table S5-S6). Some of the DNA reads, notably from *Aspergillus glaucus*, *Cutibacterium acnes*, *Malassezia restricta* and *Staphylococcus aureus* showed typical misincorporation patterns that are considered indicative of these sequences being authentically old (Fig. S7).

We additionally sequenced the swab taken from the unstained paper sample. In this case, only 96,252 pairs of reads were obtained (56,616 merged, 25,712 non-merged, 35,216 deduplicated and filtered combined merged and non-merged), with 52% of the reads that could be classified with metaMix into 66 species and 36% with KrakenUniq into 374 species, respectively (see Methods and Table S4). Although very little DNA could be retrieved from the section of the document that had not been blood-stained, we tried to identify microorganisms that were statistically significantly over represented in the blood stain relative to the unstained paper. Among these and besides, as expected, *Homo sapiens*, different species of *Aspergillus* and *Acinetobacter* were significantly overrepresented in the blood stain (Table S5). It remains questionable however whether the unstained paper represents a suitable negative control given that the newspaper had been extensively manipulated by Marat. Significant over representation of *Aspergillus* spp. and *Acinetobacter* spp. in the blood stain relative to the rest of the document could also be due to the blood providing better conditions for the growth of iron-limited microbes. Indeed, *Aspergillus spp.* and *Acinetobacter* spp. are commonly found in the environment but are also grown in blood agar. As such, it is plausible that these represent post-mortem contaminants. Indeed, for *Acinetobacter* spp. we identified no post-mortem damage pattern.

Metagenomic analysis of historical samples can be challenging as the resulting microbial communities typically comprise an unknown mixture of endogenous species as well as contaminants, both contemporary and modern. To mitigate this problem, we relied on a 'differential diagnostics' approach (Table 1), where we specifically tested for the presence of reads from pathogens that could plausibly have led to Marat's symptoms, most of which have been previously hypothesised in the literature (Bayon, 1945; Cohen and Cohen, 1958; Dotz, 1979; Jelinek, 1979; Dale, 1952). Such a differential approach is more stringent than the standard approach in clinical diagnostics aiming to identify the full list of microbes present in the samples after enforcement of a read-number threshold (Wilson et al., 2019; Miller et al., 2019). Our approach allows limiting the number of species to be tested to a small list of plausible candidates. Second, the lack of detection of even one read from a focal microbial species by direct mapping falsifies the null hypothesis that it was not involved in the disease.

Table 1
List of Diseases test, associated agents and presence in the blood stain and control samples. The following symbols denote the abundance of reads for each infectious agent tested ✓: present; ✓✓: top ten; ✓✓✓: top hit; ✗: absent.

Disease	Pathogen	Blood	Unstained paper
Syphilis	*Treponema pallidum*	✗	✗
Scrofula (tuberculosis)	*Mycobacterium tuberculosis*[a]	✗	✗
Leprosy	*Mycobacterium leprae*	✗	✗
Diabetic candidiasis (thrush)	*Candida* sp.	✗	✗
Scabies	*Sarcoptes scabiei*	✗	✗
Seborrheic dermatitis	*Malassezia* sp.	✓✓	✓
Atopic dermatitis	*Staphylococcus aureus*	✓	✗
Severe acneiform eruptions	*Cutibacterium acnes*	✓✓✓	✓✓

[a] Scrofula can also be caused by other Mycobacteria in particular *M. scrofulaceum* and *M. avium intracellulare*, which are also absent from both samples.

We did not identify a single sequencing read in either the blood stain or the unstained paper for the agents of syphilis, leprosy, scrofula (tuberculosis) and diabetic candidiasis (thrush) (Table 1, Table S5). We additionally tested for scabies, which is caused by burrowing of the mite *Sarcoptes scabiei* under the skin. Since the metagenomic reference database did not include arthropod genomes, this was tested separately by blasting all the non-human reads against the *Sarcoptes scabiei* genome (GCA_000828355.1). Again, we detected not a single read matching to *Sarcoptes scabiei*, which makes scabies an implausible cause for Marat's skin disease (Table 1, Table S5).

Conversely, metaMix recovered 15,926 and 83 filtered DNA reads from the blood stain and the unstained paper respectively, assigned to *Malassezia restricta* a fungal pathogen causing seborrheic dermatitis, which has been previously hypothesised as one of the most plausible causes for Marat's condition (Bayon, 1945; Cohen and Cohen, 1958; Dotz, 1979; Jelinek, 1979). Direct mapping of all reads to *M. restricta* (GCA_003290485.1) resulted in 19,194 reads from the blood stain dataset mapping over 17.17% of the reference genome. KrakenUniq failed to identify *M. restricta*, instead assigning 627 reads sequenced from the blood stain to *M. sympodialis*. However, further analysis of the *Malassezia* reads based on genome mapping pointed to most (80.3%) being uniquely assigned to *M. restricta* rather than *M. sympodialis* (Fig. S8). This allowed us to reconstruct a complete *M. restricta* mtDNA genome at 0.84× coverage. The *Malassezia* reads were evenly distributed along the full genome assembly supporting no mixing or misclassification of the species (Fig. S9).

We placed our Marat *M. restricta* mitochondrial genome in phylogenetic context by building a maximum likelihood phylogeny including our historical strain and available present-day mtDNA *M. restricta* genomes. Although the total number of samples is small, the fact that the *M. restricta* mtDNA molecule recovered from Marat's blood is placed basal to modern strains (Fig. S10) and exhibits some post-mortem damage (Fig. S5) further support its authenticity.

We also recovered 587 filtered reads assigned by metaMix to *Staphylococcus aureus* in the blood stain but none in the reads obtained from the unstained paper. The differential representation in the two samples is not significantly different due to the far lower number of reads in the unstained sample (Table S5). Although a common commensal, *S. aureus* is also a frequent human pathogen and the leading cause of atopic dermatitis. In order to confirm the metagenomic assignments to *S. aureus*, we mapped the raw microbial reads to a series of reference genomes from various species in the *Staphylococcus* genus. This allowed us to identify 888 reads mapping against the *S. aureus* reference genome, out of which 758 uniquely mapped to *S. aureus* (Fig. S11). The presence of *S. aureus*, but with a relatively low number of reads, may be compatible with a secondary infection by *S. aureus* rather

than *S. aureus* being the initial cause of Marat's condition. Alternatively, Marat, or someone who also handled the newspaper, could have carried *S. aureus* as a skin commensal.

The most prevalent microbial species in the blood stain was *Cutibacterium acnes* (formerly *Propionibacterium acnes (*Scholz and Kilian, 2016*)*), which was also present in the unstained paper (Table S5). *C. acnes* is largely a commensal and part of the normal skin biota present on most healthy adult humans' skin, including in association with *S. epidermis* which we also observe in our sample (Table S5-S6, Fig. S11) (Dreno et al., 2017). *C. acnes* is also a frequent contaminant in metagenomic samples (Salter et al., 2014; Mollerup et al., 2016). However, *C. acnes* can also be involved in severe acneiform eruptions (Platsidaki and Dessinioti, 2018) and we cannot exclude the possibility that it could have contributed to Marat's condition. 86,019 reads mapped to the *C. acnes* reference genome (GCF_000008345.1), yielding an alignment of 3.4× average coverage (Fig. S12) and exhibiting modest post-mortem damage (Fig. S7).

A phylogeny of Marat *C. acnes* with a collection of publicly available modern strains (Gomes et al., 2018; Mollerup et al., 2016) places our historic genome on a short branch falling basal to Type I strains, supporting its age and authenticity (Fig. S13). This phylogenetic placement suggests our Marat strain falls into *C. acnes* phylotype I (*C. acnes* subsp. *acnes*) rather than II (*C. acnes* subsp. *defendens*). Whilst our Marat strain does not cluster with phylotype Ia, the type more commonly associated with skin surface associated acne vulgaris (Lomholt and Kilian, 2010), its position, basal to Type Ib strains cannot exclude its involvement in soft or deep tissue infections (Nazipi et al., 2017).

Delineating contaminants and commensals from plausible pathogens remains challenging from this type of data source, in particular due to the absence of a suitable control. To alleviate this issue, we conducted full taxonomic assignments of two ancient metagenomes generated from historical parchment samples dating to the 17th and 18th centuries (PA1 and PA2 respectively) (Teasdale et al., 2015a). Although these samples were obtained from livestock (ruminant) skins whereas we are working with cellulose paper, we anticipate that they may have been used and handled in a comparable way to the newspaper Marat was annotating. In this way they represent what can be considered as the most biologically comparable ancient metagenomes available to date. An equivalent metaMix analysis applied to these filtered sequencing reads (Table S7) identified not a single read assigned to *M. restrica*, *S. aureus* or *C. acnes* (Table S8). We therefore do not systematically expect a significant number of reads for the three species we suggest as most plausible candidates for Marat's condition.

4. Discussion

Over the last decade, ancient-pathogen genomics has made great progress by borrowing technological advances originally developed for the study of human ancient DNA (Gelabert et al., 2016; Rasmussen et al., 2015). Although most microbial data has been secondarily generated from the sequencing of ancient human bones or teeth (Rasmussen et al., 2015; Rascovan et al., 2019; Margaryan et al., 2018; Mühlemann et al., 2018) other, rare samples, such as preserved tissues (Marciniak et al., 2016; Devault et al., 2014) or microscope slides from antique medical collections have been analysed (Gelabert et al., 2016; De-Dios et al., 2019). We are aware of no previous attempt to leverage ancient DNA technology to diagnose infections in historical characters, despite previous sequencing of remains from other prominent historical figures such as King Richard III and the putative blood of Louis XVI (King et al., 2014; Olalde et al., 2014).

In this work we analysed both human and 'off-target' microbial reads to shed light on an important historical figure of the French Revolution and his skin condition. Due to the loss of Marat's remains after their removal from the Panthéon in February 1795, the paper stained with his blood likely represents the only available biological material to study both his ancestry and the cause of his skin condition. Although second-generation sequencing techniques have been applied to the analysis of ancient parchments (Teasdale et al., 2015b) our work represents the first instance where this methodological approach has been applied to old cellulose paper.

The presence and relative abundance of different microorganisms in the documents Marat was annotating is affected by their endogenous presence as well as contemporary and modern contamination both for the blood and unstained sample. Some microorganisms present in the samples might reflect skin microbiome signatures. Whilst some other microorganisms represent environmental contaminants and are likely unrelated to Marat's condition. In order to identify the most likely candidates for Marat's condition we tested a set of proposed diagnoses, which we considered as plausible if we detected at least one read assigned to the causative infectious agent (Table 1).

Potential conditions for which we detected not a single supportive read included syphilis, tuberculosis (scrofula), leprosy, diabetic candidiasis or scabies. We appreciate that absence of evidence for an infectious agent does not constitute incontrovertible evidence of its absence. Moreover, it is not uncommon for metagenomic diagnostics applied to clinical samples to fail to identify reads from the likely infectious agent above the predefined diagnostic threshold, or even fail to detect any read at all (Miller et al., 2019; Scholz and Kilian, 2016). As such, the absence of reads from a putative pathogen makes it less plausible as the agent of Marat's condition but does not definitely rule them out.

Conversely, we detected and validated microbial reads for two of the conditions we tested, seborrheic dermatitis (*Malassezia spp.)*, atopic dermatitis (*Staphylococcus aureus*) and cannot exclude severe acneiform eruptions (*Cutibacterium acnes*) as a third, given the age and phylogenetic position of the *C. acnes* genome we obtained. For all three cases, the number of reads would have exceeded the threshold suggested for detection in clinical metagenomic diagnostic (Miller et al., 2019; Scholz and Kilian, 2016), even when considering the swab from the unstained paper as a control.

The presence of *Malassezia restricta* is of particular interest because this fungus is specialized to live on the skin (Saunders et al., 2012). Although also a common commensal and contaminant in metagenomic studies, *Malassezia* has been described in various skin conditions, including dandruff, atopic dermatitis, folliculitis and seborrheic dermatitis (Ashbee, 2010; Gupta et al., 2004). Interestingly, the latter symptoms would fit those described in Marat (Dale, 1952). The *M. restricta* reads we identified were not statistically significantly overrepresented in the blood's stain relative to the unstained paper, although they could be expected to be present in both samples if someone heavily infected was holding the newspaper. Although we cannot confidently claim the reads in Marat's blood are directly associated with Marat himself, we do identify post-mortem damage in these reads and a phylogenetic placement in a modern mitochondrial DNA phylogeny consistent with these reads being indeed old (Fig. S7, Fig. S10). We also do not systematically expect the presence of *M. restricta* on parchment of a similar age (Table S8).

Also of possible interest is the widespread presence of *Cutibacterium acnes* subsp. *acnes*, which although a common commensal or contaminant can also be implicated in severe acneiform eruptions, which constitutes the top hit in the blood sample and falls basal to phylotype I strains currently in circulation. As with *M. restricta*, we do not observe a single *C. acnes* read in two biologically equivalent historic parchment metagenomes (Table S8). *Staphylococcus aureus*, which is frequently detected in cases of atopic dermatitis, is also present in reads obtained from the blood's stain, although in fairly low number.

Whilst our results do not allow us to reach a definite diagnosis of Marat's condition, they allowed us to cast doubt on several previous hypotheses and provide, using all the available evidence, some plausible aetiologies. We suggest that Marat could have been suffering from an advanced fungal or polymicrobial infection, either primary or secondary to another condition. Future metagenomic analysis of additional

documents in Marat's possession during his assassination could help confirm the microbial composition found in this study and strengthen these observations.

Our work further illustrates the potential of sequencing technologies for the generation of (meta-)genomic information from difficult, singular samples and opens new avenues to address medical hypotheses of major historical interest.

Author contributions

P.C., L.v.D., F.B. and C.L.-F. conceived and designed the study; P.C. and C.L.B sourced the newspaper; C.B., M.A-E. and E.L. developed and performed laboratory analysis; T.d.-D., L.v.D., S.M. analysed data and performed computational analyses; T.d.-D., L.v.D., S.M., F.B., and C.L.-F. wrote the paper with inputs from all co-authors.

Acknowledgments and funding

This work was supported by Obra Social “La Caixa” and Secretaria d'Universitats i Recerca (GRC2017-SGR880) (T.M.-B. and C.L.-F.), BFU2017-86471-P and PGC2018-101927-B-I00 (MINECO/FEDER, UE) (T.M.-B) and PGC2018-095931-B-100 (MCIU/AEI/FEDER, UE) (C.L.-F.). T.M-B. is also supported by a U01 MH106874 grant and Howard Hughes International Early Career and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya. S.M is funded by a Welcome Trust post-doctoral fellowship (206478/Z/17/Z). L.v.D and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (grant MR/P007597/1) and the BBSRC (equipment grant BB/R01356X/1). Full metagenomic reads are available at NCBI under BioProject ID PRJEB32319.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104209.

References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19 (9), 1655–1664.
Ashbee, R., 2010. Pathogenic yeasts. In: Bignell, E. (Ed.), The Yeast Handbook. Springer-Verlag, Berlin Heidelberg, pp. 209–230.
Bayon, H.P., 1945. The medical career of Jean-Paul Marat. Proc R Soc Med. 39 (1), 39–44.
Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., Ramachandran, S., 2016. Pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 32 (18), 2817–2823.
Breitwieser, F.P., Baker, D.N., Salzberg, S.L., 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 19 (1), 198.
Broad Institute. Picard. http://broadinstitute.github.io/picard/
Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., et al., 2018. Single-tube library preparation for degraded DNA. Johnston S, editor. Methods Ecol Evol. 9 (2), 410–419.
Cohen, J.H.L., Cohen, E.L., 1958. Doctor Marat and his skin. Med. Hist. 2 (4), 281–286.
Coto-Segura, C., Coto-Segura, P., Santos-Juanes, J., 2011. The skin of a revolutionary. JAMA Dermatology. 147 (5), 539.
Dale, P.M., 1952. In: Norman (Ed.), Medical Biographies. The Ailments of Thirty-Three Famous Persons. Press, University of Oklahoma.
de-Dios, T., van Dorp, L., Gelabert, P., Carøe, C., Sandoval-Velasco, M., Fregel, R., et al., 2019. Genetic affinities of an eradicated European *Plasmodium falciparum* strain. Microb Genomics. 5 (9) (mgen000289).
Devault, A.M., Golding, G.B., Waglechner, N., Enk, J.M., Kuch, M., Tien, J.H., et al., 2014. Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849. N. Engl. J. Med. 370 (4), 334–340.
Dotz, W., 1979. Jean Paul Marat. His life, cutaneous disease, death, and depiction by Jacques Louis David. Am. J. Dermatopathol. 1 (3), 247–250.
Dreno, B., Martin, R., Moyal, D., Henley, J.B., Khammari, A., Seite, S., 2017. Skin microbiome and acne vulgaris: Staphylococcus, a new actor in acne. Exp. Dermatol. 26 (9), 798–803.
García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., et al., 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 28 (20), 2678–2679.
Gelabert, P., Sandoval-Velasco, M., Olalde, I., Fregel, R., Rieux, A., Escosa, R., et al., 2016. Mitochondrial DNA from the eradicated European plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. Proc. Natl. Acad. Sci. U. S. A. 113 (41), 11495–11500.
Gomes, A., van Oosten, M., Bijker, K.L.B., Boiten, K.E., Salomon, E.N., Rosema, S., et al., 2018. Sonication of heart valves detects more bacteria in infective endocarditis. Sci. Rep. 8 (1), 12967.
Gupta, A.K., Batra, R., Bluhm, R., Boekhout, T., Dawson, T.L., 2004. Skin diseases associated with Malassezia species. J. Am. Acad. Dermatol. 51 (5), 785–798.
Jelinek, J.E., 1979. Jean-Paul Marat. The differential diagnosis of his skin disease. Am. J. Dermatopathol. 1 (3), 251–252.
Jónsson, H., Ginolhac, A.A., Schubert, M., Johnson, P.L.F.F., Orlando, L., Jonsson, H., et al., 2013. mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. In: Bioinformatics. England, England, pp. 1682–1684.
King, T.E., Fortes, G.G., Balaresque, P., Thomas, M.G., Balding, D., Delser, P.M., et al., 2014. Identification of the remains of King Richard III. Nat. Commun. 5, 5631.
Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: analysis of next generation sequencing data. BMC Bioinformatics. 15 (1), 356.
Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35 (21), 4453–4455.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with bowtie 2. Nat. Methods 9 (4), 357–359.
Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., et al., 2018. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. Mol. Ecol. 27 (1), 182–195.
Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al., 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 513 (7518), 409–413.
Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., et al., 2016. Genomic insights into the origin of farming in the ancient near east. Nature. 536 (7617), 419–424.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 25 (14), 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al., 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25 (16), 2078–2079.
Lomholt, H.B., Kilian, M., 2010. Population genetic analysis of Propionibacterium acnes identifies a subpopulation and epidemic clones associated with acne. PLoS One 5 (8), e12277.
Marciniak, S., Prowse, T.L., Herring, D.A., Klunk, J., Kuch, M., Duggan, A.T., et al., 2016. Plasmodium falciparum malaria in 1st–2nd century CE southern Italy. Curr. Biol. 26 (23), R1220–R1222.
Margaryan, A., Hansen, H.B., Rasmussen, S., Sikora, M., Moiseyev, V., Khoklov, A., et al., 2018. Ancient pathogen DNA in human teeth and petrous bones. Ecol Evol. 8 (6), 3534–3542.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17 (1), 10.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al., 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (9), 254–260.
Miller, S., Naccache, S.N., Samayoa, E., Messacar, K., Arevalo, S., Federman, S., et al., 2019. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. Genome Res. 29 (5), 831–842.
Mollerup, S., Friis-Nielsen, J., Vinner, L., Hansen, T.A., Richter, S.R., Fridholm, H., et al., 2016. Propionibacterium acnes: disease-causing agent or common contaminant? Detection in diverse patient samples by next- generation sequencing. Burnham C-AD, editor. J Clin Microbiol. 54 (4), 980–987.
Morfopoulou, S., Plagnol, V., 2015. Bayesian mixture analysis for metagenomic community profiling. Bioinformatics. 31 (18), 2930–2938.
Mühlemann, B., Jones, T.C., Damgaard P de, B., Allentoft, M.E., Shevnina, I., Logvin, A., et al., 2018. Ancient hepatitis B viruses from the bronze age to the medieval period. Nature. 557 (7705), 418–423.
Nazipi, S., Stodkilde-Jorgensen, K., Scavenius, C., Bruggemann, H., 2017. The skin bacterium Propionibacterium acnes employs two variants of hyaluronate Lyase with distinct properties. Microorganisms. 5 (3), 57.
Olalde, I., Sánchez-Quinto, F., Datta, D., Marigorta, U.M., Chiang, C.W.K., Rodríguez, J.A., et al., 2014. Genomic analysis of the blood attributed to Louis XVI (1754–1793), king of France. Sci. Rep. 4, 4666.
Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2 (12), e190.
Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al., 2012. Ancient Admixture in human history. Genetics. 192 (3), 1065–1093.
Pierre, L., 2015. JVarkit: java-based utilities for bioinformatics. Figshare. https://github.com/lindenb/jvarkit.
Platsidaki, E., Dessinioti, C., 2018. Recent advances in understanding *Propionibacterium acnes* (Cutibacterium acnes) in acne. F1000Research 7 (F1000 Faculty Rev-1953).
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3), 559–575.
Rascovan, N., Sjögren, K.-G., Kristiansen, K., Nielsen, R., Willerslev, E., Desnues, C., et al., 2019. Emergence and spread of basal lineages of Yersinia pestis during the Neolithic

decline. Cell. 176 (1), 295–305.e10.
Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., et al., 2011. An aboriginal Australian genome reveals separate human dispersions into Asia. Science (80-). 334, 94–98.
Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., et al., 2015. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell. 163 (3), 571–582.
Renaud, G., Slon, V., Duggan, A.T., Kelso, J., 2015. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biol. 16 (1), 1–18.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., et al., 2011. Integrative genomics viewer. Nat. Biotechnol. 29 (1), 24–26.
Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., et al., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12 (1), 87.
Saunders, C.W., Scheynius, A., Heitman, J., 2012. Malassezia Fungi are specialized to live on skin and associated with dandruff, eczema, and other skin diseases. PLoS Pathog. 8 (6), e1002701.
Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27 (6), 863–864.
Scholz, C.F.P., Kilian, M., 2016. The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. Nov., Cutibacterium gen. Nov. and Pseudopropionibacterium gen. Nov. Int. J. Syst. Evol. Microbiol. 66 (11), 4422–4432.
Schubert, M., Lindgreen, S., Orlando, L., 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 9, 88.
Skoglund, P., Storå, J., Götherström, A., Jakobsson, M., 2013. Accurate sex identification of ancient human remains using DNA shotgun sequencing. J. Archaeol. Sci. 40 (12), 4477–4482.
Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., Pääbo, S., Krause, J., et al., 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. U. S. A. 111 (6), 2229–2234.
Taron, U.H., Lell, M., Barlow, A., Paijmans, J.L.A., 2018. Testing of alignment parameters for ancient samples: evaluating and optimizing mapping parameters for ancient samples using the TAPAS tool. Genes (Basel) 9 (3).
Teasdale, M.D., van Doorn, N.L., Fiddyment, S., Webb, C.C., O'Connor, T., Hofreiter, M., et al., 2015a. Paging through history: parchment as a reservoir of ancient DNA for next generation sequencing. Philos Trans R Soc B Biol Sci. 370 (1660), 20130379.
Teasdale, M., Doorn, N., Fiddyment, S., Webb, C., O'Connor, T., Hofreiter, M., et al., 2015b. Paging through history: parchment as a reservoir of ancient DNA for next generation sequencing. Philos Trans R Soc B Biol Sci. 370.
van Oven, M., 2015. PhyloTree build 17: growing the human mitochondrial DNA tree. Forensic Sci Int Genet Suppl Ser. 5, e392–e394.
Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York.
Wilson, M.R., Sample, H.A., Zorn, K.C., Arevalo, S., Yu, G., Neuhaus, J., et al., 2019. Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. N. Engl. J. Med. 380 (24), 2327–2340.

Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743-1793)

Toni de-Dios, Lucy van Dorp, Philippe Charlier, Sofia Morfopoulou, Esther Lizano, Celine Bon, Corinne Le Bitouzé, Marina Álvarez-Estapé, Tomas Marquès-Bonet, François Balloux, Carles Lalueza-Fox

Supplementary Tables

Supplementary Table S1: Mapping metrics of human DNA reads obtained from Marat's blood stain against the human mitochondrial genome (rCRS) and whole human genome (hg19). Unmapped reads following deduplication were taken forward for quality control and filtering (**Table S2**).

Mapping Statistics	Sequenced reads	Mapped reads	RM duplicates	Quality 30 reads	% duplication	Reads with damage	Coverage Final	Mapped Bases
mtDNA	568,623,176	57,654	3,485	3472	93.96	915	4.04X	66,911
Nuclear	568,623,176	74,244,610	4,686,746	4,461,919	93.69	1,245,497	0.03X	90,301,707

Supplementary Table S2 (external Excel document): Human mtDNA haplogroup assignments. Alleles and depth are characterised for describing haplogroups positions. Positions with more than two folds of allele depth are highlighted and considered as likely to be part of a defined haplogroup. The majority haplogroup is estimated using the number of likely positions. Contamination estimates are both calculated using schmutzi and from discordant positions in the haplogroups.

Table S3: f4 statistics of the form *f4*(Mbuti, Marat; X. Y) where X and Y are tested for combinations of possible ancestral sources: Sardinian, French, Basque, French, Italian_North, Basque. Positive *f4* values indicate a relatively higher affinity of Marat to Y relative to X. Negative *f4* values indicate a relatively higher affinity of Marat to X relative to Y. Z-scores following block jack-knife resampling are provided with a value >|2| considered statistically significant (highlighted in red).

Combination				*f4*	Z Score
Mbuti	Marat	Sardinian	French	-0.000017	-0.061
Mbuti	Marat	French	Sardinian	0.000017	0.061
Mbuti	Marat	English	Basque	-0.000049	-0.127
Mbuti	Marat	Basque	English	0.000049	0.127
Mbuti	Marat	French	Italian_North	-0.000043	-0.159
Mbuti	Marat	Italian_North	French	0.000043	0.159
Mbuti	Marat	Sardinian	Italian_North	-0.000061	-0.189
Mbuti	Marat	Italian_North	Sardinian	0.000061	0.189
Mbuti	Marat	English	Sardinian	-0.000297	-0.754
Mbuti	Marat	Sardinian	English	0.000297	0.754
Mbuti	Marat	Basque	Sardinian	-0.000247	-0.799
Mbuti	Marat	Sardinian	Basque	0.000247	0.799
Mbuti	Marat	English	Italian_North	-0.000357	-0.928
Mbuti	Marat	Italian_North	English	0.000357	0.928
Mbuti	Marat	Basque	Italian_North	-0.000308	-0.969
Mbuti	Marat	Italian_North	Basque	0.000308	0.969
Mbuti	Marat	English	French	-0.000314	-0.972
Mbuti	Marat	French	English	0.000314	0.972
Mbuti	Marat	Basque	French	-0.000265	-1.077
Mbuti	Marat	French	Basque	0.000265	1.077
Mbuti	Marat	Italian_North	Spanish	-0.000713	-2.697
Mbuti	Marat	Spanish	Italian_North	0.000713	2.697
Mbuti	Marat	Sardinian	Spanish	-0.000774	-2.888
Mbuti	Marat	Spanish	Sardinian	0.000774	2.888
Mbuti	Marat	English	Spanish	-0.001071	-3.094
Mbuti	Marat	Spanish	English	0.001071	3.094
Mbuti	Marat	Basque	Spanish	-0.001021	-3.986
Mbuti	Marat	Spanish	Basque	0.001021	3.986
Mbuti	Marat	French	Spanish	-0.000757	-4.019
Mbuti	Marat	Spanish	French	0.000757	4.019

Supplementary Table S4: Number of DNA reads from the blood-stained sample and unstained sample at each filtering steps prior to taxonomic classification. The final read counts for KrakenUniq and metaMix are provided in blue and red respectively.

Blood stained sample			
Filtering step	**Merged into longer single end reads**	**Non-merged**	**Combined (merged and non-merged)**
Deduplicated (bbdedupe)	11,336,155	2,240,430	13,576,585
After human removal (bowtie2)	8,617,745	1,505,994	10,123,739
After rRNA removal (bowtie2)	8,578,821	1,485,366	10,064,187
After QC (prinseq)	8,486,659	1,444,856	9,931,515
After human removal (megaBLAST)	8,370,312	1,437,288	9,807,600
After rRNA removal (megaBLAST)	8,356,389	1,432,558	9,788,947
Reads mapping to nucl DB (megaBLAST)	1,012,140	252,247	1,264,387
Unstained sample			
Filtering step	**Merged into longer single end reads**	**Non-merged**	**Combined (merged and non-merged)**
Deduplicated (bbdedupe)	18,539	111,432	129,971
After human removal (bowtie2)	17,593	18,550	36,143
After rRNA removal (bowtie2)	17,488	18,550	36,038
After QC (prinseq)	17,352	18,454	35,806
After human removal (megaBLAST)	17,276	18,410	35,686
After rRNA removal (megaBLAST)	17,156	18,060	35,216
Reads mapping to nucl DB (megaBLAST)	8,476	9,308	17,784

Supplementary Table S5 (external Excel document): Read classification and species assignment by metaMix of reads obtained from the swab of the unstained paper and blood-stained paper. *P*-values provide those species significantly over-represented in the bloodstained paper compared to the control.

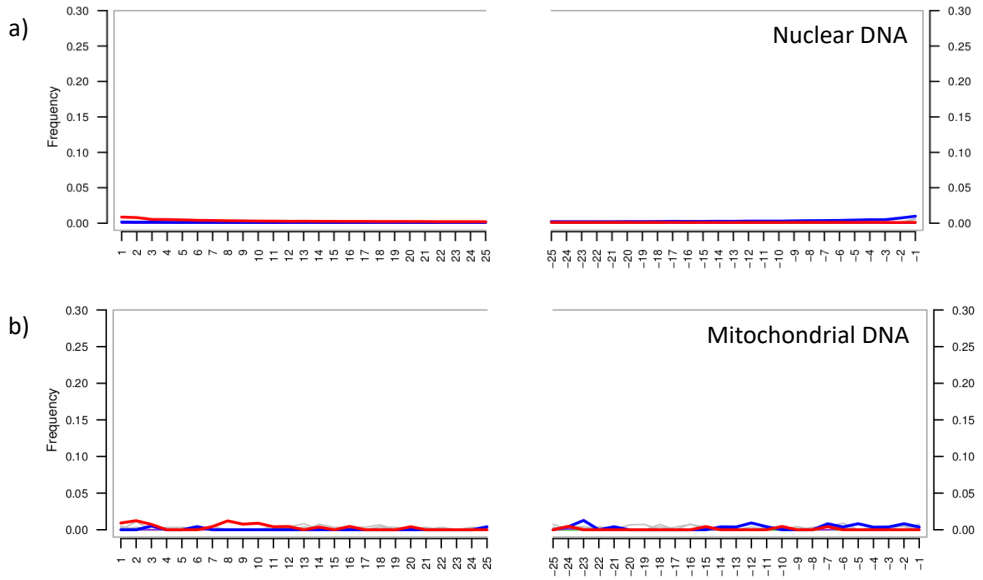
Supplementary Table S6 (external Excel document): The number of quality filtered reads mapping to candidate species using a Bowtie2 and BWA pipeline. The number of reads classified by metaMix and KrakenUniq are also provided.

Supplementary Table S7: Number of DNA reads from the two historic parchment samples at each filtering steps prior to taxonomic classification. The final read counts for metaMix are provided in red.

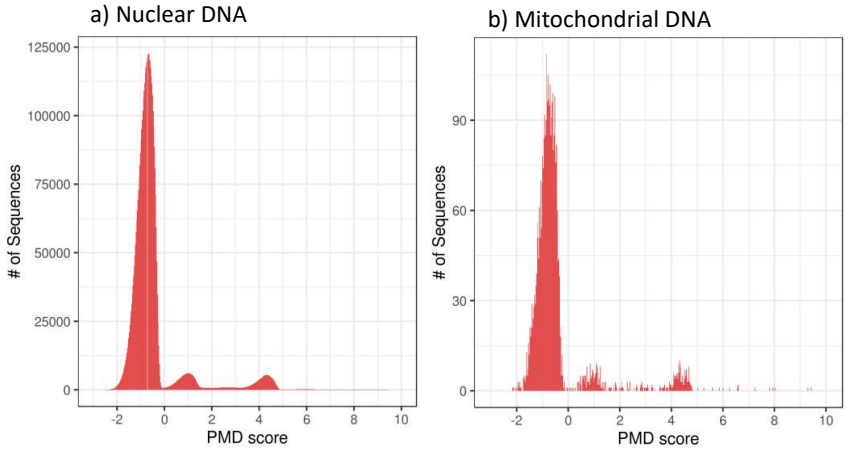
Historic Parchment Samples		
Filtering step	**ERR466100 (PA1)**	**ERR466101 (PA2)**
Raw data	17006629	31493502
After human removal (bowtie2)	16921807	31410793
After rRNA removal (bowtie2)	16916091	31403665
After ruminant removal (bowtie2)	6232580	22161979
After QC (prinseq)	6157782	21951980
After human removal (megaBLAST)	6093733	21753485
After rRNA removal (megaBLAST)	5056204	18517414
After ruminant removal (megaBLAST)	1557368	6618990
Reads mapping to nucl DB (megaBLAST)	9367	43597

Supplementary Table S8 (external Excel document): Read classification and species assignments by metaMix of reads obtained from two historic parchment samples published in *Teasdale et al.* 2015.

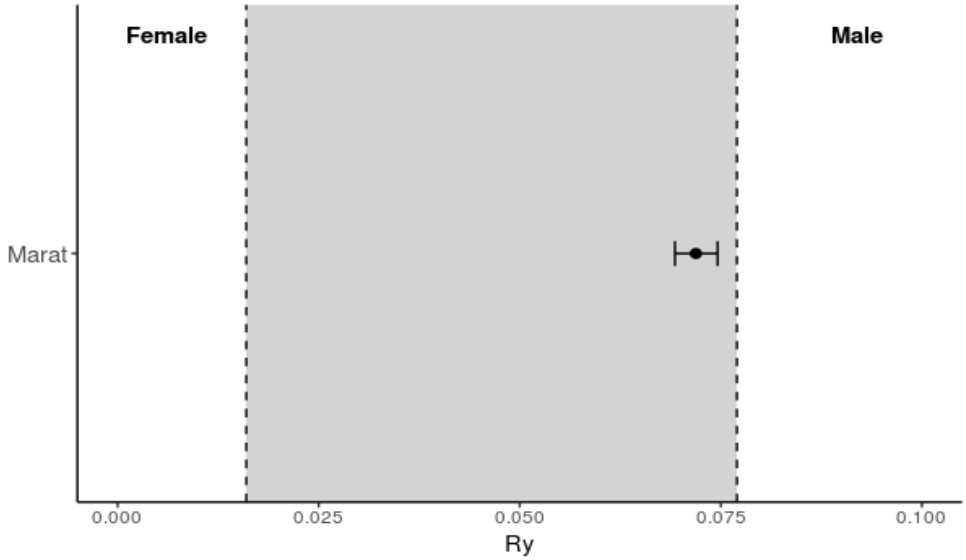
Supplementary Figures



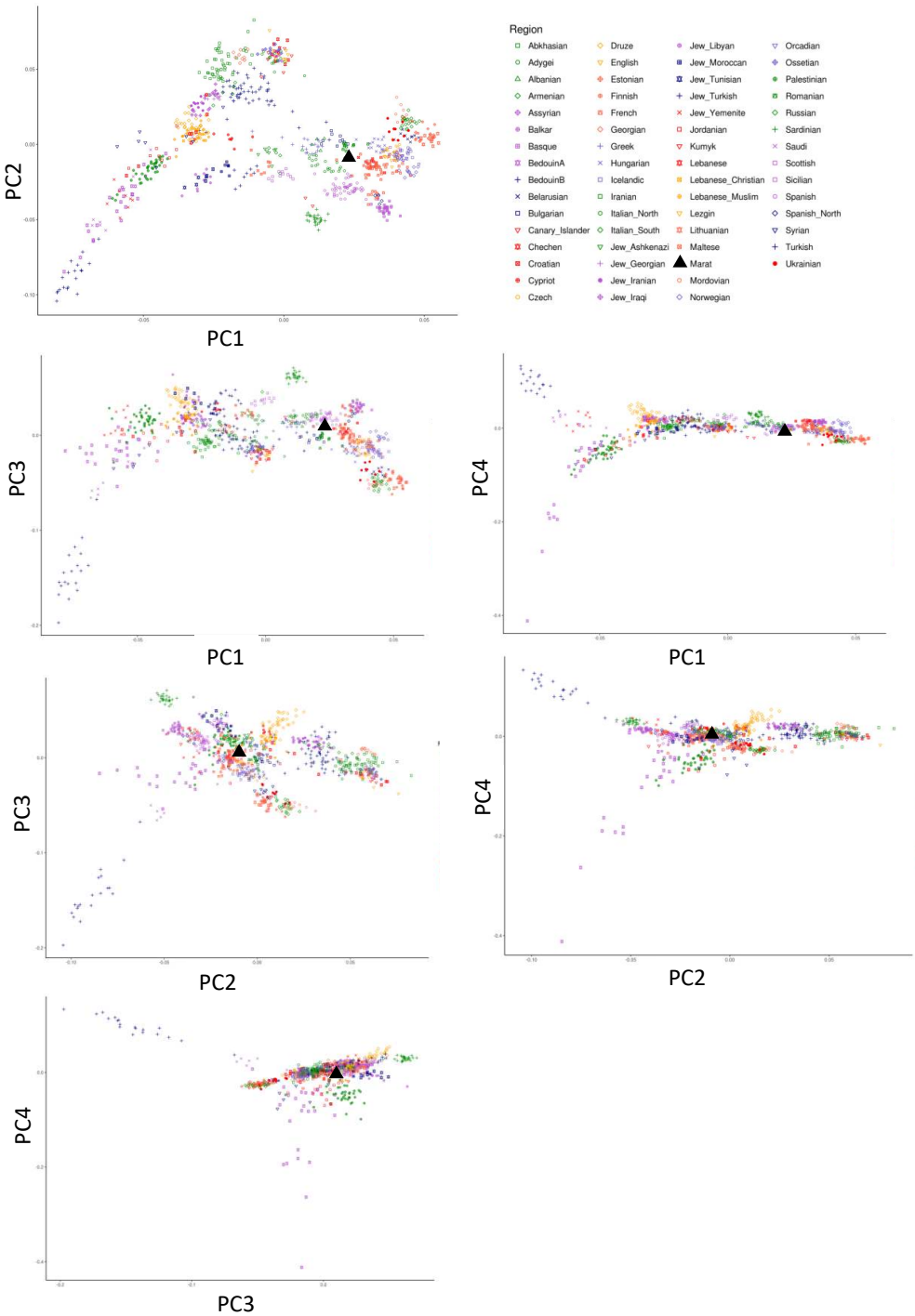
Supplementary Figure S1: a) Nucleotide misincorporation patterns at the ends of the human reads obtained from the blood stain sample (0.98% G>A at 3' and 0.86% C>T at 5'). b) Nucleotide misincorporation patterns for human mitochondrial reads only (0.38% G>A at 3' and 0.92% C>T at 5'). In both cases the red line provides the C to T substitution frequency and the blue line provides the G to A substitution frequency from 5' (left) to 3' (right). While the deamination detected is small, we observed that the post-mortem score for each read closely follows that found in a similarly old sample, a 100 year old aboriginal Australian hair sample.



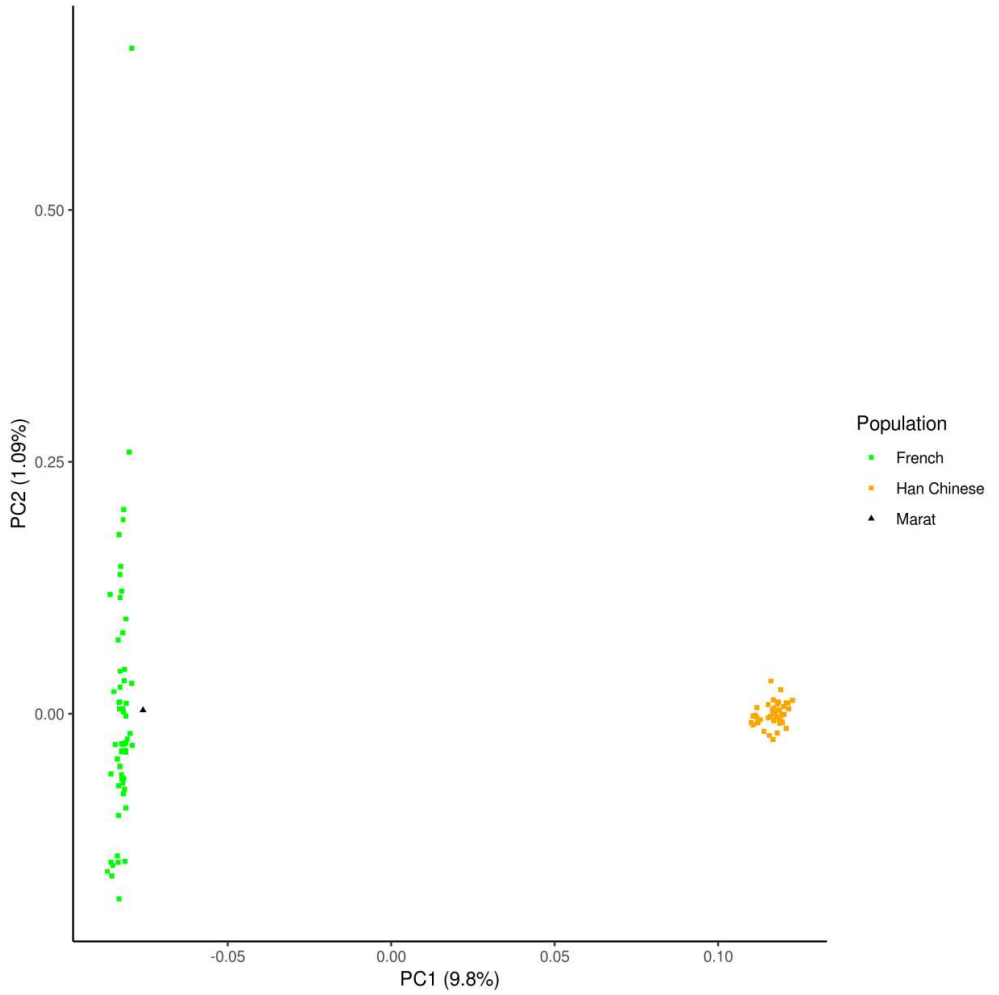
Supplementary Figure S2: Post-mortem damage score distributions for the human sequencing reads obtained from the blood-stained paper. a) provides the distribution of scores for the nuclear DNA, b) provides the distribution for the mitochondrial DNA.



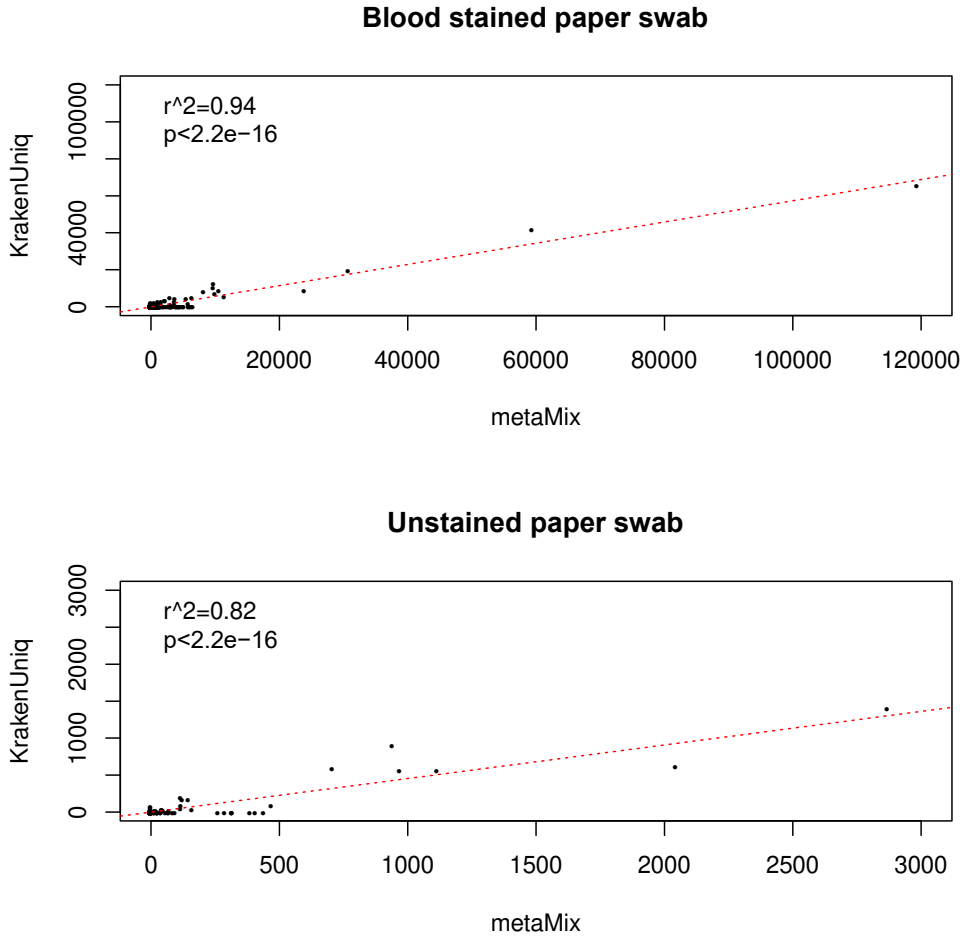
Supplementary Figure S3: Sex determination of the human reads inferred from the ratio between the number of reads aligned to the Y chromosome and the total number of reads aligned to both sex chromosomes (Ry). Error bars provide the 95% confidence value. The grey area represents the area where the sex cannot be determined. The sample is incompatible with being a female.



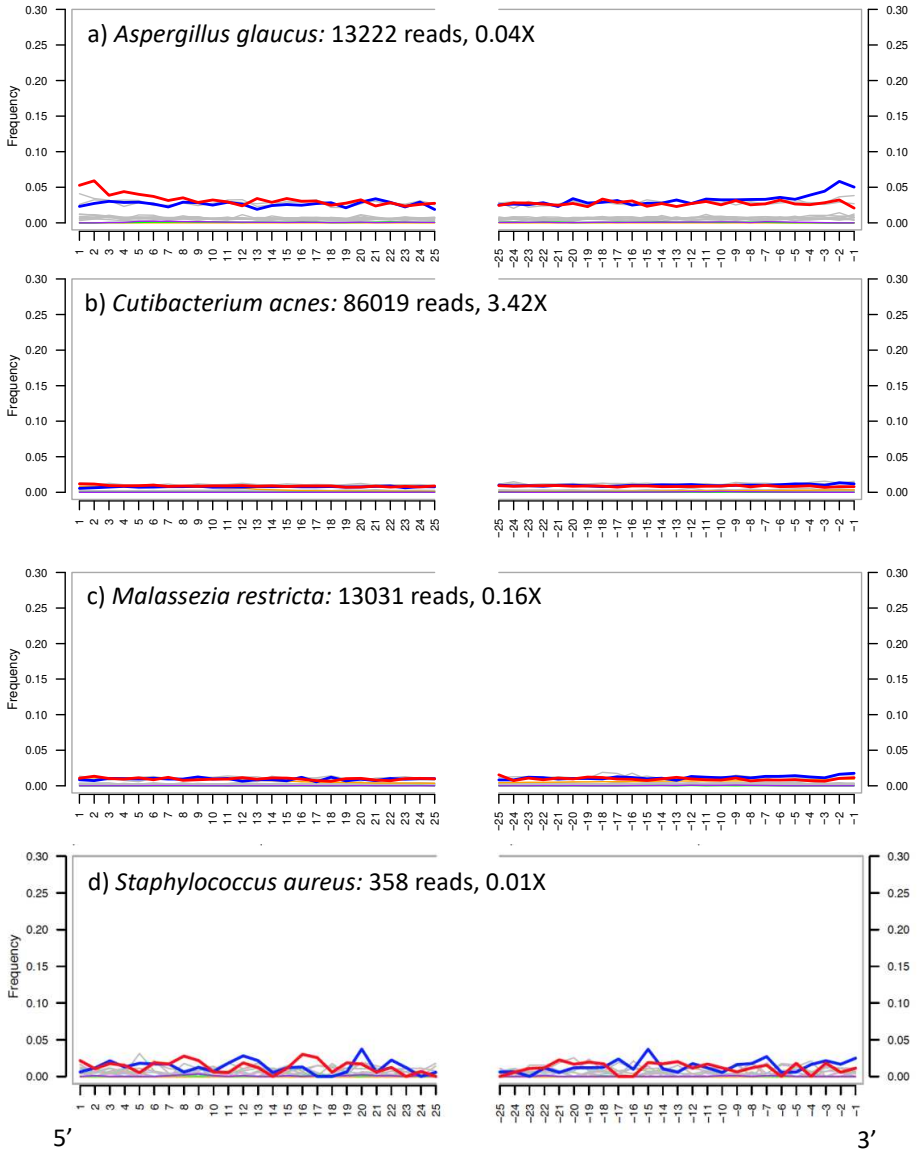
Supplementary Figure S4: Principal components plots provided for PC1-4 for West Eurasian populations coloured as per the legend (top-right). The Marat sample (triangle) is projected into PC space in each case using *lsqproject* from *SmartPCA* in *EIG v6.0.1*.



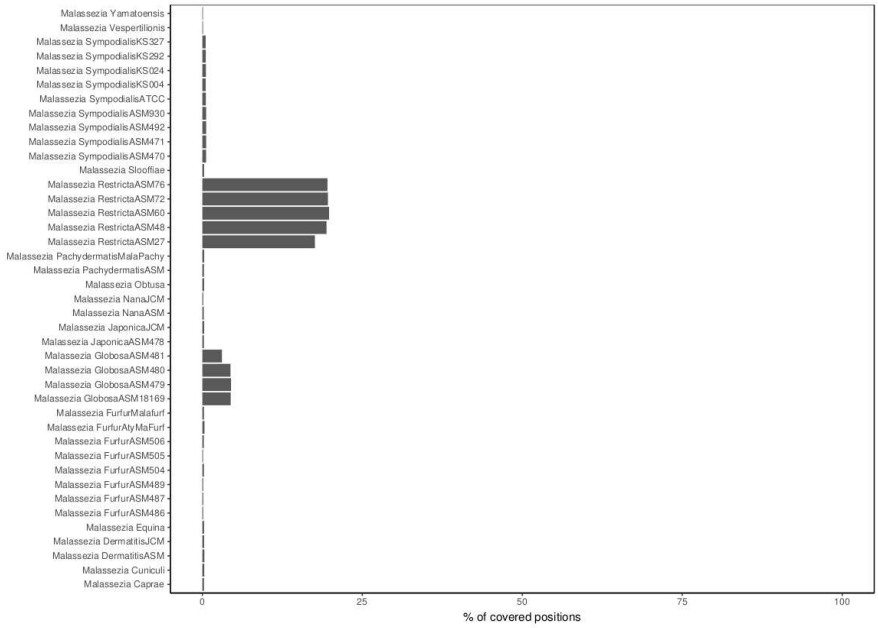
Supplementary Figure S5: Principal Component Analysis (PCA) of Marat's sample together with current French and Han Chinese individuals. Marat (triangle) is located at the lower extreme of PC1, clustering with the modern-day French.



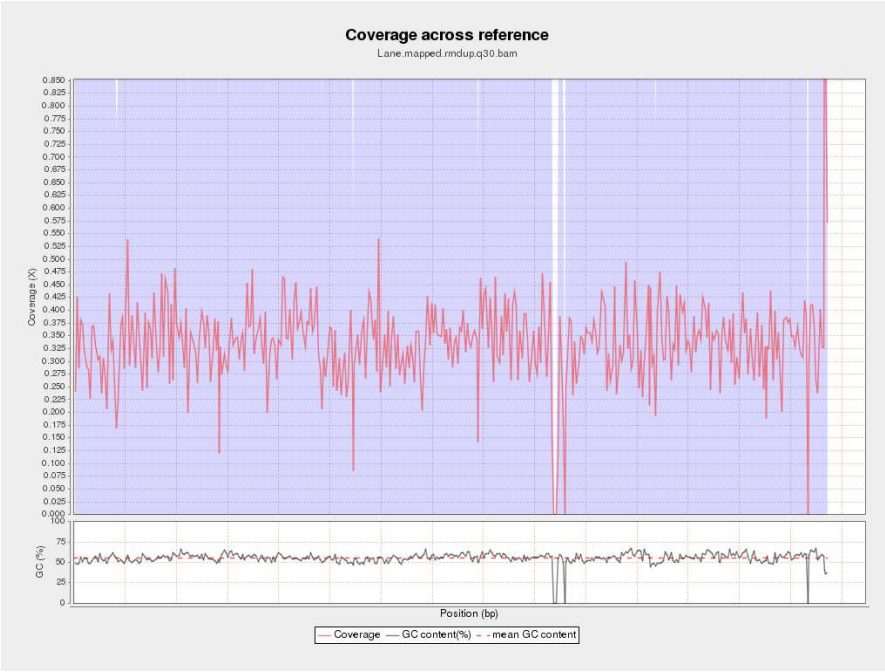
Supplementary Figure S6: Linear correlations between the number of reads assigned to individual species by the metaMix (x-axis) and KrakenUniq (y-axis) metagenomic assignment tools for the blood stain (upper panel) and unstained paper (lower panel).



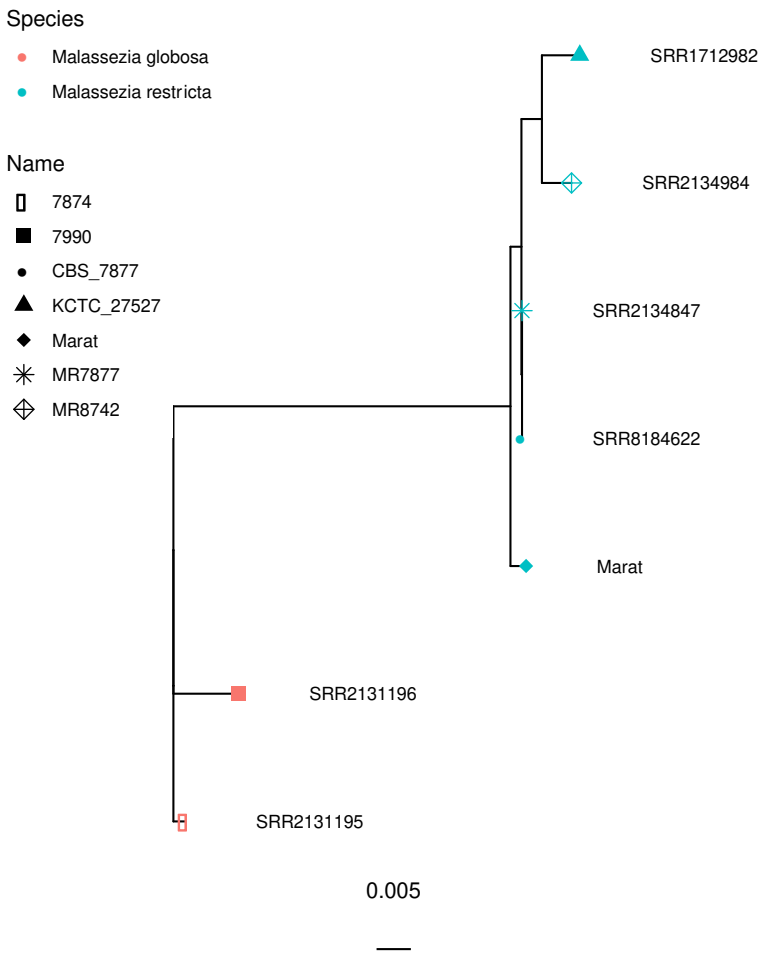
Supplementary Figure S7: From top to bottom; nucleotide misincorporation patterns at the ends of the DNA reads for *Aspergillus glaucus*, *Cutibacterium acnes*, *Malassezia restricta* and *Staphylococcus aureus*. The number of reads and coverage using the BWA ancient mapping pipeline is provided (see **Table S6**). Despite the low numbers of reads in some cases, the existence of this *post-mortem* deamination pattern suggests that these microbes are old. As in Supplementary Figure S2, the red line provides the C to T substitution frequency and the blue line provides the G to A substitution frequency from 5' (left) to 3' (right).



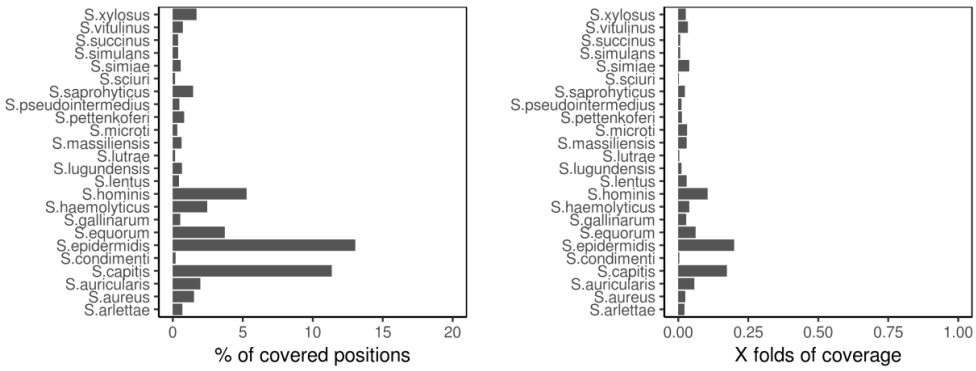
Supplementary Figure S8: Comparison of the percentage of genomic positions covered (x-axis) across different species and strains of the genera *Malassezia* following mapping independently to each species. Reads from the blood stain were mapped against a set of *Malassezia spp*. assemblies (y-axis).



Supplementary Figure S9: Average coverage and GC content of reads mapped against the full (nuclear + mtDNA) *Malassezia restricta* ASM 48 reference assembly GCF_003290485.1.

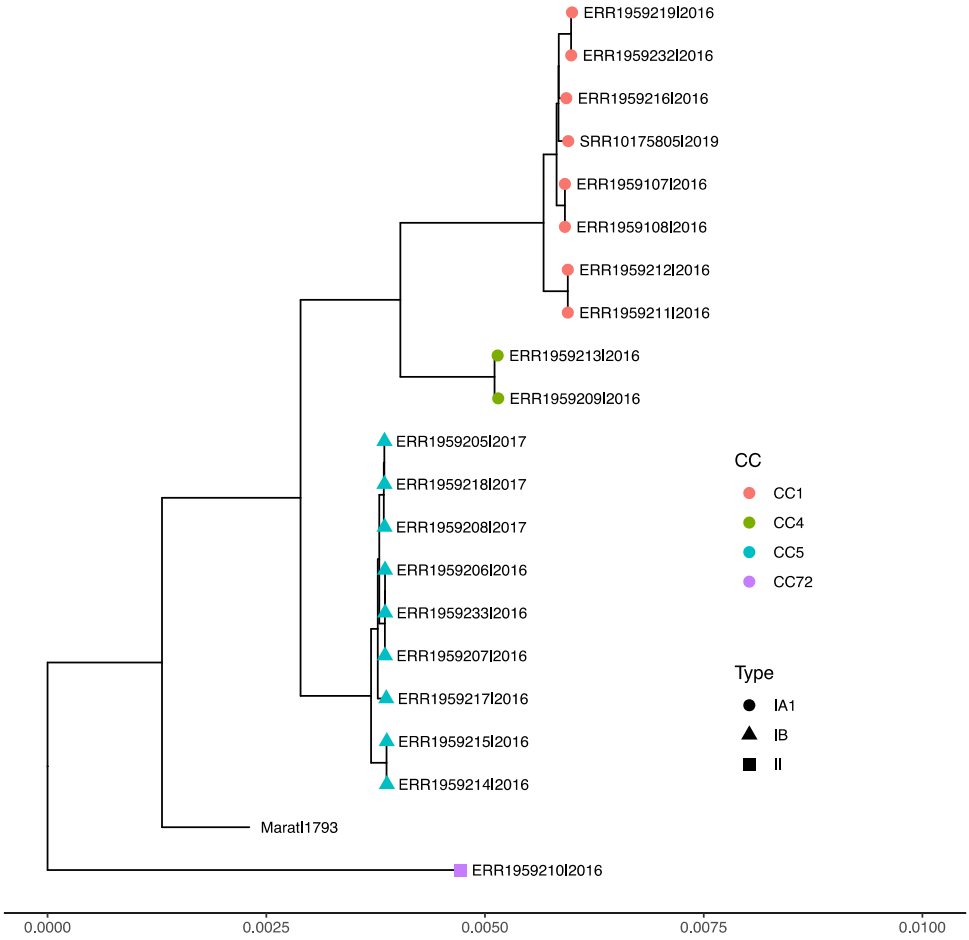


Supplementary Figure S10: Maximum likelihood (ML) phylogenetic tree of the *Malassezia restricta* mtDNA genome retrieved from Marat's blood stain (Marat) and four modern strains (accessions). The tree is rooted with two strains from the related *M. globosa* species.



Supplementary Figure S11: Comparison of the percentage of genomic positions covered (x-axis) across different species and strains of the genera *Staphylococcus*. Reads from the blood stain were mapped against a set of *Staphylococcus spp*. assemblies (y-axis).

Supplementary Figure S12: Average coverage and GC content of all reads mapped against the *Cutibacterium acnes* ASM834v1 reference assembly.



Supplementary Figure S13: Maximum likelihood (ML) phylogeny of the *Cutibacterium acnes* nuclear genome retrieved from Marat blood's stain together with all publicly available sequenced modern strains with median genome coverage >10X. The tree is rooted with *C. namnetense* as an outgroup (SRR9222443), with the outgroup branch not shown in the figure.

4.3 Ancient *Salmonella enterica* from a soldier of the 1652-siege of Barcelona (Spain) confirms historical epidemic contacts across the Atlantic.

Toni de-Dios, Pablo Carrión, Iñigo Olalde, Laia Llovera Nadal, Esther Lizano, Dídac Pàmies, Tomas Marques-Bonet, François Balloux, Lucy van Dorp, Carles Lalueza-Fox

In preparation

Ancient *Salmonella enterica* from a soldier of the 1652-siege of Barcelona (Spain) points to historical epidemic contacts across the Atlantic

Toni de-Dios[1], Pablo Carrión[1], Iñigo Olalde[1], Laia Llovera Nadal[1], Esther Lizano[1], Dídac Pàmies[2], Tomas Marques-Bonet[1,3,4,5], François Balloux[6], Lucy van Dorp[6*], Carles Lalueza-Fox[1*]

[1]Institute of Evolutionary Biology (CSIC-UPF), 08003 Barcelona, Spain

[2]Antequem. Arqueologia-Patrimoni Cultural, 08301 Mataró, Spain

[3]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain

[5]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, 08193 Cerdanyola del Vallès, Barcelona, Spain

[6]UCL Genetics Institute, University College London, London WC1E 6BT, UK

***Correspondence**: lucy.dorp.12@ucl.ac.uk; carles.lalueza.fox@gmail.com

Summary

Ancient pathogen genomics is an emerging field allowing reconstruction of the origins and spread of past epidemics. We generated and analyzed genome-wide data from two Spanish soldiers presumed to have died of Plague while besieging the city of Barcelona in 1652, during the Reapers' War. We found that one soldier's ancestry likely derived from the Basque region while the other was most related to present-day people from Sardinia, which at the time was part of the Kingdom of Spain. We could not find convincing evidence for the presence of *Yersinia pestis* DNA - the causative Plague agent. Instead, we retrieved from one individual a substantial fraction of the same *Salmonella enterica* Serovar *Paratyphi C* lineage that has been linked to paratyphoid fever in colonial period Mexico (1545-1550). Paratyphoid fever

introduced in the Americas by European colonizers has been previously suggested as a major cause for the demise of local American populations during the post-contact period. The same lineage of Paratyphi C spanning the Atlantic over at least a century adds to a growing body of evidence that Paratyphi C enteric fever was likely a far more prevalent and global disease than it is today.

Subject Areas

Genomics, ancient DNA, metagenomics, phylogenetics

Introduction

Enteric fevers are a group of similar infectious diseases caused by the bacteria *Salmonella enterica* serotypes *Typhi* (or typhoid fever, which accounts for 75% of cases) and Paratyphi A, B and C (or paratyphoid fevers)[1]. The former is globally prevalent among paratyphoid fevers[2,3], while Paratyphi B and C are relatively scarce today[1]. Typhoid and paratyphoid fevers affect up to 14 million people and cause the death of 135,900 people annually[2]. The primary route of transmission of the disease is orofecal, through ingestion of water or food contaminated by chronic carriers[4]. Although a non-life-threating disease with appropriate antibiotic treatment, without it mortality rates can reach 10–20%[5,6]. Typhoid and paratyphoid fevers are particularly prevalent in developing countries of Sub-Saharan Africa, South East Asia and South Asia, where they represent one of the leading causes of death and disability[7,8].

Although nearly absent from Europe today, recent genomics work that retrieved eight ancient Paratyphi C strains from prehistoric human remains suggest that Paratyphi C has likely been a major pathogen since Neolithic times, responsible for past outbreaks and potentially widespread epidemics[9]. Together with other *Salmonella enterica* serotypes, Paratyphi C belongs to what is defined as the Ancient Eurasian Super Branch (AESB), a cluster of phylogenetically close serotypes which infect different wild animals, livestock and humans. The diversification of AESB lineages was possibly abetted by the Neolithic transition, when changes in lifestyle and closer interactions with domesticated animals may have led to repeat exposure and infections[9].

Phylogenetic studies support the close relationship of Paratyphi C to serovar Choleraesuis, a swine pathogen which may also rarely infect humans[9–11]. This has led to the suggestion of a possible host-jump of a Choleraesuis like pathogen from pigs to humans, though the observation of host generalist strains in humans predating 4kya posits the possibility of independent host-adaptive or anthroponotic events[9,12]. Evidence for the past presence of Parathyphi C in Europe is not limited to prehistory, as the pathogen has been recovered from the 800-year-old skeleton of a young woman in Trondheim, Norway[12].

More recent evidence for the past epidemic potential of this bacterium comes in the form of the discovery of Paratyphi C in human burials from colonial México associated to the *Cocoliztli* epidemics[13]. The *Cocoliztli* (*disease* or *plague* in Nahuatl) were a series of epidemics that devastated the native populations of New Spain after the arrival of Spanish colonizers during the early 16th century[14]. These epidemics were believed to have been caused by the introduction of infectious diseases by Europeans including measles, smallpox, malaria and unknown haemorrhagic fever *i.a.*[14–16]. Detection of Paratyphi C in association with victims of the *Cocoliztli* lends support to the hypothesis that *Salmonella enterica* was a contributing agent to epidemics during recent historical times.

S. enterica has also been suggested as the causative agent of the Plague of Athens (430-426 BCE) following the amplification of two DNA fragments from individuals from the ancient *Kerameikos* mass grave, dating to that period[17]. However, subsequent phylogenetic assessment of these sequences could not authenticate them as *Salmonella enterica*[18]. This discrepancy can be partly explained by the difficulties of retrieving significant portions of ancient microbial genomes before the advent of the second-generation sequencing technologies. Therefore, although it has been suggested that Paratyphi C was a "globally" distributed pathogen, more evidence, based on genome-wide data, from different locations and time periods is needed to assess its putative role as an epidemic pathogen.

In this study we retrieved and analysed a large portion of the *Salmonella enterica* Paratyphi C genome from the tooth of a putative Spanish soldier from the siege of Barcelona that took

place between August 1651 to October 1652[19]. The accurate dating of this site will help us in understanding the spatio-temporal breadth of this disease across Europe and the Americas.

Results

We obtained 246,473,297 and 99,018,404 DNA reads from the remains of two soldiers of the Spanish army, labelled F1691-1810 and F1364-1436, respectively (Figure 1A). The end of their DNA reads displayed typical ancient DNA (aDNA) damage patterns, in a ratio consistent with the estimated age of the site (Figure S1). The molecular sex was assigned to males (Table S1). Based on the presence of different mitochondrial DNA (mtDNA) haplogroups and on heterozygosity levels of the X chromosome, we found that one of the libraries was partially contaminated (around 9% of human reads). Therefore, we only considered the non-contaminated libraries for the subsequent human population genetics analysis. The mtDNA haplogroups were assigned as U5b1f1a and H2a5a for F1691-1810 and F1364-1436, respectively (Table S1). Y chromosome haplogroups were determined with moderate confidence to R1 and R1b1a1b1 for F1691-1810 and F1364-1436, respectively. Projection of the diversity detected in both individuals as a Principal Components Analysis (PCA) plot conducted on modern West Eurasians placed F1364-1436 neighbouring the genetic makeup of present day Basques and F1691-1810 falling near present day Sardinians (Figure 1B)[20,21].

Although contemporaneous assessments suggested the soldiers died of Plague[22], metagenomic screening of the non-human DNA content from both individuals was unable to identify the presence of *Yersinia pestis* DNA reads with only 0.00003%-0.000003% of all DNA reads being assigned to the *Yersinia pestis* species (Table S2). However, X% of sequencing reads were assigned to *Salmonella* in one of the individuals F1691-1810, though only X% in the other. After mapping all reads in F1691-1810 against a comprehensive set of *Salmonella enterica* reference genomes, *Paratyphi C* was determined as the best representative based on the mean coverage, percentage of the genome covered by at least a single read and the mean read edit distance (Table S3)[10]. Mapping against the *Paratyphi C* reference chromosome alone (see Methods), we obtained 24,225 uniquely mapped DNA reads,

accounting for 30% of the reference genome at an average depth of coverage of 0.38X (Table S4).

In addition, we mapped all sequencing reads in F1691-1810 against a comprehensive dataset of bacterial plasmids. We obtained 59.02% and an average depth of coverage of 0.91X over the ~54kb Paratyphi C virulence plasmid (VirP) supporting its vertical inheritance and co-evolution with the Para C lineage[12]. The expected coverage for both the chromosome and the plasmid matched the expected theoretical value (Figure 2A and 2B) (see Methods). *Post-mortem* damage at the end of the *Salmonella* DNA reads (up to 15%) supported these sequences were authentically old (Figures 2C and 2D).

After filtering out sequences by depth, heterozygosity and the presence of low coverage transitions, we generated 1,146,808 confidently covered positions, of which 10,250 are single nucleotide polymorphisms (SNPs). Finally, we explored the depth of coverage in virulence-related regions associated to the pSPCV plasmid[10] for which we had at least X mean coverage. No apparent variation in coverage over these genes was detected, suggesting that virulence gene composition of pSPCV has been largely maintained to the present-day.

To place our strain into the wider context of *Salmonella enterica* we built a Maximum Likelihood (ML) tree over the chromosomal alignment that included 424 isolates (both ancient and modern), encompassing the AESB clade, and including *S.enterica* sr. *enteridtis* as an outgroup. This placed our 17th-century Barcelona strain within the diversity of Paratyphi C. Repeating the ML tree reconstruction using only Paratyphi C strains and Typhisuis as an outgroup, F1691-1810 was positioned in a clade (with 100% bootstrap support) falling basal to all modern Paratyphi C diversity and including the colonial Mexican Paratyphi C strains attributed the *Cocoliztli* epidemics [ref]. The only strain basal to this clade is the Medieval sample from Trondheim (Norway) dated back to 1,200 CE[12]. Under the expectation of accumulation of mutations with time, the terminal branch lengths of the ancient samples included are longer than seen in modern strains, suggesting the presence of recombination events or intrinsic features in the ancient dataset affecting branch length (Figure S2).

Due to the high genetic affinity and close temporal range of F1691-1810 individual and the ancient lineage associated to the Mexican *Cocoliztli*, we tested if the split of those branches could explain an introduction of *S. enterica* Paratyphi C into the Americas by Spanish colonizers, or if the strain from Barcelona was imported into Spain from the American continent. To create a robust alignment suitable for the application of phylogenetic tip-dating we first set out to prune from the alignment all sites in conflict with clonal evolution. To do so, we applied ClonalFrameML[23] to identify all putative homologous recombination events between sets of donors and recipients in the Paratyphi C phylogeny. We identified 338 such regions totalling 29.2Kb in length (Figure S3)(Table S5).

In addition, we enforced a further strict alignment filtering technique, identifying all homoplasic positions in the alignment given the phylogeny which may reflect recombination events undetected by ClonalFrameML or spurious sites included as a result of low coverage or DNA damage. Applying HomoplasyFinder [ref] we identify homoplasies by calculating the consistency of index, which, for each site considers the measure of the observed number of changes divided by the minimum number of changes needed to achieve a certain state at tip of a tree. We detected a further 623 positions in the alignment with a consistency index<1 and subsequently pruned them from the alignment[24] (Figure S4). The resulting filtered alignment comprised 5,180 variant sites fully covered across all of the 127 Paratyphi C strains. We generated a further maximum likelihood phylogeny using only those positions that passed the filtering criteria. The new tree was topologically congruent; however, the terminal branch length of the ancient strains was reduced, suggesting the alignment filtering procedure removed recombinant and/or spurious sites (Figure 3).

To estimate the age of the Most Recent Common Ancestor (MRCA) (the split time) we calibrated the resultant phylogenetic tree by time. To do so we first confirmed the existence of a significant temporal signal over the alignment, by computing the correlation between the root-to-tip genetic distance and the estimated time of sample collection. For our ancient sample we set this to X to reflect the mean estimate of the archaeological dating of the site. The temporal regression was significant following 10,000 randomizations of the tip sampling date (Figure 4A). We then applied BactDating to formally estimate the mutation rate over the alignment. Following convergence of the the MCMC algorithm[25] (Figure S5), we estimate

a mutation rate for this clade over the clonal frame to $1.86e^{-1}$ mutations per genome per year. The estimated mutation rate is significantly lower that those reported for analyses of modern genomes from *S. enterica* serovars[26–28], being more similar to *Y. pestis* mutation rate[29]. Based on a strict clock assumption, this rate leads to a inferred date of the split between the colonial Mexican clade and the current diversity of Paratyphi C to have occured around 280 CE (CI: 1.39 CE – 556.2 CE). A time to the MRCA of 189 BCE (CI: 586.54BCE – 206.64 CE) was estimated between the medieval Trondheim sample and the Mexican isolates (Figure 4B). Those figure postdate existing estimates of the age of the split between Typhisuis and Paratyphi C of 3655 (5147 – 2348) YBP[9].

Discussion

We performed whole genome sequencing on the remains of two Spanish soldiers who were besieging Barcelona (1651-1652) during the Reapers' War and likely died during a disease outbreak, generally believed to have been caused by the plague (*Yersinia pestis*). While we found no evidence for infection by *Y. pestis*, we recovered a substantial amount of ancient DNA from the *Salmonella enterica* Serovar *Paratyphi C* lineage previously implicated in paratyphoid fever epidemics of native Americans form Mexico during the colonial period (1545-1550). We were also able to assign a likely ancestry to the two Spanish soldiers, one to the Basque region and the other to Sardinia.

The remains from the two Spanish soldiers we sequenced are part of a far wider collection of skeletons uncovered during archaeological excavations of the Spanish army encampment in Sant Martí de Provençals. In total 576 skeletons of soldiers grouped in shallow graves have been recovered to date. Most of these contain between one and ten individuals, but two large pits with 69 and 79 individuals, respectively, have also been identified.

Little is known about the geographic origins of soldiers enlisted in the Spanish army. Our genetic analysis, which includes present-day data from human populations from Western Eurasia, the Middle East and North Africa modern populations places F1364-1436 closest to the diversity observed in present-day Basques and F1691-1810 with present-day Sardinians. What is now the Basque Country and the island of Sardinia were under dominion

of the Spanish crown during the 17th century, implying that all the regions of the Spanish realm contributed men to the war effort. In fact, it is recorded that the fall of Barcelona in 1652 was celebrated across the Spanish possessions, including Cagliari in Sardinia[30]. The analysis of further individuals could provide additional information on the heterogeneous composition of professional armies during the 17th century; for example, contemporaneous chronicles report the presence of a contingent of Irish mercenaries[31].

Based on the archaeological context of the burials, the soldiers were hastily buried - sometimes still dressed and with boots on- without signs of war injuries, and within a short period of time. Archaeologists tend to associate such burial procedures with disease outbreaks[32,33]. In contemporaneous chronicles it is mentioned that both the defenders - within the city - and the besiegers suffered a bout of "pestilence" that has been traditionally considered to be a Plague outbreak[22].

Next generation sequencing techniques now allow the recovery of historic pathogens from a variety of sources including medical collections[34,35], ancient parchments[36,37], ancient "chewing gum"[38] and human remains[9,12,13,39,40]. In the case of Plague, numerous samples have been analysed from different periods and locations[41–45]. Despite the prior suspicion that Plague was the likely cause of death of the two soldiers' remains we analysed[22], we detected no traces of *Yersinia pestis*. Instead, we recovered the partial genome of the pathogen *Salmonella enterica* Paratyphi C in one individual that may have contributed to his death. Outbreaks of enteric fever are plausible considering the sanitary conditions of a cramped military camp during a siege, as well as the surrounding marshes (that were drained in modern times[46]). Both of these factors and the presence of contaminated water may favour the rapid transmission of disease[4]. Nonetheless, it is not possible to rule out the co-existence of a *Y. pestis* outbreak given that so far we have only analysed two individuals, and the fact that Plague was a well-known disease at the time with clear symptoms[22].

Some recent studies have demonstrated the epidemic character of *S. enterica* Paratyphi C in historical times. The first time that *Salmonella* Paratyphi C DNA was found in ancient remains from the Americas was through the analysis[13] of mass burials in Mexico dating back to the mid-16th century and attributed to the *Cocoliztli* epidemics. The discovery of the pathogen in

a 12th century woman from Trondheim, Norway further supported the idea that *S. Paratyphi C* could have noticeably contributed to past pandemics in recent human history[12] despite being rare in Europe and the Americas today.

The phylogenetic placement of our bacterium sample from Barcelona within the diversity of the Mexican strains supports one of two scenarios: i) a reintroduction of *S. Paratyphi C* strains back into Europe or ii) an extensive geographic range of S. Paratyphi C prior to the colonial period. Given other cases of introduction of pathogens in the Americas during the colonial period have been recently reported aided by the analysis of ancient microbial genomics, including parvovirus and Hepatitis B virus, leprosy, syphilis and malaria[34,47–50] it is plausible the former scenario contributed to the distribution of past Parathyphi C observed. Some putative cases, such as syphilis, have been debated for decades with little progress in addressing disease introductions from the analysis of modern strains alone. Syphilis was once thought to have been introduced into Europe from the Americas; however, strains likely predating Columbian times have been recently discovered in Northern European skeletal remains[49,51].

The presence of *S.* Paratyphi C in mid-17th century Spain is particularly interesting in the context of these historical epidemics due to the fact these strains are no longer present in the current world diversity of *S.* Paratyphi C. The dating of the colonial Mexican strains - with or without the addition of the Medieval Norwegian strain - clearly predates Columbian times. Despite uncertainties associated to the dating method and to the partial data from ancient skeletal remains, these results support that *S. enterica* Paratyphi C strains were present in Europe for centuries, meaning they are conceivably a further plausible candidate for some historical pandemics of debated cause - including the one suffered by Spanish besiegers of Barcelona in 1652.

Nevertheless, it is likely that the complex pattern of trans-Atlantic connectivity detected here could only be clarified with the study of further ancient samples, both from the Americas and the Old World. However, the lack of osteological signals associated to typhoid fevers means that this survey will need to be made blindly by analysing, like we did in this study,

historical mass graves potentially attributed to other, common past pandemics such as the Plague.

Limitations of the Study

It has not been possible to retrieve the whole *Salmonella* genome which could have some impact on our temporal estimates. In the future we would like to attempt analysing additional individuals from this site and to merge all *Salmonella* DNA reads to achieve a high-coverage genome of this ancient strain.

Resource Availability

Lead contact: Carles Lalueza-Fox carles.lalueza.fox@gmail.com

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Sequences from the human genomes and *Salmonella enterica* are deposited at the European Nucleotide Archives under accession number XXXXX.

Methods

All methods can be found in the accompanying Transparent Methods Supplemental file.

Acknowledgements:

C.L.-F is supported by a PGC2018-0955931-B-100 grant (MCIU/AEI/FEDER, UE) of Spain to. TMB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement

No. 864203), BFU2017-86471-P (MINECO/FEDER, UE), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M), Howard Hughes International Early Career and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880).

Author Contributions

Conceptualization, C.L.-F. and L-v.D.; Investigation, D.P., L.L., E.L. and T.d-D.; Resources, T.M.-B. and C.L-F.; Supervision, F.B., L-v.D., T.M.-B. and C.L-F.; Writing-Original Draft, T.d-D., L-v.D., F.B. and C.L-F.; Writing -Review & Editing, all authors.

Declaration of Interests

The authors declare no competing interests.

Supplemental Information

Table S1. Mapping statistics of the human DNA reads from the two analysed individuals, separated by nuclear genome and mitochondrial DNA genome.

Table S3. Mapping stats of the 839 *Salmonella enterica* assemblies used to find the closest serovar to the F1691-1810 based on genome mapping. The top five closest assemblies correspond to Cholerasuis and Paratyphi C serovars. The closest serovar was Paratyphi C, with the highest number of mapped reads, highest mean depth of coverage, highest percentage of the total reference recovered, and lowest mean edit distance.

Table S5. Recombination events in the Paratyphi C phylogeny. The table include events exclusive of a leaf (marked using the leaf name), or events including a whole node (marked with the node number).

References

1. Cash-Goldwasser, S. & Barry, M. *CDC Yellow Book 2018: Health Information for International Travel. Clinical Infectious Diseases* vol. 66 (2018).

2. Stanaway, J. D. *et al.* The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Infect. Dis.* **19**, 369–381 (2019).

3. Ochiai, R. L. *et al.* Salmonella paratyphi A rates, Asia. *Emerg. Infect. Dis.* **11**, 1764–1766 (2005).

4. Gunn, J. S. *et al.* Salmonella chronic carriage: Epidemiology, diagnosis, and gallbladder persistence. *Trends Microbiol.* **22**, 648–655 (2014).

5. Crump, J. A., Ram, P. K., Gupta, S. K., Miller, M. A. & Mintz, E. D. Part I. Analysis of data gaps pertaining to Salmonella enterica serotype Typhi infections in low and medium human development index countries, 1984-2005. *Epidemiol. Infect.* **136**, 436–448 (2008).

6. World Health Organisation. Typhoid vaccine: WHO position paper - March 2018. *Wkly. Epidemiol. Rec.* **13**, 153–172 (2018).

7. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1211–1259 (2017).

8. Naghavi, M. *et al.* Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1151–1210 (2017).

9. Key, F. M. *et al.* Emergence of human-adapted Salmonella enterica is linked to the Neolithization process. *Nat. Ecol. Evol.* **4**, 324–333 (2020).

10. Liu, W. Q. *et al.* Salmonella paratyphi C: Genetic divergence from salmonella choleraesuis and pathogenic convergence with salmonella typhi. *PLoS One* **4**, e4510 (2009).

11. Nair, S. *et al.* Genetic markers in s. Paratyphi c reveal primary adaptation to pigs. *Microorganisms* **8**, 657 (2020).

12. Zhou, Z. *et al.* Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* **28**, 2420-2428.e10 (2018).

13. Vågene, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).

14. Malvido, E. & Viesca, C. Epidemia de cocoliztli de 1576. in *Historias* vol. 11 26–33 (1985).

15. Somolinos d'Ardois, G. La epidemia de Cocoliztli de 1545 señalada en un codice. *Trib. Medica* **15**, 85 (1970).

16. Marr, J. S. & Kiracofe, J. B. Was the Huey Cocoliztli a haemorrhagic fever? *Med. Hist.* **44**, 363–388 (2000).

17. Papagrigorakis, M. J., Yapijakis, C., Synodinos, P. N. & Baziotopoulou-Valavani, E. DNA examination of ancient dental pulp incriminates typhoid fever as a probable cause of the Plague of Athens. *Int. J. Infect. Dis.* **10**, 206–214 (2006).

18. Shapiro, B., Rambaut, A. & Gilbert, M. T. P. No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). *International Journal of Infectious Diseases* vol. 10 334–335 (2006).

19. Monguiló, E., Hernandez, J. & Molinas, R. *Intervenció arqueològica a l'espai delimitat pels carrers d'Espronceda, ronda de Sant Martí, carrer de Josep Soldevila i passeig de la Verneda (Ave Sagrera).* (2012).

20. Álvarez-Iglesias, V. *et al.* New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS One* **4**, e5112–e5112 (2009).

21. Cardoso, S. *et al.* The Expanded mtDNA Phylogeny of the Franco-Cantabrian Region Upholds the Pre-Neolithic Genetic Substrate of Basques. *PLoS One* **8**, e67835 (2013).

22. Luis, J. & Moya, B. Sociedad y peste en la Barcelona de 1651. *Manuscrits Rev. d'història Mod.* **0**, 255–282 (1990).

23. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

24. Crispell, J., Balaz, D. & Gordon, S. V. Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. *Microb. Genomics* **5**, e000245 (2019).

25. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134–e134 (2018).

26. Hawkey, J. *et al.* Evidence of microevolution of Salmonella Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. *BMC Genomics* **14**, 800 (2013).

27. Okoro, C. K. *et al.* Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nat. Genet.* **44**, 1215–1221 (2012).

28. Octavia, S., Wang, Q., Tanaka, M. M., Sintchenko, V. & Lan, R. Genomic variability of serial human isolates of salmonella enterica serovar typhimurium associated with prolonged carriage. *J. Clin. Microbiol.* **53**, 3507–3514 (2015).

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

30. Caredda, S. Un agente de la Corona hispánica en Cerdeña : Pedro Martínez Rubio (1614-1667) y la relación de las fiestas calaritanas por la rendición de Barcelona (1652). *Stud. aurea monográfica* 259–269 (2015).

31. Recio Morales, Ó. 'Una nación inclinada al ruido de las armas' La presencia irlandesa en los ejércitos españoles, 1580-1818: ¿La historia de un éxito? *Tiempos Mod.* **10**, 1–15 (2004).

32. Signoli, M. *et al.* Découverte d'un charnier de la Grande Armée en Lituanie (Vilnius, 1812). *Comptes Rendus - Palevol* **3**, 219–227 (2004).

33. Raoult, D. *et al.* Evidence for louse-transmitted diseases in soldiers of Napoleon's Grand Army in Vilnius. *J. Infect. Dis.* **193**, 112–120 (2006).

34. Van Dorp, L. *et al.* Plasmodium vivax Malaria Viewed through the Lens of an Eradicated European Strain. *Mol. Biol. Evol.* **37**, 773–785 (2020).

35. De-Dios, T. *et al.* Genetic affinities of an eradicated european plasmodium falciparum strain. *Microb. Genomics* **5**, (2019).

36. Piñar, G., Tafer, H., Schreiner, M., Miklas, H. & Sterflinger, K. Decoding the biological information contained in two ancient Slavonic parchment codices: an added historical value. *Environ. Microbiol.* **22**, 3218–3233 (2020).

37. de-Dios, T. *et al.* Metagenomic analysis of a blood stain from the French revolutionary Jean-Paul Marat (1743–1793). *Infect. Genet. Evol.* **80**, (2020).

38. Jensen, T. Z. T. *et al.* A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat. Commun.* **10**, 5520 (2019).

39. Mühlemann, B. *et al.* Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* **557**, 418–423 (2018).

40. Maixner, F. *et al.* The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**, 162–165 (2016).

41. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* **176**, 295-305.e10 (2019).

42. Bos, K. I. *et al.* Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *Elife* **5**, e12994 (2016).

43. Andrades Valtueña, A. *et al.* The Stone Age Plague and Its Persistence in Eurasia. *Curr. Biol.* **27**, 3683-3691.e8 (2017).

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

45. Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506–510 (2011).

46. Martín-Vide, J. P. Restauración del río Besòs en Barcelona. Historia y lecciones. *Ribagua* **2**, 51–60 (2015).

47. Guzmán-Solís, A. *et al.* Ancient viral genomes reveal introduction of HBV and B19V to Mexico during the transatlantic slave trade. *bioRxiv* 2020.06.05.137083 (2020) doi:10.1101/2020.06.05.137083.

48. Schuenemann, V. J. *et al.* Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science* **341**, 179–183 (2013).

49. Majander, K. *et al.* Ancient Bacterial Genomes Reveal a High Diversity of

Treponema pallidum Strains in Early Modern Europe. *Curr. Biol.* **30**, 3788-3803.e10 (2020).

50. Yalcindag, E. *et al.* Multiple independent introductions of Plasmodium falciparum in South America. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 511–516 (2012).

51. Schuenemann, V. J. *et al.* Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Negl. Trop. Dis.* **12**, e0006447–e0006447 (2018).

Figure Captions

Figure 1: Genetic affinities of the soldiers analysed. (A) Image of some of the soldiers' remains found in the site. (B) Genetic data for both individuals (see legend at top right) are projected onto the two main principal components (PC) defined by 1431 present-day West-Eurasian individuals genotyped by Human Origins array (grey-points). Modern populations more closely related to the siege of Barcelona soldiers are highlighted (present-day Basques (green); Spanish (yellow); Sardinians (cyan)).

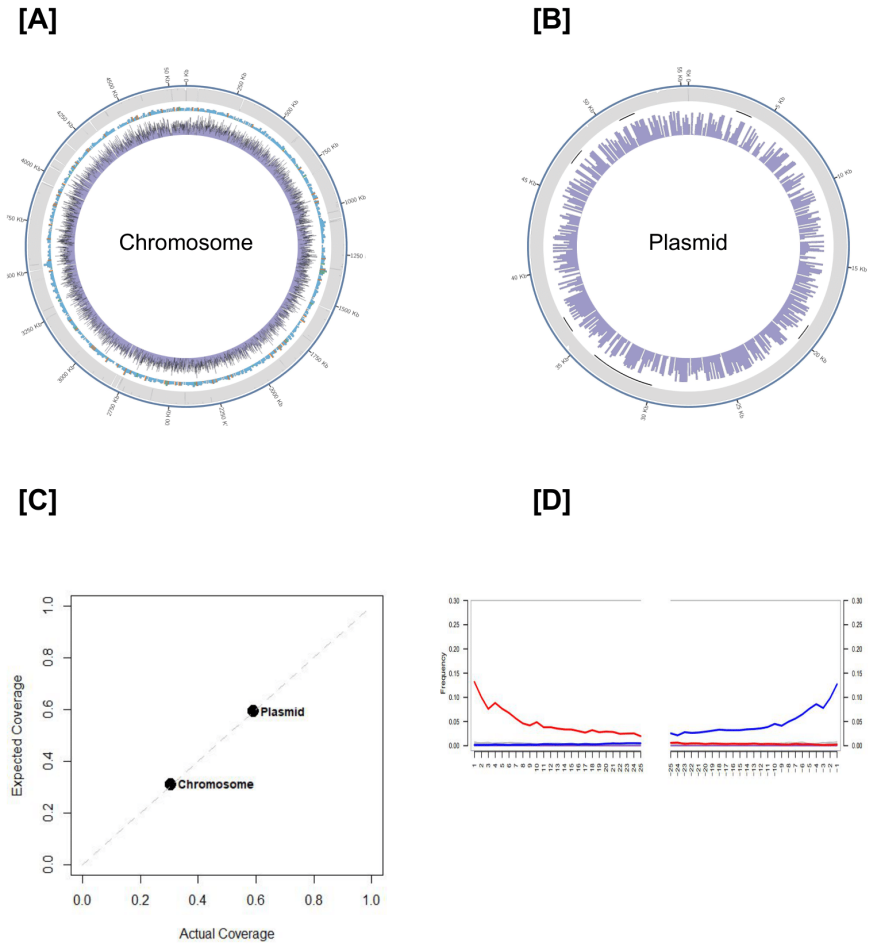


Figure 2: Characteristics of the *S. enterica* Paratyphi C recovered sequences. Coverage plots of the Paratyphi C chromosome (A) and plasmid (B); the outer blue circle provides the genomic position, the second outermost circle in grey provides the reference mappability, the third most outermost ring represents the presence of genes (blue), pseudogenes (orange) and RNAs (green) along the reference, and the innermost purple circle provides the mean depth (binned) of coverage. (C) Comparison between the calculated and expected coverage and the actual observed coverage in both plasmid and chromosome. (D) Damage patterns at both ends of the DNA reads mapped against Paratyphi C reference genome.

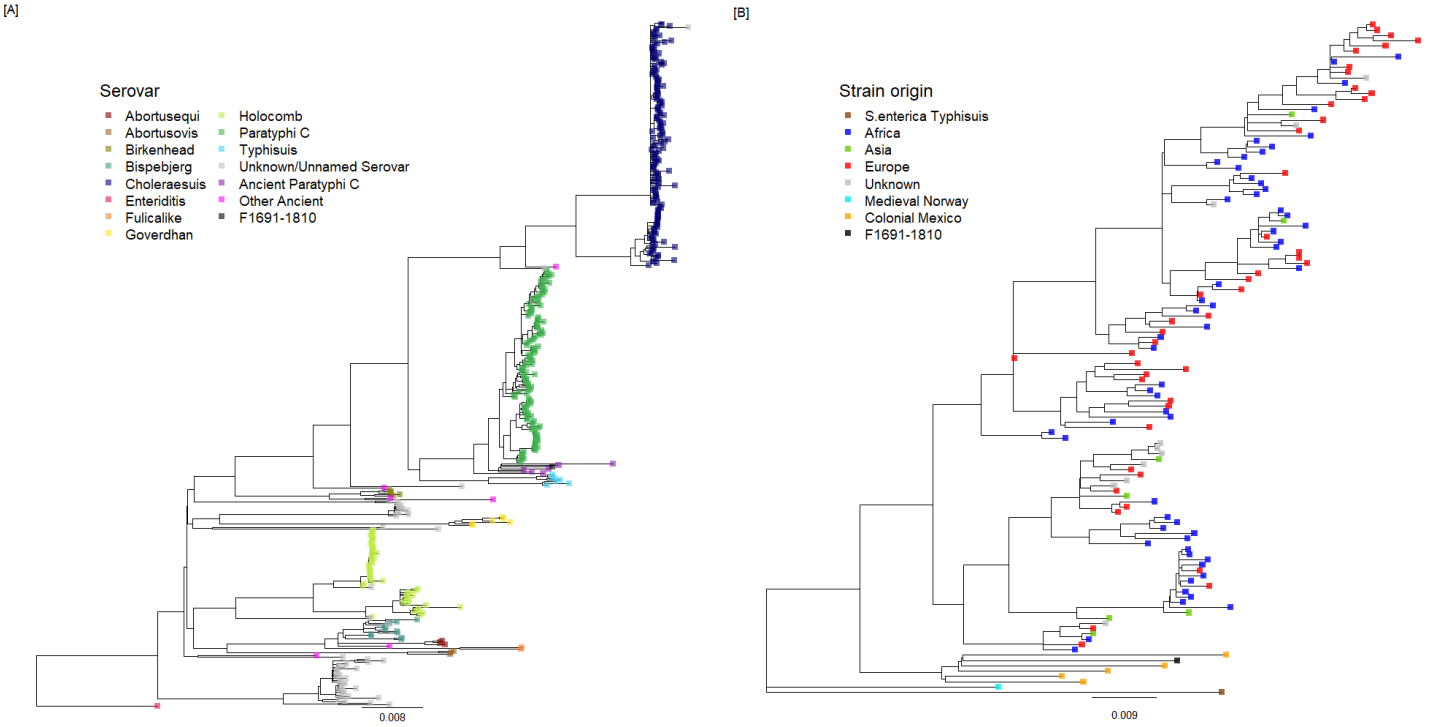


Figure 3: Maximum likelihood phylogeny of *S. enterica* falling within the Ancient Eurasian Super Branch (AESB). (A) ML tree of 424 strains representing the ancient and modern diversity of the AESB. The tree is rooted using *S. enterica* ser. Enteriditis. Serovars are coloured at the branch tip. Recombination events and homoplasies have been removed from the alignment. (B) Maximum likelihood phylogeny of the Paratyphi C clade including 127 Paratyphi C modern and ancient strains. Tip points are coloured according to their origin. Recombination events and homoplasies have been removed from the alignment.

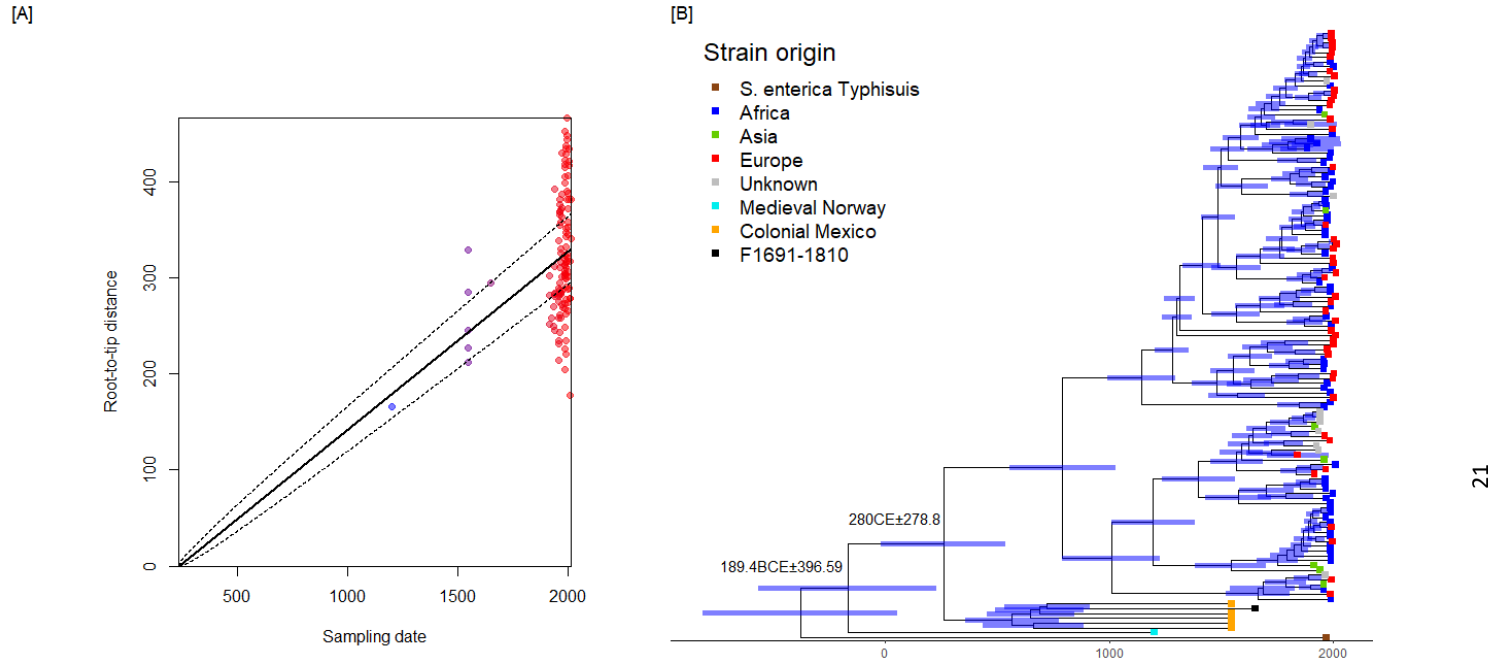


Figure 4: Temporal signal in the Paratyphi C clade. (A) Regression of the root to tip distance and collection date of the samples present in the tree, with 10,000 permutations. R^2=0.11, p<1.00e^{-4}. Dotted lines represent the 95% CI (B). ML tree of the curated Paratyphi C dataset. X axes displays the time of collection date. Blue lines at each node represent the 95% CI of the splitting date. The split of the Medieval Norway and rest of Paratyphi C diversity is dated in 189BCE. The split of the Colonial Mexico/Spain clade and current diversity of Paratyphi C is dated back to the 280CE.

Supplementary Materials

Ancient *Salmonella enterica* from a soldier of the 1652-siege of Barcelona (Spain) confirms historical epidemic contacts across the Atlantic

Toni de-Dios, Pablo Carrión, Iñigo Olalde, Laia Llovera Nadal, Esther Lizano, Dídac Pàmies, Tomas Marques-Bonet, François Balloux, Lucy van Dorp, Carles Lalueza-Fox

Transparent Methods

The site

A recently excavated archaeological site located in La Sagrera -currently a north-eastern quarter of the city of Barcelona- dates from 1651-1652 CE, during the conflict known as the Reapers' War (*Guerra dels Segadors,* 1640-1659), in the context of the Thirty Years' War[1]. During 1651, the city was under siege by Spanish forces commanded by Juan José de Austria. Despite the efforts of the local garrison and the arrival of French reinforcements, the city capitulated in the spring of 1652. The site consists of multiple mass graves containing mostly males (118 out of 140 studied by physical anthropologists so far), with ages ranging from 16 to 40 years and signals in their skeletons of having undertaken intense physical activities. The analysed individuals come from the burial F1691 (individual 1810) and F1364 (individual 1436). An upper canine was extracted from each of the skulls and used for genetic analyses.

DNA extraction and library preparation

All DNA extraction and initial library preparation steps (prior to amplification) were performed in a dedicated ancient DNA laboratory, physically isolated from the laboratory used for post-PCR analyses. Strict protocols were followed to minimize the amount of human DNA in the ancient DNA laboratory, including the existence of positive air pressure in the clean rooms, the wearing of a full body suit, sleeves, shoe covers, clean

shoes, facemask, hair net and double gloving. All lab surfaces, consumables, disposables, tools and instruments were wiped with bleach and ethanol, and UV irradiated before and after use.

First, the teeth samples were UV irradiated (245 nm) for 10 minutes and the outermost surface of the teeth was scrapped off with a drill engraving cutter, followed by another UV irradiation for 10 more minutes, in order to exclude the surface DNA contamination. Second, approximately 30 mg of tooth cementum were drilled into a fine powder by a Dremel drilling machine at low speed (5000 rpm).

DNA extraction from tooth powder was performed following the method proposed by Dabney et al. 2013[2]. The teeth-powder samples, including an extraction blank, were added to 1ml of extraction buffer (final concentrations: 0.45 M EDTA, 0.25 mg/mL Proteinase K, pH 8.0), resuspended by vortexing and incubated at 37 °C overnight (24h) on rotation (750-900 rpm). Remaining tooth powder was then pelleted by centrifugation in a bench-top centrifuge for 2 min at maximum speed (16,100 × g). The supernatant was added to 10mL of binding buffer (final concentrations: 5 M guanidine hydrochloride, 40% (vol/vol) isopropanol, 0.05% Tween-20, and 90mM sodium acetate (pH 5.2)) and purified on a High Pure Extender column (Roche). DNA samples were eluted with 45µl of low EDTA TE buffer (pH 8.0).

A total of 35µl of each DNA extract were converted into Illumina sequencing libraries following the BEST protocol[3]. Each library was amplified by PCR using two uniquely barcoded primers. After index PCR, libraries were purified with a 1.5x AMPure clean (Beckman Coulter) and eluted in 25µl of low EDTA TE buffer (pH 8.0). Libraries were quantified using BioAnalyzer and sequenced by HiSeq 4000 (Illumina).

Human DNA mapping

Reads were trimmed of sequencing adapters, filtered for reads of less than 30bp and merged using *AdapterRemoval* with default parameters[4]. Clipped reads were then mapped against the human reference genome hg37/19 and the Revised Cambridge mitochondrial DNA reference sequence using *BWA aln/samse* with the seeding option disabled[5–9]. Next, duplicated reads were removed using *Picard* and only reads with a mapping quality equal or

above 30 were considered for the following analysis[10]. Mapping statistics were calculated using *SAMtools* and *Qualimap2* [9,11]. Finally, to assess the authenticity of the DNA reads, *post-mortem* associated DNA damage was estimated using *mapDamage2* and *PMDtools*[12,13].

Sex determination and uniparental markers analysis

Molecular sex was assigned using *Ry_compute*, a script designed to determine the sex of individuals sequenced at low coverage based on the ratio of reads mapping to each sex chromosome[14]. Mitochondrial haplogroups were determined using haplogrep2[15]. Y chromosome haplogroup determination was performed by manually annotating variants from the International Society of Genetic Genealogy (http://www.isogg.org) version 15.73.

Contamination estimates

Modern human contamination was estimated using two approaches. For mitochondrial contamination, we used *Schmutzi,* which calculates the modern contamination by the profile of aDNA associated deamination in the sample[16]. For the nuclear DNA contamination, and considering that both individuals are compatible with being males, we estimated the exogenous DNA contamination based on the heterozygosity of the X chromosome sites using *angsd* [17].

Human population genetics analysis

To analyse these 17th-century individuals in the context of present-day human genetic diversity, their genomic data was merged with 1,134 West-Eurasian individuals genotyped in the Human origins (HO) array [18]. Principal Component Analysis (PCA) was computed using the modern HO individuals, and the ancient samples were projected onto the first two components (PC1 and PC2) using options 'lsqproject: YES' and 'shrinkmode: YES' of *smartpca* built-in module of EIGENSOFT (v. 7.2.1) (https://www.hsph.harvard.edu/alkes-price/software/)[19,20].

Pathogens' screening and Salmonella sequences mapping

To explore the presence of relevant microbial organisms in the samples we collapsed unique reads from the human-free sequences and removed from the dataset low complexity sequences using *Prinseq*[21]. Afterwards, we applied *kraken2* to assign reads against a standard databases (bacteria, archaea, fungi, protozoa and viral)[22]. We found almost no evidence of *Yersinia pestis* DNA reads, but one of the samples indicated a significant presence of *Salmonella enterica*.

To validate this signal, and to identify the closest representative of our sample amongst a diverse set of *S. enterica*, we downloaded 839 *Salmonella* published assemblies from NCBI (as for 03/04/2020) and created a custom database. Human-free reads were then mapped against this database using the local alignment algorithm *mem* of BWA[7]. We mapped the sequences resultant of the local alignment of the free DNA reads against each *Salmonella* assembly downloaded independently using BWA's global alignment algorithm *aln*. The settings used were then optimised for mapping ancient genomes by disabling the seeding option, setting an edit distance value of 0.01 and a gap open penalty value of 2. After that, duplicated sequenced were removed using *Picard* tools and DNA reads with mapping quality >25 were retained[8,10]. The mapping statistics and presence of *post-mortem* aDNA damage were determined as described for the human sequences. Due to the low coverage of the sample and its high ratio of aDNA damage, we decided to trim 10 bases from the ends of each of the DNA reads.

We determined that the most suitable assembly to map against was the *Salmonella enterica subsp. enterica* ser. Paratyphi C reference genome[23]. The number of mapped reads, the mean coverage, the fraction of the genome recovered, and the mean edit distance are provided in Suppl. Table 3.

Coverage and mappability of Salmonella enterica Paratyphi C

We determined the mappability of the reads to the Paratyphi C reference genome by mapping *k-mers* of 40 to 100 base pairs. The coverage of the sample was determined using *bedtools* specifying windows of 1kb for the chromosome and 100 bases for the plasmid[24]. To ascertain if the coverage found adjusted the expected, we assumed a random distribution matching the actual coverage if the sample is authentic, following a previously described

approximation[25] which calculates the probability of a position being covered given the presence of *r* reads of length *l* using the following formula: $c = 1 - \prod_{i=1}^{N}(1 - \frac{l_i}{g})^{r_i}$, in which *N* are the different l_1 to l_N, read lengths with counts r_1 *to* r_N. This value must be corrected by multiplying *c* for the mappability of the reference genome, otherwise the true coverage may not match the expected. An additional scenario leading to the expected and actual coverage not matching is when the reference has a region with no coverage in the ancient sample[25].

Variant calling and phylogenetics dataset creation

Additionally, a dataset of 411 modern *Salmonella enterica* representative of the *AESB* [26–28], as well as 11 historical samples[26,29,30] was curated. For all modern samples raw read data processing was as described as for F1691-1810 except for employing the *BWA mem* algorithm with default settings instead of *aln*, the latter being advised for ancient DNA mapping pipelines. Historical samples were processed as described for our novel ancient samples. Variants were called from processed published sequences using *GATK 3.7* algorithm *UnifiedGenotyper*. For each sample, positions were filtered for minimum coverage of 10X and indels were excluded.

Maximum Likelihood tree

All positions satisfying the filtering criteria were used to create a consensus genome fasta using the reference genome as template. Those positions that were filtered out were masked in the resultant fasta file. This resulted in a phylogenetic dataset comprising 424 *Salmonella enterica* AESB and an outgroup (ERS217420).

An initial Maximum Likelihood (ML) tree was build using RAxML 8.2.4 with the nucleotide substitution model GTRCAT and a strain of *Salmonella enterica* ser. Enteriditis as an outgroup. The resultant tree was visualised using the R package ggtree[31,32]. Following this, we used ClonalFrameML to correct the subsequent tree for the presence of homologous recombination events[33]. ClonalFrameML can infer the location of a recombination event in each branch of a ML tree, which can affect to the resultant branch length. The transition / transversion ratio used ($\varkappa$) was 2.997.

After removing the recombination events present in the Paratyphi C branch we proceeded to remove all remaining homoplasies from the dataset using HomoplasyFinder[34]. A homoplasy is a substitution event which has arisen independently in different separate linages, and which can alter the tree topology, hence these may arise from undetected recombination or spurious variant calls which may appear in low coverage samples. Homoplasies are determined by their index value which ranges between 0 and 1, with values closer to zero denoting more homoplasic variants[35]. A final ML tree based on the filtered alignment of 5,180 variant sites was generated using RAxML with the parameters described above. As before, the phylogeny was visualised using ggtree.

Temporal signal exploration

In order to perform phylogenetic tip-calibration it is necessary to first confirm the presence of significant temporal signal in the alignment. A temporal regression of sampling date (years) against root-to-tip distance was performed using the R package BactDating, which additionally conducts a date randomization test of significance[36]. Samples without an assigned collection date were treated as missing. We established the most suitable model by running all models and compare their *DIC* value. We performed the analysis using the model *arc* inference model and 10,000 iterations.

Supplementary References

1. Monguiló, E., Hernandez, J. & Molinas, R. *Intervenció arqueològica a l'espai delimitat pels carrers d'Espronceda, ronda de Sant Martí, carrer de Josep Soldevila i passeig de la Verneda (Ave Sagrera).* (2012).
2. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15758–15763 (2013).
3. Carøe, C. *et al.* Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2018).
4. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
5. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**,

e1001091 (2011).

6. Andrews, R. M. *et al.* Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA [5]. *Nat. Genet.* **23**, 147 (1999).
7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
8. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
10. Broad Institute. Picard. http://broadinstitute.github.io/picard/.
11. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
12. Jónsson, H. *et al.* mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. in *Bioinformatics* vol. 29 1682–1684 (2013).
13. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2229–2234 (2014).
14. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013).
15. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
16. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, (2015).
17. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
18. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
19. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
20. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

21. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
22. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
23. Liu, W. Q. *et al.* Salmonella paratyphi C: Genetic divergence from salmonella choleraesuis and pathogenic convergence with salmonella typhi. *PLoS One* **4**, e4510 (2009).
24. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
25. Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* **163**, 571–582 (2015).
26. Key, F. M. *et al.* Emergence of human-adapted Salmonella enterica is linked to the Neolithization process. *Nat. Ecol. Evol.* **4**, 324–333 (2020).
27. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
28. Alikhan, N. F., Zhou, Z., Sergeant, M. J. & Achtman, M. A genomic overview of the population structure of Salmonella. *PLoS Genet.* **14**, e1007261 (2018).
29. Zhou, Z. *et al.* Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* **28**, 2420-2428.e10 (2018).
30. Vågene, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).
31. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
32. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
33. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
34. Crispell, J., Balaz, D. & Gordon, S. V. Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. *Microb. Genomics* **5**, e000245 (2019).
35. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.* **20**, 406 (1971).

36. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134–e134 (2018).

Supplementary Tables

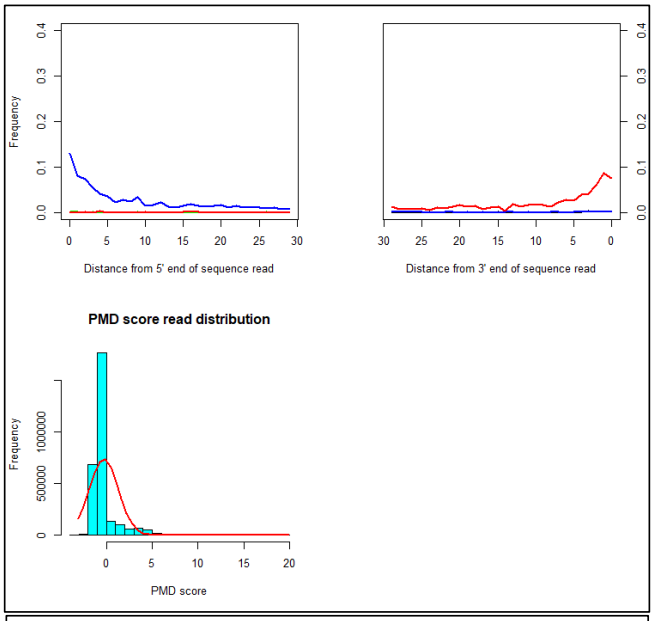
Sample	Species	Summative Taxon Reads	Exact Taxon Reads	% of all Reads in sample
F1364-1436	*Y. enterocolitica*	244	175	0,000246419
	Y. pestis	3	2	3,02974E-06
	Y. similis	34	34	3,43371E-05
	Y. pseudotuberculosis	1,827	169	0,001845112
	S. enterica	1,500	731	1.514E-05
F1691-1810	*Y. enterocolitica*	447	361	0,000181358
	Y. pestis	75	63	3,04293E-05
	Y. similis	138	138	5,59898E-05
	Y. pseudotuberculosis	945	177	0,000383409
	S. enterica	14,326	9,006	0.01%

Table S2. Metagenomic read assignation to *Yersinia* species and *Salmonella enterica* using kraken2. There is no substantial amount of *Yersinia pestis*-related bacteria.

Sample Name	Sequenced Paired Reads	Unique Reads	Quality 25 Reads	Average Depth of Coverage	% of Covered Positions
F1364-1436	99,018,404	1,572	1,452	0.003X	0.006%
F1691-1810	246,473,297	24,327	24,225	0.3852X	30.78%

Table S4. Mapping statistics of the analysed samples against *Salmonella* Paratyphi C reference genome.

A



B

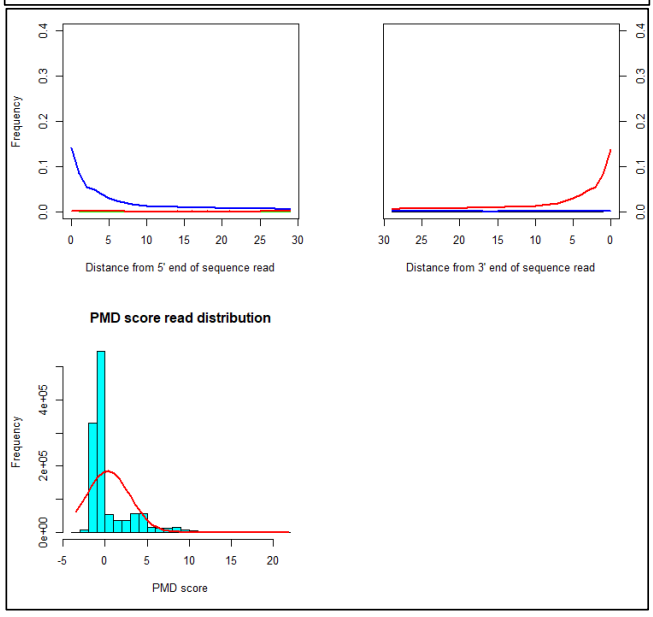


Figure S1. ancient DNA authenticity damage patterns observed in human reads from F1691-1810 (A - top) and F1364-1436 (B- bottom). Top plots provide the frequency of post-mortem damage associated substitutions at the 25 last positions of the reads in the 5' end (top left; blue) and 3' end (top right; red). Bottom plots provide the distribution of reads based on their PMD score with the associated density distribution shown by the red line.

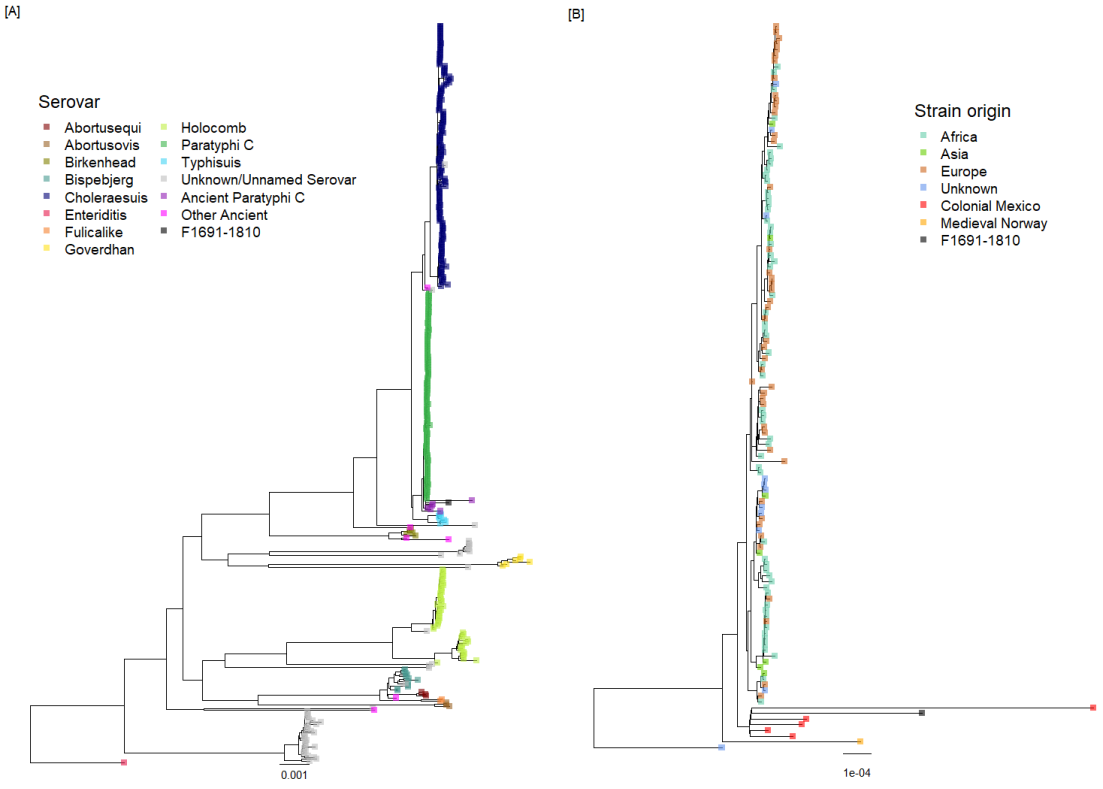


Figure S2. **:** Maximum likelihood phylogeny of *S. enterica* falling within the Ancient Eurasian Super Branch (AESB). (A) ML tree of 424 strains representing the ancient and modern diversity of the AESB. The tree is rooted using *S. enterica* ser. Enteriditis. Serovars are coloured at the branch tip. (B) Maximum likelihood phylogeny of the Paratyphi C clade including 127 Paratyphi C modern and ancient strains. Tip points are coloured according to their origin. Ancient samples display long terminal branches given their age.

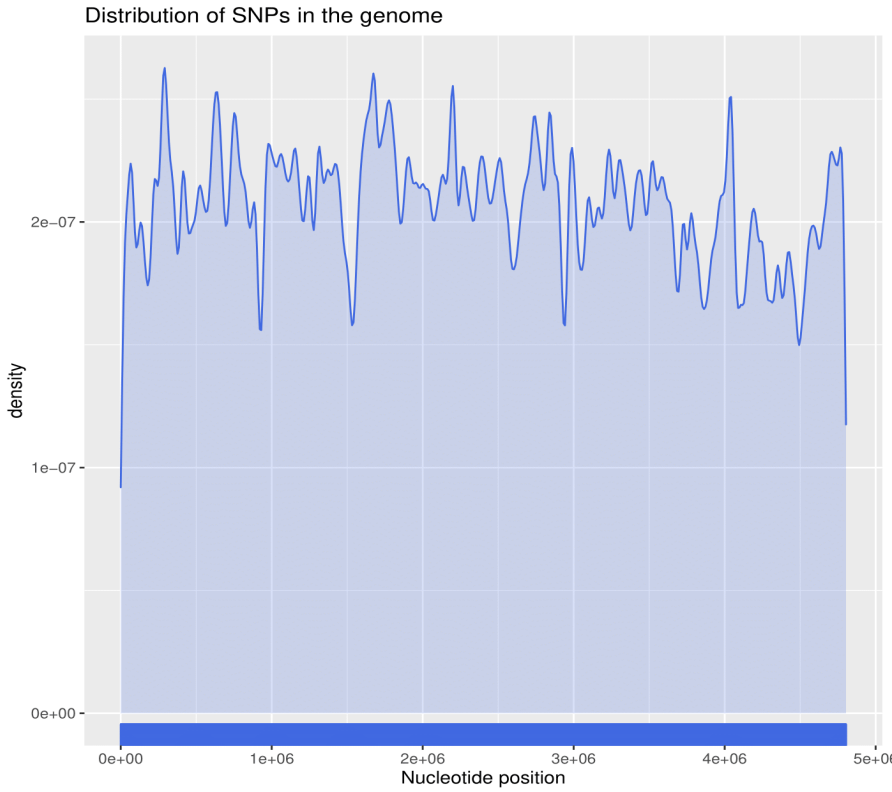


Figure S3. Distribution of the 21,793 SNPs remaining along the *Paratyphi C* genome after the removal of putative recombination events inferred by ClonalFrameML. This dataset represents a 92% of the original panel.

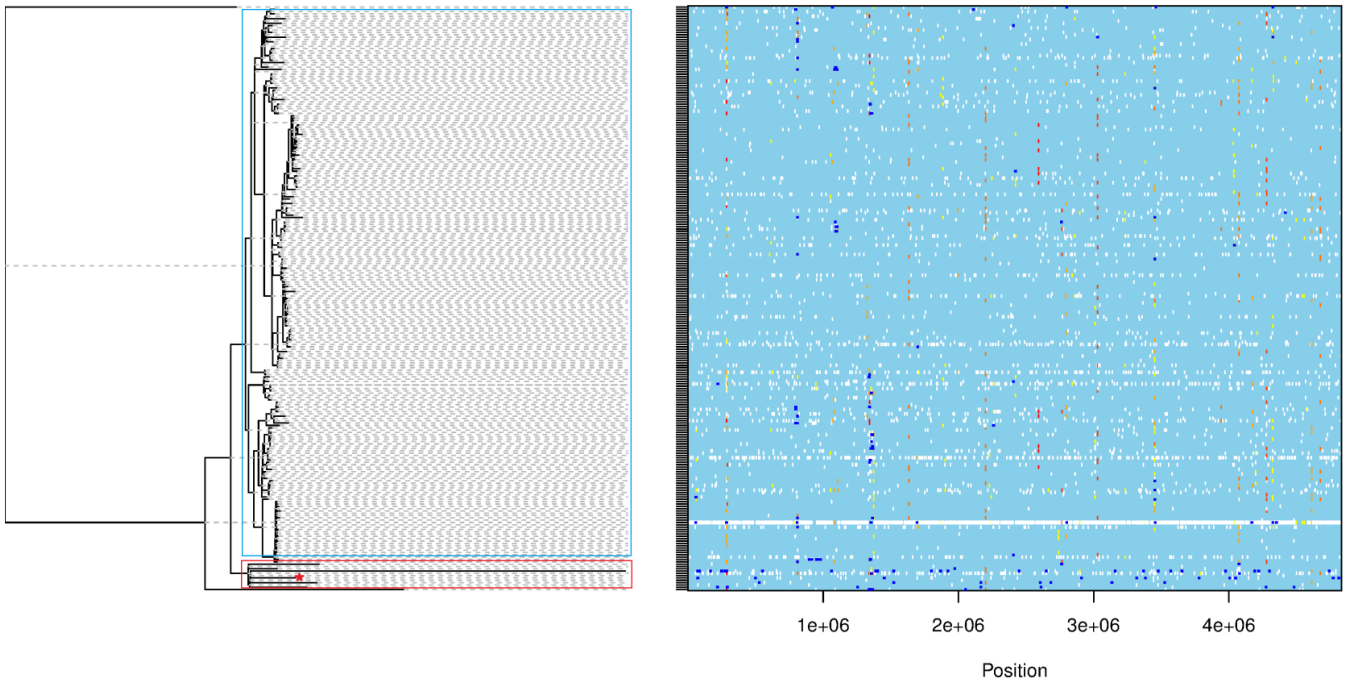


Figure S4. ClonalFrameML inferred recombinant tracts with recombination corrected ML tree of *S. enterica* Paratyphi C. Samples identified in Colonial Mexico are highlighted in red. F1691-1810 is marked with a red star. The heatmap at right provides a representation of genomic events along the Paratyphi C chromosome. Recombination tracts are marked as dark blue bars. White bars give non-homoplasic substitutions, while the bars ranging from yellow to red represent homoplasic sites (with a red value denoting a more homoplasic position).

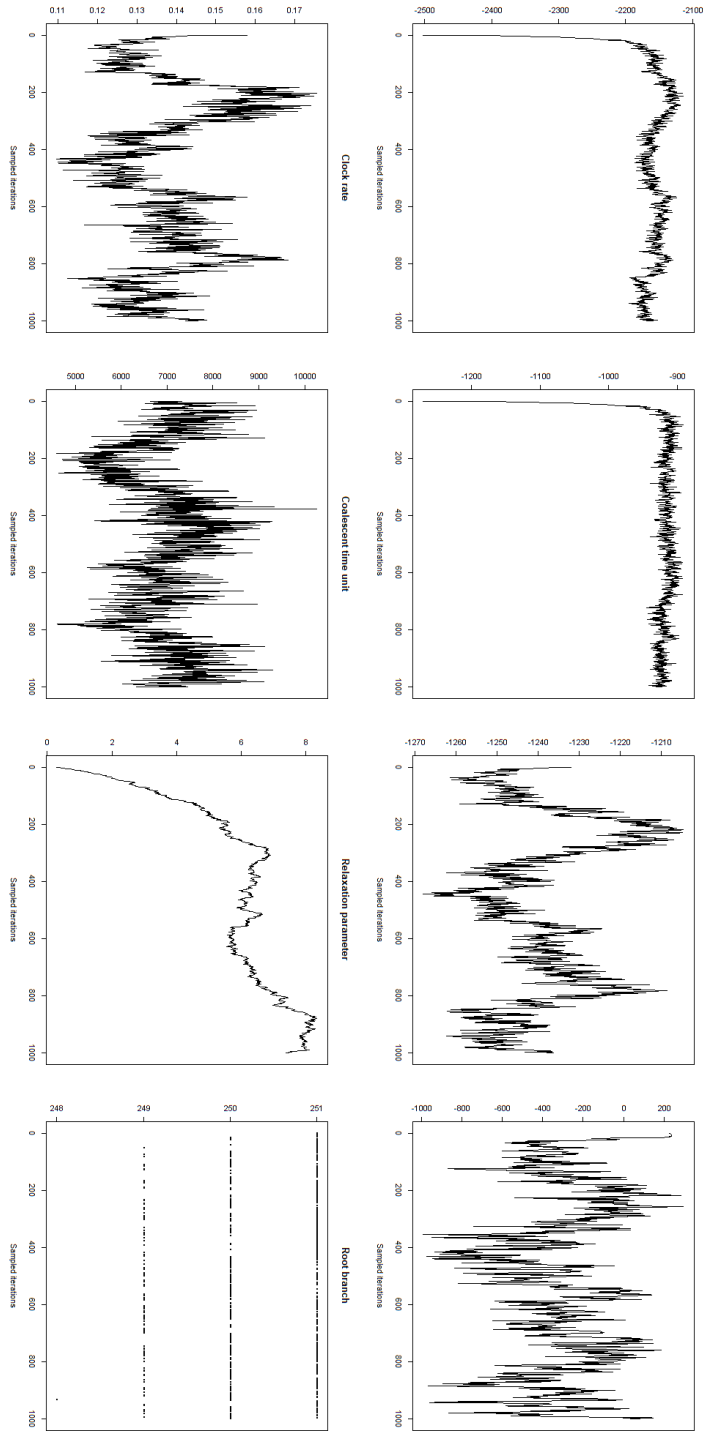


Figure S5: MCMC convergence traces for key phylogenetic parameters following 10,000 iterations.

5 Discussion

The selective pressure driven by pathogens represent one of the strongest evolutive forces in humans. Diseases and pandemics have reshaped human populations, both demographically and genetically. Due to the close relationship between pathogens and their hosts, humans, different population movements and demographic changes in the last 5 millennia, especially those occurring after the Neolithic transition, have also affected to the genomic landscape of different pathogenic agents. NGS techniques have allowed us to recover aDNA from ancient pathogens and observe those genomic imprints caused by the aforementioned events. Furthermore, we have used those same techniques to describe a historically documented clinical case from a metagenomic perspective.

In this thesis I present the result of 3 studies. The first is the characterisation and affinities of an eradicated European *Plasmodium falciparum* strain. The second is the recovery of DNA attributed to the French revolutionary Jean-Paul Marat recovered from bloodstain found in a late 18th century. The last newspaper is the recovery of a European *Salmonella enterica* Paratyphi C strain recovered from a putative *plague* epidemic during the siege of Barcelona in 1651 – 1652.

The topics I will discuss in this section are the recovery of ancient pathogen DNA and its difficulties, the genomic affinities of ancient European pathogens, the revision of historical cases using aDNA. As a remark, this thesis contributes with 2 novel findings. One is the first time *Plasmodium falciparum* ancient genomic DNA is analysed. The other is the first time human aDNA has been recovered from an ancient cellulose manuscript.

5.1 Ancient pathogen recovery

In the last years multiple ancient pathogens have been retrieved for genetic studies. The most prolific pathogen in this regard is *Yersinia pestis,* usually extracted from teeth*,* from which historical strains attributed to the first[166–169], second[56,175–178,180–182] and third pandemic[185–188] have been recovered, but furthermore, a new Neolithic[412] and 2 Bronze Age[29,163,164] lineages have been discovered. Other pathogenic bacteria retrieved from teeth are *Salmonella enterica* strains, which in the las 2 years have raised their interest as had being potentially pandemic agents in the past[30,134,219]. For other pathogens which their life cycle does not involved a bloodborne systemic disease, other samples sources are preferable, such as *Vibrio cholerae* recovered from the preserved intestines of a 19th outbreak victim[94], *Helicobacter pylori* from the stomach of a 5000 year old mummy (the Tyrolean Iceman)[92], *HIV* from paraffin preparations[90] and conserved blood plasma[413], or

Plasmodium vivax and *falciparum* in medical slides dating back to 1940s[91,283].

The recovery strategy of those pathogens usually has two radically different approaches. For the retrieval of microbial aDNA from teeth, usually tens to hundreds of different individuals are scanned before a partial genome is found, afterwards, the best candidates are used to capture the genomes of the organisms of interest. The samples are readily available, but only few of them contain pathogen DNA. In the other hand, those other pathogens with a non-conventional tissue of origin are more subject to serendipity; those samples are scarce, usually in excellent preservation conditions and usually displaying signs or have recorded history of infection before being analysed using aDNA.

At the search of Yersinia pestis DNA

Taking the first approach method I have described above; this would be the case for the *Salmonella enterica* Paratyphi C genome from La Sagrera. The initial objective was to recover *Y. pestis* sequences. From several individuals screened, only two yielded enough human DNA to be analysed, and only one of them contained pathogen DNA. I would like to highlight the problematic of finding pathogen DNA.

First, from the individual positive for *S. enterica*, 2 independent

samples were extracted from the same individual. One contained the partial Paratyphi C genome at 0.37X of average coverage and 30% of the reference recovered, displaying patterns of aDNA damage. In contrast, the other sample taken only contained endogenous DNA. Although *a priori* this could seem to be improbable, is not uncommon for different samples coming from the same individual to display different amounts of DNA as seen from other studies[24,30].

Second, is the fact that despite the historical knowledge that a Plague epidemic occurred during the 1651 siege of Barcelona[334], no traces of the *Y. pestis* were found. The characteristics of the site point to some sort of disease outbreak; the remains belonged to young men, presumably soldiers, with no signs of violent death[333]. We cannot rule out that plague was the cause of the epidemic, since the lack of *Y.pestis* sequences could be attributed to the low number of individuals that we could screen. Additionally, the presence of *S. enterica* Paratyphi C is not mutually exclusive with the presence of *Y. pestis.*

When evidence is already present

Going back to the approximation to find pathogen DNA, the other 2 studies presented in this thesis will fall upon the second category, scarce samples for which evidence of infection are visible or have been recorded. For the case of the recovery of

P. falciparum DNA from 1942 microscope slides, the parasites could be seen in the preparation. Furthermore, 3 of the 4 slides from which DNA has been recovered were previously analysed by Gelabert and colleagues, demonstrating the presence of the parasite[91]. The only issue which could affect the study regarding the actual presence of *P. falciparum* is the possibility that *P. vivax* sequences spuriously mapped against *falciparum* reference genome due to the confirm presence of co-infection. Nonetheless, this does not seem to be the case, since the number of shared reads between the 2 species is below 1% of the total reads mapped against *P. falciparum* reference genome. This resulted in a genome with 0.67X of average coverage and roughly 41% of the reference covered. Is important to mention that the obtention of *P. falciparum* from bone remain is possible as demonstrated by Marciniak and colleagues[263,414], but far more inefficient and again, with a need for extensive samples scans.

The case of the blood of Jean-Paul Marat is more complex since this was the first time the retrieval of DNA from a blood stain in an old cellulose manuscript was attempted. The most similar studies are the recovery of animal DNA from skin parchments[95,415]. Additionally, we wanted to analyse both his ancestry and the cause of his disease. We were able to recover human sequences with a 0.03X average depth of coverage for the nuclear genome, and 4X average depth of coverage for the mitochondrial genome. Marat had a skin condition which

severely affected him during the last years of his life[416]. This disease was attributed to syphilis, scabies or dermatitis of different ethology. The screening using two metagenomic profiling software (*KrakenUniq*[417] and *metamix*[418]) showed the presence of several microbial species which could explain Marat's symptoms; *Acinetobacter junii* (59.96% reference recovered), *Aspergillus glaucus* (1.65%), *Staphylococcus aureus* (1.03%), *Cutibacterium acnes* (94.79%) and *Malassezia restricta* (14.21%). All those microbes but *A. junii* showed patterns characteristics of aDNA damage. Nevertheless, after comparing the microbial profiles of Marat's manuscript with historical parchments (the most similar object we could compare our sample)[95], only *C. acnes, S. aureus* and *M. restricta* were considered as validate candidates to explain Marat's symptoms. Here the difficulty lies on discerning which microbes are contributing to the infection and which could be expected to be found as commensals in a healthy skin biota.

The problem of contamination

From the 3 studies presented, the least susceptible to be contaminated is the recovery of the *P. falciparum* European strain; the samples were handled in a laboratory where other Malaria samples were not processed at the moment, and the sequences showed aDNA damage patterns demonstrating their authenticity.

In the other hand, samples from la Sagrera were used to analyse both human and pathogen DNA. Since archaeologist had handled the remains, human contaminant sequences were to be expected. One of the samples was in fact contaminated, but since we had an additional sample of this individual, we could discard the contaminated data and use the non-contaminated for the population genetics analysis. Fortunately, *S. enterica* Paratyphi C contamination could be ruled out since the retrieved sequences showed post-mortem damage patterns. Moreover, infections of this type are nowadays relatively scarce in developed countries[419].

Again, Marat's sample was difficult to deal with, since there were multiple focusses of possible contamination, both for endogenous and for microbial DNA. The manuscript was handled for years without any precaution, and modern and ancient contaminants were to be expected. This was the case for the human DNA, from an original genome sequences at 0.1X and 16X for the nuclear and mitochondrial genomes respectively, after filtering for the presence of possible aDNA damage in the reads, we got the aforementioned genomes at 0.03X and 4X. Despite in this case modern contamination has been removed from the sample, certain fraction of ancient contaminants seems to be left in the sample, at least for mitochondrial DNA sequences, but not for nuclear DNA. Nonetheless, similar methods for filtering contaminants have been successfully applied, being the resultant data apt for its

use in population genetic analysis[355]. For the microbial contamination, *A. junii* was automatically considered contaminant because it did not display aDNA damage. *Aspergillus* were also considered as probable contaminants since their ubiquitous environmental distribution and their ability to grow in different substrates. Finally, although *M. restricta, S. aureus* and *C. acnes* could be originated from a contaminant source, their absence in historical parchments[95] suggest that they may be actually linked to Marat's condition.

5.2 Pathogens in historical Europe

As we have seen thought all the thesis, different diseases ravaged Europe for centuries, their origin being linked to the change in demographics resultant of the Neolithic transition, which for example, allowed the *Plasmodium* parasites to expand worldwide[138], or allowed *Salmonella enterica* serovars to infect humans through different zoonotic jumps[134]. The other event which allowed those pathogens to reach every corner of the globe is the Age of Discoveries. During this period of time, European powers gain control of different areas in the world, propitiating again the conditions for the spreading of different diseases.

Malaria, the case of a locally eradicated cosmopolitan disease

The *P. falciparum* strain which we have retrieved is of great importance since it provide us of a snapshot the genetic characteristics of a pathogen which is now eradicated from Europe[420]. Previous studies by Gelabert and colleagues demonstrate the plausible origin of the eradicated European strain in the Indian subcontinent using mitochondrial genomes[91]. The retrieval of wide genome data from this strain has allowed us the usage of more informative population genetic analysis.

The genetic affinities of the European *P. falciparum* as viewed through a PCA show us a close relationship between Bangladeshi strains and the European strain, confirming the observations made by Gelabert and colleagues[91]. Besides this, in the PCA is also exhibited the close affinities to South East Asian and Melanesian strains. The usage of a formal test, F4 statistic, demonstrate that the European *P. falciparum* strain is genetically closer to Asian strains than to African or American ones. Nevertheless, admixture events of the European and other strains cannot be discarded. It could be argued that its position in the PCA, at the centre of diversity, is due to some genetic exchange between African and Asian strains. This can be observed in an unsupervised Admixture analysis (not published but added to the thesis as Figure 17),

which show us that the European strain is in fact composed of different global components, including African, Asian and even American sources. This is also evidenced using *Chromopainter*, where the European *P. falciparum* share haplotype chunks with Asian strains, but also a considerable amount with African and American.

In the context of the dispersion of *P. falciparum*, the parasite in Europe is recorded only during the antiquity[421–424] and due to its genetic affinity to Indian strains, an introduction from Asia. Maybe this occurred during the stablish of the Achaemenid Empire or Alexander's conquest during the Hellenistic period, rather than an African origin through Mediterranean contacts during the Roman Empire. In regard the effects of colonialism on the dispersal of the parasite, we might consider if *P. falciparum* shares a similar history to *P. vivax.* In the case of *P. vivax*, van Dorp and colleagues demonstrate a European origin of the American strains. In the case of *P. falciparum*, previous studies suggest an African origin of the American diversity, linked to the Atlantic slave trade[265]. Despite our data also denotes a close genetic affinity between African and American strains(as seen in a PCA), different event of admixture between American and European *P. falciparum* strains may had happened to a certain degree and should not be excluded from consideration.

In addition to the genetic to the study of this *P. falciparum*

strain and its relationship to other existing strains in the context of their global dispersal, the date of this sample provides us of a “naïve” genome which predate the introduction of most antimalarial policies[425]. This is reflected in the presence of described SNPs which confer drug resistances. From 117 interrogated positions, only 2 display an allele compatible with drug resistance. Those mutations are in the *pfmrp1* gene and provide resistance to quinine and chloroquine[239,245,250]. The presence of those mutations could be explained by the natural occurrence of the polymorphisms in *P. falciparum* strains, or by the interaction with some kind of antimalarial drug. The later seems plausible with the fact that quinine is the oldest known antimalarial compound and at that time had been already used by the Spanish for 3 centuries[426]. Despite this, and given the data retrieved, it seems that this ancient strain has not been touch by the selective pressure pursued by Malaria treatment policies. The same fact has also been observed in historical European *P. vivax* strains, which only display a couple derived alleles in positions linked to drug resistance[269,283].

Finally, I would like to remark the limitations of the usage of a genome at this coverage (0.67X and 41% of positions covered). The ventage of the population genetic analysis used (PCA and F4) is that they are fairly robust and not that much affected by the lack of data to a certain point. What is true that if more data, hence more coverage, were available, a set more precise results could been have given. This would have

allowed to determine with more exactitude the admixture proportions of the European strain or use phylogenetic approaches to infer the split date of the European linage from other lineages. Furthermore, this would have allowed for a better drug resistance mutation scanning, since from the original 117 positions analysed, only 62 were covered, and 23 have 2 or more depth of coverage. We had to rely on imputation to assess the authenticity of the 2 present SNPs.

Paratyphi C and its forgotten role in past epidemics

The recovery of *S. enterica* Paratyphi C from a putative Spanish soldier dating back to 1652 opens the way for questions regarding the past relevance of Paratyphi C in ancient and historical epidemics. Recent studies have demonstrated the emergence of *Salmonella enterica* serovars tied to the Neolithic transition and animal domestication[134]. What is more, this same serovar Paratyphi C, despite nowadays being relatively scarce, was present in Medieval times[219] and has attributed to be one of the causal agent of the *Cocoliztli* epidemics during the mid-16th century in Mexico[30].

From a phylogenetic perspective, the Paratyphi C Spanish strain cluster with the Colonial strains from Mexico, being basal to all diversity within the Paratyphi C clade but the Medieval strain. The presence of those genetically similar Spanish and Mexican strains highlights the importance of the population

and goods movements between America and Europe during this period. In this case, this suggests a reintroduction of Paratyphi C strains back to Europe. Due to scarcity of the available data, we could not accurately date the split between the American and European strains.

I would like to mention that one of the main limitations present in this study is the amount of data available, a genome at 0.38X and covering roughly 30% of the reference genome. Due to this, we had to be very careful when calling variants and had to apply relatively astringent filtering parameters to evade spurious calls caused by aDNA damage. This resulted in a low number of confident calls in la Sagrera strain which could be used in the phylogeny. The software used to create the ML tree can tolerate missingness in the data up to a point. Missingness in la Sagrera manifested for example in an abnormally high branch length (which could be also attributed to aDNA damage lesions erroneously called despite the filtering parameters used) and low support within the Colonial Mexico clade. Branch length could be partially solved by removing recombinant and homoplasic sites. At the end, this inconsistencies difficulty the assessment of a correct dating.

Other of the main issues arisen from this finding is the fact that previous to this study, it was assumed that a Plague epidemic affected the besiegers of Barcelona in 1652. As previously mentioned, only few individuals were screened, when similar

studies screen from tens to hundreds of different candidate individuals, and then capture in positive samples is needed to recover complete *Y. pestis* genomes[29,164,169,181]. Additionally, the presence of Paratyphi C is not mutually exclusive with the presence of Plague. What those findings suggest is that due to the fact that enteric fevers do not leave physical evidence of infection in the remains, their past outbreaks and epidemics may have been overlooked. A way to tackle this problem would be to massively screen for the presence of *S. enterica* serovars in mass burials associated to other epidemics or outbreaks.

5.3 Revision of a documented historical disease case

aDNA has been used as a tool to analyse the remain of various famous historical figures. This has been the case for the remains of king Richard III of England[31] or the blood of the king Louis XVI of France[33]. Here we retrieve DNA attributed to the French revolutionary Jean-Paul Marat to both characterise its ancestry and try to elucidate the cause of his disease. Marat suffered from a itching skin condition which has previously been identified by historical revisionist as scabies, syphilis, atopic dermatitis, seborrheic dermatitis or dermatitis herpetiformis[427–431].

As mention in the first section of the discussion, after filtering the data to extract only the reads considered as truly ancient,

we are left with only a tiny fraction of the genome (0.03X average depth for the nuclear genome, 4X for the mitochondrial genome). The major mitochondrial haplogroup was determined to be H2a2a1f, with traces of a possible contaminant sequence of the haplogroup K1a15. Both haplogroups are found in Europe. Nevertheless, the presence of this possible contaminant does not seem to be apparent in the nuclear data since the ratio of heterozygosity in the X chromosome does not suggest the possibility of contaminant sequences. Is important to remark that the genetic sex of the sample is compatible with male but not female, hence that contamination could be measured using X chromosome frequencies. The data at our disposal was enough to perform basic population genetic analyses; PCA, ADMIXTURE and F statistics. In the PCA, the sample falls within current French diversity, somewhat displaced to the North Italian – Spanish cluster. The ADMIXTURE analysis also displays a profile compatible with a partial French ancestry. Finally, the F4 statistic denote that the sample is genetically closer to French, Basque, North Italian and English populations rather than to Spanish, South Italian or Sicilian ones. Altogether, the genetic profile of the sample is compatible with the mixed ancestry that Marat is thought to had have; his father was of Sardinian origin while his mother was a French Huguenot. Due to the low abundance of data, and the possibility of ancient contamination, the results have to be taken with caution, but they seem compatible with truly belonging to Marat. An ideal

improvement for this study would have been to have genetic data of a Marat's relative in order to prove the authenticity of the retrieved genomic sequences.

The microbial profile left after discarding possible contaminants is composed by *M. restricta, S.* aurerus and *C. acnes.* Those three microorganisms seem to be truly ancient; they display aDNA damage pattern and when placed in a phylogeny including modern day strains, they fall basal to those diversity. Furthermore, when compared with comparable historical objects, they do not appear to be present in their metagenomic profile. Of particular interest is *M. restricta,* a fungus which exclusively grows in the skin and which although could be present as commensal in a healthy microbial flora, has been linked to skin conditions such as seborrheic dermatitis, atopic dermatitis, dandruff and folliculitis[432]. Seborrheic dermatitis matches the symptoms which Marat seemed to suffer[431]. Additionally, *S. aureus* and *C. acnes* had also been linked to opportunistic skin infections. Although sequences of those microbes have been also found in the swab coming from the unstained part of the document, this could be expected if the paper had been handled by a heavily infected individual. Is for this reason that to confirm the present results, other documents handled by Jean-Paul Marat should be analysed.

Given the presence of those specific organisms, those results

seem to indicate that Marat probably suffered from a fungal or polymicrobial infection, either primary or secondary, in an advanced stage. The usage of NGS techniques has thus allowed to shed light in a historical case of disease, and given the present data, suggest a possible cause for Marat's pathology,

5.4 Conclusions

In this thesis we focus on the power of NGS techniques to retrieve ancient pathogen DNA from different context: the eradicated strain of a globally distributed disease, an ancient epidemic and a historically documented case of an individual suffering from a disease.

Stating with the recovery of the European *Plasmodium falciparum*, is important to remark that this has been the first time that wide genome data of an ancient strain of *P. falciparum* has been analysed. Using this genome data, we provide evidence that corroborate the hypothesis of an Asian origin of European *P. falciparum*. As expected, this strain does not show signals of being selected by the usage of antimalarial drug. To better understand the dynamics of *P. falciparum* malaria in Europe and its genetic relationship with other circulating strains, it would be necessary to sequence more data from either other medical collections or from bone remains. This would be particularly interesting to understand if

American or African strains have contributed to genomic diversity of European strains, which with the given amount data, we cannot properly test.

Following with the *Salmonella enterica* Paratyphi C in remains from the 1652 Barcelona siege, these findings suggest that although this pathogen is nowadays scarce, it could have been fairly common in the past and could substantially contribute to outbreaks and epidemics. Besides, the Spanish Paratyphi C showed genetic affinity to Colonial Mexico strains, highlighting the importance trans-Atlantic exchange of pathogens between Europe and America. Due to the limited data at our disposal, we could not properly date the split of this Spanish and Mexico strains to confirm the European origin of the *Salmonella* epidemics which ravaged the colonial Mexico. Is for this reason that future sample's screenings at a grand scale are necessary to retrieve more data which would allow to understand the arrival of the disease to the New World and the role that Paratyphi C played in historical epidemics.

Finally, with the use of NGS techniques we have been able to take a deeper look at the medical case of Jean-Paul Marat, which has remained unanswered for almost 230 years. With this study we present the first evidence of aDNA recovery from an ancient paper sheet. This could open the gate for the study of other historical cases. Furthermore, using metagenomic techniques we proposed a diagnosis given the historical

records of Marat's disease and the microbial data present in the sample. Due to that other hypothesis about his disease could not still be ruled out, other documents handled by Marat should be analysed in the search of more data.

To conclude, in this thesis we have seen the relevance of ancient DNA for the study of diseases throughout human history. This information is of remarkable importance to understand the spread of diseases and pathogens across the world, providing knowledge on how they affect us humans, and also how we affect them.

6 Contribution in Other publications

Gelabert P, Olalde I, **de-Dios T**, Civit S, Lalueza-Fox C. Malaria was a weak selective force in ancient Europeans. Sci Rep. 2017 May 3;7(1):1377. doi: 10.1038/s41598-017-01534-5.

Gelabert P, Ferrando-Bernal M, **de-Dios T**, Mattorre B, Campoy E, Gorostiza A, Patin E, González-Martín A, Lalueza-Fox C. Genome-wide data from the Bubi of Bioko Island clarifies the Atlantic fringe of the Bantu dispersal. BMC Genomics. 2019 Mar 6;20(1):179. doi: 10.1186/s12864-019-5529-0.

Gelabert P, Sandoval-Velasco M, Serres A, de Manuel M, Renom P, Margaryan A, Stiller J**, de-Dios T**, Fang Q, Feng S, Mañosa S, Pacheco G, Ferrando-Bernal M, Shi G, Hao F, Chen X, Petersen B, Olsen RA, Navarro A, Deng Y, Dalén L, Marquès-Bonet T, Zhang G, Antunes A, Gilbert MTP, Lalueza-Fox C. Evolutionary History, Genomic Adaptation to Toxic Diet and Extinction of the Carolina Parakeet. Curr Biol. 2020 Jan 6;30(1):108-114.e5. doi: 10.1016/j.cub.2019.10.066.

van Dorp L, Gelabert P, Rieux A, de Manuel M**, de-Dios T**, Gopalakrishnan S, Carøe C, Sandoval-Velasco M, Fregel R, Olalde I, Escosa R, Aranda C, Huijben S, Mueller I, Marquès-Bonet T, Balloux F, Gilbert MTP, Lalueza-Fox C. Plasmodium

vivax Malaria Viewed through the Lens of an Eradicated European Strain. Mol Biol Evol. 2020 Mar 1;37(3):773-785. doi: 10.1093/molbev/msz264.

Ferrando-Bernal M, Morcillo-Suarez C**, de-Dios T**, Gelabert P, Civit S, Díaz-Carvajal A, Ollich-Castanyer I, Allentoft ME, Valverde S, Lalueza-Fox C. Mapping co-ancestry connections between the genome of a Medieval individual and modern Europeans. Sci Rep. 2020 Apr 22;10(1):6843. doi: 10.1038/s41598-020-64007-2.

7 Bibliography

1. Darwin, C. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. (John Murray, 1859).
2. Mendel, G. Experiments on Plant Hybrids (1866) - Translation and commentary by Staffan Müller-Wille and Kersten Hall. *Br. Soc. Hist. Sci. Transl. Ser.* (2016).
3. Morgan, T. H., Sturtevant, A. H., Muller, H. J. & Bridges, C. B. *The mechanism of Mendelian heredity*. (Holt, 1915).
4. Sutton, W. S. On the morphology of the chromosome group in Brachystola magna. *Biol. Bull.* 24–39 (1902).
5. Boveri, T. *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns.* . (G. Fischer, 1904).
6. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *J. Exp. Med.* **79**, 137–158 (1944).
7. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
8. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356

(1961).

9. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
10. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).
11. Saiki, R. K. *et al.* Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. U.S. Department of Energy. Human Genome Project Information Archive (1990-2003): Available at https://web.ornl.gov/hgmis. https://web.ornl.gov/hgmis (2019).
14. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365 (1998).
15. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
16. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
17. Wheeler, D. A. *et al.* The complete genome of an

individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).

18. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Genereux, D. P. *et al.* A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
20. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).
21. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13**, 278–289 (2015).
22. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 2016-09-05. www.genome.gov/sequencingcostsdata/ (2016).
23. Murray, G. G. R. *et al.* Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *bioRxiv* **954**, 951–954 (2017).
24. Green, R. E. *et al.* A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
25. Fry, E. *et al.* Functional Architecture of Deleterious Genetic Variants in the Genome of a Wrangel Island Mammoth. *Genome Biol. Evol.* **12**, 48–58 (2020).
26. Schiffels, S. *et al.* Iron Age and Anglo-Saxon genomes

from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016).

27. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
28. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
29. Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* **163**, 571–582 (2015).
30. Vågene, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).
31. King, T. E. *et al.* Identification of the remains of King Richard III. *Nat. Commun.* **5**, 5631 (2014).
32. Olasz, J. *et al.* DNA profiling of Hungarian King Béla III and other skeletal remains originating from the Royal Basilica of Székesfehérvár. *Archaeol. Anthropol. Sci.* 1345–1357 (2019) doi:10.1007/s12520-018-0609-7.
33. Olalde, I. *et al.* Genomic analysis of the blood attributed to Louis XVI (1754-1793), king of France. *Sci. Rep.* **4**, 4666 (2014).
34. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. & Wilson, A. C. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**, 282–284 (1984).

35. Pääbo, S., Gifford, J. A. & Wilson, A. C. Mitochondrial DNA sequences from a 7000-year old brain. *Nucleic Acids Res.* **16**, 9775–9787 (1988).
36. Hagelberg, E., Sykes, B. & Hedges, R. Ancient bone DNA amplified. *Nature* vol. 342 485 (1989).
37. Rollo, F., Amici, A., Salvi, R. & Garbuglia, A. Short but faithful pieces of ancient DNA. *Nature* vol. 335 774 (1988).
38. Höss, M., Kohn, M., Pääbo, S., Knauer, F. & Schröder, W. Excrement analysis by PCR. *Nature* **359**, 199 (1992).
39. Thomas, R. H., Schaffner, W., Wilson, A. C. & Pääbo, S. DNA phylogeny of the extinct marsupial wolf. *Nature* **340**, 465–467 (1989).
40. Hagelberg, E. *et al.* DNA from ancient mammoth bones. *Nature* **370**, 333–334 (1994).
41. Woodward, S. R., Weyand, N. J. & Bunnell, M. DNA sequence from cretaceous period bone fragments. *Science* **266**, 1229–1232 (1994).
42. Cano, R. J., Poinar, H. N., Pieniazek, N. J., Acra, A. & Poinar, G. O. Amplification and sequencing of DNA from a 120-135-million-year-old weevil. *Nature* **363**, 536–538 (1993).
43. Hedges, S. B. *et al.* Detecting dinosaur DNA. *Science* **268**, 1191–1194 (1995).
44. Austin, J. J., Ross, A. J., Smith, A. B., Fortey, R. A. & Thomas, R. H. Problems of reproducibility - Does geologically ancient DNA survive in amber-preserved

insects? *Proc. R. Soc. B Biol. Sci.* **264**, 467–474 (1997).

45. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
46. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
47. Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5581–5585 (1999).
48. Drancourt, M., Aboudharam, G., Signoli, M., Dutour, O. & Raoult, D. Detection of 400-year-old Yersinia pestis DNA in human dental pulp: an approach to the diagnosis of ancient septicemia. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 12637–12640 (1998).
49. Salo, W. L., Aufderheide, A. C., Buikstra, J. & Holcomb, T. A. Identification of Mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 2091–2094 (1994).
50. Arriaza, B. T., Salo, W., Aufderheide, A. C. & Holcomb, T. A. Pre-Columbian tuberculosis in Northern Chile: Molecular and skeletal evidence. *Am. J. Phys. Anthropol.* **98**, 37–45 (1995).
51. Poinar, H. N. *et al.* Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311**, 392–394 (2006).
52. Rogaev, E. *et al.* Complete Mitochondrial Genome and

Phylogeny of Pleistocene MammothMammuthus primigenius. *PLoS Biol.* **4**, e73 (2006).

53. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
54. Rasmussen, M. *et al.* An aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
55. Orlando, L. *et al.* Recalibrating equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
56. Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506–510 (2011).
57. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* **468**, 1053–1060 (2010).
58. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894–897 (2010).
59. Chen, F. *et al.* A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature* **569**, 409–412 (2019).
60. Douka, K. *et al.* Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature* **565**, 640–644 (2019).
61. Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps Article Reconstructing

Denisovan Anatomy Using DNA Methylation Maps. *Cell* **179**, 180–192 (2019).

62. Olalde, I. *et al.* The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019).
63. Fernandes, D. M. *et al.* The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean. *Nat. Ecol. Evol.* **4**, 334–345 (2020).
64. Narasimhan, V. M. *et al.* The Genomic Formation of South and Central Asia. *bioRxiv* 292581 (2018) doi:10.1101/292581.
65. Olalde, I. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018).
66. Demarchi, B. *et al.* Protein sequences bound to mineral surfaces persist into deep time. *Elife* **5**, (2016).
67. Meyer, M. *et al.* Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**, 504–507 (2016).
68. Welker, F. *et al.* Enamel proteome shows that Gigantopithecus was an early diverging pongine. *Nature* **576**, 262–265 (2019).
69. Brown, S. *et al.* Identification of a new hominin bone from Denisova Cave, Siberia using collagen fingerprinting and mitochondrial DNA analysis. *Sci. Rep.* **6**, 23559 (2016).
70. Darzynkiewicz, Z. *et al.* Cytometry in cell necrobiology: Analysis of apoptosis and accidental cell death

(necrosis). *Cytometry* **27**, 1–20 (1997).

71. Höss, M., Jaruga, P., Zastawny, T. H., Dizdaroglu, M. & Pääbo, S. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.* **24**, 1304–1307 (1996).
72. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings. Biol. Sci.* **279**, 4724–4733 (2012).
73. Glocke, I. & Meyer, M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res.* **27**, 1230–1237 (2017).
74. Paabo, S. Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 1939–1943 (1989).
75. Dizdaroglu, M. Oxidative damage to DNA in mammalian chromatin. *Mutat. Res.* **275**, 331–342 (1992).
76. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659–674 (2017).
77. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15758–15763 (2013).
78. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3**, 698 (2012).
79. Günther, T. *et al.* Population genomics of Mesolithic

Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* **16**, e2003703 (2018).

80. Frisch, T., Sørensen, M. S., Overgaard, S., Lind, M. & Bretlau, P. Volume-referent bone turnover estimated from the interlabel area fraction after sequential labeling. *Bone* **22**, 677–682 (1998).
81. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
82. Pinhasi, R. *et al.* Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLoS One* **10**, 1–13 (2015).
83. Sirak, K. *et al.* Human auditory ossicles as an alternative optimal source of ancient DNA. *Genome Res.* **30**, 427–436 (2020).
84. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* **5**, 11184 (2015).
85. Hansen, H. B. *et al.* Comparing ancient DNA preservation in petrous bone and tooth cementum. *PLoS One* **12**, 1–18 (2017).
86. Margaryan, A. *et al.* Ancient pathogen DNA in human teeth and petrous bones. *Ecol. Evol.* **8**, 3534–3542 (2018).
87. Rascovan, N. *et al.* Tracing back ancient oral microbiomes and oral pathogens using dental pulps from

ancient teeth. *npj Biofilms Microbiomes* **2**, 6 (2016).

88. Schuenemann, V. J. *et al.* Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science* **341**, 179–183 (2013).
89. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494–497 (2014).
90. Gryseels, S. *et al.* A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue. *Proc. Natl. Acad. Sci. U. S. A.* **117**, (2020).
91. Gelabert, P. *et al.* Mitochondrial DNA from the eradicated European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11495–11500 (2016).
92. Maixner, F. *et al.* The 5300-year-old Helicobacter pylori genome of the Iceman. *Science* **351**, 162–165 (2016).
93. Duggan, A. T. *et al.* 17th Century Variola Virus Reveals the Recent History of Smallpox. *Curr. Biol.* **26**, 3407–3412 (2016).
94. Devault, A. M. *et al.* Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849 . *N. Engl. J. Med.* **370**, 334–340 (2014).
95. Teasdale, M. D. *et al.* Paging through history: Parchment as a reservoir of ancient DNA for next generation sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20130379 (2015).
96. Søe, M. J. *et al.* Ancient DNA from latrines in Northern

Europe and the Middle East (500 BC–1700 AD) reveals past parasites and diet. *PLoS One* **13**, 1–17 (2018).

97. Willerslev, E. *et al.* Diverse plant and animal genetic records from holocene and pleistocene sediments. *Science* **300**, 791–795 (2003).
98. Slon, V. *et al.* Neandertal and Denisovan DNA from Pleistocene sediments. *Science* **356**, 605–608 (2017).
99. Jensen, T. Z. T. *et al.* A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nat. Commun.* **10**, 5520 (2019).
100. Gilbert, M. T. P. *et al.* Recharacterization of ancient DNA miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Res.* **35**, 1–10 (2007).
101. Hofreiter, M., Jaenicke, V., Serre, D., Von Haeseler, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
102. Binladen, J. *et al.* Assessing the fidelity of ancient DNA sequences amplified from nuclear genes. *Genetics* **172**, 733–741 (2006).
103. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 14616–14621 (2007).
104. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* **7**, (2012).

105. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20130624 (2015).
106. Lindahl, T. & Andersson, A. Rate of Chain Breakage at Apurinic Sites Double-Stranded Deoxyribonucleic Acid. *Biochemistry* **11**, 3618–3623 (1972).
107. Lindahl, T. & Nyberg, B. Rate of Depurination of Native Deoxyribonucleic Acid. *Biochemistry* **11**, 3610–3618 (1972).
108. Pääbo, S. & Wilson, A. C. Miocene DNA sequences - a dream come true? *Curr. Biol.* **1**, 45–46 (1991).
109. Poinar, H. N. *et al.* Molecular coproscopy: Dung and diet of the extinct ground sloth Nothrotheriops shastensis. *Science* **281**, 402–406 (1998).
110. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).
111. Richter, D. *et al.* The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature* **546**, 293–296 (2017).
112. Moodley, Y. *et al.* Age of the association between Helicobacter pylori and man. *PLoS Pathog.* **8**, e1002693–e1002693 (2012).
113. Gurven, M. & Kaplan, H. Longevity among hunter-gatherers: A cross-cultural examination. *Popul. Dev. Rev.* **33**, 321–365 (2007).
114. Houldcroft, C. J., Ramond, J. B., Rifkin, R. F. &

Underdown, S. J. Migrating microbes: what pathogens can tell us about population movements and human evolution. *Ann. Hum. Biol.* **44**, 397–407 (2017).

115. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).
116. Carter, R. Speculations on the origins of Plasmodium vivax malaria. *Trends Parasitol.* **19**, 214–219 (2003).
117. Joy, D. A. *et al.* Early origin and recent expansion of Plasmodium falciparum. *Science* **300**, 318–321 (2003).
118. Chebloune, Y. *et al.* Structural analysis of the 5' flanking region of the β-globin gene in African sickle cell anemia patients: Further evidence for three origins of the sickle cell mutation in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 4431–4435 (1988).
119. Ohashi, J. *et al.* Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am. J. Hum. Genet.* **74**, 1198–1208 (2004).
120. Lapouniéroulie, C. *et al.* A novel sickle cell mutation of yet another origin in Africa: the Cameroon type. *Hum. Genet.* **89**, 333–337 (1992).
121. Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96**, 2358–2363 (2000).
122. Hoberg, E. P., Alkire, N. L., De Queiroz, A. & Jones, A. Out of Africa: Origins of the Taenia tapeworms in

humans. *Proc. R. Soc. B Biol. Sci.* **268**, 781–787 (2001).

123. Cockburn, T. A. the Evolution and Eradication of Infectious Diseases. *Perspect. Biol. Med.* **36**, 498–499 (1964).
124. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).
125. Kittler, R., Kayser, M. & Stoneking, M. Molecular evolution of Pediculus humanus and the origin of clothing. *Curr. Biol.* **13**, 1414–1417 (2003).
126. Cardona, P. J., Català, M. & Prats, C. Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance. *Sci. Rep.* **10**, 42 (2020).
127. Monot, M. *et al.* Comparative genomic and phylogeographic analysis of Mycobacterium leprae. *Nat. Genet.* **41**, 1282–1289 (2009).
128. Diavatopoulos, D. A. *et al.* Bordetella pertussis, the causative agent of whooping cough, evolved from a distinct, human-associated lineage of B. bronchiseptica. *PLoS Pathog.* **1**, 0373–0383 (2005).
129. Armelagos, G. J., Barnes, K. C. & Lin, J. Disease in Human Evolution: The Reemergence of Infectious Disease in the Third Epidemiological Transition. *AnthroNotes Museum Nat. Hist. Publ. Educ. Natl. Museum Nat. Hist. Bull. Teach. Notes A Newsl. Teach.* **18**, 1 (2014).

130. Kristiansen, K. The formation of tribal systems in northern europe, 4000-500 bc. in *Social Transformations in Archaeology: Global and Local Perspectives* (eds. Renfrew, C., Rowlands, M. J. & Segraves, B. A.) 64–102 (Academic Press, 2005). doi:10.4324/9780203984550-11.
131. Bellwood, P. & Oxenham, M. The expansions of farming societies and the role of the Neolithic demographic transition. in *The Neolithic Demographic Transition and its Consequences* (eds. Bocquet-Appel, J.-P. & Bar-Yosef, O.) 13–34 (Springer Netherlands, 2008). doi:10.1007/978-1-4020-8539-0_2.
132. Fuchs, K. *et al.* Infectious diseases and Neolithic transformations: Evaluating biological and archaeological proxies in the German loess zone between 5500 and 2500 BCE. *Holocene* **29**, 1545–1557 (2019).
133. Barrett, R., Kuzawa, C. W., McDade, T. & Armelagos, G. J. Emerging and Re-emerging Infectious Diseases: The Third Epidemiologic Transition. *Annu. Rev. Anthropol.* **27**, 247–271 (1998).
134. Key, F. M. *et al.* Emergence of human-adapted Salmonella enterica is linked to the Neolithization process. *Nat. Ecol. Evol.* **4**, 324–333 (2020).
135. Byun, R., Elbourne, L. D. H., Lan, R. & Reeves, P. R. Evolutionary relationships of pathogenic clones of VIbrio cholerae by sequence analysis of four housekeeping

genes. *Infect. Immun.* **67**, 1116–1124 (1999).

136. Krause-Kyora, B. *et al.* Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *Elife* **7**, e36666 (2018).
137. Robbins, G. *et al.* Ancient skeletal evidence for leprosy in India (2000 B.C.). *PLoS One* **4**, e5669–e5669 (2009).
138. Otto, T. D. *et al.* Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).
139. Düx, A. *et al.* Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science* **368**, 1367–1370 (2020).
140. Furuse, Y., Suzuki, A. & Oshitani, H. Origin of measles virus: Divergence from rinderpest virus between the 11th and 12th centuries. in *Virology Journal* vol. 7 52 (2010).
141. Ferrari, G. *et al.* Variola virus genome sequenced from an eighteenth-century museum specimen supports the recent origin of smallpox. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20190572 (2020).
142. Truman, R. W. *et al.* Probable Zoonotic Leprosy in the Southern United States. *N. Engl. J. Med.* **364**, 1626–1633 (2011).
143. Yang, R. Plague: Recognition, treatment, and prevention. *J. Clin. Microbiol.* **56**, e01519-17 (2018).
144. World Health Organization. WHO Plague Fact Sheet. *Plague* Plague http://www.who.int/mediacentre/factsheets/fs267/en/

(2016).

145. Achtman, M. *et al.* Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14043–14048 (1999).
146. Chain, P. S. G. *et al.* Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13826–13831 (2004).
147. Hinnebusch, B. J. Evaluation of the role of the yersinia pestis plasminogen activator and other plasmid-encoded factors in temperature-dependent blockage of the flea. *J. Infect. Dis.* **178**, 1406–1415 (1998).
148. Hinnebusch, B. J. *et al.* Role of Yersinia murine toxin in survival of Yersinia pestis in the midgut of the flea vector. *Science* **296**, 733–735 (2002).
149. Sun, Y. C., Jarrett, C. O., Bosio, C. F. & Hinnebusch, B. J. Retracing the evolutionary path that led to flea-borne transmission of Yersinia pestis. *Cell Host Microbe* **15**, 578–586 (2014).
150. Zimbler, D. L., Schroeder, J. A., Eddy, J. L. & Lathem, W. W. Early emergence of Yersinia pestis as a severe respiratory pathogen. *Nat. Commun.* **6**, 1–10 (2015).
151. Derbise, A. & Carniel, E. YpfΦ: A filamentous phage acquired by Yersinia pestis. *Front. Microbiol.* **5**, 701 (2014).
152. Derbise, A. *et al.* A horizontally acquired filamentous

phage contributes to the pathogenicity of the plague bacillus. *Mol. Microbiol.* **63**, 1145–1157 (2007).

153. Sebbane, F., Devalckenaere, A., Foulon, J., Carniel, E. & Simonet, M. Silencing and reactivation of urease in Yersinia pestis is determined by one G residue at a specific position in the ureD gene. *Infect. Immun.* **69**, 170–176 (2001).
154. Vogler, A. J., Keim, P. & Wagner, D. M. *A review of methods for subtyping Yersinia pestis: From phenotypes to whole genome sequencing*. *Infection, Genetics and Evolution* vol. 37 (2016).
155. CDC. Plague: Maps and Statistics. https://www.cdc.gov/plague/maps/index.html (2019).
156. Kutyrev, V. V. *et al.* Phylogeny and classification of Yersinia pestis through the lens of strains from the Plague Foci of commonwealth of independent states. *Front. Microbiol.* **9**, 1106 (2018).
157. Papagrigorakis, M. J., Yapijakis, C., Synodinos, P. N. & Baziotopoulou-Valavani, E. DNA examination of ancient dental pulp incriminates typhoid fever as a probable cause of the Plague of Athens. *Int. J. Infect. Dis.* **10**, 206–214 (2006).
158. Shapiro, B., Rambaut, A. & Gilbert, M. T. P. No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). *International Journal of Infectious Diseases* vol. 10 334–335 (2006).
159. Sabbatani, S. & Fiorino, S. La peste antonina e il declino

dell'Impero Romano. Ruolo della guerra partica e della guerra marcomannica tra il 164 e il 182 d.C. nella diffusione del contagio. *Infez. Med.* **17**, 261–275 (2009).

160. Demeure, C. E. *et al.* Yersinia pestis and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination, and diagnostics. *Genes Immun.* **20**, 357–370 (2019).
161. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* **176**, 295-305.e10 (2019).
162. Andrades Valtueña, A. *et al.* The Stone Age Plague and Its Persistence in Eurasia. *Curr. Biol.* **27**, 3683-3691.e8 (2017).
163. Yu, H. *et al.* Paleolithic to Bronze Age Siberians Reveal Connections with First Americans and across Eurasia. *Cell* **181**, 1232–1245 (2020).
164. Spyrou, M. A. *et al.* Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. *Nat. Commun.* **9**, 2234 (2018).
165. Sarris, P. The Justinianic Plague: Origins and effects. *Contin. Chang.* **17**, 169–182 (2002).
166. Harbeck, M. *et al.* Yersinia pestis DNA from Skeletal Remains from the 6th Century AD Reveals Insights into Justinianic Plague. *PLoS Pathog.* **9**, e1003349 (2013).
167. Wagner, D. M. *et al.* Yersinia pestis and the Plague of Justinian 541-543 AD: A genomic analysis. *Lancet Infect. Dis.* **14**, 319–326 (2014).

168. De Barros Damgaard, P. *et al.* 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018).
169. Keller, M. *et al.* Ancient Yersinia pestis genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12363–12372 (2019).
170. Yue, R. P. H., Lee, H. F. & Wu, C. Y. H. Trade routes and plague transmission in pre-industrial Europe. *Scientific Reports* vol. 7 12973 (2017).
171. DeWitte, S. N. Age patterns of mortality during the Black Death in London, A.D. 1349-1350. *J. Archaeol. Sci.* **37**, 3394–3400 (2010).
172. Galanaud, P., Galanaud, A., Giraudoux, P. & Labesse, H. Mortality and demographic recovery in early post-black death epidemics: Role of recent emigrants in medieval Dijon. *PLoS One* **15**, e0226420 (2020).
173. Wood, J. W., Ferrell, R. J. & Dewitte-Aviña, S. N. The Temporal Dynamics of the Fourteenth-Century Black Death: New Evidence from English Ecclesiastical Records. *Hum. Biol.* **75**, 427–448 (2003).
174. Waldron, H. A. Are plague pits of particular use to palaeoepidemiologists? *Int. J. Epidemiol.* **30**, 104–108 (2001).
175. Cui, Y. *et al.* Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 577–582 (2013).

176. Haensch, S. *et al.* Distinct clones of Yersinia pestis caused the black death. *PLoS Pathog.* **6**, e1001134 (2010).
177. Seifert, L. *et al.* Genotyping Yersinia pestis in historical plague: Evidence for long-term persistence of Y. pestis in Europe from the 14th to the 17th century. *PLoS One* **11**, e0145194 (2016).
178. Spyrou, M. A. *et al.* Historical Y. pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe* **19**, 874–881 (2016).
179. Namouchi, A. *et al.* Integrative approach using Yersinia pestis genomes to revisit the historical landscape of plague during the Medieval Period. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11790–E11797 (2018).
180. Susat, J. *et al.* Yersinia pestis strains from Latvia show depletion of the pla virulence gene at the end of the second plague pandemic. *Sci. Rep.* **10**, 1–10 (2020).
181. Bos, K. I. *et al.* Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *Elife* **5**, e12994 (2016).
182. Spyrou, M. A. *et al.* Phylogeography of the second plague pandemic revealed through analysis of historical Yersinia pestis genomes. *Nat. Commun.* **10**, (2019).
183. Ford, D. C., Joshua, G. W. P., Wren, B. W. & Oyston, P. C. F. The importance of the magnesium transporter MgtB for virulence of Yersinia pseudotuberculosis and

Yersinia pestis. *Microbiology* **160**, 2710–2717 (2014).

184. Rang, C. *et al.* Dual role of the MgtC virulence factor in host and non-host environments. *Mol. Microbiol.* **63**, 605–622 (2007).
185. Morelli, G. *et al.* Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nat. Genet.* **42**, 1140–1143 (2010).
186. Vogler, A. J. *et al.* Temporal phylogeography of Yersinia pestis in Madagascar: Insights into the long-term maintenance of plague. *PLoS Negl. Trop. Dis.* **11**, e0005887 (2017).
187. Vogler, A. J. *et al.* A decade of plague in Mahajanga, Madagascar: Insights into the global maritime spread of pandemic plague. *MBio* **4**, e00623-12 (2013).
188. Riehm, J. M. *et al.* Diverse genotypes of Yersinia pestis caused plague in Madagascar in 2007. *PLoS Negl. Trop. Dis.* **9**, e0003844 (2015).
189. Ollich Castanyé, I. La necròpolis medieval de l'Esquerda (segles VIII-XIV dC). Cronologia i noves perspectives de recerca. in *Arqueologia funerària al nord-est peninsular (segles VI-XII),* 13–24 (2012).
190. Mondal, M. *et al.* Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.* **48**, 1066–1070 (2016).
191. Raghavan, M. *et al.* Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature* **505**, 87–91 (2014).

192. Moreno-Mayar, J. V. *et al.* Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018).
193. Moreno-Mayar, J. V. *et al.* Early human dispersals within the Americas. *Science* **362**, (2018).
194. Posth, C. *et al.* Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185-1197.e22 (2018).
195. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
196. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
197. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
198. Scheib, C. L. *et al.* Ancient human parallel lineages within North America contributed to a coastal expansion. *Science* **360**, 1024–1027 (2018).
199. Cuervo, B. La conquista y colonización española de América. *Hist. Digit. Colab. con la Fund. ARTHIS* **28**, 103–149 (2016).
200. Guerra, F. Origen de las epidemias en la conquista de América. *Quinto Cent.* **14**, 43–51 (1988).
201. Somolinos d'Ardois, G. La epidemia de Cocoliztli de 1545 señalada en un codice. *Trib. Medica* **15**, 85 (1970).
202. Malvido, E. & Viesca, C. Epidemia de cocoliztli de 1576.

in *Historias* vol. 11 26–33 (1985).

203. Marr, J. S. & Kiracofe, J. B. Was the Huey Cocoliztli a haemorrhagic fever? *Med. Hist.* **44**, 363–388 (2000).
204. Crosby, A. W. & Cook, N. D. *Born to Die: Disease and New World Conquest, 1492-1650*. *Population and Development Review* vol. 24 (1998).
205. de Sahagun, B. *Historia General de las Cosas de Nueva España. (1545-1580). Available at: https://www.wdl.org/es/item/10096/*. doi:10.1017/cbo9780511792892.
206. Center for Disease Control and Prevention. Smallpox; Clinical Disease. https://www.cdc.gov/smallpox/clinicians/clinical-disease.html (2016).
207. Nobles, G. H. & Diamond, J. *Guns, Germs, and Steel: The Fates of Human Societies*. *Environmental History* vol. 4 (Norton, 1999).
208. Martin, D. L. & Goodman, A. H. Health conditions before Columbus: paleopathology of native North Americans. *West. J. Med.* **176**, 65–68 (2002).
209. Castro, M. M. *et al.* Thoracic aortic aneurysm in a pre-Columbian (210 BC) inhabitant of Northern Chile: Implications for the origins of syphilis. *Int. J. Paleopathol.* **13**, 20–26 (2016).
210. Mendieta, fray J. de. *Historia eclesiástica indiana: obra escrita a fines del siglo XVI*. (Forgotten Books, 2018). doi:036625443X.

211. McAfee, B. & Barlow, R. Anales de San Gregorio Acapulco 1520-1606. *Tlalocan* **3**, 103–141 (1952).
212. Peñafiel, A. *Anales de Tecamachalco: Crónica local y colonial en idioma nahuatl, 1398 y 1590*. (Of. tip. de la Secretar{\'\i}a de Fomento, 1903).
213. Paso y Troncoso, F. del. *Relaciones geográficas de la Diócesis de Oaxaca; manuscritos de la Real Academia de la Historia de Madrid y del Archivo de Indias en Sevilla. Años 1579-1581*. (Archivo General de Historia, Real Academia de la Indias, 1905).
214. Acuña, R. *Relaciones geográficas del siglo XVI: Tlaxcala*. (Universidad Nacional Autonoma de Mexico, 1984).
215. Zinsser, H. *Book Review Rats, Lice and History . Being a study in Biography, which, after twelve preliminary chapters indispensable for the preparation of the Lay Reader, deals with the life history of Typhus Fever. Hans Zinsser. 301 pp. Boston: Little, Brown & Compa*. *New England Journal of Medicine* vol. 213 (Printed and Pub. for the Atlantic Monthly Press by Little, Brown, 1935).
216. Diffie, B. W. & Gibson, C. *The Aztecs Under Spanish Rule: A History of the Indians of the Valley of Mexico 1519-1810*. *Ethnohistory* vol. 12 (Stanford University Press, 1965).
217. Acuna-Soto, R., Calderon Romero, L. & Maguire, J. H. Large epidemics of hemorrhagic fevers in Mexico 1545-1815. *Am. J. Trop. Med. Hyg.* **62**, 733–739 (2000).

218. Acuna-Soto, R., Stahle, D. W., Therrell, M. D., Griffin, R. D. & Cleaveland, M. K. When half of the population died: The epidemic of hemorrhagic fevers of 1576 in Mexico. *FEMS Microbiol. Lett.* **240**, 1–5 (2004).
219. Zhou, Z. *et al.* Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* **28**, 2420-2428.e10 (2018).
220. Stanaway, J. D. *et al.* The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Infect. Dis.* **19**, 369–381 (2019).
221. Ochiai, R. L. *et al.* Salmonella paratyphi A rates, Asia. *Emerg. Infect. Dis.* **11**, 1764–1766 (2005).
222. Crump, J. A., Ram, P. K., Gupta, S. K., Miller, M. A. & Mintz, E. D. Part I. Analysis of data gaps pertaining to Salmonella enterica serotype Typhi infections in low and medium human development index countries, 1984-2005. *Epidemiol. Infect.* **136**, 436–448 (2008).
223. World Health Organisation. Typhoid vaccine: WHO position paper - March 2018. *Wkly. Epidemiol. Rec.* **13**, 153–172 (2018).
224. Gunn, J. S. *et al.* Salmonella chronic carriage: Epidemiology, diagnosis, and gallbladder persistence. *Trends Microbiol.* **22**, 648–655 (2014).
225. Various Authors. *Cartas de Indias*. (Ministerio de Fomento, 1877).

226. Guzmán-Solís, A. *et al.* Ancient viral genomes reveal introduction of HBV and B19V to Mexico during the transatlantic slave trade. *bioRxiv* 2020.06.05.137083 (2020) doi:10.1101/2020.06.05.137083.
227. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5**, 3346 (2014).
228. Gilabert, A. *et al.* Plasmodium vivax-like genome sequences shed new insights into Plasmodium vivax biology and evolution. *PLoS Biol.* **16**, e2006035 (2018).
229. Ashley, E. A., Pyae Phyo, A. & Woodrow, C. J. Malaria. *Lancet* **391**, 1608–1621 (2018).
230. World Health Organization. *World Malaria Report: 20 years of global progress and challenges*. *World Health* vol. WHO/HTM/GM (2020).
231. Laporta, G. Z. *et al.* Malaria vectors in South America: Current and future scenarios. *Parasites and Vectors* **8**, 426 (2015).
232. Rénia, L. *et al.* Cerebral malaria Mysteries at the blood-brain barrier. *Virulence* **3**, 193–201 (2012).
233. Armah, H. *et al.* Cytokines and adhesion molecules expression in the brain in human cerebral malaria. *Int. J. Environ. Res. Public Health* **2**, 123–131 (2005).
234. Haldar, K., Bhattacharjee, S. & Safeukui, I. Drug resistance in Plasmodium. *Nat. Rev. Microbiol.* **16**, 156–170 (2018).
235. Price, R. N. *et al.* The pfmdr1 gene is associated with a multidrug-resistant phenotype in Plasmodium falciparum

from the western border of Thailand. *Antimicrob. Agents Chemother.* **43**, 2943–2949 (1999).

236. Reed, M. B., Saliba, K. J., Caruana, S. R., Kirk, K. & Cowman, A. F. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium falciparum. *Nature* **403**, 906–909 (2000).
237. Sidhu, A. B. S., Valderramos, S. G. & Fidock, D. A. pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in Plasmodium falciparum. *Mol. Microbiol.* **57**, 913–926 (2005).
238. Pickard, A. L. *et al.* Resistance to antimalarials in Southeast Asia and genetic polymorphisms in pfmdr1. *Antimicrob. Agents Chemother.* **47**, 2418–2423 (2003).
239. Dahlström, S. *et al.* Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *J. Infect. Dis.* **200**, 1456–1464 (2009).
240. Foote, S. J. *et al.* Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in Plasmodium falciparum. *Nature* **345**, 255–258 (1990).
241. Callaghan, P. S., Siriwardana, A., Hassett, M. R. & Roepe, P. D. Plasmodium falciparum chloroquine resistance transporter (PfCRT) isoforms PH1 and PH2 perturb vacuolar physiology. *Malar. J.* **15**, 186 (2016).
242. Durrand, V. *et al.* Variations in the sequence and expression of the Plasmodium falciparum chloroquine

resistance transporter (Pfcrt) and their relationship to chloroquine resistance in vitro. *Mol. Biochem. Parasitol.* **136**, 273–285 (2004).

243. Miotto, O. *et al.* Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nat. Genet.* **47**, 226–234 (2015).
244. Pelleau, S. *et al.* Adaptive evolution of malaria parasites in French Guiana: Reversal of chloroquine resistance by acquisition of a mutation in pfcrt. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11672–11677 (2015).
245. Mu, J. *et al.* Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol. Microbiol.* **49**, 977–989 (2003).
246. Setthaudom, C. *et al.* Role of Plasmodium falciparum chloroquine resistance transporter and multidrug resistance 1 genes on in vitro chloroquine resistance in isolates of Plasmodium falciparum from Thailand. *Am. J. Trop. Med. Hyg.* **85**, 606–611 (2011).
247. Ross, L. S. *et al.* Emerging Southeast Asian PfCRT mutations confer Plasmodium falciparum resistance to the first-line antimalarial piperaquine. *Nat. Commun.* **9**, 25–28 (2018).
248. Echeverry, D. F. *et al.* Short report: Polymorphisms in the pfcrt and pfmdr1 genes of Plasmodium falciparum and in vitro susceptibility to amodiaquine and desethylamodiaquine. *Am. J. Trop. Med. Hyg.* **77**, 1034–1038 (2007).

249. Dahlström, S., Veiga, M. I., Mårtensson, A., Björkman, A. & Gil, J. P. Polymorphism in Pfmrp1 (Plasmodium falciparum multidrug resistance protein 1) amino acid 1466 associated with resistance to sulfadoxine-pyrimethamine treatment. *Antimicrob. Agents Chemother.* **53**, 2553–2556 (2009).
250. Gupta, B. *et al.* Plasmodium falciparum multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east Myanmar. *J. Antimicrob. Chemother.* **69**, 2110–2117 (2014).
251. Veiga, M. I. *et al.* Novel polymorphisms in plasmodium falciparum ABC transporter genes are associated with major ACT Antimalarial drug resistance. *PLoS One* **6**, e20212 (2011).
252. Boullé, M. *et al.* Artemisinin-resistant Plasmodium falciparum K13 mutant alleles, Thailand-Myanmar border. *Emerg. Infect. Dis.* **22**, 1503–1505 (2016).
253. Kobasa, T. *et al.* Emergence and spread of kelch13 mutations associated with artemisinin resistance in plasmodium falciparum parasites in 12 Thai provinces from 2007 to 2016. *Antimicrob. Agents Chemother.* **62**, (2018).
254. Mishra, N. *et al.* Emerging polymorphisms in falciparum Kelch 13 gene in Northeastern region of India. *Malar. J.* **15**, 4–9 (2016).
255. Ye, R. *et al.* Distinctive origin of artemisinin-resistant

Plasmodium falciparum on the China-Myanmar border. *Sci. Rep.* **6**, 1–9 (2016).

256. Zaw, M. T., Emran, N. A. & Lin, Z. Updates on k13 mutant alleles for artemisinin resistance in Plasmodium falciparum. *J. Microbiol. Immunol. Infect.* **51**, 159–165 (2018).
257. Happi, C. T. *et al.* Polymorphisms in Plasmodium falciparum dhfr and dhps genes and age related in vivo sulfadoxine-pyrimethamine resistance in malaria-infected patients from Nigeria. *Acta Trop.* **95**, 183–193 (2005).
258. Rottmann, M. *et al.* Spiroindolones, a potent compound class for the treatment of malaria. *Science* **329**, 1175–1180 (2010).
259. Vaidya, A. B. *et al.* Pyrazoleamide compounds are potent antimalarials that target Na+ homeostasis in intraerythrocytic Plasmodium falciparum. *Nat. Commun.* **5**, 1–10 (2014).
260. Wang, Z. *et al.* Genome-wide association analysis identifies genetic loci associated with resistance to multiple antimalarials in Plasmodium falciparum from China-Myanmar border. *Sci. Rep.* **6**, 1–12 (2016).
261. Takala-Harrison, S. *et al.* Genetic loci associated with delayed clearance of plasmodium falciparum following artemisinin. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 240–245 (2013).
262. Shetty, A. C. *et al.* Genomic structure and diversity of

Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns. *Nat. Commun.* **10**, 2665 (2019).

263. Marciniak, S. *et al.* Plasmodium falciparum malaria in 1st–2nd century CE southern Italy. *Curr. Biol.* **26**, R1220–R1222 (2016).
264. Amato, R. *et al.* Genomic epidemiology of artemisinin resistant malaria. *Elife* **5**, e08714 (2016).
265. Yalcindag, E. *et al.* Multiple independent introductions of Plasmodium falciparum in South America. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 511–516 (2012).
266. Rodrigues, P. T. *et al.* Human migration and the spread of malaria parasites to the New World. *Sci. Rep.* **8**, 1–13 (2018).
267. Baird, J. K., Valecha, N., Duparc, S., White, N. J. & Price, R. N. Diagnosis and treatment of plasmodium vivax malaria. *Am. J. Trop. Med. Hyg.* **95**, 35–51 (2016).
268. Cogswell, F. B. The hypnozoite and relapse in primate malaria. *Clin. Microbiol. Rev.* **5**, 26–35 (1992).
269. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* **48**, 953–958 (2016).
270. Menegon, M., Majori, G. & Severini, C. Genetic variations of the Plasmodium vivax dihydropteroate synthase gene. *Acta Trop.* **98**, 196–199 (2006).
271. Hawkins, V. N. *et al.* Assessment of the origins and spread of putative resistance-conferring mutations in

Plasmodium vivax dihydropteroate synthase. *Am. J. Trop. Med. Hyg.* **81**, 348–355 (2009).

272. Korsinczky, M. *et al.* Sulfadoxine resistance in Plasmodium vivax is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site. *Antimicrob. Agents Chemother.* **48**, 2214–2222 (2004).
273. Ganguly, S., Saha, P., Chatterjee, M. & Maji, A. K. Prevalence of polymorphisms in antifolate drug resistance molecular marker genes pvdhfr and pvdhps in clinical isolates of plasmodium vivax from Kolkata, India. *Antimicrob. Agents Chemother.* **58**, 196–200 (2014).
274. Imwong, M. *et al.* Novel point mutations in the dihydrofolate reductase gene of Plasmodium vivax: Evidence for sequential selection by drug pressure. *Antimicrob. Agents Chemother.* **47**, 1514–1521 (2003).
275. Eldin De Pécoulas, P., Basco, L. K., Tahar, R., Ouatas, T. & Mazabraud, A. Analysis of the Plasmodium vivax dihydrofolate reductase-thymidylate synthase gene sequence. *Gene* **211**, 177–185 (1998).
276. Leartsakulpanich, U. *et al.* Molecular characterization of dihydrofolate reductase in relation to antifolate resistance in Plasmodium vivax. *Mol. Biochem. Parasitol.* **119**, 63–73 (2002).
277. Huang, B. *et al.* Molecular surveillance of pvdhfr, pvdhps, and pvmdr-1 mutations in Plasmodium vivax isolates from Yunnan and Anhui provinces of China. *Malar. J.* **13**,

346 (2014).

278. Suwanarusk, R. *et al.* Chloroquine resistant Plasmodium vivax: In vitro characterisation and association with molecular polymorphisms. *PLoS One* **2**, 1–9 (2007).
279. Brega, S. *et al.* Identification of the Plasmodium vivax mdr-like gene (pvmdr1) and analysis of single-nucleotide polymorphisms among isolates from different areas of endemicity. *J. Infect. Dis.* **191**, 272–277 (2005).
280. Barnadas, C. *et al.* Plasmodium vivax resistance to chloroquine in Madagascar: Clinical efficacy and polymorphisms in pvmdr1 and pvcrt-o genes. *Antimicrob. Agents Chemother.* **52**, 4233–4240 (2008).
281. Orjuela-Sánchez, P. *et al.* Analysis of single-nucleotide polymorphisms in the crt-o and mdr1 genes of Plasmodium vivax among chloroquine-resistant isolates from the Brazilian Amazon region. *Antimicrob. Agents Chemother.* **53**, 3561–3564 (2009).
282. Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**, 1669–1679 (2000).
283. Van Dorp, L. *et al.* Plasmodium vivax Malaria Viewed through the Lens of an Eradicated European Strain. *Mol. Biol. Evol.* **37**, 773–785 (2020).
284. Guerra, F. The European-American. *Hist. Philos. Life Sci.* **15**, 313–327 (1993).
285. Giacani, L. & Lukehart, S. A. The endemic

treponematoses. *Clin. Microbiol. Rev.* **27**, 89–115 (2014).

286. O'Byrne, P. & Macpherson, P. Syphilis. *BMJ* **365**, l4159 (2019).
287. World Health Organization. *WHO Guidelines for the Treatment of Treponema pallidum (Syphilis)*. (NCBI Bookshelf, 2016).
288. Marks, M., Solomon, A. W. & Mabey, D. C. Endemic treponemal diseases. *Trans. R. Soc. Trop. Med. Hyg.* **108**, 601–607 (2014).
289. Stamm, L. V. Syphilis: Antibiotic treatment and resistance. *Epidemiol. Infect.* **143**, 1567–1574 (2015).
290. Mitjà, O. *et al.* Global epidemiology of yaws: A systematic review. *Lancet Glob. Heal.* **3**, e324–e331 (2015).
291. Kojima, N. & Klausner, J. D. An Update on the Global Epidemiology of Syphilis. *Curr. Epidemiol. Reports* **5**, 24–38 (2018).
292. Domino, E. F. *Neurobiology of phencyclidine--an update. NIDA research monograph* vol. (suppl. 21 (U.S. Government Printing Office, 1978).
293. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545–1602 (2016).
294. De Melo, F. L., De Mello, J. C. M., Fraga, A. M., Nunes,

K. & Eggers, S. Syphilis at the crossroad of phylogenetics and paleopathology. *PLoS Negl. Trop. Dis.* **4**, e575–e575 (2010).

295. Tampa, M., Sarbu, I., Matei, C., Benea, V. & Georgescu, S. R. Brief history of syphilis. *J. Med. Life* **7**, 4–10 (2014).
296. Henneberg, R. J. & Henneberg, M. Possible occurrence of treponematosis in the ancient Greek colony of Metaponto. *Am J Phys Anthr.* **16**, 107–108 (1995).
297. Rissech, C. *et al.* A Roman Skeleton with Possible Treponematosis in the North-East of the Iberian Peninsula: A Morphological and Radiological Study. *Int. J. Osteoarchaeol.* **23**, 651–663 (2013).
298. Mays, S., Crane-Kramer, G. & Bayliss, A. Two probable cases of treponemal disease of medieval date from England. *Am. J. Phys. Anthropol.* **120**, 133–143 (2003).
299. Schuenemann, V. J. *et al.* Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Negl. Trop. Dis.* **12**, e0006447–e0006447 (2018).
300. Majander, K. *et al.* Ancient Bacterial Genomes Reveal a High Diversity of Treponema pallidum Strains in Early Modern Europe. *Curr. Biol.* **30**, 3788-3803.e10 (2020).
301. Furin, J., Cox, H. & Pai, M. Tuberculosis. *Lancet* **393**, 1642–1656 (2019).
302. Golden, M. P. & Vikram, H. R. Extrapulmonary tuberculosis: An overview. *Am. Fam. Physician* **72**, 1761–1768 (2005).

303. Churchyard, G. *et al.* What We Know about Tuberculosis Transmission: An Overview. *J. Infect. Dis.* **216**, S629–S635 (2017).
304. Bastos, M. L., Lan, Z. & Menzies, D. An updated systematic review and meta-analysis for treatment of multidrug-resistant tuberculosis. *Eur. Respir. J.* **49**, 1600803 (2017).
305. World Health Organization. *Treatment of Tuberculosis: Guidelines*. (World Health Organization, 2010).
306. World Health Organization. *Global Tuberculosis Report*. *Blood* (2020). doi:978 92 4 156450 2.
307. Daniel, T. M. *The Bioarchaeology of Tuberculosis: A Global View on a Reemerging Disease. The American Journal of Tropical Medicine and Hygiene* vol. 73 (University Press of Florida, 2005).
308. Comas, Ĩ. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
309. Sabin, S. *et al.* A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biol.* **21**, 201 (2020).
310. Hershberg, R. *et al.* High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biol.* **6**, 2658–2671 (2008).
311. Wirth, T. *et al.* Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog.* **4**,

e1000160–e1000160 (2008).

312. Kay, G. L. *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* **6**, 6717 (2015).
313. Müller, R., Roberts, C. A. & Brown, T. A. Genotyping of ancient Mycobacterium tuberculosis strains reveals historic genetic diversity. *Proc. R. Soc. B Biol. Sci.* **281**, 20133236 (2014).
314. Fischer, M. Leprosy – an overview of clinical features, diagnosis, and treatment. *JDDG - J. Ger. Soc. Dermatology* **15**, 801–827 (2017).
315. Chehl, S., Job, C. K. & Hastings, R. C. Transmission of leprosy in nude mice. *Am. J. Trop. Med. Hyg.* **34**, 1161–1166 (1985).
316. Krause-Kyora, B. *et al.* Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat. Commun.* **9**, 1569 (2018).
317. Zhang, F. *et al.* Evidence for an association of HLA-DRB115 and DRB109 with leprosy and the impact of DRB109 on disease onset in a Chinese Han population. *BMC Med. Genet.* **10**, 133 (2009).
318. Liu, H. *et al.* Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. *Nat. Genet.* **47**, 267–271 (2015).
319. Abel, L. *et al.* Susceptibility to leprosy is linked to the human NRAMP1 gene. *J. Infect. Dis.* **177**, 133–145 (1998).

320. Lockwood, D. N. J. & Kumar, B. Treatment of leprosy. *Br. Med. J.* **328**, 1447–1448 (2004).
321. Cambau, E. *et al.* Antimicrobial resistance in leprosy: results of the first prospective open survey conducted by a WHO surveillance network for the period 2009–15. *Clin. Microbiol. Infect.* **24**, 1305–1310 (2018).
322. Guinto, R. S., Doull, J. A. & de Guia, L. Mortality of persons with leprosy prior to sulfone therapy, Cordova and Talisay, Cebu, Philippines. *Int. J. Lepr.* **22**, 273–284 (1954).
323. Hulse, E. V. Leprosy and Ancient Egypt. *Lancet* **300**, 1024–1025 (1972).
324. Bloomfield, M. *Hymns of the Atharva-Veda*. (BiblioBazaar, 2009).
325. Salisbury, J. R. The Cambridge Encyclopedia of Human Paleopathology. *J. Clin. Pathol.* **51**, 879 (1998).
326. Neukamm, J. *et al.* 2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals. *BMC Biol.* **18**, 108 (2020).
327. Monot, M. *et al.* On the origin of leprosy. *Science* **308**, 1040–1042 (2005).
328. Deps, P. D. *et al.* Contact with armadillos increases the risk of leprosy in Brazil: a case control study. *Indian J. Dermatol. Venereol. Leprol.* **74**, 338–342 (2008).
329. Patterson, K. B. & Runge, T. Smallpox and the Native American. *Am. J. Med. Sci.* **323**, 216–222 (2002).
330. Mühlemann, B. *et al.* Diverse variola virus (smallpox)

strains were widespread in northern Europe in the Viking Age. *Science* **369**, (2020).

331. Behbehani, A. M. The smallpox story: Life and death of an old disease. *Microbiol. Rev.* **47**, 455–509 (1983).
332. Acuna-Soto, R., Stahle, D. W., Cleaveland, M. K. & Therrell, M. D. Megadrought and megadeath in 16th century Mexico. *Emerg. Infect. Dis.* **8**, 360–362 (2002).
333. Monguiló, E., Hernandez, J. & Molinas, R. *Intervenció arqueològica a l'espai delimitat pels carrers d'Espronceda, ronda de Sant Martí, carrer de Josep Soldevila i passeig de la Verneda (Ave Sagrera)*. (2012).
334. Luis, J. & Moya, B. Sociedad y peste en la Barcelona de 1651. *Manuscrits Rev. d'història Mod.* **0**, 255–282 (1990).
335. Rohland, N. & Hofreiter, M. Ancient dna extraction from bones and teeth. *Nat. Protoc.* **2**, 1756–1762 (2007).
336. Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42**, 343–352 (2007).
337. Yang, D. Y. & Watt, K. Contamination controls when preparing archaeological remains for ancient DNA analysis. *J. Archaeol. Sci.* **32**, 331–336 (2005).
338. Gansauge, M. T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8**, 737–748 (2013).
339. Sequencing, H. & Kit, P. O. Multiplexed Sequencing with the Illumina Genome Analyzer System. *Analyzer* vol.

2010 1–4 http://www.ncbi.nlm.nih.gov/pubmed/20516186 (2008).

340. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6**, 287–303 (2013).
341. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2009).
342. Andrews S.; S Bittencourt a. FastQC: a quality control tool for high throughput sequence data – ScienceOpen. *Babraham Institute* citeulike-article-id:11583827%0Ahttp://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
343. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
344. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
345. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
346. Burrows, M. & Wheeler, D. *A block-sorting lossless data compression algorithm*. *Algorithm, Data Compression* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.

1.141.5254 (1994) doi:10.1.1.37.6774.

347. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
348. Kircher, M. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol. Biol.* **840**, 197–228 (2012).
349. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. vol. 00 http://arxiv.org/abs/1303.3997 (2013).
350. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
351. Broad Institute. Picard. http://broadinstitute.github.io/picard/.
352. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
353. Schmidt, S. Measuring absorptive capacity. *Proc. Int. Conf. Intellect. Capital, Knowl. Manag. Organ. Learn.* **20**, 254–260 (2009).
354. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
355. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2229–

2234 (2014).

356. Team, R. R: A language and environment for statistical computing (Version 3.4. 2)[Computer software]. *Vienna, Austria: R Foundation for Statistical Computing* (2017).
357. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. in *Bioinformatics* vol. 29 1682–1684 (2013).
358. Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E. & Orlando, L. mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153–2155 (2011).
359. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013).
360. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, (2015).
361. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
362. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
363. Schiffels, S. & Peltzer, A. Sequence Tools. (2019).

364. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-33 (2013).
365. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–501 (2011).
366. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017) doi:10.1101/201178.
367. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
368. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
369. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
370. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
371. Li, H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
372. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).

373. Martiniano, R. *et al.* The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* **13**, e1006852 (2017).
374. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
375. van Oven, M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. Int. Genet. Suppl. Ser.* **5**, e392–e394 (2015).
376. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
377. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
378. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
379. Alexander, D. H. & Lange, K. Admixture 1 . 23 Software Manual. 3–4 (2013).
380. Valero-Mora, P. M. *ggplot2: Elegant Graphics for Data Analysis* . *Journal of Statistical Software* vol. 35 (Springer-Verlag New York, 2010).
381. Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P. & Ramachandran, S. Pong: Fast analysis and visualization of latent clusters in population genetic data.

Bioinformatics **32**, 2817–2823 (2016).

382. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).
383. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3321–3323 (1973).
384. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
385. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
386. Lipson, M. *et al.* Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol. Biol. Evol.* **30**, 1788–1802 (2013).
387. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
388. Raghavan, M. *et al.* Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature* **505**, 87–91 (2014).
389. Peter, B. M. Admixture, population structure, and f-statistics. *Genetics* **202**, 1485–1501 (2016).
390. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
391. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–

376 (1981).

392. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
393. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
394. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
395. Stamatakis, A. RAxML. *Manual/tutorial* 1–5 (2014).
396. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
397. Arenas, M. & Posada, D. The effect of recombination on the reconstruction of ancestral sequences. *Genetics* **184**, 1133–1139 (2010).
398. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
399. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
400. Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst. Zool.* **20**, 406 (1971).

401. Crispell, J., Balaz, D. & Gordon, S. V. Homoplasyfinder: A simple tool to identify homoplasies on a phylogeny. *Microb. Genomics* **5**, e000245 (2019).
402. Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. Measurably evolving populations. *Trends Ecol. Evol.* **18**, 481–488 (2003).
403. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134–e134 (2018).
404. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).
405. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
406. Madden, T. The BLAST Sequence Analysis Tool. in *The NCBI Handbook [Internet]* 1–10 (National Center for Biotechnology Information (USA), 2013).
407. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
408. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
409. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

410. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
411. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).
412. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. *Cell* **176**, 295–305 (2019).
413. Zhu, T. *et al.* An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597 (1998).
414. Marciniak, S., Herring, D. A., Sperduti, A., Poinar, H. N. & Prowse, T. L. A multi-faceted anthropological and genomic approach to framing Plasmodium falciparum malaria in Imperial period central-southern Italy (1st–4th c. CE). *J. Anthropol. Archaeol.* **49**, 210–224 (2018).
415. Piñar, G., Tafer, H., Schreiner, M., Miklas, H. & Sterflinger, K. Decoding the biological information contained in two ancient Slavonic parchment codices: an added historical value. *Environ. Microbiol.* **22**, 3218–3233 (2020).
416. Coto-Segura, C., Coto-Segura, P. & Santos-Juanes, J. The Skin of a Revolutionary. *Arch. Dermatol.* **147**, 539 (2011).
417. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: Confident and fast metagenomics

classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).

418. Morfopoulou, S. & Plagnol, V. Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics* **31**, 2930–2938 (2015).
419. Cash-Goldwasser, S. & Barry, M. *CDC Yellow Book 2018: Health Information for International Travel*. *Clinical Infectious Diseases* vol. 66 (2018).
420. De Zulueta, J. The end of malaria in Europe: An eradication of the disease by control measures. *Parassitologia* **40**, 245–246 (1998).
421. Jones, W. H. S., Ross, R. & Ellett, G. G. *Malaria: a neglected factor in the history of Greece and Rome*. (Bowes & Bowes, 1920).
422. De Zulueta, J. Malaria and Mediterranean history. *Parassitologia* **15**, 1–15 (1973).
423. Sallares, R., Bouwman, A. & Anderung, C. The spread of malaria to Southern Europe in antiquity: New approaches to old problems. *Med. Hist.* **48**, 311–328 (2004).
424. Sallares, R. & Gomzi, S. Biomolecular archaeology of malaria. *Ancient Biomolecules* vol. 3 195–213 (2001).
425. Cano, N. B., Gordillo, M. A. M. & Añón, R. B. Campañas sanitarias en españa frente al paludismo a partir de los trabajos publicados en dos revistas científicas: Medicina de los países cálidos y la medicina colonial(1929-1954). *Rev. Esp. Salud Publica* **90**, 1–13 (2016).

426. Achan, J. *et al.* Quinine, an old anti-malarial drug in a modern world: Role in the treatment of malaria. *Malar. J.* **10**, 144 (2011).
427. Bayon, H. P. The Medical Career of Jean-Paul Marat. *J. R. Soc. Med.* **39**, 39–44 (1945).
428. Cohen, J. H. L. & Cohen, E. L. Doctor marat and his skin. *Med. Hist.* **2**, 281–286 (1958).
429. Dotz, W. Jean Paul Marat. His life, cutaneous disease, death, and depiction by Jacques Louis David. *Am. J. Dermatopathol.* **1**, 247–250 (1979).
430. Jelinek, J. E. Jean-Paul Marat. The differential diagnosis of his skin disease. *Am. J. Dermatopathol.* **1**, 251–252 (1979).
431. Dale, P. M. *Medical Biographies. The Ailments of Thirty-Three Famous Person*. (Press, University of Oklahoma, 1952). doi:10.1086/399828.
432. Pathogenic yeasts. in *The yeast handbook* (eds. Ashbee, R. & Bignell, E.) 209–230 (Springer-Verlag, 2010).