Paleogenomics applied to the study of ancient infectious diseases: Tracing the signals of the eradicated European malaria

Pere Gelabert Xirinachs

TESI DOCTORAL UPF / 2018

DIRECTOR DE LA TESI

Dr. Carles Lalueza-Fox

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

A les qui han compartit cada moment
que m'ha dut fins aquí

El millor camí és el del llaurador,
o el del caminant, o el del somiador,
perquè per curts són grans,
arrelen profunds i moren.
Però tu, sàvia naturalesa,
saps que s'obriran nous camins
per sers que estimen la terra,
per sers que creen bellesa,
per sers que gaudeixen endins.
Vindrà dia de camins sense voreres
I tot el cel serà un camí d'estrelles

Pau Vallhonrat, 1981

Agraïments

El període del doctorat, com qualsevol altre etapa de la nostra vida comprèn de forma quasi indestriable experiències professionals i personals. M'agradaria esmentar-les totes ja que és en el conjunt que aquestes conformen el moment transcendent que recordem.

D'entrada haig d'agrair la tasca de formació i mentoratge que he rebut de l'Iñigo. Ell em va ensenyar molt del que després he aplicat de forma recurrent. Tanmateix No puc deixar d'agrair la feina, suport i moments compartits amb el Toni i el Manu. Moltes vegades fent més del que tocaria i aguantant no sempre el millor humor. Faig també extensiu l'agraïment per tots els moments viscuts a totes les persones de l'Institut de Biologia Evolutiva amb qui hem compartit moltes hores.

Agrair de forma especial i sincera a la Família Canicio per cedir-nos les preuades mostres de sang del doctor Ildefons Canicio, i amb elles una part del passat material de Sant Jaume d'Enveja, de les Terres de l'Ebre i del Paludisme a Catalunya. Agrair-los no només la cesio física de les mostres, àdhuc per permetre endinsar-me en un passat recent i desconegut al qual jo em sento especialment lligat, tant per la simpatia que sento per les Terres de l'Ebre com pel coneixement que m'ha aportat sobre quelcom del que en sóc proper. La meva família durant segles i fins a la generació dels meus avis ha llaurat la terra del Delta del Llobregat, exposant-se a la mateixa malària que jo he pogut estudiar en aquest treball, a aquesta malària, al tifus i al còlera. Aquest treball també m’ha fet recordar-los i endinsar-me a la vida d’aquells qui em precediren al Poble de les Febres.

Agrair al Dr. Carles Aranda la recerca de les mostres i el seu coneixement sobre mosquits. Gràcies per ensenyar-me Can Comas i la feina de control de mosquits i aprofundir en un camp totalment desconegut per mi. Agrair a la Dra. Assumpció Malgosa i a la Dra. Cristina Santos per endinsar-me al món de l’Antropologia Biològica. Segurament sense l’estada a la UAB res del que ha vingut després hagués estat possible. Gràcies als qui han permès aquesta tesi tant conceptualment com metologica. Gràcies al Prof. Sergi Civit pel seu paper en l’anàlisi estadístic, un fonament imprescindible. Agrair a la Lucy i a l’Adrien la paciència que han tingut i la feina que han fet contribuint a la tesis en via de les seves publicacions. Agrair també al Prof. Thomas Gilbert, al Prof. Francois Balloux i al Shyam els modelatges fets i la col·laboració imprescindible a les publicacions. També agrair a la resta de persones que han format part de les diverses publicacions i han permès la tesis.

Gràcies al meu director de Tesis, en Carles Lalueza-Fox. Gràcies per l'oportunitat de poder fer tot el que he fet durant aquests anys. Ha estat un no parar il·lusionant de projectes diferents. Puc dir que, malgrat el caos en alguns moments, tot el que he fet ha estat repte del qual n'he après moltíssim, i he disfrutat encara més amb la majoria dels projectes que m’han caigut a les mans. Gràcies, Carles per fer confiança i pel guiatge.

Donar gràcies a la meva família; a la meva mare, al meu pare i les meves germanes Joana i Júlia. Elles han hagut de suportar els meus moments de tensió i de vegades escoltar de forma estoica els meus assajos de presentacions. Els agraeixo la seva presència, l’afecte i l’humor. D’altra manera tot hauria pogut ser diferent però mai millor, ni més personal ni autèntic. Aquest agraïment també és extensiu a totes les persones per les que sento efecte i a les que m’han fet sentir el seu afecte.

Un agraïment enorme a l’Albert per tots els moments que hem passat lligats d’una corda escalant parets d’arreu (més o menys temps) o esquiant els cims del Pirineu, Gràcies pels magnífics dies a bord de cotxes explotant i furgonetes resseguint ports i planes amb el millor humor possible i amb estats d’atenció i consciència variables. I també al Josep, el Javi i tants altres companys de natura, muntanya i aventures arreu del món. L’evasió als espais més remots de la nostra geografía, posar cos i ment a prova, no sempre amb totes les garanties, ha estat quelcom fonamental per mantenir el meu estat d’ànim. Gràcies als meus imprescindibles del Prat, els qui mai no m’han fallat; Laura, Xavi, Marta, Aleix, Júlia, Albert. Els qui sempre estan disposats a sortir a sopar, a recórrer els nostres pobles i viles buscant imatges mentals i emocions. Gràcies també als altres del Prat, amb els qui he compartit quelcom més que amistat i ens hem trobat en projectes il·lusionants, decepcionants i molt costosos tant materialment com anímica, hem patit, hem treballat junts, hem exprimit l'ingeni i ho faríem tants cops com calgués. Gràcies a la Maria i a la Clara, bones amigues des de l'escola amb amb qui sempre és una alegria compartir estones.

Gràcies a les *Bioguais*: L'Helena, la Silvia, l'Anna, la Jèssica, la Marta i la Paula. Amb elles hem discutit sobre biologia, ètica, política, filosofia i també sobre coses més importants. Amb elles hem rigut i per sort ho continuem fent. Elles són el millor que m'enduc del meu grau de Biologia i ha estat un privilegi seguir al seu costat aquests anys.

I finalment gràcies a la Paula, per ser-hi i per aguantar. Han estat molts els moments en que t'he hagut de dir no, que he hagut de cancel·lar plans, que no tenia ganes de quedar o simplement no tenia temps. No ha estat fàcil i no ha estat curt però esperem que hagi valgut la pena, d'altra manera segurament no ho escriuria. Ho hem fet com millor hem sapigut i continuarem esforçar-nos-hi les vegades que calguin. Mai agrairé prou ni l'estabilitat que m'has donat ni tot el suport imprescindible que he rebut de part teva. Ens veurem al capvespre a la plaça, en ambients més foscos o a casa però hi seràs i jo també hi seré.

Resum

La malària és una malaltia infecciosa causada per diverses espècies de protozous del gènere *Plasmodium*, capaces d'infectar eritròcits humans. La malària és probablement la patologia infecciosa que ha causat més morts al llarg de tota la història de l'espècie humana. Encara avui dia és un problema de salut pública global, agreujat per la creixent aparició de soques de *Plasmodium* resistents als tractaments farmacològics. L'origen de la majoria d'espècies de *Plasmodium* és africà. Espècies com *P. vivax* s'han expandit arreu per mitjà de moviments migratoris complexes, els quals romanen encara parcialment desconeguts degut a la manca de seqüències de *Plasmodium* europeus. Aquesta expansió probablement originada en el Neolític, ha seguit sempre moviments migratoris humans, deixant variants de resistència al Plasmodium en les poblacions humanes que hi han estat exposades. En aquest treball es presenta la seqüència de les soques europees de *P. vivax* i *P. falciparum* que han estat usades per traçar els moviments migratoris dels patògens així com per datar l'expansió del *P. vivax*. No gens menys el genotipat de variants de resistència en individus europeus antics mostra un impacte genètic limitat de la malària, suggerint amb una introducció recent del *Plasmodium* al continent.

Abstract

Malaria is an infectious disease caused by several protozoa species of the *Plasmodium* genus, capable to infect human erythrocytes. Malaria is probably the infectious pathology responsible of the larger amount of deaths among all human history. Still nowadays it is a major public health concern, which is aggravated due to the emergence and spread of *Plasmodium* strains resistant to current drug treatments. Most of *Plasmodium* species have an African origin. Parasites like *P. vivax* have colonized the world following complex migrating movements, partially unclear due to the lack of European *Plasmodium* genomes. The *Plasmodium* expansion, probably associated with the Neolithic onset, has been a strong selective pressure for the exposed human populations. Here we present the genomes of eradicated European strains of *P. vivax* and *P. falciparum*, which have been used to trace the migrating movements of these pathogens, as well as for dating the *P. vivax* dispersal. A genetic screen of malaria resistance variants in ancient European populations has revealed very low rates of genetic adaptive variants, which might be explained by a very recent introduction of malaria in Europe.

Preface

Malaria is the clinical manifestation of the *Plasmodium* infection. *Plasmodium* has accompanied humans for thousands of years, in Africa as well as when humans left it. The proofs of this remarkable contact are genetically conspicuous, with multiple marks of selection in human immunity genes, and also from a historical point of view, with a broad variety of historical records focused on malaria, some of these nearly as ancient as the scripture.

The present thesis includes the publication of the genome of an eradicated strain of European *P. vivax*. Additionally, the complete mitogenomes of both European *P. falciparum* and *P. vivax* are also presented. This genomes have been used to dissect the migratory movements that have spread *Plasmodium* all around the world. The P. vivax genome has also been used to estimate the first specific *P. vivax* mutation rate. The determination mutation rate is especially relevant, it has been used to date the *P. vivax* dispersal in the Americas.

The impact of malaria in ancient European population genetics has never been specifically studied. The analysis of genetic variants linked with malaria resistance in ancient European populations has revealed a limited selective effect of the pathogen. The lack of clear signs of resistance, and the dated phylogeny are reliable evidences of a recent *Plasmodium* introduction in Europe.

Index

Abbreviations

A: Adenine
aDNA: Ancient Deoxyribonucleic Acid
ACT: Artemisinin Combined Therapies
BCE: Before Common Era
BP: Before present
C: Cytosine
CQ: Chloroquine
DNA: Deoxyribonucleic Acid
G: Guanine
LNBA: Late Neolithic-Bronze Age
mtDNA: Mitochondrial Desoxyribonucleic Acid
PQ: Primaquine
T: Thymine
PCA: Principal Components Analysis
PCR: Polymerase chain reaction
PvCRT: *Plasmodium vivax* Chloroquine Transporter
PvDHFR: *Plasmodium vivax* Dihydrofolate reductase
PvDHPS: *Plasmodium viva*x Dihydrofolate synthase
PvMDR1: *Plasmodium vivax* multidrug resistance protein
SGS: Second generation sequencing
VCF: Variant Calling Format
WHO: World health Organisation

1. INTRODUCTION

1.1 Ancient DNA

Ancient DNA (aDNA) sequencing has changed the focus of study in Anthropology and Paleontology. This revolution must be placed in the global transformation that biological disciplines have experienced with the emergence of the PCR, and afterwards SGS techniques. These innovations have been used to define individuals and populations by genetic parameters, magnifying the knowledge obtained by physical anthropological methods and cultural studies of ancient remains.

The study of aDNA started in 1984 with the sequencing of few mtDNA bases from the Quagga (Higuchi et al. 1984), an old African equid extinct in the XIXth century. From this moment the discipline has evolved through time, driven by technical and biological advances. At present time the combination of high-throughput sequencing technologies and the ultimate informatics tools has allowed the study of whole genome sequences of large population datasets (e. g., Stoneking et al. 2011, Allentoft et al. 2015, Fu et al. 2016, Olalde et al. 2018).

1.1.1 aDNA particularities

Ancient DNA can be described as any genomic sequence retrieved from dead organisms. In the living tissues the DNA damage reactions are faced by mechanisms that preserve the integrity of the genetic material, in dead organisms such mechanisms are no longer active and the sequences are found in variable stages of degradation.

A) aDNA degradation and decay

Once an organism dies the endogenous, bacterial and fungal endonucleases degrade the DNA with a variable rate (Lindahl et al. 1993). The ratio and speed of the DNA degradation is influenced by atmospheric conditions such as temperature or humidity, and other variables linked with the burial environment like salt concentration, soil pH or the chemical composition of the ground. The degradation process reduces the amount of endogenous aDNA present in the samples that is usually situated in fractions below 1% of the total sequenced reads (Fu et al. 2013). There are some exceptions in which retrieved endogenous DNA have been found in exceptionally high proportions, which in some cases have exceeded the 70% (Prüfer et al. 2014, Meyer et al. 2012, Raghavan et al. 2014, Rasmussen et al. 2010, Gamba et al. 2014, Keller et al. 2012, Carpenter et al. 2013, Lazaridis et al. 2014, Olalde et al. 2014). The described climatic and natural agents act upon the DNA producing chemical reactions such as deamination, depurination or hydrolytic damage that cause breaks in the DNA structure and degrades it (Figure 1) (Höss et al. 1996).

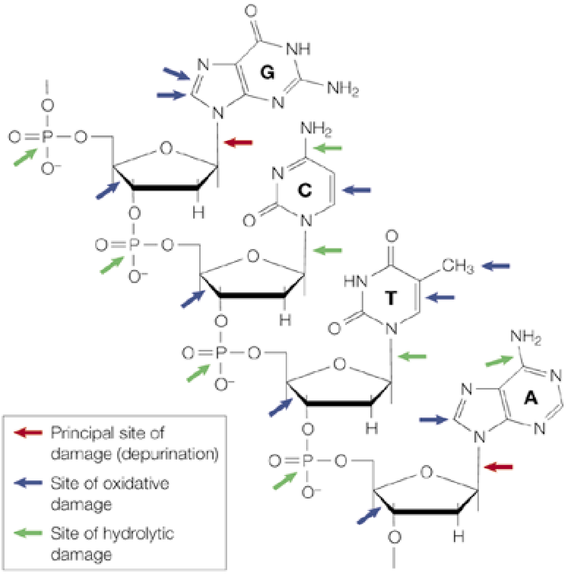

G
C
T
A
Principal site of damage (depurination)
Site of oxidative damage
Site of hydrolytic damage

Figure 1: Principal sites where damage is likely to affect ancient DNA. Depurination causes breaks in the DNA chain, Hydrophilic damage also leads to DNA chain breaks, oxidative damage modifies the nitrous bases and the sugar-phosphate backbone of the DNA (Hofreiter et al. 2001).

B) Cytosine deamination

DNA sequences suffer structural changes due to the post-mortem breaks of DNA molecule bounds. The most characteristic and abundant degradation pattern is the hydrolytic deamination of cytosine (C) to uracil (U) (Gilbert et al. 2007), this causes the incorporation of a complementary adenine (A) during DNA replication, which is observed by C to Thymine (T) substitutions in the 5' ends of the sequences (Figure 2) (Hofreiter et al. 2001).The C to T substitutions in the 5' ends of the DNA fragments, appear in an elevated ratio of G to A substitutions in the 3' ends of the complementary strands (Briggs et al. 2007, Rasmussen et al. 2014). Deaminated cytosines represent the most abundant substitution pattern in aDNA, being especially prevalent in the read ends where the fraction of deaminated cytosines can exceed the 40% (Briggs et al. 2007). This is mainly explained because the rate of degradation in the single-stranded overhanging ends is at least twice than in double-stranded ends (Lindahl et al. 1993)

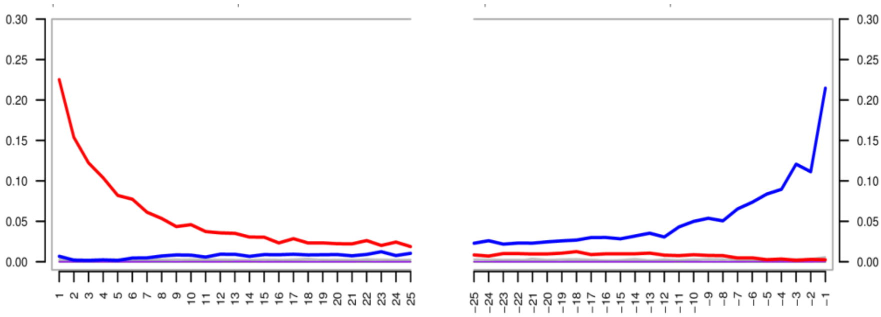


Figure 2: Damage plots generated with MapDamage. Red lines represent the C to T transitions frequency, blue lines represent the G to A transitions frequency. Y-axis indicates the percentage of sites presenting substitutions, X-axis indicates the base position in the DNA read.

Some specific techniques have been implemented to minimize the complications that aDNA damage produces in sequencing and mapping procedures. Single stranded library building protocol can be an efficient method to analyse poor quality samples, with low rates of endogenous DNA and highly degraded strains (Gansauge et al. 2013), however the efficiency increase of this method compared with the classical double stranded library performance, do not always compensates the time and effort used (Wales et al. 2015). Other methodologies used to face degradation are based on nucleic acid capture (Bos et al. 2011, Carpenter et al. 2013) or partial Uralic-DNA-glycosidase treatment (Rohland et al. 2015), that will be properly explained in the Methods section of this thesis. Despite the inconvenient, the presence of Cytosine deamination can also be used to differentiate between real aDNA and modern contaminant reads (Rohland et al. 2009).

C) Fragment length

The first published study describing the aDNA particularities already evidenced that almost all the aDNA sequences were present with very small fragments (from 50 to 100 bp). (Pääbo 1989). This evidence was subsequently reported by the first studies performed with SGS applied to aDNA (Green et al. 2010). This feature difficulties both DNA extraction and read mapping.

Sequence fragmentation is mainly caused by depurination, which has been considered the most relevant chemical damage that degrades aDNA structure. The effect of depurination is observed by an overrepresentation of purines; Guanine (G) and Adenine (A) in the 5’ fragment ends. This pattern has been reported in diverse publications (Briggs et al. 2007, Orlando et al. 2011 and Meyer et al. 2012). Moreover recent publications have demonstrated that depurination occurs in both ends of the aDNA fragments (Meyer et al. 2012), the reason why most of the publications report it only in the 5’ ends is that most library building protocols require a pre-processing step that consist in a blunt-end repair, an enzymatic process that extends recessed and degrades overhanging 3′ ends of DNA fragments (Briggs et al. 2007). Thus only with the development of single-stranded DNA libraries, which preserve both ends, could this pattern be reported in both ends of the aDNA fragments (Meyer et al. 2012).

In the depurination process an N-glycosil bond between a purine and the sugar of the DNA chain is broken. The result is a chain with an abasic site. The DNA chain is afterwards fragmented through β elimination leaving 3′-aldehydic and 5′-phosphate ends. (Figure 3) (Briggs et al. 2007).

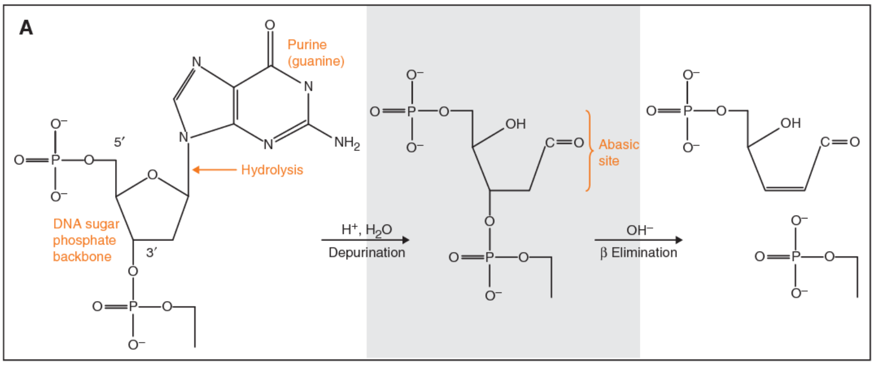


Figure 3: Depurination. Chemical teaction in which an N-glycosyl bond is broken resulting in an abasic site. The abasic site is later on removed, fragmenting the DNA through β elimination (Dabney et al. 2013).

1.1.2 Brief history of Paleogenomics

After the publication of the mtDNA from the Quagga (Higuchi et al. 1984) the first ancient human bases were sequenced. Bacterial cloning protocol was used to sequence DNA from an Egyptian mummy skin tissue (Pääbo et al. 1985). Nowadays this achievement is widely considered as an artefact of contamination. The huge amount of DNA required for sequencing in the pre-PCR era implied an arduous effort to retrieve very short fragments of DNA. The technology of the Polymerase chain reaction was developed by Kary Mullis in 1986 (Mullis et al. 1986) and represented a gigantic improve in terms of time and sample amount required for DNA sequencing. This revolutionary technique coupled with the firsts evidences of DNA retrieval from bone tissue (Hagelberg et al. 1989) allowed the direct sequencing of very tiny amounts of DNA, which avoided the destruction of large amounts of fossil material to obtain few bases sequenced (Bon et al. 2008). Additionally the evidence of the DNA preservation in bone tissue end up with the limitation of using only soft tissues in aDNA recovery (Hagelberg et al. 1989). Most of the studies based on PCR approach and using Sanger sequencing strategy (Sanger et al. 1975) target mtDNA. This approach requires the design of specific primers to amplify the desired genetic region, this feature difficulties the recovery of whole genome sequences. Nevertheless, there are some aDNA studies using PCR techniques that have targeted and sequenced nuclear regions (Krause et al. 2007, Lalueza-Fox et al. 2007).

Next generation sequencing techniques (Bentley et al. 2008) changed and improved radically sequencing technical procedures, opening the path to the effective sequencing of whole nuclear genomes. This improvement resulted particularly relevant in the field of aDNA, where the benefits linked with SGS go beyond the cost and time reduction. In SGS platforms, DNA fragments are directly sequenced over their full length. This represents a clear advantage compared to PCR. The usage of PCR platforms requires the ligation of sequencing primers to the DNA fragments, discarding sequences shorter than

100bp, which represent a great proportion of aDNA fragments. Additionally, PCR-primer ligation also hinders the evaluation of the amount of endogenous aDNA. As it has been described in section 1.1, aDNA shows specific damage patterns such as Cytosine deamination; if the ends of the sequenced reads correspond to the ligated primers this evaluation cannot be performed. The first published ancient genomic sequence obtained with SGS platforms were few Mb of a Mammoth nuclear genome (Poinar et al. 2006). Although most of aDNA studies have targeted DNA in bone/teeth tissue, is outstanding to remark that the first draft genome of an ancient species was obtained from DNA preserved in Mammoth hair (Miller et al. 2008). This genome was published using 454 data and provided significant information regarding to phylogenetics and functional genetics, once compared with the Elephant genome.

After the sequencing of the Mammoth genome fraction quickly different projects followed the path of using SGS technologies to sequence aDNA. Examples of this new period are such outstanding as the complete mitochondrial genome of a Neanderthal individual (Green et al. 2008) or the complete sequencing of a Mammoth genome (Miller et al. 2008). Later on a paleo-Eskimo individual belonging to Saqqaq culture that lived in Greenland more than 40000 years ago was published (Rasmussen et al. 2010), becoming the first modern human ancient genome published. After this achievement it’s remarkable to highlight the sequencing projects of the whole genomes of ancient human species; Neanderthal (Green et al. 2010, Prüfer et al. 2014), Denisovan (Meyer et al. 2012) as well as the outstanding discovery of an ancient-human hybrid Neanderthal-Denisovan (Slon et al. 2018). Other relevant projects are the sequencing of the oldest known modern human genome from Siberia (Fu et al. 2014), and more recently, the publication of a high coverage genome from a Neanderthal specimen from Croatia (Prüfer et al. 2017).

In the last ten years the number of sequenced human genomes has increased dramatically and nowadays hundreds of them have been published. Some of this genomes have been obtained with shotgun sequencing approaches (Skoglund et al. 2012, Skoglund et al. 2014, Seguin-Orlando et al. 2014, Olalde et al. 2014, Lazaridis et al. 2014, Cassidy et al. 2016, Jones et al. 2015, Martiniano et al. 2016, Martiniano et al. 2017) and others with hybridization capture methods, producing high amount of genomic data that has allowed the tracking of population movements around the world and the identification of selective and adaptive processes in the human genome evolution through time.

Most of the studies related with ancient population dynamics have been focused on European Remains, describing the history and peopling of Europe (Haak et al. 2015, Allentoft et al. 2015, Mathieson et al. 2015, Fu et al. 2016, Lazaridis et al. 2016, Martiniano et al. 2016, Mathieson et al. 2018, Olalde et al. 2018, Mühlemann et al. 2018, Damgaard et al. 2018, Valdiosera et al. 2018). This is partially explained because the temperate climate of most European latitudes allows a good preservation of DNA in archaeological remains and also because the vast amount of archaeological remains that have been so far studied in Europe compared to other continents. Nevertheless some studies have also been published covering the peopling of the Americas (Rasmussen et al. 2014, Raghavan et al. 2015, Rasmussen et al. 2015 and Skoglund et al. 2015). Africa (Garcia-Llorente 2015, Skoglund et al. 2017, Schlebusch et al. 2017, Schuenemann et al. 2017, Fregel et al. 2018) East Asia and Melanesia (Skoglund et al. 2016, Yang et al. 2017, Lipson et al. 2018a, Lipson et al. 2018).

Despite the fact that most of the aDNA studies have addressed topics related with human evolution dispersal and adaptation, some have attempted to describe and analyse extinct species and environments. These studies have often looked forward to define the phylogenetic relationships between the extinct animals and its present-day sister species, as well as identifying particular adaptive genomic events. Some as the New Zealand moas (Cooper et al. 2001), Giant lemurs (Orlando et al. 2008) or Carolina Parakeet mtDNA (Kirkman et al. 2012) using PCR sequencing approaches. While others like the sequencing of Balearic goats (Ramirez et al. 2009), the Passenger pigeon (Murray et al. 2017), a ~120 Kya cave bear genome (Dabney et al. 2013) or a Middle Pleistocene horse found in the permafrost (560-780 Kya) (Orlando et al. 2013) with SGS techniques. This last publication is the oldest aDNA sequence ever published, while the cave bear one is the oldest recovered genome from a fossil not preserved in the permafrost. There are also studies that have targeted flora DNA such as Holocene plants from lake sediments (Willerslev et al. 2003) or ancient wood (Wagner et al. 2013).

The oldest published DNA sequences are 700-800 Kya old (Willerslev 2007, Orlando et al. 2013). These sequences already exhibit strong signals of DNA degradation that could be an indicator that these samples are at the edge of the preservation time. This limit has been an issue under controversy since no correlation between time and degradation has been demonstrated (Sawyer et al. 2012) and probably a huge proportion of the DNA damage such as depurination occurs shortly after dead, mediated by endogenous enzymes (Briggs et al. 2007, Krause et al. 2010, Orlando et al. 2011, Meyer et al. 2012). Nevertheless, seems obvious that even frozen, DNA survives within a time limit, so we cannot retrieve genetic information from deep time in the past. Additionally, the preservation of DNA in warm environments is not comparable to the listed discoveries.

Some proteins have harder structures and resist longer and in harsher environments compared with DNA molecules. In this sense the sequencing of ancient proteins up to 80 MY (Schweitzer et al. 2007, Schweitzer et al. 2009, Wadsworth et al. 2014; Buckley et al. 2014) opens a new field that can end up with DNA time limitation. This approach has already permitted the reconstruction of partial proteomes of extinct species as Mammoth (Capellini et al. 2011) and surely be a future approach for other works.

In recent years the demands for a higher speed in sequencing pipelines as well as an increase in the length of the sequencing products are being accomplished. Third generation sequencing techniques (Schadt et al. 2010) target single DNA molecules enabling a real time sequencing. This is translated in an increase of the read length, a time reduction and the elimination of PCR biases. However, these sequencing techniques present higher error rates compared with SGS platforms and most of the clear advantages may not be useful in the sequencing of highly degraded samples.

The future of aDNA disciplines will necessarily have to look for sequences in unexplored environments and think about alternative approaches such as protein sequencing to unravel genetic adaptation through time. The discoveries in the field have been always related with the ultimate technical improvements. The advent of third generation sequencing methods can be a basis for new studies that may allow unreached goals such a de-novo assembly of aDNA remains. However, in terms of data amount, the advent of third generation sequencing methods will not have the impact that SGS had. An alternative and promising future might be found in the targeting of ancient Metagenomics data that could lead us the recovery of complete extinct environments.

1.2 Infectious diseases

The history of humanity is inconceivable without explaining the impact that infectious diseases have had in human peopling. It is well documented that Infectious diseases have been the main cause of death of human populations since first modern humans emerged in Africa more than 200,000 ago (Dye 2014), probably up to 300,000 (Scally et al. 2012). This coexistence between modern humans and Infectious diseases is neither new nor exclusive. There are described pathogens that affect the totality of the animal species.

Farming was originated in the Middle East 11,000 years ago and started expanding towards Europe 8,500 years BP (Skoglund et al. 2012). This technological revolution lead to massive migrating movements from the Near-East that expanded the agriculture and also shaped the genetic basis of European populations (Bramanti et al. 2009, Lazaridis et al. 2014, Haak et al. 2015, Skoglund et al. 2014, Mathieson et al. 2015, Gamba et al. 2014, Hofmanová et al. 2016, Lazaridis et al. 2016). The available genetic data indicates that most of the European Neolithic populations emerged from the Anatolian ones (Hofmanová et al. 2016) which expanded to all European regions during 5,000 years (Skoglund et al. 2012). Other Neolithic expansions also took place expanding farming in distinct regions of the globe, surprisingly, Neolithic populations from SW-Asia do not share genetic affinities with the Anatolian ones (Broushaki et al. 2016). Together with to the onset of farming, animal domestication took place at the same time and in the same place (Zadar et al. 2008).

Cattle was domesticated 9,000 years BP in the Near East (Scheu 2015). Pigs were primary domesticated in East Asia and introduced in Europe from Anatolia during the Neolithic (Ottonie et al. 2013) and goats were also introduced from the Near East (Fernandez et al. 2006). The farming technical revolution implied dramatic demographic changes in terms of population density and population structure which also had repercussions on the spread, virulence and evolution of infectious diseases (Diamond et al. 2002, Bocquet-Appel et al. 2011). Simply, the increase in population density linked with a close relation to livestock conduits to the spread of zoonotic diseases (Pearce-Dauvet et al. 2006, Wolfe et al. 2007). Impressively, more than 800 out of the 1400 described human pathogens are zoonotically transmitted. Commonly, zoonotic pathogens are less transmissible between humans than those that do not have animal reservoirs, however there exist relevant exceptions such as *M. tuberculosis*, *Y. pestis* or *B. burgdorferi* (Woolhouse et al. 2005). Despite being less transmissible between humans, the spread capability of the zoonotic pathogens is compensated because the introductions from the reservoir are constant (Woolhouse et al. 2005).

Most of present day infectious diseases have been infecting and killing humans for long time, but the effect of pathogens in populations is not always uniform. There are special events called Outbreaks were pathogens cause devastating effects in human populations and in its economic structures (Andam et al. 2016). Some of the most dramatic outbreaks are well known, being relevant events that have determined the history of human populations and cultures. Some examples of these epidemics are: the Black Death *(Y. pestis)*, Cholera pandemics in the XIXth Century *(V. cholerae)* or the Spanish flu *(Influenza virus)*. The genetic outcome of these outbreaks in human genome has often been a topic under debate, several studies have tempted to detect signals of selection in human genome related with these events (Laayouni et al. 2014, Mathieson et al. 2015) with variable rates of success. A deletion of 32 bases in CCR5 gene, a chemokine receptor, known as CCR5-Δ32 was postulated for long as a mark of selection attributable to the bubonic plague (Stephens et al. 1998). This variant is especially interesting because is present in European populations in frequencies up to ~10% while is absent in the rest of human populations, furthermore it is demonstrated that this variant confers resistance against HIV-1 among other pathogens (Schliekelman et al. 2001). The arose of the variant was dated in ~700 years BP and because of HIV has not been a selective pressure in European human populations, *Y. pestis* appeared as a credible candidate. This theory was later on refuted by genetic modelling of the allelic frequencies. The present day frequencies are better explained by an endemic infection such as Smallpox (Galvani et al. 2003), with a reduced death rate and constant exposition, rather than being the result of a specific devastating outbreak.

Between 1900 and present times, developed societies have experienced a drastic decline of childhood mortality linked to infections. This has resulted in a population growth that has been compensated by a decrease in births. This process, occurring nowadays in developing countries, is known as demographic transition (Figure 4) (Population Reference Bureau Staff. 2004) and is mainly related with the control of pandemics. In present days there still exists societies in which infectious diseases are the principal cause of death, especially with elevated rates of child mortality. Despite the impact of infectious disease in developed countries is moderate (WHO 2017), still nowadays there is no country in the World where infectious diseases can be considered a negligible cause of death (Woolhouse et al. 2005). For this reason, there is wide consensus that infectious diseases are one of the strongest forces that have shaped human genome (Fumagalli et al. 2011, Mathieson et al. 2015).

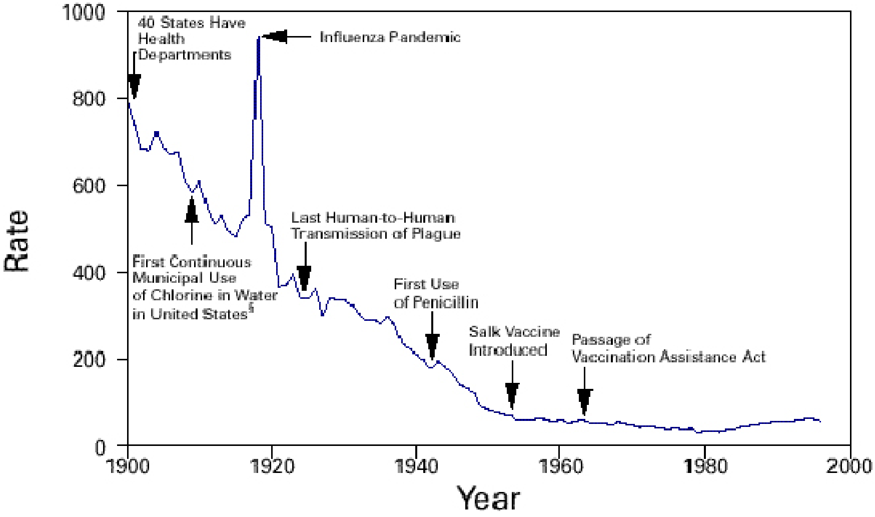


Figure 4: Evolution of deaths caused by infectious diseases in the United States of America between 1900 and 2000. The Rate is expressed for a 100,000 individual population per year (CDC).

1.2.1 Palaeomicrobiology

Palaeomicrobiology is defined as the study of microorganisms in ancient remains that were naturally present in healthy organisms as well as those that were responsible for infectious diseases.

Until the development of genetic techniques, the identification of infectious pathologies in ancient remains was restricted to the visual identification of bone injuries, added to the correlation of these identifications with ancient written proofs (Kousoulis et al. 2012). This kind of studies are very limited and often inconclusive as there are few pathologies that result in identifiable bone injuries, and to make things worse, different pathogens can produce similar injuries (Ortner et al. 2011). There are many pathogen casualties in which it is impossible to visually identify the etiological cause of the reported mortality. Some examples of this dubious lesions are Syphilis, caused by *Treponema pallidum*, which can be mistaken with skeletal lesions (Rothschild et al. 2005) or Brucellosis caused by *Brucella melitensis* that can be confused with Tuberculosis, caused by *Mycobacterium tuberculosis* (Mutolo et al. 2012, Kay et al. 2014). These identifications can only be performed through genetic markers. A paradigmatic example of this limitation is that Medieval Black Death could be attributed to a viral pandemic with an aerosol transmission pattern based on the descriptions from historical records (Bossak et al. 2007).

The first genetic studies that targeted ancient pathogens were performed with PCR techniques (Kolman et al. 1999, Gernaey et al. 2001, Drancourt et al. 2003, Zink et al. 2003 and Nguyen-Hieu et al. 2010). This methods require a prior pathological diagnosis, since specific PCR primers must be designed to amplify each possible pathogen (Willerslev et al. 2007), however as bacteria and virus are ubiquitous, the specificity of these tests is usually under-rated and false positive results are frequent (Päabo et al. 2004, Gilbert et al. 2005, Gilbert et al. 2006). The usage of SGS methods in paleomicrobiology; paleogenomics, also requires a previous pathologic evidence. The present standard protocols include selective pathogen DNA capture methods like in-solution capture (Schuenemann et al. 2011) or array-hybridization (Bos et al. 2011). Alternative Metagenomics approximations (Kay et al. 2014, Weyrich et al. 2017) can be useful to describe environments, but are a potential source of false positives, when the study is designed to identify one specific pathogen, in this case a comparative analysis becomes crucial to reduce the soil contaminants false positives (Campana et al. 2014).

In the history of paleogenomics the publication of the draft genome of *Y. pestis* from British individuals of the XIVth century (Bos et al. 2011) is a landmark becoming the first draft genome of a pathogen obtained from ancient human remains. After this publications others have characterized other strains of *Y. pestis* from the Bronze Age (Rasmussen et al. 2015, Spyrou et al. 2018) to the XIXth century pandemics (Stenseth et al. 2008) and from China (Cui et al. 2013).

Other pathogens have also been analysed and recovered, it is important to remark the publication of the sequences of ancient *Mycobacterium tuberculosis* (Wood et al. 1992, Monot et al. 2005, Bouwman et al. 2012, Campana et al. 2014, Bos et al. 2014), XVIth century Hepatitis B from a Korean mummy tissue (Kahla Bar-Gal 2012), a virus that according to a recent screening has been deeply present in Eurasia since the Bronze Age (Mühlemann et al. 2018), a XIXth Century *Vibrio cholerae* (Devault et al. 2014), the recovery of *Brucella melitensis* from medieval ages (Kay et al. 2014), historical sequences of *Mycobacterium leprae* (Schuenemann et al. 2013, Adler et al. 2013; Warinner et al. 2014, Krause-Kyora et al. 2018, Schuenemann et al. 2018), genetic sequences from the Spanish flu in the XXth century (Kay et al. 2015), Roman era *P. falciparum* genomic reads (Marciniak et al. 2016), historical American *T. pallidum* genomes (Schuenemann et al. 2018), Variola virus from the XVII century (Duggan et al. 2016) and *Helicobacter pylori* from the Ötzi Iceman (Maixner et al. 2016).

The oldest published microbiological data is microbiota DNA isolated from dental calculus from Neanderthal specimens that lived 40ky BP. The sequenced Neanderthal microbiota once compared with the dental bacteria of modern humans and chimpanzee showed strong affinities with the one found in the chimpanzee. This indicates a correlation between the dietary habits and the corresponding oral microbiota. (Figure 5) (Weyrich et al. 2017). The tissue that was previously studied by Adler and colleagues in 2013 and Warinner and colleagues in 2014 also contained strains of *Streptococcus mutans*, the bacteria responsible of causing caries, thus becoming the oldest sequenced pathogen until present days.

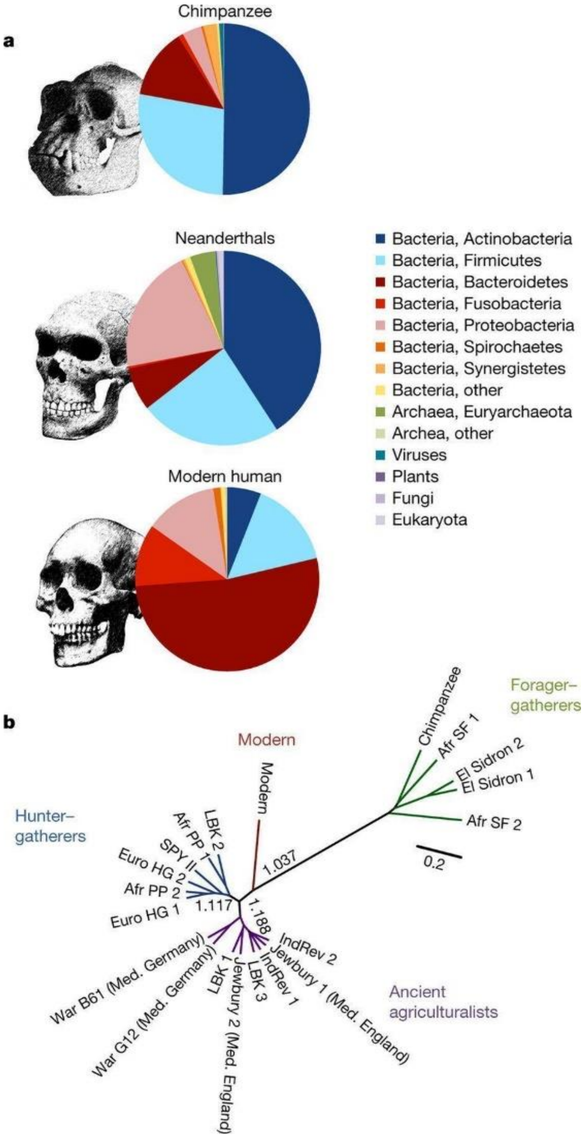


Figure 5: Comparison of oral microbiota of a wild-caught chimpanzee, Neanderthals and modern humans in the bottom of the figure an UPGMA clustering is displayed for analysed metagenomes revealing a strong correlation between diet and the oral microbiota (Weyrich et al. 2017).

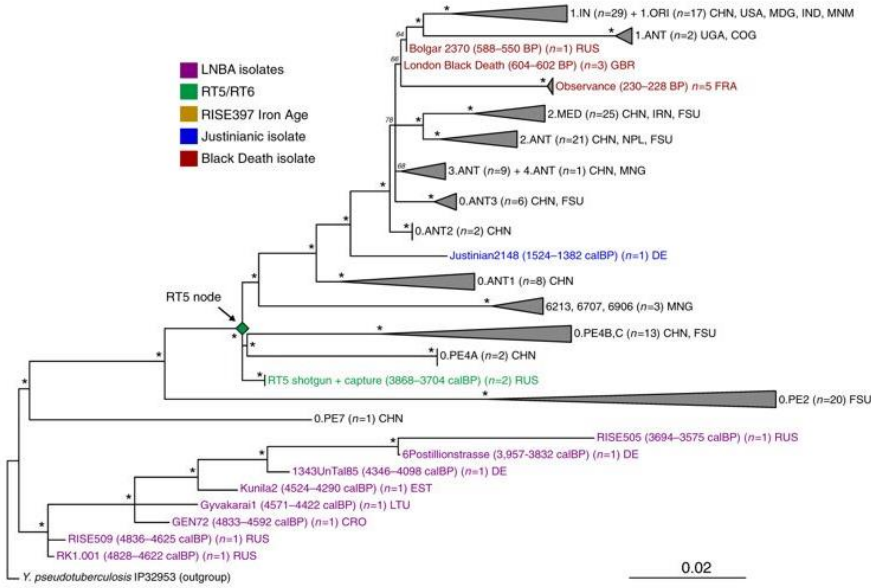
1.2.2 The Plague

Y. pestis has attracted the attention of a great proportion of the studies related with palaeomicrobiology. Because of that, is the pathogen with the largest number of published strains recovered from ancient remains, as well as the one with the most complete timeline. The interest that this bacteria attracts is the consequence of its famous outbreaks (Figure 6) (Bos et al. 2011, Wagner et al. 2014). All together defines *Y. pestis* as the paradigmatic pathogen in the field of infectious diseases in past populations research.

Figure 6: The Triumph of Death (Pieter Bruegel the Elder, 1562). Oil panel painting showing an allegory of the Last Judgment influenced by the medieval plague scenes.

Y. Pestis infection is the cause of bubonic plague. This bacteria originated and evolved from *Y. pseudotuberculosis,* much less pathogenic than *Y. pestis* (Achtman et al. 1999). Nowadays is still present in rodent reservoirs and it is endemic in 17 countries. Three main *Y. pestis* epidemics have affected Europe in historical times. The oldest documented one is the Plague of Justinian, from the 6th to the 8th Century AD (Russell et al. 1968). Probably the most famous one is the pandemic that devastated Europe in the XIVth century. Named the Black Death, this epidemic could have killed up to the 40% of the European population and was present in Europe until the XVIIIth century (Zietz et al. 2004; Benedictow et al. 2004). The most recent plague pandemic occurred between XVIIIth and XIXth centuries (Cohn et al. 2008; Stenseth et al. 2008). Based on literal records, possibly earlier *Y. pestis* outbreaks occurred in Europe prior to the Justinian plague, such as the Plague of Athens (Vth century BC) and Antonine plague (IIth Century AD). The lack of concluding DNA evidence do not allows neither the confirmation of such events nor the identification of the pathogen linked with the historical records (Drancourt et al. 2002). We can only count on strains from the Justinian outbreak (Wagner et al. 2014; Feldman et al. 2016), XIV century Black Death strains (Bos et al. 2011, Schuenemann et al. 2014, Spyrou et al. 2016) and XVIII Century pandemic (Bos et al. 2016).

The earliest evidence of *Y. pestis* DNA presence in human remains has been detected in Late Neolithic and Bronze Age individuals from the steppe and eastern Europe (5000-3500 BP) (Rasmussen et al. 2015) and in some LNBA individuals from Siberia (Valtueña et al. 2017). Interestingly the comparison of these strains has revealed that most recent common ancestor of all European *Y. pestis* strains lived up to ~6000 years BP, however the oldest recovered *Y. pestis* strains do not harbour the pMT1 plasmid. This plasmid encodes the ymt gene, which is needed for the flea transmission, without these gen *Y. pestis* cannot cause bubonic plague. The oldest *Y. pestis* strain capable to be transmitted by fleas was retrieved from Andronovo ancestry Bronze Age individuals (3800 BP) (Figure 7) (Spyrou et al. 2018). This discovery indicates that the Black Death causal strain has been present in Europe at least since the Bronze Age.



LNBA isolates
RT5/RT6
RISE397 Iron Age
Justinianic isolate
Black Death isolate
1.IN (n=29) + 1.ORI (n=17) CHN, USA, MDG, IND, MNM
1.ANT (n=2) UGA, COG
64
Bolgar 2370 (588–550 BP) (n=1) RUS
66
London Black Death (604–602 BP) (n=3) GBR
Observance (230–228 BP) n=5 FRA
78
2.MED (n=25) CHN, IRN, FSU
2.ANT (n=21) CHN, NPL, FSU
68
3.ANT (n=9) + 4.ANT (n=1) CHN, MNG
0.ANT3 (n=6) CHN, FSU
0.ANT2 (n=2) CHN
Justinian2148 (1524–1382 calBP) (n=1) DE
0.ANT1 (n=8) CHN
6213, 6707, 6906 (n=3) MNG
RT5 node
0.PE4B,C (n=13) CHN, FSU
0.PE4A (n=2) CHN
RT5 shotgun + capture (3868–3704 calBP) (n=2) RUS
0.PE2 (n=20) FSU
0.PE7 (n=1) CHN
RISE505 (3694–3575 calBP) (n=1) RUS
6Postillionstrasse (3,957-3832 calBP) (n=1) DE
1343UnTal85 (4346–4098 calBP) (n=1) DE
Kunila2 (4524–4290 calBP) (n=1) EST
Gyvakarai1 (4571–4422 calBP) (n=1) LTU
GEN72 (4833–4592 calBP) (n=1) CRO
RISE509 (4836–4625 calBP) (n=1) RUS
RK1.001 (4828–4622 calBP) (n=1) RUS
Y. pseudotuberculosis IP32953 (outgroup)
0.02

Figure 7: *Y. pestis* Phylogenetic tree. The tree is built with current strains of *Y. pestis*, Late Neolithic samples, and Bronze Age samples, a Justinian plague sample and Black Death samples. The Andronovo sample shows more phylogenetic relationship with the Black Death strains that with the LNBA samples. (Spyrou et al. 2018).

1.3 Malaria

Malaria is an infectious disease caused by one of the six known Plasmodium species capable to infect human erythrocytes: *P. ovale curtisi*, *P. ovale wallikeri*, *P. knowlesi*, *P. malariae, P. falciparum* and *P. vivax.* It is estimated that up to 455,000 people died from paludism in 2016. The 91% of these deaths occurred in Africa. The limited access that African populations have to antimalarial treatments and the deficient malaria control policies are the main causes of such mortality. (Figure8). (World malaria report 2017, WHO)

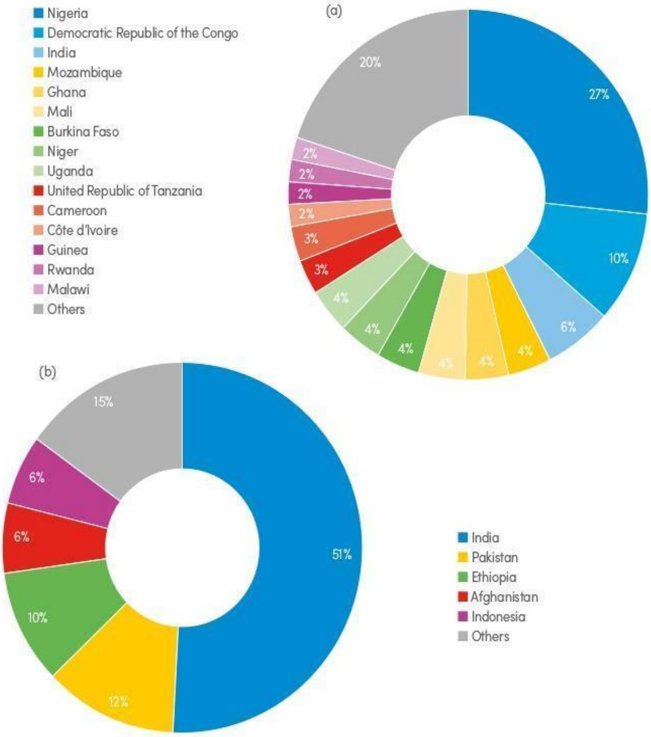


Figure 8: Present distribution of malaria cases worldwide. A) Global malaria cases by country, expressed in %, B) Cases of *P. vivax* malaria by country, expressed in %. (World malaria report 2017, WHO)

P. vivax is the species of *Plasmodium* genus with a broader distribution. While *P. falciparum* is the one responsible of most deaths associated with malaria (World malaria report 2017, WHO).

Nowadays, malaria is endemic in 90 countries (World malaria report 2017, WHO). The propagation capacity of *Plasmodium spp*. is basically defined by the presence of a proper vector, however there exists also other factors such as the temperature, humidity of the environment and the density of human populations which mediate in the *Plasmodium* reproduction and expansion. There are evidences that link the exposure of the human populations to malaria with the capacity of such populations to develop immune resistance (Cowman et al. 2016).

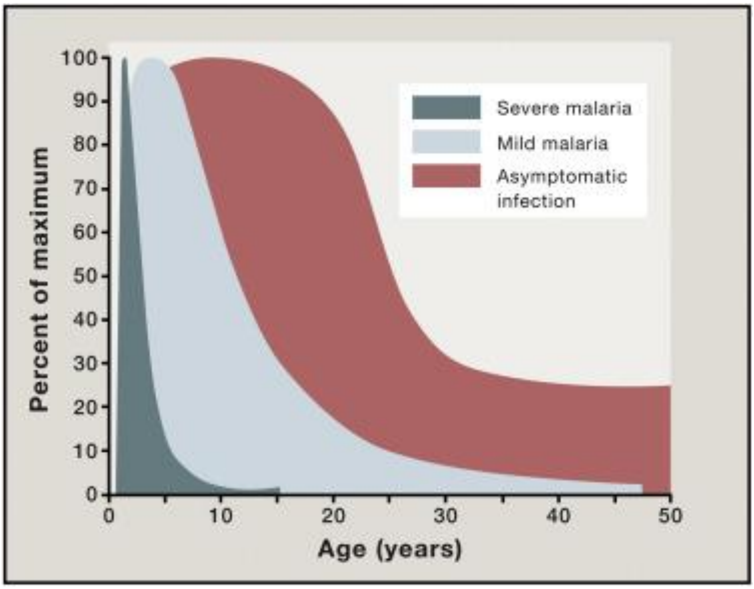


Figure 9: Distribution of severe, mild and asymptomatic malaria. The plot shows the distribution of malaria types per population, in the malaria endemic countries (Cowman et al. 2016).

As it is illustrated in Figure 9, most of the causalities associated with malaria correspond to child mortality, mortality rates decrease rapidly with age. The simplest explanation to this evidence would be that individuals acquire immunity during their life so as older they are, less capacity the pathogen has to kill the individual. Unfortunately, the explanation seems to be quite more complex. In endemic malaria areas most of individuals are constantly exposed to *Plasmodium* strains and normally present asymptomatic parasitemia, children as well. Children, who are less immunized suffer the most severe malaria. According to this continuous parasitemia, individuals acquire specific strain-resistance, so the immunity increase over time is due to the acquisition of particular resistance to particular strains. This explains why the rates of symptomatic and asymptomatic malaria also decrease through life in adult individuals (Cowman et al. 2016).

1.3.1. *Anopheles* mosquito

Plasmodium is transmitted by *Anopheles spp*, although there are more than 400 known species of the genus *Anopheles*, only 25 are proper vectors of *Plasmodium*. (Figure 10) (Sinka et al. 2012). The distribution of *Anopheles* mosquito defines the maximum extension of Malaria. Out to the present endemic areas malaria was present in most of the environments of the globe. Malaria was present in North America and in Eurasia, from Western Europe to the Eastern Euro Asiatic steppe. Malaria was eradicated from Europe during the XXth century while in recent years other countries such as Turkmenistan

(2006), United Arab Emirates (2007), Morocco (2011) or Paraguay (2018) have been declared Malaria transmission-free. There are others waiting for the WHO certification (World malaria report 2017, WHO 2017).

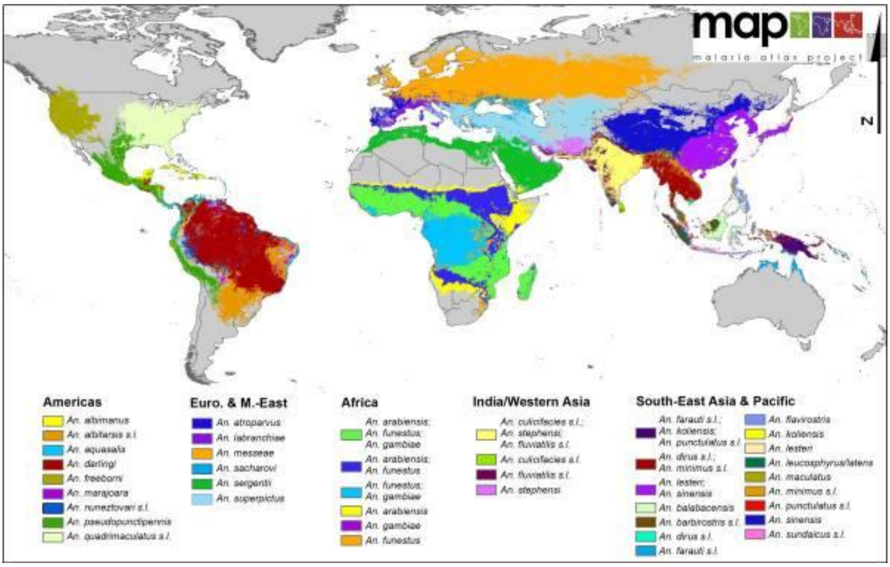


Figure 10: Distribution of the *Anopheles* mosquito species: Each colour is referred to a specific species of *Anopheles* genus. The Iberian Peninsula and part of the west Mediterranean shore are mainly colonized by *A. atroparvus* (Sinka et al. 2012).

Anopheles mosquito is a very old taxon, which originated in Pangea, and nowadays is present in all the continents of the world (Figure 10). *Anopheles* species have evolved for long and each species is highly adapted to its biological niche (Reidenbach et al. 2009). *Anopheles* mosquito presents two main characteristics that makes this species unique in terms of malaria transmission: The female mosquito is able to handle with the complete life cycle of *Plasmodium spp*. and the mosquito has a special predilection for human feeding (Fontaine et al. 2015). The predilection of some *Anopheles* species such as

Anopheles gambiae for human feeding seems to be the result of an adaptive process linked with the growth of human settlements in the Neolithic (Carter et al. 2002).

The control of mosquito population is one of the most common and useful strategies that administrations adopt to control the spread of *Plasmodium spp.* The massive use of insecticides can bring about the selection of resistant mosquito populations, which is an issue of major concern for public administrations (World malaria report 2017, WHO 2017).

1.3.2 *Plasmodium falciparum*

P. falciparum is the most lethal human malaria parasite. Its closest species are the African Lavernia parasites, of which it derived through a speciation process that is still under debate (Liu et al. 2010, Prugnolle et al. 2010). Lavernia is a subgenus of *Plasmodium* that comprises 8 other species that infect Gorillas and Chimpanzees (Boundenga et al. 2015, Liu et al. 2016). *P. falciparum* is the only Lavernia parasite with the ability to infect humans (Liu et al. 2010). The split of *P. falciparum* from its closest species, the Gorilla pathogen *P. parafalciparum* occurred 40,000-60,000 years BP. (Otto et al. 2018). Some authors have proposed that *P. falciparum* have been infecting humans at least for 60,000-100,000 following the out of Africa migrations towards Asia (Hughes et al. 2010, Tanabe et al. 2010). However, it is widely accepted nowadays that he present day distribution is the result of recent human population growth and migrations about 10,000 years ago, essentially explained by the Neolithic onset (Carter et al. 2002, Joy et al. 2003, Sundararaman et al. 2016). Farming led to a dramatic boost of both human populations and densities. The disparity between human and great apes population size turn into a selective pressure for *Anopheles* to feed on humans (Carter et al. 2002), which motivated the selection of human feeding genotypes. Therefore, *P. falciparum* would have also been forced to select those genotypes that conferred a major success both for human and mosquito infection. This process promoted an important population growth of *P. falciparum* that genetically appears as a bottleneck dated in 5,000 BP (Otto et al. 2018).

The present day genetic diversity of *P. falciparum* reveals a strong differentiation between the Asian-Melanesian isolates and the African-American strains (Figure 11), (Amato et al. 2016). American *P. falciparum* mtDNA diversity points out a close relationship of American *P. falciparum* strains and the African ones. (Joy et al 2003). The levels of genetic diversity found in America are sensibly lower than the Asian and the African ones, although the levels of population differentiation are higher in American populations than in African or Asian ones. American *P. falciparum* diversity seems to be the result of successive and diverse migrations from Africa linked with the Slave trade. The recurrent introductions explain the unexpected amount of genetic differentiation observed between present day American *P. falciparum* populations. (Yalcindag et al. 2012). The introduction of *P. falciparum* in Europe likely followed the spread of farming from Anatolia and could present strong similarities with the present day Indian populations. These populations, based on mtDNA analysis, appears closely related with the present day African populations (Tyagi et al. 2014).

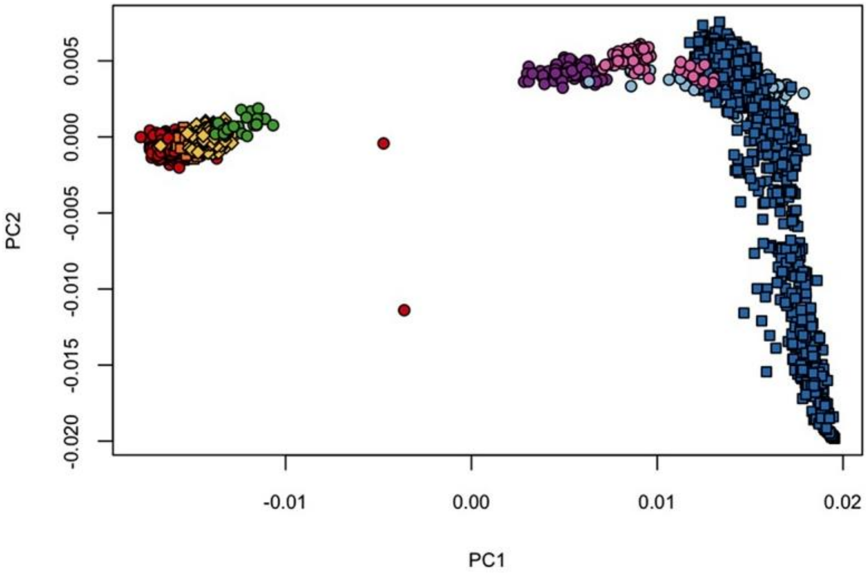


Figure 11: Principal components analysis. PCA computed with 3,411 clinical samples of *P. falciparum* from 43 countries, the PCA reveals a clear differentiation between African-American (left) and Asian-Melanesian samples (rigth) of *P. falciparum* (Amato et al. 2016).

P. falciparum has a complex life cycle, a feature shared with all species of *Plasmodium* genus. The life cycle of these parasites comprises one mosquito stage and one intra-human phase. Once the mosquito female bites the host, the sporozoites get into de the blood circulation reaching the hepatocytes. This process is quite complex and sporozoites need to cross the sinusoidal barrier, composed by endothelial and macrophage cells, (Tavares et al. 2013) by the mediation of a diverse sets of proteins (Cowman 2016). In the hepatocytes the sporozoites reproduce asexually, releasing the merozoites in the blood circulation (Sturm et al. 2006). Those merozoites invade erythrocytes, by a quite fast mechanism (Figure 12) (Weiss et al. 2015). Once inside the erythrocytes, merozoites reproduce asexually. The resultant merozoites egress the erythrocyte after several hours, precipitating the cell destruction. The merozoite release triggers a chronic cycle of asexual schizogony in the bloodstream (Dvorin et al. 2010).

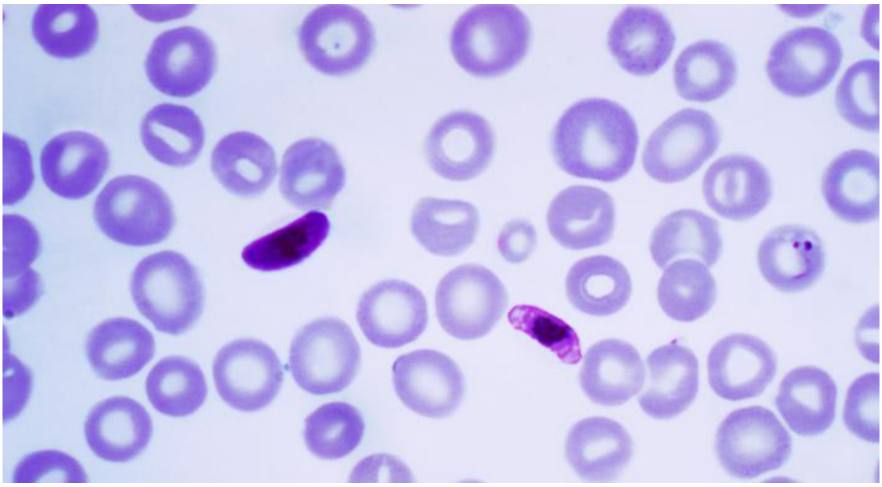


Figure 12: Gametocyte stage of *P. falciparum*. Gametocytes show the typical “banana shape”, in this stage the gametocytes can be ingested by the bite of an *Anopheles* female (CDC)

A proportion of those released merozoites are programed to undergo gametocytes. When these gametocytes mature they are ready to be ingested by a mosquito. In the mosquito gut the gametocytes emerge as extracellular female and males gametes. Here in the gut, mating occurs and once the zygote is formed this form migrates through the mosquito gut epithelium, where it reproduces asexually producing sporozoites that migrate towards the salivary gland where they are later on able to infect human again. (Cowman et al. 2016).

The genome of *P. falciparum* is about 24.4 Mb and has an extremely low GC content, below the 20%. There are 5,362 annotated coding genes. As it is common in *Plasmodium* genomes, the subtelomeric regions allocate highly extensive and diverse gene families like: var, rifin and stevor genes. (Gardener et al 2002). The encoded proteins mediate most of the *P. falciparum* important functions and are proper candidates for drug treatments and vaccines. One of the most extensively studied is PfEMP1 family. PfEMP1 proteins are exported to the *P. falciparum* infected erythrocytes membranes. The parasite is able to switch the expression of the proteins, promoting the evasion of the human immune response. This capacity explains the chronicity of the *Plasmodium* infections. The different PfEMP1 proteins have different receptor-binding selectivity which added to the sequestration of *P. falciparum* infected erythrocytes contributes to the mortality (Nunes-Silva et al. 2015)

1.3.3 *Plasmodium vivax*

Plasmodium vivax is one of the six *Plasmodium* species that routinely infects human erythrocytes (Singh et al. 2004, Sutherland et al. 2010 and Calderaro et al. 2013). The parasite is mostly absent in Sub-Saharan Africa due to the presence of Duffy Negative Haplotype (Miller et al. 1976). However, it is the *Plasmodium* species with the broader distribution, and the most prevalent one out of Africa, causing more than 16 million malaria clinical cases annually (World malaria report 2017, WHO 2017). The enormous distribution of *P. vivax* is explained by two unique features of the species. i) elevated infectious capacity of *P. vivax* to *Anopheles species* (Mueller et al. 2009) ii) ability to generate dormant hypnozoites, that can remain in a latent state in the hepatic cells for months. Hypnozytes became a reservoir that extends clinical attacks across seasons inhospitable to *Anopheles,* expanding the natural range of *P. vivax* into temperate zones (White 2011, Gething et al. 2012).

The low levels of *P. vivax* parasitemia do not appear to make *P. vivax* a benign parasite as it was previously believed. (Price. et al 2007, Mueller et al. 2009, Naing et al. 2014). *P. vivax* is responsible of severe malaria and death, and causes significant morbidity (Baird 2013). The control and elimination of *P. vivax* cannot be restricted to follow the programs designed for *P. falciparum.* The unique biology of *P. vivax* requires a deep comprehension and the definition of particular control programs (Howes et al. 2016).

***A).Plasmodium vivax* genome**

The genome of *P. vivax* is about 29Mb length distributed in 14 chromosomes, (Figure 13), plus a mitochondrial circular genome and an apicoplast circular genome (Carlton et al. 2008). The most used *P. vivax* reference genome is Salvador 1 assembly (Carlton et al. 2003). This reference genome has 5.400 genes annotated. With most of the genetic material properly assembled in chromosomes, despite having more than 2700 unassigned scaffolds. A great proportion of the unassigned scaffolds belong to repetitive and subtelomeric regions of the Chromosomes. This incomplete assembly especially difficulties the annotation of vir genes family, the broadest *P. vivax* protein family as most of its genes are located in these regions. The subtelomeric and telomeric regions exhibit higher rates of Tymines and Adenines compared to the core region, which has a higher GC content that reaches the 45% (Cunningham et al. 2010, Auburn et al. 2016)

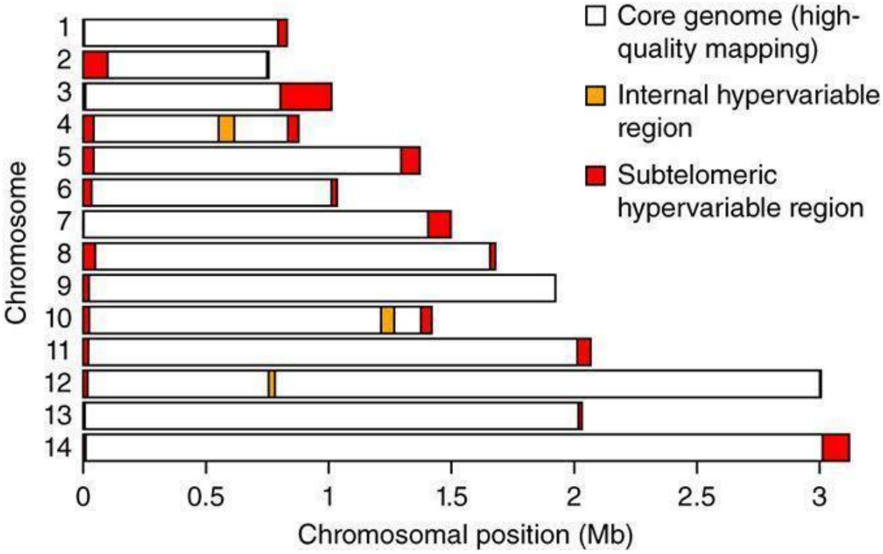


Figure 13: *P. vivax* genome map. The coloured regions highlight the variable fragments of the 14 chromosomes. The core genome of *P. vivax* is the uncoloured fraction (Pearson et al. 2016).

Recently a new *P. vivax* assembly (PvP01) has been build. This new assembly extracted from a Papuan monoisolate sample (Auburn et al. 2016) has improved the previous Sal1 assembly. In the PvP01 assembly there are only 227 unassigned scaffolds and the number of annotated genes is up to 6,642, with a particular important increase in the number of vir genes identified and properly annotated; 1,212 genes versus the 346 annotated in the Sal1 assembly. Vir gene family is the largest gene family present in the *P. vivax* genome, and remains widely unstudied. The family was firstly defined by 32 members distributed on 6 different subfamilies, but nowadays we know that at least there are 1,221 vir encoded proteins (Pir) that can be clustered in 27 different major groups (del Portillo et al. 2001, Auburn et al. 2016), and which represents the 5% of the *P. vivax* genome (Cunningham et al. 2010). Because of the size of the family, vir genes show extreme levels of genetic variability (Pearson et al. 2016), which suggest that different families mediate in differentiated vital functions for the life of the parasite (Carlton et al. 2008). In vitro studies have already demonstrated blood-cell binding properties (Yam et al. 2016), meditation of the adherence to endothelial cells (Bernabeu et al. 2012) and many other hypothesized functions that able the host immune evasion and proliferation (Hughes et al. 2004, Auburn et al. 2016).

Other relevant *P.* vivax proteins are the antigenic merozoite surface proteins (MSP), that englobe different protein families. Those proteins have a great potential as vaccine candidates as they are expressed in the merozoites surface (Rice et al. 2014). Members such as MSP3, MSP7 or MSP10 are allocated in some of the most divergent regions of *P. vivax* among the different parasite populations, which suggests local adaptations to human immune evasion (Rice et al. 2014, Hupalo et al. 2016). Other genes such as

SERA or MAEBL also play important roles in human immune evasion (Arisue et al. 2007).

B) *Plasmodium vivax* life cycle

P. vivax infection begins when an *Anopheles* female bits the skin and the delivered sporozoites reach the hepatocytes (Figure 14) (Frevert et al. 2005, Amino et al. 2006). In the hepatocytes *P. vivax* sporozoites can either differentiate into tissue schizonts or remain as dormant hypnozoites inside the hepatocytes for months (Krotoski et al. 1985). The molecular mechanisms that promote the activation of dormant hypnozoites are still completely ununderstood; both stress and a release pattern linked with a seasonal distribution of the mosquito look as possible variables (Carter et al. 2003). Both tissue schizonts and the activated hypnozytes (precursors of tissue schizonts), which after multiple mitotic replications, build the merosomes. This merosomes are vesicles filled of merozoites that are liberated into the liver sinusoids. This structure, which is generated from host hepatocytes membranes, allows the merozoites to avoid the surveillance of the immune system Kupffer cells to reach the blood vessels (Sturm et al. 2006). Once in the blood vessels *P. vivax* merozoites generally infect reticulocytes. This propensity is a characteristic of *P. vivax*, a feature that is not shared with *P. falciparum*, that primary infects mature erythrocytes (Sturm et al. 2006). This also allows a differential visual identification of *P. vivax* and *P. falciparum* infections based on the deformation of the infected blood cells; *P. vivax* deforms the erythrocytes, which present vesicular complexes build by the mediation of *P. vivax* expressed proteins (Suwanarusk et al. 2003, Barnwell et al. 1990).

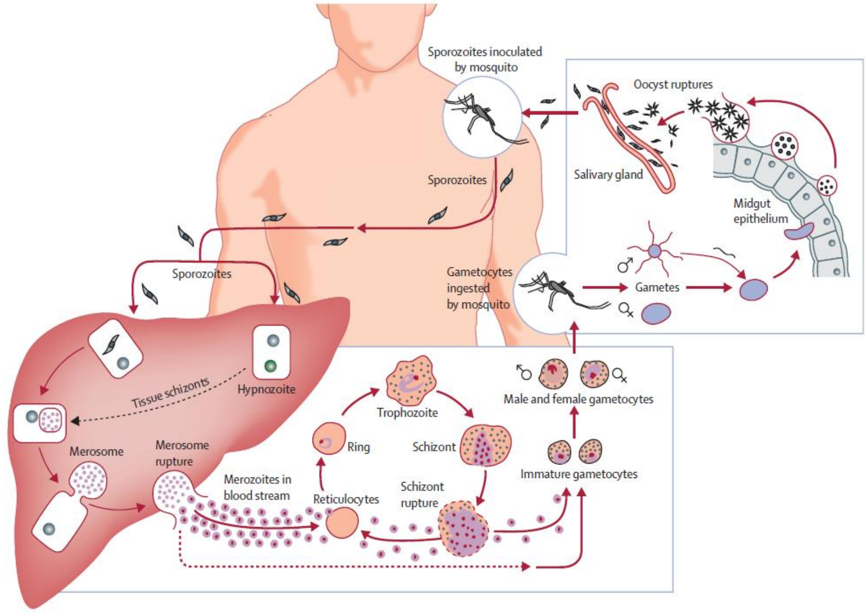


Figure 14: Life cycle of *Plasmodium* vivax. The cycle combines both human and mosquito stages. (Mueller et al. 2009).

The predilection of *P. vivax* for the reticulocytes, which represents less than the 2% of the total amount of erythrocytes, is proposed as one of the reasons for the dubious benign forms of malaria associated to *P. vivax* infection. Albeit some authors have demonstrated that *P. vivax* as a mono infection is able to produce infections as virulent as the ones caused by *P. falciparum* (Tjirta et al. 2008; Kochar et al. 2006; Barcus et al. 2007).

Some of the released merozoites can subdiferenciate into gametocytes that can be ingested by the mosquitoes, dispersing the pathogens before the first clinical symptoms appear. *P. vivax* gametocytes have round shapes while *P. falciparum* ones exhibit the genuine "*banana shape*" (Mueller et al. 2009). Inside the mosquito the gametes follow the typical pattern common to other *Plasmodium* species by migrating to the gut endothelium, where they reproduce sexually. The resultant squizonts cross the gut epithelium inside a motile ookinet that differentiates to an oocyst which, once broken, liberates the squizonts in the salivary glands, starting the cycle again (Krotoski et al 1985, Muller et al. 2009).

C) *Plasmodium vivax* origin and dispersal

Plasmodium vivax has an African origin, although for a long time the most standing hypothesis suggested an Asian origin based on the similarities of *P. vivax* with *P. cynomolgi* (Figure 15) (Tachibana et al. 2012) and other Asian macaque infecting species. This theory suggested that *P. vivax* arose several hundred thousand years ago following a cross-species process from a macaque parasite in South-east Asia, and infected modern humans when they arrive in Asia 60,000 years BP. (Mu et al. 2005, Escalante et al. 2005, Cornejo et al. 2006, Carlton et al. 2013 and Tachibana et al. 2012). Nowadays the African origin is confirmed as *P. vivax*-like species have been found in African great apes endemically parasited (Liu et al. 2010, Liu et al. 2014 and Gilabert et al 2018). These *P. vixax*-like parasites are capable to infect humans (Prugnolle et al. 2013) and they do not show specific Chimpanzee and Gorilla lineages, as phylogenetic analyses show interspersed lineages, suggesting a continuous cross-speciation (Liu et al. 2014). The speciation date has been inferred with mtDNA phylogenies, estimating it in 400,000 years (Cornejo

2006), which has been confirmed with genomic data (Neafsey et al. 2012), pointing that possibly P. vivax originally infected multiple human populations (Gilabert et al 2018).

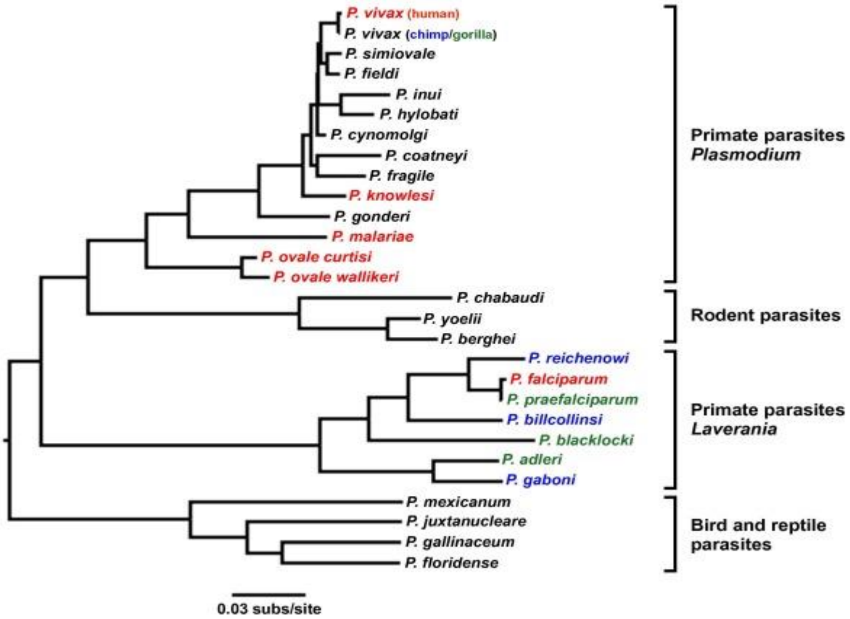


Figure 15: *Plasmodium* genus phylogeny: *Plasmodium vivax* clusters with its sister Gorilla and Chimpanzee *Plasmodium species*. The human parasites are highlighted in red, the chimpanzee in blue and the gorilla infecting *Plasmodium* in green. (Loy et al. 2017).

The parasite constitutes a monophyletic clade within the radiation of ape parasites. This suggests a single speciation event for *P. vivax*, however this explanation is at odds with the evidence that ape *P. vivax* can infect humans and they have a lack of specificity (Liu et al. 2014, Prugnolle et al. 2013). What seems to be more in concordance with the observed lack of genetic variation in human *P. vivax* is that the current *P. vivax* strain is lineage that survived the bottleneck caused by Duffy negative expansion in Africa (Conway et al. 2000, Tanabe et al. 2010 and Loy et al. 2015). The differential distribution of this polymorphism supports this theory (Howes et al 2011). Afterwards, *P. vivax* could spread to all latitudes following the Neolithic expansion (Carter et al. 2002).

The present-day genetic diversity of *P. vivax* can be fitted into four main differentiated clusters. Worldwide populations are clearly separated by population genetics software like PCA or Admixture. This distribution includes one African-Indian bundle that appears in an intermediate position between American and Asian clusters. The fourth, groups the Melanesian isolates that clearly differentiate to the rest (Hupalo et al. 2016, Pearson et al. 2016). The present day African strains show the highest level of genetic affinity with the Indian samples as the result of a Colonial African reintroduction after the spread of the Duffy negative phenotype (Mullis et al. 1976, Culleton et al. 2011 and Taylor et al. 2013). The current African and Indian strains show elevated levels of heterogeneity once compared with other specimens indicating that probably are the result of admixtures between European and Asian ones, suggesting the presence of human population connections occurring in the Indian subcontinent involving Asian, middle-East and Mediterranean populations which could be supported by human genetic data (Reich et al. 2009). The absence of European strains prevents to elucidate if the present day American strains have an African or a European origin. Once eradicated in Europe, the endemic areas of *P. vivax* are mostly placed in the Americas and in Asia. (Figure 16) (Price et al. 2007).

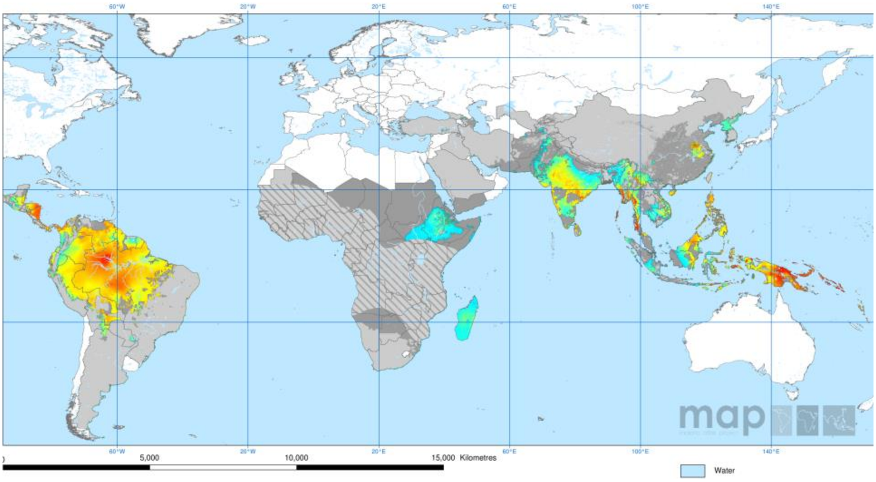


Figure 16: Spatial distribution of *P. vivax.* Nowadays *P. vivax* is endemic in Central-South America, Asia and Oceania added to some East-African regions (Howes et al. 2016).

1.3.4 Malaria treatment and resistances

The control of malaria is based on three main strategies that must be coordinated: vector propagation control, drug treatment and the research based on the development of an efficient malaria vaccine.

A).Vector Control

Mosquito control is an effective approach to prevent malaria dispersion. Reduction or even elimination of *Anopheles* populations avoids the propagation of the infection. The implemented techniques consist in the drainage of wetlands and in the usage of insecticides to kill the *Anopheles* mosquitoes. However, the massive usage of insecticides has induced the appearance of *Anopheles* populations resistant to some of these compounds (WHO 2017).

Mosquito surveillance also includes the control of its reproduction environments. An example of the importance of mosquito control can be found in historical records of World War II. During the Italian Campaign (1943-1945) the German Army tried to hamper Allied Amphibious landings near Rome by flooding the Pontine Marshes near Anzio, which had been drained by the Italian government before the war. Millions of larvae of *Anopheles labranchiae,* a malaria vector, were introduced as an act of biological warfare. This action, probably the unique documented case of Malaria used as a Biological Weapon did not have much effect because of the availability of antimalarial drugs to the Allied armies.

B) Drug treatment

The main purpose of all antimalarial treatments is to erase the parasite at the erythrocytic stage. Since the life cycles of *P. vivax* and *P. falciparum* are not identical, the treatments are diverse and adapted to each etiological agent. The first reported antimalarial treatment is Quinine, a natural alkaloid extracted from a native tree in South America and introduced in Europe from Peru in the 1600s. Quinine is still nowadays an effective antimalarial treatment (IOM 2004). The first synthesized antimalarial compound was Chloroquine in 1945 and quickly turn into the essential worldwide antimalarial drug. The first signals of CQ resistance emerged in *P. falciparum* strains in Papua New Guinea in 1989 (Rieckman et al. 1989) in 1991 there were also reported resistant strains of *P. vivax* in Indonesia (Baird et al.

1991). In the present days the most recommended treatment of *P. falciparum* infections are the Artemisinin combined therapies (ACT), whereas in the case of *P. vivax* infection, the most effective treatment consists in a combined therapy of Chloroquine (CQ) and Primaquine (PQ) (Nosten et al. 2007). It has been shown that only PQ acts efficiently depleting hypnozoites (Wells 2010), unfortunately is contraindicated in pregnant women, child and G6PD deficient individuals because causes hemolysis. These contraindications severely limit the capacity to treat *P. vivax* (Carter et al. 2011). In the listed cases, and if the patient has infected in a place where the resistance to Chloroquine is widely spread WHO recommends to switch to ACT as first line treatment (WHO 2017).

The genetic basis of drug resistance in *P. vivax* remains partially unknown. For years the strategy has relied on assigning homologous gene annotations detected in *P. falciparum* to explain the acquired *P. vivax* resistance, strategy that has not showed the expected results (Nomura et al. 2001). Genetic polymorphisms leading to increased level of expression of the PvCRT gene have been linked with the resistance to CQ (Sa et al. 2006) Several mutations in gene Pvmdr1 have also been linked to a resistant CQ phenotype in *P. vivax* isolates (Carlton et al. 2008, Joy et al. 2018). Modern approaches have compared *P. vivax* population allelic frequencies to determine genes of *P. vivax* that have been affected by positive selection, which partially can be due to antimalarial treatments. Genes like PvDHFR or PvDHPS were shown to be under high selective pressure as well as being located in genomic regions with high levels of linkage disequilibrium, which would be indicating the presence of recent selective shifts (Hupalo et al. 2016). Pvmdr1 gene was observed to be duplicated in several Thailand strains (Pearson et al. 2016). Despite these main evidences, still nowadays most of the molecular basis of the drug resistance acquisition and dispersal remain unknown.

Gene	Resistance to
PvMDR1	CQ, Mefloquine, Amodiaquine, Sulfadixine-PQ
PvDHPS	Antifolate, Amodiaquine, PQ
PvDHFR	Antifolate
PvCRT	CQ
Kelch12	Artemisinin

Table 1: principal genes related with *P. vivax* resistance to antimalarial drugs. Specific mutations in each gene confers resistance to specific antimalarial compounds. .

C) Malaria vaccine

An effective malaria vaccine has been one of the major objectives of the Malaria research for decades. Nowadays there are multiple lines of research in different stages of development. One of the projects that currently appear as more promising is a Vaccine that targets the sporozoite stage of *P. falciparum*. This vaccine that has been tested with African populations has showed to be effective in reducing the clinical cases of malaria in the test populations (RTS 2015).

1.3.5 Malaria in Europe

P. falciparum would have spread from Africa to all tropical and subtropical climates within the last 6,000 years (Carter et al. 2003, Otto et al. 2018). The expansion towards Europe seems to be much more recent, probably occurred in historical times. The spread of *P. vivax* is suggested to be quite comparable, *P. vivax* has also an African origin and after a complex path would have colonized Europe in recent times. The proposed estimations are around 10,000 years BP (Culleton et al. 2011). Nevertheless other publications suggest an older expansion dated up to 40,000 BP, that would have followed the first peopling of Europe (Escalante et al. 2005; Mu et al. 2005), however these theory seems to be poorly supported by the present estimations that link the arrival of *P. vivax* in Europe with the onset of Farming (Liu et al. 2014).

Hippocrates of Cos (460–370 BC), the Greek medical doctor of the Pericles era, considered as one of the fathers of medical science, described episodes of fevers in the Classical Greece. Being the most ancient record that depicts malaria symptoms. The illustrated symptoms are probably describing episodes of tertian malaria. (Bogdonoff et al. 1985). Tertian cyclic Malaria pattern is characteristic of *P. vivax* and *P. malariae* infections. Some authors have augmented that the impact of Malaria in Ancient Greece was restricted to *P. vivax* infections. The *P. falciparum* expansion along the Mediterranean shore was probably later, not before the Roman Republic or even in Imperial ages (Bruce-Chwatt et al. 1980). In addition an Egyptian papyrus known as the Papyrus of Ebers (Ebers 1875) that was drafted during the reign of Pharaoh Amenhotep I, about 3500 BP, could be describing the presence of malaria in Ancient Egyptian times, 1000 years before Hippocrates did in his reports. This Egyptian document is a collection of medical cases, pharmacological and care procedures which describe symptoms that can also be attributed to malaria, but symptoms of other pathologies could also fit with the described fevers. There are also texts that could be referring to malaria in Ancient China (3000 BP) and Ancient India (Dong et al. 1996). All this information concur with the hypothesis that malaria arrived in Europe between the last Ice Age and 2,500 years BP.

P. falciparum malaria is first documented in Europe in Classic Roman era remains and texts. There are biological samples of *P. falciparum* from the 5th Century of the CE (Sallarés et al. 2001) and from the 1st-2nd century of the CE in the Italian Peninsula (Marciniak et al. 2016). The roman author Celsius, differentiated between the symptoms of *P. vivax* and *P. falciparum* malaria during the reign of the Emperor Tiberius in the 3th Century of the CE. The first visual identification of *P. falciparum* and a proper description of the pathogen was drafted by Marchiafava and Celli in 1889. The etymology of Malaria derives from the Italian *mal-aria*, this reveals that until the discovery of the mosquito transmission in 1889, malaria was thought to be an infectious disease caused and dispersed by the corruption of the air.

Malaria was present in Europe until the XXth Century. In Spain, malaria was declared eradicated in 1964 (Pletsch et al. 1964), being one of the latest European countries to receive a final eradication certification. The environmental changes linked with the global warming represent thread in terms of malaria propagation in Europe. The climatic conditions of several ancient European reservoirs such as the Ebro Delta are seriously in danger to be colonized again by *Plasmodium* (Sainz-Elipe et al. 2010).

1.3.6 Human resistance against *Plasmodium* infections

Malaria has an enormous lethal potential, which added to a continued exposure for long, explains the colossal selective power of the disease (Kwiatkowski et al. 2005). The most spread and distinctive mutation linked with malaria resistance is the Duffy Blood phenotype. The phenotype is defined by the rs28814788 SNP located in the ACKR1 gene and is mostly fixed in West-African populations (Jallow et al. 2009). The mutated protein expresses a structural change that prevents the interaction with the *P. vivax* receptor and consequently inhibits the invasion of erythrocytes, which conferees a full resistant phenotype against *P. vivax* infections (Figure 17) (Tournamille et al. 1995). This variant has also been observed in low frequencies in some South-American populations. Maybe explained by an African ancestry flow introduced due to the Slave trade (Hupalo et al. 2016).

Genetic studies have identified populations of *P. vivax* that have selected variants to face the spread of Duffy negative phenotype. It is the case of Malagasy *P. vivax* populations, that present multiple copies of the Duffy binding gene (Menard et al. 2013), genotype also observed in Asian samples (Pearson et al. 2016). These variants explain the presence of *P. vivax* malaria in Duffy negative populations like Madagascar (Menard et al 2015).

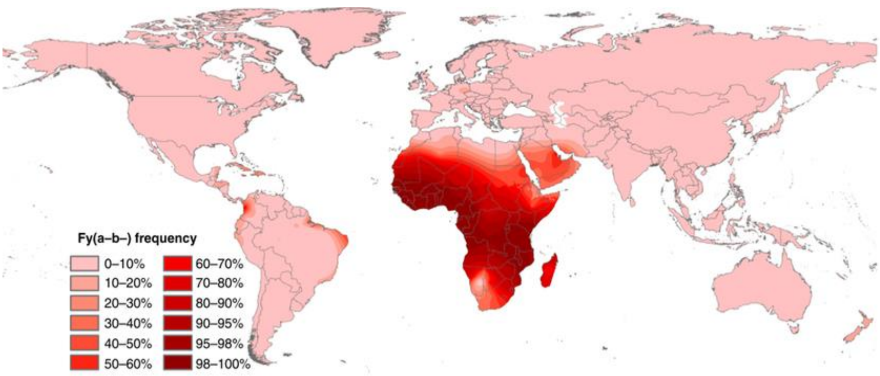


Figure 17: Global distribution of Duffy Negative Phenotype (Howes et al. 2011).

Regarding to *P. falciparum* resistance acquisition, the most paradigmatic mutation is sickle haemoglobin (HbS). A deformation of the erythrocytes that was discovered in the beginnings of the XXth century in some anemia patients (Herrick et al. 1910). The deformation is caused by a mutation in the beta globin gene, translated in a Glu>Val substitution in the 6th position of the protein (Serjeant et al. 2001). The heterozygous individuals are usually asymptomatic (Serjeant et al. 2001), but the homozygous normally die before age of 5. The elevated rates of such mutation in African populations (Figure 18) are explained because heterozygous individuals are resistant to *P. falciparum* malaria (Haldane et al. 1949). In vivo studies have confirmed such hypothesis (Min-Oo et al. 2005, Williams et al. 2006).

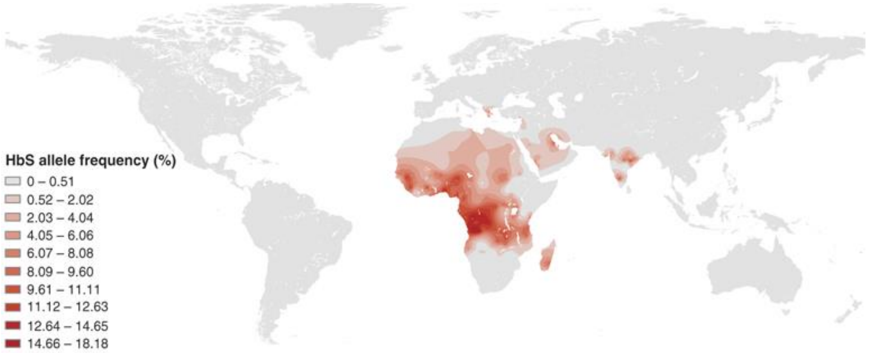


Figure 18: HbS Allele frequency distribution. The mutation is moustly found in African individuals (Piel et al. 2010).

There are other mutations located in the Haemoglobin gene (HbA) linked with malaria resistance: i) HbC chain, present both in Asian or African individuals (Gougna et al. 2010), ii) HbE chain present in Asian populations (Gougna et al. 2010. These mutations deforms the erythrocytes and prevent the infection of *P. falciparum.*

The European populations historically exposed to malaria also exhibit signals of malaria resistance. The most studied ones are found in the G6PD gene. This gene codifies an enzyme that plays a central role in the oxidative pentose pathway. If this gene is mutated, the protein is less effective and difficulties the invasion of the erythrocytes by *Plasmodium* merozoites. The G6PD B- mutation is named Mediterranean and it is found in frequencies up to 20% in some Mediterranean populations. There are other mutations present in African populations such as G6PD A- (Hirono et al. 1989).

Some putative resistant mutations have been proposed in various genes, mostly belonging to the immune system. Some of which have been functionally demonstrated as CD40LG or CD36 conferring resistance in *P. falciparum* infections (Rockett et al. 2014) while other mutations in genes like IL-10, FCGR2B, TIRAP, MARVELD3 and ATB2B4 (Khor et al. 2007; Gupta et al. 2015) are possible effective mutations deduced from GWAS studies

2. METHODS

2.1 Biological samples

The samples used in the retrieval of eradicated *P. vivax* and *P. falciparum* strains from the Ebro Delta belong to the personal collection of Dr. Ildefonso Canicio. Dr. Canicio was the head of the antimalarial centre created in 1925 by the Catalan Government in Sant Jaume d'Enveja, a village located in the middle of the Ebro Delta. The samples were collected between 1942 and 1944 and they include some information about the patients (Figure 19).

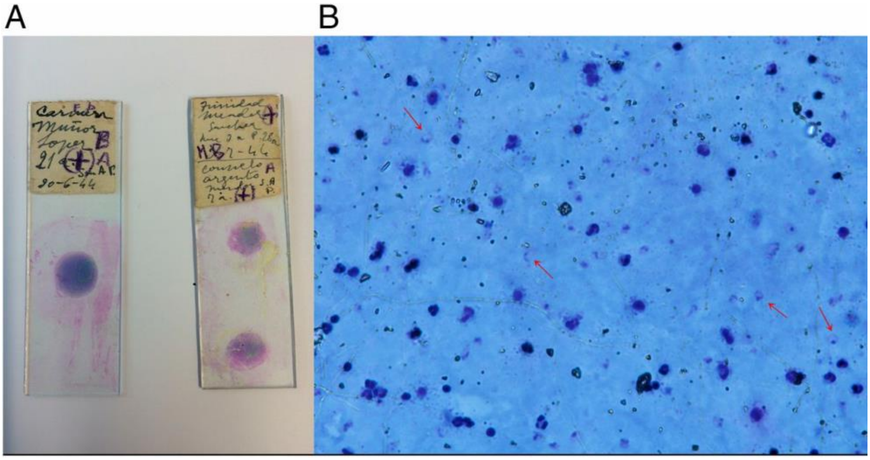


Figure 19: Giemsa stained slides used in the study of European *P. vivax* and *P. falciparum*. A) Blood slides with the patient description and the sampling date. B) Microscopic view of the sample (400x), red arrows are pointing *Plasmodium* parasites present in the blood.

2.2 DNA extraction

The percentage of endogenous DNA is the most limiting factor in aDNA retrieval and analysis. Usually the content of endogenous aDNA is lower than the 1% of the total sequenced reads, even though the rate of endogenous aDNA of different samples from the same individual can differ by orders of magnitude (Green et al. 2010). Petrous portion of the temporal bone is the densest bone in the human skeleton, its extractions usually are the ones that show the highest rates of endogenous DNA (Gamba et al. 2014, Pinhasi et al. 2015). Nevertheless in the study of pathogens from ancient remains targeting the petrous bone usually is not an option (Margaryan et al. 2018). Pathogens are targeted in different tissues depending on each particular life cycle. e. g. *Y. pestis* or *M. leprae* are often targeted in bone tissue or teeth (Schuenemann et al. 2013, Bos et al. 2011), dental cementum exhibits rates of endogenous DNA similar to those of the petrous bone (Hansen et al. 2017), furthermore the extensive blood irrigation favours the presence of systemic pathogens, as Marciniak et al. 2016 evidenced reporting the presence of *P. falciparum* DNA in Roman teeth. Other pathogens such as *M. tuberculosis* are usually targeted in the ribs, as severe pulmonary tuberculosis favours the presence of *M. tuberculosis* in these bones (Bowman et al. 2012). Those pathogens that do not infect the bone can only be targeted in soft tissues, which is restricted to rare mummified or dissected specimens: *V. cholera* (Devault et al. 2014), *H. pylori* (Maixner et al. 2016), Variola (Duggan et al. 2016).

DNA extraction protocols are typically divided in two main sections. The first one with the goal of the liberation of DNA from the rest of the cell and the second focused on the purification of such DNA (Heintzman et al. 2015). In the first stage the liberation of the genetic material is mediated by compounds like proteinase K which breaks collagen and Ethylenediaminetetraacetic acid (EDTA) which degradates hydroxyapatite (Rohland et al. 2007), the main objective is to degrade the bone tissue, liberating the DNA.

The liberated DNA is purified and separated from other organic and inorganic molecules by silica based methods or phenol-chloroform approaches (Barnett et al. 2012). Nowadays phenol-chloroform methods are outdated, being the silica based methods the most used. Within the category of silica based methods we must differentiate between the in-solution based and the column based ones. In the in-solution based method the calcified tissue is digested and the DNA captured through a cation-mediated interaction with a silica pellet (Rohland et al. 2007). Once captured, the silica pellet is washed and the DNA is liberated and recovered. This method was more recently improved by replacing the in-solution silica by silica columns (Dabney et al. 2013) developing an efficient protocol for the recovery of extra-short DNA fragments (<80 bp), which represents the broader fraction of ancient DNA reads (Orlando et al. 2015). This method was the selected for recovering the sequences presented in this thesis. The analysed blood slides were immersed in lysis solution and left in a Falcon of 5-mL tubes at 37ºC overnight. The lysis buffer was composed by the usual compounds (EDTA, Proteinase K, Tris and

SDS). The supernatant was concentrated with silica column-base. Finally there exist one third method commonly used in the aDNA recovery of long reads also based in silica columns (Yang et al. 1998). In review studies it has been demonstrated that silica column methods are the most advantageous in the recovery of short and degraded aDNA fragments, these methods are also able to recover higher fractions of endogenous aDNA reads compared with the in-solution silica method (Gamba et al. 2016).

2.3 Library preparation

Library preparation are the chemical reactions and procedures that modify the DNA fragments to able its sequencing in SGS platforms. This modifications briefly consists in the attachment of little DNA fragments known as adapters in the read ends that lead to the platform recognition (Figure 20) (Bentley et al. 2008). The most used protocol, and the one that has been used to recover *Plasmodium* genomes is the double stranded library preparation protocols. However aDNA library preparation and sequencing protocols are not standard and are usually adapted to the needs and objectives of each project

The first step of these protocol consists in fragmenting the DNA into little bits to be processed. When working with ancient samples this step can be avoided because usually aDNA is already fragmented. The ends of the tiny fragments are repaired. The reparation consists basically in the degradation of the overhanging 3' ends and the filling of 5’ overhanging ends. This reparation creates paired reads with the same read length for the two strands, which are called blunt ends.

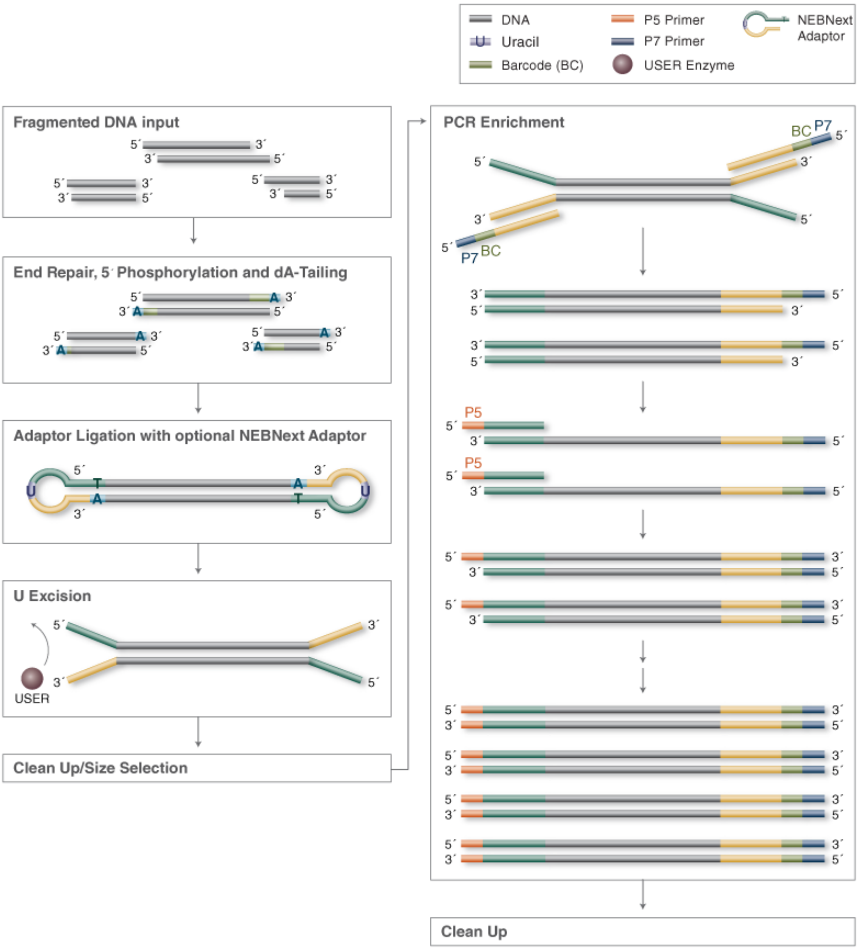


Figure 20: Library preparation method for double strand libraries. First the DNA is fragmented, and the read ends are repaired by adding A bases, afterwards the adapters are ligated to the repaired ends. Finally reads are amplified with PCR technique. (Ilumina).

Some protocols designed for aDNA sequencing treat the DNA fragments with UDG glycosylase before the adaptor ligation. This enzyme removes the Uracil bases of the DNA fragments. The resulting abasic sites are afterwards removed by the endonuclease VIII, which is also added to the solution (Seguin-Orlando 2013). The elimination of Uracil deamination products reduces the number of mismatches and the genotyping errors but also clears away the specific patterns that can be used to validate the presence of aDNA. The next step is the adapter ligation which is common for both treated and non-treated reads. The ligation can be accomplished directly by attaching two different adapters (blunt-end ligation) to the read ends or otherwise only using a single Y shaped adapter with a T-overhang that it is ligated to both ends of DNA, in this protocol read ends have previously been modified to bring A-overhangs in a process called A-tailed ligation (Fig 18). This Y shaped adapter has the disadvantage of leading to the misincorporation of T in the read ends (Seguin-Orlando 2013). After it, DNA sequences are replicated with several PCR cycles. It is important to limit, as possible, the number of PCR cycles to maintain the variability of the library. In this step the choice of an adequate PCR polymerase will be also decisive to avoid GC and read length biases (Dabney and Meyer 2012). To differentiate different libraries sequenced in the same run, barcodes are attached to the adapters during the amplification process (Craig et al. 2008) (Figure 20).

2.3.1 Single-stranded DNA library preparation

There exists one alternative SGS library preparation method, used in some aDNA analyses, known as single-stranded library. The main advantage that Single-stranded method presents in comparison with double stranded is that this method enables the incorporation of damaged reads and very short DNA reads that are normally lost in dsDNA preparation methods (Gansauge et al. 2013). The first step of this approach consist in denaturalize the DNA by heating it. Adapters are afterwards attached to the 3’ ends of the single-stranded molecules, which bind beats. Once the single stranded reads are fixed in the beats the reverse complement strand is synthetized and the sequence become double stranded.

When the efficiency of ssDNA library preparation method has been compared with dsDNA preparation methods the results have shown a higher efficiency for the ssDNA libraries (Prüfer et al. 2014). This efficiency differences can be up to 30 fold increase of endogenous DNA in ssDNA libraries (Bennett et al. 2014). Other comparative studies have confirmed the previous efficiency differences and also have arisen other relevant differences between both strategies as a lower GC content bias in ssDNA libraries (Wales et al. 2015). These benefits have to face with an elevated cost and an extended preparation time (Bennet et al. 2014) that makes this approach little selected.

2.4 *Plasmodium* targeted capture and human depletion

Ancient pathogens commonly represent very low proportions of the total sampled reads. In order to efficiently sequence the pathogen genome the prepared libraries are not usually directly sequenced. The most used techniques concentrate the endogenous pathogen reads using an array capture approach (Bos et al. 2011) or in solution baits capture (Schuenemann et al. 2011). As it was already discussed in section 1.2.1 this requires a previous pathogenic evidence as the targeted capture methods are selective for each species or targeted region. Moreover this approach can also be carried in parallel with proved efficiency in terms of economic cost and sensitivity, showing low rates of false negatives (Bos et al. 2015).

To concentrate the endogenous *Plasmodium* reads we have used an in solution whole genome capture method. We followed the MYbaits 3.0.1 protocol (Figure 21). The first step of the process consist in a denaturalization of the DNA library in the presence of adapter-specific blocking oligonucleotides. The Library and blockers are then hybridized by heating the mixture. Biotinylated RNA baits are introduced in the mixture and hybridize the targeted sequence. The Bait-target hybrids are pulled out of the solution with streptavidin-coated magnetic beads. Beads are washed several times to remove the non-hybridized molecules. The captured DNA library is released from the beads and then amplified.

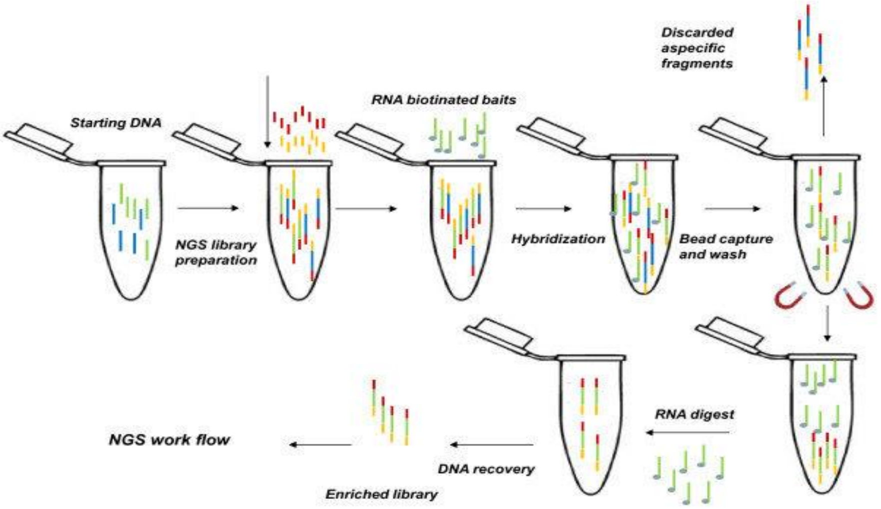


Figure 21: Diagram of an in-solution capture method: The diagram describes the steps used to concentrate the targeted sequence. (Rizzi et al. 2012).

In the present thesis, this method has been used both to enrich the *Plasmodium* reads by a classical usage with P. falciparum baits and also was used to selectively deplete the Human reads. We used human Mybaits to capture the Human reads and proceed with the sequencing of which in a normal protocol would be the waste.

2.5 DNA sequencing

There exists different commercial DNA sequencing platforms, in this thesis only Illumina has been used and all the coming section will be focused in this platform. In the Illumina sequencing platform the amplification and sequencing reactions are conducted in a flow cell. Once the library has been prepared it is introduced in the flow cell, the surface of the cell has multiple copies of forward and reverse primers that are specifically complementary to the adapters attached to the DNA fragments in the library preparation step. (Buermans et al. 2014). The linkage of these primers with the library adapters requires a previous denaturalization step of double stranded libraries, only single-stranded molecules are able to hybridize with the primers. Afterwards, the hybridized fragments are amplified in the flow cell by a reaction known as bridge amplification. This procedure consists in the ligation of a primer, located in the flow cell with the 3' free end of the fragment, allowing the synthesis of the complementary strand. Next, the double strand is denaturalized and the fragment is amplified again.

Once the amplification process has produced enough amount of fragment copies to be detected by the machine the sequencing starts. The sequencing system is based on fluorescence. First of all, a polymerase incorporates a fluorescent nucleotide which represents the complement of the template strand, the growing strand. In a second step, unincorporated nucleotides are washed and the incorporated base is identified. In order to continue with the next base and restart the cycle, the fluorescent dye of the incorporated base is removed and the growing strand incorporates one nucleotide again (Figure 22).

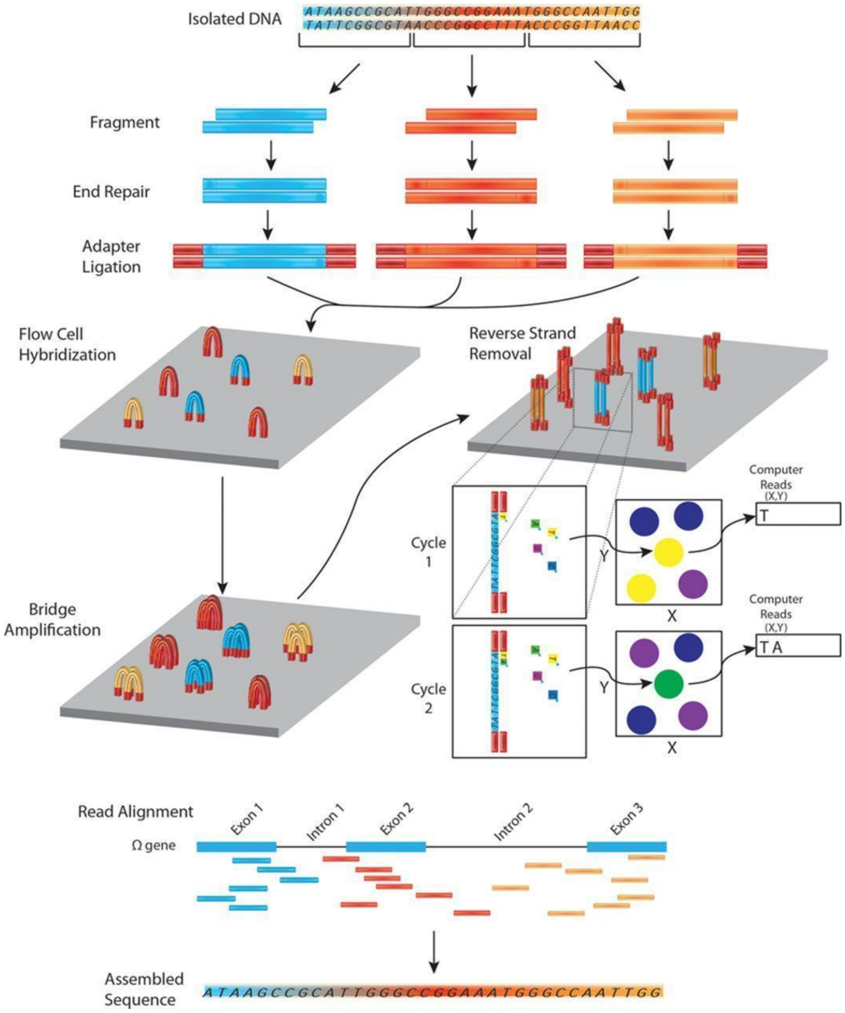


Figure 22: Illumina sequencing workflow: End repair and adapter ligation, Bridge amplification process in the flow cell and DNA sequencing by fluorescence. (Churko et al. 2013).

2.6 Read Mapping

Once the data is obtained from the sequencing platforms there are two possible approximations to reconstruct the genome. The resulting reads can be aligned to a close reference genome or can be de novo assembled. In de-novo assembly approaches the sequenced genome is built by the superposition of the sequenced reads, this method is rarely applicable in aDNA studies due to the low amount of sequence and read lengths.

Before initiating the read mapping procedure it is convenient to evaluate the quality of the libraries and to identify any problem that could imperil the quality of future analyses. FastQC (Andrews 2010) is an available software environment designed to detect possible errors in the high-throughput sequenced libraries. This platform provides easy control checks such as the presence of adapters, the presence of duplicate sequences, the quality distribution, the GC content or the sequence length distribution.

All the steps of read mapping procedures are mainly carried out by programs that have not been specifically designed for aDNA, so several options must be optimized. There are currently efforts in building specific pipelines such as EAGER (Peltzer et al. 2016) designed to face the usual aDNA mapping troubles and maximize the quality of analyses.

2.6.1 Adapter removal

The usual length of an Illumina platform sequenced read can be up to 300 bases. In the case of aDNA due to the fragmentation promoted by the post-mortem damage the length of the DNA fragment is very short, often shorter than the sequencing reads of the platform, this can lead to the partial or entire sequencing of the adapters. Which will result in misalignments or the presence of mismatches (Lindgreen et al. 2012).

There are several software products used for the adapter trimming. One of the most used is Cutadapt (Lindgren et al. 2012). This software performs the search of a given sequence of an adapter and removes it from the sequenced reads, both adapters present in the 5’ and 3’ ends. The specificities of the algorithm depend on single or paired end reads (Figure 23).

If the sequencing insert to be sequenced is shorter than the read length, the read will include part of the adapter sequence in the 3’ end (Fig 21 B and E). In the case of the paired end data (Figure 23) the two sequence reads will be identical and the adapter identification will be easier. Additionally when reads overlap, two reads are collapsed in a single one which increases the sequence quality.

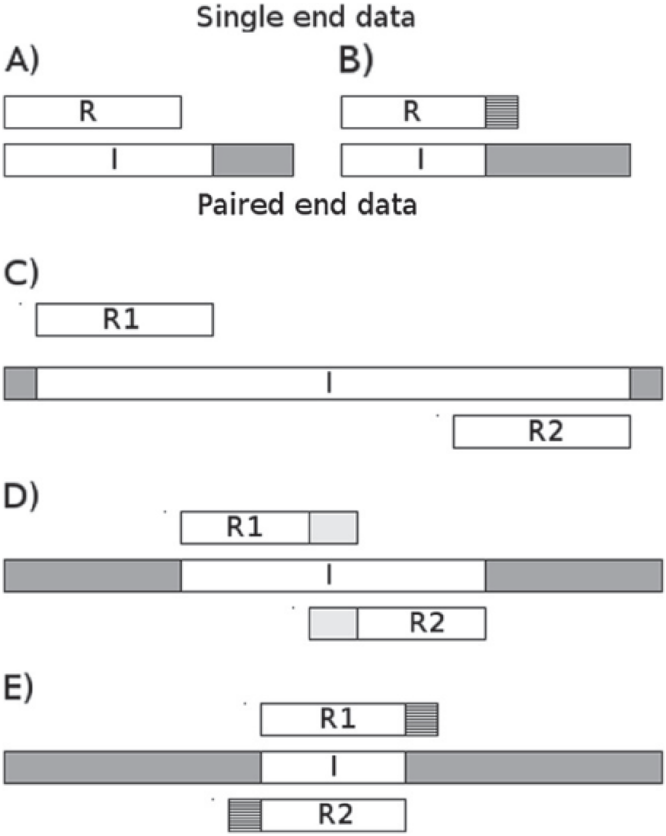


Figure 23:Cutadapt function: inserts are denoted I and R refers to single end reads; A, B, R1 and R2 refer to paired end reads; C,D,E. Read length is denoted L_R and insert length is denoted L_I. A) $L_I \geq L_R$, no contamination. B) $L_R < L_I$ adapter contamination occurs. C) $L_I \geq 2 * L_R$, no adapter contamination and no read overlap. D) $L_R < L_I < 2 \cdot L_R$, no adapter contamination but the two reads overlap. E) $L_I < L_R$, adapter contamination in 3' ends of both reads, overlap between 5' ends of reads. (Lindgreen et al. 2012).

2.6.2 Mapping Software

As it occurs in all the steps, mapping algorithms are not optimised to deal with the particularities of aDNA. To construct a usable genome, an efficient and precise alignment is needed. The most used software product to map the trimmed reads is BWA (Li et al. 2009).

BWA normally performs an initial alignment using few based of the 5’ end. This procedure is called seeding and it is performed to speed up the mapping process. Working with aDNA, this option must be disabled as the deamination patterns, accumulated in the read end (Briggs et al. 2009, Green et al. 2010) would be a source of misalignments (Schubert et al. 2012). Additionally it is also recommended to eliminate short reads (<30 bp) and to increase the tolerated edit distance to avoid an extreme loss of genetic material due to DNA degradation modifications. This relaxed mapping threshold will produce an increase of mismatches that must be afterwards removed from the variant data.

2.6.3 Post-mapping procedures

Mapped reads must be processed after being aligned with a reference genome, in terms of quality filtering and mapping accuracy. Library preparation methods include the use of PCR to amplify the DNA fragments. This amplification produces multiple copies of the same DNA fragment. Usually the sequencing platforms sequence multiple times the same DNA fragments. The fraction of duplicated reads is higher in aDNA libraries as the original amount of DNA is significantly lower than in modern DNA samples. To remove these duplicated reads Picard tools MarkDuplicates has been presented as the most efficient tool (Schubert et al. 2012).

Mapping software products assign a mapping quality score to each mapped read. This numeric value indicates the probability that the mapped read is misplaced. This value is calculated by the sum of mismatches in the alignment. Normally in this scales as higher is the value, lowest is the probability of the misplacing. (Li et al. 2008). In the case of aDNA, reads exhibiting mapping qualities below 25 should be removed (Schubert et al. 2012).

2.7 Data quality software

There are software products designed to assess the quality of the sequenced mapped reads, most of which have been previously introduced. It is also relevant to highlight those that are specifically designed to evaluate the aDNA particularities.

2.7.1 FastQC

FastQC (Andrews 2010) is a framework that integrates diverse quality analysis that provides an integrative and easy representation of the raw sequencing data. The analysis is divided in modular sets that lead to the identification of specific parameters that can be problematic, some examples are: the distribution of the fragments length, the GC content or the presence of sequencing adapters, etc.

2.7.2 PMDtools

PMDtools (Skoglund et al. 2014) is a framework specifically designed and conceived to evaluate the likelihood of a sequence to be a post-mortem degraded read. This software allows the identification of true aDNA reads by the recognition of patterns of aDNA damage in the analysed sequence. Additionally, the software also allows the calculation of the contamination present in the sample. The software calculates the likelihood for each read, which is translated in a PMD score, for which positive values indicate support for the sequence being genuinely ancient. As higher the threshold is settled, the confidence of the authenticity of the filtered reads increases. However the possibilities of losing authentic endogenous reads with little damage also increases.

2.7.3 Mapdamage

Mapdamage (Ginolhac et al. 2011) is a software with an improved version called Mapdamage2 (Jonsson et al. 2013) that reports the DNA damage patterns in SGS reads. This software allows the quantification and easily visualization of cytosine deamination and other kinds of substitutions. This software also allows visualizing other misincorporations such as deletions or the presence of adapters in the sequence. As it indicates the fraction of cytosine deamination it can be used to test the degradation state of the sample and to predict the possible arrangements during the variant calling process to minimize the effect of the post-mortem damage.

2.8 Variant Calling

Variant calling is the process to discover and annotate polymorphisms in the sequenced data. The input data for most of the variant calling pipelines are the mapped reads, for example BAM files, and the output data are files with the identified variants plus quality indicators and the desired annotations. The most used format is Variant Calling Format files (VCF) files.

One of the most commonly used variant calling pipelines is the Genome Analysis Toolkit (GATK) (McKenna et al. 2010). This software is specifically designed to identify SNP and little indels from genome sequencing data using a reference genome. There are other software products that are also designed for this purpose such as VarScan (Koboldt et al. 2009). In this Thesis only GATK has been used.

GATK is not a unique utility, is a complex platform that includes multiple tasks such as pre-processing, data analysing or post-processing quality control and annotations. For the calling variants step there are two main algorithms included in the GATK platform that can be used; UnifiedGenotyper and HaplotypeCaller. HaplotypeCaller is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region. UnifiedGenotyper uses a Bayesian likelihood model to estimate the most likely genotypes in the variant positions.

The output VCF files show the variant files and other information such as the quality of the positions or the mapped reads that carry the reference and the derived allele. The platform provides multiple annotations that can be used to tag the discovered variants such as allele depth, allele frequency, etc. These annotations are used to filter the variants with the desired criteria

The natural degradation and post-mortem damage characteristic of aDNA are not incorporated in most of the Variant Calling analysis. Recently this mutational pattern has been integrated in a python script named AntCaller (Zhou et al. 2017). The method uses the damage information to compute the posterior probabilities of each genotype reducing the false discovery rate caused by misincorporations. It has been demonstrated that AntCaller outperforms the accuracy of GATK, nevertheless this improvement is not significant in samples with high coverages, were the misincorporations can be filtered.

When analysing whole genome sequence samples with very low coverages usually GATK is not used. In this cases the genotypes are assessed generating pseudo-haploid calls by randomly selecting one sequenced read for each position (Skoglund et al. 2012; Allentoft et al. 2015; Haak et a. 2015; Mathieson et al. 2015; Fu et al. 2016; Lazaridis et al. 2016).

2.9 Imputation

Genotype imputation methods use genotype data from a panel of reference samples, to infer the genotypes of missing variants in the target samples. There are multiple available reference panels such as the 1000 genomes genotypes or the HapMap datasets. The imputation procedure uses Linkage disequilibrium to detect the presence of haplotypes in a population. Afterwards this haplotypes are used to assess the missing variants of the samples included in those detected haplotypes. This pipeline is integrated in software such as Beagle (Browning et al. 2016).

Imputation has been already used in aDNA studies showing accuracy rates above the 99% of the imputed genotypes, in individuals with depth coverage of 1x (Gamba et al. 2014). Genotypes of samples with mean depth of coverage much below 1x are no suitable to be assessed by imputation, although there is not a clear critical value. A recent study involving 67 ancient Europeans genomes has situated the threshold for obtaining confident genotypes in only 0.85x (Martiniano et al. 2017). Imputation approaches have been mainly used to improve the outcomes of population genetics analysis, however, genotypes of ancient samples have been also used to detect selection in ancient individuals (Ye et al. 2017).

The usage of this approximation in low coverage samples reduces the reference bias compared with the traditional pseudo-haploid calling method, which simply means to choose random reads, and which produces much more genotyping mistakes that the imputation of those genotypes. Nevertheless, it is important to remark that although the rate of missasigments is lower in imputed genotypes, it still exists. The rate of genotyping errors is dependant of the MAF of the SNPs present in the reference panel. SNPs with very low MAF cannot be accurately imputed (Martiniano et al. 2017) (Figure 24).

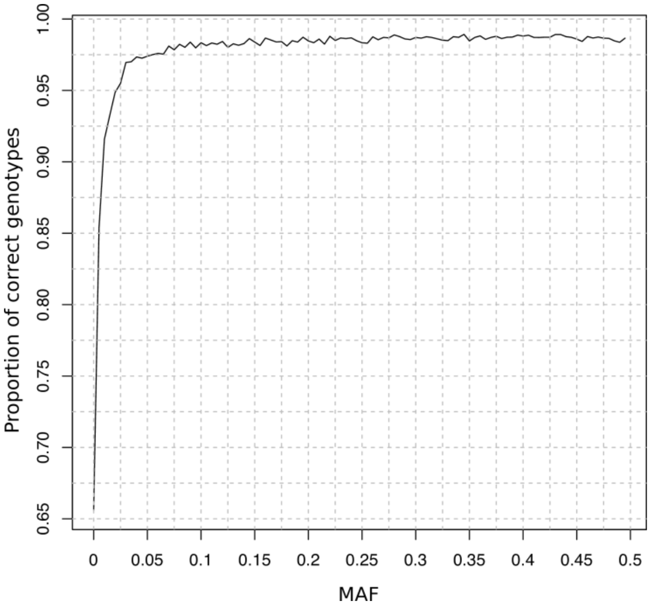


Figure 24: Imputation efficiency. The proportion of correct imputed genotypes is dependent on the allelic frequencies of the SNPs in the panel. (Martiniano et al. 2017).

2.10 Population genomics analyses

2.10.1 Phylogenies

Phylogenies are one of the most common representations of the biological diversity. Phylogenetic trees show the relations of biological samples or species using branches, nodes and taxa. The branching pattern of a tree is known as topology. Nodes define clades, which is the name given to the group of descendants of a node. If a clade is composed by all the descendants of a common ancestor, this clade is monophyletic. If not, the clade is paraphyletic.

Phylogenetic trees can be built with genetic or morphological diversity. One of the most used methods to build a genetic phylogenetic trees is Maximum Likelihood (ML). In Statistical terms ML is a method for estimating unknown parameters in a probability model, in other words ML provides a probability of the sequences given a model of their evolution on a particular tree. The resultant likelihood value is the probability of an observed sequence assuming a specific evolutionary model. The probability is expressed in logarithm, so as less negative is, the greater the probability is. This method is used to reconstruct phylogenies and is implemented in multiple software products (Figure 25).

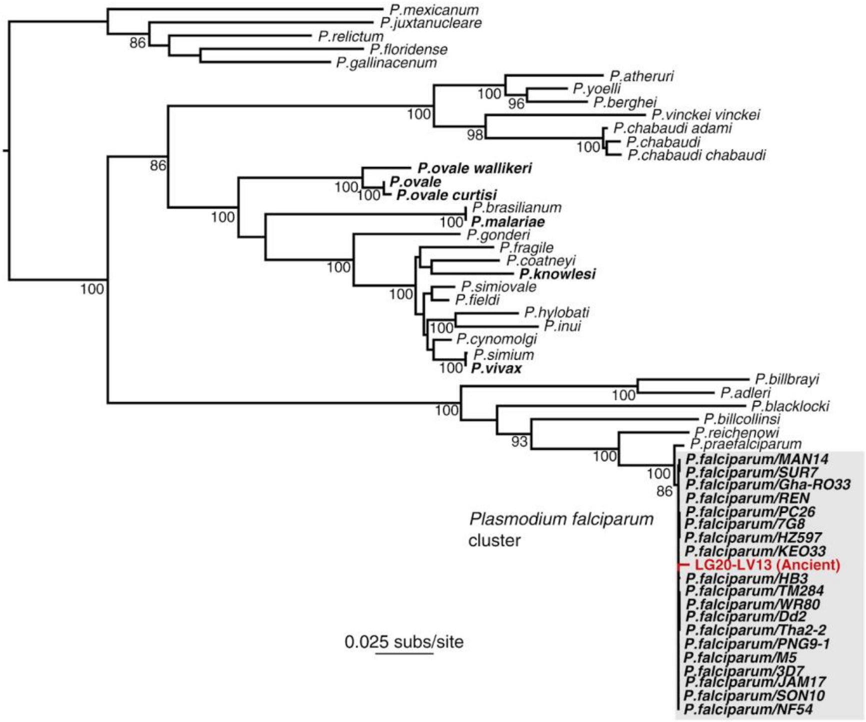


Figure 25: Maximum likelihood tree of *P. falciparum*. The tree has been rooted using mammalian species of *Plasmodium* (Marcinak et al. 2016)

Phylogenetic trees can be presented rooted or unrooted. To root a tree it is required to know the ancestor of all the taxa present in the phylogeny. If an outgroup is added, it is possible to identify the ancestral node of the phylogeny and root it. If the tree is rooted, the branch lengths can be interpreted as time estimates.

2.10.2 Network tree

Median Joining network algorithm (Bandelt et al. 1999) combines features of the minimum spanning algorithm and maximum parsimony. This method allows the construction of phylogenetic networks that present very low levels of reticulation, using polymorphic data such as SNPs or STRs (Figure 26). Reticulation is the graphical representation of biological process such as recombination or parallel mutations, these process produce closed structures between samples in the represented phylogenies.

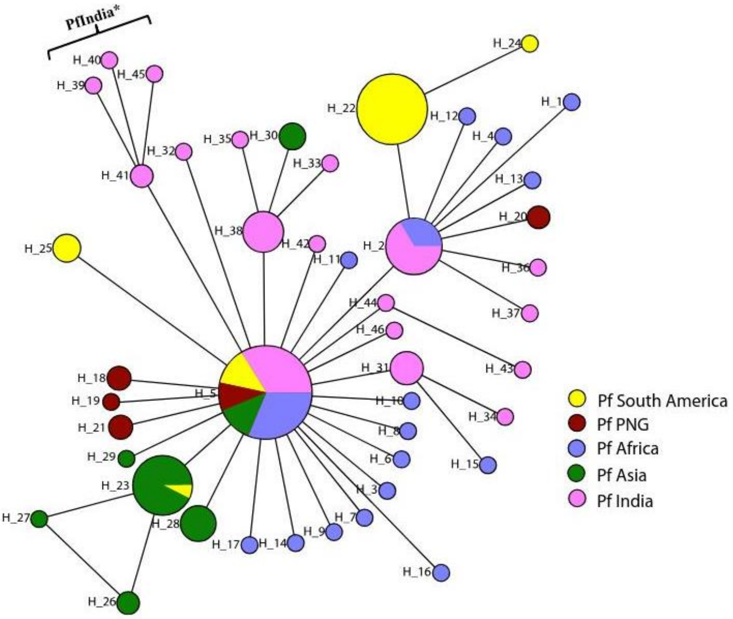


Figure 26: Median joining Network tree built with *P. falciparum* mitogenomes. Each colour represents a population. The dot size is proportional to the number of samples clustered. The distances of the branches between dots are proportional to the number of substitutions between clusters. (Tyagi et al. 2014).

Network is a software product used to build phylogenetic trees. Also is designed to estimate the dating of the ancestors in the trees. This software uses multiple alignments of nucleotide sequences as input. These alignments are usually generated with software products such as Muscle (Edgar et al. 2004), Clustal Omega (Sievers et al. 2011) or MEGA (Tamura et al. 2007).

Network software has been previously used to illustrate diverse alignments the software allows the building of networks based on two differentiated methods; the reduced median and the median-joining (http://www.fluxus-engineering.com/). This is an adequate method to visualize the relation between samples when the number of markers used to illustrate those relations is limited such as in the case of mtDNA, including *Plasmodium* (Rodrigues et al. 2014, Tyagi et al. 2014, Rodrigues et al. 2018).

2.10.3 Principal Components Analysis

Principal components analysis (PCA) is a mathematical procedure used to reduce the number of variables that define a dataset by correlating small variables in major ones called principal components. These components define most of the genetic diversity. The components are sorted in descending order according to the rate of diversity explained.

PCA is built using genotype data with software products such as SmartPCA developed by Patterson et al. 2006 and integrated in the EIGENSOFT software. Otherwise it is also possible to refine the PCA by using haplotype data substituting the genotype data. This kind of analysis is implemented in fineSTRUCTURE software (Lawson et al. 2010).

PCA is a useful and powerful method for the genetical analysis of population although there are several cases or specific population characteristics that makes PCA analysis difficult. Particularly, in the case of highly structured samples, large populations, datasets with a very distant population or closely related subpopulations, the method could face difficulties in properly assigning the individual in a subpopulation (Intarapanich et al. 2009).

2.10.4 Admixture

Admixture is a population genetics software product based on maximum likelihood estimations, which makes this software extremely fast compared with other Bayesian approximations. The software models the probability of the observed genotypes by the usage of ancestry proportions, simultaneously estimating population allele frequencies along with ancestry proportions (Figure 27). Admixture is a useful tool to define and plot the genetic components of populations, even though it is not an adequate software product to infer in ancestry relations.

The admixture input are the genotype data from the sampled individuals as well as the estimation of the genetic components (K) that define the populations (Liu et al. 2013).

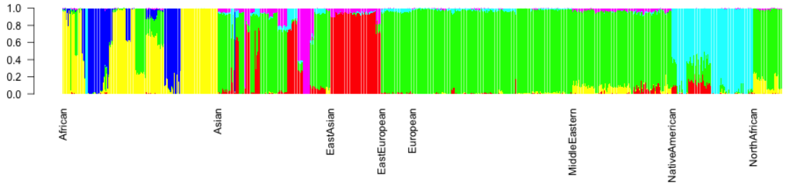


Figure 27: Admixture plot. The plot shows the genomic components of individuals from diverse populations. Each colour represents a genomic component. Each bar represents one individual. The Y axis shows the percentatge of each component in each individual (Liu et al. 2013).

Dense marker sets must be pruned before being analysed in the Admixture software, sites in Linkage Disequilibrium (LD) can be a source of errors in the ancestry block building (Alexander et al. 2009), as this would lead to the detection of false components.

2.10.5 Fixation index

Fixation index is statistical tool used in population genetics to differentiate populations based on small polymorphisms data such as; SNPs or Microsatelites. This statistic tool has been developed from the Wright F-statistic (Whright, 1951). The computation of the F_{st} is based on the correlation of randomly selected alleles within the same subpopulation relative to the whole population. This concept extrapolated in terms of population comparison means the proportion of diversity between populations due to the allele frequency diversity between populations (Holsinger et al. 2009). Is used to detect variants with differential allelic frequencies that may be under the pressure of selective forces.

2.10.6 F-statistics

Wright's fixation index (Wright 1951) also referred as F-statistic is a measurement of population differentiation. It is a common used population genetic statistic to compare individuals or populations. The source data used in this kind of studies is normally SNP but also other genetic polymorphisms such as microsatellites can be used in this approach. The term of F-statistic refers to a specific framework developed by Reich et al. in 2009. This approach developed both an admixture test for populations and a proof of population relationship. Other applications proposed and developed by authors include the elucidation of the number of funders of a population (Reich et al. 2012; Lazaridis et al. 2014), a quantitative approach in terms of admixture proportions in a population (Green et al. 2010; Haak et al. 2015) or studies focused on explaining complex demographic histories according to splits and mixture events (Patterson et al. 2012; Lipson et al. 2013). The models used to solve the explained applications are called admixture graphs (Figure 28).

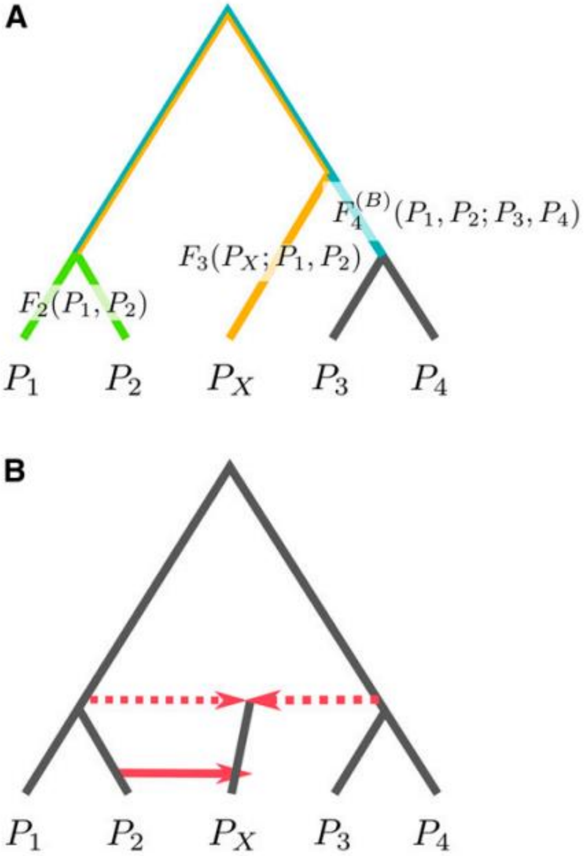


Figure 28: F-statistics: A) Population phylogeny with branches corresponding to F2 (green), F3 (yellow) and F4 (blue). B) Admixture with phylogeny by allowing gene flow (red, solid line) and admixture events (red, dotted line) (Benjamin et al. 2016).

We can differentiate diverse statistics according to Reich et al. 2009 depending on the number of populations analysed.

- F2: Corresponds to the phylogeny from P1 to P2.
- F3: Defines the relation of Px to P1 and P2.
- F4: The branch connecting P1 and P2 to P3 and P4

F3 or three populations method is frequently used as an admixture test for the test population (Px). The F3 statistic calculates the difference of allele frequency shared between Px and P1 with Those shared between Px and P2: f3 = (Px-P1)(Px-P2) (expressed in terms of shared allele frequencies). In a admixture event between P1 and P2 that results in the advent of Px the product of the two comparisons will be negative as the position of Px is intermediate between P1 and P2. An particular application of F3 statistics is outgroup F3 statistics that uses the statistic to determine measure the genetic drift shared between two test populations from an outgroup population. In Most of the implemented tests use a weighted block jackknife (Busing et al. 1999) to provide significance of the static.

F4 statistic was also introduced by Reich et al. in 2009. Is a very powerful tool for detecting introgression and as all the other F-static is based on the allele frequencies of four studied populations. F4 (P1-P2)(P3-P4). The statistic is the product of differences between the two pairs of populations. In a non-related situation, the frequencies of the fourth populations should be independent so the product of the differences should be 0. If there existed introgression between analysed populations the f4 results would be no zero. In this analysis also block jackknife is applied for statistical significance.

2.10.7 Haplotype Sharing Based softwares

Chromopainter is a software product designed for the study of admixtures based on SNP phased data. The software uses haplotype data from donor chromosomes to paint the target chromosomes. These donor chromosomes represent the source of admixture of the target populations. The main basis of the software is that each individual is defined by genetic components form all other individuals. The output of chromopainter can be the sample haplotypes series known as "chromosome chunks" o the number of recombination events at all sites. Normally Chromopainter output data is used to generate haplotype based Principal Component Analysis (PCA), for individual clustering (FineSTRUCTURE) or for dating admixture events (Globetrotter).

FineSTRUCTURE (Lawson et al. 2011) uses the output data generated with Chromopainter to cluster the individuals both based on the shared haplotypes and the recombination pattern.

Admixed populations should present segments of DNA from the contributing population. This principle is the basis that Globetrotter (Hellenthal et al. 2014) uses. The sizes of the segments inherited decreases over generations as a result of recombination. This approximation is shared with other popular softwares designed for

Admixture dating as ALDER (Loh et al. 2013). The haplotype chunks generated with Chromopainter to determine if any of the target populations derives from any sampled population.

2.10.8 BEAST

The evolutionary history of live can be inferred by the distribution of the populations or species, by dating fossil records or by accessing dates based on molecular diversity. One of the challenges for evolutionary biologists is to merge in a proper way all these information and combine it. Bayesian Evolutionary Analysis by Sampling Trees (BEAST) (Drummond et al. 2007) is a software specifically designed to answer these questions. This software uses a Bayesian approximation to combine the prior known information with the one obtained from Bayesian probabilities performed with molecular data.

The algorithm implemented in BEAST software is based on
Bayesian Markov chain Monte Carlo (MCMC) method, which has been deemed appropriate for phylogeny reconstruction in software products such as Mrbayes (Huelsenbeck et al. 2001). The typical posterior probabilities of Bayesian probabilities in BEAST are the evolutionary parameters of the Given molecular sequence. The most known utility of this software is to add an time-scale to rooted trees

providing calibrated phylogenies and genealogies (Drummond et al. 2007) .

To run the software the needed input is a Newick format tree, a format that summarized the information into comas and points. Additionally in order to calibrate the desired phylogeny a set of priors must be established. These priors are parameters such as a predefined mutation rate, a substitution model or population growth model. An important issue related with the proper running of Beast relays on the choice of such values. Before implementing the selected parameters into the phylogenetic model, the accurateness of the prior values must be checked with Tracer. This software is used to evaluate the convergence of the settled parameters across the multiple replicates of the Bayesian estimations. (Nylander et al. 2007).

2.10.9 Selection software

Selection software (Schraiber et al. 2016) is a useful tool specifically designed to test the presence of selection driving allelic frequencies in past populations time series. The software, based on Bayesian inference, models the evolutionary trajectory of an allele in a defined demographic history (Ye et al. 2017). Variations in allelic frequencies across multiple sampling times are used to estimate the selection coefficients for both heterozygous and homozygous haplotypes (Figure 29).

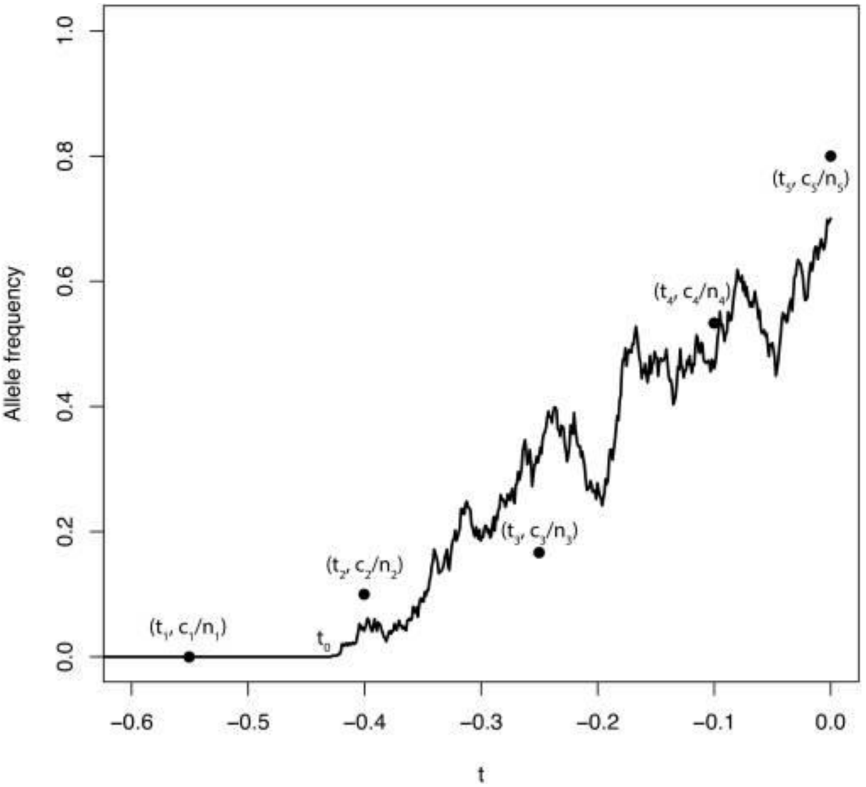


Figure 29: Simulation of the variation of the allelic frequency in a timeline. In each observation (t_x) a c observed counts of derived alleles in a n size population (Schraiber et al. 2016).

This method presents a clear advantage in comparison with the classical approaches based on the comparison of allelic frequencies between two different periods (Mathieson et al. 2015), as is capable to study multiple sampled populations from different sampling times. The framework is also capable to estimate the age of the allele under selective pressure, if this estimation is also implemented the value of the selective coefficients represent the selective forces driving the allele frequencies since the arose of the variant to the last sampling time. Some information must be provided to adjust the estimations:

The sample size, the generation time, the sampling times and the number of derived alleles per observation time.

This approach is based on the idea of population continuity, but few tests and applications have been proposed to examine population continuity. There are tests that restrict the analysis to a single locus in large samples sizes (Sjödin et al. 2014) or to a single individual with genome wide data (Rasmussen et al. 2014). Recently a relevant publication has presented a new approach designed for the most usual aDNA situation: multiple samples and low coverage (Schreiber et al. 2018). Relevantly when this approach was tested, continuity was rejected in all the European populations analysed.

3. OBJECTIVES

The main objective of all the work presented in this Thesis is to study the evolutionary history of malaria in Europe both focusing on the pathogen and in humans.

And specifically:

- Retrieve *Plasmodium* DNA from ancient blood slides from medical collections.
- Sequence the nuclear and mitochondrial DNA genomes of ancient European *Plasmodium vivax* and *Plasmodium falciparum*
- Use the recovered genetic material to asses migrating movements of both *P. vivax* and *P. falciparum*
- Explore the genome of both *P. vivax* and *P. falciparum* in the direction to provide relevant information concerning adaptive features.
- Use the sequence of Ancient European *Plasmodium* to calculate a mutation rate.
- Define the genetic landscape of ancient European populations by genotyping genetic variants related with the resistance against malaria.
- Identify signals of malaria adaptation in ancient European published genomes
- Use all the generated data to elucidate the age for the presence of *Plasmodium* in Europe and America.

4. RESULTS

4.1 Mitochondrial DNA from the eradicated European *Plasmodium* vivax and P. *falciparum* from 70-year-old slides from the Ebro Delta in Spain

Gelabert P, Sandoval-Velasco M, Olalde I, Fregel R, Rieux A, Escosa R, 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. 2016 Oct 11;113(41):11495–500. DOI: 10.1073/pnas.1611017113

4.2 Malaria was a weak selective force in ancient Europeans

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

4.3 An eradicated European *Plasmodium vivax* strain retrieved from antique medical slides sheds light on its dispersal

Pere Gelabert, Lucy van Dorp, Adrien Rieux, Toni de-Dios, Shyam Gopalakrishnan, Christian Carøe, Marcela Sandoval-Velasco, Rosa Fregel, Iñigo Olalde, Raül Escosa, Carles Aranda, Silvie Huijben, Ivo Mueller,. Francois Balloux, Thomas P. Gilbert and Carles Lalueza-Fox.

In revision

An eradicated European *Plasmodium* vivax strain retrieved from antique medical slides sheds light on its dispersal

Pere Gelabert[1], Lucy van Dorp[2], Adrien Rieux[3], Toni de-Dios[1], Shyam Gopalakrishnan[4], Christian Carøe[4], Marcela Sandoval-Velasco[4], Rosa Fregel[5], Iñigo Olalde[6], Raül Escosa[7], Carles Aranda[8], Silvie Huijben[9], Ivo Mueller[10,11,12], François Balloux[2], M. Thomas P Gilbert[4,13] and Carles Lalueza-Fox[1]*

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

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

[3] CIRAD, UMR PVBMT, St. Pierre de la Reunion, France

[4] EvoGenomics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen, Denmark

[5] Department of Genetics, Stanford University, Stanford, California, United States

[6] Department of Genetics, Harvard Medical School, Boston, 02115 MA, United States

[7] Consorci de Polítiques Ambientals de les Terres de l'Ebre (COPATE), 43580 Deltebre, Spain

[8] Servei de Control de Mosquits, Consell Comarcal del Baix Llobregat, 08980 Sant Feliu de Llobregat, Spain

[9] Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, 85281 AZ, United States

[10] ISGlobal, Barcelona Institute for Global Health, Hospital Clínic-Universitat de Barcelona, 08036 Barcelona, Spain

[11] Population Health and Immunity Division, Walter & Eliza Hall institute, Parkville, 3052 VIC, Australia

[12] Department of Medical Biology, university of Melbourne, Parkville, 3052 VIC, Australia

[13] Norwegian University of Science and Technology (NTNU) University Museum, N-7491 Trondheim, Norway

Abstract

Plasmodium vivax, the most widely geographically distributed malaria parasite, was eradicated from Europe in the second half of the 20th century. Although several studies have tried to reconstruct how it spread, the lack of genomic information from past European strains has prevented a clear understanding of its phylogeographic origin and structure. We therefore exploited a recently discovered set of medical microscope slides prepared in 1944 from malaria-affected patients in Spain's Ebro Delta to generate a first complete genome from a now eradicated European strain. We find this strain falls basal to a cluster including the most common New World strains, and take advantage of its known age to estimate a new P. vivax pan-genome mutation rate. Using this rate, we estimated the Last Common Ancestor for the cluster of European and American strains to be around 1455 CE. Together with historical accounts, this suggests a post-Columbus introduction of the pathogen to the American continent from a European source. We furthermore estimated that present, continental-wide diversity of P. vivax is only about 2,000 years old. Together with the observation that some known variants for resistance to anti-malaria drugs were already present in a strain predating their use, the availability of a genomic mutation rate opens up the opportunity to model the emergence and spread of resistance mutations. We conclude that the future recovery of P. vivax genomes from other antique medical collections, ancient bones will help reconstruct the ancient genomic diversity of this parasite and understand its complex dispersal.

Introduction

Malaria inflicts a dramatic disease burden with an estimated 200 million people infected annually, and around 429,000 fatal cases [1]. The disease is caused by several species of parasitic protozoans from the genus *Plasmodium,* which is transmitted by various species of mosquitoes from the genus *Anopheles*. Two species in particular - *P. falciparum* and *P. vivax* - are responsible for the majority of human infections worldwide, and although *P. falciparum* causes 99% of malaria deaths globally [1], *P. vivax* is the aetiological agent of 42% of all cases outside of Africa. Furthermore, in contrast to *P. falciparum, P. vivax* is capable of producing recurrent malaria episodes due to its resistant latent forms known as hypnozoites [2,3]. This capacity allows *P. vivax* to maintain itself in temperate climates, resting in a dormant state in the cold months where anopheline population is in diapause, and creating a persistent presence of parasite reservoirs [4].

Today, the endemicity of genus *Plasmodium* is restricted to tropical and subtropical latitudes, spanning large regions of East and South-East Asia, Sub-Saharan Africa, Central and South America and Melanesia [5,6]. However, malaria was historically present in most of Europe, from the Mediterranean to the southern shores of the Baltic Sea, and from southern Britain to European Russia [7], transmitted by different local mosquito species belonging to the genus *Anopheles*. Malaria was eradicated from all European countries during the second half of the 20th century [8], with Spain being one of its last footholds from which it was only declared officially eradicated in

1964 [9]. Nevertheless, even though *Plasmodium* is currently extinct in Europe, its re-emergence has been identified as a plausible consequence of climate change [10,11].

P. vivax is widely considered to have emerged in sub-Saharan Africa, a region in which it is now extinct [12,13]. From here, it is believed to have spread globally through a complex pattern of migration events by hitchhiking with its human host, as humans spread out of Africa [14,15]. For example, analysis of a geographically diverse sampling of 941 *P. vivax* mitochondrial DNA (mtDNA) genomes detected genetic links between strains from the Americas with African and South Asian isolates, although also identified possible contributions from Melanesia into the Americas [13]. It is also likely that given Europe's role in facilitating global movement, it played a central role in the dispersion of *P. vivax* in historical times. The recovery of genome sequences from the now-extinct European strains is critical for improving our understanding of the biology and phylogenetic history of *P. vivax*. The recent discovery of a set of microscope slides with bloodstains from malaria-affected patients from an area where the disease was transmitted by *A. atroparvus*, the Ebro Delta (Spain), dated between 1942-1944, allowed the first retrieval of genetic material from the historical European *P. vivax* strains. The complete mtDNA genome of this strain showed a genetic affinity to the most common present-day South and Central American haplotypes, suggesting their introduction into the Americas was linked to Spanish colonial-driven transmission of European strains [16]. However, mtDNA is a maternally inherited single locus and, in comparison to the entire genome, has a limited power to reconstruct complex evolutionary histories.

In this paper, we report the complete genome of an extinct European *P. vivax* obtained from the antique Ebro Delta microscopy slides. This European *P. vivax* genome, together with the recent publication of a reference *P. vivax* genome from a Papua-Indonesian patient isolate (PvP01) [17] opens up new opportunities to resolve the historical dispersal of this parasite using genome-wide data. The availability of this historical genome allows us to compute an accurate evolutionary rate for *P. vivax* and date the clustering of the European and the American strains. Furthermore, the availability of an old genome allows us to ascertain the presence of some resistance-alleles prior to the introduction of most anti-malaria drugs. This information, together with the new evolutionary rate, is critical for further investigation of the evolution of the parasite as well as prediction of the future emergence of drug-resistance mutations.

Results

We generated shotgun Illumina sequence data from four archival blood slides derived from malaria patients sampled between 1942 and 1944 in Spain's Ebro Delta, although the overwhelming majority of reads that mapped to P. vivax (90.18%) derived from a single slide (1a). In total, 446,529 DNA reads mapped to P. vivax, yielding genomic data at 1.62x coverage, and spanning 66% of the Salvador1

(Sal1) reference. The mtDNA genome was recovered at 32x coverage (Supplementary Tables 1 and 2).

We initially conducted several population genomics data analyses to explore the phylogeographic affinities of the eradicated European strain (Supplementary Table 3). A principal component analysis (PCA) on the P. vivax nuclear genome data (Figure 1b) showed strong geographic structure with at least three main clusters separating 1) South East Asian/East Asian samples, 2) Oceanian samples and 3) those from India, Africa and Central/South America. The European strain (labelled Ebro-1944) falls within the diversity of the latter cluster, being related to strains from Mexico, Brazil and also Mauritania.

Unsupervised ADMIXTURE model-based clustering (Figure 2) provided qualitatively consistent inferences to those observed in the PCA, with one ancestry component maximized in South East/East Asian samples, one maximised in Oceanian samples and another one that is prevalent in Central/South Asia, Africa and Central/South American samples. Ebro-1944 shares 100% of the latter ancestry component. Some samples from Western Thailand, Myanmar, China and Vietnam have an inferred component also shared with American, African and our European strain (blue) possibly reflecting shared ancestry or admixture between these strains.

To explore these relationships more formally, we calculated f4 statistics of the form (P.cynomolgi, Ebro-1944, Papua-New Guinea, X), where X is tested for data from [17] worldwide strains (Figure 3), and P. cynomolgi represents the closest out-group genome sequence available. Specifically we tested which populations share significantly more drift with Ebro-1944 relative to Papua New Guinea, the most genetically extreme population for which we have a good number of samples. Figure 3 shows that Ebro-1944 shares significantly more genetic drift with Central and South American samples than those from South East and East Asia, suggesting the presence of a cline of ancestry stretching from North-Western Europe to the Americas. Additional f4 statistics were generated combining strains from all continents (Supplementary Figure 8); in these tests Mexico (n=5) and Brazil (n=11) consistently have the most significantly positive f4 scores. Comparing the scores under f4 (P.cynomolgi,Ebro-1944,Mexico,Brazil) and f4 (P.cynomolgi,Ebro-1944,Brazil,Mexico), Ebro-1944 shows a greater affinity to Mexico. This supports a topology where Ebro-1944 is closer to Central rather than South American strains, although the precise relationship remains contentious.

Additional evidence for this relationship was also obtained using an independent method designed to cluster samples based on inferred patterns of haplotype sharing between global strains. Specifically using fineSTRUCTURE, Ebro-1944 was again found to cluster with samples from Brazil and Mexico, consistent with a Central/South American-like ancestry. However, it also supports the existence of commonality in the patterns of haplotype sharing with other strains from the Americas, Africa and India (Figure 4), suggesting shared common ancestry relative to the other strains sampled. This result is robust when ignoring haplotype information (Supplementary Figure 9).

Genomes obtained from historical or ancient materials provide unique opportunities to calibrate phylogenetic trees by associating sampling dates directly with the sequences representing the tips (terminal nodes) of a phylogenetic tree. These in turn enables inference of divergence times and mutation rates without the need for any other age-related external data [18]. We first used our historic sample in conjunction with 14 other closely related public genomes to demonstrate that this dataset had sufficient temporal signal for tip-dating inferences to be performed (Supplementary Table 5). We observed a significant positive correlation between root-to-tip distances and sampling times suggesting the presence of detectable temporal accumulation of de novo mutations within the timescale of our dataset (Figure 5a). Mutation rates were subsequently estimated with the Bayesian phylogenetic tool BEAST [19] for each partition of nucleotides (non-exonic, exonic synonymous and exonic non-synonymous). Non-exonic substitution rates were found to be 1.30x and 2.34x higher than exonic synonymous and non-synonymous rates, respectively (Figure 5b). Averaged at the whole-genome scale, we estimate a mutation rate of 4.90E-7 substitution/site/year (HPD 95% 1.07E-7 – 1.02E-6), which results in the emergence of 11.3 substitution per genome/year (HPD 95% 2,4-13,5). We obtain low values (<0.2) of ucld.stdev (the standard deviation of the uncorrelated log-normal relaxed clock) for each partition, suggesting very little variation in rates amongst branches. To our knowledge, no other direct mutation rates exist for P. vivax, although a rate of 5.07E-9 substitution/site/year has been proposed for the pathogen's mtDNA [20]. Our analyses also enabled us to estimate that the historical Ebro-1944 genome shares a common ancestor with the South-American cluster at ca 1455 (HPD 95% 1012-1750) CE (Figure 5c) and the most recent common ancestor (TMRCA) of a geographically diverse set of P. vivax strains that encompasses its global diversity (Supplementary Table 8) to 2011 years (95% HPD: 400-3,000). Our mutation rate is considerably higher than that obtained in P. falciparum cell lines (1.7E-9 substitution/site/generation) [21]; however, the differences in the life cycle of both malaria parasites makes difficult to know if both mutation rates can be compared.

A number of P. vivax genes have been identified to confer resistance to antimalarial drug treatments developed in the later decades of the 20th century. For instance, mutations in the pvdhfr gene are known to be involved in resistance to pyrimethamine whilst the pvdhps gene confers resistance to sulfadoxine. Other genes, differing from those described in P. falciparum, could be related to chloroquine resistance. The characterization of multiple loci (N=516) showing strong signals of recent natural selection in a geographically diverse set of modern strains [22] points to a diverse response of P. vivax to malaria drugs as well as local differentiation processes. Some of the detected regions encompass genes previously known to be involved in drug resistance, including three with strong experimental validation of resistance phenotypes: pvmdr1, dhfr and dhfps [23]. Our historical genome has

355 of these positions covered, of which Ebro-1944 exhibits the ancestral allele in 349 (Supplementary Table 4). Therefore, only six of the 355 variants with high FST values (including the Met205Ile variant at DHPS gene) were present in the historical European sample suggesting the relatively rapid accumulation of resistance conferring mutations in more recent strains.

Discussion

Our genome-wide analyses of a historic European P. vivax nuclear genome both confirm the previous observation that European P. vivax mtDNA clusters with the most common current Central and South American strains [16], and suggest a divergence time of the two clades in ca. the 15th century. This is consistent with a post-Columbus origin of the dominant New World strains, as even when accounting for uncertainties associated with the mutation rate, the estimates are much more recent than the date required should the parasite have been introduced into the Americas alongside the first humans to colonize the continent ca. 15,000 years ago. The fact that historical accounts indicate the existence of malaria in Europe since at least Classical Greece[4] suggests that the pathogen spread from Europe into the Americas and not the other way around.

The extremely high mtDNA diversity reported in American P. vivax strains has been interpreted as an indication that different migratory waves and admixture events took place in this continent, maybe through contacts from South Asia, Africa and even Melanesia [13].

Moreover, genomes of the American strains exhibit a shared ancestry with the few African strains currently available (from Mauritania and Madagascar) but also with small strain collections from India and Sri Lanka. Therefore, beyond the plausible transmission of at least some American strains from Europe in post-Columbus times during the European settlement of the continent, the general dispersal patterns of the parasite along different continents remains unsolved due to the complexity of its evolutionary history. Moreover, an additional inference from our analyses is that all P. vivax dispersals, including those of the most-divergent Melanesian strains, seem to have occurred in recent times, probably in the last 2,000 years. This recent date is relatively close to some previous works that place the TMRCA around 5,000 or 10,000 years ago [4,20] but in any case conflicts with much older estimates, including those around 45,000-81,000 years 24 or 53,000-265,000 years [25]. We caution that the lack of historical P. vivax samples between 1944 and 1991 can influence our mutation rate estimate. However, the post-Neolithic spread of P. vivax strains is not a unique feature among malaria pathogens. A recent genomic study has revealed a recent population bottleneck, around 4,000-6,000 years for P. falciparum current diversity [26].

Malaria is believed to have represented one of the strongest selective forces to have shaped the human genome [27]. Well known examples of selection against P. falciparum include mutations at the HBB gene that exhibit resistant isoforms of proteins such as HbS and HbE in African and Asian populations, respectively, and mutations at the G6PD gene which are broadly spread in African populations and are also present in the Mediterranean [28]. One of the best-known examples of directional selection is the Duffy blood negative genotype that confers natural resistance to P. vivax. The derived allele is essentially fixed in Sub-Saharan Africa, yet virtually absent elsewhere. This mutation is credited for the near absence of the pathogen in sub-Saharan African populations, where the species appears to have originated [12]. There are also several other putative malaria-resistance variants present in modern European populations, although their specific roles and importance in malaria resistance through history remains unclear [29]. Our evidence, pointing to a recent origin of modern P. vivax strains, suggests that selection did not have much time to act against this pathogen.

There is increasing concern about the rapid spread of P. vivax strains resistant to antimalarial drugs. Chloroquine has been established as the main therapy against P. vivax infections since 1946 [30]. It has been a well-tolerated and effective treatment until chloroquine resistances appeared and spread through the entirety of the endemic range of P. vivax [31]. Despite extensive drug resistance within present-day P. vivax populations, caused by a variety of resistance loci in multiple genes [15], chloroquine-primaquine combined therapy remains the most commonly prescribed treatment [32]. The increase in drug resistant strains is thus a subject of public health concern due to its major human and economic costs.

Our historical strain predates all modern anti-malaria drugs, with the exception of quinine that was introduced in Europe as early as 1683

[33]. Ebro-1944 shows the ancestral allele in an overwhelming amount (99.3%) of SNPs known to have undergone selection in modern strains, including all those associated with drug resistance. Interestingly, one of the six Ebro-1944 derived alleles corresponds to the Met205Ile amino-acid change [34] of the DHPS gene, that appears to be fixed or at very high frequencies in modern Asian and American isolates and has been implicated in sulfadoxine resistance [34–36]. The fact that these alleles were present in a sample predating the use of this drug points to the presence of standing variation in this gene in historical P. vivax strains and could help explain the rapid appearance and spread of this gene in resistant strains. In other drug-resistant genes such as DHFR-TS [37–41] and MDR1 [42–45] the historical sample carries the ancestral alleles. A precise assessment of the current as well as ancient genomic diversity of P. vivax could help elucidate the adaptive mechanisms of this parasite and increase our knowledge on how resistant mutants spread, particularly in the early phase. Our finding on the DHPS gene has important applications for the development of new treatments and intervention schemes. For instance, if resistant mutations are already present in the parasites prior to the drug even been introduced to a population, an "aggressive treatment" (defined as a high dose treatment aimed to kill parasites as fast as possible) might actually select for resistance strongest [46,47].

Our study stresses the value of old microscopy slides and, more generally, of antique medical collections, as unique tools for retrieving genomic information on past pathogens, including eradicated strains that could not be studied from contemporary specimens. The slides analysed here were stained but not fixed and it remains to be seen what additional DNA damage is exerted by different fixation methods. Our slides dated from the years 1942-1944; however, it is likely that older slides are available in both public and private collections given the popularity of microscopy in Victorian times. Therefore, a future goal will be to ascertain if massive genomic data retrieval is possible from even older slides.

Our results also offer new opportunities to further clarify the path of P. vivax dispersals through retrieving ancient *Plasmodium* sequences from either additional antique medical slides or directly from bones. The recent retrieval of P. falciparum sequences from ancient Roman human skeletal remains [48] demonstrates this approach is technically feasible, and as P. vivax infection is more prevalent than P. falciparum, it is plausible that further ancient strains could be reported in the near future. An additional possibility would be to directly retrieve *Plasmodium* sequences from Anopheles remains preserved for instance in ancient lake sediments. This, together with additional sequencing of current P. vivax strains from under-sampled areas, could help elucidate the evolutionary history of this important parasite.

Methods

Samples

The slides analyzed belong to the personal collection of the descendants of Dr. Ildefonso Canicio, who worked in the antimalarial center established by the Catalan Government at Sant Jaume d'Enveja (Ebro Delta, Spain) in 1925. Five samples have been analyzed, three of them included in a previous study [16]. One of the two new samples had only an unstained, anonymous drop of blood and the second one was a drop of blood from a double slide, stained with Giemsa (Figure 1a).

DNA extraction

DNA extraction was performed by incubating the slide with 20 L of extraction buffer (10 mM Tris-HCl (pH 8), 10 mM NaCl, 5 mM CaCL, 2.5 mM EDTA, 1% SDS, 1% Proteinase K, 0.1% DTT (w/v)) in an oven at 37C for 20 minutes for a total of 3 rounds. The resulting dissolved bloodstain and buffer was collected in 1.5 mL Lobind Eppendorf and then incubated for 1 hour at 56C and subsequently added to 10x volume of modified binding buffer [49] and passed through a Monarch silica spin column (NEB) by centrifugation (Supplementary Methods and Figure 1). The column was washed with 80% ethanol and DNA was subsequently released with EBT buffer to a final volume of 40L (see Supplementary Material for details). All the analyses have been done in dedicated ancient DNA laboratories where no previous genetic work on *Plasmodium* has been carried out, both in Barcelona (extraction of slides in 2016) and Copenhagen (extraction of slides in 2017).

Library preparation and DNA sequencing

Sequencing libraries for the Illumina platform were prepared using a single-tube protocol for double-stranded DNA [50], with minor modifications and improvements as detailed in Mak et al. (2017) [51] (see Supplementary Materials for details). Sequencing was performed at the Natural History Museum of Denmark on one lane of an Illumina Hiseq 2500 instrument in paired end mode running 125 cycles (Supplementary Methods and Table 7).

Sequence mapping

The sequenced reads were analyzed with FastQC to determine the quality prior and after adapter removal. The 3' read adapters and the consecutive bases with low quality scores were removed using AdapterRemoval and reads shorter than 30 bp were eliminated. To increase the final coverage, all *Plasmodium* reads were pooled and an in-house script was used to discriminate between P. vivax and P. falciparum sequences. As in Gelabert et al. (2016)[16]; the former were initially mapped to Salvador1 (Sal 1) P. vivax reference genome [52] and the later to P. falciparum 3D7 reference genome [53]. Mapping was performed with Burrows-Wheeler Aligner (BWA) [54]. Duplicated reads were deleted using Picard-tools MarkDuplicates. Reads with a quality score lower than 30 were further excluded. MapDamage was used to check for signatures of postmortem damage at the ends of the reads. C to T and G to A substitutions at the 5' ends and 3' ends, respectively, were found to be present at a frequency of about 2.5% (Supplementary Figures 2 and 3), coherent with the date of the sample and in agreement with the damage detected at the mtDNA reads [16]. To exclude this potential DNA damage, the first three nucleotides of each read were trimmed.

The newly generated P. vivax reads (labeled Ebro-1944) were merged with the 2016 reads [16]. Genotypes were subsequently called with GATK v 3.7 UnifiedGenotyper 55 run with default parameters (Supplementary Figure 4 and 5).

Population genetics dataset

The population genetics dataset comprised 228 samples from the P. vivax Genome Variation Project6 from Brazil (3), Cambodia (40), China (1), India (1), Indonesia (55), Laos (2), Madagascar (1), Malaysia (6), Myanmar (1), Papua New Guinea (11), Sri Lanka (1), Thailand (92) and Vietnam (14). We included 20 additional samples from Brazil (10)56, Mexico (5)56, Mauritania (Africa) (2)[56] and Madagascar (3)[57], as well as our ancient European sample Ebro-1944 (1). The dataset was generated and SNPs called relative to the Salvador 1 reference genome published in 2008 52. Across these datasets the 303,616 high-quality SNPs [6] were extracted and filtered for a quality score of at least 20 (Supplementary Figure 6).

In order to identify samples with single predominant genotypes as opposed to those which are from mixed infections we calculated the metric FWS as originally described in Manske et al. (2012)[58] on all samples with >1 sample per population. Only those samples with FWS >0.95 were kept for further analysis. We also excluded all sites that were heterozygous in Ebro-1944 (198 sites) across this dataset. This resulted in a dataset of 152 samples (including Ebro-1944) as described in Supplementary Table 3 across 302,762 sites after removing multi-allelic positions with a total genotyping rate of 0.86.

The missingness per sample ranged from 0.01-24% (1st-3rd quantiles) with the ancient sample Ebro-1944 missing 94,935 of the 302,762 sites (31%) in this dataset and containing 2,585 non-reference SNPs.

Allele-frequency based measures of population structure

This dataset described in Supplementary Table 3 was filtered for SNPs in high LD using a 60 SNP sliding window advancing each time by 10 steps and removing any SNP with a correlation coefficient ≥0.1 with any other SNP within the window 59. This left a pruned dataset of 180,348 SNPs for analyses relying on independent sites (unlinked).

We performed PCA implemented using the prcomp function in R to the LD pruned global dataset. We ran unsupervised ADMIXTURE [60] for values of K between 1-15. K = 3 provided the lowest cross-validation error. We also performed out-group f4 analysis implemented in qpDstat of AdmixTools 61 to measure correlated drift between populations using an unpruned dataset. In particular, using Ebro-1944 as a target, we explored which populations share more drift with Ebro-1944 relative to every combination of modern strain/strains in our reference dataset and using P. cynomolgi as an out-group.

Inferring patterns of allele and haplotype sharing

In addition we also used an unrelated method to explore patterns of allele (unlinked) and haplotype (linked) sharing implemented in

CHROMOPAINTER [62]. Unlike f-statistics, this approach does not rely on proposing a tree-like topology and can consider the relationship of all samples to all others collectively. Additionally the use of haplotype information has been shown to greatly aid the ability to resolve fine-scale population structure [63]. We first implemented the CHROMOPAINTER unlinked approach (-u switch), which considers sites as independent and thus does not reply on our ability to confidently call haplotypes. As this approach requires low levels of missingness across comparisons we pruned the previously described global dataset for only the positions present in Ebro-1944 and retained only those samples with ≤10% missing data (207,827 sites, 106 samples). A schematic of the workflow is provided in Supplementary Figure 7. Briefly, this method compares the DNA patterns in a "recipient" chromosome to that in a collection of "donor" chromosomes. Under the unlinked model, CHROMOPAINTER calculates, separately for each position, the probability that a "recipient" chromosome is most closely related to a particular "donor" in the dataset. Here, we use all samples in our reduced global dataset as donors and the equivalent samples as recipients in an "all-versus-all" painting approach so that for each recipient sample r, we define yrd to be the total amount (measured in matching allele counts) of DNA for which sample r is inferred to be most closely related to a donor chromosome from group d (the ".chunkcounts.out" output from the CHROMOPAINTER model).

To the resultant painting profiles we applied unsupervised clustering in fineSTRUCTURE [62] to group samples based on each of their inferred yrd in relation to every other sample. fineSTRUCTURE was run using for 2,000,000 sample MCMC iterations (-y) with a thin interval of 10,000 (-z). The tree building step was run after discarding the burn-in using 100,000 hill climbing iterations (-x) and 10,000 tree comparisons (-t). The fineSTRUCTURE inferred counts and hierarchical clustering is given in Supplementary Figure 9.

In order to include more samples from diverse populations we extracted another reduced dataset from our original global collection but now allowing for samples to have ≤40% of missingness. This allowed the incorporation of additional samples from Brazil, India, North Cambodia and Laos together with additional samples from other global populations (207,827 sites, 120 samples). In order to account for this higher level of missingness, we followed the protocol of Samad et al. (2015)64 to impute missing sites using BEAGLE v3.3.2 65.

We then performed chromosome painting as before to infer the genome of each sample (recipient) as a mosaic of every other sample (donors). However rather than calculating the probability at each position independently we now use a linked model assuming a uniform recombination map of constant rate 14kb per cM. Under the linked model yrd can be thought of as the length or proportion (rather than counts) of DNA shared between each recipient sample r and each donor d (the “chunklengths.out” output from linked CHROMOPAINTER model). To do this we first inferred the switch rate (Ne) and mis-copying parameter (theta) as advocated by Lawson et al. (2012)[62] by running the CHROMOPAINTER expectation-maximisation algorithm on every sample for 10 iterations. The inferred values were weight-averaged across all chromosomes to give mean estimates of Ne=1680:67 and =0:0055 across the 120 samples.

CHROMOPAINTER was then run with these values fixed using the -N and -M switches. As implemented for the unlinked analysis, we additionally performed clustering on the inferred painting profiles (yr'ds) across all samples in fineSTRUCTURE. fineSTRUCTURE inferred many more clusters (44) highlighting how using linkage information can increase resolution of finer-scale population structure (Figure 5).

Drug resistance variants analysis

We have screened the European P. vivax for the list of alleles that are confirmed or suspected to confer resistance to different antimalarial drugs [6]. We have considered all loci for which we have overlapping reads but note that some alleles may not be present due to the limited coverage of our historic sample (Supplementary Table 4)

Estimating a substitution rate for P. vivax

The P. vivax reads were subsequently mapped against the PvP01 reference assembly and the mapped genome was filtered as described for the population genetics dataset. The main motivation for using the PvP01 reference genome for this analysis was as this assembly has better definition of sub-telomeric genes and repetitive pir genes that are in recombinant regions, thus allowing more refined filtering out of these regions [17]. After calling variants with GATK version 3.7 UnifiedGenotyper the dataset was further filtered by selecting those SNPs with a coverage >2, and by removing heterozygous positions and sub-telomeric genes. The resulting core genome was aligned to 20 P. vivax genomes, prioritizing those with decades-old sequencing dates. The variant positions identified in the alignments were classified as exonic and non-exonic according to the annotations of PvP01. The exonic positions were then classified as synonymous and non-synonymous with an in-house script.

Selected isolates

A subset of 15 samples (Supplementary Table 5) with recorded sampling times were selected based on two criteria: 1) the isolation time needed to be safely described and 2) strains should not have been in-vivo or in-vitro maintained for too long (long-term experiments) before having been sequenced, Unfortunately, condition 2 resulted in exclusion of three P. vivax genomes isolated in 1944, 1972, and 1980 respectively (SRR575087, SRR575089 & SRR828528) but maintained in monkey reservoirs for several years before sequencing. Based on the results from a previous study 16 comparing our European historical mtDNA genome with a panel of strains representative of the global diversity, we selected strains originating from Central America, South America, Melanesia and Papua New Guinea so as to maximize the tempo-geographic spread. Including our 1944 European historical genome, we compiled a total dataset of 16 genomes spanning 73 years (1940-2013) of evolution (see Supplementary Table 6 for more details).

Investigating the temporal signal

To investigate the extent of temporal signal existing in our dataset we concatenated SNPs amongst all chromosomes selecting only the core, non-recombining region of the P. vivax genome and built a maximum-likelihood phylogenetic tree without constraining tip-heights to their sampling times using PhyML [66]. We then rooted the tree using a P. falciparum genome and computed genetic distances between every single tip and the MRCA of all P. vivax isolates using the distRoot function included in the adephylo R package [67]. A linear regression between isolates sampling year and the previously calculated genetic distances was then performed using the lm function implemented in R68.

Estimating substitution rates

Providing sufficient temporal signal, substitution rates were estimated by running a tip-calibrated inference using Markov chain Monte Carlo (MCMC) sampling in BEAST 1.8.3 [19]. Based on an annotation of the SNPs, we defined 3 partitions composed of non-exonic, exonic-synonymous and exonic-non-synonymous SNPs, respectively. The best-fit nucleotide substitution model for each partition was estimated using the software BModelTest 69(synonymous: GTR, non-synonymous: HKY and non-exonic SNPs: GTR) and was used henceforth. Rate variation among sites was modelled with a discrete gamma distribution with four rate categories. We assumed an uncorrelated lognormal relaxed clock to account for rate variation among lineages. To minimize prior assumptions about demographic history, we adopted an extended

Bayesian skyline plot (EBSP) approach in order to integrate data over different coalescent histories. In order to calibrate the tree using tip-dates only, we applied flat priors (i.e., uniform distributions) for the substitution rate (1.10-12 - 1.10-2 substitutions/site/year) as well as for the age of any internal node than in the tree. We ran five independent chains in which samples were drawn every 10,000 MCMC steps from a total of 100,000,000 steps, after a discarded burn-in of 10,000,000 steps. Convergence to the stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size >200). Parameter estimation was based on the samples combined from the different chains.

Estimating the Time of the Most Common Recent Ancestor (TMRCA)

We have built a specific dataset (Supplementary Table 6) by selecting the most diverse samples from the population genetics dataset (Supplementary Methods and Table 8). A robust topology was first estimated from this alignment using PhyML (Supplementary Figure 10). In a second step, we used BEAST [19] to calibrate this tree using the previously estimated synonymous rate of evolution. We used as prior for the rate of substitution a normal distribution with mean 6.0E-7 substitution/site/year and a standard deviation of 2.0E-7 (all other parameters being unchanged as compared with the previous tip-dating inference).

References

1. World Health Organisation. World Malaria Report 2017. (2017).doi:http://www.who.int/malaria/publications/world-malaria-report-2017/report/en/.

2. Gonzalez-Ceron, L. et al. Molecular and epidemiological characterization of *Plasmodium* vivax recurrent infections in southern Mexico. Parasit. Vectors 6, 109 (2013).

3. Adekunle, A. I. et al. Modeling the Dynamics of *Plasmodium* vivax Infection and Hypnozoite Reactivation In Vivo. PLoS Negl. Trop. Dis. 9, 1–18 (2015).

4. Carter, R. Speculations on the origins of *Plasmodium* vivax malaria. Trends Parasitol. 19, 214–219 (2003).

5. 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. 23, 2975–2987 (2014).

6. Pearson, R. D. et al. Genomic analysis of local variation and recent evolution in *Plasmodium* vivax. Nat. Genet. 48, 959–964 (2016).

7. Huldén, L., Huldén, L. & Heliövaara, K. Endemic malaria: An ‘indoor’ disease in northern Europe. Historical data analysed. Malar. J. 4, 1–13 (2005).

8. Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M. & Snow, R. W. The global distribution and population at risk of malaria: past, present, and future. Lancet Infect Dis 4, 327–336 (2004).

9. Pletsch, D. Report on a mission carried out in Spain in September-November 1963 for verification of the erradication of malaria. Rev. Sanid. Hig. Publica (Madr). 39, 309–367 (1965).

10. Petersen, E., Severini, C. & Picot, S. *Plasmodium* vivax

malaria: A re-emerging threat for temperate climate zones? Travel Med. Infect. Dis. 11, 51–59 (2013).

11. Zhao, X., Smith, D. L. & Tatem, A. J. Exploring the spatiotemporal drivers of malaria elimination in Europe. Malar. J. 15, 1–13 (2016).

12. Liu, W. et al. African origin of the malaria parasite *Plasmodium* vivax. Nat. Commun. 5, 3346 (2014).

13. Rodrigues, P. T. et al. Human migration and the spread of malaria parasites to the New World. Sci. Rep. 8, 1–13 (2018).

14. Culleton, R. et al. The Origins of African *Plasmodium* vivax; Insights from Mitochondrial Genome Sequencing. PLoS One 6, e29137 (2011).

15. 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).

16. 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).

17. Auburn, S. et al. A new *Plasmodium* vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 1, 4 (2016).

18. Rieux, A. & Balloux, F. Inferences from tip-calibrated phylogenies: A review and a practical guide. Mol. Ecol. 25, 1911–1924 (2016).

19. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 1–8 (2007).

20. Leclerc, M. C. et al. Meager genetic variability of the human malaria agent *Plasmodium* vivax. Proc. Natl. Acad. Sci. U. S. A. 101, 14455–14460 (2004).

21. Bopp, S. E. R. et al. Mitotic Evolution of *Plasmodium* falciparum Shows a Stable Core Genome but Recombination in Antigen Families. PLoS Genet. 9, 1–15 (2013).

22. Pearson, R. D. et al. Genomic analysis of local variation and recent evolution in *Plasmodium* vivax. Nat. Genet. 48, 959–964 (2016).

23. Haldar, K., Bhattacharjee, S. & Safeukui, I. Drug resistance in *Plasmodium*. Nat. Rev. Microbiol. 16, 156–170 (2018).

24. Escalante, A. A. et al. A monkey's tale: The origin of *Plasmodium* vivax as a human malaria parasite. Proc. Natl. Acad. Sci. U. S. A. 102, 1980–1985 (2005).

25. Mu, J. et al. Host Switch Leads to Emergence of *Plasmodium* vivax Malaria in Humans. Mol. Biol. Evol. 22, 1686–1693 (2005).

26. 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).

27. Hedrick, P. W. Resistance to malaria in humans: the impact of strong, recent selection. Malar. J. 11, 349 (2012).

28. Kwiatkowski, D. P. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am J Hum Genet 77, 171–192 (2005).

29. Gelabert, P., Olalde, I., De-Dios, T., Civit, S. & Lalueza-Fox. Malaria was a weak selective force in ancient Europeans. Sci. Rep. 7, 1377 (2017).

30. Most, H. & London, I. M. Chloroquine for treatment of acute attacks of vivax malaria. J. Am. Med. Assoc. 131, 963–967 (1946).

31. Rieckmann, K. H., Davis, D. R. & Hutton, D. C. *Plasmodium* vivax resistance to chloroquine? Lancet 2, 1183–1184 (1989).

32. Phillips, E. J., Keystone, J. S. & Kain, K. C. Failure of

combined chloroquine and high-dose primaquine therapy for *Plasmodium* vivax malaria acquired in Guyana, South America. Clin. Infect. Dis. 23, 1171–1173 (1996).

33. 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).

34. 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).

35. 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 Sulfadoxine Resistance in *Plasmodium* vivax Is Associated with a Specific Amino Acid in Dihydropteroate. Antimicrob. Agents Chemother. 48, 2214–2222 (2004).

36. Menegon, M., Majori, G. & Severini, C. Genetic variations of the *Plasmodium* vivax dihydropteroate synthase gene. Acta Trop. 98, 196–199 (2006).

37. 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).

38. de Pecoulas, P. E., K. Basco, L., Tahar, R., Taoufik, O. & Mazabraud, A. Analysis of the *Plasmodium* vivax dihydrofolate reductase–thymidylate synthase gene sequence. Gene 211, 177–185 (1998).

39. 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).

40. Leartsakulpanich, U. et al. Molecular characterization of dihydrofolate reductase in relation to antifolate resistance in *Plasmodium* vivax. Mol. Biochem. Parasitol. 119, 63–73 (2002).

41. 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).

42. 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).

43. Sá, J. M. et al. *Plasmodium* vivax: Allele variants of the mdr1 gene do not associate with chloroquine resistance among isolates from Brazil, Papua, and monkey-adapted strains. Exp. Parasitol. 109, 256–259 (2005).

44. 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).

45. 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).

46. Read, A. F., Day, T. & Huijben, S. The evolution of drug resistance and the curious orthodoxy of aggressive chemotherapy. Proc. Natl. Acad. Sci. U. S. A. 108, 10871–10877 (2011).

47. Huijben, S. et al. Aggressive Chemotherapy and the Selection of Drug Resistant Pathogens. PLOS Pathog. 9, e1003578 (2013).

48. Marciniak, S. et al. *Plasmodium* falciparum malaria in 1st–2nd century CE southern Italy. Curr. Biol. 26, R1220–R1222 (2016).

49. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. Nature 522, 167 (2015).

50. Christian, C. et al. Single-tube library preparation for degraded DNA. Methods Ecol. Evol. 9, 410–419 (2017).

51. Mak, S. S. T. et al. Comparative performance of the BGISEQ-

500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience 6, 1–13 (2017).

52. Carlton, J. M. et al. Comparative genomics of the neglected human malaria parasite *Plasmodium* vivax. Nature 455, 757–763 (2008).

53. Gardner, M. J. et al. Genome sequence of the human malaria parasite *Plasmodium* falciparum. Nature 419, 498–511 (2002).

54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

55. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1298–1303 (2010).

56. Carlton, J. M. R. et al. Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species. Mol. Biochem. Parasitol. 118, 201–210 (2001).

57. Chan, E. R. et al. Whole Genome Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity in *Plasmodium* vivax. PLoS Negl. Trop. Dis. 6, e1811 (2012).

58. Manske, M. et al. Analysis of *Plasmodium* falciparum diversity in natural infections by deep sequencing. Nature 487, 375–379 (2012).

59. 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).

60. Alexander, D. H. & Novembre, J. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. 19, 1655–1664 (2009).

61. Patterson, N. et al. Ancient Admixture in Human History. 192, 1065–1093 (2012).

62. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. PLoS Genet. 8, e1002453 (2012).

63. Leslie, S. et al. The fine-scale genetic structure of the British population. Nature 519, 309–314 (2015).

64. Samad, H. et al. Imputation-Based Population Genetics Analysis of *Plasmodium* falciparum Malaria Parasites. PLoS Genet. 11, e1005131 (2015).

65. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459–471 (2013).

66. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321 (2010).

67. Jombart, T., Balloux, F. & Dray, S. adephylo: New tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26, 1907–1909 (2010).

68. Ross, I., Robert, G., Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. Journal of computational and graphical statistics 5, 299–314. (1996).

69. Posada, D. jModelTest: Phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256 (2008).

Acknowledgements

We are grateful to Thomas D. Otto (University of Glasgow) and Thomas Lavsten (University of Copenhagen) for helpful comments and suggestions and to the descendants of Dr Canicio, Miquel and Ildefons Oliveras for sharing with us their slides. We additionally would like to thank Dr. George Busby and Dr. Jacob Almagro for useful discussions on chromosome painting of *Plasmodium* sp. This research was supported by a grant from FEDER and Ministry of Economy and Competitiveness (BFU2015-64699-P) of Spain to C.L.-F and ERC Consolidator Grant (681396-Extinction Genomics) to MTPG. P. vivax genomes are deposited at the European Nucleotide Archive under accession numbers xxx.

Author contributions

P.G., M.T.P.G., F.B. and C.L.-F. conceived and designed the study; R.E. and C.A. discovered the slides; C.C. extracted DNA; P.G., L.v.D., A.R., T. d.-D., S.G., R.F., I.O., analysed data and performed computational analyses; S.H., and I.M. provided comments and suggested analyses; P.G., L.v.D., A.R., F.B., M.T.P.G. and C.L.-F. wrote the paper with inputs from all coauthors.

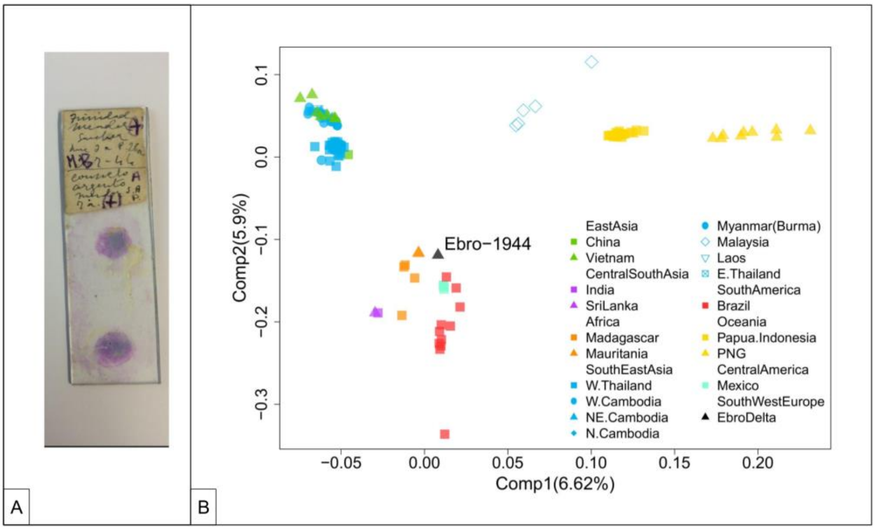


Figure 1: a) Microscope slide with two blood drops from two different patients from the Ebro delta used in the retrieval of *Plasmodium* vivax. Most of the data was retrieved from the above stain b): Principal components analysis (PCA) of the ancient Ebro-1944 sample in a geographically diverse set of modern P. vivax strains.

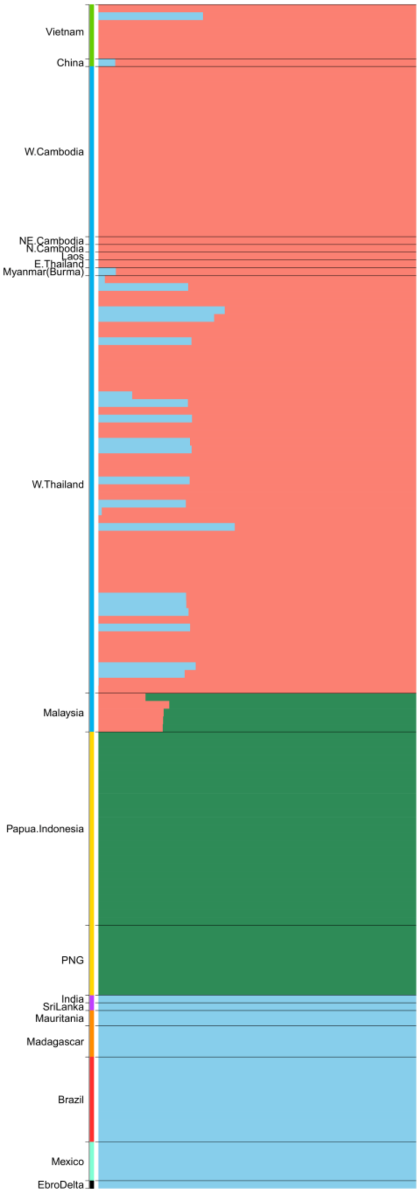


Figure 2: Unsupervised ADMIXTURE clustering analysis at K=3. Samples are arranged by geographic region and colored as in Figure 1.

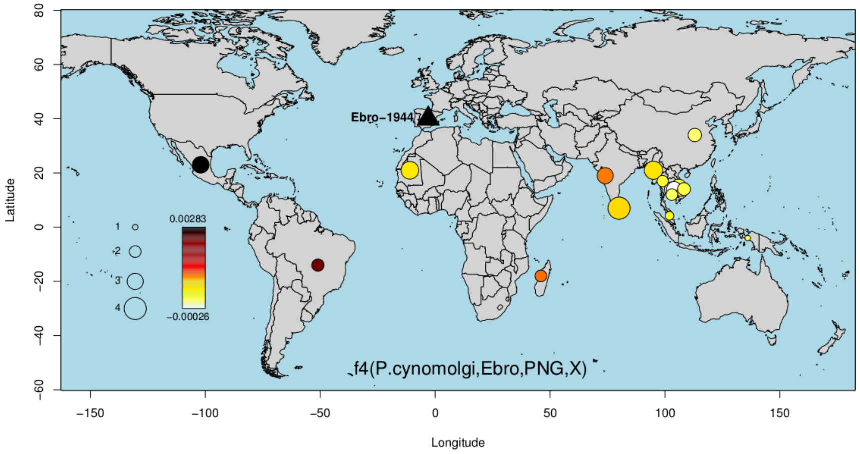


Figure 3: f4-values inferred under the test relationship (P.cynomolgi, Ebro, Papua-New Guinea (PNG),X), where X iterates through global sampling locations. The color scale provides the value of the f4 statistic with the significance (absolute z score), assessed through block jack-knife resampling, provided by the circle size.

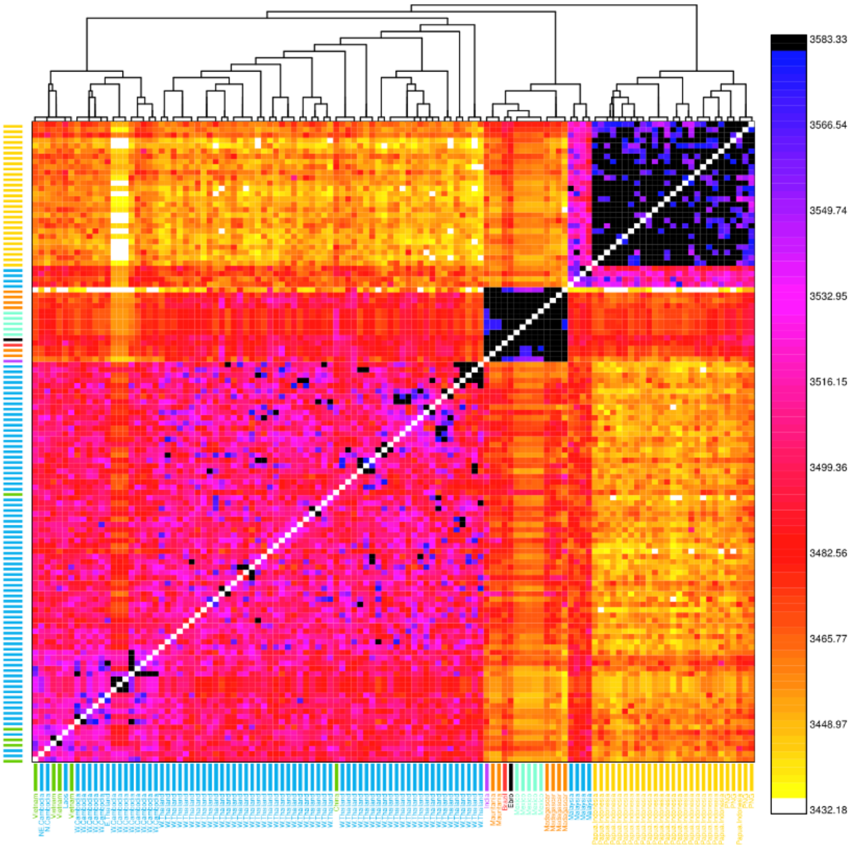
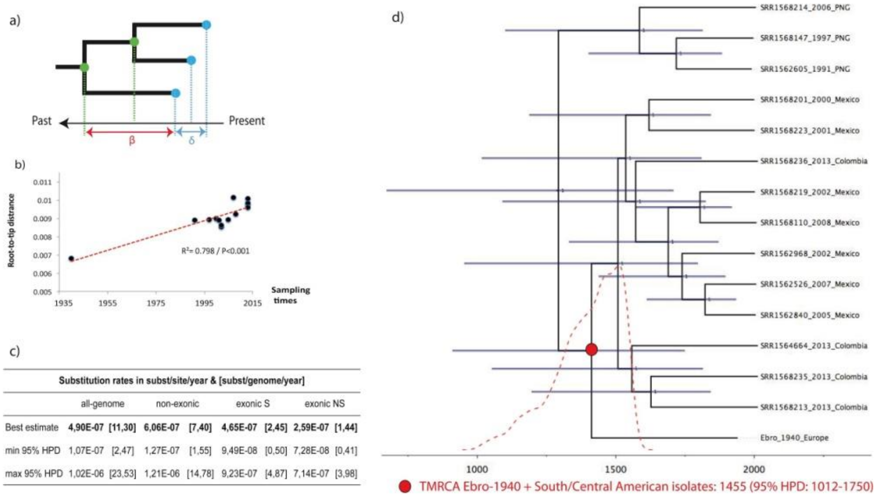


Figure 4: CHROMOPAINTER co-ancestry matrix with population structure assignment based on fineSTRUCTURE analysis. CHROMOPAINTER. 's inferred counts of matching DNA genome wide that each of the 44 inferred clusters (columns) is painted by each of the 44 clusters (rows). The tree at top shows fineSTRUCTURE's inferred hierarchical merging of these 44 clusters and the colors on the axes give the continental region and population to which samples in each cluster are assigned. Ebro-1944 is depicted in black and clusters with the sample from Brazil and Mexico.



Substitution rates in subst/site/year & [subst/genome/year]

	all-genome		non-exonic		exonic S		exonic NS	
Best estimate	**4,90E-07**	**[11,30]**	**6,06E-07**	**[7,40]**	**4,65E-07**	**[2,45]**	**2,59E-07**	**[1,44]**
min 95% HPD	1,07E-07	[2,47]	1,27E-07	[1,55]	9,49E-08	[0,50]	7,28E-08	[0,41]
max 95% HPD	1,02E-06	[23,53]	1,21E-06	[14,78]	9,23E-07	[4,87]	7,14E-07	[3,98]

Figure 5: a) Calibration of phylogenetic tree using dated tips. From a collection of sequences sampled over various times (blue dots), if the time d between the youngest and oldest sequence represents a significant proportion of the time b since all the sequences last shared a common ancestor, then it is possible to jointly estimate the phylogenetic tree topology (black branches), the rate of evolution and the age of any internal node in the tree (gree dots). b) Root-to-tip distances correlate with isolate date (P<0.001), indicating that the data have sufficient temporal signal and predictive power for tip-dates to be used as phylogenetic calibration points. c) Substitution rate estimates obtained with tip-calibration from different nucleotide subsets. Both best (mean) and 95%HPD are given either in yearly number of substitutions per site or per genome. d) Tip-dated phylogenetic tree obtained with BEAST. Blue bars refer to 95%HPD for internal node ages. The posterior probability distribution for time to the most recent common of the historical European strain and American isolates is indicated in red.

5. DISCUSSION

The coming section will summarize the methodological approaches and results presented in the included scientific publications. Furthermore, the section aims to place the highlighted novelties in the global context of malaria and ancient pathogens knowledge. Using the presented discoveries to suggest future directions of these disciplines.

The discussion will be divided in four main sections. The first one will be focused on the technical procedures implemented in the *Plasmodium* analysis, with special regard to *Plasmodium* DNA extraction and sequencing pipelines. The second will discuss the origin of *Plasmodium* species and its dispersals, with special attention to the position of both European *P. vivax* and *P. falciparum* isolates. The third one will evaluate the outcomes regarding the European *P.* vivax genome sequencing. Finally the fourth section will review the selective signals that malaria has left in European populations, as well as the methodological approaches that have been implemented in the genotyping process of ancient individuals.

For the first time ever, we have successfully sequenced the genomes of two eradicated European *Plasmodium* strains. These sequences have been used to outline relevant questions with regard to *Plasmodium* dispersal and adaptation. This achievement is a qualitative step in the definition of the global genetic mechanisms involved in malaria dispersal and virulence.

5.1 *Plasmodium* DNA retrieval

There are few examples of pathogenic ancient DNA extracted from medical collections. Probably the most notorious achievement has been the sequencing of HIV-1 virus from African plasma, sampled in 1959 (Zhu et al. 1998). The same virus was also found in a paraffin block, dated from 1960 (Worobey et al. 2008). This two publications evidenced that HIV-1 virus was infecting humans at least some years before AIDS was initially described. The comparison between the present-day strains of *V. cholerae* and one extracted from a XIXth century intestine (Devault et al. 2014) is also an outstanding example of ancient pathogens recovery. In this sense, the rareness and the technical challenges, turn the sequencing of the eradicated European *Plasmodium* into a unique achievement, in the field of ancient pathogens recovery. To empathise the accomplishment, not only there are no others published eradicated *Plasmodium* strains, also there is a complete lack of other evidences of ancient DNA recovered from giemsa-stained slides, although protozoa DNA has been recovered from modern giemsa-slides (Motazedian et al. 2002).

Capture and Depletion approaches

The recovery of DNA from medical samples is rare in the field of paleogenomics. The absence of described protocols has enforced us to experiment with the existing approaches, as well as to introduce new methods to enrich the targeted sequences. *Plasmodium* are protist, which are neither naturally present in the organism of health humans, nor in the European environment. This makes the cross contamination from environmental organisms to become an issue of little concern. An advantageous and unusual situation when working with aDNA samples. Additionally not a single *Plasmodium* sample had never been in the laboratory where the extractions were carried out, which dramatically neglected any possible source of contamination from other similar organisms.

In order to recover the *Plasmodium* from the slides we have taken advantage of the described pathogen capture pipelines (Schuenemann et al. 2011, Bos et al. 2011), and also we have used this methods with the inverse intention. Pathogens are usually targeted in bone or dental tissues. As bacterial DNA is ubiquitous, and sequences between different species can be extremely similar, false positives are a matter of major concern when pathogens are recovered (Campana et al. 2014). Selective Capture enrichment is commonly the only approximation that can be implemented to selectively separate targeted sequence from the rest of naturally present microbial data, and also to enrich the sequencing libraries. (Schuenemann et al. 2011, Bos et al. 2011, Wagner et al. 2014, Bos et al. 2014, Rasmussen et al. 2015, Spyrou et al. 2016, Andrades-Valtueña et al. 2017, Spyrou et al. 2018). Remarkably *P. vivax* research faces the same difficulty. Due to the biological cycle of *P. vivax* (1.3.3 (B)), the pathogenesis is reduced and the recovery very arduous. The used techniques to concentrate *P. vivax* sequences from blood samples consist in: i) Hybrid selection capture (Bright et al. 2012), ii) growth in splenectomised monkeys (Carlton et al. 2008), iii) ex-vivo culture (Auburn et al. 2013), iv) single cell sequencing

(Nair et al. 2014), v) Leukocyte depletion (Venkatesan et al. 2012). The usage of such techniques make the sequencing of high quality *P. vivax* samples a very laborious and complex task.

We have implemented and therefore compared two different approaches to recover *Plasmodium* sequences from the antique slides. A first one based on the selective capture of the *Plasmodium* reads by in-solution DNA baits. And a second one consisting in a selective depletion of the human DNA, discarding the product of an in-solution human DNA capture. We have observed that *Plasmodium* capture method produces lot of clonality that was not observed in the human-depleted libraries. The amount of *Plasmodium* reads has been observed to be extremely low in proportion to the human DNA, which has promoted the sequencing of huge amount of duplicated sequences, in the capture libraries. Nevertheless, the final performance of both approaches, after removing the duplicated sequences, is very similar. The number of unique sequences and in consequence the depth coverages of the retrieved *P. vivax* and *P. falciparum* (Gelabert et al. 2016) are almost identical. The evidence is that we have recovered the mtDNA genome of *P. vivax* in the CM sample with a depth coverage of 3x in the Capture library, while we have reached a 3.06x in the human depleted library.

We have also observed that the efficiency of the human depletion approach has not been perfect. Although the clonality was reduced and the final efficiency was comparable with the capture approach, the amount of endogenous reads was very little. This must be mainly attributed to the sequence of human reads not efficiently eliminated in the depletion process.

According to the exposed results we estimate that *P. falciparum* capture approach did not represent a quantitative improvement compared to depletion method. As the capture method requires the specific design of DNA-baits, and was comparable to human depletion, in terms of efficiency, we therefore did not use it in the libraries prepared in 2017. In the future, the sequencing of aDNA from medical slides should experiment with the usage of more powerful tools to concentrate the DNA such as flow cytometry or laser capture. Both in the direction of selecting the infected erythrocytes only.

Coinfection, false positive or sensitivity?

Surprisingly we have found mixed *P. vivax* and *P. falciparum* infections in all the blood slides that we have sampled. No previous work has ever presented a pipeline to process SGS libraries with three genomes present in the same sample, two of it being the targeted ones. Therefore, we designed a strategy that classified each sequencing read based on the substitutions that presented with both reference sequences. The results that we have obtained, verify that we have been able to discriminate both pathogens with a high standard of accuracy, excepting those reads with identical edit distances with both assemblies. This limitation has possibly affected the mean depth of coverage calculation. The difference in the GC content between both genomes (Gardner et al. 2002, Carlton et al.

2008) have also been used to demonstrate the authenticity of the obtained data.

The rates of coinfected patients in present day Malaria endemic countries are variable. Because of *P. vivax* is mainly absent in Africa, mixed infections are primarily reported in Asian and American individuals. Although no global rates have been appraised, mixed infections are observed in the 13% India malaria patients (Siwal et al. 2018) and in the 10% of the malaria patients of the French Guiana (Ginouves et al. 2015). The dramatic ubiquity observed in the Ebro Delta samples (100% of coinfection) could be strongly related with a bias in the health care attendance. Regarding the recurrence patterns of both infections (Douglas et al. 2011), Ebro Delta population was probably infected recursively by both pathogens, and only when people showed acute symptoms, mainly caused by *P. falciparum* (Carter et al. 2002) were treated in the medical centre. This assumption is supported by the reported extreme deficiency in the *P. vivax* diagnose (Pattanasi et al. 2003, Howes et al. 2016). *P. vivax* blood-stage low parasite densities lead to high rates of false negative diagnoses by microscope (Mueller et al. 2009). Also observed in mixed infections, were blood-stage *P. vivax* is underdiagnosed (Carlton et al. 2011). In this sense the coinfection rates shown by Ebro Delta samples could be defining the landscape Mediterranean malaria infections, as well as supporting the evidence that *P. vivax* malaria is under detected.

We have also reported variability in the coinfection patterns between sampled slides: Two slides (CA, POS) presented more unique *P. falciparum* reads than *P. vivax* reads, one slide presented similar outcomes (CM), while the samples sequenced in 2017 presented almost only *P. vivax* reads. Although we cannot extrapolate this fluctuations at a populational level, we interpret it as a signal of variability in the recurrence of both pathogens in the environment, added to the discussed diagnose limitations.

One of the particularities of *P. vivax* infections, as it has been described in section 1.3.3, are the genuine hypnozytes that can be releasing mersomes for months, causing longstanding infections (White 2011). This promotes the presence of mixed-strain *P. vivax* infections. Multiple strains can be inoculated by the same mosquito in one single bite, and others can be inoculated in posterior bites. The recurrence of multiple-strain infections has been estimated to be nearly to the 50% in most of the populations exposed, finding variable number of strains in each single sample (Pearson et al. 2016). The low genomic depth coverage of our *P. vivax* genome (1.6x) has not permitted us to evaluate the presence of different *P. vivax* isolates in the same sample. It would have required the analysis of the heterozygous calls that have been excluded in all the analyses (Nair et al. 2014). We therefore have assumed the presence of a single strain, taking in account only the homozygous “consensus” positions. A hypothetical mixed-strain sample could have been extremely interesting as could have been used to estimate the intrapopulation diversity in the Ebro Delta.

5.2 *Plasmodium* origin and dispersals

The Plasmodium vixax arrival in the Americas

The peopling of America was the last great human expansion of Late Pleistocene (Skoglund et al. 2016). Modern humans colonized the Americas moving from Siberia trough Beringia (Reich et al. 2012), there are evidences of human settlements in Siberia dated by 28,000 years BP. The oldest American human remains (Gilbert et al. 2008) and the climatic conditions indicate that such migration occurred 14,000-15,000 years BP, at the end of the last Glacial Maximum (Rasmussen et al. 2014), however a recent archaeological finding proposes that such migration might have occurred up to 20,000 years BP (Williams et al. 2018).

Anopheles mosquito split from its closest taxon more than 200 MY BP in Pangea supercontinent (Reidenbach et al. 2003). Very little is known about the dispersal dynamics and speciation of the *Anopheles* taxa across America, Africa and Asia. Nowadays there are 9 *Anopheles* spp. capable of transmitting malaria in America. Two present in North America, and the rest with distributions that encompass areas of Mesoamerica and South America (Orfano et al. 2016). Despite the mosquito was already present long before America was peopled, a *Plasmodium* spread in the Americas linked with the firsts human migrations seems non-credible (Carter et al. 2003), since the settlers of Beringia were not in contact with any *Plasmodium*.

The American anopheline mosquitoes belong to the subgenus *Nyssorhynchus,* this taxon split from the European anopheline vectors approximately 100 MY BP (Moreno et al. 2010). Recent studies have presented high genetic divergence in Pfs47/Pvs47 (*P. falciparum/P. vivax*) genes between American and Asian/African populations (Molina-Cruz et al. 2014, Hupalo et al. 2016), remarkably, in the case of *P. vivax*, the American isolates show a reduced haplotype diversity in Pvs47, which has been interpreted as a mark of non-recent selection. Pvs47 gene plays an important role in the evasion of the mosquito immune response (Molina-Cruz et al. 2013), the genetic divergence observed in American *P. vivax* seems to be an adaptation to New-World mosquitoes. We have failed to recover the sequence of the Pvs47 gene in the Ebro Delta strain. The comparison of this sequence with the American ones could be an evidence of such adaptation.

American *P. vivax* populations are extremely diverse from a genetic point of view. Including high genome-wide diversity, presented with little geographic clusterization (de Oliveira et al. 2017, Cowell et al. 2017, Cowell et al. 2018). However, the existent structure can be observed by f-statistic analysis, showing correlations of genetic and geographical distances (de Oliveira et al. 2017). The huge American diversity is supposed to be the result of multiple recent introductions of *P. vivax,* as well as various admixture events that have occurred in the continent (Taylor et al. 2013, Rodrigues et al. 2014, de Oliveira et al. 2017, Rodrigues et al. 2018, Cowell et al. 2018, Rodrigues. et al 2018). The position of the Ebro-1944 sample, in relation to the American strains, is consistent with a European ancestry of the

American *P. vivax* populations. The totality of the implemented test suggests that the American samples share a great proportion of the genotypes with the European sample. Additionally, the American samples are the ones with the highest genetic affinities with Ebro-1944 *P. vivax*.

A single sample is not enough to quantify the amount of genetic American diversity that can be explained by European introductions. However, our results demonstrate that European strains played a crucial role in the formation of the American *P. vivax* genetic basis. The European contribution can also be complemented by other hypothesis that implicate Melanesian and African sources to the present days American *P. vivax* diversity. Regarding to these hypothesis, is important to contextualize that almost all the publications that support the Melanesian contribution are based on mtDNA analyses (Taylor et al. 2013, Rodrigues et al. 2018). This approach could have biased the results, as the phylogenies were built with very few SNPs. Our analyses indicate that Melanesian samples are the most distant compared with the American ones, as it has also been reported previously (Hupalo et al. 2016, Pearson et al. 2016). Although particular contacts cannot be discarded, and less when human Pacific migrations have contributed to the America peopling (Skoglund et al. 2015), most of the present day American diversity seems to have a European origin.

The dataset that we have used is quite limited in American isolates, and lacks for isolates from some *P. vivax* endemic areas such as Peru or Colombia. This limited American sample size may explain why we have not been able to extract concluding results regarding to single population affinities. Nevertheless, a bigger and more complete dataset may would not be as informative as it could be expected. The little American populational clustering has been demonstrated when *P. vivax* barcoding (Baniecki et al. 2015) has proved to be unspecific with American samples (de Oliveira et al. 2017), due to the populational extreme diversity. More European samples would be much more informative and probably would help to quantify the intrapopulational European diversity and confer more robustness to the discussed hypothesis.

The mitochondrial *P. vivax* genome network also situates the European strain closely related to all the present American diversity. The use of mtDNA has some advantages as compared with genomic DNA studies, despite the fact that has much less polymorphic sites than the nuclear genome. The mitogenome of *P. vivax* has only 6.000 bases compared with the 26.8Mb of the nuclear genome (Carlton et al. 2008), making it easier to be retrieved completely. Another advantage is that mtDNA SNPs appear to be neutral, which allows a reliable tracking of *P. vivax* demographic history (Cornejo et al. 2006, Taylor et al. 2013). Most of the current American mtDNA diversity is grouped in two different clusters that appear in a central position in the worldwide *P. vivax* diversity (Culleton et al. 2011). The finding that the European mtDNA isolate differs by only one substitution from one of these clusters indicates that at least a proportion American diversity must be partially attributable to the

European diversity.

All the studies that have analysed the position of the American strains in the context of the global diversity have evidenced that the closest *P. vivax* strains are the African and Indian ones (Hupalo et al. 2016, Pearson et al. 2016, Rodrigues et al. 2017). This has been used to support the theory of an African origin of the American *P. vivax*, usually explained with the same populational history as the *P. falciparum* one, implicating a trans-Atlantic migration related with slave trade (Rodrigues et al. 2018). The Ebro-1944 strain appears to be genetically close to the American samples, additionally also appears to share genetic diversity with the African ones. Possibly because they share a quite close recent common ancestor, but the absence of *P. vivax* in the Gulf of Guinea, due to the fixation of Duffy- mutation, makes an African origin of the American *P. vivax* be poorly credible (Culleton et al. 2011, Taylor et al. 2013).

The Origin of the European Plasmodium falciparum

P. falciparum infection in humans originated from cross-species transmission of Lavernia parasites from Gorillas in West Africa. This transfer occurred 10,000-100,000 years ago (Liu et al. 2010), and possibly expanded following the human Neolithic spread (Tanabe et al. 2010, Otto et al. 2018). The introduction of P. *falciparum* in the Americas would have followed an independent movement, with different and independent migrations from Africa related with slave trade (Yalcindag et al. 2012, Rodrigues et al. 2018).

The presence of the pathogen in Europe since Classical times is accredited by historical records (Sellares et al. 2004) and recently by the discovery of *P. falciparum* sequences in Ancient Roman remains from the I-II Century of the CE (Marciniak et al. 2016). Unfortunately, the mtDNA retrieved fragment was not long enough to compare it with the present-day diversity. Here we provide the evidence that *P. falciparum* was still present in Southern Europe, being a malaria pathological agent, at least until 20 years before the eradication of the pathology in Spain (Pletsch et al. 1965).

Surprisingly, the retrieved *P. falciparum* mitogenome from the Ebro Delta exhibits three mutations previously described only in Indian strains. This mutation combination has been named PfIndia (Tyagi et al. 2014). This finding suggests a close relation between the European and Indian strains, possibly because both have a common origin. The introduction of *P. falciparum* in Europe is thought to be recent, possibly in historical classical times. (de Zulueta et al. 1973, Carter et al. 2002). The Indian Campaign of Alexander the Great (IV c. BC) could be linked with the dispersal of P. falciparum from India to Greece. However, most the hypothesis link the dispersal with the Roman Empire, as no Greek records describe it.

The present day variation of *P. falciparum* displays that the American strains are closely related with the African ones. The diversity showed by the American populations is elevated, consistent with multiple *P. falciparum* African populations introductions (Yalcindag et al. 2012, Rodrigues et al. 2018), and possibly complemented by

Europeans ones, as well, (Hume et al. 2003), although cannot be quantified.

5.3 *Plasmodium vivax* Genome

We have retrieved 446,529 *P. vivax* unique reads, representing the 66% of the *P. vivax* Sal1 reference sequence, with a mean read depth of 1.62X. After a restrictive variant calling process we have been able to define a set of “confident” SNPs. The genotypes have allowed us the setting of a mutation rate for *P. vivax* and an evaluation of the described signals of adaptation in *P. vivax* populations (Hupalo et al. 2016, Pearson et al. 2016). As it has been previously explained in this section, the genomic diversity of *P. vivax* is notable and remarkably higher than the *P. falciparum* one (Neafsey et al. 2012, Hupalo et al. 2016). Our sample has also contributed to support this evidence. Only in 2,585 out of the 207.827 *P. vivax* variable positions, Ebro-1944 carries the derived allele. This demonstrates that most of the variability of *P. vivax* populations is only shared by little fraction of individuals that differ many from the others.

A mutation rate for Plasmodium vivax

No specific mutation rate has been previously published for *P. vivax*. The rate that has been constantly used is the general eukaryotic mutation rate of 1E^{-9} (Paget-McNicol et al. 2001) which has been widely accepted in the field of *P. falciparum* genetics (Neafsey et al. 2012). We therefore have taken advantage of the Ebro-1944 isolate to place it in a time-scale of *P. vivax* isolates. The resultant phylogeny has revealed an accumulation of mutations in the most recent isolates of the dataset. This results are presumably indicating an increased acquisition of mutations due to selective pressure against drug treatments, as it has been presented previously (Neafsey et al. 2012). Our estimation has resulted in a substitution rate of 4.90E-7^{-7} substitution/site/year (HPD 95% 181 1.07E^{-7} – 1.02E^{-6}). This mutation rate is sensibly higher than the 1.7E^{-9} proposed for *P. falciparum*, obtained from cell lines (Bopp et al. 2013) and the 0,9-1,5 substitution/genome/year obtained with in-vitro *P. falciparum* lines (Otto et al. 2018). Nevertheless as *P. vivax* exhibits higher genetic diversity compared with *P. falciparum*, a higher *P. vivax* mutation rate could be expected (Parobek et al. 2016).

This mutation rate has been used to date the dispersal of *P. vivax*. The calibrated phylogeny indicates that the global *P. vivax* diversity is very recent, probably most of the present strains would have expanded in the last 2,000 years. This TMRCA estimation is quite similar to other proposed (5,000-10,000) (Carter et al. 2003, Leclerc et al. 2004) but faces with estimations that situate the TMRCA of all present day *P. vivax* in 45,000-81,000 years (Escalante et al. 2005) or 53,000-265,000 year (Mu et al. 2005). The dispersal of *P. falciparum* probably occurred linked with farming migrating movements, between 4,000-6,000 BP (Otto et al. 2018), which seems to be comparable to the *P. vivax* history.

Standing variation in the resistance against Antimalarial drugs

Populations tend to adapt to the novel selective agents by selecting existing variants or otherwise expanding novel mutations (Rowan et al. 2008). The Ebro-1944 sample carries the Met205Ile mutation in the DHPS gene, which conferees resistance to Sulfadoxine treatment. (Korsinczky et al. 2004, Hawkins et al. 2009). The Ebro-1944 sample also carries 5 derived alleles in genetic regions with high diversity between populations (Hupalo et al. 2016). This discovery has two main implications. The first one is the report of a drug resistance variant in an isolate that was never in contact with these compounds, evidencing the presence of standing variation enhancing the spread of *P. vivax* resistant strains. The second refers that according to the evidence of this standing variation, aggressive treatments without combined compounds can easily led to the selection of resistant isolates (Read et al. 2011, Huijben et al. 2013). The elevated ratios of genetic diversity observed in *P. vivax* suggest that the situation found in the European *P. vivax* strain is not an exception and is just another evidence of the gigantic diversity of the species.

5.4 Selective signals of malaria in European populations

Variant calling efficiency

Variant calling is a crucial step in any variant analysis pipeline. Ancient samples particularities: i) low rates of endogenous DNA, ii) post-mortem damage, iii) contamination, limit seriously genotype calling confidence (Wall et al. 2007, Shapiro et al. 2014). Usually, when working with aDNA samples, each project evaluates the best calling strategy, with the aim to find the equilibrium between genotyping accuracy and analysis performance. The selected pipeline will condition both the accuracy of the results and the typology of the tests that will be implemented.

Very degraded samples are sequenced with capture approaches, in this situations clustering the samples into populations, and estimating allelic frequencies only at a population level has been previously used to screen ancient samples for signals of selection. In this cases, population allele frequencies can be computed with a maximum-likelihood estimation based on counting the sequences covering each SNP (Mathieson et al. 2015). When analysing shotgun sequencing samples with moderate values of depth coverage (>1X) the genotypes can be called individually, taking into account severe filtration thresholds to certify the reliability of the observed genotypes (Martiniano et al. 2017). An intermediate approach between the previously presented is to analyse selected variants in shotgun low-coverage individuals genotyped with Bayesian based pseudo-haploid callers (Hofmanová et al. 2016).

To maximize the amount of variants obtained, while ensuring to get confident calls we implemented differential calling strategies depending on the technique used in sample sequencing. Shotgun sequencing samples genotypes were called with a regular variant calling using GATK, a hard filtering process and genotype Imputation (Gamba et al. 2014), while capture samples genotypes were recovered implementing pseudo-haploid callings. This approach allowed us to outperform the implementation of pseudo-haploid callings in all the samples of the dataset.

The implemented strategy has successfully recovered a great proportion of the targeted genotypes, especially in shotgun sequenced samples. In this samples the ratios of recovered genotypes are near the 100% of the targeted SNPs. The rates of recovered pseudo-haplotypes in the capture samples were conditioned by the presence of the targeted SNPs in the 1240K dataset (Mathieson et al. 2015). Which has been used in other publications with large sample-size (Lazaridis et al. 2016, Lipson et al. 2017, Olalde et al. 2018). This dataset bias explains why positions like; rs5030868, rs137852314,rs137852328 in G6PD gene, rs33930165, rs33950507, rs334 in HBB gene, rs8176719 in ABO gene, rs2230345 in GRK5 gene, rs1800890 in Il-10 gene, rs3092945 in CD40LG gene and rs201346212 in CD36 were only genotyped in shotgun sequenced individuals. Otherwise SNPs such as rs33950507, present in the SNP array, was genotyped in almost all the samples of the dataset.

As most of the published ancient samples have been sequenced with capture sequencing, the bias that we have observed will be a constant in any study that pretends to genotype a list of alleles selected by prior functional evidences. This limitation will be accompanied by the lack of samples from some specific periods and locations. Nevertheless, if the SNPs are present in the genotyping arrays, the rates of recovered alleles are elevated (Ye et al. 2017).

Malaria footprints in European genetics?

Malaria is known to be one of the strongest selective pressures that have driven human genomes. There are well described genetic variants conferring resistance against *P. vivax* as Duffy negative phenotype (Nagao et al. 2002) or G6PD deficiency (Luzzatto et al. 1969). And also *P. falciparum* driven mutations in HBB gene like HbC or HbE haemoglobins (Ohashi et al. 2004). The existence of different mutations in HBB gene; HbC in African populations and HbE in Asiatic populations, both conferring resistance against *P. falciparum* infections, supports the hypothesis of a recent *Plasmodium* out of Africa, as different populations exposed to the same selective agent have developed different adaptive patterns (Kwiatkowski et al. 2005).

The screened European individuals illustrate a powerless selective impact of Malaria in European human genomes. The impact, if it exists, cannot be comparable with the one of HBB and Duffy (ACKR1 gene) variants, in African populations. We have not found any ancient European individual carrying one of the three traits that clearly protects against *P. vivax* infection: Duffy negativity (ACKR1 gene), Ovalocytosis (in South East Asian groups) and G6PD deficiency (Nagao et al. 2002, Miller et al. 1976, Luzzatto et al. 1969). The recent introduction of malaria in Europe, (Section 3.3), (Carter et al. 2003, Leclerc et al. 2004, Rich et al. 2009, Otto et al. 2018) is probably the main cause of such absence.

There are two mutations in the G6PD gene, which confer resistance against Plasmodium infections that are present in present-day European populations, especially in the Mediterranean ones: G6PD A- (Vulliamy et al. 1991) and G6PD Med (Kirkman et al. 1964). Statistical modelling has suggested that both mutations emerged recently, meaning that the present day frequencies can only be explained by the impact of malaria since the Neolithic (Tishkoff et al. 2001). We have failed in finding a single individual carrying any of these mutations. Although the lack of ancient Mediterranean samples has excluded the most exposed populations, which difficults to find such variants. In future studies it will be necessary to revaluate this data with other possible sequenced ancient Mediterranean individuals. The differences observed between North-South populations in rs8177374 SNP at TIRAP gene, evidence the presence of a geographical clustering.

In contrast, we have been able to detect some statistically significant allele frequency changes that may be explained by malaria presence. Three out the 20 screened variants (rs1050501 in FCGR2B, rs4951074 in ATP2B4, rs8176746 in ABO) showed statistically significant higher allelic frequencies in the present day Europeans compared with the past populations. Regarding to these three variants is also relevant to state that all these three alleles were already detected in Upper Palaeolithic individuals.

Malaria arrival linked with the Neolithic?

The genotyping of large human datasets has permitted the detection of differences in allele frequencies between populations (Sabeti et al. 2007). This differences once combined with the records of different ecological pressures have led to the validation of selective forces and the identification of locus under selective pressure. Moreover, this estimations obtained with the analysis of present populations are little sensible to distinguish between selection of standing variation or novel allele arose, this impediment difficulties the capacity of timing new mutations (Peter et al. 2012).

The sequencing of ancient individuals, and specifically large datasets from different locations and time periods allows direct estimations of mutation timing and allele frequency fluctuations across time (Wilde et al. 2014, Mathieson et al. 2015). The best supported selective signals include: i) metabolic genes: FADS1, DHCR7, lactase (LTC) (Mathieson et al. 2015, Ye et al. 2017), ii) Pigmentation (SLC24A5), iii) Immunity TLR-1, TLR-6, TLR-10, MHC (Mathieson et al. 2015).

We have applied a different strategy to assess the same question. Although we could not identify strong differences of allelic frequencies between ancient and modern populations, and the time series did not show extreme variations, we implemented a Bayesian framework (Schraiber et al. 2016) to infer on the presence of selective forces driving Malaria genetic variants, which is a novel strategy to detect signals of pathogen selection in aDNA time series. The application of scale series has been previously implemented in horse evolution (Librado et al. 2017) as well as in the evolution of human genes, highlighting metabolic adaptations (Ye et al. 2017).

We have determined that all the screened variants were already present in human populations long before the Neolithic onset, with the exception of FCGR2B rs1050501 mutation, that has an estimated age of only 9,500 years. Albeit, we have not been able to detect fluctuations in the allelic frequencies that could be showing a clear tendency of Neolithic impact. Probably the little number of Palaeolithic and Mesolithic individuals as well as the impossibility to cluster the samples into geographical populations is one the main causes of such results. Also an issue that will have to be addressed in future publications is the population selection and comparisons. The development of effective approaches to test continuity (Schreiber et al. 2018) will allow to determine if observed differences can fairly be related with genetic selection instead of being clues of the presence of admixtures or genetic drift.

5.5 Conclusions and future directions

The principal aim of the thesis was to integrate both human genotyping data and European *Plasmodium* genome sequences to illustrate and uncover the history of malaria in Europe. Both focusing on the impact of the disease on human populations as well as defining the migrating movements and evolutionary history of the European *Plasmodium* strains

In this thesis we have presented the first whole genome sequence of a European eradicated *P. vivax* strain. Furthermore, we also provide the mitogenomes of both European *P. vivax* and *P. falciparum.* This achievement is extremely relevant, as this sequences are unique and no other European *Plasmodium* genomes have been retrieved, except few mitochondrial *P. falciparum* bases.

The sequencing of the first European *P. vivax* isolate is particularly outstanding. The analyses have proved a European source for most of the *P. vivax* diversity in the Americas. The phylogeny date calibration has revealed that *P. vivax* colonized America 500 years ago. This dating strongly supports the genomic evidences of a European source of the American *P. vivax* diversity. Hypothetical future European isolates, coming from other medical collections, from bone tissue or even from mosquito preserved in wetlands sediments, will probably confirm the results that we present, and will also allow the study of intra-European population movements.

We provide a mutation rate for *P. vivax* genome that has been used to reveal an extreme recent worldwide expansion of its populations. The high value of this rate is in accordance with the elevated values of genetic variability observed in *Plasmodium* populations.

It is an evidence that the expansion and emergence of *Plasmodium* strains resistant to antimalarial treatments can sonly represent a serious thread for malaria cure. In this sense a deep comprehension of the genetic dynamics of drug resistance development, can be a possible source novelties in the treatment and eradication of Malaria around the world. Here we report that Ebro-1944 *P. vivax* carries yet one mutation linked with Antimalarial drug resistance. This finding should be interpreted as an evidence of high standing variation in *P. vivax* genomes that would dramatically facilitate the selection of pre-existing variants. More sequences and further research in this direction can be notably interesting

The study of ancient pathogens is extremely interesting. It brings us relevant information related with human migrations and genetic adaptations. Pathogens are usually strong selective pressures that have shaped our genome and have conditioned us, its study allows a major comprehension of our evolution. Soon more pathogenic sequences will be studied, and more genetic mechanisms will be understood, and hopefully in a near future, malaria research will only be placed in the category of ancient pathogens.

6. OTHER PUBLICATIONS

Gelabert P, Ferrando-Bernal M, de-Dios T, Mattore B, Campoy E, Gorostiza A, Patin E, González-Martín A, Lalueza-Fox C. Genome-wide data from the Bubis from Bioko Island represents the Atlantic fringe of the Bantu dispersal. 2018. under revision in Genome Biology and Evolution.

7. REFERENCES

Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E, et al. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci USA. 1999;96(24):14043–14048.

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522(7555):167-172.

Alonso PL, Brown G, Arevalo-Herrera M, Binka F, Chitnis C, Collins F, et al. A research agenda to underpin malaria eradication. PLoS Med. 2011; 8(1): e1000406.

Amato R, Miotto O, Woodrow CJ, Almagro-Garcia J, Sinha I, Campino S, et al. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5:pii:e08714.

Amino R, Thiberge S, Martin B, Celli S, Shorte S, Frischknecht F, et al. Quantitative imaging of *Plasmodium* transmission from mosquito to mammal. Nat Med. 2006;12(2):220-224.

Andam CP, Worby CJ, Chang Q, Campana MG. Microbial Genomics of Ancient Plagues and Outbreaks. Trends Microbiol. 2016;24(12):978-990.

Andrades-Valtueña A, Mittnik A, Key FM, Haak W, Allmäe R, Belinskij A, et al. The Stone Age Plague and Its Persistence in Eurasia. Curr Biol. 2017;27(23):3683-3691.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available online at:http://www.bioinfoe1000406.rmatics.babraham.ac.uk/projects/fastqc

Arisue N, Hirai H, Arai M, Matsuoka H, Horii T. Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*. J Mol Evol. 2007;65(1):82–91.

Auburn S, Marfurt J, Maslen G, Campino S, Ruano Rubio V, Manske M, et al. Effective preparation of *Plasmodium* vivax field isolates for high-throughput whole genome sequencing. PLoS One. 2013;8(1):e53160.

Auburn S, Böhme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, et al. A new *Plasmodium* vivax reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. Wellcome Open Res. 2016;1:4.

Baird JK. Evidence and implications of mortality associated with acute *Plasmodium* vivax malaria. Clin Microbiol Rev. 2013;26(1):36-57.

Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution. 1999;16:37–48.

Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF, et al. Development of a single nucleotide polymorphism barcode to genotype *Plasmodium* vivax infections. PLoS Negl Trop Dis. 2015;9(3):e0003539.

Barcus MJ, Basri H, Picarima H, et al. Demographic risk factors for severe and fatal vivax and falciparum malaria among hospital admissions in northeastern Indonesian Papua. Am J Trop Med Hyg. 2007;77(5):984-991.

Barnwell JW, Ingravallo P, Galinski MR, Matsumoto Y, Aikawa M. *Plasmodium* vivax: Malarial proteins associated with the membrane-bound caveola-vesicle complexes and cytoplasmic cleft structures of infected erythrocytes. Exp Parasitol. 1990;70(1):85-99.

Barrett RDH, Schluter D. Adaptation from standing genetic variation. Trends in Ecology & Evolution. 2008;23(1):38-44.

Bennett EA, Massilani D, Lizzo G, Daligault J,Geigl EM, Grange T. Library construction for ancient genomics: single strand or double strand? BioTechniques. 2014;56:289–298.

Bernabeu M, Lopez FJ, Ferrer M, Martin-Jaular L, Razaname A, Corradin G, et al. Functional analysis of *Plasmodium* vivax VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. Cell Microbiol. 2012; 14(3):386-400.

Bocquet-Appel JP. When the world's population took off: the springboard of the Neolithic Demographic Transition. Science. 2011. 29;333(6042):560-561.

Bogdonoff MD, Crellin JK, Good RA, McGovern JP, Nuland SB, Saffon MH. 1985. The Genuine Work of Hippocrates (Hippocrates, Epidemics 1.6,7,24–26; Aphorisms 3.21,22;4.59,63; On Airs, Waters and Places c. 10). Classics of Medicine Library, Birmingham, AL. 5. Ebers G. 1875.

Bopp SE, Manary MJ, Bright AT, Johnston GL, Dharia NV, Luna FL, et al. Mitotic evolution of *Plasmodium* falciparum shows a stable core genome but recombination in antigen families.PLoS Genet. 2013;9(2):e1003293.

Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of Yersinia pestis from victims of the Black Death.Nature. 2011;478(7370):506-510.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514(7523):494-497.

Bos KI, Jäger G, Schuenemann VJ, Vågene AJ, Spyrou MA, Herbig A, et al. Parallel detection of ancient pathogens via array-based DNA capture. Philos Trans R Soc Lond B Biol Sci. 2015;370(1660):20130375.

Boundenga L, Ollomo B, Rougeron V, Mouele LY, Mve-Ondo B, Delicat-Loembet LM, et al. Diversity of malaria parasites in great apes in Gabon. Malaria Journal. 2014;14(111).

Bouwman AS, Kennedy SL, Müller R, Stephens RH, Holst M, Caffell AC, et al. Genotype of a historic strain of Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2012;109(45):18511-18516.

Bright AT, Tewhey R, Abeles S, Chuquiyauri R, Llanos-Cuentas A, Ferreira MU, et al. Whole genome sequencing analysis of *Plasmodium* vivax using whole genome capture. BMC Genomics. 2012;13:262.

Bramanti B, Thomas MG, Haak W, Unterlaender M, Jores P,Tambets K, et al. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. Science. 2009;326(5949):137–140.

Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, et al. Early Neolithic genomes from the eastern Fertile Crescent. Science. 2016;353(6298):499-503.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084-1097.

Bruce-Chwatt L, de Zulueta J. The rise and fall of Malaria in Europe: a historico-epidemiological study. Oxford university press. 1980. Pp 18-25.

Buermans HP, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. 2014;1842(10):1932-1941.

Calderaro A, Piccolo G, Gorrini C, Rossi S, Montecchini S, Dell'Anna ML, et al. Accurate identification of the six human *Plasmodium* spp. causing imported malaria, including *Plasmodium* ovale wallikeri and *Plasmodium* knowlesi. Malar J. 2013;12:321

Campana MG, Robles N, Ruhli F, Tuross N. False positives complicate ancient pathogen identifications using high-throughput shotgun sequencing. BMC Research Notes.

2014;7(111):10.1186/1756-0500-7-111.

Cappellini E, Jensen LJ, Szklarczyk D, Ginolhac A, da Fonseca RA, Stafford Jr TW, et al. Proteomic Analysis of a Pleistocene Mammoth Femur Reveals More than One Hundred Ancient Bone Proteins. J Proteome Res. 2012;11(2):917-926.

Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium* vivax.Nature. 2008; 455(7214):757-763.

Carlton JM, Das A, Escalante AA. Genomics, population genetics and evolutionary history of *Plasmodium* vivax. Adv Parasitol. 2013;81:203-222.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, et al. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. Am J Hum Genet. 2013;93(5): 852–864.

Carter R, Mendis KN. Evolutionary and historical aspects of the burden of malaria. Clin Microbiol Rev. 2002;15(4):564–594.

Carter N, Pamba A, Duparc S, Waitumbi JN. Frequency of glucose-6-phosphate dehydrogenase deficiency in malaria patients from six African countries enrolled in two randomized anti-malarial clinical trials. Malar J. 2011;10:241.

Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, Bradley DG, et al. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. Proc Natl Acad Sci U S A. 2016 ;113(2):368-373.

Churko JM, Mantalas GL, Snyder MP, Wu JC. Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases. Circulation Research. 2013;112(12):1613-1623.

Contacos PG, Coatney GR, Orihel TC, Collins WE, Chin W, Jeter MH. Transmission of *Plasmodium* schwetzi from the chimpanzee to man by mosquito bite. Am J Trop Med Hyg. 1970;19(2):190-195.

Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature. 2001;409(6821):704-707.

Cornejo OE, Escalante AA. The origin and age of *Plasmodium* vivax. Trends Parasitol. 2006;22(12):558–563.

Cowell AN, Loy DE, Sundararaman SA, Valdivia H, Fisch K, Lescano AG, et al. Selective Whole-Genome Amplification Is a Robust Method That Enables Scalable Whole-Genome Sequencing of *Plasmodium* vivax from Unprocessed Clinical Samples. MBio. 2017;8:1.

Cowell AN, Valdivia HO, Bishop DK, Winzeler EA. Exploration of *Plasmodium* vivax transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. Genome Med. 2018;10(1):52.

Cui Y, Yu C, Yan Y, Li D, Li Y, Jombart T, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci USA. 2013; 110(2):577–582.

Culleton R, Coban C, Zeyrek FY, Cravo P, Kaneko A, Randrianarivelojosia M, et al. The origins of African *Plasmodium* vivax; insights from mitochondrial genome sequencing. PLoS One. 2011;6(12):e29137.

Culleton R, Carter R. African *Plasmodium* vivax: distribution and origins. Int J Parasitol. 2012;42(12):1091-1907.

Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J. The pir multigene family of *Plasmodium*: antigenic variation and beyond.Mol Biochem Parasitol. 2010;170(2):65-73.

Dabney J, Meyer M, Pääbo S. Ancient DNA damage. Cold Spring Harb Perspect Biol. 2013;5(7):1–7.

Damgaard PdB, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018:557(7705):369-374.

de Oliveira TC, Rodrigues PT, Menezes MJ, Gonçalves-Lopes RM, Bastos MS, Lima NF, et al. Genome-wide diversity and differentiation in New World populations of the human malaria parasite *Plasmodium* vivax. PLoS Negl Trop Dis. 2017;11(7):e0005824.

Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, Tien JH, et al. Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849. N Engl J Med. 2014;370(4):334-340.

De Zulueta J. Malaria and Mediterranean history. Parassitologia. 1973;15(1):1–15.

Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418(6898):700-707.

Díaz J, Ballester F, López-Vélez R, Impacts on Human Health. In The Preliminary Assessment of the Impacts in Spain due to Effects of Climate Change; Project ECCE (Evaluación de los Impactos del Cambio Climático en España), Ministry of the Environment: Madrid, Spain, 2005; pp. 699–741.

Dvorin JD, Martyn DC, Patel SD, Grimley JS, Collins CR, Hopp CS, et al. A plant-like kinase in *Plasmodium* falciparum regulates parasite egress from erythrocytes. Science. 2010;328(5980):910-912.

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 USA. 1998;95(21):12637–12640.

Drancourt M, Raoult D. Palaeomicrobiology: current issues and perspectives. Nature Reviews Microbiology. 2005;3(1):23-55.

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7 Molecular Biology And Evolution. 2012;29(8):1969-1973.

Dong L, Xi M, Thann F. Les maux épidémiques dans l'empire chinois. Paris. L'Harmattan. 1996. P 95-96.

Duggan AT, Perdomo MF, Piombino-Mascali D, Marciniak S, Poinar D,. Emery MV, et al. 17th Century Variola Virus Reveals the Recent History of Smallpox. Curr Biol. 2016;26(24):3407–3412.

Douglas NM, Nosten F, Ashley EA, Phaiphun L, van Vugt M, Singhasivanon P, et al. *Plasmodium* vivax recurrence following falciparum and mixed species malaria: risk factors and effect of antimalarial kinetics. Clin Infect Dis;52(5):612-620.

Dye C. After 2015: infectious diseases in a new era of health and development. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1645):20130426.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32(5):1792-1797.

el Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez CP, et al. A superfamily of variant genes encoded in tdhe subtelomeric region of *Plasmodium* vivax. Nature. 2001; 410(6830):839-42.

Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, et al. A monkey's tale: the origin of *Plasmodium* vivax as a human malaria parasite.Proc Natl Acad Sci U S A. 2005;102(6):1980-1985.

Feachem RGA, Phillips AA, Targett GA, Snow RW. Call to action: priorities for malaria elimination. Lancet. 2010;376(9752): 1517–21.

Fernández H, Hughes S, Vigne JD, Helmer D, Hodgins G, Miquel C, et al. Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. Proc Natl Acad Sci U S A. 2006;103(42):15375-15379.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, et al. Mosquito genomics.Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015;347(6217).

Frevert U, Engelmann S, Zougbédé S, Stange J, Ng B, Matuschewski K, et al. Intravital observation of *Plasmodium* berghei sporozoite infection of the liver. PLoS Biol. 2005;3(6):e192.

Fregel R, Méndez FL, Bokbot Y, Martín-Socas D, Camalich-Massieu MD, Santana J, Morales J, et al. Ancient genomes from North Africa evidence prehistoric migrations to the Maghreb from both the Levant and Europe. Proc Natl Acad Sci U S A. 2018:115(26):6774-6779

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, et al. DNA analysis of an early modern human from Tianyuan Cave, China. PNAS. 2013;110(6): 2223-2227.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. The genetic history of Ice Age Europe. Nature. 2016;534(7606):200-205.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R.Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7(11):e1002355.

Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science. 2015;350(6262):820-822.

Galvani AP, Slatkin M. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. Proc Natl Acad Sci U S A. 2003;100(25):15276–15279.

Gamba C, Jones ER, Teasdale MD, McLaughlin RL, Gonzalez-Fortes G, Mattiangeli V, et al. Genome flux and stasis in a five millennium transect of European prehistory. Nat Commun. 2014;5:5257.

Gamba C, Hanghøj K, Gaunitz C, Alfarhan AH, Alquraishi SA, Khaled AS, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol Ecol Resour. 2016;16:459–469.

García-Garcerà M, Gigli E, Sanchez-Quinto F, Ramirez O, Calafell F, Civit S, et al. Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics. PLoS One. 2011;6(8):e24161.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al.Genome sequence of the human malaria parasite *Plasmodium falciparum*.PMC. 2002;419(6906):498-511.

Gansauge MT, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–748.

Gernaey AM, Minnikin DE, Copley MS, Dixon RA, Middleton JC, Roberts CA. Mycolic acids and ancient DNA confirm an osteological diagnosis of tuberculosis. Tuberculosis. 2001;81:259–265.

Gerszten E, Allison MJ, Maguire B.Paleopathology in South American Mummies: A Review and New Findings. Pathobiology. 2012;79(5):247-256.

Gething PW, Elyazar IR, Moyes CL, Smith DL, Battle KE, Guerra CAA, et al. Long Neglected World Malaria Map: *Plasmodium* vivax Endemicity in 2010. PLoS Negl Trop Dis. 2012;6(9):e1814.

Gilabert A, Otto TD, Rutledge GG, Franzon B, Ollomo B, Arnathau C, et al. *Plasmodium* vivax-like genome sequences shed new insights into *Plasmodium* vivax biology and evolution. Plos Biol. 2018;16(8):e2006035.

Gilbert MTP, Rudbeck L, Willerslev E, Hansen AJ, Smith C, Penkman KEH, et al. Biochemical and physical correlates of DNA contamination in archaeological human bones and teeth excavated at Matera, Italy. J Arch Sci. 2005;32:785–793.

Gilbert MTP, Hansen AJ, Willerslev E, Turner-Walker G, Collins M. Insights into the processes behind the contamination of degraded human teeth and bone samples with exogenous sources of DNA. Int J Osteoarchaeol. 2006;16:156–164.

Gilbert MTP. Postmortem Damage of Mitochondrial DNA. In: Human Mitochondrial DNA and the Evolution of Homo sapiens. Berlin Heidelberg: Springer; 2006. pp. 91– 115.

Gilbert MTP, Binladen J, Miller W, Wiuf C, Willerslev E, Poinar H, et al. Recharacterization of ancient DNA miscoding lesions: Insights in the era of sequencing-by-synthesis. Nucleic Acids Res. 2007;35(1):1–10.

Gilbert MTP, Jenkins DL, Gotherstrom A, Naveran N, Sanchez JJ, Hofreiter M, et al. DNA from pre-Clovis human coprolites in Oregon, North America. Science. 2008;320(5877):786-789.

Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics. 2011;27(15):2153-2155.

Ginouves M, Veron V, Musset L, Legrand E, Stefani A, Prevot G, et al. Frequency and distribution of mixed *Plasmodium* falciparum-vivax infections in French Guiana between 2000 and 2008.Malar J. 2015;10:446.

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PLF, Uhler C, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008; 134(3):416–426.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-722.

Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, et al. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science. 2018;360(6384):111-114

Haldane JBS. The rate of mutation of human genes. Hereditas. 1949;35(1):267-273.

Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC.Extreme mutation bias and high AT content in *Plasmodium* falciparum. Nucleic Acids Res. 2017;45(4): 1889–1901.

Hawkins VN, Suzuki SM, Rungsihirunrat K, Hapuarachchi HC, Maestre A, Na-Bangchang K, et al. Assessment of the origins and spread of putative resistance-conferring mutations in *Plasmodium* vivax dihydropteroate synthase.Am J Trop Med Hyg. 2009;81(2):348-355.

Heintzman PD, Zazula GD, Cahill JA, Reyes AV, MacPhee RDE, Shapiro B, et al. Genomic Data from Extinct North American Camelops Revise Camel Evolutionary History. Molecular Biology and Evolution. 2015;32(9):2433–2440.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history.Science. 2014; 343(6172):747–751.

Herrick JB. Peculiar, elongated and sickle-shaped red blood corpuscles in a case of severe anemia. Arch Intern Med. 1910;6:517–552.

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC.DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312(5991):282-284.

Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. Nat Rev Genet. 2001;2(5):353-359.

Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009;10(9):639-650.

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

Höss M, Jaruga P, Zastawny TH, Dizdaroglu M, Pääbo S. DNA damage and DNA sequence retrieval from ancient tissues. Nucleic Acids Res. 1996;24(7): 1304–1307.

Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria C, Gething P, et al. The global distribution of the Duffy blood group. Nat. Commun. 2011;2(266).

Howes RE, Battle KE, Mendis KN, Smith DL, Cibulskis RE, JK Baird, et al. Global Epidemiology of *Plasmodium* vivax. Am J Trop Med Hyg. 2016;95(6):15–34.

Hughes AL. The evolution of amino acid repeat arrays in *Plasmodium* and other organisms. J. Mol. Evol 2004;59:528–535.

Hughes AL, Verra F. Malaria parasite sequences from chimpanzee support the co-speciation hypothesis for the origin of virulent human malaria (*Plasmodium falciparum*). Mol Phylogenet Evol. 2010;57(1):135-143.

Huijben S, Bell AS, Sim DG, Tomasello D, Mideo N, Day T, et al. Aggressive Chemotherapy and the Selection of Drug Resistant Pathogens. PLoS Pathog. 2013;9(9): e1003578.

Hume JC, Lyons EJ, Day KP. Review Human migration, mosquitoes and the evolution of *Plasmodium* falciparum. Trends Parasitol. 2003; 19(3):144-9.

Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29(13):1682-1684.

Joy S, Mukhi B, Ghosh SK, Achur RN, Gowda DC, Surolia N.Drug resistance genes: pvcrt-o and pvmdr-1 polymorphism in patients from malaria endemic South Western Coastal Region of India. Malar J. 2018;17(1):40.

Karlsson EK, Kwiatkowski DP, Sabeti CP. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014;15(6): 379–393.

Kay GL, Sergeanta MJ, Giuffra V, Bandiera P, Milanesed M, Bramanti B, et al. Recovery of a Medieval Brucella melitensis Genome Using Shotgun Metagenomics. mBio. 2014;5:4.

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 2012;3(698).

Khor CC, Chapman SJ, Vannberg FO, Dunne A, Murphy C, Ling EY, et al. A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. Nature Genet. 2007;39(4):523-528.

Kirchman JJ, Schirtzinger EE, Wright TF. Phylogenetic Relationships of the Extinct Carolina Parakeet (Conuropsis carolinensis) Inferred from DNA Sequence Data. The Auk. 2012;129(2):197-204.

Kirkman HN, McCurdy PR, Naiman JL. Functionally abnormal glucose-6-phosphate dehydrogenases. Cold Spring Harbor Symp Quant Biol. 1964;29:391-398.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. Timing the Ancestor of the HIV-1 Pandemic Strains. Science. 2000;288(5472):1789-1796.

Korsinczky M, Fischer K, Chen N, Baker J, Rieckmann K, Cheng Q. Sulfadoxine resistance in *Plasmodium* vivax is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site. Antimicrob Agents Chemother. 2004;48(6):2214-22.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283-2285.

Kochar DK, Saxena V, Singh N, Kochar SK, Kumar SV, Das A. *Plasmodium* vivax malaria. Emerg Infect Dis. 2005;11:132–134.

Kolman CJ, Centurion-Lara A, Lukehart SA, Owsley DA, Tuross N. Identification of Treponema pallidum subspecies pallidum in a 200-year-old skeletal specimen. J Infect Dis. 1999;180(6):2060–2063.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Pääbo S. The derived FOXP2 variant of modern humans was shared with Neanderthals. Curr Biol. 2007;17(21):1908-1912.

Krause-Kyora B, Nutsua M , Boehme L, Pierini F, Pedersen DD, Kornell SC, et al. Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. Nat Commun. 2018;9(1):1569.

Krotoski WA. Discovery of the hypnozoite and a new theory of malaria relapse. Trans R Soc Trop Med Hyg. 1985;79(1):1–11.

Kwiatkowski DP. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am. J. Hum. Genet. 2005;77(2):171–192.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409-413.

Laayouni H, Oosting M, Luisi P, Ioana M, Alonso S, Ricaño-Ponce I, et al. Convergent evolution in European and Rroma populations reveals pressure exerted by plague on Toll-like receptors. PNAS. 2014;111(7):2668-2673.

Lalueza-Fox C, Römpler H, Caramelli D, Stäubert C, Catalano G, Hughes D, et al. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. Science. 2007;318(5855):1453-1455.

Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;536(7617):419-424.

Leclerc MC, Durand P, Gauthier C, Patot S, Billotte N, Menegon M, et al. Meager genetic variability of the human malaria agent *Plasmodium* vivax.Proc Natl Acad Sci U S A. 2004;101(40):14455-14460.

Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362(6422):709-715.

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research. 2008;18:1851-1858.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009; 25(14):1754-1760.

Li Y, Willer C, Sanna S, Abecasis G. Genotype Imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.

Librado P, Gamba C, Gaunitz C, Der Sarkissian C, Pruvost M, Albrechtsen A, et al. Ancient genomic changes associated with domestication of the horse. Science. 2017;356(6336):442-445.

Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes. 2012;5:337.

Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M. et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science. 2018; pii: eaat3188.

Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R6, Buckley H, et al. Population Turnover in Remote Oceania Shortly after Initial Settlement. Curr Biol. 2018 2;28(7):1157-1165.

Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, Keele BF. Origin of the human malaria parasite *Plasmodium* falciparum in gorillas. Nature. 2010. 23;467(7314):420-425.

Liu W, Sundararaman SA, Loy DE, Learn GH1, Li Y1, Plenderleith LJ, et al. Multigenomic Delineation of *Plasmodium* Species of the Laverania Subgenus Infecting WildLiving Chimpanzees and Gorillas. Genome biology and evolution. 2016;8(6):1929-1239.

Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics. 2013;193:1233–1254.

Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, et al. Out of Africa: origins and evolution of the human malaria parasites *Plasmodium* falciparum and *Plasmodium* vivax. Int J Parasitol. 2017;47(2-3):87-97.

Luo Z, Sullivan SA,Carlton JM. The biology of *Plasmodium* vivax explored through genomics. Ann N Y Acad Sci. 2015; 1342(1): 53–61.

Luzzatto L, Usanga FA, Reddy S. Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites. Science. 1969;164:839–842.

Margaryan A, Hansen HB, Rasmussen S, Sikora M, Moiseyev V, Khoklov A, et al. Ancient pathogen DNA in human teeth and petrous bones. Ecol Evol. 2018;8(6):3534-3542.

Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. Nat Commun. 2016;7:10326.

Martiniano R, Cassidy LM, Ó'Maoldúin R, McLaughlin R, Silva

NM, Manco L, et al. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. Plos Genet. 2017; 13(7):e1006852.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians.Nature. 2015;528(7583):499-503.

Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018;8;555(7695):197-203.

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 Research. 2010;20(9):1297-1303.

Menard D, Chan ER, Benedet C, Ratsimbasoa A, Kim S, Chim P, et al. Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium* vivax strains. PLoS Negl Trop Dis. 2013;7(11):e2489.

Menard D, Barnadas C, Bouchier C, Henry-Halldin C, Gray LR, Ratsimbasoa A, et al. *Plasmodium* vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. Proc Natl Acad Sci USA. 2010;107:5967–5971.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012;338(6104):222-226.

Miller LH, Mason SJ, Clyde DF, McGinniss MH. The resistance factor to *Plasmodium* vivax in blacks. The Duffy-blood-group genotype, FyFy. N Engl J Med. 1976;295(6):302-304.

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, et al. Sequencing the nuclear genome of the extinct woolly mammoth. Nature. 2008;456(7220):387-90.

Min-Oo G, Gros P. Erythrocyte variants and the nature of their malaria protective effect. Cell Microbiol. 2005;7(6):753-763.

Molina-Cruz A, Garver LS, Alabaster A, Bangiolo L, Haile A, Winikor J, et al. The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. Science. 2013;340(6135):984-987.

Molina-Cruz A, Barillas-Mury C. The remarkable journey of adaptation of the *Plasmodium* falciparum malaria parasite to New World anopheline mosquitoes. Mem Inst Oswaldo Cruz. 2014;109(5):662-667.

Moreno M, Marinotti O, Krzywinski J, Tadei WP, James AA, Achee NL, et al. Complete mtDNA genomes of Anopheles darlingi and an approach to anopheline divergence time. Malar J. 2010;9:127.

Motazedian H, Karamian M, Noyes HA, Ardehali S. DNA extraction and amplification of leishmania from archived, Giemsa-stained slides, for the diagnosis of cutaneous Leishmaniasis by PCR. Ann Trop Med Parasitol. 2002;96(1):31-34.

Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J. Host switch leads to emergence of *Plasmodium* vivax malaria in humans. Mol Biol Evol. 2005;22(8):1686-1893.

Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso P et al. Key gaps in the knowledge of *Plasmodium* vivax, a neglected human malaria parasite. Lancet Infect Dis. 2009:9:555-566.

Mühlemann B, Jones TC, Damgaard PdB, Allentoft ME, Shevnina I, Logvin A. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. Nature. 2018:557(7705):418-423.

Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, et al. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science. 2017;358(6365):951-954.

Mutolo MJ, Jenny LL, Buszek AR, Fenton TW, Foran DR. Osteological and molecular identification of brucellosis in ancient Butrint, Albania. Am J Phys Anthropol. 2012;147:254–263.

Nagao E, Seydel KB, Dvorak, JA. Detergent-resistant erythrocyte membrane rafts are modified by a *Plasmodium* falciparum infection. Exp Parasitol. 2002;102:57–59.

Naing C, Whittaker MA, Wai VN, Mak JW. Is *Plasmodium* vivax Malaria a Severe Malaria?: A Systematic Review and Meta-Analysis. PLoS Negl Trop Dis. 2014;8(8):e3071.

Nair S, Nkhoma SC, Serre D, Zimmerman PA, Gorena K, Daniel BJ, et al. Single-cell genomics for dissection of complex malaria infections. Genome Res. 2014;24(6):1028-1038.

Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite *Plasmodium* vivax exhibits greater genetic diversity than *Plasmodium* falciparum.Nat Genet. 2012;44(9):1046-1050.

Nguyen-Hieu T, Aboudharam G, Signoli M, Rigeade C, Drancourt M, Raoult D. Evidence of a louse-born outbreak involving typhus in Douai, 1710–1712 during the War of Spanish Succession. PLoS One. 2010;5:e15405.

Nosten F, White NJ. Artemisinin-based combination treatment of falciparum malaria. Am J Trop Med Hyg. 2007. 77(6):181-92.

Nunes-Silva S, Dechavanne A, Moussiliou N, Pstrąg JP ,Semblat S, Gangnard N, et al. Beninese children with cerebral malaria do not develop humoral immunity against the IT4-VAR19-DC8 PfEMP1 variant linked to EPCR and brain endothelial binding. Malar J. 2015;14:493.

Nylander JA., Wilgenbusch JC, Warren DL, Swofford DL. AWTY (Are We There Yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 2007;24(4):581-583.

Ohashi J, Naka I, Patarapotikul J, Hananantachai H, Brittenham G, Looareesuwan S, et al. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. Am. J. Hum. Genet. 2004;74(6):1198–1208.

Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CW, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507(7491):225-8.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555(7695):190-196.

Orfano AS, Duarte AMP, Molina-Cruz A, Pimenta PF, Barillas-Mury C. *Plasmodium* yoelii nigeriensis (N67) Is a Robust Animal Model to Study Malaria Transmission by South American Anopheline Mosquitoes. PLoS One. 2016;11(12):e0167178.

Orlando L, Calvignac S, Schnebelen C, Douady CJ, Godfrey LR, Hänni C.DNA from extinct giant lemurs links archaeolemurids to extant indriids.BMC Evol Biol. 2008;8:121.

Orlando L, Calvignac S Schnebelen C, Douady CJ, Godfrey LR, Intarapanich A, Shaw PJ, Assawamakin A, Wangkumhang P, Ngamphiw C, Chaichoompu K, et al. Iterative pruning PCA improves resolution of highly structured populations. BMC Bioinformatics. 2009;10:382.

Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, Magnussen K, et al. True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res;2011;21(10):1705-1719.

Orlando L, Gilbert MT, Willerslev E. Reconstructing ancient genomes and epigenomes. Nat Rev Genet. 2015;16(7):395-408.

Ortner DJ. Human skeletal paleopathology. Int. J Paleopathol. 2011;1:4-11.

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):687-697.

Ottoni C, Flink LG, Evin A, Geörg C, DeCupere B, Van Neer W, et al. Pig domestication and human-mediated dispersal in western Eurasia revealed through ancient DNA and geometric morphometrics. Mol. Biol. Evol. 2013;30(4):824-832.

Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci U S A. 1989;86(6):1939-1943.

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, et al. Genetic analyses from ancient DNA. Annu Rev Genet. 2004;38:645–679.

Paget-McNicol, S., and A. Saul. 2001. Mutation rates in the dihydrofolate reductase gene of *Plasmodium* falciparum. Parasitology 122:497-505.

Papyros Ebers. Das hermetische Buch über die Arzneimittel der Alten Ägypter. W. Engelmann Verlag, Leipzig, Germany.

Parobek CM, Lin JT, Saunders DL, Barnett EJ, Lon C, Lanteri CA, Balasubramanian S, et al. Selective sweep suggests transcriptional regulation may underlie *Plasmodium* vivax resilience to malaria control measures in Cambodia.Proc Natl Acad Sci U S A. 2016;113(50):E8096-E8105.

Pattanasin S, Proux S, Chompasuk D, Luwiradaj K, Jacquier P, et al. Evaluation of a new *Plasmodium* lactate dehydrogenase assay (OptiMAL-IT) for the detection of malaria. Trans R Soc Trop Med Hyg. 2003;97:672–674.

Patterson N, Price AL, Reich D. Population structure and eigenanalysis.PLoS Genet. 2006;2(12):e190.

Pearce-Duvet JM. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. Biol. Rev. Camb. Philos. Soc. 2006:81(3):369–382.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, et al. EAGER: efficient ancient genome reconstruction. Genome Biol. 2016;17:60.

Perkins SL. Molecular systematics of the three mitochondrial protein coding genes of malaria parasites: corroborative and new evidence for the origins of human malaria. Mitochondrial DNA. 2008;19(6):471-474.

Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet. 2012;8(10):e1003011.

Peter MP. Admixture, Population Structure, and F-Statistics. Genetics. 2016;202(4):1485-1501.

Pitulko V, Nikolsky P, Girya E, Basilyan A, Tumskoy V, Koulakov S, et al. The Yana RHS site: humans in the Arctic before the last glacial maximum. Science;303(5654):52-56.

Pletsch D. Informe sobre una misión efectuada en España en septiembre-noviembre de 1963 destinada a la certificación de la erradicación del paludismo. Rev Sanid Hig Publica. 1965;39(7/9):309–367.

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006;311(5759):392-394.

Price RN, Tjitra E, Guerra CA, Yeung S, White NJ, Anstey NM. Vivax malaria: neglected and not benign. Am J Trop Med Hyg. 2007;77(6):79-87.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505(7481): 43-49.

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358(6363):655-658.

Prugnolle F, Rougeron V, Becquart P, Berry A, Makanga B, Rahola N, et al. Diversity, host switching and evolution of *Plasmodium* vivax infecting African great apes. Proc Natl Acad Sci U S A. 2013;110(20):8123–8128.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2014;505(7481):87-91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Pleistocene and recent population history of Native Americans. Science. 2015;21;349(6250):aab3884.

Ramírez O, Gigli E, Bover P, Alcover JA, Bertranpetit J, Castresana J, et al. Paleogenomics in a Temperate Environment: Shotgun Sequencing from an Extinct Mediterranean Caprine. PLoS One. 2009;4(5):e5670.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010; 463(7282):757-762.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506(7487):225-229.

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren KG, et al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell. 2015;163(3):571-582.

Read AF, Day T, Huijben S.The evolution of drug resistance and the curious orthodoxy of aggressive chemotherapy.Proc Natl Acad Sci U S A. 2011;108 Suppl 2:10871-10877.

Rees DC, Williams TN, Gladwin MT.Sickle-cell disease.Lancet.

2010;376(9757):2018-2031.

Rehkopf DH, Gillespie DE, Harrell MI, Feagin JE. Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial mRNAs. Mol Biochem Parasitol. 2000; 105(1): 91–103.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489–494.

Reidenbach KR, Cook S, Bertone MA, Harbach RE, Wiegmann BM, Besansky NJ. Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. BMC Evol. Biol. 2009;9:298.

Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA.The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium* vivax and related parasites. Mol Phylogenet Evol. 2014;78:172–184.

Rich SM, Leendertz FH, Xu G, LeBreton M, Djoko CF, Aminake MN, et al. Proc Natl Acad Sci U S A. 2009;106(35):14902-149027.

Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D. Ancient DNA studies: new perspectives on old samples. Genet Sel Evol. 2012;44:21.

Rodrigues PT, Alves JMP, Santamaria AM, Calzada JE, Xayavong M, Parise M, et al. Using Mitochondrial Genome Sequences to Track the Origin of Imported *Plasmodium* vivax Infections Diagnosed in the United States. Am J Trop Med Hyg. 2014; 90(6): 1102–1108.

Rodrigues PT, Valdivia HO, de Oliveira TC, Alves JMP, Duarte AMRC, Cerutti-Junior C. Human migration and the spread of malaria parasites to the New World. Sci Rep. 2018;8(1):1993.

Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. Nature Protocols. 2007; 2:1756–1762.

RTS, S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. Lancet. 2015;386(9988):31–45.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-918.

Sainz-Elipe S, Latorre JM, Escosa R, Masià M, Fuentes MV, Mas-Coma S, et al. Malaria resurgence risk in southern Europe: climate assessment in an historically endemic area of rice fields at the Mediterranean shore of Spain. Malar J. 2010;9:221.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441-448.

Saving Lives, Buying Time: Economics of Malaria Drugs in an Age of Resistance. Institute of Medicine (US) Committee on the Economics of Antimalarial Drugs. Washington (DC): National Academies Press (US); 2004. pp. 126-128.

Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PLoS One. 2012;7(3):e34131.

Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nature reviews. Genetics. 2012;13(775):745-753.

Scheu A, Powell A, Bollongino R, Vigne JD, Tresset A, Canan C, et al. The genetic prehistory of domesticated cattle from their origin to the spread across Europe. BMC Genet. 2015;16:54.

Schliekelman P, Garner C, Slatkin M. Natural selection and resistance to HIV. Nature. 2001; 411(6837):545-546.

Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358(6363):652-655.

Schubert M, Ginolhac A, Lindgreen S, Thompson JF, AL-Rasheid KAS,Willerslev E, et al. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics. 2012;13:178.

Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, Mittnik A, et al. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. Proc Natl Acad Sci U S A. 2011;108(38):E746-52.

Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang CC, et al. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nat Commun. 2017;8:15694.

Schuenemann VJ, Kumar Lankapalli A, Barquera R, Nelson EA, Iraíz Hernández D, Acuña Alonzo V, et al. Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. PLoS Negl Trop Dis. 2018;12(6):e0006447.

Schuenemann VJ, Avanzi C, Krause-Kyora B5, Seitz A, Herbig A, Inskip S, et al. Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. PLoS Pathog. 2018;14(5):e1006997.

Schweitzer MH, Suo Z, Avci R, Asara JM, Allen MA, Arce FT, et al.Analyses of soft tissue from Tyrannosaurus rex suggest the presence of protein. Science. 2007;316(5822):277-280.

Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, et al. Biomolecular characterization and protein sequences of the Campanian hadrosaur B. canadensis. Science. 2009;324(5927):626-631.

Schraiber JG, Evans SN, Slatkin M. Bayesian Inference of Natural Selection from Allele Frequency Time Series. Genetics. 2016;203(1):493-511.

Schraiber JG. Assessing the Relationship of Ancient and Modern Populations. Genetics. 2018;208(1):383–398.

Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, Prado JL. et al. Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. Plos One. 2013;8(10):e78575.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346(6213):1113-1138.

Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19:227-240.

Shapiro B, Hofreiter M. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. Science. 2014;343(6169):1236573.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539

Singh B, Kim Sung L, Matusop A, Radhakrishnan A, Shamsul SS, Cox-Singh J, et al. A large focus of naturally acquired *Plasmodium* knowlesi infections in human beings. Lancet. 2004;363(9414):1017-1024.

Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, Coetzee M, et al. A global map of dominant malaria vectors. Parasit Vectors. 2012;5:69.

Siwal N, Singh US, Dash M, Kar S, Rani S, Rawal C, et al. Malaria diagnosis by PCR revealed differential distribution of mono and mixed species infections by *Plasmodium* falciparum and P.

vivax in India. PLoS One. 2018;13(3):e0193046.

Sjödin P, Skoglund P, Jakobsson M. Assessing the maximum contribution from ancient populations. Mol Biol Evol. 2014;31(5):1248-1260.

Skoglund P,Northoff BH, Shunkov MV, Derevianko A, PääboS, Krause J, Jakobsson M. Separating ancient DNA from modern contamination in a Siberian Neandertal, Proceedings of the National Academy of Sciences USA. 2014;111(6):2229-2234.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015;525(7567):104-108.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. Nature. 2016;538(7626):510-513.

Skoglund P, Reich D. A genomic view of the peopling of the Americas. Curr Opin Genet Dev. 2016;41:27-35.

Slon V, Mafessoni F, Vernot B, Cesare de Filippo, Grote S, Viola B, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature. 2018.

Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Beltrán de Heredia J, et al. Historical Y. pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. Cell Host Microbe. 2016;19(6):874-881.

Spyrou MA, Tukhbatova RI, Wang CC, Valtueña AA, Lankapalli AK, Kondrashin VV, et al. Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. Nat Commun. 2018;9(1):2234.

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, et al. Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J

Hum Genet. 1998;62(6):1507-1515.

Stoneking M, Krause J.Learning about human population history from ancient and modern genomes. Nat Rev Genet. 2011;12(9):603-614.

Sturm A, Amino R, van de Sand C, Regen T, Retzlaff S, Rennenberg A, et al. Manipulation of Host Hepatocytes by the Malaria Parasite for Delivery into Liver Sinusoids. Science. 2006;313(5791):1287-1290.

Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.

Suwanarusk R, Cooke BM, Dondorp AM, Silamut K, Sattabongkot J, White NJ, et al. The deformability of red blood cells parasitized by *Plasmodium* falciparum and P. vivax. J Infect Dis. 2004;189(2):190–194.

Sutherland CJ, Tanomsing N, Nolder D, Oguike M, Jennison C, Pukrittayakamee S, et al. Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium* ovale occur globally. J Infect Dis. 2010;201(10):1544-1550.

Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, et al. *Plasmodium* cynomolgi genome sequences provide insight into *Plasmodium* vivax and the monkey malaria clade. Nat Genet. 2012;44(9):1051-1055.

Tajebe A, Magoma G, Aemero M, Kimani F. Detection of mixed infection level of *Plasmodium* falciparum and *Plasmodium* vivax by SYBR Green I-based real-time PCR in North Gondar, north-west Ethiopia. Malar J. 2014;13:411.

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007;24(8):1596-1599.

Tanabe K, Mita T, Jombart T, Eriksson A, Horibe S, Palacpac N, et al. *Plasmodium* falciparum accompanied the human expansion out of Africa. Curr Biol. 2010;20(14):1283-1289.

Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RL, Fairhurst RM, et al. The evolutionary history of *Plasmodium* vivax as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. Mol Biol Evol. 2013;30(9):2050-2064.

Titra E, Anstey NM, Sugiarto P, et al. Multidrug-resistant *Plasmodium* vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia. PLoS Med 2008;5(6):e128

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(12):2975-2987

Valdiosera C, Günther T, Vera-Rodríguez JC, Ureña I, Iriarte 6, Rodríguez-Varela R, et al. Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia. Proc Natl Acad Sci U S A. 2018;115(13):3428-3433.

Venkatesan M, Amaratunga C, Campino S, Auburn S, Koch O, Lim P, et al. Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium* falciparum-infected whole blood samples. Malar J. 2012;11:41.

Vulliamy TJ, Othman A, Town M, Nathwani A, Falusi AG., Mason, et al. Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variants A and A-. Proc Nat Acad Sci U S A. 1991;88:8568-8571.

Wadsworth C, Buckley M.Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. Rapid Commun Mass Spectrom. 2014;28(8):605–615.

Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, et al. High-Throughput DNA sequencing of ancient wood. Mol Ecol. 2018;27(5):1138-1154.

Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA1, Madrigal JR, Orlando L, Gilbert MTP. New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. Biotechniques. 2015;59(6):368-371.

Wall JD, Kim SK. Inconsistencies in Neanderthal Genomic DNA Sequences. PLoS Genet. 2007;3(10):e175.

Wells TNC, Burrows JN, Baird JK: Targeting the hypnozoite reservoir of *Plasmodium* vivax: the hidden obstacle to malaria elimination. Trends Parasitol. 2010;26:145-151.

White NJ. Determinants of relapse periodicity in *Plasmodium* vivax malaria. Malar J. 2011;10:297

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A. 2014;111(13):4832-4837.

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MT, Shapiro B, et al. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science. 2003;300(5620):791-795.

Willerslev E, Cappellini C, Boomsma W, Nielsen R, Hebsgaard MB,1 Brand TB, et al. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science. 2007;317(5834):111–114.

Williams TN. Human red blood cell polymorphisms and malaria. Curr Opin Microbiol. 2006;9(4):388-394.

Williams TJ, Collins MB, Rodrigues K, Rink WJ, Velchoff N, Keen-Zebert A, et al. Evidence of an early projectile point technology in North America at the Gault Site, Texas, USA. Sci Adv. 2018;4(7):eaar5954.

Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279–283.

Woolhouse ME, Gowtage-Sequeria S. Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005;11(12):1842-1847.

World Health organization (WHO). World malaria report 2017. Geneva. 2017. Available from: http://www.who.int/malaria/publications/world-malaria-report-2017/en/.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 2008;455(7213):661-664.

Wright S. The genetical structure of populations. Ann Eugen. 1951;15(4):323–35.

Xiang H, Gao J, Yu B, Zhou H, Cai D, Zhang Y, et al. Early Holocene chicken domestication in northern China. Proc Natl Acad Sci U S A. 2014;111(49):17564-17569.

Yalcindag E, Elguero E, Arnathau C, Durand P, Akiana J, Anderson TJ, et al. Multiple independent introductions of *Plasmodium* falciparum in South America. Proc Natl Acad Sci U S A. 2012; 109(2):511-516.

Yam XY, Brugat T, Siau A, Lawton J, Wong DS, Farah A, et al. Characterization of the *Plasmodium* Interspersed Repeats (PIR) proteins of *Plasmodium* chabaudi indicates functional diversity. Sci Rep. 2016;21:23449.

Yang DY, Eng B, Waye JS, Dudar JC, Saunders SR. Technical note: improved DNA extraction from ancient bones using silica-based spin columns. Am J Phys Anthropol. 1998;105(4):539-543.

Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, et l. 40,000-Year-Old Individual from Asia Provides Insight into

Early Population Structure in Eurasia. Curr Biol. 2017;27(20):3202-3208.

Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. Dietary adaptation of FADS genes in Europe varied across time and geography. Nat Ecol Evol. 2017;1:167.

Zeder MA. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci U S A. 2008;105(33):11597-11604.

Zeder MA. The Origins of Agriculture in the Near East. Curr Anthropol. 2011;52(4):221-235.

Zhao X, Smith DL, Tatem. Exploring the spatiotemporal drivers of malaria elimination in Europe. Malar J. 2016;15:122.

Zhou B, Wen S, Wang L, Jin L, Li H, Zhang H. AntCaller: an accurate variant caller incorporating ancient DNA damage. Molecular Genetics and Genomics. 2017;292(6):1419–1430

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998;391(6667):594-597.

Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, Nerlich AG. Characterization of Mycobacterium tuberculosis complex DNAs from Egyptian mummies by spoligotyping. J Clin Microbiol. 2003;41(1):359–367.