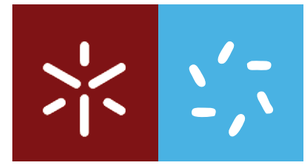
Christina Casal Ribeiro Branco

mtDNA and the evolutionary history of the Iberian Peninsula

UMinho | 2018



Universidade do Minho
Escola de Ciências

Christina Casal Ribeiro Branco

mtDNA and the evolutionary history of the Iberian Peninsula

Janeiro de 2018

Universidade do Minho
Escola de Ciências

Christina Casal Ribeiro Branco

mtDNA and the evolutionary history of the Iberian Peninsula

Tese de Mestrado
Mestrado em Bioquímica Aplicada
Área de Especialização em Biomedicina

Trabalho efetuado sob a orientação do
Professor Doutor Pedro Alexandre Dias Soares

Janeiro de 2018

DECLARAÇÃO

Nome: Christina Casal Ribeiro Branco
Endereço eletrónico: christinabranco@gmail.com
Telefone: +351 914978667
Bilhete de Identidade/Cartão de Cidadão: 13802108

Título da dissertação: mtDNA and the evolutionary history of the Iberian peninsula
Orientador: Doutor Pedro Alexandre Dias Soares

Ano de conclusão: 2018

Mestrado em Bioquímica Aplicada, Área de Especialização em Biomedicina

DE ACORDO COM A LEGISLAÇÃO EM VIGOR, NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA TESE.

Universidade do Minho, 31 de Janeiro de 2018

Assinatura:

We are survival machines – robot vehicles blindly programmed to preserve the selfish molecules known as genes. This is a truth which still fills me with astonishment.

Richard Dawkins, *The Selfish Gene*

Then the Lord God said, "Behold, the man has become like one of us in knowing good and evil. Now, lest he reach out his hand and take also of the tree of life and eat, and live forever—therefore the Lord God sent him out from the garden of Eden to work the ground from which he was taken.

Genesis 3, 22-23

Acknowledgements

One of the many good things about finishing a thesis is that you are forced to think about all the people behind your work and wonder where you would be without them – both academically and personally.

First off, I would like to thank Dr Pedro Soares for accepting to mentor me during this journey. I could not have wished for a kinder, more patient, more hardworking, more knowledgeable advisor.

Dan: how coincidental is it that the first person I meet at the beginning of my academic journey is also the person who I could not have done without when finishing it? Whoever is in charge doesn't play dice.

Marina, despite the channel between us: thanks for the initial tutoring!

João Guimarães. Pedro Fernandes and the most recent addition, Mafalda: working alone is boring. Thank you for being in this with me.

Dr Isabel João and Dr Tony Collins: thank you for putting up with me at yours and Pedro's office. I know I'm very distracting.

Joana, Mafalda and Pedro: thank you for making my first year in Braga much friendlier.

Alexandra, Eduardo, Helena, Margarida, Pedro, Sara, Vitor: I could not have done without our lunch/coffee/cigarette break group. Thanks for the scientific discussion, the not-so-scientific discussion, the venting, the ranting, the troubleshooting, the believing in me and my work even when I myself didn't.

To João, my academic "godson": thanks for the people-watching, gossip, movie and book browsing sessions.

Rita, Carla, and more recently, Ru: I gladly got in a train at the end of an exhausting day and rode an hour-something to have tapioca with you, several times. I wouldn't have done it if you weren't excellent company.

Aida, Diana, and fellow Pandoras: sometimes the only thing I had to look forward to all month was our little bookclub. Thanks for it. Thanks for making me read books I'd never have known otherwise. *Juntas fazemos acontecer*.

Dr Manuela Brito and Dr Manuel Esteves: I could not have done this without you. Thank you for taking care of the major work tool of this thesis.

To my grandmother: thanks for always being excited to see me, even if we live in the same house. The past few years with you were a blessing.

To my late grandfather: thanks for instilling my love of geography and maps. It certainly came in handy.

To my uncle/godfather: thank you for understanding exactly why my line of work is important, why we need to look at the past to understand the present.

To my father, thank you for imparting me from a young age with the knowledge that we're all brothers and sisters from the Near East.

Alcino: thanks for, well, everything, really. Love, companionship, friendship, laughter, and all that good stuff. *Es lo que hay*.

Mummy, thank you for being the strongest person I know. Maybe this work is proof some of it rubbed off on me.

This thesis is dedicated to Adaltiva, Edite, Ana, Rosa, Ana Francisca, Maria Rosa, Teresa... and all the women in the unbreakable chain that connects Christina in Portugal to Eve in Africa.

This work was performed on the scope of the FCT project PTDC/EPH-ARQ/4164/2014 partially funded by FEDER funds (COMPETE 2020 project 016899).

mtDNA and the evolutionary history of the Iberian Peninsula

Abstract

Mitochondrial DNA (mtDNA) is widely used in population studies due to its high copy number, lack of recombination, higher mutation rate and maternal inheritance. In archaeogenetics, "the application of molecular genetics to study the human past", it assumes utmost importance, and is the most used genetic marker.

Archaeogenetics and by consequence, mtDNA, has been used to unravel the genetic makeup of various ancient migrations. Founder analysis, an underused method in which potential founder sequence types in a source population are chosen and lineages clusters deriving from them in the settlement zone of interest are dated, was used in this work. It has never been done before with such a large amount of whole mtDNA sequences.

In this study, we focused on the prehistoric genetic heritage of Iberia and by consequence, Europe. This prehistory can be defined in broad terms into a first colonization circa 45000 years ago, a recolonization from Southern refugia after the Late Glacial Maximum (LGM), the postglacial re-colonization by Mesolithic groups at the end of the Younger Dryas, dispersals of Near Easterners with the Neolithic and lastly, exchanges from the Copper Age onwards, including a probable major expansion from the Steppes in the Bronze Age. There is a longstanding debate as to how much of each of these events impacted the genetic diversity of Europe. To assess the impact of these expansions in Iberia, Europe and its fringes (British Isles and Sardinia), 17542 published and unpublished whole mtDNA samples were incorporated into 7 different models of founder analysis, with different source and sink populations. After an initial scan, specific event migration models were created for each test, attending not only to the peaks visualized but also to archaeological, palaeoclimatic and palaeontological evidence.

Our work reveals that much of European and Iberian mitochondrial diversity is in fact fruit of post-LGM expansions, not from Southern European refugia as previously suggested, but from the Near East. This includes haplogroup H, the major lineage in Europe and formerly thought to have entered in Europe pre-LGM. Iberia is not as much of what we call a "fringe" of Europe as with Sardinia and the British Isles , and forms an almost continuum with the Mediterranean. The next largest component to the Iberian maternal variance is the Neolithic, where lineages that entered Europe from the Near East in the Mesolithic spread through the Mediterranean with farming.

Bronze age expansions are most important for the British Isles, however in Iberia they also represent a large partition of diversity.

mtDNA e a história evolutiva da Península Ibérica

Resumo

O ADN mitocondrial (mtDNA) é amplamente utilizado em estudos populacionais devido ao seu alto número de cópias, não-recombinação, maior taxa de mutação e herança exclusivamente materna. Na arqueogenética, "a aplicação da genética molecular para estudar o passado humano", este assume a maior importância e é o marcador genético mais utilizado. A arqueogenética e, por consequência, o mtDNA, tem sido usada para desvendar a composição genética de várias migrações antigas. A análise do fundador, um método subutilizado em que são escolhidos os possíveis tipos de sequencias de fundadores numa população fonte, e os *clusters* de linhagens que deles derivam na zona de recepção de interesse são datados, foi usado neste trabalho. Nunca antes foi feito com uma quantidade tão grande de sequências completas de mtDNA.

Neste estudo, concentramo-nos na herança genética pre-histórica da Ibéria e, consequentemente, da Europa. Esta pré-história pode ser definida em termos gerais numa primeira colonização há cerca de 45000 anos atrás, uma recolonização do refúgios do Sul após o Último Máximo Glacial (LGM), a recolonização pós-glacial por grupos mesolíticos no final do Dryas recente, dispersões de populações do Próximo Oriente com o Neolítico, e por fim, os intercâmbios da Idade do Cobre, incluindo uma provável expansão das estepes na Idade do Bronze. Há um longo debate sobre quanto impacto cada um desses eventos teve sobre diversidade genética da Europa. Para avaliar o impacto destas expansões na Ibéria, na Europa e nas suas franjas (Ilhas Britânicas e Sardenha), 17542 amostras de mtDNA inteiras, publicadas e não publicadas, foram incorporadas em 7 modelos diferentes de análise de fundador, com diferentes populações-fontes e população-receptora. Após um *scan* inicial, foram criados modelos de migração de eventos específicos para cada teste, atendendo não apenas aos picos visualizados, mas também a evidências arqueológicas, paleoclimáticas e paleontológicas.

O nosso trabalho revela que grande parte da diversidade mitocondrial europeia e ibérica é de facto fruto das expansões pós-LGM; não dos refúgios do sul da Europa, como sugerido anteriormente, mas do Oriente Próximo. Isto inclui o haplogrupo H, a linhagem principal na Europa e anteriormente assumido como tendo entrado na Europa pré-LGM. A Ibéria não é tanto o que chamamos "franja" da Europa quanto a Sardenha e as Ilhas Britânicas, e forma um quase

contínuo com o Mediterrâneo. O próximo componente mais importante da variância materna ibérica é o Neolítico, onde as linhagens que entraram na Europa do Próximo Oriente no Mesolítico se espalharam pelo Mediterrâneo com a agricultura. As expansões da idade de bronze são mais importantes para as Ilhas Britânicas, no entanto, na Península Ibérica, elas também representam uma grande parte da diversidade.

Table of contents

List of figures

List of abbreviations

aDNA – ancient DNA

AMH – Anatomically Modern humans

ANE – Ancient Northern Europeans

BBC – Bell Beaker Culture

bp – base pair

CRS – Cambridge Reference Sequence

CWC – Corded Ware Culture

D-loop – displacement loop

FA – Founder analysis

hg - haplogroup

HVS – hypervariable segment

indel – insertions and deletions

kb – Thousand base pair

kya – thousand years ago

LBK – *Linearbandkeramik*

LGM – Last Glacial Maximum

ML – Maximum Likelihood

MP – Maximum Parsimony

mtDNA – mitochondrial DNA

Mya – Million years ago

NCBI – National Center for Biotechnology Information

NJ – Neighbour joining

PCR – Polymerase chain reaction

rCRS- revised Cambridge Reference Sequence

RM – Reduced median

rRNA – ribossomal RNA

RSRS – Reconstructed Sapiens Reference Sequence

SNP – Single nucleotide polymorphism

SSA – Sub-Saharan Africa

sub/nucl/year – substitution per nucleotide per year

tRNA – transfer RNA

UPGMA – Unweighted Pair Group Method With Arithmetic Mean

XML- eXtended Markup Language

Y-DNA – Y chromosome DNA

Chapter 1: Introduction

1.1.MtDNA

1.1.1Mitochondria

Mitochondria are organelles commonly known as the "powerhouse of the cell"[1], due to their role in the production of metabolic energy in the form of ATP from sugars and fatty acids – a process known as oxidative phosphorylation. However, they also play an important part in synthetizing steroid and lipids[2], apoptosis, and storage of calcium ions[3]. They are present in virtually all eukaryotic organisms, but notably absent in mature mammalian red blood cells. Mitochondria are implied in various diseases, such as myopathies[4], neuropathies[5] and metabolic disorders[6].

Structure-wise, they are commonly depicted as solitary and rod shaped despite being quite flexible and even fusing and dividing from one another, creating large interconnected networks [7]. They are relatively large (between 0.5 and 2μm in diameter) and consist of a double membrane that separates four subcompartments: the outer membrane, intermembrane space, inner membrane, and the matrix. This inner membrane folds into structures known as cristae where oxidative phosphorylation takes place in mega Dalton sized *F0F1* complexes.

The most widely accepted theory of mitochondrial origin is the endosymbiotic theory. It proposes that mitochondria and other organelles such as chloroplasts originate from energetically efficient bacteria engulfed by a host cell. As proof, these organelles contain their own genome which is not bound by histones and has its own genetic code, are formed by a double membrane, and show independent growth and division from the rest of the cell. In the case of mitochondria, the engulfed bacteria in question seem to be endocellular parasites known as α-proteobacteria[8]. There is some

debate over the nucleus being incorporated first into a protoeukaryote (in the Archezoa hypothesis[9]) or the nucleus and the mitochondria being formed as part of the same evolutionary event (the hydrogen hypothesis[10]).

1.1.2 The mitochondrial genome

Most mitochondrial proteins are coded by genes from the nucleus. However, mitochondria have their own genome, hereby referred to as mitochondrial DNA (mtDNA). It is circular, not wound around histones and has its own genetic code with slight differences in relation to the universal genetic code[11]: UGA codes tryptophan and not the stop codon, AUG codes methionine and not isoleucine, and AGA and AGG codes the stop codon and not arginine.

Human mtDNA was the first mitochondrial genome to be sequenced, in 1981[12]. It is formed by 16569bp and distributed in two strands: one "heavy", rich in guanine, and one "light", rich in cytosine and in relation to which the sequence is numbered[13]. It contains 37 genes: 22 code for tRNA, 2 for rRNA and 13 for proteins involved in the respiratory chain (Figure 1)

All 16569 nucleotides are numbered according to the revised Cambridge Reference Sequence (rCRS)[14], an updated version of the original CRS[12]. There are some criticisms in relation to it, as it represents a derived rather than an ancestral state. An alternative notation – Reconstructed Sapiens Reference Sequence, or RSRS – has been proposed[15], however, a replacement of rCRS in the scientific community is unlikely to occur[16] completely considering the historical use of the rCRS in databases, software tools and publications.

The mtDNA sequence is extremely compact – there are very few non-coding bases between genes[17]. The exception is the control region, a 1.1kb zone between bp 16024 and 576 that contains transcription and regulation genes and the origin of replication of the heavy strand (O_H). Most of the control region is comprised of the so-called D-loop (for "displacement loop"). Although a common error, these terms are not interchangeable[18]. This D-loop is a third linear strand formed during replication. The exact mechanism of mitochondrial replication is unknown[17].

In the control region, the mutation rate is up to ten times higher than in the coding region[19], possibly due to selective pressure acting over the latter. However, that does not fully explain the pattern, as protein-coding genes' synonymous mutations under little selection have lower rates than control region mutations. Of particular interest in the control region are the two hypervariable

sections, or regions, HVS-I and HVS-II, fundamental for studying human population genetics. Some sources, mainly in forensic science, consider an additional hypervariable region HVS-III[20].

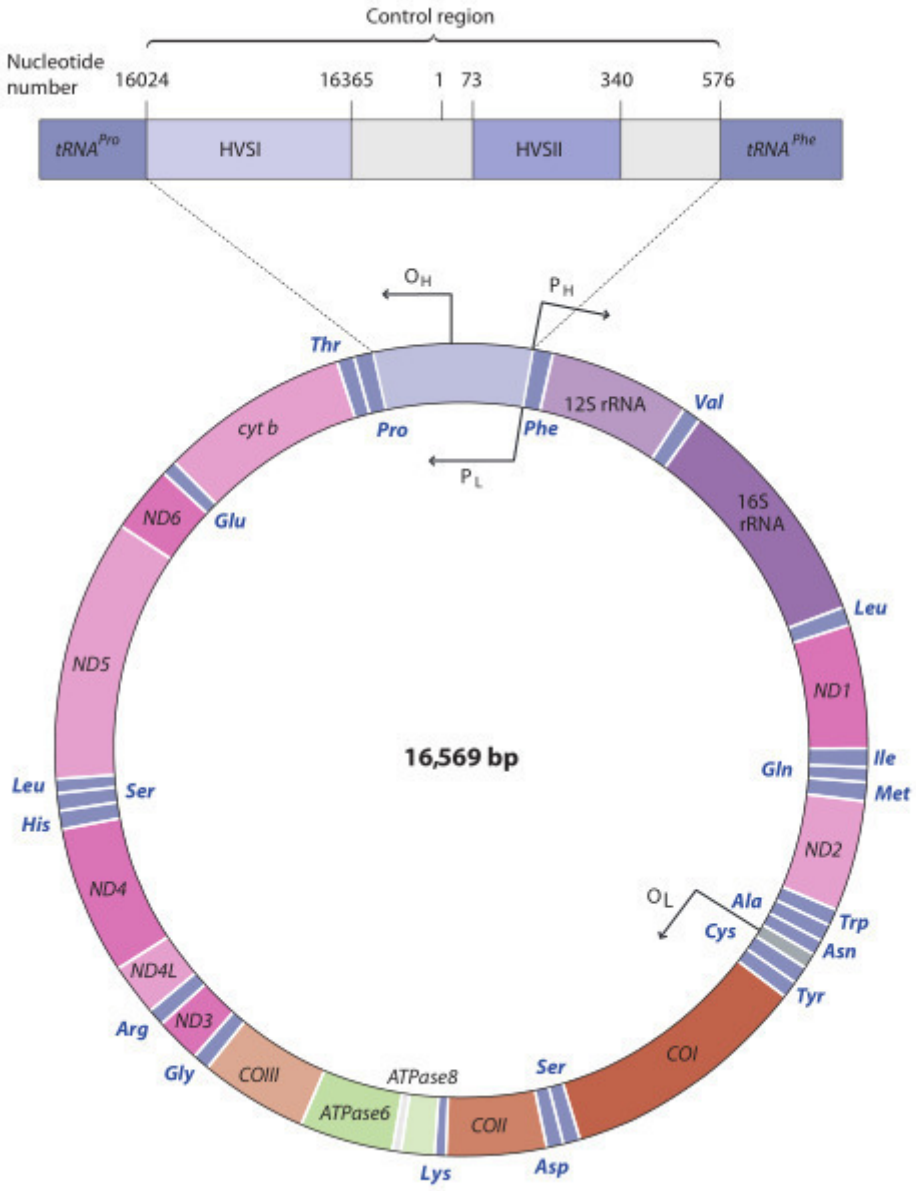


Figure 1 - A representation of the human mitochondrial genome, with emphasis on the control region and its two hypervariable regions, HVS-Iand HVS-II[1]

1.1.3 Why mtDNA?

There are several reasons why mtDNA is used in population studies: high copy number, maternal inheritance, lack of recombination and higher mutation rate.

High copy number

While most cells only contain one copy of DNA per nucleus, on average a cell contains several hundreds[22] to thousands [23]of copies of mtDNA. This is particularly important in ancient DNA studies, as this type of DNA is typically degraded and in low concentration [24]. It also plays a role in mitochondrial disease – Ye et al. demonstrated that up to 90% of the population carries a heteroplasmy (coexisting mtDNA variants in a single cell), and 20% one implicated in disease [23]. However, there is a dramatic variability in the phenotypic presentation of mitochondrial mutation – there seems to be a threshold of the number of mutant mtDNA over which the disease manifests itself[25]. Furthermore, during oogenesis, there is a genetic bottleneck which means that the degree of heteroplasmy, and thus phenotype, can vary wildly between each child.(Figure 2)

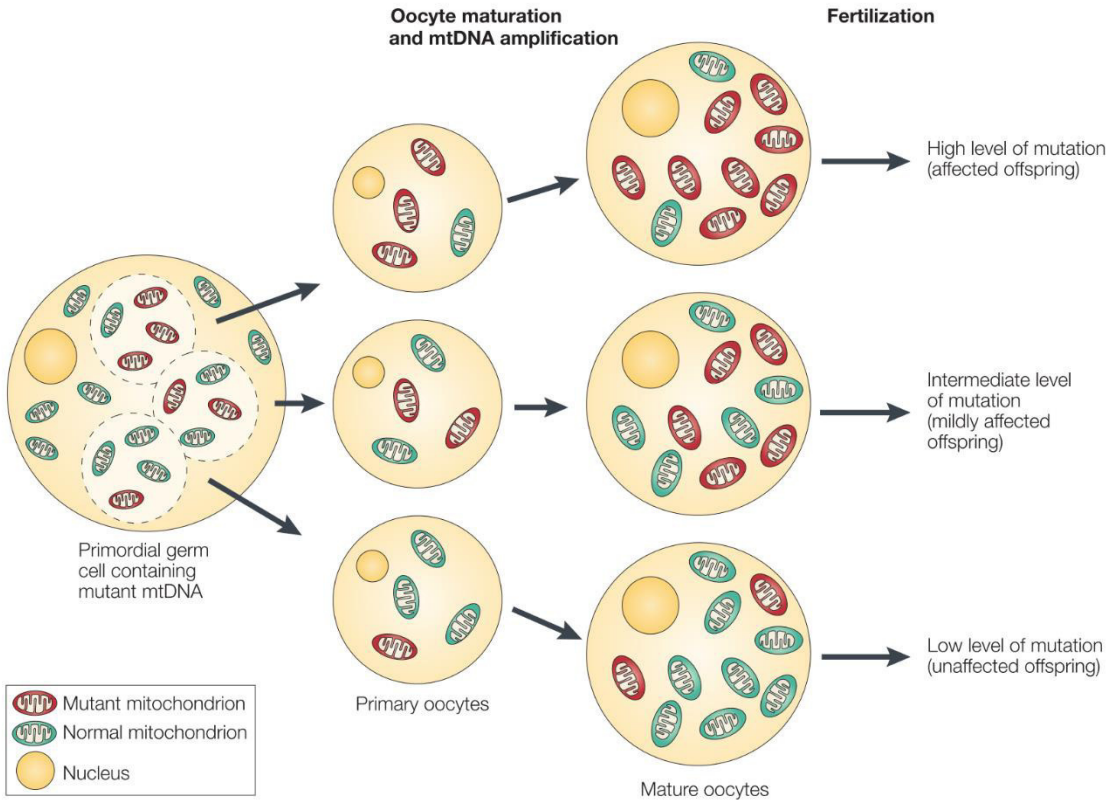


Figure 2 -The mitochondrial genetic bottleneck [26]

Maternal inheritance

Despite being put in question several times, through linkage disequilibrium studies [27] and a single documented case of paternal inheritance [28], maternal inheritance of mtDNA is still given as certain.

This is assured not only by the mitochondria in sperm being outnumbered by those in ova[29] but also by a mechanism that destroys paternal mitochondria after ubiquitin tagging [30].

Haploid marker systems are useful for studying sex-biased events. mtDNA is passed on only from the mother, and thus only tells the story of the female line; likewise for the Y chromosome and the male line. Males or females may contribute differently to the genetic makeup of a population – for example, European traders and missionaries tended to be male, and thus more European lineages in the sink population will be detected, but not in mtDNA, as is the case in Brazil[31,32] and Polynesia[33].

Lack of recombination

Despite mitochondria containing all the necessary machinery for recombination[34] and the report o at least one case of paternal inheritance [28], mitochondrial recombination is considered to be in inexistent in human evolution studies[21], as it was never detected.

Higher mutation rate

The mtDNA mutation rate seems to be 10 times higher than that of nuclear DNA, which allows for diversity to accumulate faster in the timeframe of human evolution. There are many explanations, such as a higher exposure to reactive oxygen species from the mitochondrion's role in oxidative phosphorylation, a higher turnover rate (i.e. more replications) and lack of histones.

1.2 Reconstructing the past

1.2.1 Phylogeography

Phylogeography has been defined as "the relationship between phylogenetics and geography"[35], or that is, the fusion between concepts typical of systematic biology such as clades with regional specificity to "see genes through time and space"[36]. To do this, elements such as phylogenetic trees, geographic distribution, molecular clocks and additional evidence from disciplines like linguistics, palaeoclimatics, archaeology and demographic history are employed.

1.2.1.1 Phylogenetic trees

Phylogenetic trees are structures that represent the evolutionary relationship between entities, that can be molecular sequences or organisms (taxa). They consist of branches that connect through nodes. These nodes represent theoretical or real common ancestors. Trees can be rooted or unrooted: rooted trees have a common ancestor. This root defines the directionality of the tree.

There are two types of methods that can used for building phylogenetic trees: distance and characters. For distance-based methodologies, matrices are created by measuring the similarity between taxa, and examples of such methods include UPGMA (unweighted pair-group method with arithmetic mean) and neighbour-joining (NJ). Despite its slower computational speed, character-based methodologies are more suitable for mtDNA phylogeography, as they rely on evolutionary changes as the common substitutions that create SNPs, and it better reflects a real evolutionary process.

The most used character based methodologies are Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian methods.

Maximum parsimony (MP)

In this method, the length of each branch is the number of evolutionary changes (polymorphisms) along it. These can often be weighted by their relative weight of mutation, for example, if certain sites are known to be hypermutable. The best tree is the one with the least evolutionary changes, that is, the shortest. It requires lower computational power as it only considers polymorphic sites where at least two alleles are present, but is prone to a statistical error called long branched attraction, in which two independent lineages with fast evolving characters are presented erroneously associated in the tree.

Maximum likelihood (ML)

Maximum likelihood generates all combinations of unknown parameters, such as tree topology and branch lengths, and then estimates the tree most likely to represent the data according to a predefined evolutionary model. This is a computationally intensive methodology.

Bayesian inference

With the availability of increased computing power, Bayesian inference is now widely used in phylogenetics. Rather than present a single tree, it presents a set of them and their probabilities, after repeated Markov Chain Monte Carlo simulations. While ML approaches determine the probability of data (such as sequence alignment) given a tree, Bayesian approaches determine the probability of a tree given the data and a prior probability over the possible trees.

1.2.1.2. Median networks

In some cases, such as when recombination, or parallel gene flow occurs[37], a single phylogenetic tree is not an adequate representation of data, as loops, called reticulations are generated. The different parsimonious evolutionary routes are thus represented in structures called networks. Due to its common parallel or recurring mutations (homoplasy), it is also useful in mtDNA. The most commonly used type of network in mtDNA are called median networks[38].

In these, variant sites are converted to binary characters. Then, samples are linked by their distance to one another. Reticulations sometimes produce large hyperdimensional cubes when the

number of taxa is large(Figure 3), and as such, a set of rules has been devised to eliminate some of the less likely links.

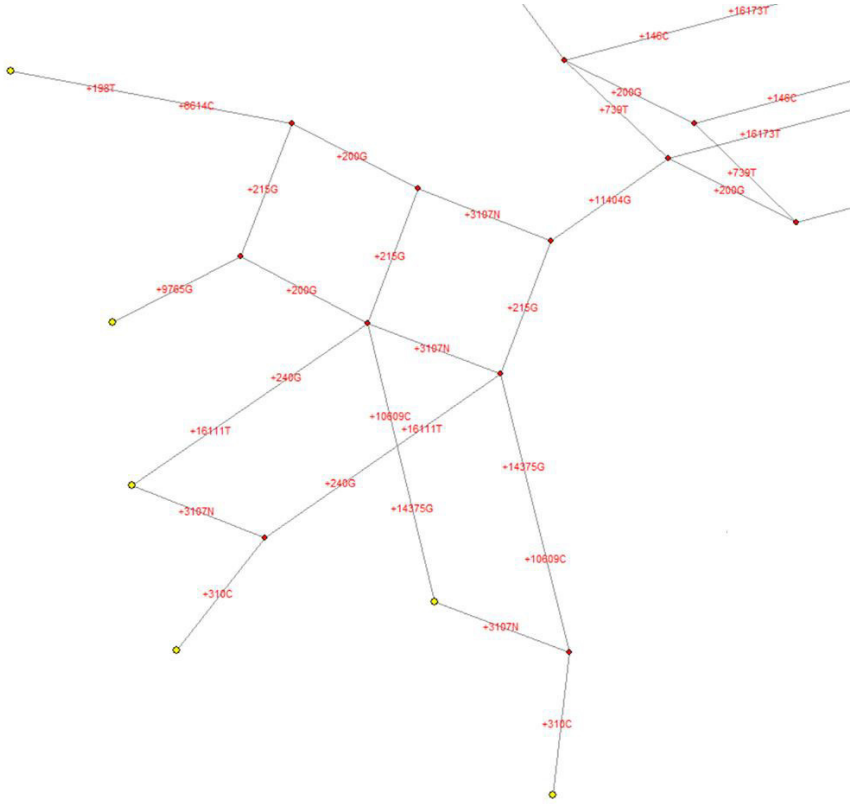


Figure 3 – A network diagram rendered with Network[68] and Network Publisher software (www.fluxus-engineering.com). The yellow dots represent individuals and the red dots represent reticulations.

1.2.1.3 Founder analysis

Founder analysis is a method for analysis aiming at the identification, scaling and dating of migrations into a new territory It was first proposed by Richards et al. in 2000[39] for studying European founder lineages in the Near East and has since been used for the colonization of Sardinia[40] and that of of Southeast Asia[41,42], African migrations[43], determining maternal lineages of Ashkenazis[44], and colonization from Glacial refugia[45]. The method picks out founder sequence types in potential source populations (the "source") and dates lineage clusters deriving from them in the settlement zone of interest (the "sink").

First, after displaying data in phylogenetic trees and/or networks, haplotypes either existing in the data or inferred in common between the hypothetical source and sink populations are identified.

Finally, the genetic diversity of the founder clade in the sink population is used to estimate the time of the migratory events associated with the founder.

The recurrence of common mutations in mtDNA and reverse gene flow (from the sink to the source populations) may hinder the signal of the founder analysis. Therefore, several different criteria were created to select what is and what is not considered a founder group.

The *f0* criterion is the least restrictive, as it considers as founders all the haplotypes present in both source and sink. In *f1* and *f2* criteria, for a clade to be considered a founder, it should not have matches at the tip of the phylogeny: either one or two further derived branches (containing variation within the founder clade), respectively, should be present in the source population (Figure 4).

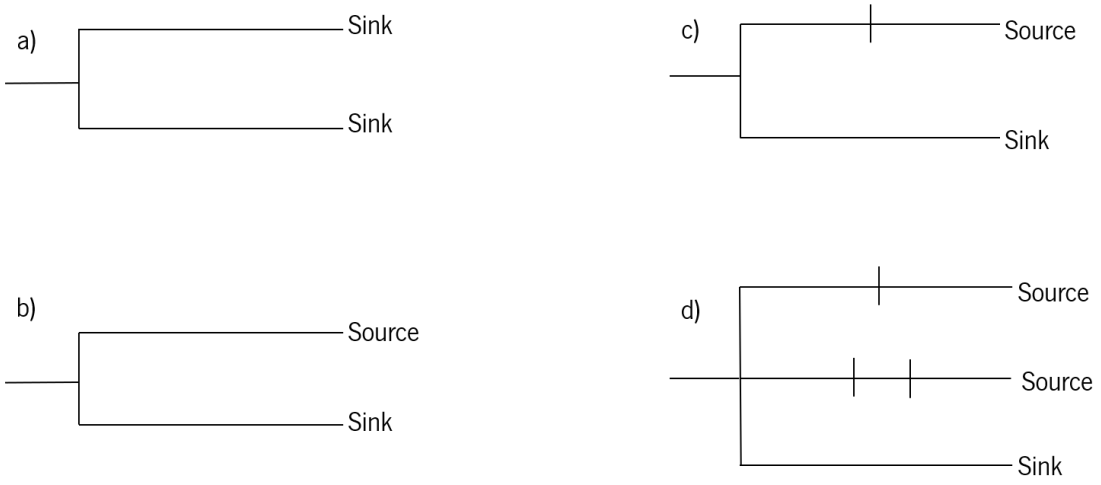


Figure 4 – A schematic representation of the application of founder criteria. Variants are represented by a vertical line. Clade a is not accepted as a founder in any criterion; clade b is considered a founder in f0 but not f1 or f2 criteria; clade c is considered a founder in f0 and f1, but not in f2; clade d is considered a founder in f0, f1 and f2 criteria.

1.2.1.4 Molecular clock

Molecular clocks are functions that allow us to estimate dates of divergence of branches in a phylogenetic tree, assuming a constant mutation rate. To calibrate this molecular clock it is necessary to date these lineages resorting to palaeontology, for example. The clocks can be used not only for distance between species but also for nonrecombining regions of the genome of a species, for example, mtDNA.

In the past, it had been argued that climate adaptation was a major driving force behind the mutation rate[46], however, further studies disproved this, stating that the high number of non-synonymous polymorphisms in younger branches of the mtDNA tree is spread worldwide, due to purifying selection[47][48].

Until 2009, several mutation rates were proposed, covering either the coding or non coding regions but never whole mtDNA. Forster et al (1996) presented a HVS-I only mutation rate of 1.80×10^{-7} substitutions per nucleotide per year (sub/nucl/year)[49] and Mishmar et al (2003) 1.26×10-8 sub/nucl/year for the coding region[46]. Not only did these rates not take into account important variation along the mtDNA genome, they also assumed that the mutation rate was linear through time, failing to consider the role of purifying selection. Soares et al (2009) devised a time-dependent clock with both coding and non-coding regions. The mutation rate now currently accepted is thus 1.665×10^{-8} sub/nucl/year, or one mutation every 3624 years[50], however this rate is further mathematically corrected for the effect of purifying selection.

One process for mutation rate calculation is by analysing familial pedigrees. The mutation rates tend to be 10× faster in this method.[51][52] Another approach is by calibrating the molecular clock with the use of ancient genomes[53]

1.2.1.5. Phylogeography of mtDNA

For descriptive purposes, mtDNA variation is assigned to different haplogroups, or monophyletic clusters of haplotypes. The first haplotypes, christened A, B, C and D were described in Native Americans[54]. From then on, the discovered haplogroups were given subsequent letters of the alphabet. Haplogroups are designated with capital letters and their derived subclades are named by intercalating lower-case letters with numbers (e.g.: U5b1c is a subclade of U5b1, which is in turn a subclade of U5b). Note that sometimes a haplogroup designated by a letter is in fact an offshoot of another haplogroup, e.g. haplogroup K is a subclade of U8b. When a lineage contains polymorphisms that don't belong to any described subclade, it is represented with an asterisk next to it (e.g. "H*", pronounced H-star); when a lineage is missing some key polymorphisms that would make it a derived subclade but contains others from that same clade, the pre- prefix is added (e.g. pre-U7). Monophyletic clades that are composed of two or more previously named haplogroups

are labeled by joining their names (e.g. “HV”) and separating them by apostrophe if the name includes a number (e.g., U2’3’4’7’8’9)[15].

1.2.1.6 Phylogeny of mtDNA in Europe

It is to be assumed that all mtDNA in the human gene pool descend from one common matrilineal ancestor who lived 200,000 years ago in Sub-Saharan Africa – the so-called “Mitochondrial Eve”[55]. However, despite its importance as the cradle of humanity and the main location of modern humans for most of their existence, the initial population dynamics and dispersal routes remain poorly understood. The human mtDNA phylogeny can be divided into two daughter branches, L0 and L1-6. Non-African mtDNA lineages all diverge from clades M and N, themselves descended from L3 – supporting a single exit model for all non-Africans[56].Of these, all modern European lineages derive from N, and most from R, its subclade that diverged shortly after its own birth[50] (Figure 5).

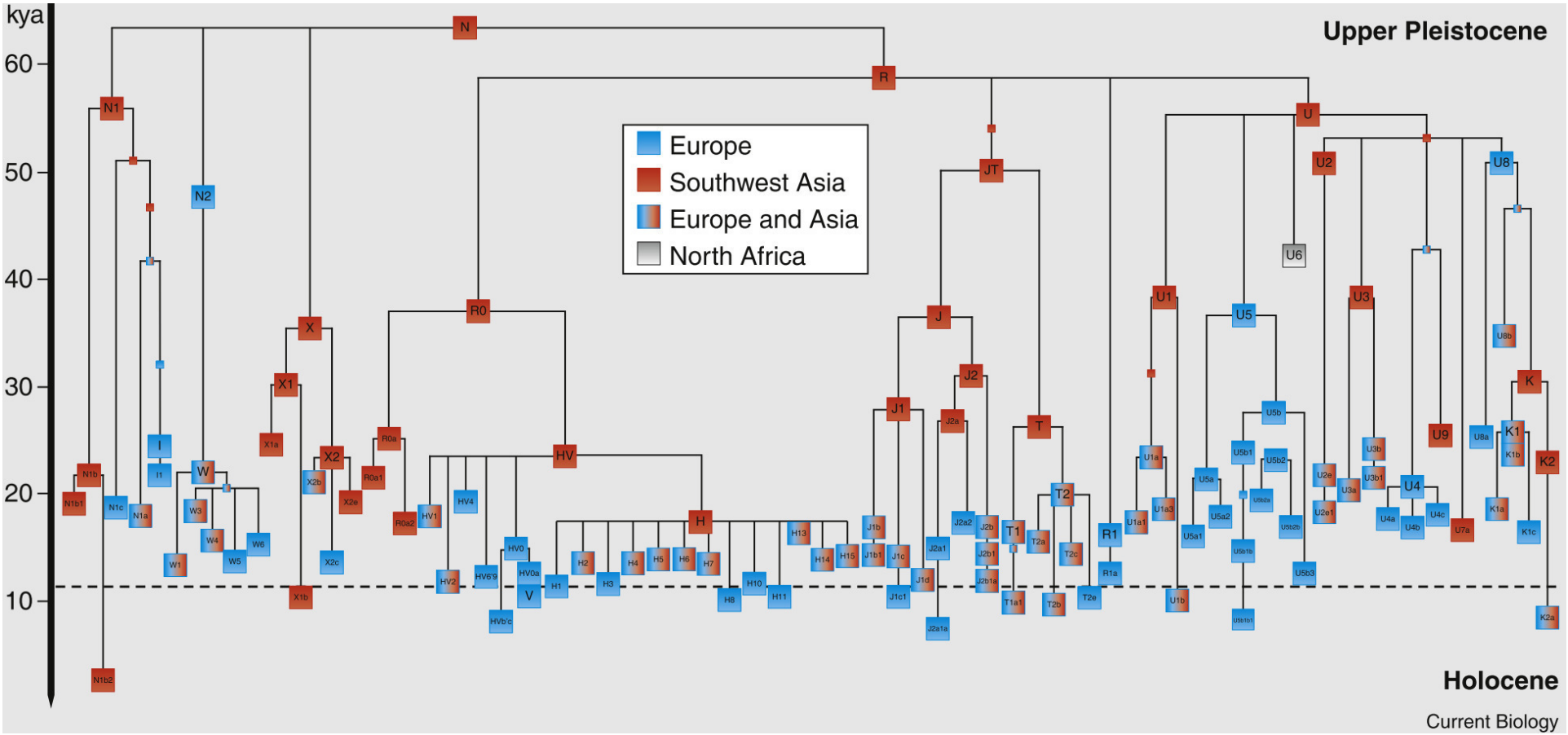


Figure 5 - Phylogeny of European mtDNA lineages, with respective ages[67]

While today haplogroup M is almost absent in European populations, a 2016 study found 3 hunter-gatherer individuals from current day France and Belgium, dating from between 28 and 25kya, carrying this haplogroup – suggesting a possible climate-influenced genetic bottleneck event[58] that led to its disappearance in Europe.

Currently, most of haplogroup N diversity is found in Asia[50], some subclades of N, were important in the peopling of Europe most likely form a South-western Asian source (Near East). 19,6% of individuals associated with LBK culture (*Linearbandkeramik*, or Linear Pottery, connected with Neolithic central European farmers) described by Haak et al. carried haplogroup N1a[59,60], however it was not found in the remains of adjacent Mesolithic hunter-gatherer populations[61] nor is it found in such numbers in modern Europeans (0,2%)[60]. This data challenges the notion that most of today's European genetic diversity is owed to farmers from the Near East[60]. Haplogroup I, by far the most frequent subclade within N1, is more frequent in Europe, however most diversity is found in the Gulf region, Anatolia and Southeast Europe, indicates that its probable origin is Near Eastern[62]. Haplogroup I (particularly its subclade I1a), was detected in 9% of a southern population in Portugal[63], which Santos et al. consider characteristic of Mediterranean Iberia[64]. W, a subclade of N2, reaches a frequency larger than 10% in some Eastern European regions such as Finland, however its diversity peaks in southeast Europe, northwest Africa and the Arabian Peninsula. Nevertheless, its Near Eastern lineages are nested among European clades, hence a probable European origin[62]. X, a haplogroup nested directly in N, is subdivided into X1, of overall lower frequency and restricted to North and East Africa and the Near East, and X2, which is found in Europe, western and central Asia, Siberia, the Near East, and interestingly, in North America[65]. The latter has been found in LBK, Alföld Linear Pottery culture and in Iberian Chalcolithic remains (3 individuals)[66,67]. Two subclades of X2, X2b3 and X2c2a, are considered unique to Sardinia, and the latter is implied to have originated in Iberia[40].

Major haplogroup R which comprises nearly all of European mitochondrial diversity and includes haplogroups R0 (which comprises of R0a'b, a relatively rare subclade present in the Near East[68], and its greatly diverse sister clade HV), JT , U and the very rare R1.

HV likely had its origin between Central and Western Asia[39], and while not particularly common in Europe if H is excluded (0-10%[69]), due to a strong founder effect it represents 93.3% of Belmonte Jews' maternal ancestry[70]. Nested into HV0a is haplogroup V[71], controversial in regards to its source: Torroni et al. (1996) implies a post Last Glacial Maximum (LGM) expansion from Iberia[72], while

Behar et al. (2012) argues a later Neolithic origin[15]. Despite its low expression in general[72], it reaches a frequency of 41,1% in the Saami[73] in Northern Scandinavia, 16% in the Tuaregs of the African Sahel Belt[74], 16% in Berbers from Matmata (Tunisia)[75] and 21% in Cantabrians[76].

H is an extremely diverse haplogroup spanning across Europe, the Near East and Central Asia[77], and at a frequency of 40% is the most frequent haplogroup in Europe. Up to 106 subclades have been described, and more recent studies on the full mtDNA genome have revealed that 71% of its diversity is located outside the control region[71]. It is generally accepted that H originated in the Near East[78] and most of the lineages arrived in Europe pre-LGM[57], expanding from Ice Age refugia in Iberia[44] and the Near/East Caucasus region[77]. In spite of this phylogeographic evidence, only one pre-Neolithic individual bearing H was found (in Cantabria[79]). Even in the Early Neolithic it seems less frequent[60], a possible indication that it expanded with late Neolithic pan-European cultures such as the Bell Beaker Culture (BBC)[14].

U clusters directly within R, and, like H, shows a wide dissemination across Europe and Asia[80]. It is subdivided into 4 subclades, U1, U5, U6 and U2'3'4'7'8'9. The latter subdivides into U2, U3, U4'9, U7, U8 (which contains K) and U9. The subclades show region specificity: U1 and U3 to the Near East[81–83], U2 has a frequency and diversity peak in South Asia with some subclades confined to Europe and the Near East[39], U4 ,U5[84] and U8[44] to Europe, U6 to North Africa and the Canary Islands[85], U7 has low frequencies in general but is widely distributed from Europe to India[80] and finally U9 is found only sporadically in Arabia, Ethiopia and India[86–88]. U most likely originated in Southwest Asia about 55kya shortly after the arrival of anatomically modern humans (AMH) from Africa[62] with U2 and U8 diverging soon after (50kya[44]). These lineages were possibly taken to Europe with its earliest AMH settlers: a U2 individual was found in Kostenki (southern Russia) dating to 38kya[89]; and pre-U5 and pre-U8b 32ky old remains were found simultaneously in present day Czech Republic[90]. U5 likely emerged in Europe 37kya[57] and together with U4 forms the bulk of lineages associated with hunter-gatherers[59,61]. One particular branch of U5b, U5b1b, common both in the Saami (together with haplogroup V it accounts for 90% of their mtDNA diversity) and the Berbers [91] has been cited as evidence of pre-Neolithic expansion from Franco-Cantabrian refugia, while at the same time U5b3 serves as a signal of expansion from the Italian peninsula[92] and U4 from the Ukranian plains[93]

Haplogroups U and R0's sister clade, JT (including the rare South-western Asian R2 branch), likely arose 58kya, followed by the parting of J and T, 40 and 30kya respectively, associated with

the peopling of the Fertile Crescent[94]. Both J and T now extend across Eurasia and North Africa, together making up to 20% of European mitochondrial diversity[95]. Pereira et al. demonstrated, with the aid of founder analysis, that the majority of these lineages entered the Mediterranean after the LGM but only expanded to the rest of Europe with the Neolithic.

1.2.2 Archaeogenetics

Archaeogenetics is a relatively new field of both archaeology and genetics and can be defined as "the application of molecular genetics to study the human past"[96]. It began with the study of inherited traits such as human blood groups[97], lactase persistence[98] and even earwax[99], their relationships with linguistic and ethnic groupings, and started becoming what today we know it as with the work of Luca Cavalli-Sforza. In the 1960s, this population geneticist pioneered statistical methods such as maximum likelihood to estimate evolutionary trees[100] and principal component maps to evaluate the distribution of genetic variation in space[101], focusing primarily on the Neolithic transition.

In the mid-1980s to 90s, the advent of polymerase chain reaction (PCR)[102] and Sanger DNA sequencing[103] initiated a new wave in this discipline. Together, these techniques allow short fragments of DNA to be sequenced in multiple individuals, and more importantly, from infinitesimal amounts of fragmented DNA, as is it many times the case with ancient DNA (aDNA). At the same time, mtDNA studies acquire an utmost importance: firstly because of its aforementioned high mutation rate, lack of recombination, hypervariable regions; and secondly, due to these attributes it is more lilely to be retrieved from aDNA..

1.2.2.1 Ancient DNA

Ancient DNA (aDNA) is DNA obtained from specimens not deliberately preserved for genetic analysis[104], like archaeological specimens (palaeogenetics), for example. Not by chance, mtDNA was the first[105] – and is the most[106] – used genetic marker in palaeogenetics as it has a high copy number relative to the nuclear genome and thus a greater per-locus chance of being recovered. This is an important feature when the concentration of DNA in artefacts is a few orders of magnitude lower than modern samples[107]. aDNA has some other very special characteristics: it undergoes modification by oxidative damage and hydrolytic processes[108] and is broken into very small fragments (nuclear DNA ~150bp, mtDNA ~400bp[24]). The advent of PCR and next generation

sequencing (NGS) are able to overcome these obstacles, however, they are two-edged swords: it can also amplify tiny contaminating material[24].In fact, the results of some high profile studies have turned out to be results of contamination [109]. The upper limit for aDNA conservation seems to be 100kya to 1 Mya[110], depending on the climate.

1.3. Human origins

1.3.1 Out of Africa

It is now widely accepted that AMH originated in the African continent. Previously, some regarded a possible multiregional hypothesis: a transition from Homo erectus to Homo sapiens took place in a number of areas outside Africa. This has since been disproved with the aid of fossil evidence [111],of genetic analysis (genome-wide, Y-chromosome and mtDNA). Indeed, both mtDNA and Y-DNA phylogenetic trees divide into African and non-African lineages, with the former being very deep rooted, contrasting with the latter being a star-like structure.

As to where in the African continent did our species arise, the debate is still on. Based on fossil evidence, Eastern Africa seems to be the most likely origin[111–113]. Other fossils point to Morocco or Israel[114,115], albeit their dating being controversial. Genetics-wise, the results are even more confusing. Two genome-wide studies indicated a southern origin [116,117] and a phylogenetic analysis of the Y-chromosome located its root in central/west Africa[118]. mtDNA, as previously stated, splits into two main lineages, L0 and L1-6. L1-6 has a central/eastern origin[119], while L0 probably has a southern origin, as it has a predominance in these areas[119]

Another source of contention is how many significant migration waves were there: a single or multiple exits. The multiple exit model is based on cranial findings in the 90s in Sub-Saharan Africa (SSA) that showed isolated populations with morphological differences between them. It was therefore postulated that these populations each exited in turn to form the different non-African populations[120]. The Y chromosome seemed to not contradict this, as it does not present one clear founder outside Africa[121]. The single exit model stated that only one significant wave of migration occurred, is responsible for the dispersal of AMH around the world, and relies strongly on mtDNA.

Studies show haplogroup L3, the ancestor of all non-SSA haplogroups is around 59-70ky old, and haplogroups M and N diverged around 50-60kya, which rules out any dispersal before the Toba eruption [43,50,62]. This model is further corroborated by increased resolution Y phylogenies that suggest a male equivalent of the L3 haplogroup (M168)[122].

Finally, which route was taken out of Africa? Looking at a map, only two routes seem plausible: one through the Sinai peninsula, and another via the Bab al-Mandab strait into the Arabian peninsula[123]. This last hypothesis seems likely given that haplogroups N and R seem to have originated in the Arabian peninsula[62].

1.3.2. Into Europe

One can divide the prehistory of modern humans in Europe into more or less five probable episodes: the Out of Africa colonization; the re-colonization from Southern refugia (after the Late Glacial Maximum (LGM); the postglacial re-colonization by Mesolithic groups at the end of the Younger Dryas (a sharp decline in temperature that ended around 11 kya[124]); dispersals of Near Easterners with the Neolithic and lastly, exchanges from the Copper Age onwards, including a probable major expansion from the Steppes in the Bronze Age.

The presence in Europe of the genus *Homo* seems to date back to 1.1Mya [125]. From 350 on it was occupied by *Homo neanderthalensis* or *Homo sapiens neanderthalensis*[126], a cold adapted species that diverged around 550 kya[50] from Homo heidelbergensis. Neanderthals did not survive in most of Europe/Caucasus after 39kya[127], however interbreeding between AMH and Neanderthals has been strongly suggested by archaeological[128] and genetic[129] approaches, despite no signal remaining in either mtDNA or Y-DNA, and few Neanderthal variants remaining in the X chromosome or in genes activated in the testes, indicating that the hybrids possibly suffered from reduced fertility [130]. Nevertheless, around 2% of the genome of non-Africans is Neanderthal[131], suggesting that interbreeding occurred mostly before the split into different non-African populations..

While Asia was likely colonised 60-70 kya [132] through a "Southern Coastal Route", the route (assuming a southern Out of Africa hypothesis) from the Arabian Gulf to the Levant was blocked by desert until 50kya[133] . As such, AMH only arrived in Europe 45kya[134]. Some important archaeological findings from this time period include the 45-43ky old Italian Grotta del Cavallo remains associated with Uluzzian culture[135] and the 44.2 – 41.5 ky old Kent's Cavern remains associated with Aurignacian culture [136]. Along with Proto-Aurignacian, these industries expanded

from a source in the Near East, as evidenced by findings from Ksar Akil, Lebanon[137]. The next major cultural diffusion in Europe, the Gravettian, appeared 28kya associated to a sharp degradation of climate. It is argued whether it arose from the Near East (based on genetic evidence[39]) or central/Eastern Europe[138]; nevertheless, Gravettian populations were characterized by their relatively high mobility and interaction between dispersed demes, as demonstrated by their use of non-local raw material and distribution of very similar figurines across wide geographical areas[139].

The Last Glacial Maximum (LGM) was a period of time between 25 to 19.5 kya, when human populations concentrated themselves the so-called "Southern refugia", such as south-west Europe, the Mediterranean, Balkans, Levant and the Eastern European plain[139]. Ice sheets covered Scandinavia, Northern Europe, the Alps and the Pyrenees, incorporating a large amount of the world's water and decreasing sea levels[140], while landscapes shifted from forests to open tundras in non-refugial Europe[139]. Notwithstanding, Gravettian culture subsisted in North/Central Europe until the early LGM[141]. There has been strong genetic evidence that most of the mtDNA signal in modern Europeans comes from recolonization of central and Northern Europe after a warming phase 15kya. Important European haplogroups such as V[142], H1, H3[143], H5[57], U4[93,144] and U5[91,92] appear to have originated in refugia and expanded into the rest of Europe, although recent aDNA studies have contradicted this view[59]. Recently, Pala et al. have presented support for the expansion of J and T haplogroups from a Near Eastern refuge[94]. It is however possible that all these expansions occurred not with the end of the LGM but with the end of the Younger Dryas – a short (12-11kya) "cold snap" during the warming period after the LGM, during which the climate conditions briefly became fully glacial once more[57]. This stabilization of climate marks the end of the Pleistocene and the start of the Holocene.

With the stabilization of climate, farming emerged independently in different locations of the world within a restricted timeframe[145], possibly linked with the extinction of various large mammals[21]. At 11kya, the Natufian people of the Near East began cereal agriculture[146] and animal husbandry[147]. This new way of life spread throughout Europe 9 to 5kya[148], not in a single continuous dispersal but through maritime and land movements along coastal routes and river valleys[149]. One route went from the Balkans to central Europe, associated with dairying culture[150] and lactase persistence alleles[151]; another along the Mediterranean coastline, splitting further into Iberia and the Paris basin, carrying sheep, goats and Cardial pottery[152]. Despite the fast waves of expansion, the British Isles and Scandinavia show no signs of the Neolithic dated more than 4-6kya[153,154].

Ancient DNA reveals that Neolithic hunter-gatherers and farmers coexisted closely for over 2000 years, with little to no admixture between them[155,156]. While the farmers showed a genetic affinity to today's Near Easterners[59], the hunter-gatherers were closer to Northern Europeans[155]. Is the current European genetic landscape a result of these Neolithic expansions? Two opposing models of expansion exist: a demic diffusion, involving gene flow, and the cultural diffusion, or acculturation.

The demic hypothesis was the first model proposed, by Cavalli-Sforza[157] in the 80s. In it, the population growth associated with larger availability of food propelled a wave of expansion into Europe, where they replaced foraging populations. This is evidenced by analysis of allele frequencies[158] and nuclear polymorphisms[159] that form clinal patterns oriented from Southeast to Northwest.

In the cultural diffusion hypothesis, the farmers did not move in great numbers but instead the hunter-gatherers adopted their way of life, ideas and technology. Founder analysis based on the control region of mtDNA contradicts the notion that most of European diversity is due to the Neolithic, supporting acculturation. Richards et al.[39] propose that most of extant mtDNA lineages are Palaeolithic in origin, and thus the diversity is due to post-LGM expansion. Ancient DNA also seems to support this, as is the case of the LBK remains that contain a large percentage of a haplogroup rarely seen in Europe today[59,60]; analysis of Neolithic pigs suggest they were imported from the Near East, while today's European pigs show more in common with the European wild boar[159].

Interactions between Palaeolithic hunter-gatherers and Near Eastern farmers do not seem to fully explain the genetic history of Europe. The Late Neolithic Corded Ware Culture (CWC) showed a significant (75%) affinity to the Yamnaya[160], herders with wheeled vehicles[161]from the Pontic steppe who carried ancestry from Ancient Northern Europeans (ANE) and Caucasian hunter-gatherers[162]. It is assumed that these invaders carried with them proto-Indo-European languages, and contribute in varying amounts to Europe's genes: from 67,8% of the Finns all the way to the 7,1% of the Sardinians[160], considered to be a genetic outlier and the closest to the Neolithic farmers[67,163]. All 7 of the male Yamnaya skeletons analysed by Haak and colleagues carried Y haplogroup R1b, now the most frequent haplogroup in Europe[164]. This suggests it was a male-dominated invasion: indeed, X-chromosome studies showed that for every Yamnaya woman, there were 5 to 14 men[164]. Surprisingly, Basques have the highest frequency of R1b in Europe (87,1%), despite not speaking an Indo-European language[165].

The Bell Beaker culture (BBC) was a widely scattered archaeological culture from the late Neolithic/Early Bronze age, so named because of the distinctive bell-shaped pottery artefacts. It is understood not only as a particular pottery type, but as a complete and complex cultural phenomenon. Radiocarbon dating seems to support that the earliest "Maritime" Bell Beaker design style is encountered in Iberia, specifically in the copper using communities of the Tagus estuary in Portugal around 4.8kya[166]. Central European BBC remains show a striking genetic resemblance to today's Iberian populations, with Brotherton et al. suggesting that the expansion of hg H to its current first-place in mtDNA diversity in Europe is due to migrations carrying this industry, and as such the expansion of BBC was not only a trading but a population expansion phenomenon, linked to the spread of the Celtic language family throughout Western Europe[14].

1.4 Aims

Despite being pioneered in the early 2000s, founder analysis is an underused method. It has only been used regarding solely the control region[39,167], or using the whole mtDNA of a few haplogroups of interest[40–44]. With today's large databases of whole mtDNA from all over the world and increased computer processing power, it seems only logical to use a large dataset of complete mtDNA sequences to settle some longstanding debates in Europe's history.

The goal of this project is to characterize the maternal population history of the Iberian Peninsula. In order to understand routes of Neolithic, postglacial or Bronze age migrations the population history of the Peninsula will need to be contextualized in the overall European scenario. Aims include to confirm if European maternal lineages are fruit of Neolithic[53] or post-LGM expansions[39], or more plausibly, a mix of both. Subsequent questions focus on the origin of certain lineages: is JT Neolithic or the echo of a post-glacial migration from the Near East[95]? Is H pre-Neolithic[57] or did it arrive with farming and expand with BBC[14]? Is V a marker of post-glacial expansion[15,72]?

One cannot think of Europe as a whole monolith peopled in a single wave. In all likelihood, evolving climate and burgeoning technologies meant that humans expanded, retreated and intermingled repeatedly over the last 50kya. As such, it makes sense to divide Europe in time and space and apply different founder analysis hypotheses. Certain areas of interest include the Mediterranean, as it possibly received the first Neolithic dispersal; Iberia, as an ice age refuge and westernmost tip of Europe; Sardinia, as the great outlier and Neolithic sink; the British Isles, who only received the Neolithic later; and the Near East, where it all began.

Chapter 2: Materials and methods

2.1 Dataset

A dataset was compiled containing both published and unpublished complete mtDNA sequences, the first obtained from NCBI (http://www.ncbi.nlm.nih.gov/nuccore/). Unpublished data was generated by the laboratories of Professor Martin Richards (University of Leeds and University of Huddersfield) and the supervisor. The total number of samples was 19159. Of these, 1617 were filtered out, due to being from aDNA, cancer studies (known for its mtDNA mutations), mitochondrial disorders, or simply bad quality sequences (e.g.: missing key polymorphisms). A full dataset of published samples ins available in Suplemmentary file 2.

The variations in the complete mtDNA samples were scored as variants when compared with the revised Cambridge Reference Sequence (rCRS)[13] and then the sequence was primarily apportioned to a specific haplogroup, considering the sequence variation, using the online software HaploGrep 2.0.

2.2 Phylogenetic reconstruction

A phylogenetic tree was constructed using the a reduced-median (RM) algorithm[38] of the software Network 5 and the Network Publisher add-on from Fluxus Technology Ltd. To avoid networks with an excess of reticulations, difficult to represent or interpret, fast sites were down-weighted. By default, the weighting value is 10 for each position, but for this analysis the weight of the very fast mtDNA positions (146, 150, 152, 195, 16129, 16189, 16193 16311 and 16362)[50] was dropped to 6. The polymorphisms 16182C, 16183C, 16518T and 16519C were removed altogether, along with indels. The reduction threshold was 1.6, except in haplogroup N1 in which it was 1. To resolve the reticulations inherent to this method, the tree chosen was the most likely to occur given the variation of mutability of the various polymorphic sites[50]. When the reticulations were constituted by polymorphisms with similar probability, other criteria was considered in the conversion to the phylogenetic tree, namely the diversity of the clades involved, how many individuals were present in each node and the recognition that the occurrence of specific mutations can vary between the mtDNA clades, according to Phylotree[38].

2.3 Founder analysis

After assemblage of the phylogenetic trees in Network, they were transcribed to Extensible Markup Language (XML). Samples were assigned as source, sink or undefined in accordance with the 7 different tests we devised. Samples were geographically classified as:

Near East: For the purpose of this study, samples were included in "Near East" if labelled as Turkey, Iran, Iraq, Kurd, Palestine, Israel, Druze, Bedouin, Jordan, Lebanon, Syria, Kuwait, Saudi Arabia, Yemen, Oman, Bahrain, Qatar or United Arab Emirates.

Caucasus: Samples were included in "Caucasus" if labelled as Armenian, Georgia, Azerbaijan, Dagestan, Chechnya, North Ossetia, Kabardian, Karachay, Adygei, Abkhasia, Artsakh, Ingush, Kalmyk, Lezgin, Dargin or Circassian.

Mediterranean: Samples were assigned to this class if labelled Greece, Bulgaria, Macedonia, Albania, Macedonia, Montenegro, Serbia, Bosnia, Herzegovina, Croatia, Slovenia or Italy (*sans* Sardinia). France was excluded, as more precise geographic location within the country was not assured. Sardinia was excluded for this test as it seems to be an outlier in the panorama of European genetic diversity[44].

Iberia: Portugal (including Madeira and Azores), Spain (including the Basque country, Balearic Islands but excluding the Canary Islands)

Europe: France, Germany, Switzerland, Belgium, Netherlands, Austria, Czech Republic, Slovakia, Hungary, Poland, Ukraine, Romania, Moldova, Belarus, Russia (if not specified as one of the Caucasian ethnicities above or Asian), Lithuania, Latvia, Estonia, Finland, Sweden, Norway, Denmark, Iceland

British Isles: Ireland, Northern Ireland, England, Wales, Scotland, Orkney Islands

Sardinia: Sardinia.

Excluded samples: For the purpose of the various analyses some samples were considered unclassified. Canary Islanders were excluded as there seems to be a greater genetic affinity to North Africa[168]. Europeans carrying haplogroup U6 were also excluded, as it is clearly a North African lineage[85] which would not be part of the selected source. Ashkenazi Jews were considered undefined as there is conflicting evidence in relation to their maternal ancestry being Near Eastern[169] or European[44]. Samples that showed clear recent migration events were classified as undefined.

2.3.1 Migration models

The following migration models were considered:

Test 1

Source: Near East and Caucasus

Sink: Europe, Iberia, Mediterranean and British Isles

Test 2

Source: Near East and Caucasus

Sink: Mediterranean

Test 3

Source: Near East, Caucasus and Mediterranean

Sink: Europe and British Isles

Alternative Test 3

Source: Near East, Caucasus

Sink: Europe and British Isles

Test 4

Source: Near East, Caucasus and Mediterranean

Sink: Iberia

Test 5

Source: Near East, Caucasus, Mediterranean, Iberia, Europe and British Isles

Sink: Sardinia

Test 6

Source: Near East, Caucasus, Europe, Mediterranean, Iberia

Sink: British Isles

The founder analysis was performed using an in-house software developed by Daniel Vieira. Migration clusters were computed using an effective number of samples per cluster in order to properly take into account the uncertainty in each cluster[43] and a time range of 100 to 55000 years (attending to the time of entry of AMH in Europe[57]) and intervals of 100 years. Mutation rate was defined as 2593y/mutation, estimated taking into account the curve described by Soares et al.[50] and a time range in the Bronze Age.

After a the founder analysis scan through equally distributed intervals that will allow to obtain a general distribution of the founder clades without imposing any specific model on the data, we defined specific migration event models as done traditionally for the Founder Analysis approach[39]. These give a) the percentage of lineages entering in each predefined migration event and b) the probability that a founder lineage entered during said migration event. Models were devised according to the peaks viewed in the founder analysis scan but also archaeological, palaeoclimatic and paleontological evidence. Time of arrival of the Neolithic was defined for each region according to Silva and Vander Linden[153]. The Mesolithic/Younger Dryas was evidentially allocated to the time of the Younger Dryas that established current climatic/resources conditions. We opted for not including a postglacial partition as the scan hardly indicated any entrance of lineages at that point. Nevertheless, lineages entering in that period would be allocated to this last migration (Mesolithic/Younger Dryas) that could account for all the period between the Ice Age and the Neolithic. The first settlement was defined from the first evidence of modern humans in Europe[134]

.

The Bronze Age migration is well defined archaeologically and genetically corresponding for most of Europe to a massive migration from the Steppes[160]. One migration value of 200 years was included to account for recent, historical migrations.

Given this, the defined migration times were the following:

2.3.2 Migration times

Test 1
Recent migrations: 200
Bronze age: 4500
Neolithic: 8000
Mesolithic/Younger Dryas: 11500
First settlement: 45000

Test 2
Recent migrations: 200
Bronze age: 4500
Neolithic: 8000
Mesolithic/Younger Dryas: 11500
First settlement: 45000

Test 3 and alternative Test 3:
Recent migrations: 200
Bronze age: 4500
Neolithic: 7000
Mesolithic/Younger Dryas: 11500
First settlement: 45000

Test 4:
Recent migrations: 200
Bronze age: 4500

Neolithic: 7200

Mesolithic/Younger Dryas: 11500

Test 5:

Recent migrations: 200

Neolithic: 7800

Mesolithic/Younger Dryas: 11500

Test 6:

Recent migrations: 200

Bronze age: 4500

Neolithic: 6400

Mesolithic/Younger Dryas: 11500

Chapter 3: Results

Test 1:

Source: Near East and Caucasus

Sink: Europe, Mediterranean, Iberia and British Isles

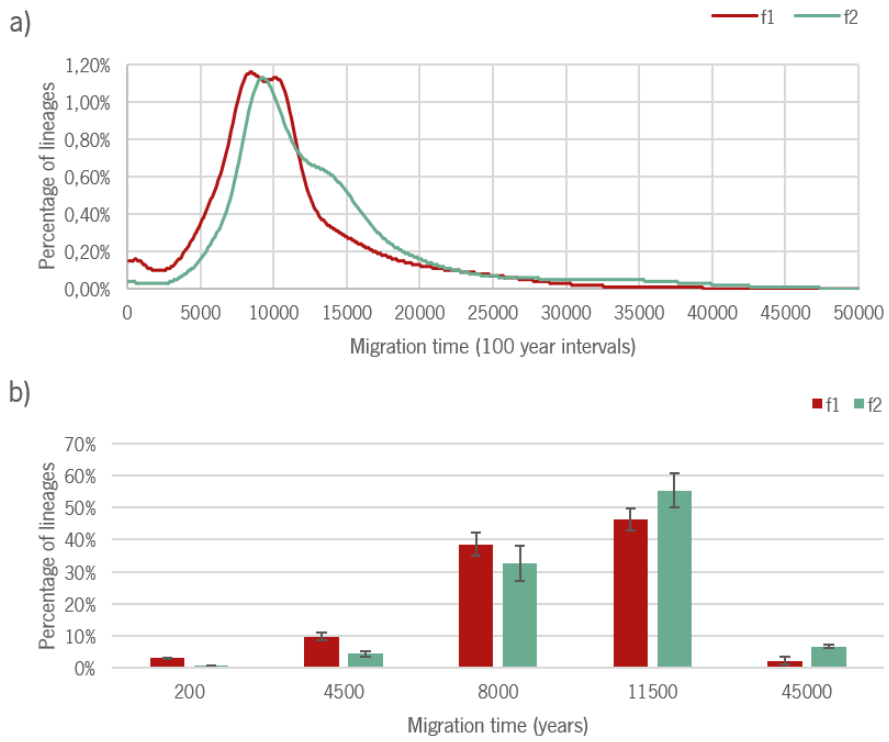


Figure 6 – Founder analysis results for test 1. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of first colonization, Late Glacial, Neolithic, Bronze Age and recent founder lineages in a five-migration model.

f1 criterion shows two very similar peaks at 8400 and 10000 while *f2* shows a slight hump at 13800 years ago (possibly Late Glacial) and a peak at 9200 (possibly Neolithic)(Figure 6a). Additionally, *f1* also detects a slight peak in migration in the last 1000 years. Overall, about half (46% if considering *f1*, 55% if considering *f2*) of lineages date to the Mesolithic (Figure 6b), echoing a previous study centred on JT lineages[95] . Indeed, as previously described, J1c, T2b and to a lesser extent J2b1 and J1c2 probably entered Europe post-glacially; the remaining lineages described in Pereira et al. are not detected in both criteria(Figure 7, Supplementary File 1) . Other post-glacial founders include HV, H, H2, H1u, K1, K1a, N1a3a, T2f, U5a2, U5a1 and U5b. These last subclades of U5 might represent back-migration into the Near East, as U5 is thought to have

originated in Europe[57]. Nevertheless, we should not exclude a back-migration into the Near East followed by a re-expansion into Europe.

The second largest (*f1* – 32%, *f2* – 39%) component is the Neolithic component, dated to 8000 years. Lineages such as HV+16311, H1, H3, U2e1a, and I2 hypothetically entered Europe during this period. Interestingly, N1a1a, associated with Neolithic dispersals in Europe[60], shows a mainly (81%) 45000 year old signal. Other lineages that entered during this period are U8 and X2+225 (only in *f1*) and U2'3'4'7'8'9 and U5 (in *f2*). U8 and U2'3'4'7'8'9 have been confirmed to be among the oldest European lineages[89,170]

Bronze age migrations represent between 4 (*f1*) and 10% (*f2*) of the lineages, with T1a1 being a main founder. The low frequency of this component is expected as the hypothetical source of Bronze Age migrations in not the Near East. A recent migration comprised mostly of U1a1c1c was also detected, which makes sense given that U1 is a mostly Near Eastern clade[81].

These general values of colonization into Europe are very misleading in relation to different geographic regions. Given this we performed more geographically focused analyses.

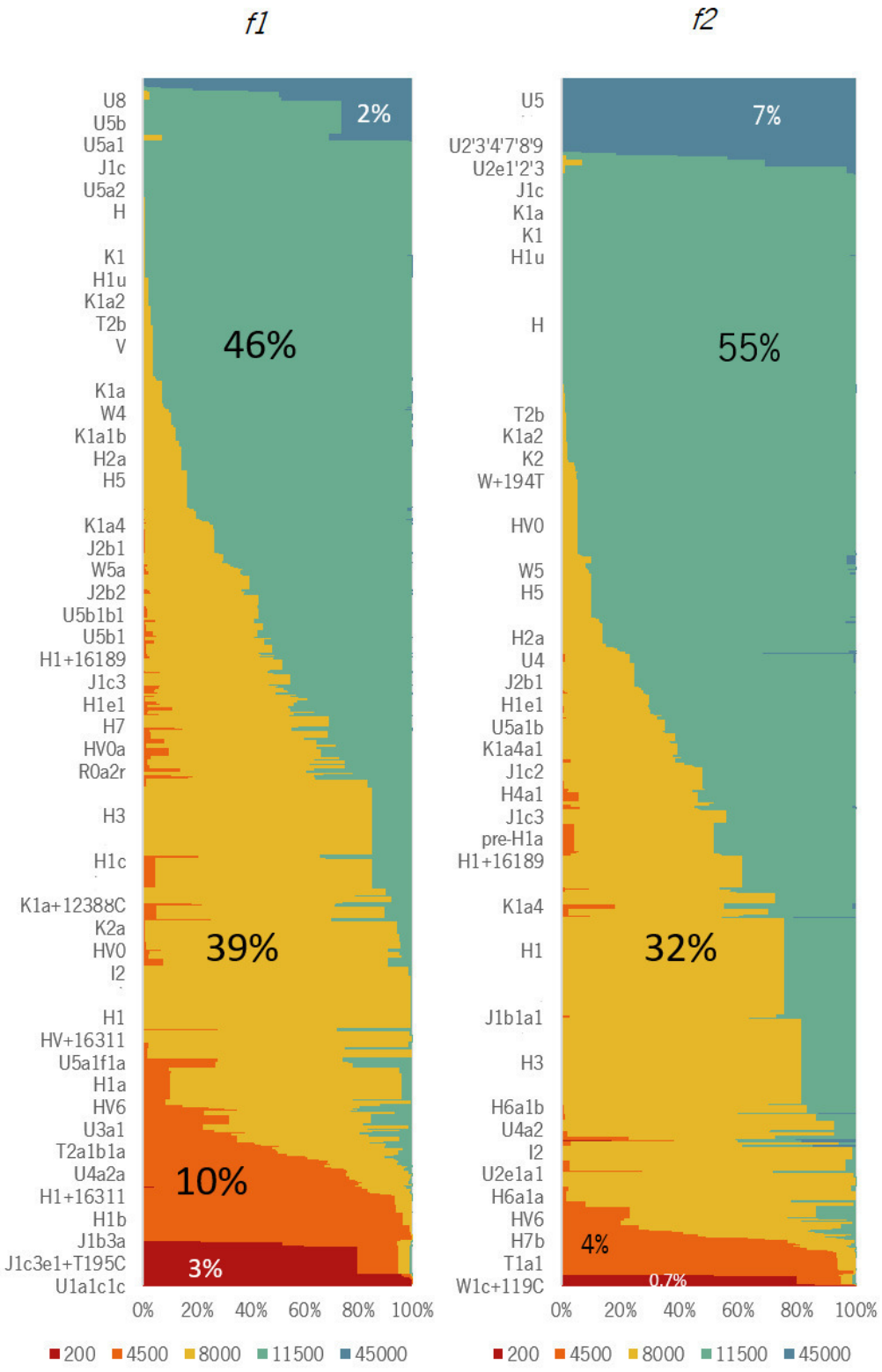


Figure 7 – Probabilistic distribution by migration time for some common founder clades in test 1: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Test 2

Source: Near East and Caucasus

Sink: Mediterranean

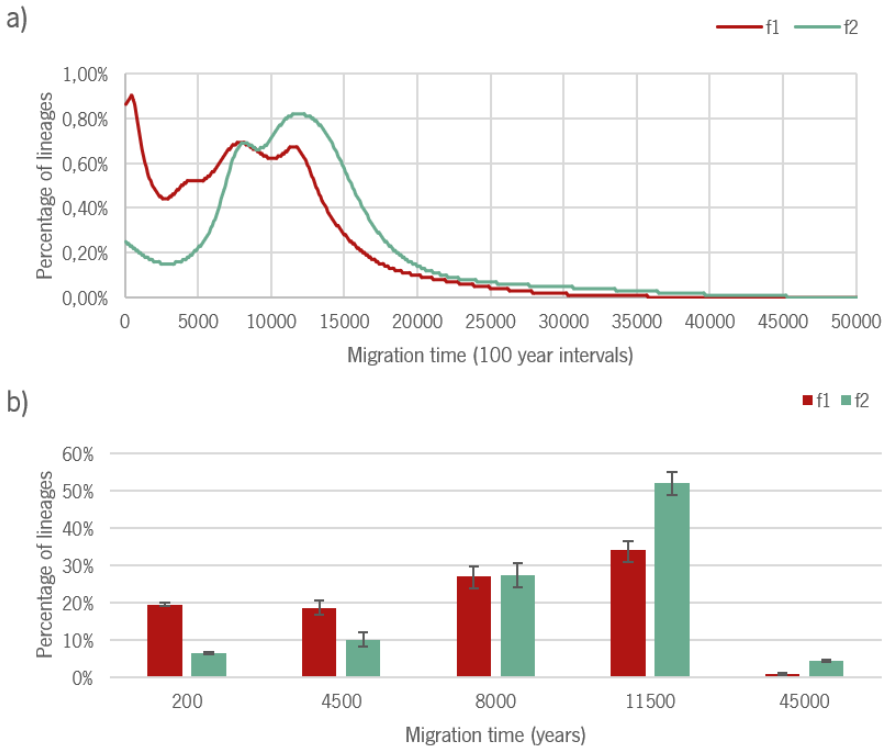


Figure 8 - Founder analysis results for test 2. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of first colonization, Late Glacial, Neolithic, Bronze Age and recent founder lineages in a five-migration model.

To better discern routes of migration, a founder analysis was performed using the Europe's Mediterranean coast (excluding Iberia) as a sink. Analyses of haplogroup JT suggested that Mediterranean Europe was the main receptacle of lineages from the Near East at different time periods.

f1 shows an almost constant wave of migration beginning circa 15kya. Both criteria show peaks around 12kya and 8.2kya (Figure 8a), perhaps tied to post-glacial and Neolithic expansions, respectively. Most lineages date to the Mesolithic (34% in *f1*, 52% in *f2*) (Figure 8b, Figure 9, Supplementary File 1) , particularly evident in *f2*. H, H5, U5a, U5a1, U5a2, U5b, J1a, J2b1 and T2b all entered in this time period. The Neolithic component accounts for 27% in both criteria, with J1c2, T1a1, H6a1a, H1, H3 and V being the main founders. The last three have been accounted

as postglacial lineages in the past[57]. Haplogroup V seems wrongly allocated here as it descends from a lineage (HV0) that was probably in Europe earlier.

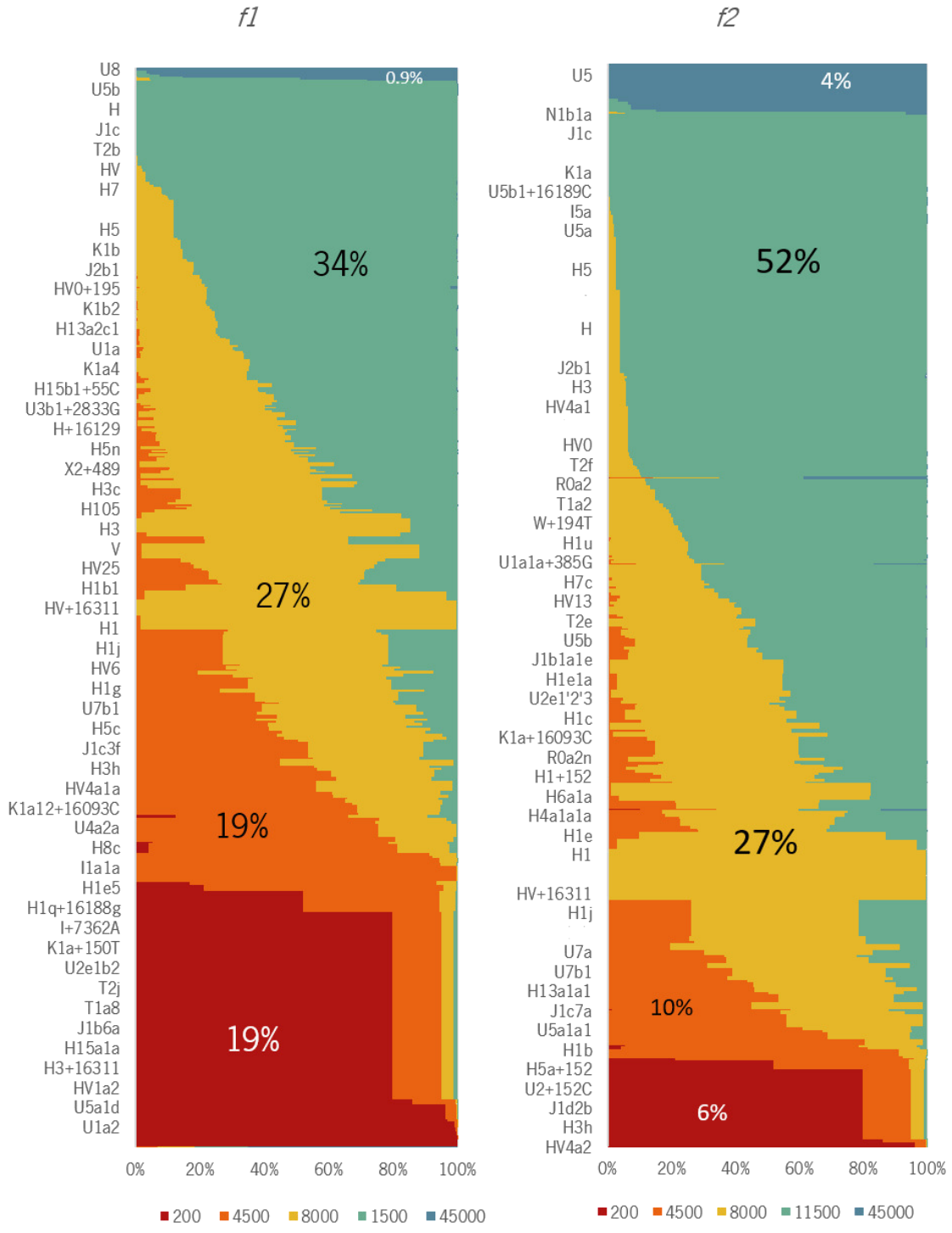


Figure 9 - Probabilistic distribution by migration time for some common founder clades in test 2: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Test 3 *vs.* Alternative Test 3

Test 3

Source: Near East, Caucasus and Mediterranean

Sink: Europe, British Isles

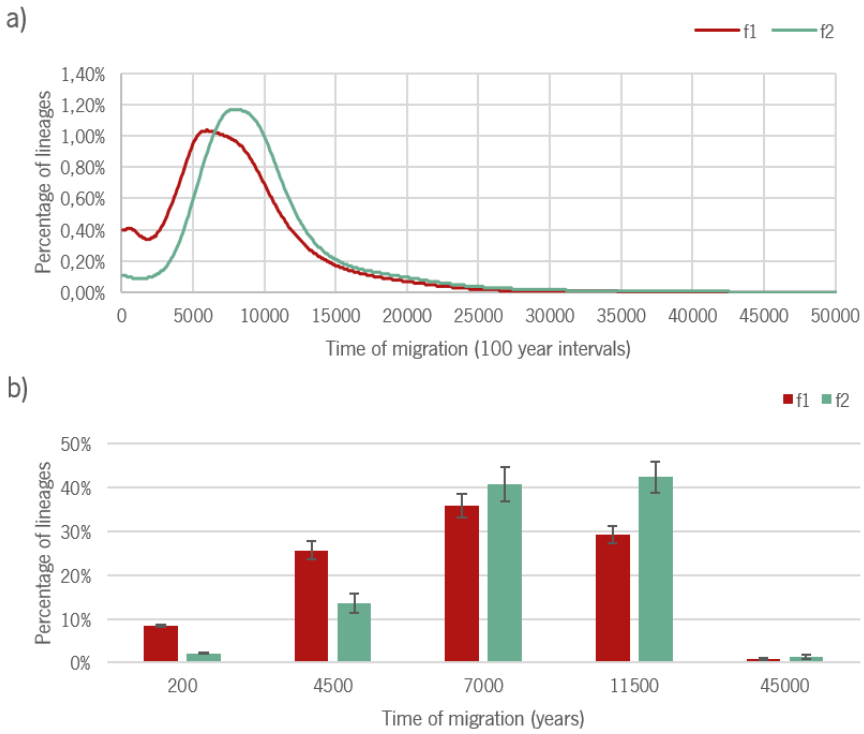


Figure 10 - Founder analysis results for test 3. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of first colonization, Late Glacial, Neolithic, Bronze Age and recent founder lineages in a five-migration model.

Pereira and colleagues[95] (2017) suggested that JT lineages expanded to Mediterranean Europe mostly in the Mesolithic, however they reached most of Europe only in the Neolithic. That established a scenario where the expansion in the Neolithic to most of Europe contained a large component of the Mesolithic lineages from the Mediterranean which we will test with full population data. Considering this we tested the colonization of Europe in two modes. In one the Mediterranean is considered in the source (Test 3) and in the second the Mediterranean is

excluded from the source (Alternative Test 3). We expect that in the alternative test some lineages that are clustered in the Mesolithic will be Neolithic if we consider the Mediterranean in Test 3.

In Test 3, *f1* showed a peak with a small plateau at 6kya and *f2* a sharper peak at 8kya (Figure 10a), coinciding with the Neolithic from archaeologic data. *f1* showed that slightly more lineages entered with the Neolithic than with the Mesolithic (36 *vs* 26%)for *f2* they are roughly the same (41 *vs* 42%). U5a1, U4b1, K1a and V, associated with post-glacial expansions, are confirmed to have expanded in this time period; a signal from N1a1a1a, commonly associated with the LBK culture of the Neolithic, is also detected (Figure 11 and Supplementary file 1).

Main lineages that expanded with the Neolithic comprise of H1, H3, T2b, I2, H6a1a and K2a.

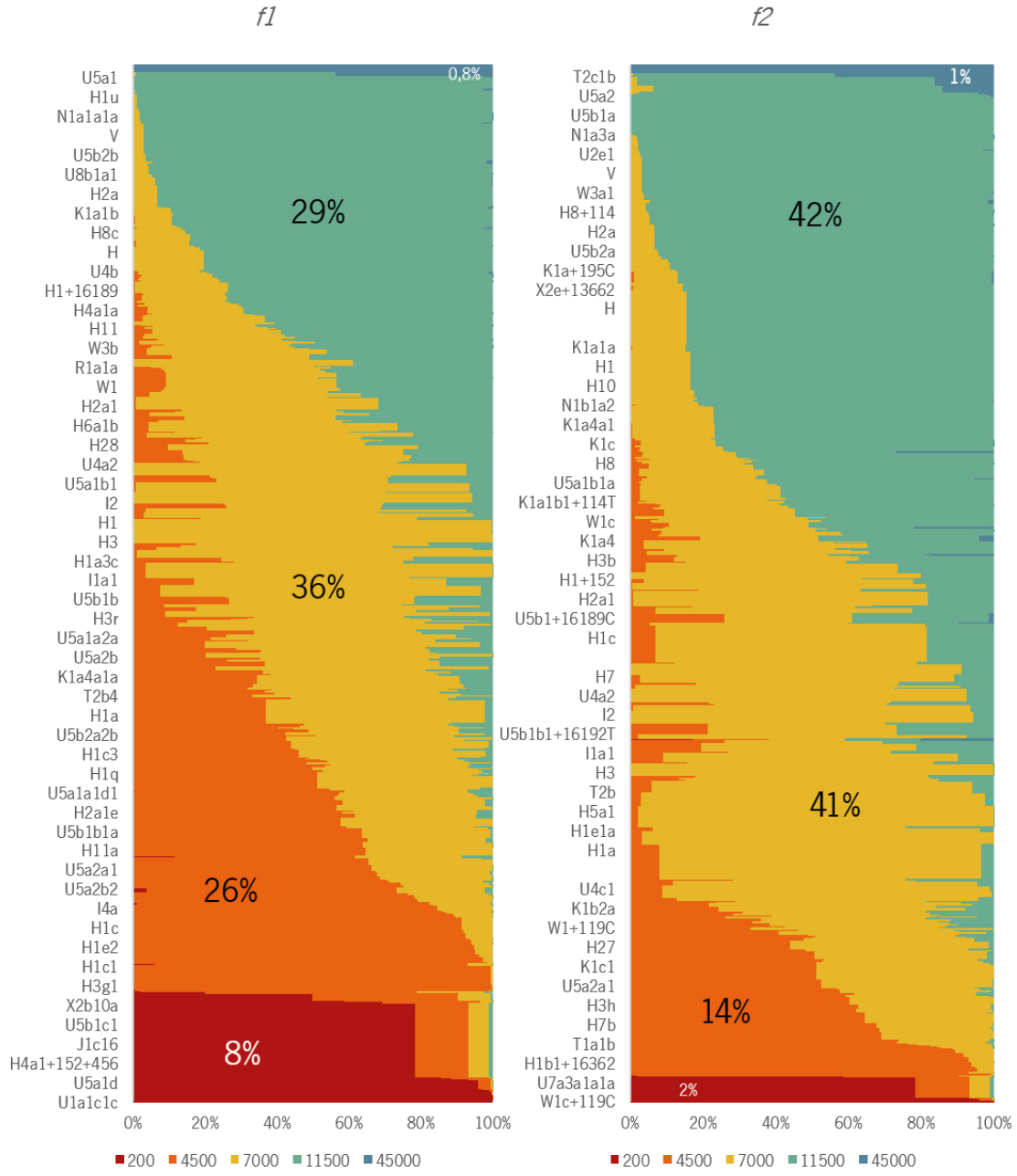


Figure 11 - Probabilistic distribution by migration time for some common founder clades in test 3: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Alternative Test 3

Source: Near East and Caucasus

Sink: Europe and British Isles

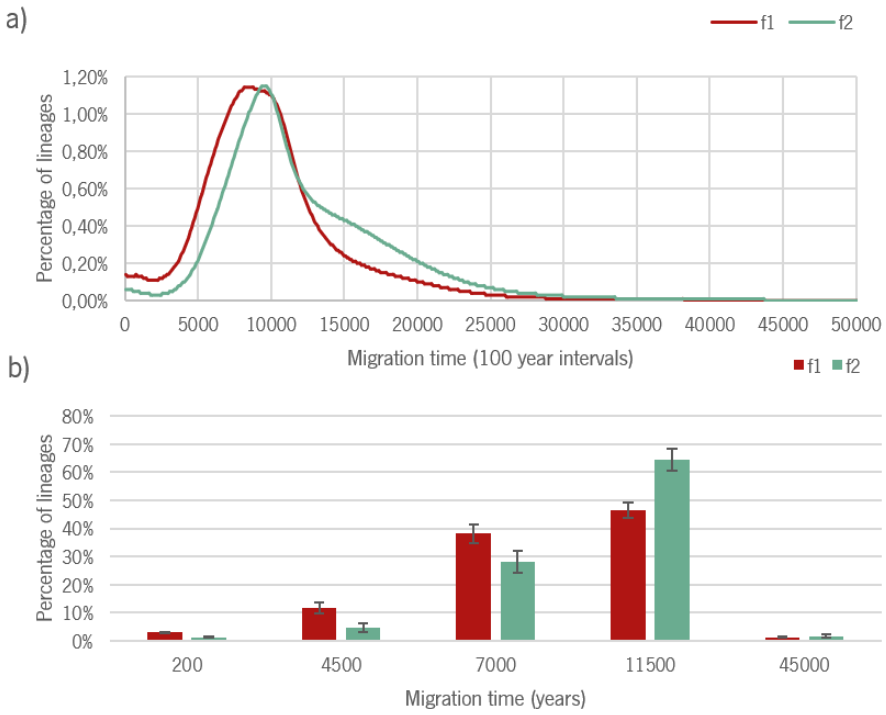


Figure 12 - Founder analysis results for alternative test 3. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of first colonization, Late Glacial, Neolithic, Bronze Age and recent founder lineages in a five-migration model.

In this test two clear peaks can be distinguished in each criterion: *f1* at 8800ya and a slightly humped *f2* peak at 9300ya (Figure 12a). This may be due to the geographic context of the Mediterranean in relation to the Near East. Very little lineages entered in the near past or in the first migrations. Lineages such as U2'3'4'7'8'9, U8, and N1a1a entered 45000 years ago according to the analysis (Figure 13 and supplementary file 1). The main Mesolithic lineages (*f1*- 46%, *f2* – 64%)were H and V, alongside H1u (that seems awkwardly placed here but again this is a less probable scenario than the previous one), H5, H2a, J1c, K1a and U5a2 (Figure 13 and Supplementary file 1). The second highest percentage of lineages (*f1*- 38%, *f2* – 28%) (Figure

12b) entered with the Neolithic. These include T2b (with a much lesser impact in *F2*), U4a2, K2a, I2, H3, H1, H6a1a and H5a1.

A comparison between both Tests 3 shows a shift of around 20% in both criteria from Mesolithic (almost 50% in *f1* and over 60% in *f2*) in the alternative test 4 to less than 30% (*f1*) and just above 40% (*f2*) in Test 4. This drop of frequency for Test 4 corresponds to an increment of the Neolithic or Bronze Age components when considering the Mediterranean in the source, some results much more in agreement with aDNA studies.

Following the establishment of the migration patterns into Continental Europe, founder analyses were performed into more peripheral areas namely Sardinia to the South, Britain to the North and Iberia to the West.

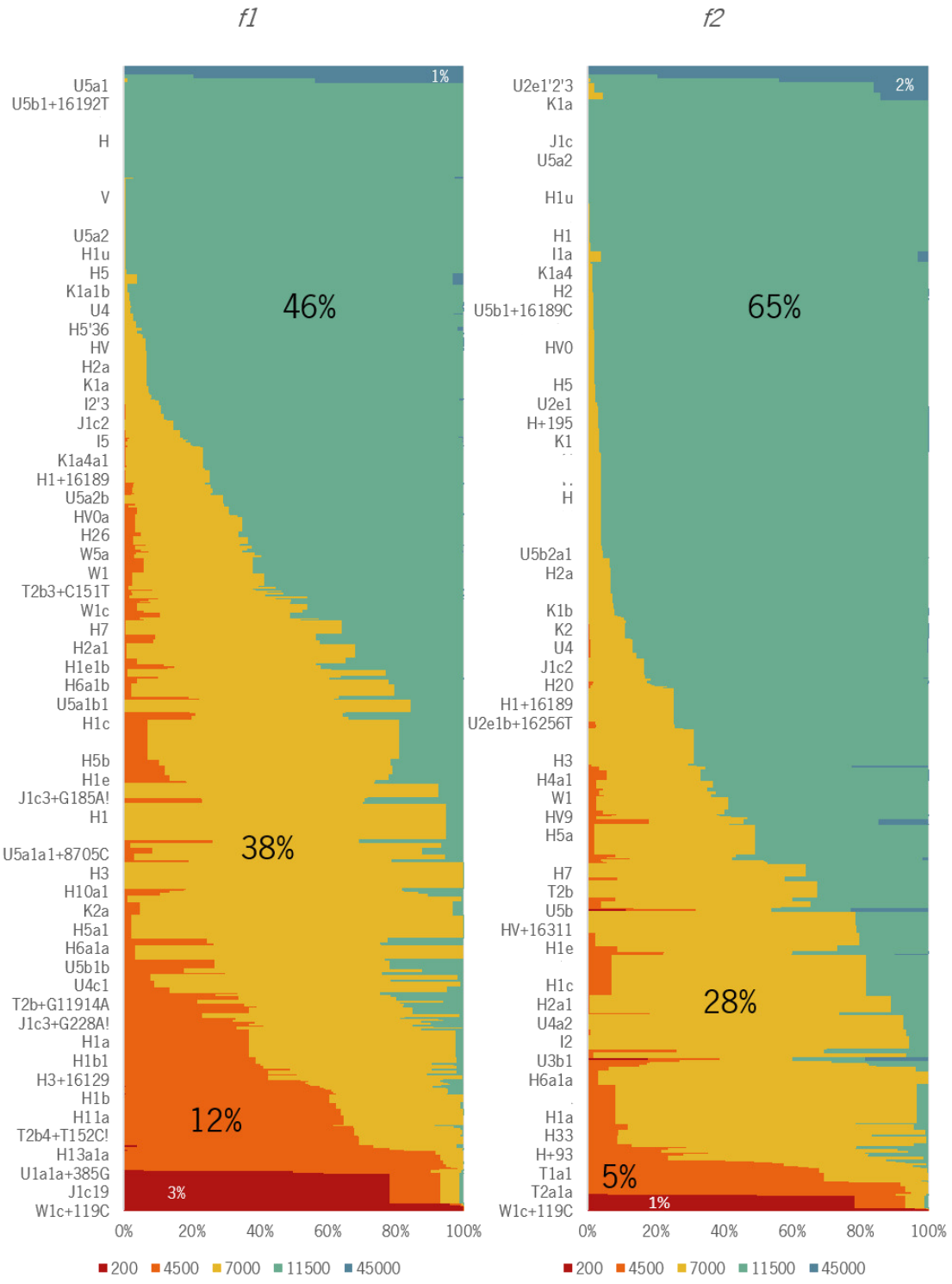


Figure 13 - Probabilistic distribution by migration time for some common founder clades in alternative test 3: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Test 4:

Source: Near East, Caucasus and Mediterranean

Sink: Iberia

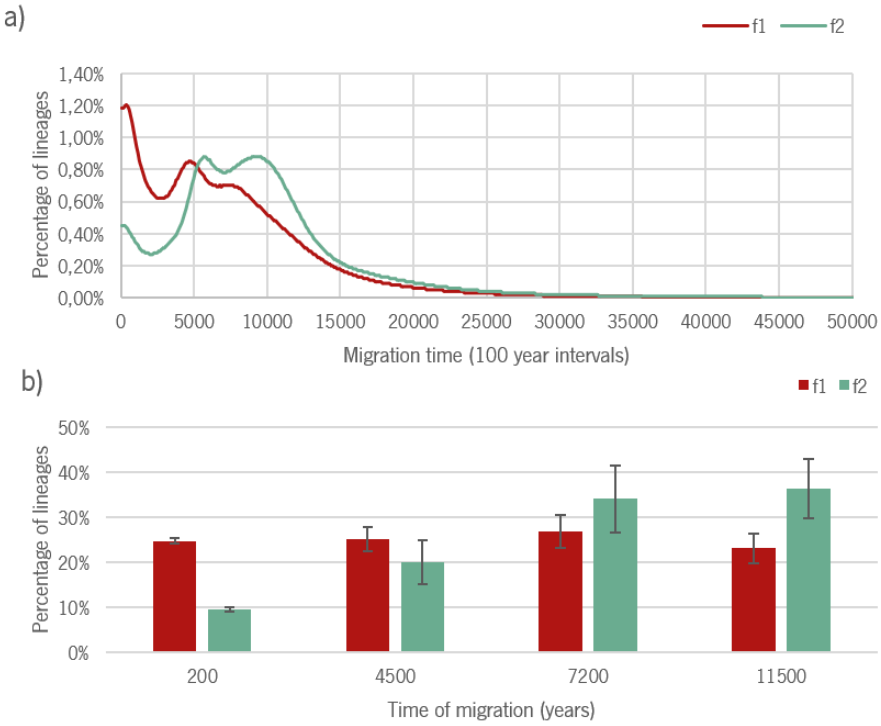


Figure 14 - Founder analysis results for test 4. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of Late Glacial, Neolithic, Bronze Age and recent founder lineages in a four-migration model.

In the founder analysis scan to the Iberia Peninsula, *f1* and *f2* appear slightly different: while the first shows an almost crescendo of lineages with subtle peaks at 200, 5000 and 7200 ya, almost perfectly matching the Bronze Age and the Neolithic, the latter shows two peaks at 5600 and 9100 somewhat older or younger than expected (Figure 14a). Regarding the migration event models, while *f1* partitions lineages into a quarter at each migration event, *f2* shows that most lineages entered at similar frequencies between the Mesolithic and Neolithic, with each event thereafter decreasing in percentage (Figure 14b).

Main Mesolithic founders were X2i, U2'3'4'7'8'9, U5a1, U5b1, U5b2, N1a1a, HV1b, K2b and H. As for the Neolithic, only HV+16311 appears as a very good founder (>90% probability) for this time period. Other potential candidates for founders in this time period are T1a1a, K1a4a1, U2e1a1, T2b3 and H6a1a (Figure 15 and Supplementary file 1) .

Regarding the Bronze Age, I1a1a only appears as an important founder in *f1*; H3, H1b and T2a1a appear as equally probable founders in both. H1e seems to be a recent (200y) lineage in both criteria.

The results seem to suggest that the Mesolithic wave of migration in the Mediterranean expanded until the Iberia Peninsula, followed by a strong migration proportion in the Neolithic and a substantial one in the Bronze Age.

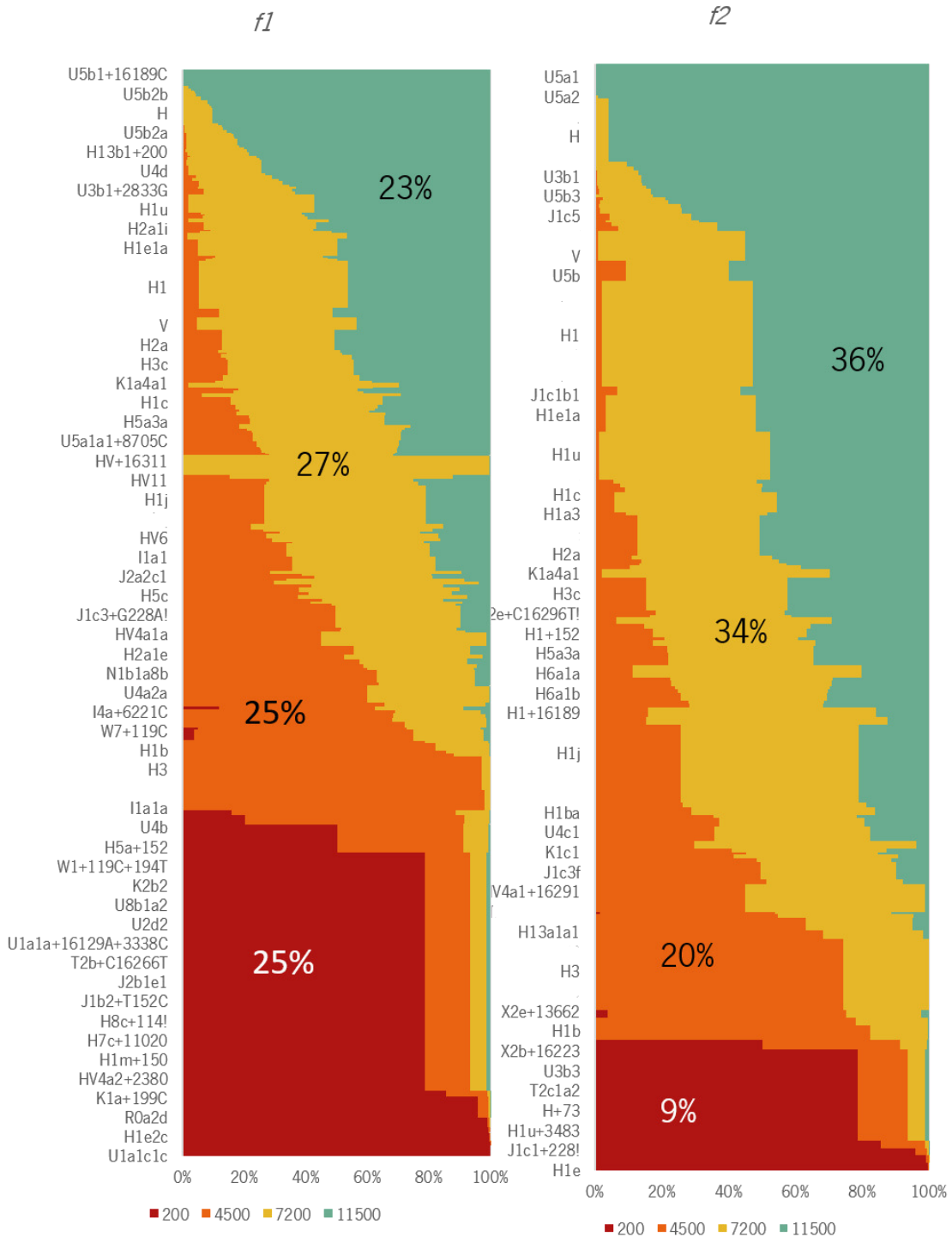


Figure 15 - Probabilistic distribution by migration time for some common founder clades in test 4: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Test 5

Source: Near East, Caucasus, Mediterranean and Europe

Sink: Sardinia

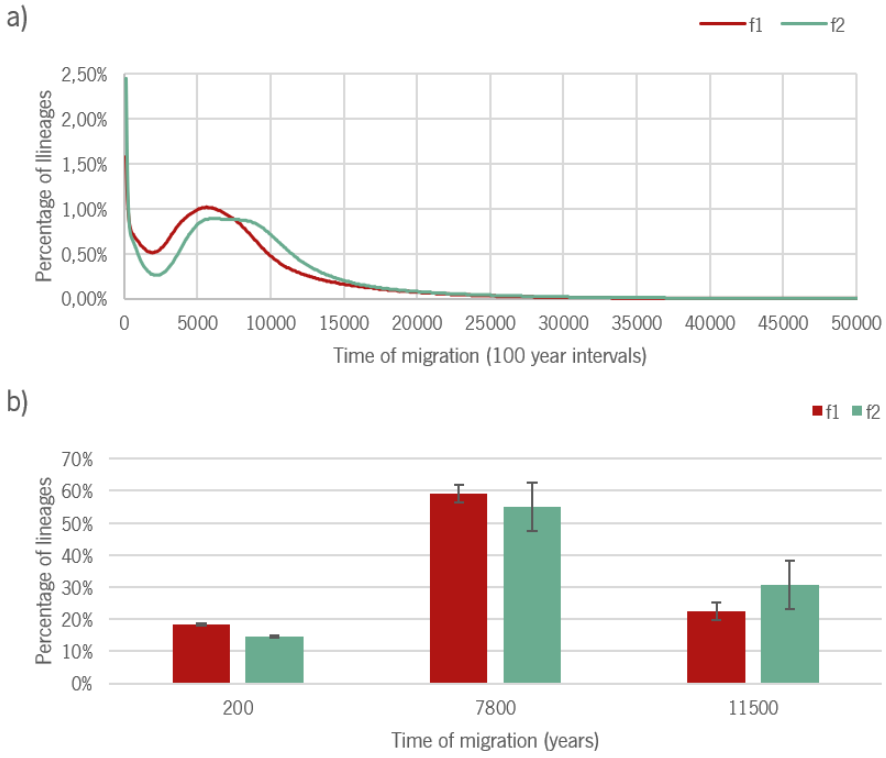


Figure 16 - Founder analysis results for test 5 a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of Late Glacial, Neolithic and recent founder lineages in a three-migration model.

Apart from a sharp increase in lineages ca. 2000 years ago, clear peaks appear at 6kya and 7kya (*f1* and *f2* respectively), corresponding to the Neolithic period (Figure 16). This is confirmed by the time of migration model: between 50 and 60% of Sardinian lineages appeared in this time period (Figure 16b). These include H1c1, H6a1a, H5a1, U5b3a1a, H1, I2, H27, U4a2, H1b and H3 (Figure 17 and Supplementary file 1). Of these, only U5b3a1a is considered a Sardinian-specific haplogroup per Olivieri et al.[40], with a BEAST estimated age of 8.73kya, which coincided with our data. Mesolithic lineages are mostly comprised of T2c and T2b3+151T, with a contribution from

U4b1a and U5a2a. Sardinia has been described on genome-wide patterns as the most representative living European population of the Neolithic wave and the results corroborate this. Two Mesolithic samples (with 10,000 years) retrieved from Sardinia were from haplogroups I3c and J2b1[171]While the first was not detected in the Island in the present day, the latter has founder ages of around 10,000 years matching the expectation.

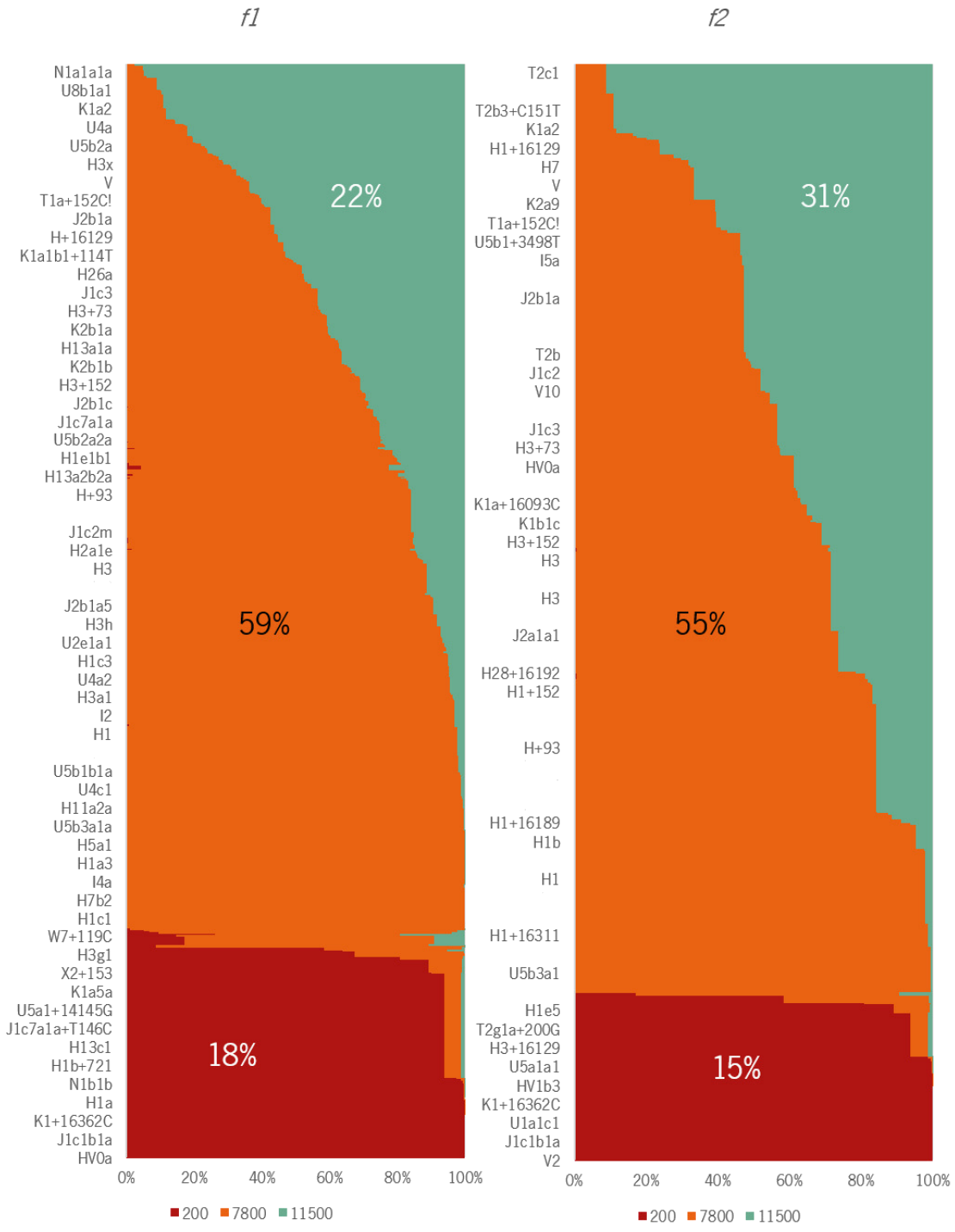


Figure 17 - Probabilistic distribution by migration time for some common founder clades in test 5: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Test 6

Source: Near East, Caucasus, Mediterranean, Iberia and Europe

Sink: British Isles

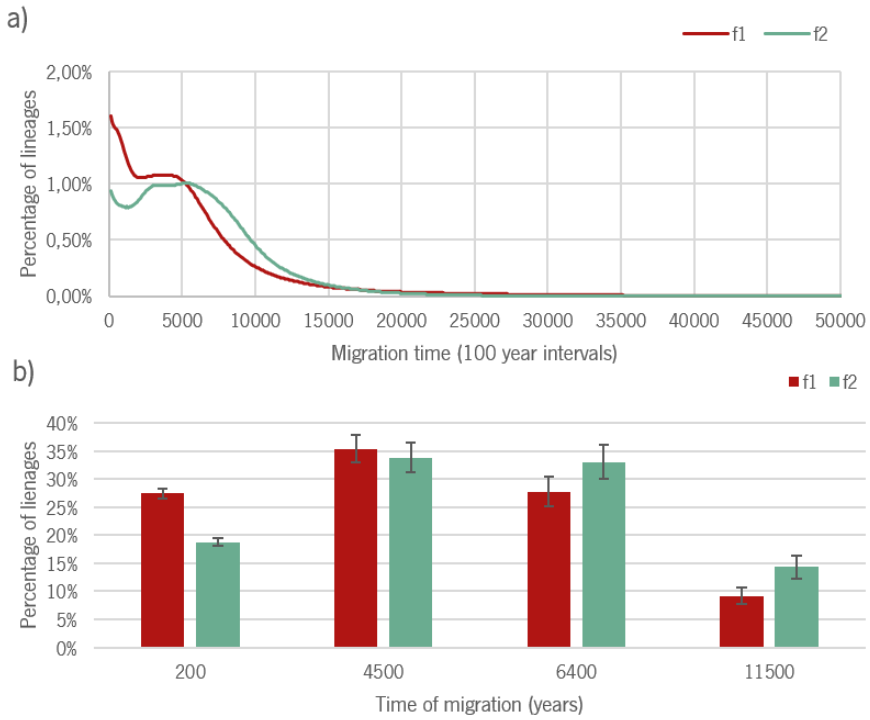


Figure 18 - Founder analysis results for test 6. a) Probabilistic distribution across migration times scanned at 100 year intervals from 0 to 55 kya; b) proportion of Late Glacial, Neolithic, Bronze age and recent founder lineages in a three-migration model.

In the founder analysis into the British Islands, two plateauing peaks appear at 3500 and 4900 years agos (*f1* and *f2*) respectively (Figure 18a), indicating that the majority of lineages entered the British Isles probably with the Neolithic and Bronze age expansions, a situation further confirmed in Figure 18b. Contrary to the previous results in this work, there are no lineages that are totally Neolithic, however H, H3 and I2 show that they mostly expanded in this timeframe. Haplogroups H1c1, H1c3b, K1c2+9006G, H4a1a1a, H27 and I4a expanded in the Bronze Age (Figure 19 and Supplementary File 1), falling in line with previous aDNA studies on Beaker culture remains.

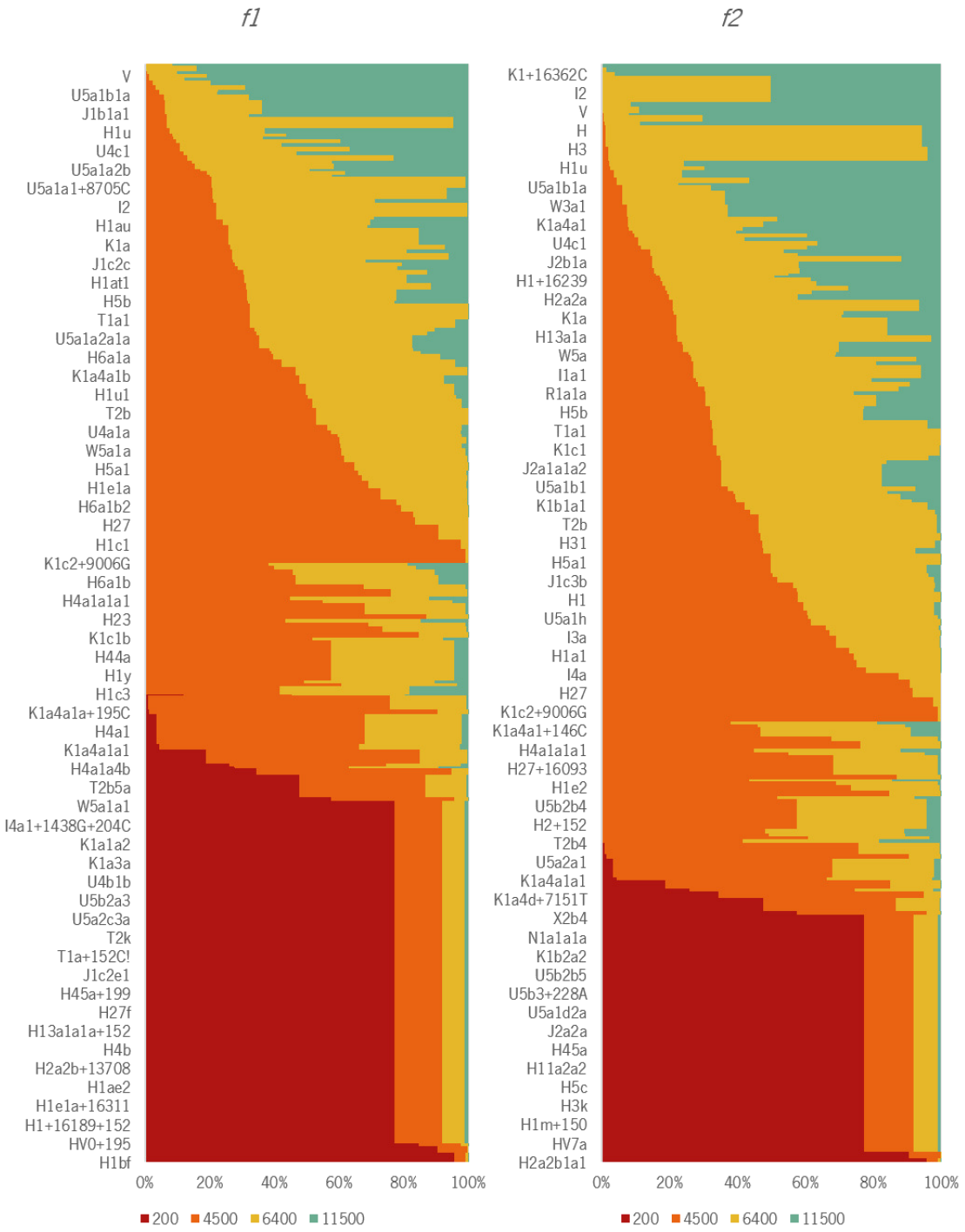


Figure 19 - Probabilistic distribution by migration time for some common founder clades in test 6: on the left, for criterion f1 and on the right, for criterion f2. Full data is available in supplementary file 1.

Discussion

Nature rarely deals in absolutes, and the maternal history of Europe – and Iberia in particular – is not an exception. Indeed, looking at our founder analysis it is impossible to say that in general most lineages are Neolithic or Mesolithic as populations display a complex history with various layers of demographic occurrences.

Overall, the first peopling of Europe seems to have contributed less than 5% to the current mitochondrial diversity. For the Palaeolithic, some first lineages appearing in Europe are U2'3'4'7'8'9, U8, U2 and U5, a finding corroborated by archaeology[58,90,170]. More puzzling is the finding of subclades of X2(X2i, X2+225 and X2b'd) and N1a1a in this time frame, as they appear only in remains related to Neolithic cultures such as LBK, Alföld Linear Pottery and the Iberian Chalcolithic[59,60,66,67]. This may be an error due to the extremely low amount of these lineages in the population sampled or it is such a case where the lineages in the source disappeared completely by drift. That seems certainly the case for N1a1a as published and unpublished aDNA shows the presence of the clade in the Near East 10,000 years ago.

A clear signal of expansion from the Near East into Europe after the LGM is detected. Previous studies concerning JT[94,95] had pointed to this result: we indeed find J1c, T2b and J2b1 lateglacial founders in Europe as a whole and in the Mediterranean. Some discrepancies are noted in relation to Pereira et al: T2e, is mostly Neolithic; J1c2 entered central Europe in the Mesolithic but dates to the Neolithic in the Mediterranean, representing a possible back migration.

It is clear that other haplogroups also represent a migration from the Near East in this timeframe. While a significant amount of lineages such as U5a1, U5b1, U5a2 appear, even when applying the *f2* criterion, it is clear from not only aDNA (citation) but also phylogeography that U5 arose in Europe – this signal is possibly due to a back migration into the Near East and posterior

migration into Europe with the end of the LGM. Another lineage that enters in this timeframe is H, now the most frequent in Europe, and previously thought to have entered pre-LGM. Together with HV, H2a, H5'36, H5, H7, H8 and sister-clade V, these previously suggested markers of expansion from European refugia show a signal of expansion from the Near East into Europe alongside the climatic improvement, and show some regional specificity: Iberia only receives basal H and H5; Sardinia H5'36, H5 and H2+152C; the British Isles receive relatively little, mostly V, H2a1a and H1u. This pattern might account for what was thought to be fruit of Bronze Age expansions. H3 and U4a appear Mesolithic in test 2, but only when applying the *f2* criterion, and later appear clearly in test 3 as Neolithic; Whole genome aDNA studies confirm a turnover of genetic diversity in Europe around the LGM: genetic clusters of the remains of ancient Mediterraneans start to resemble present day Near-Easterners[1]; the HERC2 lighter pigmentation variant appears at the same time in the Middle East and Mediterranean[170].

One of the goals of this work was to find if Near Eastern lineages had expanded in the Mediterranean before colonizing the rest of Europe, particularly in the Neolithic, hence tests 3 and alternative test 3, in which the Mediterranean was added to the source in the latter. It is now clear that some lineages that entered the Mediterranean in the Mesolithic later expanded into the rest of Europe in the Neolithic, an example of which are H3, T2b, H7, U5a1a1 and U5a1b1 deriving from U5, H5a1 deriving from H5. T2b has been associated in numerous aDNA studies[66,160,173] with Neolithic cultures such as LBK. Pereira et al argues for a postglacial entry in the Mediterranean, followed by a Neolithic expansion into Iberia. Lineages such as I2, I4a and I1a1 also appear in Europe in this time period, however, given that they do not appear in the Mesolithic in the Mediterranean, it is therefore possible that they represent a migration leaving from the Caucasus and entering Europe through the East. This wave of expansion might explain the slight masked peak present in the alternative test 3 but absent in test 3.

T1a1a1 has been described as a marker of the Neolithic[94,95], however, in our analyses it is the ancestral T1a1 that appears, spreading relatively quickly across the Mediterranean, Iberia and Sardinia with the advent of farming. In the British Isles, the signal is weaker, meaning it could have been brought with Bronze Age expansions, as is the case in Europe.

In our analyses, H6a1a is a clear Neolithic marker: it enters the Europe through the Mediterranean and through the steppes in the Neolithic. In Sardinia, it is clearly Neolithic, with the presence of two variants (H6a1a+152C, H6a1a+16482); in the British Isles and Iberia, the signal is partially (50 to 60% in the latter) Neolithic with a 30% probability Bronze age expansion. It has been found in CWC[14] and post-Bell Beaker remains[174].

Along with along with T1a1, the most definite signal of the Neolithic in Iberia is HV+16311, which has been found in aDNA from this time period in Spain[175]. While basal HV, H and some of its subclades and V appeared in the Mesolithic, this lineage spreads throughout Europe, with no discrimination between the Mediterranean and the Eastern point of entry. In Sardinia and the British Isles, however, it is a recent entry.

Much has been debated about Bronze age migrations, associated with cultural phenomena such as the Bell Beakers. Coinciding with previous suggestions[14], many lineages deriving from H flourish during this period. A major founder, ubiquitous throughout the Mediterranean, Europe, Iberia and British Isles is H4a1a1a. aDNA confirms that it has been found in remains dating to a relatively short timeframe, from cultures as vast as the Central European Unetice[14,176], Bell Beaker[14], Corded Ware[177], Bulgaria[172], and Lithuania[178]. Regarding Iberia, T2a1a, as predicted in Pereira et al.[95], entered in this time. Alongside it, subclades such as H3, H5f, H1b, H24 enter from the Mediterranean, or in the case of I1a1, a more central European source. In the British Isles the scenario seems different. Many lineages that are subclades of K, such as K2a6, K1c1b,

K1a4a1+195C and K1c2 – lineages derived from subclades present in the Mesolithic in Europe, appear, alongside the typical H-subclades, T2b4, T2a1a and I4a. All of these lineages have appeared associated with the Bell Beaker package, alongside older European lineages such as U5.[174,177]

A significant part of our overall lineages seem to have entered recently (in the past 200 years). While this might be the case with some lineages, like U1a1c1c, with a clear recent Near Eastern origin, others are probably older; e.g., W1c+119 is a recent lineage in Europe according to the tests, yet it has been found in German Neolithic remains[177]. When only one founder for a lineage exists (the so called "singletons"), the algorithm tends to dramatically underestimate the age of it.

Conclusion and further perspectives

The original goal of this work was to characterize the maternal history of Iberia. When plotting it, it became evident that one cannot study Iberia without studying Europe as a whole as well as its fringes (British Isles, Sardinia). These provide context for the story to unfold. We tried to approach this study as from a blank slate – without leaning toward a Neolithic or post-glacial major component for the history of different region of Europe.

What we found was that the current mitochondrial diversity in Iberia cannot be said to be Neolithic, post-glacial, or pre-glacial; it is rather what Gamble called a "bio-tidal zone"[179], where numerous populations "wash up", than a refugia. The major source for post-glacial refugia seems to have been the Near East, and not Iberia, the Italian Peninsula and the Ukranian plains, as previously described[57]. To further assess Iberia's maternal history, more whole mtDNA studies of the population must be done, especially for Portugal and non-Basque Spain. Compared to some populations such as the Sardinians[40], very little sampling of the populations has been done, and when available only the control region has been sequenced[63,180]. We know that a substantial amount of mtDNA diversity is present in the coding region (especially in the H haplogroup[71] – the most common haplogroup not only in Iberia but in Europe as well, and containing important markers of post-glacial, Neolithic and Bronze age expansions).

mtDNA only tells us one side of the story – the maternal, female side. As such, male-driven migrations, such as the invasion from the Eurasian steppes, remain uncharacterized in Iberian populations. A further founder analysis, identical to this one but with the Y-chromosome, would be an interesting endeavour. It could possibly uncover other previously unknown, sex-driven expansions. Not less exciting is the possibility of expanding the database of genome wide sampling, in Iberia and the rest of Europe, and comparing it with genome-wide aDNA.

The amount of studies on aDNA continues to be relatively small, especially for remains dating to the pre-Neolithic. While we can obviously not control the amount of remains that survive to this day, an increase in aDNA studies in Iberia would certainly help to corroborate findings.

It is clear that in the extremes of the time frame studied (in the recent history and in the first colonization of Europe), founder analysis under or overestimates the age of lineages. This needs to be revisited in further work, not only with the aforementioned increase in mtDNA sampling but also of with a possible revision of the FA software.

References

1. Siekevitz, P. Powerhouse of the Cell. *Sci. Am.* **197,** 131–144 (1957).

2. McBride, H. M., Neuspiel, M. & Wasiak, S. Mitochondria: More Than Just a Powerhouse. *Curr. Biol.* **16,** R551–R560 (2006).

3. Yi, M., Weaver, D. & Hajnóczky, G. Control of mitochondrial motility and distribution by the calcium signal. *J. Cell Biol.* **167,** 661–672 (2004).

4. Harding, A. E. & Holt, I. J. Mitochondrial myopathies. *Br. Med. Bull.* **45,** 760–71 (1989).

5. Chaturvedi, R. K. & Flint Beal, M. Mitochondrial Diseases of the Brain. *Free Radic. Biol. Med.* **63,** 1–29 (2013).

6. Jitrapakdee, S., Wutthisathapornchai, A., Wallace, J. C. & MacDonald, M. J. Regulation of insulin secretion: role of mitochondrial signalling. *Diabetologia* **53,** 1019–1032 (2010).

7. Westermann, B. Mitochondrial fusion and fission in cell life and death. *Nat. Rev. Mol. Cell Biol.* **11,** 872–884 (2010).

8. Gray, M. W., Burger, G. & Lang, B. F. The origin and early evolution of mitochondria. *Genome Biol.* **2,** REVIEWS1018 (2001).

9. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11,** 209 (2010).

10. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392,** 37–41 (1998).

11. Barrell, B. G., Bankier, A. T. & Drouin, J. A different genetic code in human mitochondria. *Nature* **282,** 189–94 (1979).

12. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).

13. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23,** 147 (1999).

14. Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the genetic

origins of Europeans. *Nat. Commun.* **4,** 1764 (2013).

15. Doron M. Behar, M. van O. S. R. M. M. E.-L. L. N. M. S. T. K. A. T. R. V. *et al.* A 'copernican' reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).

16. Bandelt, H.-J., Kloss-Brandstätter, A., Richards, M. B., Yao, Y.-G. & Logan, I. The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *J. Hum. Genet.* **59,** 66–77 (2014).

17. Nicholls, T. J. & Minczuk, M. In D-loop: 40 years of mitochondrial 7S DNA. *Exp. Gerontol.* **56,** 175–181 (2014).

18. Holt, I. J. & Reyes, A. Human mitochondrial DNA replication. *Cold Spring Harb Perspect Biol* **4,** 1–16 (2012).

19. Howell, N., Elson, J. L., Howell, C. & Turnbull, D. M. Relative rates of evolution in the coding and control regions of African mtDNAs. *Mol. Biol. Evol.* **24,** 2213–2221 (2007).

20. Brandst[illegible]er, A. *et al.* Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *Int. J. Legal Med.* **118,** 294–306 (2004).

21. Jobling, M., Hollox, E., Kivisild, T. & Tyler-Smith, C. *Human Evolutionary Genetics*. **1,** (Garland Science, 2015).

22. Robin, E. D. & Wong, R. Mitochondria1 DNA Molecules and Virtual Number of Mitochondria per Cell in Mammalian Cells. *J. Cell. Physiol.* **136,** 507–513 (1988).

23. Ye, K., Lu, J., Ma, F., Keinan, A. & Gu, Z. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci.* **111,** 10654–10659 (2014).

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

25. Rossignol, R. *et al.* Mitochondrial threshold effects. *Biochem. J.* **370,** 751–62 (2003).

26. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nat. Rev.*

Genet. **6,** 389–402 (2005).

27. Awadalla, P., Eyre-walker, A. & Smithz, J. M. Linkage Disequilibrium and Recombination in Hominid. **286,** 1–4 (1999).

28. Schwartz, M. & Vissing, J. Paternal Inheritance of Mitochondrial DNA. *N. Engl. J. Med.* **347,** 576–580 (2002).

29. Wai, T. *et al.* The role of mitochondrial DNA copy number in mammalian fertility. *Biol. Reprod.* **83,** 52–62 (2010).

30. Sutovsky, P. *et al.* Development: Ubiquitin tag for sperm mitochondria. *Nature, Publ. online 25 Novemb. 1999; | doi10.1038/10.1038/46466* **402,** 371 (1999).

31. Carvalho-Silva, D. R., Santos, F. R., Rocha, J. & Pena, S. D. The phylogeography of Brazilian Y-chromosome lineages. *Am. J. Hum. Genet.* **68,** 281–6 (2001).

32. Alves-Silva, J. *et al.* The Ancestry of Brazilian mtDNA Lineages. *Am. J. Hum. Genet.* **67,** 444–461 (2000).

33. Hurles, M. E. *et al.* European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. *Am. J. Hum. Genet.* **63,** 1793–806 (1998).

34. Thyagarajan, B., Padua, R. A. & Campbell, C. Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.* **271,** 27536–43 (1996).

35. Avise, J. C. *et al.* Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annu. Rev. Ecol. Syst.* **18,** 489–522 (1987).

36. Hewitt, G. M. Speciation, hybrid zones and phylogeography - Or seeing genes in space and time. *Mol. Ecol.* **10,** 537–549 (2001).

37. Bandelt, H.-J., Macaulay, V. & Richards, M. B. Median Networks: Speedy Construction and Greedy Reduction, One Simulation, and Two Case Studies from Human mtDNA. *Mol. Phylogenet. Evol.* **16,** 8–28 (2000).

38. Bandelt, H.-J. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141,** 743–753 (1995).

39. Richards, M. B. *et al.* Tracing European Founder Lineages in the Near Eastern mtDNA

Pool. *Am. J. Hum. Genet* **67,** 1251–1276 (2000).

40. Olivieri, A. *et al.* Mitogenome Diversity in Sardinians: A Genetic Window onto an Island's Past. *Mol. Biol. Evol.* **34,** 1230–1239 (2017).

41. Soares, P. A. *et al.* Resolving the ancestry of Austronesian-speaking populations. *Hum. Genet.* **135,** 309–326 (2016).

42. Soares, P. *et al.* Climate change and postglacial human dispersals in Southeast Asia. *Mol. Biol. Evol.* **25,** 1209–1218 (2008).

43. Soares, P. *et al.* The expansion of mtDNA haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* **29,** 915–927 (2012).

44. Costa, M. D. *et al.* A substantial prehistoric european ancestry amongst ashkenazi maternal lineages. *Nat. Commun.* **4,** 1–10 (2013).

45. Pereira, L. *et al.* High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* **15,** 19–24 (2005).

46. Mishmar, D. *et al.* Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 171–6 (2003).

47. Kivisild, T. *et al.* The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172,** 373–387 (2006).

48. Pereira, L., Soares, P., Radivojac, P., Li, B. & Samuels, D. C. Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. *Am. J. Hum. Genet.* **88,** 433–439 (2011).

49. Forster, P., Harding, R., Torroni, A. & Bandelt, H.-J. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59,** 935–45 (1996).

50. Soares, P. *et al.* Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *Am. J. Hum. Genet.* **84,** 740–759 (2009).

51. Howell, N. *et al.* The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am. J. Hum. Genet.* **72,** 659–70 (2003).

52. Santos, C. *et al.* Understanding Differences Between Phylogenetic and Pedigree-Derived mtDNA Mutation Rate: A Model Using Families from the Azores Islands (Portugal). *Mol. Biol. Evol.* **22,** 1490–1505 (2005).

53. Cavalli-Sforza, L. L. & Minch, E. Paleolithic and Neolithic Lineages in the European Mitochondrial Gene Pool. *Am. J. Hum. Genet.* **61,** 247–251 (1997).

54. A Torroni, T. G. S. M. F. C. M. D. B. J. V. N. M. L. D. G. S. C. M. V. D. C. W. Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am. J. Hum. Genet.* **53,** 563 (1993).

55. Behar, D. M. *et al.* The Dawn of Human Matrilineal Diversity. *Am. J. Hum. Genet.* **82,** 1130–1140 (2008).

56. Oppenheimer, S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 770–784 (2012).

57. Soares, P. *et al.* The archaeogenetics of Europe. *Curr. Biol.* **20,** R174-83 (2010).

58. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in Europe. *Curr. Biol.* **26,** 827–833 (2016).

59. Haak, W. *et al.* Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biol.* **8,** (2010).

60. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310,** 1016–8 (2005).

61. Bramanti, B. *et al.* Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. *Science (80-.).* **326,** 137–140 (2009).

62. Fernandes, V. *et al.* The Arabian cradle: Mitochondrial relicts of the first steps along the Southern route out of Africa. *Am. J. Hum. Genet.* (2012). doi:10.1016/j.ajhg.2011.12.010

63. Marques, S. L. *et al.* Portuguese mitochondrial DNA genetic diversity—An update and a phylogenetic revision. *Forensic Sci. Int. Genet.* **15,** 27–32 (2015).

64. Santos, C. *et al.* Mitochondrial DNA and Y-chromosome structure at the mediterranean

and Atlantic façades of the Iberian Peninsula. *Am. J. Hum. Biol.* **26,** 130–141 (2014).

65. Reidla, M. *et al.* Origin and diffusion of mtDNA haplogroup X. *Am. J. Hum. Genet.* **73,** 1178–90 (2003).

66. Lipson, M. *et al.* Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551,** 368–372 (2017).

67. Günther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci.* **112,** 11917–11922 (2015).

68. Gandini, F. *et al.* Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. *Sci. Rep.* **6,** 25472 (2016).

69. De Fanti, S. *et al.* Fine dissection of human mitochondrial DNA haplogroup HV lineages reveals paleolithic signatures from European Glacial refugia. *PLoS One* **10,** 1–19 (2015).

70. Nogueiro, I., Teixeira, J. C., Amorim, A., Gusmão, L. & Alvarez, L. Portuguese crypto-Jews: the genetic heritage of a complex history. *Front. Genet.* **6,** 12 (2015).

71. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** 386–394 (2009).

72. Torroni, A. *et al.* Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144,** 1835–50 (1996).

73. Tambets, K. *et al.* The Western and Eastern Roots of the Saami—the Story of Genetic 'Outliers' Told by Mitochondrial DNA and Y Chromosomes. *Am. J. Hum. Genet.* **74,** 661–682 (2004).

74. Pereira, L. *et al.* Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. *Eur. J. Hum. Genet.* **18,** 915–923 (2010).

75. Fadhlaoui-Zid, K. *et al.* Mitochondrial DNA heterogeneity in Tunisian Berbers. *Ann. Hum. Genet.* **68,** 222–233 (2004).

76. Maca-Meyer, N. *et al.* Y Chromomosome and Mithocondrial DNA Characterizatin of Pasiegos , A Human Isolate from Cantabria (Spain). *Ann. Hum. Genet.* **67,** 327–339 (2003).

77. Roostalu, U. *et al.* Origin and Expansion of Haplogroup H, the Dominant Human Mitochondrial DNA Lineage in West Eurasia: The Near Eastern and Caucasian Perspective. *Mol. Biol. Evol.* **24,** 436–448 (2007).

78. Richards, M. *et al.* Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59,** 185–203 (1996).

79. Ottoni, C. *et al.* Mitochondrial Haplogroup H1 in North Africa: An Early Holocene Arrival from Iberia. *PLoS One* **5,** e13378 (2010).

80. Sahakyan, H. *et al.* Origin and spread of human mitochondrial DNA haplogroup U7. *Sci. Rep.* **7,** 1–9 (2017).

81. Derenko, M. *et al.* Complete Mitochondrial DNA Diversity in Iranians. *PLoS One* **8,** e80673 (2013).

82. Al-Zahery, N. *et al.* In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. *BMC Evol. Biol.* **11,** 288 (2011).

83. Al-Zahery, N. *et al.* Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol. Phylogenet. Evol.* **28,** 458–472 (2003).

84. Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. & Barbujani, G. Geographic Patterns of mtDNA Diversity in Europe. *Am. J. Hum. Genet* **66,** 262–278 (2000).

85. Pereira, L. *et al.* Population expansion in the North African Late Pleistocene signalled by mitochondrial DNA haplogroup U6. *BMC Evol. Biol.* **10,** 390 (2010).

86. Abu-Amero, K. K., González, A. M., Larruga, J. M., Bosley, T. M. & Cabrera, V. M. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. *BMC Evol. Biol.* **7,** 32 (2007).

87. Quintana-Murci, L. *et al.* Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74,** 827–45 (2004).

88. Fornarino, S. *et al.* Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol. Biol.* **9,** 154 (2009).

89. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464,** 894–897 (2010).

90. Fu, Q. *et al.* A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Curr. Biol.* **23,** 553–559 (2013).

91. Achilli, A. *et al.* Saami and Berbers–an unexpected mitochondrial DNA link. *Am. J. Hum. Genet.* **76,** 883–6 (2005).

92. Pala, M. *et al.* Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. *Am. J. Hum. Genet.* **84,** 814–21 (2009).

93. Malyarchuk, B. *et al.* Mitochondrial DNA Phylogeny in Eastern and Western Slavs. *Mol. Biol. Evol.* **25,** 1651–1658 (2008).

94. Pala, M. *et al.* Mitochondrial DNA signals of late glacial recolonization of europe from near eastern refugia. *Am. J. Hum. Genet.* **90,** 915–924 (2012).

95. Pereira, J. B. *et al.* Reconciling evidence from ancient and contemporary genomes: a major source for the European Neolithic within Mediterranean Europe. *Proc. R. Soc. B Biol. Sci.* **284,** 20161976 (2017).

96. Renfrew, C. & Boyle, K. *Archaeogenetics*. (McDonald Institute for Archaeological Researcj, 2000).

97. Boyd, W. C. Newer Concepts of Human Races Suggested by Blood Group Studies. *J. Natl. Med. Assoc.* **44,** 1 (1952).

98. Harrison, G. G. Primary Adult Lactase Deficiency: A Problem in Anthropological Genetics. *Am. Anthropol.* **77,** 812–835 (1975).

99. Bass, E. J. & Jackson, J. F. Cerumen types in Eskimos. *Am. J. Phys. Anthropol.* **47,** 209–210 (1977).

100. Cavalli-Sforza, L. L. & Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19,** 233–57 (1967).

101. Cavalli-Sforza, L. L. *The History and Geography of Human Genes*. (Princeton University Press, 1994).

102. Pääbo, S., Higuchi, R. G. & Wilson, A. C. Ancient DNA and the Polymerase Chain Reaction. **264,** 9709–9712 (1989).

103. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–95 (1977).

104. Ho, S. Y. W. & Gilbert, M. T. P. Ancient mitogenomics. *Mitochondrion* **10,** 1–11 (2010).

105. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. & Wilson, A. C. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312,** 282–4

106. Ermini, L. *et al.* Complete mitochondrial genome sequence of the Tyrolean Iceman. *Curr. Biol.* **18,** 1687–93 (2008).

107. Geigl, E.-M. On the circumstances surrounding the preservation and analysis of very old DNA. *Archaeometry* **44,** 337–342 (2002).

108. Höss, M., Jaruga, P., Zastawny, T. H., Dizdaroglu, M. & Pääbo, S. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Res.* **24,** 1304–7 (1996).

109. Millar, C. D., Huynen, L., Subramanian, S., Mohandesan, E. & Lambert, D. M. New developments in ancient genomics. *Trends Ecol. Evol.* **23,** 386–393 (2008).

110. Hebsgaard, M. B., Phillips, M. J. & Willerslev, E. Geologically ancient DNA: fact or artefact? *Trends Microbiol.* **13,** 212–220 (2005).

111. McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433,** 733–736 (2005).

112. White, T. D. *et al.* Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423,** 742–747 (2003).

113. Clark, J. D. *et al.* Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423,** 747–52 (2003).

114. Millard, A. R. A critique of the chronometric evidence for hominid fossils: I. Africa and the Near East 500-50 ka. *J. Hum. Evol.* (2008). doi:10.1016/j.jhevol.2007.11.002

115. Smith, T. M. *et al.* Earliest evidence of modern human life history in North African early Homo sapiens. *Proc. Natl. Acad. Sci.* **104,** 6128–6133 (2007).

116. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 5154–62 (2011).

117. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324,** 1035–44 (2009).

118. Cruciani, F. *et al.* A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am. J. Hum. Genet.* **88,** 814–818 (2011).

119. Soares, P., Rito, T., Pereira, L. & Richards, M. B. in *Africa from MIS 6-2. Population Dynamics and Paleoenvironments* (2016). doi:10.1007/978-94-017-7520-5_18

120. Lahr, M. M. & Foley, R. Multiple dispersals and modern human origins. *Evol. Anthropol. Issues, News, Rev.* **3,** 48–60 (2005).

121. Underhill, P. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65,** 43–62 (2001).

122. Richards, M. B., Bandelt, H.-J., Kivisild, T., Oppenheimer, S. & Bujnicki, J. M. A Model for the Dispersal of Modern Humans out of Africa. *Hum. Mitochondrial DNA Evol. Homo sapiens* **18,** 227–268 (2006).

123. Derricourt, R. Getting 'Out of Africa': Sea crossings, land crossings and culture in the Hominin migrations. *J. World Prehistory* (2005). doi:10.1007/s10963-006-9002-z

124. Carlson, A. E. The Younger Dryas Climate Event.

125. Carbonell, E. *et al.* The first hominin of Europe. *Nature* **452,** 465–469 (2008).

126. Klein, R. G. PALEOANTHROPOLOGY: Whither the Neanderthals? *Science (80-.).* **299,** 1525–1527 (2003).

127. Pinhasi, R., Higham, T. F. G., Golovanova, L. V & Doronichev, V. B. Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 8611–6 (2011).

128. Duarte, C. *et al.* The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 7604–9 (1999).

129. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530,** 429–433 (2016).

130. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507,** 354–7 (2014).

131. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).

132. Macaulay, V. *et al.* Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308,** 1034–6 (2005).

133. Van Andel, T. H. & Tzedakis, P. C. Palaeolithic landscapes of Europe and environs, 150,000-25,000 years ago: An overview. *Quat. Sci. Rev.* **15,** 481–500 (1996).

134. Mellars, P. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* (2006). doi:10.1038/nature04521

135. Benazzi, S. *et al.* Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* **479,** 525–528 (2011).

136. Higham, T. *et al.* The earliest evidence for anatomically modern humans in northwestern Europe. *Nature* **479,** 521–524 (2011).

137. Mellars, P. Archeology and the dispersal of modern humans in Europe: Deconstructing the 'Aurignacian'. *Evol. Anthropol.* **15,** 167–182 (2006).

138. Svoboda, J., Ložek, V. & Vlček, E. *Hunters between East and West*. (Springer US, 1996). doi:10.1007/978-1-4899-0292-4

139. Gamble, C., Davies, W., Pettitt, P. & Richards, M. B. Climate change and evolving human diversity in Europe during the last glacial. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **359,** 243-53–4 (2004).

140. Cruciani, F. *et al.* Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *Am. J. Hum. Genet.* **74,** 1014–22 (2004).

141. Verpoorte, A. Eastern Central Europe during the Pleniglacial. *Antiquity* **78,** 257–266 (2004).

142. Torroni, A. *et al.* A signal, from human mtDNA, of postglacial recolonization in Europe. *Am. J. Hum. Genet.* **69,** 844–52 (2001).

143. Achilli, A. *et al.* The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am. J. Hum. Genet.* **75,** 910–8 (2004).

144. Malyarchuk, B. *et al.* The Peopling of Europe from the Mitochondrial Haplogroup U5 Perspective. *PLoS One* **5,** e10285 (2010).

145. Price, T. D. & Bar-Yosef, O. The Origins of Agriculture: New Data, New Ideas. *Curr. Anthropol.* **52,** S163–S174 (2011).

146. Brown, T. A., Jones, M. K., Powell, W. & Allaby, R. G. The complex origins of domesticated crops in the Fertile Crescent. *Trends Ecol. Evol.* **24,** 103–9 (2009).

147. Kuijt, I. & Goring-Morris, N. Foraging, Farming, and Social Complexity in the Pre-Pottery Neolithic of the Southern Levant: A Review and Synthesis. *J. World Prehistory* **16,** 361–440 (2002).

148. Pinhasi, R., Thomas, M. G., Hofreiter, M., Currat, M. & Burger, J. The genetic history of Europeans. *Trends Genet.* **28,** 496–505 (2012).

149. Whittle, A. W. R. & Cummings, V. *Going over : the Mesolithic-Neolithic transition in north-west Europe*. (Published for the British Academy by Oxford University Press, 2007).

150. Burger, J. & Thomas, M. G. in *Human Bioarchaeology of the Transition to Agriculture* 369–384 (John Wiley & Sons, Ltd, 2011). doi:10.1002/9780470670170.ch15

151. Itan, Y., Powell, A., Beaumont, M. A., Burger, J. & Thomas, M. G. The Origins of Lactase Persistence in Europe. *PLoS Comput. Biol.* **5,** e1000491 (2009).

152. Barker, G. *The Agricultural Revolution in Prehistory: Why Did Foragers Become Farmers? Oxford: Oxford University Press* (2006).

153. Silva, F. & Vander Linden, M. Amplitude of travelling front as inferred from 14C predicts levels of genetic admixture among European early farmers. *Sci. Rep.* **7,** 31–34 (2017).

154. Rowley-Conwy, P. & Peter. Westward Ho! *Curr. Anthropol.* **52,** S431–S451 (2011).

155. Skoglund, P. *et al.* Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science (80-.).* **336,** 466–469 (2012).

156. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).

157. Ammerman, A. J. & Cavalli-Sforza, L. L. (Luigi L. *The neolithic transition and the genetics of populations in Europe*.

158. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201,** 786–92 (1978).

159. Dupanloup, I., Bertorelle, G., Chikhi, L. & Barbujani, G. Estimating the Impact of Prehistoric Admixture on the Genome of Europeans. *Mol. Biol. Evol.* **21,** 1361–1372 (2004).

160. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).

161. Anthony, D. W. *The horse, the wheel, and language : how Bronze-Age riders from the Eurasian steppes shaped the modern world*.

162. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6,** 1–8 (2015).

163. Sikora, M. *et al.* Population Genomic Analysis of Ancient and Modern Genomes Yields New Insights into the Genetic Ancestry of the Tyrolean Iceman and the Genetic Structure of Europe. *PLoS Genet.* **10,** (2014).

164. Myres, N. M. *et al.* A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* **19,** 95–101 (2011).

165. Balaresque, P. *et al.* A Predominantly Neolithic Origin for European Paternal Lineages. *PLoS Biol.* **8,** e1000285 (2010).

166. Koch, J. T. & Cunliffe, B. W. *Celtic from the West 2 : rethinking the Bronze Age and the arrival of Indo-European in Atlantic Europe*.

167. Pereira, L. *et al.* High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* **15,** 19–24 (2005).

168. Rando, J. C. *et al.* Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.* **63,** 413–28 (1999).

169. Behar, D. M. *et al.* The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. *Am. J. Hum. Genet.* **78,** 487–497 (2006).

170. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534,** 200–205 (2016).

171. Modi, A. *et al.* Complete mitochondrial sequences from Mesolithic Sardinia. *Sci. Rep.* **7,** 42869 (2017).

172. Mathieson, I. *et al.* The Genomic History Of Southeastern Europe The Genomic H istory of Southeaste rn Europe. *bioRxiv Prepr. first posted online May. 9, 2017; doi http//dx.doi.org/10.1101/135616* (2017). doi:10.1101/135616

173. Szécsényi-Nagy, A. Molecular genetic investigation of the Neolithic population history in the western Carpathian Basin. (Johannes Gutenberg University Mainz, Germany, 2015).

174. Olalde, I., Reich, D. & et al. The Beaker Phenomenon and the Genomic 2 Transformation of Northwest Europe. *BioRxiv Prepr. Serv. Biol. - Cold Harb. Spring Lab.* 1–28 (2017). doi:10.1038/NMAT3123

175. Hervella, M., Izagirre, N., Alonso, S., Fregel, R. & de-la-Rúa, C. Early Neolithic funerary diversity and mitochondrial variability of two Iberian sites. *Archaeol. Anthropol. Sci.* **8,** 97–106 (2016).

176. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536,** 419–424 (2016).

177. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342,** 257–61 (2013).

178. Mittnik, A. *et al.* The Genetic History of Northern Europe. *bioRxiv, 113241* 1–26 (2017). doi:10.1101/113241

179. Gamble, C. Human display and dispersal: A case study from biotidal Britain in the Middle and Upper Pleistocene. *Evol. Anthropol. Issues, News, Rev.* **18,** 144–156 (2009).

180. Pereira, L., Prata, M. J. & Amorim, a. Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann. Hum. Genet.* **64,** 491–506 (2000).

Supplementary data

The following data is annexed in digital format, for the sake of brevity:

Supplementary File 1 (.xls)

Probability of entry time for each founder clade in multiple migration model.

Supplementary File 2 (.xls)

Discriminated founder analysis results, including ρ-estimated ages.

Supplementary File 3 (.pdf)

List of published samples used in this work.