Inferring recent human population history from a Y chromosome perspective

Neus Solé Morata

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI

Dr. Francesc Calafell / Dr. David Comas

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Als meus pares
A Iñigo
A l'Unai

If you want to go fast, go alone. If you want to go far, go together.

African Proverb

Acknowledgments

Abans de tancar aquest capítol de la meva vida, m'agradaria dedicar unes paraules a tots els que m'heu ajudat a arribar fins aquí. No només a aquells que m'heu fet costat i acompanyat en aquest camí acadèmic, sinó també als que heu format part de la meva vida i m'heu donat el suport i l'energia necessària quan m'han faltat. M'agradaria dir-vos moltes més coses de les que he escrit en aquestes línies, però es faria molt més llarg del que ja és. Així que espero haver-vos demostrat en persona que, sense vosaltres, res d'això no hauria estat possible.

En primer lloc, vull de dedicar unes paraules d'agraïment a qui més m'ha guiat en aquest camí. Francesc, han passat més de 7 anys des que vam començar a treballar plegats i només tinc paraules d'agraïment per a tu. Has estat un gran director de tesi, però també un professor del que he après moltíssim. Gràcies per tenir paciència amb les meves decisions quan ho he necessitat, però sobretot gràcies per valorar la meva feina dia rere dia. David, moltes gràcies a tu també. Tens la capacitat d'escoltar i aportar solucions als problemes. No és fàcil trucar a la porta del teu *jefe*, però tu ho fas senzill. Moltes gràcies per la teva confiança al oferir-me un projecte amb el que he après moltíssim. Elena, a tu també t'he de donar les gràcies. Potser si t'haguessis limitat a dir-me que no tenies lloc en el teu grup, en comptes de buscar-me'n un altre, ara no seria aquí. Finalment, un agraïment cordial a en Ferran, a en Carles i a l'Òscar, per confiar en mi mostres tan delicades.

Gràcies a tots els membres de l'IBE, tant als que encara hi sou, com als que ja heu marxat. Aquests anys al vostre costat m'han ajudat a créixer tant a nivell acadèmic com a nivell personal. Gràcies especialment a en Juan, en Diego, en Nino, en Txema, en Javi, en Fede, en Guillem, en Lukas i la Clàudia, per tants moments divertits, tant dins com fora del laboratori. També gràcies a la Marina, al Marco i a la Irene per les vostres visites al 412.04, tant inesperades com indispensables. A Raquel, que mejor manera para olvidarse del estrés que compartir una clase de *spinning* con una buena amiga. Gracias por todos esos buenos ratos que tanto echo de menos. Mònica, gràcies per cuidar-me tant, espero que, tot i la distància, ens seguim cuidant l'una a l'altra. I a tu Judith, gràcies per ajudar-me a resoldre tota classe de problemes i per aquesta paciència infinita que tens. Ets imprescindible.

Després de casi 7 anys en el mateix despatx, he tingut la oportunitat de compartir el meu dia a dia amb un grapat de persones de les que he après molt més del que queda recollit en una tesi. Al Marc, gràcies per encomanar-me les ganes de fer un doctorat i confiar experiments molt delicats en una estudiant de tercer de carrera. A Koldo, gracias por escucharme, por animarme y por tantos buenos consejos. Gràcies també a l'Ignasi, a la Maria i a l'Alicia. Amb vosaltres he compartit la major part d'aquesta etapa i m'és impossible recordar alguns moments al despatx sense que se'm dibuixi un somriure a la cara. Heu estat els millors companys d'aventura que podia trobar. Finalment, vull agrair-li al nostre últim fitxatge la seva energia i positivitat. Carleta, ens vam conèixer fa molt poc, però vaig compartir amb tu moments molt decisius de la

meva vida. Gràcies per escoltar-me i preocupar-te per mi, però sobretot, gràcies per seguir-ho fent en la distància.

Vull agrair també la seva ajuda a la Núria i al Roger. No només m'heu proporcionat suport tècnic, sinó que també m'heu ajudat a encaixar els desenganys del món experimental i a buscar solucions als problemes que anaven sorgint. Gràcies també al Max, al Carlos, al Salva i al Dani, ha estat un plaer treballar al vostre costat i, tot i que el *grill* ens ha fet patir bastant, la vostra companyia ho compensava. A tu Dani, m'alegro que la vida hagi tornat a creuar els nostres camins i puguem seguir gaudint d'una bona amistat a l'altre costat de l'atlàntic.

Moltes gràcies a l'Anna, la Jèssica, l'Isaac, la Irina, la Miruna i la Neus... els meus *mindundis*. Amb vosaltres vaig aprendre que ensenyar el que hem après és igual o més important que aprendre-ho. Sobretot tu Neus, gràcies per la teva ajuda dins i fora del laboratori. Tot i la feina, sempre has intentat trobar un moment per donar-me un cop de mà. Estic convençuda que arribaràs molt lluny.

Vull agrair especialment la seva ajuda als meus companys de grup. Aquest bon rollo i complicitat que tenim són difícils de trobar. Al Gerard, moltes gràcies pel teu suport bioinformàtic i la teva predisposició a ajudar. A l'Àlex, gràcies per la teva curiositat científica, per implicar-te tant en els problemes que m'anaven sorgint i intentar aportar solucions. André, Simone, cuando he necesitado merienda o quejarme un rato a partir de las 7 de la tarde, allí estabais vosotros y vuestro buen humor. Muchas gracias. Y a ti Lara, una de las piezas claves de esta tesis. Siempre dispuesta a

echarme una mano o animarme cuando lo necesitaba. Pero además de una buena compañera, en ti encontré a una gran amiga. Junto con Núria, hemos vivido momentos increíbles y varias anécdotas que me encanta recordar. Moltes gràcies a les dues per fer aquest camí molt més agradable.

Finalmente, uno de los mejores regalos de este doctorado has sido tú, Jéssica. Eres como una hermana para mí y una de las personas más fuertes que he conocido. Gracias por tus consejos, incluso a altas horas de la madrugada, y por entenderme como nadie. Pero sobretodo, gracias por saber estar tan cerca a pesar de estar tan lejos.

En un terreny més personal vull agrair-li a la Júlia una amistat que resisteix el pas del temps i la distància. Gràcies per trucar-me quan feia temps que no sabies res de mi. A Carolina, mi vecina y mejor amiga durante muchos años. Eres la persona más alegre y positiva que he conocido, gracias por tantos buenos ratos. A les nenes de l'handbol, sobretot a l'Anne Laure, la Núria, l'Anna, la Sílvia i la Mariona. Gràcies per la vostra amistat, tant dins com fora de la pista. Ens vam fer grans juntes i sou les protagonistes d'una època que mai oblidaré. Gràcies també al Cristian, al Sergi, al Renato i a l'Isma per les hores compartides al pavelló, però també per les nostres escapades al refugi de Planoles. Finalment, gràcies al Ricard, al Roy, al Jordi, al Marcos, al David, la Marta, la Cris i la Lorena per les nits a *La Sidreria*, els carnavals i les nostres escapades d'estiu en busca de festes majors.

La biologia no només em va obrir les portes del doctorat, també em va aportar un grapat de bons amics que han estat indispensables aquests últims anys. Joan Pau, el psicòleg de la psicòloga, gràcies per ajudar-me a relativitzar quan ho he necessitat, sempre aconsegueixes que els problemes semblin menys problemes. A Irene, vivir bajo el mismo techo no solo no estropeó nuestra amistad, sino que la hizo más fuerte. I a tots dos, gràcies per tants moments increïbles al JIN. Vaig deixar de viure amb la meva família, per formar-ne una de nova. Carlos, tens la capacitat de convertir un dinar avorrit de dilluns en un moment inoblidable. Escoltar-te és un dels meus *hobbies* favorits. I a tu Joan, una de les persones més divertides que conec, gràcies per arrancar-me un somriure fins i tot el dia més gris. Marcel, el *rockstar* amb el cor més gran del món, m'encanta que m'ajudis replantejar-me tantes i tantes coses. Aida, vas marxar massa aviat i no vaig poder gaudir prou de tu i de la teva capacitat organitzativa. Però la distància no ha estat cap impediment per acostar-nos una mica més. Gràcies també a l'Alba, la *mama* del nostre clan Heredia, si formo part d'aquest grup de persones meravelloses, és gràcies a la teva insistència i ganes de conèixer-me. A la Clara, sóc increïblement feliç de saber que has trobat el teu lloc. Et mereixes totes les coses bones que et passin i totes les olives que puguis desitjar. A l'Ulises, ets únic i irrepetible, gràcies per remoure'm la consciència i, qui sap, potser al final acabaràs fent *vegana* a la filla d'una carnissera. I a tu Maria, crec que tenim moltíssimes coses en comú i si algun dia aconseguim tornar a viure a la mateixa ciutat, segur que enfortirem una amistat que sempre he valorat molt. Moltes gràcies al nostres

dos no-biòlegs del grup. Sergi, Cris, la vostra incorporació a la colla, no només ha estat un èxit, sinó que l'ha fet encara més increïble i meravellosa del que ja era. Gràcies també a l'Albert, tot i que no ens veiem tan sovint com m'agradaria, sóc molt feliç quan ens retrobem en alguna data assenyalada. I a tu Marc, gràcies pels consells i per escoltar-me sempre, sigui la hora que sigui. La complicitat que tenim costa de trobar i em sento molt afortunada de seguir-te sentint a prop tot i la distància. Maura, tot i que fa temps que ens separa la distància, m'alegra seguir compartint amb tu els moments feliços i ajudant-nos quan les coses es torcen. A tu Karen, has marcat un abans i un després a la meva vida. No només hem viscut històries increïbles dignes de ser novel·lades, sinó que hem sabut donar-nos el suport i l'energia necessària per superar les dificultats que anaven sorgint. I ara més que mai, em sento molt afortunada de tenir-te al meu costat. Gràcies també a la Marta, la teva espontaneïtat em va robar el cor des del primer dia i, ràpidament, ens vam fer bones amigues. Estic convençuda que superaràs cada dificultat, i espero ser a prop per cuidar-te, tal i com tu m'has cuidat aquests últims mesos. I a tu Laura, gràcies per una amistat que creix cada dia que passa, fins i tot a més de 5,000 km de distància. Gràcies per compartir amb mi cada problema, cada preocupació, però sobretot, gràcies per alegrar-te com ningú per cada cosa bona que m'ha passat. I finalment, moltes gràcies a TOTS per estimar-me tant, m'heu fet millor persona, però també m'heu ajudat a estimar-me més a mi mateixa. Quan un grapat de persones tan increïbles t'estima, deu ser que t'ho mereixes.

A la meva família vull agrair-li haver-me donat les forces i el suport necessari per arribar fins aquí. Als meus tiets: Albert, Montse, Josep, Carles, Susana, Aurora i Francis, gràcies pel grapat de moments feliços que he viscut al vostre costat. Gràcies sobretot al meu padrí, aquest interès que demostres per la meva feina m'ajuda a estimar molt més el que faig. Gràcies a l'Enric, al Carles i al Francesc, amb vosaltres he compartit alguns dels moments més feliços de la meva infància i, encara és avui, que retrobar-nos un cop l'any a l'hort de l'avi és un dels meus dies preferits. Gràcies també a la Marta, la Núria i l'Alba, sempre us he considerat més unes germanes que unes cosines i, cada cop més, unes molt bones amigues. Segurament per això us trobo tant a faltar. Finalment, moltes gràcies als meus avis pel seu amor incondicional i per tant de temps invertit en fer-me feliç. He après moltíssimes coses de vosaltres i em sento afortunada d'haver-vos pogut gaudir durant molts anys. Al meu avi Joan, gràcies per ensenyar-me que encara que la vida ens jugui males passades i ens aparti de qui més estimem, no hem de perdre mai el somriure i les ganes de viure. A l'avi Pere, gràcies per fer-me tant feliç. El moment més dur de la meva vida va ser haver-te de dir adéu i, quasi 5 anys més tard, encara no m'he acostumat a la teva absència. M'encantaria haver vist la teva cara al saber que series besavi. I a tu iaia, més que una àvia has estat com una mare per mi, potser per això m'agrada tant instal·lar-me a casa teva cada cop que puc. Tampoc no perdo mai la oportunitat de presumir d'àvia. Ni el dolor més intens atura les teves ganes de viure i d'aprendre coses noves. M'has ensenyat que els límits només ens els posem nosaltres. Sé que no podré agrair-te mai

tot el que has fet per mi, però aprofito aquestes línies per fer-te saber com t'estimo.

A mi familia vitoriana, gracias por todo el cariño que me habéis dado. Des del primer día, tuve la sensación de haber encontrado una nueva familia a más de 500 km de distancia y cada día que pasa lo tengo más claro. Muchas gracias a Lurdes y Honorio por acogerme en vuestra casa verano tras verano y por tratarme como a una hija. Espero compensároslo con muchas alegrías como las de estos últimos dos años. Os quiero mucho.

Al Pere, la teva arribada va ser un somni fet realitat. Ets el millor germà que algú podria desitjar i, últimament, t'has convertit en un molt bon amic. Sovint em penedeixo de no haver estat prou al teu costat, potser la diferència d'edat ens va allunyar abans del compte. Sé que a vegades el camí fa pujada, però si t'esforces i lluites trobaràs el teu lloc i , un cop el trobis, brillaràs. T'estimo moltíssim. Gràcies també a la Patri, la millor companya que hauries pogut trobar. En poc temps t'has convertit en una més de la família i els nostres interessos similars ens han apropat encara més. Espero poder-te gaudir durant molt de temps.

Als meus pares, els veritables i últims responsables d'aquesta tesi. Gràcies per la vostra confiança i suport, sou els culpables de que mai m'hagi sentit incapaç de res. A la meva mare, gràcies per ser-hi sempre; quan et necessito i quan crec que no necessito a ningú. Ets dolça, atenta i divertida però sobretot, forta i valenta com ningú. Em conformaria en ser la meitat de bona mare que tu. Al meu pare, gràcies per ensenyar-me a no conformar-me, a exigir-me més, però

també gràcies per ajudar-me a demanar ajuda quan has vist que la necessitava. El millor èxit és que et sentis orgullós de mi. Tot i que a vegades em costi demostrar-ho, sabeu de sobres que us estimo amb bogeria.

I per acabar, vull dedicar unes paraules a les dues persones més importants de la meva vida ara mateix. Iñigo, conocerte ha sido lo mejor que me ha pasado en la vida y por eso quiero dar las gracias a todas las casualidades que hicieron que nos cruzáramos en esta vida. Solo tú sabes sacar lo mejor de mí, me das la fuerza necesaria para superar cada dificultad y me haces sentir capaz de todo. Y por todo esto y mucho más, te quiero como no sabía que se podía llegar a querer. Hace casi un año nos embarcamos en una gran aventura, que no solo no nos ha alejado, sino que nos ha unido aún más. Y el éxito de esta aventura, nos dio las fuerzas y el valor necesario para embarcarnos en un nuevo proyecto, aún más arriesgado. I és a aquest petit però gran projecte bostonià, a qui vull agrair-li el seu suport inconscient i desinteressat. Unai, no oblidaré mai la primera vegada que vaig sentir el teu cor bategar impacient dins meu. Però encara em fa més feliç pensar que aquest és només el primer de molts moments inoblidables que viuré al teu costat. I qui sap, potser d'aquí uns anys, quan llegeixis aquestes línies entendràs com, sense saber-ho, m'has donat l'energia necessària per escriure cada línia d'aquesta tesi, perquè amb cada paraula que escrivia sentia que era més a prop de tenir-te entre els meus braços.

Neus Solé Morata

Abstract

The Y chromosome is the longest non-recombining DNA sequence of the human genome. This avoidance of recombination, together with its paternal inheritance, makes the Y chromosome a powerful tool with which to study population history, male genealogy, forensics and medical genetics. Besides the progress made in the field during the last two decades, the recent advent of massive parallel sequencing (MPS) has yielded the discovery of thousands of new variants that have allowed to build a more reliable phylogeny and to obtain direct estimates of its mutation rate. In the present thesis, I analyse the Y-chromosome diversity from two different perspectives and with different purposes. First, by targeting specific SNPs and STRs in ~ 2,500 men bearing one of the selected 50 Catalan surnames, we investigated the driving forces behind the origin, systematization, and diffusion of surnames. And then, by using whole Y-chromosome sequences from North African individuals belonging to the most frequent lineage in the area (E-M183), we have been able to refine the phylogeography of this lineage and to shed light on the controversial dates for its origin and spread.

Resum

El comportament únic del cromosoma Y, heretat per via paterna sense patir recombinació amb cap altre cromosoma, el converteix en un marcador excepcional amb aplicacions en àmbits com la genètica de poblacions humanes, la genealogia o la genètica forense. Tot i el progrés en l'estudi del cromosoma Y realitzat en les últimes dues dècades, el recent desenvolupament de les tecnologies de seqüenciació massiva ha permès el descobriment de milers de noves variants, mitjançant les quals s'ha obtingut una millor reconstrucció filogenètica, així com una estimació directa de la seva taxa de mutació. En aquesta tesi s'analitza la diversitat del cromosoma Y des de dues perspectives diferents i amb els següents propòsits. En primer lloc, mitjançant el genotipat de marcadors específics del cromosoma Y en ~2500 homes portadors d'un dels 50 cognoms catalans escollits, s'han investigat els processos que han donat lloc a l'origen, la sistematització i la difusió dels cognoms. Per altra banda, la seqüenciació de cromosomes Y en homes nord africans pertanyents al llinatge més freqüent en aquesta àrea (E-M183), ha permès un refinament de l'estructura filogeogràfica d'aquest llinatge, així com l'establiment temporal del seu origen i dispersió.

Preface

In the last two decades, there has been a keen interest in using variation on the NRY to examine questions about human population history. Since the discovery of the first molecular polymorphism in 1985, the number of Y-chromosome variants has accumulated at an increasing pace, and especially since the advent of massive parallel sequencing technologies. Nowadays, the characterization of Y-chromosome diversity has its application in DNA forensics, genealogical reconstruction, medical genetics and human evolutionary studies.

The use of Y chromosome in a genealogical frame has had different applications, such as the identification of historically relevant remains, the confirmation of genealogical relatedness, and the estimation of false-paternity rates. Within this context, the Y chromosome has also been used to learn about patrilineal surnames, suggesting that because of its paternal inheritance, they might be acting as cultural markers for shared ancestry. By the time when this thesis was conceived, surname systems have only been analysed in Britain and Ireland, with different results. To better understand the relationship between Y chromosome and surnames, we have investigated the driving forces behind the origin, systematization, and diffusion of surnames by using a set of 50 Catalan surnames. Catalan surnames are based on the Catalan language and diffused throughout the Catalan-speaking lands. They showed a higher diversity compared with Spanish surnames and their systematization was marked by the Council of Trent (1545–

1563). Our results support the idea that surname frequency might be explained in terms of polyphyletism.

As mentioned above, the Y chromosome has also been used to study demographical events, such as population origin or migration, from a paternal perspective. Taking advantage of the advent of MPS technologies, we aim to infer on the history of North African populations using whole Y-chromosome sequences. North African populations are distinct from Sub-Saharan Africans both culturally and genetically speaking. However, the time and the extent of genetic divergence between North African and Sub-Saharan populations are not fully understood. To gain a more complete understanding of human population movements in this region, we have focused on a particular North African lineage: E-M183 (E-M81). Due to its pattern of distribution, this lineage has been suggested as the autochthonous North African Y chromosome. By inferring the origin and spread of E-M813, we aimed to shed some light on the origin and demography of North African populations.

Index

Abbreviations

A: Adenine

AMOVA: Analysis of molecular variance

bp: Base pairs

BCE: before Common Era

BWA: Burrows-Wheeler Alignment

CE: Common Era

C: Cytosine

DNA: Deoxyribonucleic acid

G: Guanine

GATK: Genome analysis toolkit

IBD: Identity by descent

IBS: Identity by state

ISOGG: International Society of Genetic Genealogy

kb: kilo bases

kya: thousand years ago

Mb: Megabases

MJ: Median-joining

MP: Maximum parsimony

mtDNA: Mitochondrial DNA

MSY: Male specific region of the Y chromosome

MPS: Massive Parallel Sequencing

NRY: Non-recombining region of the Y chromosome

OTU: operational taxonomic units

RNA: Ribonucleic acid

SC: Sequence capture

SNP: Single nucleotide polymorphism

STR: Short tandem repeat

PAR: Pseudoautosomal region

PCA: Principal component analysis

PCR: Polymerase chain reaction

PPPG: per position per generation

T: Thymine

TMRCA: Time to the most recent common ancestor

UEP: Unique event polymorphism

WGS: Whole genome sequencing

ya: Years ago

YCC: Y chromosome consortium

1. INTRODUCTION

In the following section I will introduce the main features of the Y chromosome (section 1.1). Then, I will describe one of the applications of studying Y-chromosomal diversity: surname studies (section 1.2). Finally, I will review North African Population history and demography (section 1.3).

1.1. The human Y chromosome

The human genome is divided into 46 chromosomes. Males and females share a set of twenty-two pairs of these molecules, known as autosomes, but differ in the remaining pair of DNA segments: the sex chromosomes. While females carry two copies of the X chromosome, males have only one copy of it, but carry instead a much smaller piece of DNA known as the Y chromosome (1).

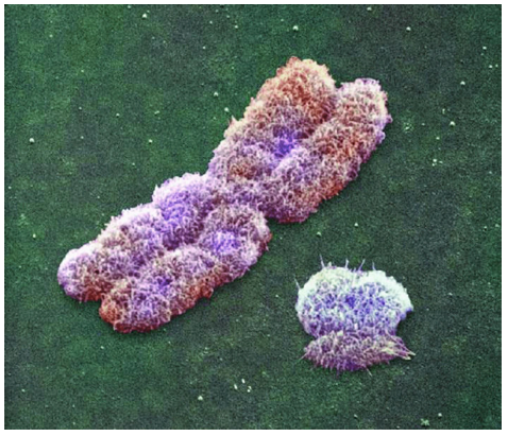


Figure 1. Electron micrograph of a human Y chromosome (right), in comparison to a human X chromosome (left) (2).

The Y chromosome is much shorter than the X, being only 2–3% of the haploid genome (Figure 1). Nevertheless, besides its small size, the Y chromosome has an important role: it determines the male sex. Only males possess the Y chromosome and thus, it is paternally inherited (3).

The sex chromosomes have evolved from a homologous ancestral pair of autosomes within the last 300 million years (4). During this process, the Y chromosome suffered a progressive degeneration that led to a reduction in size and to a loss of sequence identity with the X chromosome. Due to the sequence divergence of the sexual chromosomes, the recombination between both chromosomes (X and Y) is avoided. This is true for more than 95% of the Y-chromosome length, a region known as the non-recombining portion of the Y (NRPY or NRY). Only two short regions of the Y chromosome, the pseudoautosomal regions (PAR1 and PAR2), show sequence identity with the X chromosome and thus, behave like an autosome and recombine during meiosis.

The pseudoautosomal region 1 (PAR 1) comprises a region of 2.6 megabases (Mb) and is located at the tips of the short arms of both X and Y chromosomes. This region derives from the ancient mammalian sex chromosomes and it is required for the pairing of the X and Y chromosome during male meiosis. On the other hand, PAR 2, a much shorter region (330 kilobases (kb)) lying at the tips of the long arms of the sex chromosomes, appears to be a recent evolutionary acquisition of humans; it does not have an important role in chromosome segregation and thus, is not necessary for

fertility (5). The progressive degeneration of the Y chromosome not only has led to a reduction in size, but also in terms of gene content. While the X chromosome contains on the order of thousands of genes, the Y chromosome has kept only about a dozen and most of them are involved in male fertility. The consequence of having only one X chromosome is that males are hemizygous for genes on this chromosome. As a result, disease-causing alleles on such chromosome will behave differently in males and females: the phenotype associated with a recessive allele at an X-linked gene (on the X chromosome) will always be manifested in males with that allele. Example of this fact are the recessive alleles that cause haemophilia. Such X-linked recessive illnesses will be, therefore, more frequent in males than in females (3,6).

As mentioned above, the non-recombining region of the Y chromosome, also known as the male specific region (MSY), comprises 95% of the chromosome length. This region, flanked on both sides by the pseudoautosomal regions, is a mosaic of heterochromatic sequences and three classes of euchromatic sequences: X-transposed, X-degenerated and ampliconic (Figure 2a). The MSY includes at least 156 transcription units, all of them distributed along the 23 Mb of euchromatic sequences, half of which probably encode proteins. Of these, about 60 belong to nine different MSY-specific gene families, and the remaining are single-copy genes. Overall, the MSY encodes at least 27 distinct protein or protein families. Regarding the patterns of tissue expression of these 27 distinct proteins or protein families identified so far, 12 are expressed ubiquitously and 11 are expressed exclusively or

predominantly in testes (6).

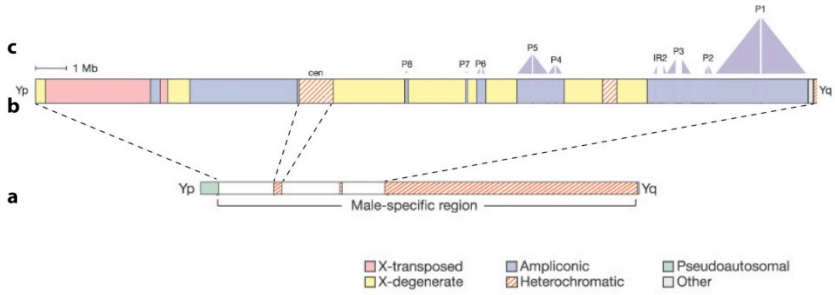


Figure 2. a) Schematic representation of the whole Y chromosome, including the pseudoautosomal and heterochromatic regions. b) Enlarged view of a 24-Mb portion of the MSY, extending from the proximal boundary of the Yp pseudoautosomal region to the proximal boundary of the large heterochromatic region of Yq. Shown are three classes of euchromatic sequences, as well as heterochromatic sequences. A 1-Mb bar indicates the scale of the diagram. c) Triangles denote sizes and locations of arms of eight palindromes (P1–P8) and of IR2 inverted repeats (whose arms exhibit 99.95% identity). Gaps between opposed triangles represent the non-duplicated 'spacers' between palindrome arms. Modified from Skaletsky *et al*. (6)

The X-transposed sequences consist in two non-contiguous DNA segments, originated from a massive X-to-Y transposition that occurred after the divergence of the human and chimpanzee lineages (3-4 million years ago) (Figure 2b) (7–9). As a result, these sequences exhibit a 99% identity to DNA sequences in Xq21, a band in the long arm of the human X chromosome. Despite the identity to the X chromosome, these X-transposed sequences do not participate in X-Y crossing over during male meiosis, which distinguishes them from the pseudoautosomal regions mentioned above. In addition, with only two genes in a length of 3.4 Mb, the X-transposed region exhibits the lowest density of genes among the three types of euchromatic sequences. By contrast, these sequences

exhibit the highest density of interspersed repeat elements, more specifically, long interspersed element 1 (LINE 1) elements account for 36% of all X-transposed sequence (6). Recent work claimed that the allelic unequal recombination between the X-transposed region Yp11.2 and Xq21.3 strongly supports the presence of a new pseudoautosomal region, named PAR3 (4).

While the X-transposed sequences are characterized by low gene density, the X-degenerate segments of the MSY contain DNA sequences that display between 60% and 96% nucleotide sequence identity to 27 X-linked genes (Figure 2b). Probably, these sequences are surviving remnants of the ancient pair of autosomes, from which the sex chromosome originated, as mentioned above. In 14 cases, the MSY homologue is a single-copy gene that seems to be transcribed, and whose product is a protein very similar to those transcribed by the X-linked gene. In the remaining cases, the MSY homologue is a pseudogene. Overall, the X-degenerate sequences encode 16 of the MSY 27 distinct proteins or protein families. Regarding the pattern of expression, while all 12 ubiquitously expressed MSY genes mentioned above, are located in the X-degenerate regions, only 1 out of 11 genes expressed in testes, the sex determining SRY, is X-degenerate (6).

Finally, the ampliconic segments are composed by large sequences that exhibit a high level of sequence identity (>99%) to other sequences in the MSY (Figure 2b). These long, MSY-specific repeat units, known as amplicons, are dispersed across the euchromatic long arm and proximal short arm and its combined

length is 10.2 Mb. In addition, the ampliconic sequences exhibit the highest density of genes among the three sequence classes in the MSY euchromatin. Overall, considering together both coding and non-coding elements, the ampliconic sequences contain 135 of the 156 MSY transcription units identified so far. Furthermore, while X-degenerated genes were mostly ubiquitously expressed, all nine protein-coding families in the ampliconic regions are expressed in testes. Finally, among the three euchromatic classes, the ampliconic regions exhibit the lowest density of interspersed repeat elements (6).

Noticeably, one quarter of the euchromatic region of the Y chromosome is occupied by eight massive palindromes, which are located within the ampliconic regions of the long arm (Yq). An MSY map highlighting all eight palindromes is shown in figure 2c. The palindromes are long, with their arms ranging from 9 kb to 1.45 Mb in length, and they show an arm-to-arm nucleotide identity of about 99.9 %. The most spectacular is palindrome P1, which shows an arm-to-arm identity of 99.97% and spans 2.9 Mb. The eight palindromes collectively comprise one-quarter of the MSY euchromatin. Six of the eight palindromes carry recognized protein-coding genes that appear to be expressed in testes. Noticeably, all genes located in the MSY palindrome sequence show identical gene copies on the opposite arm of the palindrome. As we mentioned above, nine multi-copy protein-coding gene families have been identified in the MSY, eight of which have members on palindromes. In fact, six families are located exclusively in palindromes (the DAZ and the CDY genes). In addition, the

palindromes contain at least seven families of apparently non-coding transcription units, all expressed exclusively or predominantly in testes (6).

1.1.1. Y-chromosome diversity

Both uniparental inheritance and avoidance of recombination of the male specific region of the Y chromosome, make it a superb tool for investigating human evolution from a male perspective. Y chromosomes pass almost intact from generation to generation and thus preserve a simpler record of their history. Understanding the processes that shaped Y-chromosome diversity is crucial for reconstructing population history (10).

1.1.1.a. Mutation

A mutation is any change in DNA sequence and is the only process introducing new alleles. Thus, it provides the fuel for evolutionary processes. Mutations include a broad array of events, which can range from the substitution of a single base in the genome, to the translocation of chromosomal segments and even changes in chromosomal number. However, not all mutations are passed to the next generation and contribute to evolutionary change. To do so, they must occur in the germ line and be survivable and compatible with fertility (3).

Although mutation is the cause of introducing new variation, the word *mutation* itself is mainly restricted to variation causing disease, while *polymorphism* is used to describe a sequence

difference with no apparent effect on function. Alternatively, a polymorphism has been defined as any variant with a minor allele frequency ≥1%. (3,10).

Mutations on the Y chromosome are typically neutral and constitute the only source of variation for its non-recombining part. These make the Y chromosome the best molecular clock in the human genome (11). To calibrate this clock, we must know the rate at which these changes occur. Mutation rates are widely used to estimate the time to the most recent common ancestor (TMRCA) of a group of related Y chromosomes (11) (section 2.3.4). Finally, as we will discuss below, each type of mutation has its own mutation rate.

When studying the Y chromosome from an evolutionary perspective, two common types of variation are typically used: Single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs). Since the discovery of the first molecular polymorphism in 1985 (1), the number of Y-chromosome polymorphisms has accumulated at an ever-increasing rate. Hundreds of microsatellites have been identified from the reference sequence. Genome sequencing projects have led to the discovery of many thousands of SNPs, although not all of them have been validated or placed on the Y phylogeny (12).

Single nucleotide polymorphisms

Single nucleotide polymorphisms are the simplest difference between two homologous DNA sequences. They consist on a base

substitution, in which one base (A, C, G, T) is exchanged for another. If the change occurs between a purine and another purine (for example, A for G), or between two pyrimidines (for example, C for T), it is called transition. However, when a pyrimidine is exchanged for a purine, or viceversa, it is called transversion. The insertion or deletion (indel) of one or a few bases is often pooled with SNPs in the category of binary polymorphisms (1), since both forms of variation have often exactly two alleles.

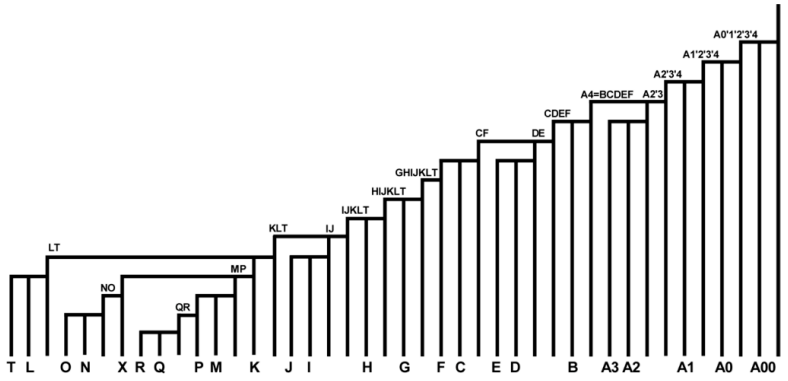


Figure 3. Phylogenetic tree illustrating the relationships of the major Y-chromosomal haplogroups. Branch lengths are not proportional to mutational differences. Adapted from van Oven *et al.* (13).

Recently, the 1000 Genomes project has reported about 65,000 SNPs and short indels on the Y chromosome (14). While some Y-SNPs may have an ancient origin and are shared by most men, others appear to be more recent and are geographically distributed. But each of these Y-SNPs first originated at some branch of the genealogical tree that unites all human Y chromosomes (15). As a result, binary polymorphisms or indels can easily be combined into haplotypes, known as haplogroups, which can be related by a single

phylogeny (section 2.3.3). The major clades of the Y-chromosomal tree are labelled from A to T (Figure 3), and each subclade is further subdivided into alphanumerically named subclades (3).

However, as the number of variants increases, the length of names does so. As a result, an alternative recently adopted is to use the name of the derived variant furthest from the root, the advantage of this alternative nomenclature is that variant names are stable, while new, phylogenetically intermediate SNPs are discovered (1). It is also worth mentioning that haplogroups show a geographical structure, thus they can be used to study population history and demography, as will be discussed below (section 1.1.2) (10). The global distribution of haplogroups is displayed in figure 4.

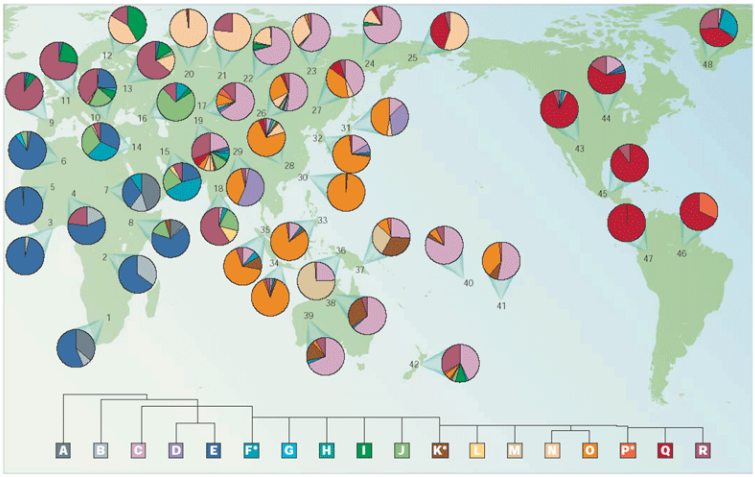


Figure 4. Global distribution of Y haplogroups. Each circle represents a population sample with the frequency of the 18 main Y haplogroups identified by the Y Chromosome Consortium (YCC) indicated by the coloured sectors (10).

In 2002, the Y Chromosome Consortium (YCC) published a single parsimony tree showing the relationships among 153 haplogroups

based on 243 binary markers and conceived a standardized nomenclature system to name lineages nested within this tree (16). Subsequently, Jobling and Tyler-Smith (10) published a modified version of the YCC tree. Then, Karafet *et al.* (17) published an extensively revised Y chromosome tree containing 311 distinct haplogroups, including two new major haplogroups (S and T), and incorporating 586 binary markers. After that, the application of new sequencing technologies (www.1000genomes.org) has yield the discovery of thousands of new markers (14). A maximum-likelihood phylogenetic tree using 60,555 biallelic SNPs derived from 10.3 Mb of accessible DNA is displayed in figure 5.

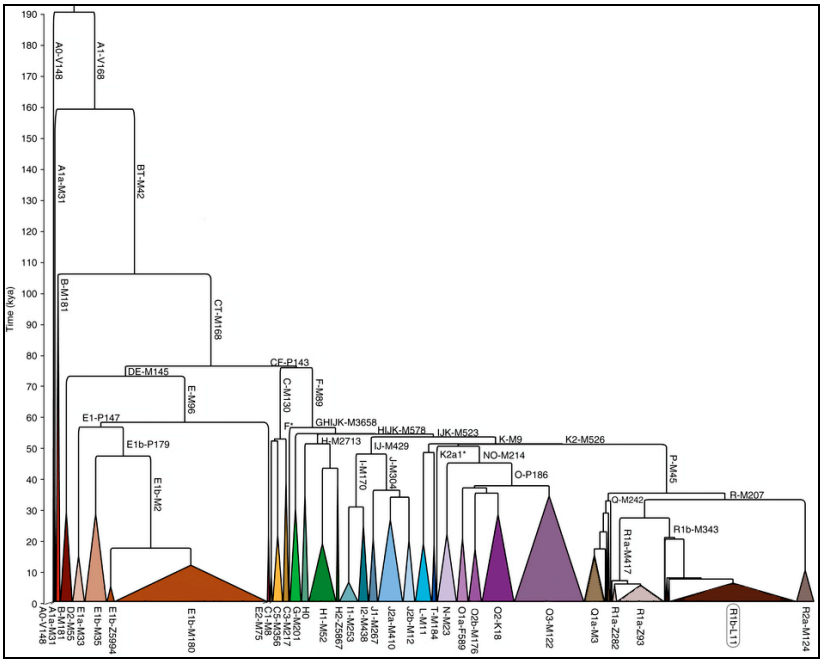


Figure 5. Y-chromosomal phylogenetic tree. Branch lengths are drawn proportional to the estimated times between successive splits, with the most ancient division occurring ~190 kya. Colored triangles represent the major clades, and the width of each base is proportional to one less than the corresponding sample size. Modified from Poznik *et al*. (18).

Finally, it is also worth mentioning the role of the International Society of Genetic Genealogy (ISOGG) in maintaining and updating the Y-chromosomal tree. Recent technological advances have enabled an extensive research on Y chromosome, yielding to continuous changes in the structure of the Y phylogeny. ISOGG aims to keep the tree updated with the latest developments in the field, but maintaining available older versions of the tree. The ISOGG web-based Y-DNA haplogroup tree is based on the YCC nomenclature (https://isogg.org/).

Mutations are the fundamental source of biological variation, and their rate is a crucial parameter for evolutionary and medical studies. As will be mentioned in section 2.3.4, by knowing the rate at which such mutations have accumulated, we can date past evolutionary events, such as the time since divergence of a set of Y-chromosomal sequences. By using genome-sequencing data, recent studies have provided improved calibrations of the Y-chromosome SNP mutation rate. Those rates can be estimated in two ways: either by calibrating genetic divergence against fossil or historical evidence for a past separation (the phylogenetic rate) or by direct comparison of DNA sequences from parents and offspring (genealogical rates) (19).

By using whole-genome sequencing data to find de novo mutations through a direct comparison of chromosomes from parents and offspring, Helgason *et al*. (20) have provided a genealogical mutation rate in the range of $0.89–1.2 \times 10^{-8}$ mutations per position per generation (PPPG). By contrast, a recent work has provided an

additional Y-chromosome mutation rate using as a calibration point a 45,000-year-old modern human male from Siberia, known as Ust'-Ishim. By measuring the number of 'missing' mutations in the Ust'-Ishim individual compared to 23 present-day non-African males, they have obtained a phylogenetic rate of 0.76×10^{-9} per site per year (95% CI 0.67×10^{-9} to 0.86×10^{-9}) (21).

In general, genealogical rates tend to be slightly higher than the phylogenetic rate estimates. An explanation for this overestimation of the true rate of molecular evolution in genealogical rates is that, not all new mutations that arise in families each generation are neutral; some of them will be deleterious and hence eliminated by selection, while others will be lost by drift. Obviously, different rates can result in temporal estimates that deviate several-fold and, as a result, it is unclear which rate to choose when calculating TMRCAs. In the case of Y chromosome, the genealogical rate seems to provide better estimates for younger haplogroups (<7000 years), whereas for older haplogroups an evolutionary rate might be a better choice (3,11).

Given their low mutation rate, SNPs are also known as unique event polymorphisms (UEPs) and, in general, they show identity by descent (IBD), rather than identity by state (IBS). A base substitution is IBD in two or more individuals if they have inherited it from a common ancestor without recombination. Generally, the direction of this mutation can be established by examining the orthologous DNA sequence in great apes, our closest living nonhuman relatives. When alleles are different between humans and

apes, the allele resembling the great ape is known as the ancestral allele, whereas the allele that differs is referred to as the derived allele. In spite of the low base substitution rate, the human population is so large that we expect the recurrence or reversion of any SNP to occur each generation. Since the population frequency of these mutation is very low, they will usually go undetected (10). Nevertheless, the problem with reversions and, to a lesser extent, with recurrences is the disruption the linearity of molecular genetic distances, that result in the underestimation of the number of differences between two molecules accumulated since their split from a common ancestor (1).

Short tandem repeats

Short tandem repeats, also known as microsatellites, are tandem repeated copies of a 2– 6 base pairs (pb) sequence and span a median of 25 base pairs (bp). They are one of the most abundant types of repeats in the human genome, with the CA dinucleotide repeats being the most common (Table 1). Approximately, 700,000 STR loci exist in the human genome and, overall, occupy about 1% of its total length (18). Although some microsatellites composed of some specific repeat units show clustering, most are distributed throughout the genome (1).

The mutational mechanism of STR differs from that of nucleotide substitutions in DNA sequences and thus, STR loci evolve in a different fashion than SNPs. It is widely accepted that mutations at an STR locus are due to replication slippage.

Table 1. Properties of microsatellites by repeat unit size. Modified from Jobling *et al.* (1).

Repeated unit/bp	Properties and distribution	Utility
1	Mostly poly(A)/poly(T), associated with Alu, LINE, and other retroelements	Not used, due to small differences in allele size and problem of allele-calling due to PCR stutter, resulting from slippage synthesis errors by the PCR polymerase
2	(AC)n/(GT)n most common, representing 0.5% of genome; (GC)n extremely rare	Widely used in early studies because of ease of discovery; stutter a problem
3	Wide range of different repeat units; some arrays are within or close to genes and can cause diseases through expansion. (AAT)n and (AAC)n most common	Widely used. Alleles easily discriminated, and little stutter
4	Wide range of different repeat units. (AAAC)n and (AAAT)n most common; (GATA)n/(GACA)n frequent, and clustered near centromeres	Widely used. Alleles easily discriminated, and little stutter; form basis of most forensic microsatellite profiling
5-7	Range of different repeat units	Not widely used because of relative scarcity

During DNA replication, once the DNA polymerase encounters the repeat tract, the two strands might dissociate. Then, if the new strand realigns out of register, it will lead to gains or losses of repeat units. As shown in figure 6, if this misalignment introduced a loop on the new strand, it would increase in repeat length. A loop that is formed in the template strand leads to a decrease in repeat length (22). Therefore, it seems that the polymerase loses track of how many repeats there are. Most of these mutations involve the loss or addition of just one repeat unit; this means that STR loci evolve under a stepwise mutation model, understanding one step as one repeat unit. One feature of stepwise mutations is their high probability of occurring independently in different individuals,

which means that alleles with the same size and sequence may not reflect IBD, but IBS. Moreover, the probability of a STR allele mutating appears to be dependent on the number of repeats in that allele; alleles with more repeats are more prone to mutation, probably because it is more likely that DNA loops will be formed (3).

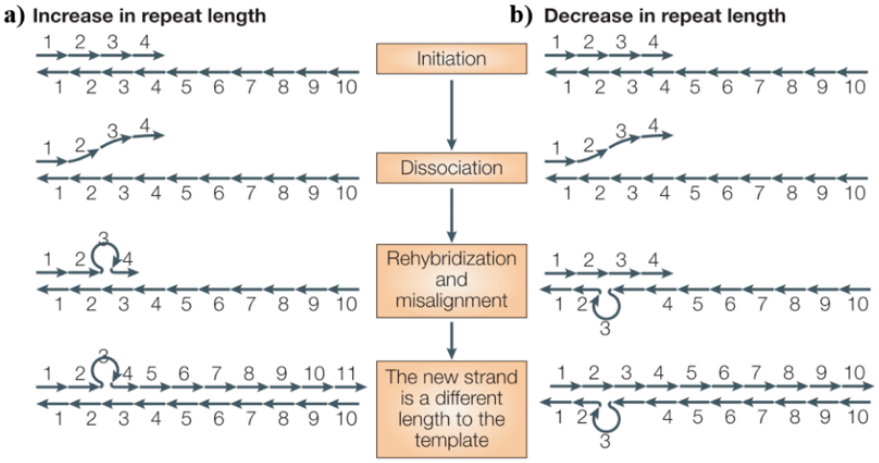


Figure 6. Replication slippage results in an a) increase and b) decrease in repeat length (22).

The most recent catalogue lists a total amount of 4,500 STRs within the Y-chromosomal sequence (18). As mentioned above, the repetitive nature has led an increased rate of mutation. The average mutation rate for all polymorphic Y-STRs is 3.83×10^{-4} mutations per generation (18), although the most commonly employed and commercially available subsets have faster mutation rates: 2.78×10^{-3} (http://www.yhrd.org, accessed April 10th, 2016) for the 16 most used Y-STRs, and 13 rapidly-mutating Y-STRs reach mutation rates $>10^{-2}$ (23). Due to their fast mutation rate, STR

markers provide insights into more recent historical events compared to SNPs and are routinely genotyped in forensic genetics settings.

Due to their higher mutation rates, Y-STRs cannot be used to build reliable phylogenies. However, given that each Y-STR mutation occurs in a chromosome that belongs to a particular Y-SNP haplogroup, Y-STR allele variation is deeply partitioned by haplogroup. As a result, combinations of Y-STRs, which are referred to as haplotypes, define more informative haplotypes within the haplogroups (24,25). In addition, Y-STR haplotypes can be used to predict Y-chromosomal haplogroups in the absence of Y-SNP genotypes (26). Relationships among Y-STR haplotypes are often displayed with median-joining networks (section 2.3.3), which can also incorporate haplogroup information.

1.1.1.b. Recombination

Besides the recombination within pseudoautosomal regions, the remaining Y-chromosomal sequence is thought to avoid recombination. However, recent work has shown extensive gene conversion between arms of palindromes in human and ape Y chromosome and, since gene conversion is considered a form of recombination, some authors claimed that the Y chromosome recombines (27). As a result, the non-recombining region of the Y chromosome was renamed as the male specific region (6).

As we mentioned before, the most pronounced structural feature of the ampliconic regions of the MSY are eight massive palindromes,

that exhibit about 99.9% intra-palindromic (arm-to-arm) sequence identity and contain many testis-specific genes (6). Using comparative sequencing in great apes, Rozen *et al*. (27) showed that at least six of these ampliconic palindromes predate the divergence of human and chimpanzee lineages, which occurred 5 million years ago. This finding means that the high identity observed between palindrome arms could be attributed to gene conversion. And indeed, analysis of MSY palindrome sequence variation in present-day human populations proves the recurrent arm-to-arm gene conversion in our species. Interestingly, human–chimpanzee divergence is significantly reduced in MSY palindrome arms as compared with other MSY sequences examined, which may suggest a directional bias in gene conversion, favouring restoration of the original sequence (27). As a result, gene conversion has been suggested as a mechanism to cope with the irreversible accumulation of deleterious mutations in the absence of recombination, known as the Muller's ratchet (28).

Finally, recombination within the MSY may also occur in males that carry two Y chromosomes (1/1,000). Since both chromosomes are identical, theoretically, Y-Y crossing over may be observed in those cases. In practice, however, most fertile XYY men eliminate one Y chromosome in their germline, so there is no possibility of recombination. Those that retain all three sex chromosomes tend to suffer spermatogenic arrest, so any potential recombinants will never enter to the next generation and thus are not observed (10).

1.1.1.c. Selection

Making past inferences from Y-chromosomal diversity usually assumes that selection has not played an important role in patterning Y-haplotype diversity in populations. However, the Y chromosome is subject to purifying selection and negative selection clearly removes Y chromosomes carrying severe defects in male sex determination or fertility from the population. For example, the loss or inactivation of Y genes can produce an XY female or male infertility (29,30). On the other hand, balancing selection seems unlikely because heterozygous advantage is impossible for this haploid locus, and frequency-dependent selection has not been found. Positive selection may also occur if an advantageous mutation arose and confer an advantage on a previously neutral variant. Because of the lack of recombination, any selection would affect the entire chromosome and produce an increase in frequency of a lineage more rapidly than would be expected by drift (10).

Although studies of nucleotide diversity in global samples of Y chromosomes have suggested an absence of recent selective sweeps or bottlenecks, a few robust associations between polymorphisms in Y-encoded proteins and phenotypes have been found. For example, Jobling *et al.* (31) reported that one class of infertile males, PRKX/PRKY translocation XX males, is more associated with haplogroup P than to any other Y chromosome lineage and thus, showed that selection is acting on Y haplotype distributions in the population.

1.1.1.d. Genetic drift

Given that only one half of the population contains one Y chromosome, its effective population size is limited to 1/4 that of autosomes and 1/3 of the X chromosome. Thus, genetic drift, which involves changes in the frequency of haplotypes due to random sampling from one generation to the next, is expected to be strongly acting upon the Y chromosome. Indeed, genetic drift accelerates the differentiation between groups of Y chromosomes in different populations, which is useful to study past migrations. However, genetic drift could also compromise the suitability of Y chromosome for making past inferences, by changing haplogroup frequencies very quickly. Finally, effective population size differs between the sexes. Most studies have shown that males have higher reproductive variance than females, and thus their effective population size is smaller (1).

1.1.2. Applications of Y-chromosome diversity

The Y chromosome contains the largest non-recombining block of the human genome and thus can be considered one of the most informative haplotyping systems. In this section, I will introduce its main applications in population genetics, forensics and genealogical reconstructions.

1.1.2.a. Population Genetics

Modern humans originated in Africa ~195 thousand years ago (kya) and then spread into the rest of the world. Since this moment

onwards, human populations have been accumulating diversity. Some of these differences might have originated as a consequence of recombination and mutation, while others may have arisen due to selection or by random drift. By discerning these differences between populations we may be able to answer questions such as when and how they arose.

Analyses of genetic diversity at non-recombining uniparental loci have provided important clues regarding human evolutionary events (32). Although mitochondrial DNA (mtDNA) and the MSY each represent only a single evolutionary path, they have been widely used to infer the distribution of allelic variants in and between populations. Indeed, phylogenetic analyses, or the study of the evolutionary history and relationships within and between populations, have provided important insights on the level of structure among populations, as well as on the order and time of their descent. The suitability of uniparental markers in population genetics is mainly driven by their geographical clustering, which is more pronounced in the Y chromosome. The reason is that most societies are patrilocal: when a woman and a man from different places marry, normally the bride moves to the groom's birthplace. As a consequence, most men live in their birthplaces than do women, enhancing the geographical specificity of Y chromosome variants (Figure 4) (10).

The male-specific region of the Y chromosome has been widely exploited in studies of human evolution and population history (10). Over the last two decades, since the report of the first Y

chromosome polymorphism (33), many studies have constructed phylogenies from Y-SNP and have been very useful in defining the geographical distribution of haplogroups (16,17,34). Nonetheless, all of them have suffered from ascertainment bias in the sequence variants studied, which is the systematic distortion in a data set that caused by the way in which markers or samples are collected (10). In spite of their ascertainment bias, these studies have provided important anthropological insights. For instance, Hammer *et al.* (35) showed that the Y phylogeny is rooted in Africa, with the first two branches (A and B) being genetically diverse with subhaplogroups geographically distinct from one another, which supports the out-of-Africa hypothesis. Furthermore, detailed studies of Y-chromosomal diversity in Europe showed that the most frequent haplogroup was R (xR1a) (36).

The application of massive parallel sequencing to large segments of the MSY has provided an unbiased ascertainment of SNPs. Several MPS-based studies (12,37,38) have been able to construct detailed phylogenies with branch lengths being proportional to time, allowing direct estimates of the TMRCAs (section 2.3.4). Nonetheless, phylogenies provided by these studies vary in their sample sizes, the number of population assessed, as well as their representation of known lineages (Table 2). Moreover, sequencing methodologies have also been heterogeneous both in the amount of DNA sequenced and in the mean coverage. By using a sequence-capture design Hallast *et al.* (39) sequenced 3.7Mb of MSY in each of 448 human males at a very high coverage of 51X. The resulting

phylogeny resolves polytomies and provides date estimates for deep nodes.

Table 2. MPS Studies of Human Y-Chromosome Diversity. WGS: whole genome sequencing; SC: Sequence-capture; N: sample size. Modified from Hallast *et al.* (39).

Study	Approach	Mb	Mean depth	N	Populations
Wei *et al.* (12)	WGS	8.97	28.4	36	5
Poznik *et al.* (37)	WGS	9.9	3.1	69	9
Francalacci *et al.* (38)	WGS	8.97	2.16	1,208	1
Scozzari *et al.* (40)	SC	1.5	50	68	unknown
Hallast *et al.*(39)	SC	3.7	51	448	19
Poznik *et al.* (14)	WGS	10.3	4.3	1,244	26

Nonetheless, since it is difficult to summarize all the anthropological insights provided by the analysis of Y chromosome, I will focus on the latest large-scale Y-chromosomal study by Poznik *et al.* (14). By analysing 1,244 human Y chromosomes from 26 worldwide populations of the 1000 Genomes Project, the authors discovered more than 65,000 variants, including single-nucleotide variants, multiple-nucleotide variants, insertions and deletions, short tandem repeats, and copy number variants. By using 60,555 biallelic SNPs they construct a maximum-likelihood phylogenetic that refines the previous phylogeny (Figure 5). The TMRCA of the tree, calibrated using a mutation rate estimate of 0.76 × 10−9 mutations/bp/year, is ~190,000 years and the TMRCA of all non-African lineages (DE and CF) is ~76,000 years. Moreover, around ~50–55 kya, there is a notable increase in the number of lineages outside Africa, suggesting a geographical expansion and

differentiation of Eurasian populations (35). The authors also suggest that according to the principle of maximum parsimony, haplogroup E, the most common African haplogroup might have arisen outside the continent. Moreover, they estimate this back-to-Africa around 58 kya, which is consistent with other hypothesis based on non-Y-chromosomal data (41).

Finally, they infer punctuated bursts in human male demography. In the Americas, an expansion of Q1a-M3 around 15 kya is linked with the time of the initial colonization of the continent (42). Moreover, two independent expansions of E1b-M180 lineages are reported in Sub-Saharan Africa around ~5 kya, both of them predating the Bantu migrations. The authors suggested that they might have been triggered by the development of iron working (1). Another remarkably expansion took place in Western Europe around ~4.8–5.5 kya and is related to R1b-L11. Although it is associated with the origin of the Bronze Age Yamnaya culture, which are thought to have replaced the Neolithic farming cultures in temperate Eastern Europe by 3000 before Common Era (BCE), the six Yamnaya individuals did not carry lineages descending from or ancestral to R1b-L11, so a Y-chromosome connection has not been established (43). In summary, all of these studies highlight the potential of Y-chromosomal data to investigate human population history.

1.1.2.b. Genealogical studies

The study of one's origin can be achieved by tracing back a genealogical tree and identifying one's ancestors. In general, this

has been done by researching archival evidence, such as documents in civil records. However, when this type of evidence is not available, either because it has been lost or because it has never existed, genetic tests can establish the genetic relatedness between individuals at a time depth comparable to that of most genealogical trees (10,15).

Whole genome data is being used to establish relatedness between individuals. However, due to the process of recombination, for close family relationships (>6th degree), the sensitivity and precision of this kind of data diminishes. Because of their uniparental inheritance, Y chromosomes and mtDNA can provide a way to trace back the paternal and maternal lines across a large number of generations, respectively. Nonetheless, in genetic genealogy, given its stronger geographic differentiation and a much wide-range of mutation rates, the NRY is more popular than mtDNA (15).

In a genealogical frame, the Y chromosome has provided a way to identify historical remains. For instance, Y-chromosome STRs are a valuable tool for the identification of human remains in mass graves from World War II (44) and the Spanish Civil War (45). Not only males who died in violent conflicts can be identified, but also historical figures. Y-SNPs and Y-STRs have been used to identify the remains of Jörg Jenatsch, a Swiss national hero in the seventeenth century, whose putative remains were exhumed in 2012 from the Chur cathedral. After comparing his Y-STR haplotype and Y-haplogroup with three Jenatsch men, Haas *et al*. (46) reported

that the remains were more likely to belong to Jörg Jenatsch than to anyone else.

In some cases, instead of a positive identification, the Y chromosome can provide historical cases of false paternity. A good example is the research on the putative remains of King Richard III. Although archaeological, osteological, radiocarbon and mtDNA data were consistent with being his remains, when a set of five paternal relatives were compared to each other, everything changed. They found that four relatives shared a Y-STR haplotype, and that the shared haplotype did not match with that of King's remnants. Given the amount evidence for the remains belonging to King Richard III, King *et al*. (47) assumed that those unmatched haplotypes were two cases of false paternity.

The identification of ancient remains not always ended in a success. In the attempt to verify two presumed remains of Bourbon kings of France, namely a mummified head attributed to Henry IV (48), and a handkerchief that may have been dipped in Louis XVI's blood (49), the identification was compromised by other factors. Both the difficulties of in extracting sufficient undegraded DNA from ancient samples and the fact that the remains of historical figures, such as kings, are very valuable, may complicate the process of identification (15).

Besides the identification of remains, the Y chromosome has proven to be a powerful tool to discern genetic connections. An example that illustrates this nicely is the famous case of US President

Thomas Jefferson. Sally Hemings, one of Jefferson's slaves, claimed that he fathered her children. As a defence, Jefferson's descendants claimed that the true father was one of Jeffereson's nephews, the sons of his sister (Samuel or Peter Carr). However, since Jefferson did not have any legitimate sons of his own, to disentangle the mystery, male descendants of his paternal uncle, Field Jefferson, were used as a source of the Jefferson Y chromosome.

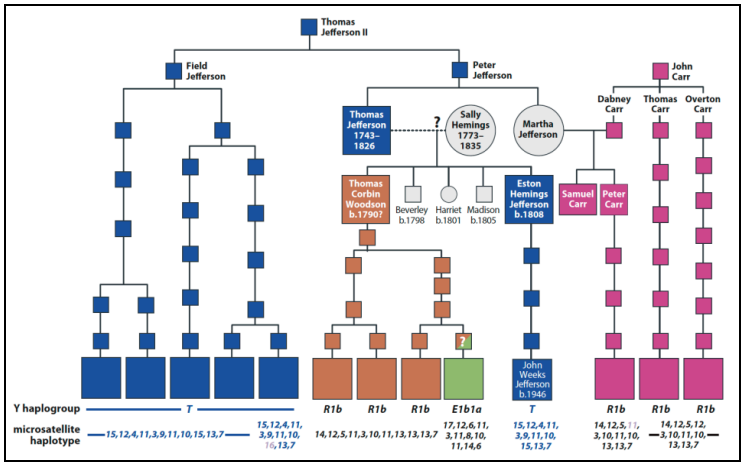


Figure 7. Y-chromosomal DNA analysis of the Thomas Jefferson paternity case (1).

By using Y-specific binary polymorphisms (mostly SNPs) and microsatellites to compare between Jefferson Nephew's and Field Jefferson descendants' Y chromosome, they showed that the descendants of one of Sally's sons matched Field Jefferson's descendants, which agreed with Thomas being his father (50) (Figure 7). Moreover, the frequency of the Jefferson haplotype in the general population appears to be low, and he also belonged to the T haplogroup, which is relatively rare (51). Although the results

do not offer formal proof of paternity, when combined with historical evidence, it seems very likely that Jefferson was the true father of one of Sally's sons (52).

Genealogical studies can be extended further back in time to include a lineages associated with a particular group. For example, the study of Y-chromosomal diversity in Asia revealed a sublineage within haplogroup C*(xC3c) defined by a 16 Y-STR haplotype and found in 16 populations throughout much of Asia, which made up ~8% of the Y chromosomes sampled from the region. However, the pattern of variation within this lineage suggested that it originated in Mongolia ~1,000 years ago and, such a rapid expansion cannot have occurred by chance. Indeed, it has been proposed that this lineage is carried by likely male-line descendants of the Mongolian emperor Genghis Khan, and thus, its frequency was increased by social selection (53).

Nevertheless, except for a few examples, rather than being used to confirm genealogical relationships, the Y chromosome has been extensively used to quantify rates of misattributed paternity or non-paternity. For example, by comparing large numbers of distantly related men in Flanders, Belgium, where documental genealogies can be obtained, Larmuseau *et al.* (54) found that the rate of non-paternity events was 0.9% per generation.

As will be deeply analyse in the following section, surname studies are another important application of the Y chromosome in a genealogical frame. In most cultures, surnames are inherited in the

same fashion as the Y chromosome, through the paternal line. So, surnames can be modelled as multiple alleles of a single locus in the Y chromosome. They have been used to investigate inbreeding and gene flow between populations (55), but more commonly, Y-chromosome diversity has been used to investigate either a particular surname or the whole surname system of a population. However, as will be discussed below, the link between surnames and Y chromosome can be broken due to false paternity, adoption and inheritance of the maternal surname.

1.1.2.c. Forensic studies

The enormous number of differences between the genomes of two individuals and the access to this variation would allow us to identify someone from a DNA sample. However, in practice, the process of individual identification is not as easy as it seems, mainly because all methods and procedures employed to identify individuals must be reliable and accurate. Two types of errors should be allowed at any price: a false inclusion (type 1 error) could lead to the conviction of an innocent person, whereas a false exclusion (type 2 error) could lead to a guilty person being considered innocent.

Autosomal microsatellites appear to be a sensitive and powerful method for proving someone's guilt or innocence. In fact, many countries have databases provided with DNA profiles, which are helping to solve previously intractable crimes. However, whereas the genetic variability in autosomes is mostly due to mutation and

recombination, the NRY change only by mutation, which greatly reduced haplotype variability and compromises the effectivity of the Y chromosome in forensic studies. Moreover, since the Y chromosome avoids recombination, patrilineal relatives are expected to share the same Y-chromosome haplotype. So, we expect that a number of relatives will be indistinguishable from the donor sample using non-recombining DNA. In addition, the geographical distribution showed by the Y chromosome can also compromise the effectivity of this marker. Despite all of these factors are not a problem for exclusions, they can increase the probability of type 1 errors. However, the analysis of Y-chromosomal markers has specialized uses (1,10).

In rape cases, sperm cells of the perpetrator are mixed with the epithelial cells of the victim. The biological evidence most often sought in specimens from alleged sexual assault victims is the isolation and profiling of sperm DNA. However, this method sometimes fails; recovery of spermatozoa by using cervicovaginal smears has been reported in 46–71% of cases. Negative results are usually due to a prolonged post coital interval or when the rapist is azoospermic and has no sperm in his semen (56). In these cases, molecular techniques using Y-chromosome-specific DNA should be reconsidered to identify Y-bearing sperm and non-sperm cells. Moreover, Y-chromosomal STR evidence is also gaining importance in parentage testing with male children and in identity testing with male relatives. Considering a paternity cases of male children where the presumed father is unavailable, Y-chromosomal markers can be tested in male- line relatives, such as brothers (57).

When a Y-chromosomal profile from a crime-scene sample matches that of a suspect, the significance of the match cannot be assessed in the same way as for an autosomal profile, because the alleles are not independently inherited and the profile is usually compared with a forensic population database, such as Y-Haplotype Reference Database (YHRD). Y-chromosomal DNA profiles normally consist on a set of between 9 and 17 (mostly tetranucleotide) STRs. However, the screening of complete Y-chromosomal DNA sequences has yielded the discovery of 166 new Y-STRs, that could be also used for this purpose (58). A key requirement for the application of Y-STR is an accurate knowledge of the mutation rate and pattern. By analysing 186 Y-STR markers in nearly 2,000 DNA-confirmed father-son pairs, Ballantyne *et al.* (59) showed extremely high mutation rates (up to 7.44×10^{-2} per transmission of the Y chromosome). As mentioned in the previous section, the factor that most strongly affects the Y-STR mutability is the total number of repeats. However, the length in base pairs of the repeated motif and the father's age also contributed to Y-STR mutability (60). Indeed, the analysis of the 13 most mutable Y-STRs in an independent sample set, empirically proved their suitability for distinguishing close and distantly related males. The discriminatory power of the so called rapidly-mutating STRs reduces the problems of shared haplotypes among patrilineal relatives, and thus increases the utility of the Y chromosome in forensic genetics (59).

1.2. Surname studies

Every language has its own words, where each of them has a different meaning and belongs to a certain category (i.e. names, adjectives, verbs, etc.). Some of these words have a special feature, they are inheritable. These words are called surnames and, in most societies, they are passed down from father to son. This paternal inheritance establishes a link between them and the Y chromosome. Recently, the study of this relationship has been enhanced by the developments in DNA analysis, but also by the public availability and easy access of surname information. In this section, I will briefly cover some historical aspects of surnames. Then, I will present the most outstanding results obtained in this field. And finally, I will cover the possible applications of surname studies.

1.2.1. History, diversity and types of surnames

Having a name has an obvious advantage, which is to be identifiable. And, although we do not normally consider its importance, a world without people's names would be nothing but chaos. The addition of a heritable name, namely a surname, further facilitates an individual's identification. There are some societies, such as Iceland, that continue to avoid surnames (they use strict patronymics, see below), but the fact is that governments like to identify filiation links, and thus, find surnames rather useful. Therefore, in some countries, such as Turkey, all citizens were forced to adopt a heritable surname in 1934, and in Mongolia

surnames were made compulsory as recently as 1997. By contrast, China is thought to own the oldest hereditary surname system, dating back 4,000 to 5,000 years (24). As discussed below, the age of the surname system has consequences in surname diversity, since, where drift has had more time to operate, surname diversity is reduced.

The first known surname tradition in Europe appeared in Ancient Greece, and it basically featured one's place of origin as part of a person's identification (i.e. Thales of Miletus) (61). However, it never became heritable. The first to establish an inheritable surname system were the Romans. Among the upper classes, each Roman citizen had a *praenomen* (similar to our given name), and a *nomen gentilicium* (name derived from the '*gens*' or clan of an individual). Later on, after the Republic, they introduced the cognomen, which acted as a family name (i.e. Gaius Julius Caesar) (62,63).

The Roman naming pattern started to change with the arrival of Christianity into Europe. The Christian naming system consisted on adopting a biblical first name, which had higher prestige and influence. Since the set of available biblical names was larger than the set of Roman's praenomen, the use of a two surname system decreased as the number of Christian citizens in the empire increased (61). In addition, by the 5th century Common Era (CE), the invasion of Germanic peoples, who had a strong single-name identification, enhanced even more the return to a single-name system (63). Although there were some exceptions (i.e. Ireland), the surname system was brought back to Europe around the 13th and

14th centuries and, until this moment, the origin of surnames as we know them today was relatively similar in all European countries (64).

The diversity of heritable surnames varies considerably depending on the country. For example, in China, surname diversity is considerably low. In fact, the world's commonest surname, Li, comes from China (65). In contrast, in most countries surname diversity is incredibly high, with surnames having an average of carriers below 100. One of the main reasons for having high levels of diversity appears to be the long history of admixture (e.g. USA), or the recent adoption of surnames (e.g., the Netherlands, Turkey). In Britain the recent immigration has tripled the number of surnames in the last years (24).

Despite the controversy around derivations of surnames, many of them can be classified into five large general categories or types, generally common in all Europe. These broad categories are patronymic, occupations or status, toponymic, topographic, and nicknames or characteristics (24).

Patronymic surnames consist on using the name of the father next to the given name, to distinguish between people with the same first name. They are generated by adding, after the given name, the father's name. It can be used as it is, excluding any sort of prefixes or suffixes, such as in the Catalan, French or German tradition (i.e. Andreu, Richard and Walter, respectively), or it can be modified by adding a particle meaning before or after the name. Different

languages have different ways to generate patronymic surnames. For example, in English it is common to add –son after the name (or –s in Wales; e.g. Robertson and Williams) and in Spanish, although many patronymic surnames were created without suffixes, most of them were formed with the suffix –ez or –z (e.g. Pérez and Díaz) (24,63,64). In Catalan-speaking territories, new patronyms were generated by phonetic translation, and, in general, –ez became –is (e.g. Peris) (61). As a curiosity, in Iceland, most individuals carry a strict patronymic rather than a surname: if the father's name is Stefán, his sons or daughters will be known by a first name plus Stefánsson or Stefánsdottir, respectively. The son's first name will generate the patronymic for the next generation (24).

Toponyms generally include surnames based on the actual name of a town, region, village, valley, etc. These include Charlesworth (England), Tsui (China), Pontecorvo (Italy) or Toledo (Spain). Toponyms can also include surnames that describe the place where the person is a native or autochthonous of (e.g. French, Welsh or Dane) (24,64). As mentioned before, this tradition was already established in Ancient Greece, where people were named according to the place where they come from, where born in, etc. (61).

Topographic surnames designate people living in or near a recognizable milestone of the land, such as relevant features of the landscape or topographical accidents of the territory. Examples of this are Bridges, Forrest, Suzuki (Japanese – pampas grass), Colina (Spanish – Hill), Rius (Catalan – Rivers), etc (24,62,63). In many

societies, both toponymic and topographic surnames are indeed the most frequent (64).

Occupational or status surnames include all names derived from the job, post, status or social position held by the person carrying the surname. Among these surnames we could find Fisher (fisherman), Wright (maker of machinery or objects), Franklin (feudal status term), Chakraborty (Indian – local landlord), Müller (German – miller), Monje (Spanish – Monk), Ferrer (Catalan – Smith), etc. This naming system was more prevalent in small villages and rural areas, which made up the majority by the time when surnames originated (24,64).

Finally, surnames from nicknames or personal characteristics generally describe physical features, personality, kinship or working instruments. Examples of these surnames are Calvo (Spanish – Bald), Darwin (dear friend), Klug (German – wise), Martell (Catalan – Hammer), etc (24,62). Surnames that arose from personal characteristics or nicknames are probably one of the oldest ways to distinguish between people having the same first name. The use of nicknames is still widespread among many different cultures, mainly in rural areas. The funny thing is that, in many of the cases, when these surnames are inherited, the personal characteristics of the actual bearer do not match with the nickname (64).

1.2.2. Y chromosome and surnames

In most societies, surnames are inherited together with the non-recombining region of the Y chromosome and so, men sharing surnames might be expected to share the same Y chromosome haplotypes. This property of surnames made them a useful tool to investigate consanguinity and genetic relations within and between populations through measurement of isonymy, which means sharing a surname (55,66–68). Nonetheless, the inference path has been recently inverted: Y chromosome polymorphisms are being used to investigate surname history (24). As we mention in section 1.1.1, two types of markers are commonly used to distinguish between Y chromosomes: STRs and SNPs.

However, the link between Y chromosome and surnames could be influenced by a number of additional factors (69). First of all, mutation will act to diversify Y chromosome haplotypes associated with a particular surname through time. Nevertheless, its impact is quite predictable and, given the time scale of surname establishment, only rapidly-mutating markers such as short tandem repeats are expected to alter Y chromosome haplotypes associated with a certain surname. The mutation rates of single nucleotide polymorphisms are so low that are not expected to undergo mutations. Hence, we expect biological descendants of the same surname's founder to have the same SNP haplotype (*haplogroup*), but that might not be the case for STR haplotypes (24).

Another factor that might have some impact on the strength and structure of the relationship between surname and Y-haplotypes are the differences in the number of founders at the time of surname establishment within a given population. Those deriving from common occupations or first names, for example, are more likely to have multiple founders than those derived from the names of small villages (70). So, this will result in more than one Y haplotype type being associated with a given surname (24). Furthermore, a group of factors referred to as nonpatrilineal transmissions (NPTs), which include non-paternity, child adoption, and matrilineal surname transmissions, will act to introduce exogenous haplotypes into a surname (71).

Finally, genetic drift, which can be defined as the random changes in haplotype frequencies through time, will also influence the link between Y chromosome and surnames. However, in contrast to all the other factors described below, it will act by reducing the diversity of haplotypes within surnames and, stochastically, some Y-chromosome lineages could increase in frequency while others could become extinct (24,71). In fact, it has been reported (72) that genetic drift is responsible for the complete extinction of some British surnames.

Many studies have focused on the Y chromosome to investigate surname history. The pioneering study of the surname Sykes (73) revealed low Y-haplotype diversity among unrelated carriers of the name, which was compatible with a single founder event. This study also indicated that the distribution of other Sykes Y

chromosome haplotypes was not significantly different from those in controls, which may be accounted for by the nonpatrilineal inheritance of the surname with an estimated rate of 1.6% per generation. Another single-name study investigated the Y chromosome of men carrying Colom and Colombo surnames with the purpose of determining whether a genetic analysis of the remains of Christopher Columbus could reveal some doubts about his geographical origin (74). They found that Y-chromosome diversity was reduced in the Iberian Colom and most of the men belonged to a few lineages. On the contrary, the Colombo samples were genetically as diverse as the general north-western Italian population. This difference is due to the fact that orphans there used to be given the Colombo surname. In this study, it was shown that by studying the Y chromosome we can identify groups of founders for names. In another study, King *et al.* (75), rather than focusing on a single name, examined 150 randomly ascertained pairs of males who each share a British surname. By genotyping a set of 17 Y-STRs, they showed that sharing a surname significantly increases the probability of sharing a Y-chromosomal haplotype, and that this probability increases as surname frequency decreases. In other words, the rarer the surname is, the strongest the association between surname and Y chromosome (75).

After that, two studies, in Britain (71) and Ireland (76) have collected and analysed larger groups of men with fewer surnames, using the same set of 17 Y-STRs, plus several haplogroup-defining SNPs. Moreover, in order to analyse the diversity within surnames, both of them used networks to display the relationship between Y

STRs haplotypes, haplogroups and surnames (Figure 8). Results showed that British control males bearing different surnames show very little haplotype sharing (Figure 8a), which is also true for men carrying the commonest surname, Smith (Figure 8b). By contrast, less common surnames (Figure 8c, d) show decreasing haplogroup diversity and extreme haplotype sharing, which might indicate that a single founder event took place**.** Taken together, these results suggest that surname frequency was driven by polyphyletism. This study also revealed that the estimated time depth is compatible with the historical time in which paternal surname inheritance is systematized (71).

By contrast, the other large Y-chromosome-surname association study carried out in Ireland (76), showed no significant correlation between the rarity of a surname and the diversity of the Y chromosomes within it. Some common Irish names such as Ryan (Figure 8e), show an extreme degree of haplotype sharing. The difference between both studies could be explained by an amplification of genetic drift in Ireland, driven by natural selection associated to a cultural marker: the surname. Indeed, males bearing some prestigious surnames could have had more social and reproductive success. However, it could also reflect other demographic historical differences, such as greater urbanization in Britain and different impacts of epidemic disease. In addition, it should be noted that using only 17 Y-STRs might be insufficient considering that Ireland most frequent haplogroup is R1b (77,78). Therefore, a subsequent high-resolution study in Ireland is required to verify the claims by McEvoy and Bradley (76). Finally, in Spain,

together with one of the main works of this thesis (section 4.2), a study on Spanish, Catalan and Basque surnames was published by Martinez-Cadenas *et al.* (79) and showed that, as proposed by the British model, the frequency of a surname is driven by how often it was founded.

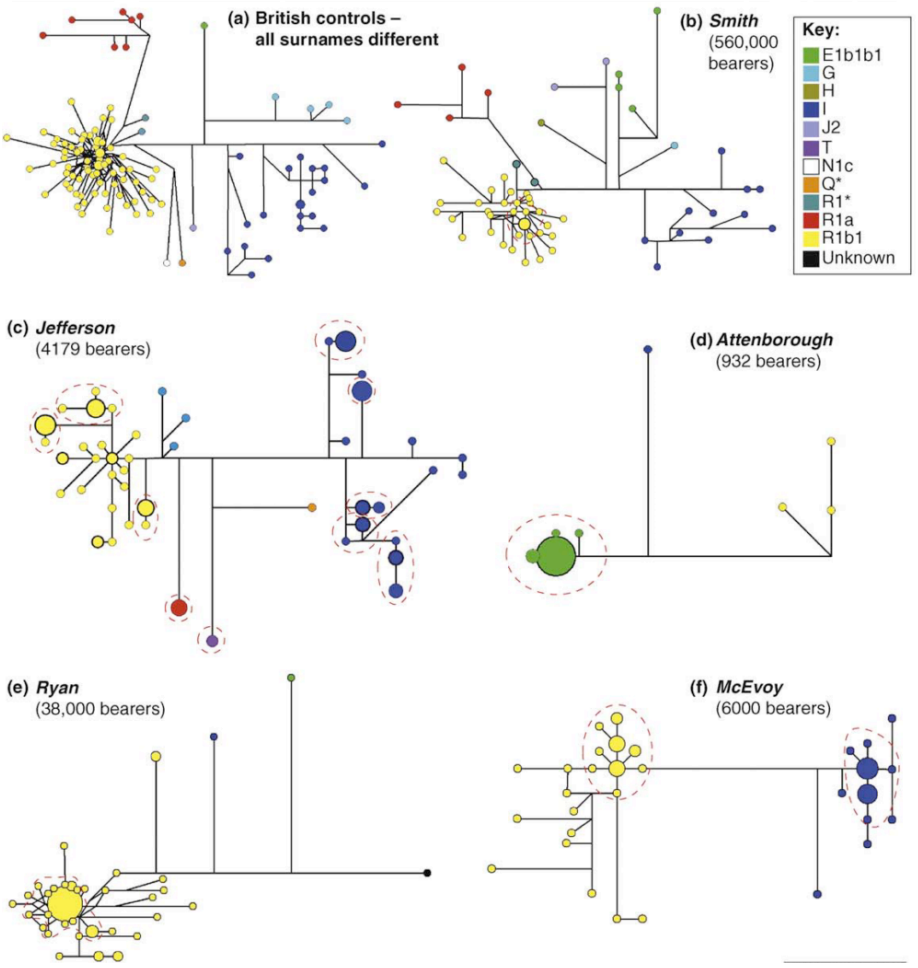


Figure 8. Diversity of Y chromosome haplotypes among control males and five surname groups is represented by median joining networks. Circles within the networks represent Y haplotypes, with area proportional to frequency and coloured according to haplogroup (24).

1.2.3. Applications of surname studies

As we introduced in the previous section, the first application of surnames in genetics was in isonymy studies, a field originated by Charles Darwin's son George. The weak point of isonymy is that it assumes that a shared surname implies shared ancestry, which, as we discussed below, is often likely to be incorrect (80). However, the field of isonymy remains active (55). Here, we will focus on main applications of surname studies.

1.2.3.a. Population genetics

The geographical specificity of surnames, together with its link with the Y chromosome, made them a useful tool for genetic studies of historical migrations and admixture. A study focused in Y chromosome variation and Irish origins (81) showed that when samples are partitioned by surname, significant differences are found in genetic frequency between those of Irish Gaelic and of foreign origin. Furthermore, another study carried out in the Irish population revealed that a particular 17-STR haplotype that peaks in frequency in the northwestern part of the island, is significantly associated with surnames supposed to have descended from the most important and enduring dynasty of early medieval Ireland, the Uí Néill. This is a powerful illustration of how Y-chromosome phylogeography can be influenced by social selection (82). Nevertheless, a subsequent study by McEvoy *et al.* (83) showed much less evidence of coancestry within two other patrilineal clans,

suggesting that the link between modern surnames and early origins has been broken in some cases.

Given the strong geographical differentiation of Y haplotypes between continents, a study by King *et al*. (84) showed an association of a clearly African Y lineage with a rare English surname. In addition, the observation of low diversity of Y haplotypes in surname groups in Colombia, suggested a founder effects caused by Spanish and Portuguese male colonizers (85). In the same study, surnames have been also used to make inferences about the past rates of non-paternity.

Furthermore, as proposed by Manni *et al.* (67), surnames can also be employed as a sampling strategy. Rather than categorize donors into local sub-populations on the basis of at least two generations of residence, which can be compromised by migration events, modern Y chromosome have been chosen on the basis of surnames. Indeed, in a study focused on the genetic legacy of the Vikings to the Wirral peninsula and West Lancashire, in northwest England (86), showed that the Y-chromosomal haplotypes of two sets of samples are significantly different, and in admixture analyses the surname-ascertained samples show markedly greater Scandinavian ancestry proportions. This, not only supports the idea that northwest England was once heavily populated by Scandinavian settlers, but strongly suggests the idea that surname-based ascertainment provides a sample that more closely reflects past populations, before immigration from elsewhere.

1.2.3.b. Forensic application

As proposed by King *et al.* (75) the link between surname and Y-chromosomal haplotype could be used for surname prediction in forensic investigation. Having a list of surnames with associated Y-STR haplotypes could prioritize a suspect list, when the "guilty" Y profile is matched with one or more surnames. Obviously, surname prediction would act only as an investigative tool to provide leads, and not serve as evidence in a trial.

This approach might be compromised by several factors. Since the link between surname and Y haplotype is weak for common names (Figure 8), we can only include those surnames with intermediate or low frequency. Moreover, it has been showed that some males bearing different surnames share the same Y haplotypes, which would result in many surnames being returned. Finally, one concern in human genetics research is maintaining the privacy of study participants and it has been shown (87) that surname prediction has an associated risk of infringing someone's privacy. The growth in genealogical registries may contribute to loss of privacy, given that genotypic information is accessible online to facilitate discovery of genetic relationships. Indeed, predictions could be made for the surnames of the CEU participants of the HapMap. Even more astonishing is the case of a 15-year old boy conceived by anonymous sperm donation who wanted to track down his genetic father (88). Although his genetic father had never had his DNA tested, by genotyping his own Y chromosome genealogical records

and exploring some public databases, the boy was able to trace his biological father by surname prediction.

1.2.3.c. Genetic genealogy

Genetic genealogy is probably the most popular application of the relationship between surnames and Y-chromosomal haplotypes (24). The increasing popularity of genetic genealogy could be attributed to the access to Y-chromosomal data and the huge availability of information afforded by the internet. Indeed, commercial kits that offer Y-DNA testing can provide valuable information to genealogical researchers. First, they can show if two men with the same surname share a haplotype and thus, a recent common ancestor (89,90). Then, these tests can also provide, with some uncertainty, estimates of the time when that ancestor lived (91).

Genetic genealogy can provide a link between families of the same village or region with the same surname and between families with different spelling variants of a similar surname, even in the absence of any archival evidence of a relationship between the families. Moreover, the link between surnames and Y chromosome appears to be a useful tool for genealogists to verify and extend their paternal lineages using haploid markers from the Y chromosome (92). Nonetheless, as we mentioned above, not everyone who shares the same surname can be linked to each other. In that sense, surname projects can provide a refinement of family trees, as well as the inclusion or rejection of branches for further genealogical

investigation. And, in spite of the possible ascertainment bias of samples, this resource should be seen as a link between academic and amateur activities (24).

Publications on the online Journal Genetic Genealogy Genetic genealogy evidence the impressive level of knowledge of amateurs ("citizen scientists") about topics such as molecular evolution, population genetics, and others. Another popular resource for genetic genealogy is the International Society of Genetic Genealogy (www.isogg.org), which provides a support network for genetic genealogists and contains the regularly updated ISOGG Y-haplogroup tree. Some DNA-typing companies host many surname projects, such as Family Tree DNA (www.familytreedna.com) and DNA Heritage (24). The interpretation of the relationships among customers' Y haplotypes depends on the haplotype resolution, which improves with the number of markers typed. However, we should keep in mind that when the number of STRs increases, the probability of detecting an STR mutation between close relatives also increases (69). Finally, an obvious risk of genetic genealogy is the detection of unexpected past non-paternities, as well as cases of adultery within family (e.g. the Thomas Jefferson case).

1.3. North African population history

1.3.1. Evidence from the archaeological and historical records

Although much of the archaeological research on the topic of early biological and behavioural modernity has focused on East and southern Africa, North Africa has yielded one of the richest and most complete hominin fossil records of early *H. sapiens* (93). Early descriptions of the hominins from Jebel Irhoud cave, located 55 km south east of Safi (Morocco), emphasized similarities with Neanderthals (94). However, study of the dental development of a juvenile from Irhoud (Irhoud 3) showed that the pattern of dental growth was similar to that exclusive of modern humans (95). Moreover, the application of X-ray synchrotron microtomography revealed that Irhoud 3 lived around 130,000–190,000 years ago (ya), suggesting that the presence of *Homo sapiens* in North Africa dates from just less than the geological ages of the earliest evidence for early *Homo sapiens* in East Africa (96).

Nonetheless, new excavations at Irhoud enabled the discovery of human fossils that showed a mosaic of features that resemble early or recent anatomically modern humans. This, in combination with an age of 315 ± 34 thousand years, places the Irhoud evidence in a new perspective and supports a complex evolutionary history of H. sapiens involving the whole African continent (97).

Indeed, the antiquity of the human remains found in North Africa highlighted the importance of the area for understanding the origins of human anatomical and behavioural modernity and has opened a discussion about the evolution of the first settlers. On the light of these results, it has been proposed that first modern humans might have originated in Northwestern Africa and then spread into the rest of the continent. Nonetheless, most researches strongly support the possibility that those hominids belonged to a group of modern humans that were not involved in the Out of Africa migrations and thus, did not contribute to the genetic heritage of current human populations (98). This controversy aside, it is undeniable that these findings suggested that the origins of modern humans are much more complex than was previously thought (Figure 9).

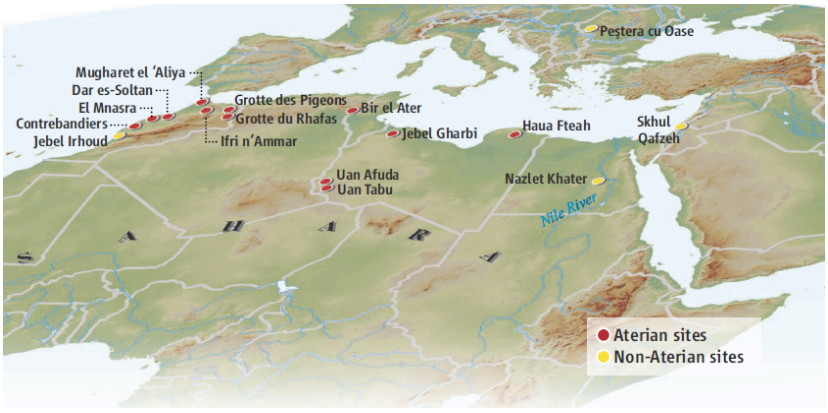


Figure 9. Hominin sites in North Africa in the Upper Pleistocene (98).

1.3.1.a. North Africa during Prehistory

Associated with significant fossil hominin remains, different cultures have been identified in archaeological records. The Upper Pleistoncene (known as the Late Stone Age in the study of African

prehistory) in North Africa is associated with either Mousterian or Aterian assemblages. The Mousterian industry includes hand axes, racloirs and points, and displays the presence of the Levallois technology (93). The Aterian, which extends from the Atlantic coast almost to the Nile Valley and from the Mediterranean coast to the southern Sahara, exhibits novel tools not seen in other Mousterian contexts, namely tanged or stemmed artifacts, some bifacial foliates, unifacial points and worked bones (99) (Figure 10a). Another fundamental innovation associated with the Aterian is the presence of elements more related to a symbolic behavior, which is one of the markers of modernity associated to *Homo sapiens*, and would thus establish the difference between early and modern humans. Examples of this behavior are shell beads, pigment use and structured fireplaces (100–102).

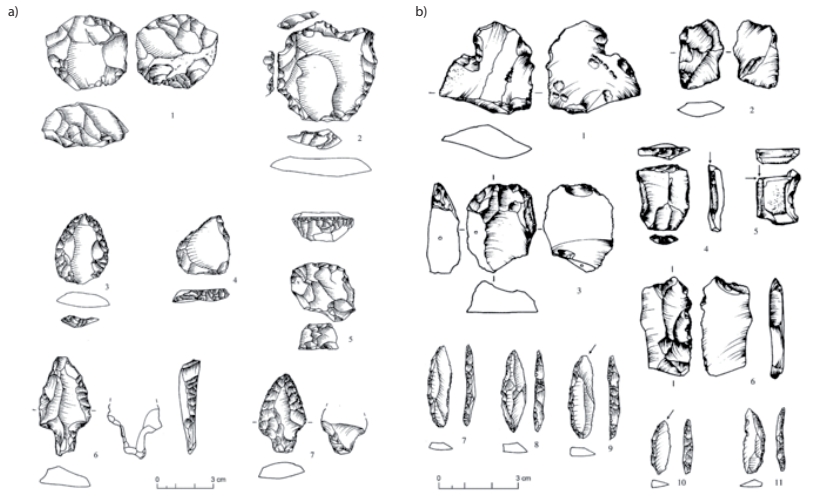


Figure 10. a) Aterian artifacts and b) Iberomaurusian artifacts. Adapted from Garcea and Giraudi (103).

New techniques to date stratigraphic layers (104,105) suggested that the Aterian industry appeared at least 80,000 ya, probably going as far as 120,000 ya. Moreover, it has been suggested that from 80,000 to 60,000 ya groups associated with Aterian technocomplexes underwent a major expansion in combination with behavioural innovations (106), which has been linked to the evolution of modern human behaviour and dispersal in the region (107) and the hypothesis that fully modern human behaviour would have taken place first in North Africa and then spread into the rest of the continent.

Finally, it has been proposed that climate conditions played a pivotal role in population mobility within North Africa (103). In fact, some authors claim that Aterian-related groups spread throughout North Africa during the Last Interglacial, between 125,000 and 74,000 ya, developing some local or regional diversification. Moreover, changes in the environmental conditions during this period would have driven the variability in the Aterian culture. While a drier climate would have favoured isolation, wetter periods would have enabled contacts between groups. Nonetheless, it is still not fully understood why Aterian tool production outlived this climate changes without any clear sign of cultural evolution.

Around 20,000 ya the Aterian culture was suddenly replaced by the Iberomaurusian and, in the beginning of the Holocene most of the technological assemblages found in North Africa were Iberomaurusian (Figure 10b). This industry is characterized by microlithic backed, partially backed obtuse-ended, and other

bladelets (108). Furthermore, Iberomaurusian assemblages have been found together with human burials, whose examination has showed that Iberomaurusian people were a very hetereogenous society (109).

None of the several hypotheses posited for the disappearance of the Aterian culture has been proven (110). In fact, it has not been possible to establish whether the origins of Iberomaurusian technology took place through population continuity or population replacement (103). Like its origins, the end of the Iberomaurusian culture is also very controversial. While in Western Maghreb and Libya it was replaced by the Neolithic period, in Tunisia and eastern Algeria it was displaced by a new industry, the Capsian.

Capsian sites contain large accumulations of ashes, burned stones, knapped flint and, especially, of land snail shells, which have given them the name of *escargotières*. An elaborated Capsian art, revealed by all sorts of decorative objects, suggested the presence of symbolic system associated with complex social interactions. Capsian assemblages, dated between 10,400 and 6,000 ya, have been divided into two categories: i) Typical Capsian, characterized by large tools fabrication, a poor bone industry and a microlithic component, and ii) Upper Capsian, with much smaller tools, abundance of bone industry and the development of geometric microliths (111,112). In addition, morphological similarities between Capsian human remains and extant North Africans, have suggested that the origin of the first Berber peoples can be traced back in this period, around 9,000 ya (113). It has also been

hypothesized that these first Berbers interacted with Iberomaurusians and Neolithic people, but no archaeological evidence has been found. Finally, the overlap (around 6,000 ya) between the latest Capsian sites, and the first Neolithic sites in Morocco suggested that the Capsian culture persisted well into the Neolithic, a fact that implies the presence of a hunter-gatherer population in the middle of an agricultural and complex society.

The Neolithic represents the beginning of food and animal domestication, the introduction of agriculture and the end of the hunter-gatherer lifestyle. A relatively comprehensive understanding of the origin and spread of agriculture and the transition from hunter-gatherer economies to farming has been achieved for several regions of the world, such as the Near East and Europe. For example, in the Nile Valley, the transition to agriculture took place around 7,500 ya, probably as a consequence of the Near Eastern influence. However, in certain zones such as North Africa the evidence is more challenging to interpret (114).

As mentioned above, in Northwest Africa and Libya, Neolithic replaces Iberomaurusian. In Morocco, the Neolithic is first evidenced by the presence of ceramics (102) and later by the presence of cereal crops. A gap of around 3,000 years has been reported between the origins of ceramic production (9,000 ya) and the origins of food domestication (6,000 ya) in North Africa (115). All of this evidence suggested that in North Africa sedentarism would actually predate agriculture. Moreover, in North Africa food domestication started with pastoralism, not agriculture. Agriculture

is thought to have been developed in the region around 4,000 ya (115,116).

1.3.1.b. North Africa during historic times.

The Neolithic period would last in North Africa until the Phoenician colonization. Indeed, the first reports of North African history refer to the Phoenician establishment around 900 BCE. The Phoenicians were a distinctive and independent civilization that dominated the Mediterranean Sea during the first millennium BCE, emerging from a coastal section of the Eastern Mediterranean. They developed a large presence throughout the Mediterranean by the establishment of settled colonies and many trading posts. One of the most important and wealthiest Phoenician cities was Carthage, in modern Tunisia (117,118) (Figure 11).

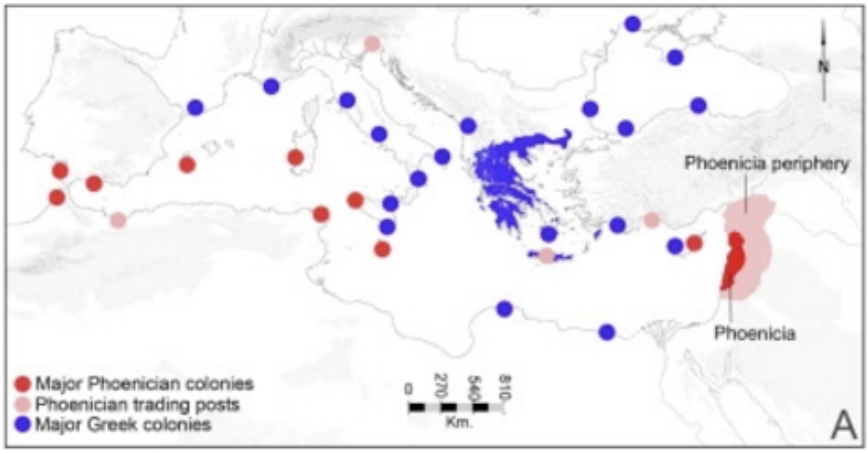


Figure 11. Maritime expansions of the Phoenicians (11th century BCE) and Greeks. Red: Phoenicia, Phoenician colonies; pink: Phoenician trading posts; blue: Greece and Greek colonies. Adapted from Zalloua *et al*. (119).

However, by 264 BCE some conflicts with Libyans, Numidians and

Mauri Berbers might have triggered the beginning of first Punic war. Then, by 146 BCE the third Punic war resulted in the complete destruction of the city of Carthage and the annexation of all remaining Carthaginian territory by Rome, which in turn gave the Romans the hegemony of the Mediterranean Sea. During the first centuries, Roman occupation had a limited demographic impact, but in 27 BCE, with the transformation of Rome to an Empire, it caused the Romanization of the region and of part of the Berber population, and the division of North Africa in two provinces. At this moment, North Africa was one of the wealthiest Roman provinces, and Christians and Jews started to settle in the region (117).

During the 7th century CE, the Arabs conquered northern Africa from east to west and spread their language and religion throughout the native Berber population, which changed the history of North West (NW) Africa profoundly. The cultural and political impact of this invasion was accompanied by a demographic contribution of the Arabs to NW Africa and, as we will see in the following section, many attempts have been made to estimate the extend of this contribution. Moreover, the Islamization of the Berbers, which included the adoption of the Arab language, implied a dramatic change in the North African culture. In fact, it is difficult to ascertain whether current Arab-speaking North Africans are from Arabic descent, and current Berber-speaking North Africans were already present in North Africa before the Islamic expansion. Finally, soon after the conquest of North Africa, a mixed group of Berbers troops under Arab leadership, known as the Moors, crossed to the Iberian Peninsula where they would rule until 1492

(120,121). The rule of the Arab dynasties ended with a decline in the 16th century, when the Ottoman Empire took control of most of North Africa (with the exception of present-day Morocco), and from the 18th to the middle of the 20th centuries Spain, France, Italy and the United Kingdom colonized North Africa (122).

In addition to the migration complexity found in North Africa, cultural diversity is characterized by the presence of two main branches of languages: Arab and Berber, both belonging to the Afro-Asiatic family. During the Arab Caliphate rule, the Arabic language started to spread replacing the use of the indigenous Berber languages. As a result, these languages, which are considered the languages spoken by the ancestral peoples of North Africa, underwent a dramatic reduction in usage (113).

1.3.2. Evidence from genetic data

1.3.2.a. Results from autosomal data

A database of allele frequencies based on classical genetic polymorphisms in North African populations was compiled by Bosch *et al.* (123), which included information on blood groups, red cell enzymes, serum proteins, and HLA antigens extracted from previous studies (124–127). Overall, a total of 62 different loci were used in an attempt to reconstruct North African population's demographic history. Principal component analysis and neighbour-joining tree based on genetic distances showed an east-to-west pattern of variation between populations from northwest Africa and

populations from Egypt and Libya. Moreover, Egypt and Libya showed the smallest genetic distances with European populations. The longitudinal genetic cline could have been produced by an isolation by distance process. However, this model could not account for the short genetic distance between Libya or Egypt and the European populations and thus, other factors such as directional migration from east to west must have taken place. In fact, they suggest that population replacement during the Neolithic from the Levant could explain the genetic similarity between Libya, Egypt and European populations. Since the Neolithic expansion originated in the Fertile Crescent, it would have easily reached the surroundings, but it would have taken more time to reach the west. Finally, in addition to the Neolithic expansion, authors also proposed that the Arabic expansion may have also contributed to the differentiation between eastern and western North Africa (123).

Later on, the analysis of autosomal STRs in several Arab- and Berber-speaking groups from northwest Africa showed a very low genetic differentiation pattern (128). In fact, grouping populations according to cultural or linguistic differences was not associated with genetic differentiation. Thus, the authors claimed that Arabisation was mainly a cultural process. By contrast, a clear genetic difference was found between northwest Africa and Iberian populations, which must be accounted to the Gibraltar Straits acting as a strong barrier to genetic exchange. Nonetheless, the finding that northwest Africans were genetically closer to Iberians and to other Europeans than to African Americans, suggested that some degree of gene flow into Southern Iberia may have existed.

Until recently, genome-wide analyses in North African populations were restricted to a sample of a Berber isolate from Algeria, the Mozabite, which exhibited a shared genetic background with European and Near Eastern populations, but also traces of sub-Saharan gene flow (129,130). Nonetheless, the use of a single proxy has ignored the genetic heterogeneity of North Africa. In 2012, Henn *et al.* (131) published the pioneer genome-wide analysis of seven North African populations (North and South Morocco, the Western Sahara, Algeria, Tunisia, Libya, and Egypt), which showed an amalgam of ancestral components (i.e., sub-Saharan, Maghrebi, European, and Middle Eastern) in North African groups. In fact, they identified a gradient of likely autochthonous Maghrebi ancestry that increases from east to west across northern Africa and suggest that this ancestry is likely derived from ‘‘back-to-Africa’’ gene flow more than 12,000 ya, in pre-Holocene times. Moreover, the high frequency autochthonous Maghrebi component showed by the single Berber sample analyzed, suggested that indigenous North African ancestry might be more frequent in populations with historical Berber ethnicity. As shown in the following section, this hypothesis challenges the lack of genetic differentiation shown by uniparental markers.

This study also evidenced that most North African populations exhibit a shared ancestry with the Near East, and to a lesser extent with Sub-Saharan Africa and Europe. Signatures of sub-Saharan African ancestry vary substantially among populations and this might be due to trans-Saharan slave trade that occurred during this period. These Sub-Saharan migrations have been estimated around

1,200 ya in southern Morocco and about 750 ya in Egypt. This is possibly due to the trans-Saharan slave trade that occurred during this period. In summary, this analysis revealed an extraordinary complex history of migrations (131).

However, the inclusion of a single Berber sample in the study compromises the conclusion regarding the differentiation between Arab and Berber groups, because it can be attributed either to a genetic singularity of Berbers or to the isolation and drift of this specific sample. In order to characterize the genetic heterogeneity of North African groups, focusing on the putative differences of Berbers and Arabs and estimate admixture dates, Arauna *et al*. (132) analyzed genome-wide autosomal data in five Berber and six Arab groups, and compared them with Middle Easterners, sub-Saharan Africans, and Europeans. By using haplotype-based methods, they showed a lack of correlation between geographical and genetic populations and a high degree of genetic heterogeneity. Moreover, the analysis of additional Berber samples supports the idea of no strong genetic differentiation between Berbers and Arabs. In fact, authors claim that not all Berber groups might be considered genetically isolated or homogeneous, whereas some groups such as the Chenini and Sened (Tunisian Berbers), showed a genetic homogeneous pattern and high inbreeding signals, other groups such as Moroccan Berbers, are genetically heterogeneous and diverse. Finally, admixture time estimates showed a strong peak around the 7th century CE, which coincides with the Arabic expansion; another admixture date around the 17th century that could be attributed to the sub-Saharan slave trade in the Modern

Era; and, a lower proportion of Sub-Saharan migration dated from the first century BCE, which could account for more ancient slave trade during the Roman or Islamic periods.

1.3.2.b. Results from uniparental data

The analysis of uniparental markers exhibited a lack of differentiation between Arab- and Berber-speaking populations, which is in agreement with autosomal data (132). The analysis of more than 2,000 North African mitochondrial DNA lineages showed moderate degree of East-West differentiation, with a genetic discontinuity between Libya and Egypt. Although sub-Saharan haplogroups are present in North Africa, most mtDNA lineages have been recently brought from Europe (H, HV0, L1b, L3b, U6) and the Near East (L0a, R0a, N1b, I, J) (133).

Despite this heterogeneous haplogroup distribution, two mtDNA lineages, U6 and M1b, are found at higher frequencies than elsewhere and thus, have been suggested to have originated from North Africa. Moreover, as shown by Fadhlaoui-Zid *et al.* (133) both haplogroups exhibit an opposite frequency gradient in North Africa: while U6 is more frequent in the West, M1 predominates in the East. Given that U6 and M1 might have a Palaeolithic origin (134), the most plausible explanation of the frequency distribution and the coalescence age estimates of both haplogroups, would be an early split in the back-to-Africa migration followed by a period of stability and a period of expansion.

Indeed, the mitogenome of a 35,000-year-old *Homo sapiens* from

Europe strongly supports that an Upper Paleaolithic back-to-Africa migration carried the U6 lineage into North Africa, where it spread and diversified into multiple subhaplotypes. Therefore, the U6 haplotypes found in modern North-African populations are derived from the Western Asia basal haplogroup (135).

As we have mentioned in section 1.1.2, most modern societies are patrilocal, which means that women tend to move to their husbands' location (136). As a consequence, it is expected that Y chromosome analysis will show more structure than mitochondrial DNA. Indeed, studies focused on Y chromosome have revealed two lineages (E1b1b1a-M78 and E1b1b1b-M81) at high frequency in North African populations. However, the origin and emergence of these lineages have been very controversial. E1b1b1a-M78 probably emerged in Northeastern Africa (137) and is today widely distributed in North Africa, East Africa, West Asia, and Southern Europe. E1b1b1b-M81 shows high frequencies in Northwest Africa and, noticeably, a high prevalence among Berber-speaking groups (i.e. the Tuareg) (138). Thus, it has been proposed as an autochthonous North African Y chromosome. Besides these two main E lineages, J1 and J2 are also common in North Africa. While J1 is found at high frequencies in the Arabic peninsula and has been previously associated with the Islamic expansion (139), J2 is very frequent in the Levant/Anatolia/Iran region (140) and its spread in the Mediterranean might have been facilitated by the Phoenicians (1550 BCE- 300 BCE) (119). Finally, a smaller contribution has been described from sub-Saharan Africa (120,141).

The analysis of 44 Y-chromosome biallelic polymorphisms in population samples from NW Africa and the Iberian Peninsula (120), suggested an Upper Paleolithic origin of north-western African Y chromosomes. By contrast, in another study by Arredi *et al.* (142), the presence of specific lineages in North Africa was attributed to the Neolithic expansion from the Middle East. Moreover, by typing 275 men from five populations in Algeria, Tunisia, and Egypt with a set of 119 binary markers and 15 microsatellites from the Y chromosome, Arredi and colleagues (142) showed an east-west cline of genetic variation that extends into the Middle East and is compatible with a hypothesis of demic expansion. In addition, given the reduction in gene diversity and the frequency increase of Y haplogroup E1b1b1b-M81 towards the West (Figure 12), this expansion might have involved a relatively small number of Y chromosomes.

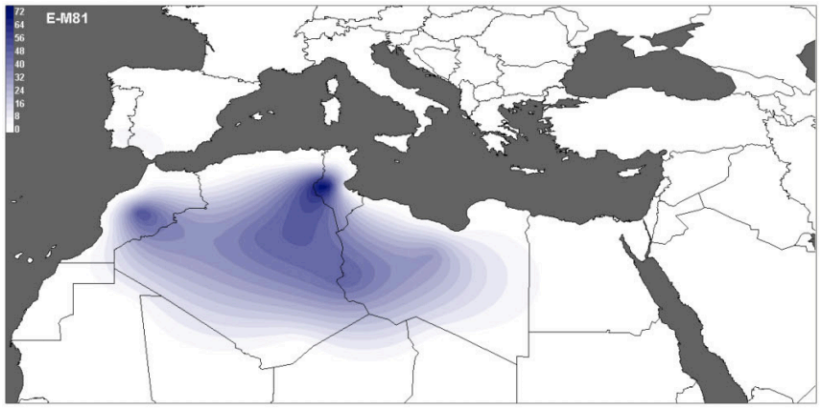


Figure 12. Frequency distribution of haplogroup E-M81 (143).

Until recently, information from paternal, maternal, and biparental molecular markers has been assessed independently. In an attempt to bring together uniparental and genome-wide data, Fadhlaoui-Zid

et al. (143) showed that both males and females in North Africa underwent a similar admixture history with slight differences in the proportions of admixture components. Furthermore, principal component analysis (PCA) using haplogroup frequencies showed that the E-M81 lineage drove the split of all North African populations (except from Egyptians) from the rest (Figure 13). Finally, they showed that the ancestors of most current North Africans emerged 15,000 years ago, after the Last Glacial maximum, with no traces of genetic continuity with the first human settlers in the region.

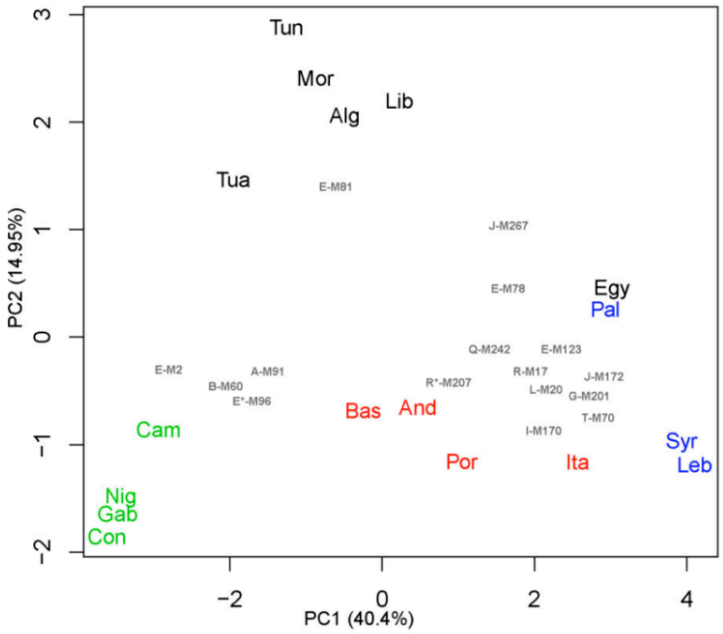


Figure 13. Principal component analysis of haplogroups frequencies. B) Multidimensional scaling plot based on RST distances between populations derived from Y-STR data. Modified from Fadhlaoui-Zid *et al*. (143)

2. METHODS

In the following section I will describe the laboratory procedures (section 2.1) and the computational processing that made this work possible (section 2.2). Finally, I will briefly cover the theoretical framework of the main analysis performed (section 2.3).

2.1. Laboratory procedures

2.1.1. DNA extraction

Before studying any kind of genetic diversity, we must find and assay this variation. Possible sources of human DNA are, among others, blood, saliva cells, semen, hair and cell lines, which mostly differ in the amount of DNA that they yield. Most of the DNA used in the present work has been obtained from cells in saliva and to a lesser extent from blood cells.

Despite blood being the most popular source of human DNA and the preferred source for massive parallel sequencing, taking blood requires appropriately qualified personnel and involves some risk of contagious diseases, such as HIV or hepatitis B. By contrast, saliva offers a non-invasive source of DNA and can be self-sampled, which has favoured the use of this method when sampling individuals for genetic profiling or recreational genomics. A 5ml of blood sample yields 50-200 μg of DNA, whereas the amount of DNA obtained by rubbing in the inside of the check using a brush

or swab is less than <1 μg. An alternative to buccal swab is a "spit kit" such as the one implemented in the Oragene™ kit, which can yield up to 100 μg of DNA and that is suitable for high-throughput SNP analysis and massive parallel sequencing (1). Nevertheless, instead of using a commercial kit, for most of the work presented in this thesis, we collected 2mL of saliva in a falcon tube filled with 2mL of stabilization buffer. This approach yielded an amount of DNA similar to those obtained with the Oragene™ kit, but reduced the costs by two orders or magnitude.

Protocols used to extract DNA from saliva and blood cells are different, but share some general steps. First, in the cell lysis step, we break the cell membranes in order to release the DNA out of the cell. This membrane degradation can be performed with detergents and surfactants (e.g. SDS), which will degrade the lipids from the cell membrane and the nucleus; or it can be mediated by proteinase K, which breaks the proteins located in the cell membranes. Once the DNA has been released, an RNAase solution is used to degrade the RNA. After that, this solution is treated with a concentrated salt solution in order to pool together degraded proteins, lipids and RNA. Then, this cluster of cellular waste is removed by centrifugation. Finally, the DNA is separated from the detergents, proteins, salts and reagents previously used during the cell lysis step. This process is called DNA purification and consists of two basic steps: First, given that DNA is insoluble in alcohols such as ethanol or isopropanol, it will precipitate in their presence and aggregate, giving a pellet upon centrifugation. Then, by adding a

phenol-chloroform solution we will extract DNA. Since phenol denatures proteins, after centrifuging the sample, denatured proteins stay in the organic phase while aqueous phase containing nucleic acid is mixed with the chloroform that removes phenol residues from solution. After isolation, DNA is dissolved in a slightly alkaline buffer, such as ultra-pure water (144) and stored at 4ºC.

2.1.2. Genotyping: STRs and SNPs

As stated in section 1.1.1, variation between Y chromosomes exists over a wide range of scales, from single bases to chromosomal rearrangements. Therefore, the methods employed to detect and assay such variation will depend on the scale of the variation under study.

As will be discussed below, using massive parallel sequencing we can obtain the genotypes of all positions in the Y chromosome sequence. However, in some cases we might be interested in the genotypes at particular positions in a DNA sequence. In general, this is because we have prior knowledge on these positions, but also because it is easier and cheaper to focus only on them. Although plenty of methods to target specific positions can be used, we will focus on those used in the present work to genotype specific Y chromosome SNPs and STRs.

Microsatellites are amplified with a polymerase chain reaction (PCR) that allows the subsequent detection of single-repeat-unit size differences between alleles. First, DNA is repeatedly denatured

at a high temperature to separate the double strand, then cooled to allow the annealing of primers and the extension of nucleotide sequences through the microsatellite. PCR primers consist of unique sequences at the flanking regions of the STR labelled at its 5' end with a fluorescent dye, so that PCR products can be detected on capillary-based sequencing platforms using laser technology. In addition, the combination of multiple fluorescent dyes allows the simultaneous amplification and efficient separation of the several Y-STR loci during automated DNA fragment analysis.

Short tandem repeats analysed in the present work have been obtained with the Yfiler™ PCR amplification kit, which simultaneously amplifies 17 Y-STR loci including the loci in the ''European minimal haplotype'' (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393), the Scientific Working Group on DNA Analysis Methods (SWGDAM) recommended Y-STR loci (DYS438 and DYS439), and the highly polymorphic loci DYS437, DYS448, DYS456, DYS458, Y GATA H4, and DYS635 (145). Results obtained are then analysed with the GeneMapper® ID or ID-X Software. The 17 Y-STR profile of an unknown male is displayed in figure 14 (1,3,145).

Despite the huge variety of existing methods, binary markers used in this project have been mainly genotyped with TaqMan probes (Applied Biosystems, Inc., Foster City, CA, USA). TaqMan assays can be used to type SNPs in single reactions, that is, one SNP per TaqMan reaction, but they can also be employed to genotype many SNPs in a single array, a procedure known as multiplexing (146).

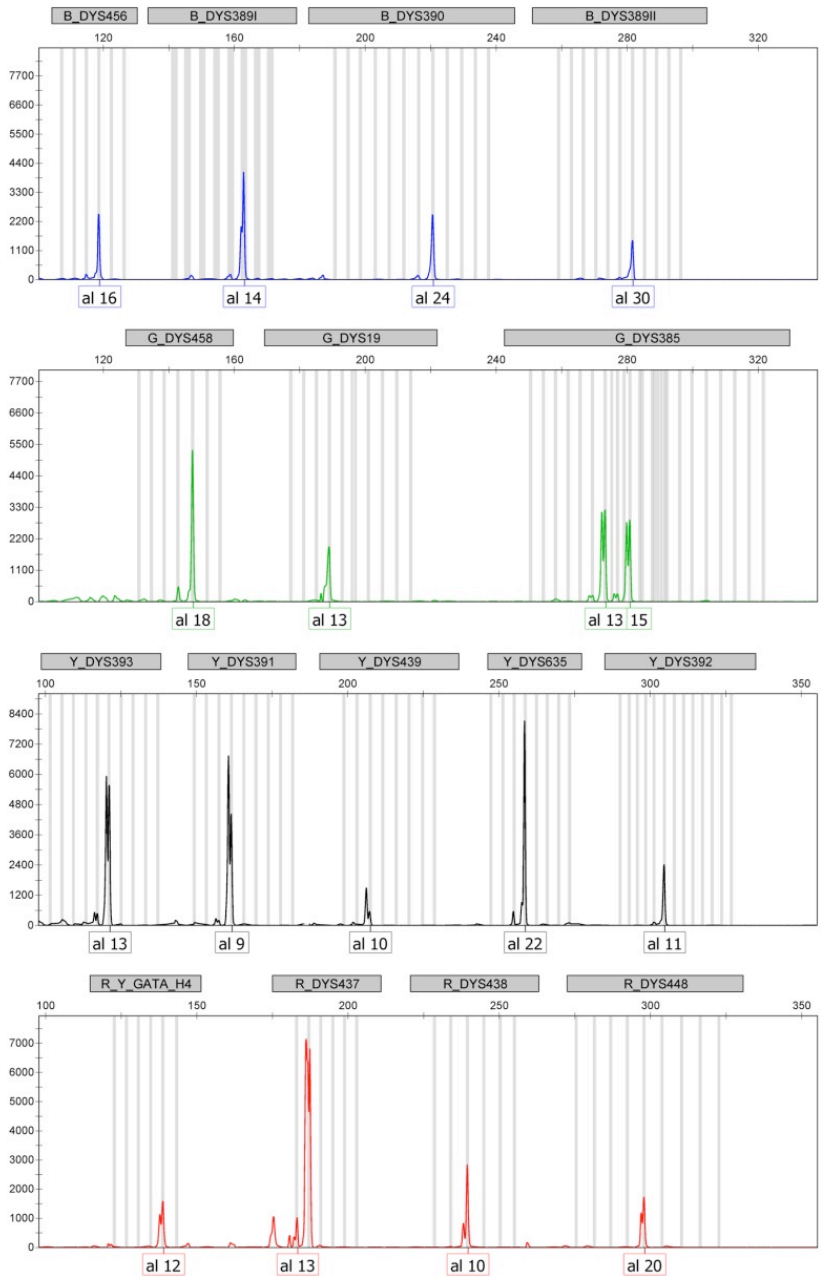


Figure 14. Representative electropherogram from GeneMappers ID software v3.2 showing the profile of a control male sample. The four panels correspond to 6-FAMTM (blue), VICs (green), NEDTM (black), and PETs (red) dye-labeled peaks. The haplotype is shown with the allele number displayed underneath each peak.

In order to genotype multiple Y-chromosome SNPs in a single reaction, we have used the Taqman OpenArray® Genotyping plates. This plates consist on a 63-mm × 19-mm plate divided into 48 subarrays. As shown in figure 15, each subarray consists of 64 through-holes. Each through-hole in a genotyping plate may contain a single assay. When you order a genotyping plate, you select the TaqMan assays to include in the plate, which are dried-down and preloaded into it. The number of assays in the genotyping plate and the number of samples you can load in the plate depend on the format you select. There are six TaqMan OpenArray® Genotyping Plate formats available. In the present project, we selected the 64 format, in which the genotyping plate is preloaded with 64 assays and you can load up to 48 samples. The samples are loaded into the OpenArray® 384-Well Sample Plate, which is a 384-well microtiter plate divided into eight areas; each sample plate area is 12 wells × 4 wells (48 wells). It is crucial to track where the samples are in the sample plate, in order to identify the genotype of each sample. (tools.thermofisher.com/content/sfs/manuals/cms_058198.pdf)

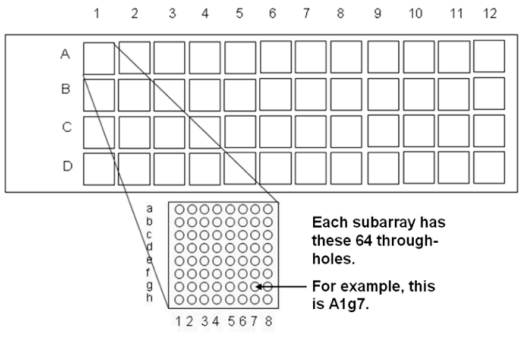


Figure 15. Distribution of a Taqman® OpenArray® Genotyping plate. (tools.thermofisher.com/content/sfs/manuals/cms_058198.pdf)

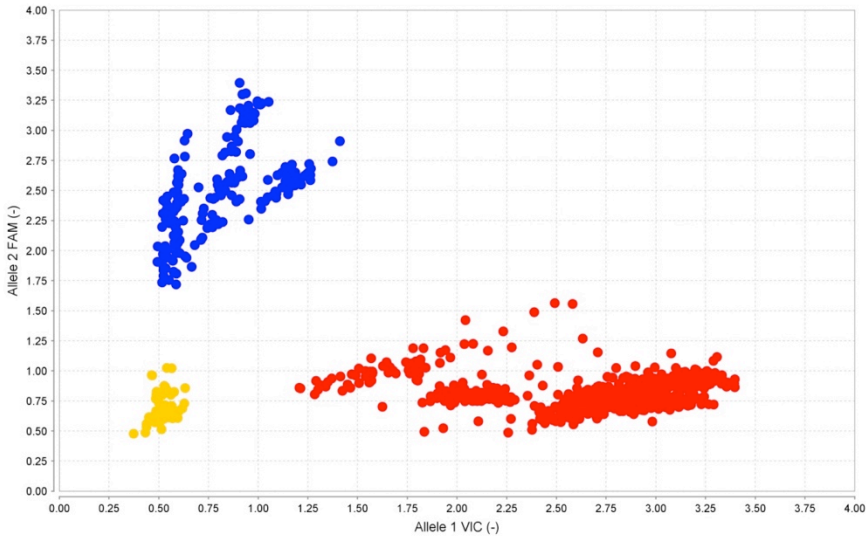


Figure 16. Plot showing the fluorescence of >200 samples for the M183 Y-SNP. Blue dots (FAM) correspond to the ancestral allele (A) and red dots (6-VIC) to the derived allele (C). In yellow, the negative controls.

Regardless of performing a single Taqman reaction or multiplexing, the principle behind this method relies on the 5´ exonuclease activity of the enzyme *Taq* polymerase. A TaqMan assay contains two probes, each one complementary to one of the two alleles (ancestral or derived) of the targeted SNP; and a specific set of primers, which amplify the region containing the polymorphic site. Each of these probes is labelled with a different fluorophore at the 5’ end (usually VIC and 6-FAM, Applied Biosystems) and a quencher at the 3' end, which prevents fluorescence of the fluorophore. During the PCR amplification step, if the allele-specific probe is not perfectly complementary to the polymorphic site, it will not bind efficiently, preventing the exonuclease activity from acting on the probe. Otherwise, if the allele-specific probe is perfectly complementary to the polymorphic site, it will bind to the

target DNA strand and be degraded by the *Taq* polymerase. This degradation releases the fluorophore, relieving the quenching effect and allowing fluorescence of the fluorophore. The amount of fluorescence observed enables the identification of the allele carried by any particular sample (146,147). Figure 16 shows a successful assay for the M183 Y-SNP.

Another approach used in this project to target binary polymorphisms is Sanger sequencing. Introduced by Sanger and colleagues (148) in the late seventies, this sequencing method became the primary technology of the so-called first-generation sequencing methods and remained as the prevailing technique for the next 30 years (1,3).

Sanger sequencing relies upon the synthesis of a new DNA strand initiated from specific primer by incorporating modified nucleotides, known as dideoxynucleotides (Figure 17a). These are labelled analogues of the four usual nucleotide bases (A, T, G, and C), but once they are incorporated, they prevent further elongation of the new DNA strand, and thus the new DNA strand terminates at that position. As a result, we end up with DNA fragments of different lengths (Figure 17b), each one terminated at one of the four specific dideoxynucleotides. Then, knowing which dideoxynucleotide has been incorporated, the DNA sequence at that position can be ascertained. Initially, fragment detection was by autoradiography of radiolabelled fragments separated by gel electrophoresis. Later on, the development of fluorescent labels that could be detected by laser excitation, and the replacement of gels by

capillary electrophoresis, enabled a better detection and automation, and increased the throughput of Sanger sequencing (Figure 17c) (1,3).

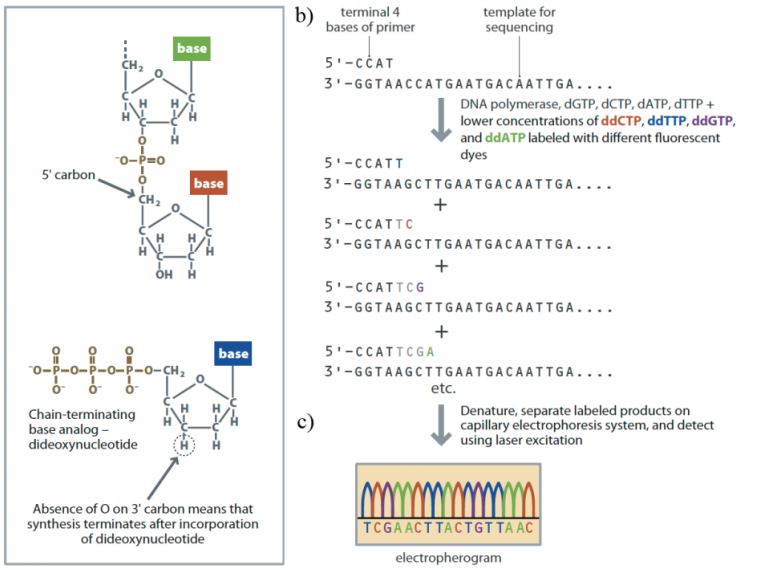


Figure 17. Principle of Sanger sequencing of DNA (1).

2.1.3. Massive parallel sequencing

The introduction of important improvements in the Sanger method enabled the completion of the first human genome sequence by the Human Genome Project in 2004 (149), which last 10 years and cost around $3 billion (150). However, the low throughput, low speed and high cost of first-generation methods led to the development and commercialization of massive parallel sequencing technologies (MPS), also known as next generation sequencing (NGS), or second generation sequencing (SGS). These methods can generate enormous amounts of data at a fraction of the cost of Sanger-sequencing (Figure 18) (3).

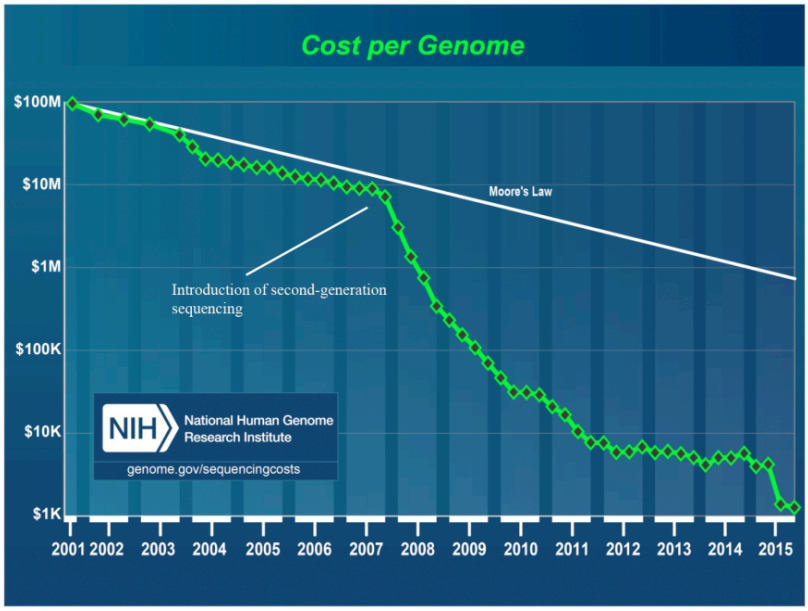


Figure 18. Cost per raw megabase of DNA sequence ($) (green line). The line labeled “Moore's Law" describes the long-term trend whereby the cost of computing halves every two years. Adapted from www.genome.gov/sequencingcosts/.

The first commercially successful MPS technology was the Roche 454 pyrosequencing method, released in 2005. One year later, Solexa (later purchased by Illumina) sequencing platform was commercialized, followed by SOLiD technology in 2007 and Ion Torrent in 2010 (151). These platforms represent the major SGS technologies, and have introduced several improvements in sequencing chemistry and base-calling software since the original releases. Finally, Illumina technology, which is currently the leader in the MPS industry, offers the highest throughput and the lowest per-base cost (1). Table 3 compares the different sequencing platforms mentioned.

Table 3. Comparison of the most widely used sequencing platforms. Adapted from Jobling *et al.* (1).

Platform	Read length (bp)	Run time (days)	Gb/run/ machine	Pros	Cons
Capillary Sanger sequencing	850	1	≤0.001	accurate; useful for validation of MPS data	low throughput, expensive
Illumina HiSeq 2000/2500	2 x 160 (paired end)	11	600	massive throughput	massive throughput and high costs can be a disadvantage for some projects
Life/APG SOLiD	2 x 50 (paired end)	11-16	100	high throughput	lengthy run times
Roche GS FLX Titanium XL + (455 sequencing)	1000	1	0.7	long reads aid mapping/assembly	inaccurate sequencing of homopolymeric tracts; expensive
Ion Torrent	2 x 300 (paired end)	0.1	1	rapid runs, cheap instrument; useful for SNP validation	inaccurate sequencing of homopolymeric tracts

Before sequencing, all currently available platforms require some level of DNA pre-processing. First, DNA molecules are fragmented to a specific average size through a process of sonication, which consists on exposing DNA in solution to ultrasound waves. The longer the exposure, the shorter the DNA fragments. Then, artificial DNA segments, called adapters, are attached to both ends of the fragments, so that they are recognized by the sequencing platform; this process is known as "library preparation".

Rather than detailing the diverse sequencing methods, we will focus on the MPS approach used in the work presented in this thesis: Illumina technology.

2.1.3.a. Library preparation

Illumina has two main library preparation methods: single-strand and double strand library preparation. In this thesis, we used the

double-stranded approach, which consists on incorporating the adapters to a double strand DNA molecule. First, during the end-repair step, 3' overhangs are removed and 5' overhangs are filled in to create blunt ends. After that, short adapters are ligated to the ends of a double-stranded DNA molecule (152). Then, PCR is used to amplify the adapter-ligated DNA molecules and to add a platform-specific adapter. Finally, if we aim to sequence multiple libraries in the same sequencing run, sample-specific indexes should be attached to each fragment of the library during amplification (153,154). These indexes will be later used to computationally assign each read to its corresponding sample.

2.1.3.b. Sequencing

All the steps of Illumina technology take place in a flow cell, which can be partitioned in different lanes. The surface of the flow cell lanes is densely coated with forward and reverse primers, which are complementary to the adapter sequences introduced during library preparation (155). The first step is denaturalizing the double-stranded library fragments into single-stranded molecules. Then, these molecules are loaded onto the flow cell and hybridized at one end to the surface primers. The free 3' end of the fragments bends over and hybridizes to a complementary primer on the surface, forming a bridge structure that allows the synthesis of the complementary strand (Figure 19).

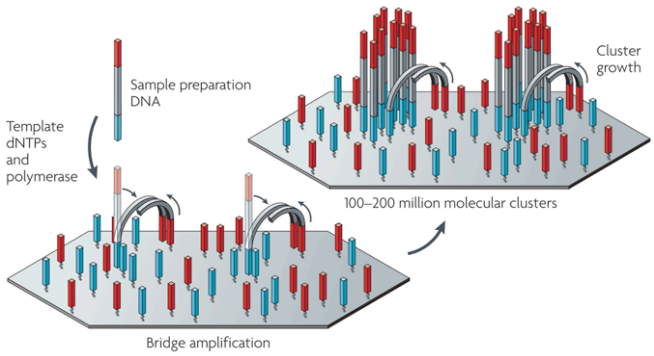


Figure 19. Cluster generation on the Illumina sequencing platform. Modified from (154).

This process of bridge amplification creates a cluster of multiple copies for each original library fragment. Then, the double-stranded bridge structure is denatured and a new amplification cycle begins. This bridge amplification process is necessary because the final detection method is not sensitive enough to detect single-molecule fluorescence. Overall, a flow cell contains millions of spatially separated clusters, each with ~1,000 copies of an original library template.

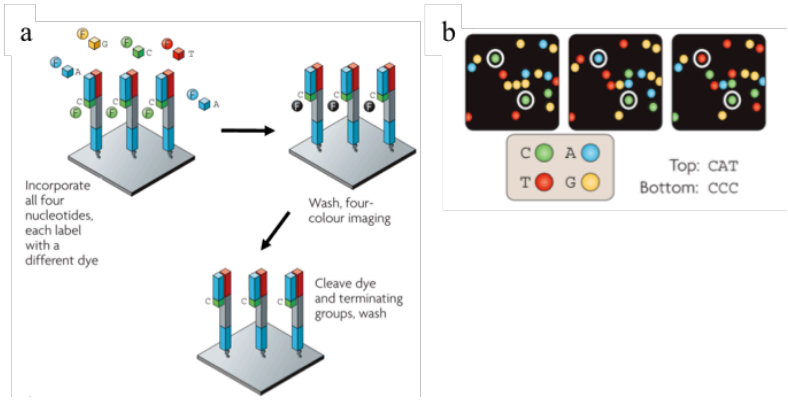


Figure 20. Sequencing on the Illumina platform. a) One cycle of nucleotide incorporation, imaging and cleavage. b) Three sequencing cycles highlighted for two clusters. Adapted from Metzker *et al.* (156).

After the cluster generation, sequencing is performed using a cyclic process that comprises three steps (Figure 20a). First, sequencing primers are annealed to each cluster template and a DNA polymerase incorporates a fluorescently labelled terminating nucleotide to the growing strand, which is complementary to the template base. During the second step, the remaining unincorporated nucleotides are washed away and fluorescent imaging is performed to determine identity of the incorporated nucleotide. Finally, a cleavage step removes the fluorescent dye and the 3' terminator, and an additional wash prepare the DNA strand for a new cycle. Since one nucleotide is incorporated in each cycle, the number of cycles determines the length of the resulting reads, which typically is about 200 bp (Figure 20b) (155).

Sequencing on the Illumina platforms can be run in single-end mode or paired-end mode. If single-end sequencing is performed, only one end of the DNA fragment is sequenced using the primer site present in one of the adapters. During paired-end sequencing, once the sequencing from one end is finished, the complementary strand is synthesized and the sequencing priming site in the second adapter is used to sequence the other end of the DNA fragment (155). The sequences analysed in the present work have been obtained by paired-end sequencing. The advantages of this approach is that it yields higher number of sequence reads, and produces pairs of sequences that are known, a priori, to be separated by the average length of each fragment (1).

The error rate associated with massive parallel sequencing platforms tends to be quite high, and around 10 times higher than the error rate per base for Sanger sequencing. To overcome this issue, each base is sequenced many times, in the so-called shotgun fashion. The average number in which each nucleotide is sequenced is called the coverage, so a 20X sequence coverage means that each nucleotide has been sequenced independently an average of 20 times. In general, the higher the coverage, the better the sequence accuracy. Noticeably, the Y chromosome exhibits half of the coverage of the rest of the genome, because it is carried as a single copy. Besides the coverage, what tends to be more problematic when sequencing Y chromosomes is the short read length, due to its repetitive regions (3). In the following section, I will describe the computational strategies used to handle MPS data.

2.2. Bioinformatic processing

2.2.1. Mapping

While generating sequencing data is becoming easier and cheaper, the computational storage and analysis of MPS data is becoming a difficult task. The first step is mapping the huge number of reads sequenced. Several software solutions have been developed for mapping sequencing reads to reference genomes (157–160). The Y-chromosome reference sequence is 59.36 Mb, but this includes a 30-Mb stretch of constitutive heterochromatin on the q arm, a 3-Mbcentromere, and 2.65-Mb and 330-kb telomeric pseudoautosomal regions (PAR). As a result, Y-chromosomal reads are mapped to the remaining 22.98 Mb of assembled reference sequence (37).

The approach used in the work presented in this thesis is the Burrows-Wheeler Alignment tool (BWA), a read alignment package based on backward search with Burrows–Wheeler Transform (BWT) (157). BWA efficiently aligns short sequencing reads against a large reference sequence such as the human genome, and allows mismatches and gaps. Moreover, BWA outputs alignment in the standard SAM (Sequence Alignment/Map) format, to take advantage of the variant calling and other downstream analyses implemented in the open source SAMtools software package. Finally, BWA always requires the full read to be aligned, therefore, for longer reads, which are likely to be interrupted by

structural variations, it is advisable to divide them into multiple short fragments and align the fragments separately.

After library preparation, DNA libraries are PCR-amplified in order to reach the amount of DNA required for sequencing. This leads to an excess of clonal molecules in the sequencing reads, which must be identified and removed as they can severely bias downstream analyses such as coverage estimation or polymorphism detection. The most common approach for identifying PCR duplicates of the same original DNA fragment is based on the detection of reads mapping on the same strand with the same outer alignment coordinates. This task can be performed with standard bioinformatics tools such as the *rmpdup* tool from Samtools (161) or the *MarkDuplicates* program from Picard tools (http://broadinstitute.github.io/picard/). Once PCR clones are detected, these tools select the read with the highest sum of base quality scores, while the rest are either removed from the alignment file or flagged as PCR duplicates.

2.2.2. Variant Calling

Several software solutions have been developed to identify genomic variants, such as SNPs and DNA insertions and deletions. Among many variant callers, the Genome Analysis Tool Kit (GATK) is one of the most widely used in genomic variant analyses (162). The GATK framework has provided several variant calling tools, such as UnifiedGenotyper and HaplotypeCaller.

In the present project, Y chromosome variants have been called using the UnifiedGenotyper tool. This approach uses a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of N samples, producing a genotype for each sample. This tool can either output only variant sites or complete genotypes. This variant caller can be very sensitive but it is prone to false positives, therefore some specific post-calling filters will be needed to eliminate most of these false positives (section 2.2.3). Currently, this tool has been superseded by HaplotypeCaller, which is more a sophisticated variant caller that produces much better calls, especially on indels. It also includes features that allow it to scale to much larger cohort sizes. However, at the moment when this work was conducted, UnifiedGenotyper was by far the best option (162,163).

Before calling variants, an indel realignment step is usually needed. Indels in sequence reads (especially near the ends) often are mapped with mismatching bases that might look like evidence for SNPs, but are actually mapping artefacts. These artefactual mismatches can compromise variant detection and therefore should be corrected. The GATK framework has provided a two-step pipeline to identify the most consistent placement of the reads with respect to the indel. First, the intervals that need to be realigned are identified by the RealignerTargetCreator tool. Then, the IndelRealigner tool determines the optimal consensus sequence and performs the actual realignment of reads (163).

2.2.3. Y chromosome filtering

After the variant calling, the raw variant calling file (vcf) obtained needs to be filtered based on certain criteria. A commonly used GATK tool for filtering variants after the calling step is the variant quality score recalibration (VQSR). However, it requires a large number of variant sites to operate properly and thus, it is not suitable for some small-scale experiments or targeted regions such as the Y chromosome itself. An alternative to VQSR is provided by the VariantFiltration tool (GATK) which allows to hard filter variants on small datasets to reduce the amount of false positive calls. Normally, this filtering focus on mapping quality (reads with ref bases vs. those with the alternate allele), strand bias (variation being seen on only the forward or only the reverse strand) and quality by depth (variant confidence divided by the unfiltered depth of non-reference sample) (163).

Another important question about MPS data is which areas of the genome or the region under study (i.e. Y chromosome) are considered callable. The callable loci tool (GATK) considers the coverage at each locus and provides a summary interval BED file with statistics on callable, uncallable or poorly mapped variants. Therefore, this tool allows to build a callability mask based on a certain coverage threshold (163). The callability mask used to filter the variants analyses in the work presented here was based on a read depth >5.

Moreover, given the high degree of self-identity within the ampliconic segments and the X-chromosome homology of the X-transposed region, we restricted our analysis to some high quality regions defined by Wei *et al.* (12) within the male specific region of the Y chromosome (MSY) (6). Wei and colleagues (12) excluded the ampliconic, pseudoautosomal and X-transposed regions from the Y chromosome, obtaining a nine-segments 8.97 Mb sequence in which read mapping and variant detection are expected to avoid complications introduced by repeated sequences.

2.3. Data analysis

Past demographical events such as population size changes and migrations, are somehow written in the DNA of current human populations and thus, can be assessed by the analysis of genetic data. In the following sections, I will describe the main analysis performed in the present work, which has been based on the works of Jobling *et al*. (1) and Stoneking *et al*. (3).

2.3.1. Summarizing genetic diversity

Several statistics can be used to summarize the amount of variation and to compare individuals or loci. Heterozygosity is probably one of the simplest ways to summarize the amount of variation and it can tell us about the structure of a population. It measures the probability that two alleles drawn at random from the population will be different from each other. Heterozygosity values range from 0, which means that everyone has the same allele, to 1, when everyone has a different allele. Here, we focused on expected heterozygosity ($H_{exp,}$ or gene diversity, D), which is equal to the observed heterozygosity as long as Hardy-Weinberg's law is met. The simplest way to calculate it for a single locus is as:

$$H_{exp} = \left(1 - \sum_{i=1}^{q} x_i^2\right)$$

where q is the number of different alleles, x_i is the frequency of the i-th allele. This statistic can be estimated either within many

populations for a single locus (differences between populations), or at many loci within a single population (differences within populations). Moreover, this measure is blind to molecular distance between alleles and thus, is applicable to both classical alleles and molecular haplotypes. A serious disadvantage of this measure is its dependence on sample sizes.

Another commonly used measure of diversity, highly correlated to the expected heterozygosity, is Nei's gene diversity. As defined below, this measure is corrected by sample size, in order to provide an unbiased estimator of population heterozygosity:

$$H = \left(1 - \sum_{i=1}^{q} x_i^2\right) n/(n-1)$$

where q is the number of different alleles, x_i is the frequency of the i-th allele and n is the number of alleles or chromosomes in the sample. This means that this equation also holds for haploid systems: for mtDNA and the NRY, n corresponds to the number of sampled individuals; and for autosomal loci, n means the number of alleles and thus, twice the number of individuals.

In some cases, where almost all alleles (or haplotypes) are different from one another and thus gene diversity is close to 1 in all populations, we must consider the molecular nature of allelic variation by using measures of diversity that take account of the distances between alleles. A common way to summarize the observed diversity within a set of nucleotide sequences, is counting

the number of nucleotide sites that vary within the entire set of aligned sequences, known as the segregating sites. However, this measure is clearly dependent on the sequence length.

One measure of genetic variation that takes sequence length into account is the mean number of pairwise differences. It can be estimated by counting the differences between every pair of sequences in a sample, then sum the number of differences across all pairs of sequences and divide by the total number of pairs. Nevertheless, this measure depends on how many bases are sequenced, which becomes an issue when we aim to compare diversity estimates for different DNA segments. The longer the DNA segment, the more polymorphic sites you expect, and hence the more pairwise differences between individuals.

Nucleotide diversity, which is analogous to Nei's gene diversity, takes all of these factors into account, and describes the probability that two copies of the same nucleotide drawn at random from a set of sequences will be different from one another. For a sequence, it is the average expected heterozygosity at the nucleotide level. The definition of is shown below:

$$\pi = \left(\sum_{ij}^{q} x_i x_j \, d_{ij} \right) n/(n-1)$$

where q is the number of different allelic sequences, x_i and x_j are the frequencies of the *i*-th and *j*-th sequences respectively, and d_{ij} the proportion of different nucleotides between them. Nucleotide

diversity is a highly useful measure of genetic diversity, not only because it can be directly compared between studies even when sample sizes and/ or DNA segment lengths differ but also because it turns out to be an estimate of $4N_e\mu$ (also known as θ), where N_e is the effective population size and μ is the neutral mutation rate.

Another statistic used in the present work is the variance, which is a standardized measure of the range of observed values relative to the average. As shown below, it is computed as the sum of the squared deviations (x_i) from the mean ($\bar{x}$) divided by one less than the sample size (n). Specifically, we have computed the mean STR variance, which is the average of STR repeat size variances over all STR loci.

$$variance\ (\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

2.3.2. Measuring genetic distance

Measures of genetic distance are statistics that allow us to compare the relatedness of populations or molecules, and their numerical value increase as the evolutionary distance is greater.

Although several methods for estimating genetic distances between populations exist, one of the most widely used method and applicable to any kind of genetic data is the F_{ST} statistic. It can be defined as the difference between the average expected heterozygosity within populations and the expected heterozygosity

across all populations. Alternately, it can be understood as the probability that two alleles sampled at random from a single subpopulation are identical given the probability that two alleles sampled from the total population are identical.

There are several advantages of using F_{ST} as a measure of genetic distance. First, it can be computed either for pairs of populations or for many populations at once. It can also be estimated for a single genetic locus or averaged over many loci. Moreover, as it can be computed for any kind of genetic data, it allows the direct comparison of different genetic markers. On top of that, this measure has a natural and straightforward interpretation, as it reflects the proportion of the total genetic variance that is due to differences among populations. And indeed, it can easily be extended to additional hierarchical subdivisions of the total genetic variance. Besides its several advantages, F_{ST} also has some drawbacks. For example, when two populations are each polymorphic for different alleles at a locus, which means that they are totally differentiated (F_{ST} =1), due to mathematical issues F_{ST} will be underestimated.

An alternative to allele frequency-based methods such as Fst was introduced in the early 1990 by Laurent Excoffier and colleagues (164). They implemented a useful statistical framework named analysis of molecular variance (AMOVA), which takes into account the molecular differences in the alleles at a given locus (number of substitutions for DNA sequences, number of repeat differences for STR alleles, etc.), rather than just their frequency, when

apportioning variance between levels of the hierarchical population structure. The AMOVA analysis can be applied to any data where genetic distances between alleles can be calculated, and it can be carried out by an extremely useful and user-friendly program called Arlequin (165).

Displaying genetic distances is not as straightforward as it seems. If we have *n* populations, we require *n* dimensions to fully display their pairwise genetic distances on a graph. Multivariate analyses allow us to reduce these multiple dimensions to only two or three dimensions, in order to represent genetic distances in a more comprehensible manner, albeit with loss of information. One of these methods is Multi-Dimensional Scaling (MDS), which uses a pairwise distance matrix as an input and place the populations in a space consisting of fewer dimensions while maintaining the distance relationships among the populations as closely as possible. Rather than going through the equations to do this, it is better to focus on how to evaluate and interpret MDS plots. The stress value compares the observed distance values between each pair of populations to those obtained from the plot and thus, it provides an evaluation of how well the resulting plot retains the structure of the data. The lower the stress statistic, the better the MDS fit to the data and, therefore the less information is lost.

Another widely used method for displaying genetic distances is Principal Component Analysis (PCA), which also extracts the most important information in multivariate data. Specifically, PCA transforms a series of correlated variables into a smaller number of

dimensions or axes called principal components (PCs), which represent relatedness between individuals. These PCs, also known as eigenvectors are extracted sequentially, with each PC being independent and encapsulating as much of the remaining variation as possible. As a result, the first PC captures most of the information. It contrasts with MDS, where axes do not have any inherent meaning. Another important difference with MDS is that, rather than using genetic distances, PCA can be performed using the raw data of allele frequencies. Finally, as shown in figure 21, PCs have been used to construct synthetic maps that summarize information from several alleles with similar geographic distributions. Nevertheless, these maps should be interpreted with caution.

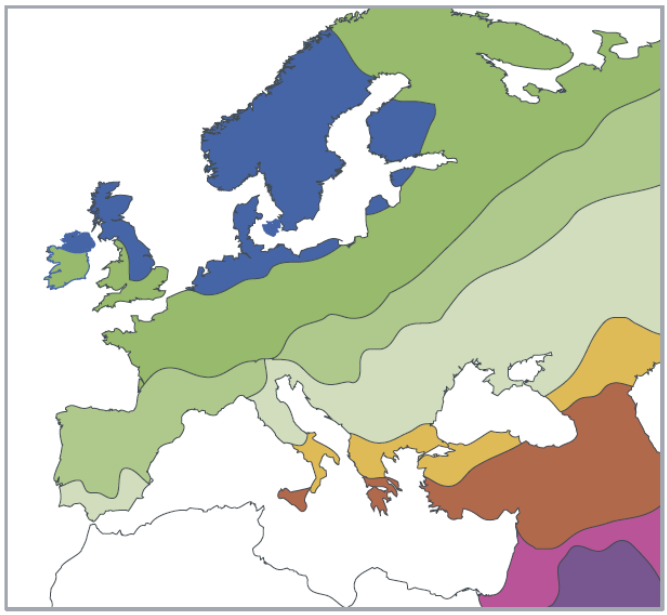


Figure 21. Synthetic map of Europe and Western Asia obtained using the first PC of classical genetic data (124).

2.3.3. Phylogenetics

Phylogenetic trees are one of the most widely used methods for displaying the relationships between different biological entities, and in some cases, such as the NRY, they also encapsulate the mechanism by which diversity arose. Despite in the present work we have focused on the human Y chromosome phylogeny, trees are actually used across a wide variety of disciplines and for a broad range of purposes.

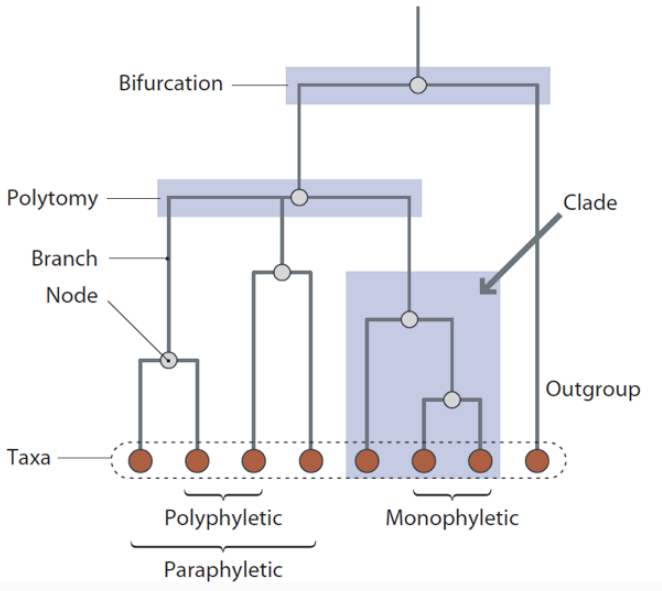


Figure 22. Terminology for phylogenetic trees (1).

Phylogenetic trees have their own distinctive terminology (Figure 22). A tree consists of branches between nodes, and the branching pattern of a tree is known as its topology. The descendants of a single node form a clade and, when only two branches descend from a node, it is called a bifurcation. Alternatively, when more

than two branches descend from the same internal node, it is known as a polytomy (or multifurcation). Sometimes branch lengths between nodes are irrelevant, and the tree only display the relationships between taxa, this is the case of cladograms. By contrast, in additive trees branch lengths reflect evolutionary distance quantitatively.

Another feature of phylogenetic trees is that they can be rooted or unrooted. A tree can be rooted when the ancestral node to all taxa is known. Moreover, since evolutionary changes are assumed to have occurred from ancestral to derived states, the root defines the directionality in the tree and thus, the proximity of a node to the root determines the relative antiquity of the divergence event. When we ignore the ancestral node (unrooted tree), we can assume that the root falls midway along the longest branch on the tree or, alternatively, we can incorporate an external outgroup to all other taxa.

Phylogenetic trees aim to establish the relationship between groups of populations or molecules, known as taxa or operational taxonomic units (OTUs). If a group of taxa fall into a single clade, it is known as monophyletic. By contrast, if this grouping excludes other members of the same clade it is paraphyletic. Finally, taxa that span multiple clades are polyphyletic.

There are many different methods for constructing phylogenetic trees, which are generally classified depending on the input data type and the means by which a tree is constructed. Input data can be

either genetic distances or characters, such as SNPs or microsatellites. The two main classes of phylogeny construction methods are clustering and searching methods. Whereas clustering methods use an iterative algorithm to combine taxa together in a hierarchical fashion (one by one), the idea behind searching methods is to find the tree that best fits the data within the whole range of possible trees, according to some optimality criteria. When choosing a method, we should take into account that not all of them are suitable for all the different types of data available. Here, we will focus on the principle of maximum parsimony (MP), a searching method based on character data.

As mentioned in section 1.1.1, unique event variants of the Y chromosome can easily be combined into monophyletic haplogroups. Due to the absence of recombination, these monophyletic haplogroups can be related by a single phylogeny using the principle of maximum parsimony (MP). MP defines the best tree as the one that requires the smallest number of evolutionary changes to account for the data, and its branch lengths correspond to the numbers of individual evolutionary changes along each branch. Since it is computationally expensive to examine all possible trees, this method only considers those polymorphic sites where at least two alleles are present in two or more individuals, known as informative sites. The searching strategy will jump between different locales within the range of possible trees and will identify the most parsimonious tree in each locale. In some cases, two (or more) trees can be equally parsimonious and thus, no unique tree can be inferred.

Noticeably, MP methods can incorporate information about the relative rate of different mutations, allowing different mutational events to be weighted accordingly, with the rarer changes carrying more influence. Finally, by a process known as branch attraction, MP methods can take account of unequal rates of evolution or undersampling of certain parts of the phylogeny.

Once a phylogeny has been obtained, to statistically assess how confident we can be about it, we use the bootstrap. The idea behind the bootstrap is that, if a dataset strongly supports a certain tree, it should also be supported by randomly chosen subsets of the data. Bootstrap values are usually displayed as percentages next to the nodes to which they refer. For example, a value of 90 means that the same node was reconstructed from 90% of all random subsets.

In general, since evolutionary time progresses toward present, branches diverge but never coalesce. Nevertheless, some biological processes are not well represented by phylogenies in which taxa split but never coalesce. Processes such as recombination, gene flow, parallel mutation, will generate four-sided closed structures known as reticulations. A type of tree that incorporate such structures are known as networks. As any other tree, networks can be either constructed from genetic distances or character data. There are a number of methods of construction, such as the minimum spanning or the median network. In the work presented in this thesis, we used the median-joining (MJ) method (Figure 23).

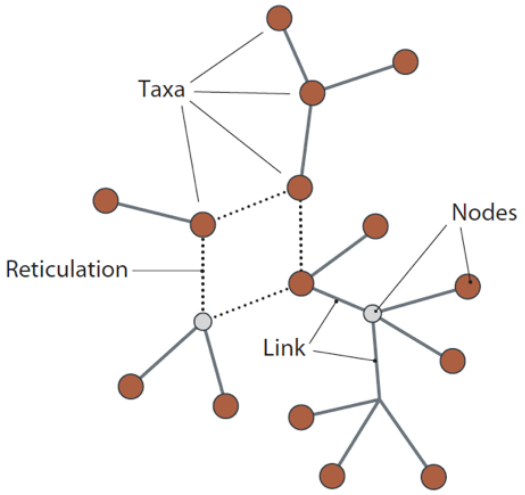


Figure 23. Terminology for network trees (1).

The phylogenetic MJ network algorithm was introduced by Bandelt and colleagues (166) and combines features of the minimum spanning algorithm for finding trees by favouring short connections, and the maximum-parsimony (MP) heuristic algorithm, which sequentially adds new vertices called median vectors, except that MJ method does not resolve ties. MJ allows constructing networks with limited levels of reticulation and is applicable to multiallelic polymorphism (i.e. STRs). Moreover, an additional feature of this method is the speed of the implemented algorithm and thus, is useful for larger datasets. MJ networks can be easily produced by using the Free Phylogenetic Network software (www.fluxus-engineering.com), which, besides generating MJ network trees from different types, computes age estimates for any putative ancestor in the tree.

2.3.4. Dating evolutionary events

The most important link between genetic diversity and human history is time and thus, dating past events such as species divergence or when a population split from another, has been an outstanding interest in evolutionary biology. Molecular dating approaches are based on the idea that all living things are descended from a single origin. In the case of species divergence, the amount of genetic change between two species then provides an estimate of when they diverged from this common ancestor. If this is true for all living things, then the alleles of a genomic segment can be also traced back to a single common ancestral copy of that genomic segment. Thus, the concept of the TMRCA can be easily extended to dating the origin of the variation within a genomic segment within a single species. All of these ideas arise from the coalescent theory, which not only provides a mathematical framework for constructing genealogies, but also aims to make inferences about population history.

In general terms, molecular dating consists in knowing the amount of variation between individuals or regions under study and the rate of genetic change over time. However, underlying these methods there is the assumption that the rate of genetic change has been constant over time. Therefore, if this rate has changed over time, the dates derived from molecular clocks will be meaningless. The relative rate test can be performed before using a clock approach in order to test whether the rate of evolution has been constant. If this is not the case, one alternative is to use a relaxed molecular clock

approach. Finally, another important point about molecular clocks is that they are always accompanied by rather large variances and confidence intervals, much larger than those associated with other dating methods, such as radiocarbon dating. The reason is that although mutations occur at a certain average rate, they are a stochastic process and thus, at a given moment, there can be more or fewer mutations than expected.

In the present work, we have focused on estimating the TMRCA of specific mutations, which indeed has the widest application in dating mtDNA and NRY haplogroups. For non-recombining haplotypes, such as the Y chromosome, mutations drive this diversification, so the haplotype itself represents a molecular clock. The basic idea behind dating specific mutations in these regions is that when a new mutation arises, it probably will be lost by drift, but alternatively it can rise in frequency and occur as a polymorphism in the population. At first, all the copies of the mutant chromosome will be identical, but over time new mutations will act diversifying those chromosomes (Figure 24). Therefore, the amount of genetic diversity associated with these chromosomes can be correlated with the age of the mutation. In the case of Y chromosome, STR loci have been widely used to estimate the age of specific NRY haplogroups, however, with the advent of massive parallel sequencing, methods focused on nucleotide substitutions for the NRY are providing valuable results.

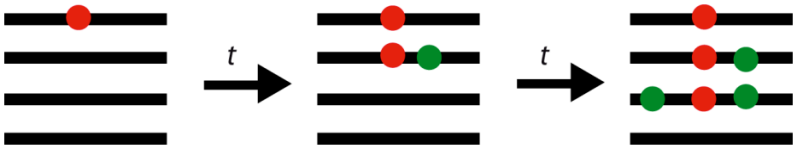


Figure 24. Diversity associated with a mutation increases over time. A new mutation (red circle) arises, by definition, on a single chromosome (black line), and then over time, new mutations (green circles) arise and become associated with that mutation. Adapted from Stoneking (3).

In the work presented here, we have estimated the TMRCA of sets of Y-STR haplotypes using the ρ (or rho) statistic. This statistic counts the average number of nucleotide changes between the root haplotype and every individual in the sample and is handily implemented in the program Network that, as mentioned above, is widely used to construct network trees. ρ is related to time by the equation:

$$\rho = \mu t$$

where μ is the mutation rate and t corresponds to the years per mutation. However, since this method is based on the average number of mutations that have occurred among all the sequences descended from a common ancestral sequence, it assumes a "starlike" phylogeny associated with a population expansion, which might not be always the case. In fact, detailed simulations showed that this method does not perform well under scenarios other than population expansion (167).

An alternative approach, which does not depend on the assumption of population expansion, is to use a Bayesian approach. The idea behind this approach is that data are simulated using several different models and then the model that best fits the real data is

chosen. This approach also requires prior estimates of the parameters, usually based upon existing diversity data. One should take special care when selecting the prior distribution of such parameters to not compromise the likelihood of a particular model fitting the data.

3. OBJECTIVES

The main goal of this thesis project is to give a broad view of the diversity of the human Y chromosome. In the first part of the project, I have used the Y chromosome as a tool to study a relatively recent event in human history: surname inheritance. More than 2,500 samples from volunteers bearing one of 50 Catalan surnames have been genotyped for 17 Y-chromosome STRs and 68 SNPs, with the following objectives:

a) Discover and quantify the processes that drive surname frequency.
b) Were the founders of surnames that are linguistically Arab or Hebrew, North Africans or Jews themselves?
c) Were the founders of Germanic patronymic surnames of a different genetic origin form the rest of the population?
d) Does an ethnonymic surname indicate a foreign origin? Some Catalan surnames (*Alemany*, *Danés*, *Guasch*) denote geographic origin (they mean German, Dane, Gascon, respectively).

In the second part of the project, I have analysed whole Y chromosome sequences from North Africa males bearing the most common paternal lineage in the area: E-M81, with the following objectives:

a) Increase the knowledge on the North African paternal lineage by discovering new variants within this branch.

b) Genotype those variants to refine the phylogeography of the EM81 branch and thus discover whether North African populations present any substructure.
c) Provide updated time estimates from both SNP and STR data.
d) Shed some light on the geographical origin of this paternal lineage in order to better understand the demographic history of North African populations.

4. RESULTS

4.1. Recent Radiation of R-M269 and High Y-STR Haplotype Resemblance Confirmed

Solé-Morata N, Bertranpetit J, Comas D, Calafell F. Recent Radiation of R-M269 and High Y-STR Haplotype Resemblance Confirmed. Ann Hum Genet. 2014 Jul;78(4):253–4. DOI: 10.1111/ahg.12066

4.2. Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency

Solé-Morata N, Bertranpetit J, Comas D, Calafell F. Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency. Eur J Hum Genet. 2015 Oct 18;23(11):1549–57. DOI: 10.1038/ejhg.2015.14

4.3. Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81)

Solé-Morata N, García-Fernández C, Urasin V, Bekada A, Fadhlaoui-Zid K, Zalloua P, et al. Whole Y-chromosome sequences reveal an extremely recent origin of the most common North African paternal lineage E-M183 (M81). Sci Rep. 2017 Nov 21;7(1):15941. DOI: 10.1038/s41598-017-16271-y

5. DISCUSSION

In the following sections, I will first comment on using the Y chromosome as an evolutionary marker (section 5.1). Then, I will discuss the relevance of the results presented herein for our understanding of the surname system in Catalonia (section 5.2) and North African population history (section 5.3).

5.1. The Y chromosome as an evolutionary marker

The Y chromosome can be viewed as the longest piece of non-recombining DNA in the human genome. Both this avoidance of recombination and its paternal inheritance make the Y chromosome a powerful tool with which to study human history. However, this male specificity means that the patterns of diversity of the Y chromosome reflect the characteristics of male behaviours. Thus, we should be aware that the inferences obtained from the Y chromosome provide only a paternal perspective of demographic processes, and hence the diversity of the entire human genome should be accounted to completely characterize human evolution (10,168). Nevertheless, since the discovery of the first polymorphisms, many studies have used the variation encapsulated in this paternal marker with different purposes, such as confirming genealogical relationships or tracing human migrations.

SNPs or STRs?

Although other types of mutations exist, single nucleotide polymorphisms and short tandem repeats are those commonly used in evolutionary studies. Since both markers differ in their mutation rates, they can be used to obtain demographic insights at different time-scales. Moreover, until recently, binary polymorphisms of the Y chromosome have suffered from ascertainment bias, which has been somehow alleviated by combining both markers (168). In the present work, we combine information both from slowly mutating SNPs and rapidly-mutating STRs. In section 4.2, SNPs, whose mutation rate is low enough not to be represented in the timescale of surname inheritance systematization, are used to assign a haplogroup to each individual. By contrast, Y-STRs, which are less affected by ascertainment bias and present a higher mutation rate, are used in order to define more informative haplotypes within the haplogroups. Moreover, on the basis of STRs we can establish closely related sets of haplotypes, known as descent clusters, and estimate their TMRCAs (section 5.2). So, depending on the information needed, we will rely on one of these markers.

Furthermore, despite haplogroups being defined by SNPs, the compartmentalization of STR haplotype by haplogroup (25) allowed Schlecht and colleagues (26) to propose an alternative method for assigning a sampled Y chromosome to a haplogroup by using STR data. The power of haplogroup prediction mostly depends on the number of STRs typed, and it might be the case that closely related haplogroups share identical Y-STR haplotypes, even

when many STRs are typed. In general, older haplogroups will harbour higher STR haplotype diversity than younger ones. Indeed, in the work presented in section 4.1, we reported an extreme homogeneity of Y-STR haplotypes across R-M269 subhaplogroups. This pattern, previously observed by Larmuseau *et al.* (77) in Flanders (Belgium) and part of the Netherlands, has been attributed to the recent, rapid radiation of this haplogroup. Nonetheless, despite this extreme haplotypic resemblance across R-M269 subhaplogroups, some exceptions were found in its R-U198 (77) and R-Z220 (78) subbranches; in which their Y-STR haplotypes carry relevant phylogenetic information and thus, could be used for haplogroup prediction.

Sequencing or genotyping?

As mentioned in section 2.1.2, in some cases, we might be interested in the genotypes at particular positions in a DNA sequence, which is mostly because we have prior knowledge on these positions, but also because it is easier and cheaper to focus only on them. In section 4.2, by targeting specific SNPs and STRs we have been able to learn about the relationships between male bearing the same surname. However, the best way to identify variation on the Y chromosome is to sequence it. The work presented in section 4.3 combines both strategies. First, by using sequence data, we have been able to discover new variants within a set of Y chromosomes belonging to particular lineage, and therefore it has allowed us to define new subhaplogroups within it. Then, by targeting 17 Y-STRs we have been able to study the level of

variation within these sublineages and to establish the time since the haplogroup-defining mutation occurred.

Besides the advantages of using MPS in variant discovery, there are also several limitations. Y chromosome sequence data can be used to build phylogenies free of ascertainment bias, with branch lengths proportional to number of mutations; however, sequencing this region has its own difficulties. First, both the repeated nature of the MSY sequence and short-read data obtained from most MPS platforms, complicate its alignment to the reference genome, which so far is only possible in the unique regions of the chromosome (168). Future studies should investigate the possibility of extracting reliable variant calls from additional regions of the chromosome, which would be facilitated by the longer reads expected as sequencing technologies improve (12). In the present project, we have sequenced the entire genome of some individuals and then, extracted bioinformatically those reads considered relevant. Alternatively, other studies have sequenced specifically those regions considered unique (39,169). This strategy, known as sequence capture, has the advantage of high coverage and good singleton representation (39), which is another of the limitations of massive parallel sequencing. Indeed, low coverage data, like that of the Sardinian population (38), yields an overrepresentation of singleton variants, so terminal branch lengths may be artificially short and thus, time depth will be overestimated.

Other technical factors, such as the huge diversity of sequencing platforms, the variant calling algorithm or the filtering strategy that

affect the phylogeny obtained from sequence data, suggest the necessity of defining a unified pipeline to deal with Y chromosome data. Finally, until recently, sample size has also been an issue; Y-trees provided by different studies vary in the number of population sampled, in sample size and in the number of Y-lineages represented. Nonetheless, this has been recently alleviated by a recent work from Poznik *et al*. (14), that has provided a phylogeny based on 1,244 Y chromosome sequences from worldwide populations.

The advent of massive parallel sequencing technologies has enabled the discovery of thousands of new SNPs, but also can be used to find new Y-STRs. Willems *et al.* (18) exploited whole-genome sequencing data to estimate the mutation rates of Y-chromosome STRs with 2–6 bp repeat units that are accessible to Illumina sequencing. They also identified Y-STRs with potential applications in forensics and genetic genealogy and assessed the ability to differentiate between the Y chromosomes of father-son pairs, which demonstrates the great potential of large-scale Y-STR studies. Nevertheless, we should be aware that short-read sequencing leads to an under-ascertainment of the longest STRs.

Which rate to use?

The Y chromosome is considered the best molecular clock in the human genome (11). However, in order to calibrate this clock, we must know its mutation rate, which will be then used to estimate the TMRCA of a group of related Y chromosomes. Mutation rate

estimates have been obtained from both STRs and SNPs; however, while estimations based on STRs markers have been used for two decades, estimates from SNPs only became possible with large-scale sequencing techniques. Given that STRs mutate much faster, they will remain as important markers for constructing fine-scale phylogenies (11,18). However, they are affected by uncertainty over the appropriate choice of mutation rate (170), and can be affected by mutation saturation over long time-scales (171,172). In section 4.2, ages for the founders of surnames have been estimated by using STR data, because their mutation rate is high enough to be represented in the timescale of surname inheritance systematization. By contrast, in section 4.3, we were interested in an older event in human history, which is the origin of a certain haplogroup. Therefore, we have also employed binary polymorphisms obtained from sequence data to estimate the TMRCA of our set of Y chromosomes.

Regardless of the type of variation used, we should discuss which approach might be used for estimating mutation rates. In section 1.1.1, we have described that mutation rates can be classified into two main categories: evolutionary or genealogical rates. Genealogical rates are directly counted from pedigree data in many cross-consistent studies, and its precision constantly increases. Nevertheless, a problem with Y-STR genealogical rates is that they differ by two to three orders of magnitude across individual STRs (18), a situation that gets even worse with evolutionary rates. For the present work, we relied on genealogical rates, specifically, on those provided by YHRD, which regularly updates a collection of

mutation rates for forensic Y-STRs (11). As in the case of STRs, genealogical Y-SNP rates are obtained by pedigree based data, whereas evolutionary rates can be obtained either by using archaeological dates or by using of ancient DNA sequences of known ages. In section 4.3, we used both approaches; the genealogical mutation provided by Helgason *et al.* (20), and an evolutionary rate obtained using ancient DNA data (21), to obtain direct estimates of the TMRCA of E-M183. And, despite all this approaches give reasonably consistent estimates, there is a still a slightly difference between both mutation rates, indicating the remaining uncertainty. Indeed, as we will expose in section 5.3, our estimates using an evolutionary rate are slightly older than those provided by a genealogical mutation rate, which in turn, are consistent with ages obtained by a genealogical STRs mutation rate. When choosing which rate to use, we should consider that the genealogical rate might be more relevant for recent expansions (14). However, regardless of which rate is chosen, we must always incorporate its confidence intervals, in order to increase the chance that the true TMRCA lies within the confidence interval provided (11).

Pitfalls of the applications of the Y chromosome

As described in section 1.1.2, the variation encapsulated in the Y chromosome has provided valuable insights into different areas, such as population genetics, genealogy, forensic and medical studies. Nevertheless, we must be aware of its limitations to correctly interpret our results.

Although the Y chromosome has been widely used in population genomics, it can be strongly influenced by genetic drift and by sex-biased behaviours, such as brides moving to the grooms' birthplaces after marriage. Thus, this region should be used to understand past social structure and potentially different behaviours of men and women, but past demographic inferences should be made with caution. For example, studies of many populations in the Americas have shown the dramatically male-biased contribution of Europeans compared with indigenous or African-derived populations (173–175). Moreover, as we mentioned in section 1.3.2, while Y-chromosome variation is geographically distributed in North Africa (142), genome-wide studies have shown little or no genetic structure (132,176).

Another important area in which the male-line inheritance of the Y chromosome can be employed is genetic genealogy. The study of family history is an enormously popular hobby that has been enhanced by the availability of commercial DNA testing kits. Therefore, amateur people who are not academics trained in population genetics can make contributions to the scientific literature (177). Nevertheless, these contributions should be interpreted carefully, because dealing with someone's own family history, might imply that results could be interpreted emotionally. An important risk of genetic genealogy is the detection of unexpected past non-paternities, as happened in the Thomas Jefferson's case described in section 1.1.2 (50–52). Finally, a collateral risk of "recreational genetics" is the discovery of disease

associations; for instance, the absence of specific Y-STRs and SNPs is associated with male infertility (178).

Finally, the Y chromosome has been used in forensic cases for male sex identification, male lineage identification and identification of the geographical origin of male lineages. The use of Y-STRs to characterize paternal lineages of unknown male donors has been especially useful in sexual assault cases, when males and females have contributed to the same trace. Nevertheless, the DNA left in a crime scene is often degraded and so, methods for individual identification should be very sensitive. In addition, is difficult to establish a uniform forensic database, because methods for identifying individual variation change quickly. Another problem with large genetic profile databases is the risk of infringing someone's privacy. Finally, as mentioned in section 1.1.2, genotyping errors can lead to false exclusions and thus, a guilty person considered innocent; or even worse, due to false inclusions, an innocent person could be imprisoned (1,179). So, all of these factors should be taken into account when using the Y chromosome with forensic purposes.

5.2. Y-chromosome diversity and Catalan surnames

The biological behaviour of the Y chromosome, which is paternally inherited, implies that males sharing the same surname may also share a similar Y chromosome. Until recently, surname systems have only been analysed in Britain and Ireland, with different results. Our aim in section 4.2, was to examine Y-chromosome haplotypes and haplogroups of men bearing one of the 50 Catalan surnames selected, in order to provide some insights into the driving forces behind the origin, systematization, and diffusion of surnames.

The main result of this work is that Y-chromosome diversity increases with surname frequency. By defining groups of men likely to share a common ancestor, known as descent clusters, we showed that the number of these clusters increases with surname frequency, suggesting that frequent surnames became abundant because they were founded multiple times, while rarer surnames tended to have fewer origins. This hypothesis was also supported by an MDS plot of genetic distances; while frequent surnames were closer to a general population sample, those less represented in the sample tended to be more peripheral. As mentioned in section 1.2.2, a previous study carried out by King and Jobling (71) have already proposed that surname frequency was mainly driven by polyphyletism. In addition, results obtained by a study of Y-chromosome variation in Spain (180), are also in agreement with

this hypothesis. Therefore, the higher levels of coancestry showed by frequent Irish surnames, may be an exception in Western Europe that could be explained in terms of social selection. Nevertheless, the geographic discontinuity of surname studies strongly evidence that sampling of a wider variety of populations is needed to solve the current geographical bias and to provide new links between surnames and demographic history.

Descent cluster ages estimated for British and Irish surnames were about 650 and 1,100 years old, which are consistent with surname establishment in those places, around the twelfth and thirteenth centuries in Britain, and in early tenth century in Ireland (71,83). In Spain, ages of descent clusters ranged from 167 to 1,310 years, but only a small percentage of those (9.5%) were estimated to be older than 800 years, the earliest possible limit for hereditary surname establishment in Spain (180). Noticeably, our descent cluster ages estimates are on average slightly more recent than does obtained in the other studies (around 500 years old). A possible factor that may have biased downward the age estimate for Catalan surnames is that, despite surname inheritance may have already been present in the Late Middle Ages, the systematic transmission of the paternal surname was not established until the sixteenth century with the mandate of the Council of Trent, to keep record of births, marriages and deaths. However, other Catholic countries should be tested in order to test this hypothesis.

Since the main conclusions of this work were built on the basis of descent clusters, its definition should be reviewed carefully.

Descent clusters have been defined following a modified version of the heuristic in Martinez-González *et al.* (74), which provides a less stringent definition than the rules defined by King and Jobling (71). The importance of the criteria used to define such descent clusters lies on its effect on age estimation. Since our criteria is less restrictive than that adopted by King and Jobling (71), the inclusion in descent clusters of some individuals that should be excluded with a more stringent definition, will yield older estimates. Although the consistency of our definition was indeed verified, taken together, these observations strongly suggest that further studies should focus on a single definition, in order to allow the direct comparison of age estimates between different studies. It is also worth to mention that, despite the reliability of descent clusters to infer the foundation of a certain surname, some of these clusters might not represent the initial founder, but rather belong to later introgressions. To differentiate between both scenarios, we considered that those descent clusters comprising more individuals were more likely to represent initial founders, and we defined major descent clusters (MDCs) as those with $n \geq 4$. It should be noticed that, despite it seems quite arbitrary, the comparison with other thresholds strongly support our definition.

Age estimates do not only depend on the definition of these descent clusters, they are also influenced by the mutation rate employed. We obtained our mutation rate from the compilation in the YHRD database (www.yhrd.org, accessed on Feb. 5th, 2014). Given the mutation rate of the set of 17 Y-STRs used and considering a generation time for men of 35 years, we obtained a mutation rate of

one mutation per haplotype per 777 years. In our case, the male generation time was estimated from the average age difference between bridegrooms and their fathers in a database of 550,000 marriages help in Catalonia between 1451 and 1905; Dr. Anna Cabré (http://dag.cvc.uab.es/projects/five-centuries-of-marriages), personal communication. This means that two men sharing a common ancestor 777 years ago (roughly at the age when surnames were established) would be separated, on average, by two mutations. This mutation rate was faster than those adopted by King and Jobling (1 mutation per 1,373 years), which was based on their own genotyping of pairs of related individuals totalling 274 transmissions of the Y chromosome. Obviously, different mutation rates will yield different results, and thus, comparisons between studies should be taken carefully.

One of the most important caveats of surname studies is that, as time goes, the link between Y chromosome and patrilineal surnames is more likely to be affected by non-paternity, adoption, multiple founding events for names and matrilineal transmission of the surname. As mentioned in section 1.2.2, these processes will yield the appearance of new chromosomes linked to a certain surname. The rate at which Y chromosome and surname are not inherited together estimated in the present work was 1.5–2.6% per generation, which is consistent with those found in Britain by King and Jobling (71), but slightly higher than those estimated in Flanders (54) and Northern Italy (181), respectively, at 0.9 and 1.2% per generation. These differences could be explained by differences on the non-paternity rate between populations, but also

they might depend on how usually the maternal surname is transmitted in a certain area. Remarkably, at the time of writing, in Spain both parents have to decide which surname is transmitted to their offspring. So, future analysis performed in this area should be aware that, for males born after June 30, 2017, it cannot be automatically assumed that he has been given his father's surname. Maternal surname transmissions did exist before that date, but they were regarded as exceptional. In conclusion, information from other geographical areas will help to obtain a broad view of how often the Y chromosome and surname are not inherited together.

An important breakthrough of this project was to address questions related to the etymology of surnames. To do so, we focused on groups of surnames in which their etymology suggested a particular origin for the founder of the surname and verified whether that was the case. The most encouraging results were obtained in surnames with a linguistic Arabic or Hebrew background. For example, two linguistically Arabic surnames, Nàcher and Massot, showed a statistically significant excesses of haplogroups common in North Africa (such as E-M81 or J1-M267), which might suggest a North African origin of their founders. In the case of linguistically Hebrew surnames, only the surname Estruch exhibited an overrepresentation of Jewish haplogroups. One caveat of this goal is that some surnames might have alternative etymologies and thus, we might be testing the wrong hypothesis. Moreover, in the case of surnames with a putative Jewish origin, it could be explained by the conversion to Christianity of Jewish people, which may have implied a change in their surnames. In conclusion, despite it was

worth testing hypothesis related with etymology, we were a priori quite aware that the outcome of this analysis might not be very satisfying.

Probably, the most controversial analysis of this work is regarding surname prediction. As proposed by Jobling et al. (69) the link between surname and Y-chromosomal haplotype suggests the idea of predicting a surname in forensic investigations. Therefore, in the absence of an autosomal DNA profile or when it yields no matches in a database, a list of Y-STR profiles and its associated surnames could be used to prioritize a suspect list. The idea behind surname prediction is that, when a Y-chromosome haplotype from a forensic sample belongs to a given MDC, we attribute that surname to the sample. Although the validity of this approach has also been confirmed by King *et al.* (75), its practical application is more controversial because, as we discussed above, MDC coverage correlates negatively with surname frequency. So, common surnames should not be considered in surname predictions.

In addition, in some cases, common haplotypes being shared across surnames could result in many surnames being returned. So, we have also estimated how often a sample that did not belong to a MDC in their own surname, which is known as the false discovery rate (FDR). This false discovery rate, which does not correlate with surname frequency, may be due to NPT factors, such as non-paternity events or the matrilineal transmission of the surname. Alternatively, it might be that the combination of 17 STRs and 68 SNPs does not have enough resolution to differentiate between such

events unrelated Y chromosomes that may have similar haplotypes, especially in the case of Y chromosomes belonging to the recent and abundant R1b-M269 haplogroup (section 4.1). Finally, though surname prediction could be useful in forensics, we should be aware of the risk of breaking the privacy of those contributing DNA anonymously for medical research. As we have already mentioned in section 1.2.3, a 15-year old boy conceived by anonymous sperm donation was able to trace back his biological father by surname prediction.

Finally, although we started the discussion of this work remarking the necessity of incorporating a high number of samples and surnames, we should notice that most important advances in the field will arise from exploitation of recent technological developments. Massive parallel sequencing techniques will enable the incorporation of whole genome data, allowing the reconstruction of genealogies that incorporate links across the sexes. Moreover, recent improvements on the ancient DNA field should provide a means of confirming genealogical links between living individuals and putative patrilineal ancestors and also among archaeological human remains (182).

5.3. A Y-chromosomal perspective of North African populations

The demographic history of North African populations is extremely complex. Past and recent historical migrations gave rise to a complex genetic landscape that has been strongly influenced by geography. Several studies have tried to assess whether early populations were replaced by later migrations or if there has been a continuous settlement of the region. Our goal in section 4.3 was to shed some light on North African population history with the aid of whole Y-chromosome sequences from individuals bearing the most frequent North African haplogroup: E-M183 (E-M81). While most of the studies have explored the paternal structure of North Africa focused on targeted SNPs and STR of the Y chromosome (123,142,143), here, by using whole Y-chromosome sequences, we have been able to discover new variation and thus, to increase the knowledge of internal new branches within E-M183. As a result, we have provided a refinement of the phylogeography of this lineage.

Despite massive parallel sequencing data enables the discovery of new variation, we should notice that, given the costs of MPS versus targeting specific variants, the sample size may be reduced and thus, population inferences have to be made more carefully. An example that clearly illustrates the effects of having a reduced sample size is the frequency of M183* in the Iberian sample. Samples from the Iberian Peninsula showed the lowest frequency of the different subhaplogroups discovered in this study and the highest frequency

of M183*. Given the low frequency values of this haplogroup in the area, this could be attributed to genetic drift acting on a low-frequency variant. Alternatively, the lack of sequence data from Iberian individuals may have prevented the discovery of specific variants of this area. Therefore, future analyses should incorporate samples from the Iberian Peninsula, as well as from Near Eastern populations, in order to build an even more reliable phylogeny of E-M183.

As we mentioned in section 1.1.1, given that each Y-STR mutation occurs in a chromosome that belongs to a particular Y-SNP haplogroup, Y-STR allele variation is deeply partitioned by haplogroup. As a result, Y-STRs haplotypes define more informative haplotypes within the haplogroups (24,25). In order to provide a better analysis of the new subhaplogroups discovered by MPS we genotyped the 17 Y-STR contained in the Yfiler^{TM} kit (Applied Biosystems) (section 2.1.2) in more than 200 North African males, and showed that Y-STR haplotypes within E-M183 individuals are strikingly similar to each other. Therefore, subhaplogroups within E-M183 cannot be distinguished from each other based on Y-STR differences. The scenario that better explains such haplotype resemblance within a particular haplogroup is a recent and rapid radiation of subhaplogroups (77). Nevertheless, this lack of differentiation between Y-STR haplotypes within specific subhaplogroups could also be attributed to a lack of resolution; that is, using only 17 microsatellites might not be enough to differentiate Y-STR haplotypes between haplogroups.

This lack of resolution could be easily solved by genotyping more STRs (e.g. Family Tree DNA genotypes more than a hundred STRs for its customers). Furthermore, massive parallel sequencing techniques are enabling the identification of microsatellites from sequence data, which opens the possibility of using such markers in population data studies increasing the level of resolution.

Besides this lack of structure between subhaplogroups and Y-STR haplotypes, we have also observed a lack of geographic structure, meaning that different North African populations do not show different subhaplogroup composition. This contrasts with the strong geographical structure of the Y chromosome reported by Arredi *et al*. (142), but is in agreement with genome-wide studies, which showed a lack of correlation between geographical and genetic populations (132). This dilution of the strong geographical structure of the Y chromosome when we go deeply into a particular Y-chromosomal branch could be attributed to the rapid radiation mentioned above, but could also be due to the sampling strategy. Therefore, further studies should consider not only the inclusion of more samples from of Northwestern Africa, but also from populations surrounding Egypt and from the Near East. Finally, despite this lack of structure within E-M183, more inland territories (Western Sahara, Algerian Reguibates), showed less diversity than those placed in the coast. This pattern has been observed also with genome wide data (132) and could be related to migrations along the coast, but again, a better sampling will be needed to test this hypothesis.

As we mentioned in section 1.3.2 the dating of this North African paternal lineage has been controversial. While Bosch *et al*. (123) proposed a Paleolithic origin, later studies have pointed to more recent origins, such as the Neolithic (142) or even more recent (143). Our results suggested that the origin of E-M183 is much more recent than was previously thought. In addition, whereas other studies have relied only on STR data to provide time estimates, here, for the first time, we have used Y-chromosomal sequence data to calculate the TMRCA for E-M183. In general, time estimates obtained from both STR or SNP are slightly similar, however, we do observe some differences between those estimates obtained depending on the mutation rate used. As we described in section 1.1.1, genealogical rates tend to overestimate the rate of mutation and thus, will provide more recent estimates than those obtained using calibration points.

Here, the TMRCA obtained using the mutation rate provided by Fu *et al*. (21), was slightly older than that obtained by the genealogical rate provided by Helgason *et al*. (20). It is not easy to decide which mutation rate should be used and probably the best option is to combine both approaches. Here, by using both approaches we showed that, despite the differences, the confidence intervals of the TMRCA obtained by both strategies overlapped. Finally, besides the strategy followed to obtain the TMRCA, it is also important to correctly interpret the age estimates of NRY haplogroups. This date does not represent the age of the migration that spread the chromosomes carrying those haplotypes, the presence of a

haplogroup of a particular age in a population provides only an upper bound as to when that haplogroup entered in that population (1).

While TMRCA estimates of a certain haplogroup and its subbranches provide some constraints on the times of their origin and spread, its current distribution and diversity can suggest their geographical origin. By analysing the patterns of genetic diversity within EM183, Arredi *et al*. (142) suggested that an expansion from the Near East could explain the observed east-west cline of genetic variation that extends into the Near East. Indeed, our results also showed a reduction in STR heterozygosity towards the West, which may be taken to support the hypothesis of an expansion from the Near East. Noticeably, a Near Esternern origin has been also supported by studies based on genome-wide SNPs (131,132), which showed a North African autochthonous component that increase towards the West. Nonetheless, our correlations should be taken carefully because our analysis includes only six locations on the longitudinal axis, none from the Near East. Therefore, future analyses should include more samples in order increase our statistical power to confirm a Near Eastern origin, but also to investigate if there is a real west-to-east cline of genetic diversity, or this correlation analysis only evidences the low genetic diversity in Western Sahara.

According to our time estimates, if E-M183 was originated in the Near East, it could have been brought by the Islamic expansion on the 7th century. This hypothesis would also fit with the patterns

observed in the rest of the genome, where an extensive male-biased Near Eastern admixture event is registered ~1,300 ya, coincidental with the Arab expansion (132). However, a Near Eastern origin is not the only hypothesis that could be envisaged. Indeed, the high frequency of E-M183 in the Maghreb would fit the clear pattern of longitudinal isolation by distance reported in genome-wide studies (131,132). If E-M183 was originated somewhere in Northwest Africa and then spread through the entire region, it could be related with the end of the third Punic War (146 BCE), when Carthage (in current Tunisia) was defeated and destroyed. Indeed, the fact that about 2,000 ya North Africa was one of the wealthiest Roman provinces, may explain the rapid expansion of E-M183, which gave rise to observed patterns of variation.

Finally, although using current populations to study past events has been a very popular strategy, recent technological advances in the field of ancient DNA have exhibit how important is to include ancient data to study population history. Therefore, despite an expansion from the Near East seems more plausible given the results obtained in this project and using genome-wide data, if available, ancient DNA data should be incorporated in the analysis to provide a better understanding of the demographic history of North African populations and to shed some light into the origin of this paternal lineage.

5.4. Concluding remarks and future directions

The main goal of this thesis was to study recent demographic and evolutionary processes from a Y-chromosomal perspective. By using high-throughput techniques, I have determined sets of descendants of a common ancestor of men carrying the same surname, as well as the population of origin of the common ancestor and the date when such a common ancestor lived. Moreover, this study strongly supports the hypothesis that surname frequency has been driven by polyphyletism. Despite other studies have attempted to understand the processes driving surname frequency, most of them have focused on surnames on the British Island (71,73,75,76), and thus, this work has contributed to understand the relationship between surnames and Y chromosome.

Another key contribution of this thesis has been eased by the latest advances in MPS techniques. By sequencing the Y chromosome of 32 North African individuals belonging to the most common haplogroup in the area (E-M183), we have discovered new variation and thus, improved the knowledge of this important but ignored Y-chromosomal branch. Moreover, we learned that this haplogroup was more recent than was previously thought, which suggest that previous hypothesis should be reviewed and that, this E-M183, thought to be autochthonous of North Africa, appeared in the Near East and then was brought to NW Africa by the Islamic expansion of the 7th century. Additionally, DNA sequences produced in this

work will be available for other researchers, and will surely contribute to future investigations addressing population genetics, selection and other topics.

Thanks to Y-chromosomal data accumulated in the last decades, our knowledge of human population history is more complete than ever before. Nevertheless, given the huge complexity of the Y-chromosomal sequence, more efforts have to be done to better understand its structure. In addition, many interesting questions are still waiting to be addressed with using this paternal piece of DNA. Although the information obtained from the Y chromosome is sex-biased, comparing the patterns of diversity between this marker and the rest of the genome can help to better understand sex-related behaviours. Moreover, sampling currently underrepresented areas will hence provide us with a more accurate vision of such processes.

Finally, improvements in MPS techniques and bioinformatic processing in Y-chromosome research, will improve the quality of the data and the amount of information obtained from this piece of DNA. Indeed, despite its limitations, the Y chromosome will continue to be a great tool with which to study human history.

6. CONTRIBUTIONS TO OTHER PUBLICATIONS

6.1. Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI

Charlier P, Olalde I, Solé N, Ramírez O, Babelon J-P, Galland B, et al. Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI (both Kings of France). Forensic Sci Int. 2013 Mar 10;226(1–3):38–40. DOI: 10.1016/ j.forsciint.2012.11.018

6.2. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations

Casals F, Anglada R, Bonet N, Rasal R, van der Gaag KJ, Hoogenboom J, et al. Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations. Forensic Sci Int Genet. 2017 Sep;30:66–70. DOI: 10.1016/j.fsigen.2017.06.006

6.3. Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ

Solé-Morata N, Villaescusa P, García-Fernández C, Font-Porterias N, Illescas MJ, Valverde L, et al. Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ. Sci Rep. 2017 Aug 4;7(1):7341. DOI: 10.1038/s41598-017-07710-x

7. REFERENCES

1. Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. Human Evolutionary Genetics. 2nd ed. Human Evolutionary Genetics, 2nd edition. New York and London: Garland Science; 2014. 670 p.

2. Willard HF. Tales of the Y chromosome. Nature. 2003;423(6942):810–3.

3. Stoneking M. An Introduction to Molecular Anthropology. New Jersey: Wiley-Blackwell; 2016. 400 p.

4. Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the y chromosome. Funct Integr Genomics. 2013;13(3):285–93.

5. Helena Mangs A, Morris BJ. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Curr Genomics. 2007;8(2):129–36.

6. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423(6942):825–37.

7. Page DC, Harper ME, Love J, Botstein D. Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. Nature. 1984 Sep 13;311(5982):119–23.

8. Mumm S, Molini B, Terrell J, Srivastava A, Schlessinger D. Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. Genome Res. 1997;7(4):307–14.

9. Schwartz A, Chan DC, Brown LG, Alagappan R, Pettay D,

Disteche C, et al. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet. England; 1998 Jan;7(1):1–11.

10. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003 Aug;4(8):598–612.

11. Balanovsky O. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. Hum Genet. Springer Berlin Heidelberg; 2017;

12. Wei W, Ayub Q, Chen Y. A calibrated human Y-chromosomal phylogeny based on resequencing Accepted Email alerting service A calibrated human Y-chromosomal phylogeny based on resequencing. 2013;388–95.

13. van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. Seeing the Wood for the Trees: A Minimal Reference Phylogeny for the Human Y Chromosome. Hum Mutat. 2013 Oct 26;1–5.

14. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016;12(9):809–809.

15. Calafell F, Larmuseau MHD. The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. Hum Genet. Springer Berlin Heidelberg; 2016;1–15.

16. Consortium TYC. A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. Genome Res. 2002;12(2):339–48.

17. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms

reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 2008;18(5):830–8.

18. Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am J Hum Genet. The American Society of Human Genetics; 2016;98(5):919–33.

19. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13(10):745–53.

20. Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, et al. The Y-chromosome point mutation rate in humans. Nat Genet. Nature Publishing Group; 2015;47(5):453–7.

21. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514(7523):445–9.

22. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435–45.

23. Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, et al. A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. Forensic Sci Int Genet. Elsevier Ireland Ltd; 2012;6(2):208–18.

24. King TE, Jobling MA. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet. 2009 Aug;25(8):351–60.

25. Bosch E, Calafell F, Santos FR, Pérez-Lezaun a, Comas D, Benchemsi N, et al. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am J Hum Genet. 1999;65:1623–38.

26. Schlecht J, Kaplan ME, Barnard K, Karafet T, Hammer MF,

Merchant NC. Machine-learning approaches for classifying haplogroup from Y chromosome STR data. PLoS Comput Biol. 2008;4(6).

27. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 2003;423(6942):873–6.

28. Muller HJ. The relation of recombination to mutational advance. Mutat Res. Netherlands; 1964 May;106:2–9.

29. Berta P, Hawkins JB, Sinclair AH, Taylor A, Griffiths BL, Goodfellow PN, et al. Genetic evidence equating SRY and the testis-determining factor. Nature. 1990 Nov 29;348(6300):448–50.

30. Sun C, Skaletsky H, Birren B, Devon K, Tang Z, Silber S, et al. An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. Nat Genet. 1999;23(4):429–32.

31. Jobling MA, Williams GA, Schiebel GA, Pandya GA, McElreavey GA, Salas GA, et al. A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol. 1998;8(25):1391–4.

32. Underhill P a, Kivisild T. Use of Y Chromosome and mitochondrial DNA Population Structure in Tracing Human Migrations. Annu Rev Genet. 2007;41:539–64.

33. Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, et al. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science. 1985;230(4732):1403–6.

34. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, et al. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000;26(3):358–61.

35. Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol. United States; 1998 Apr;15(4):427–41.

36. Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim a, et al. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet. 2000 Dec;67(6):1526–43.

37. Poznik GD, Henn BM, Yee M-CC, Sliwerska E, Euskirchen GM, Lin A a, et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 2013 Aug 2;341(6145):562–5.

38. Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, et al. Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome Phylogeny. Science. 2013 Aug 2;341(6145):565–9.

39. Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, et al. The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. Mol Biol Evol. 2015 Mar 1;32(3):661–73.

40. Scozzari R, Massaia A, Trombetta B, Bellusci G, Myres NM, Novelletto A, et al. An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. Genome Res. 2014 Mar;24(3):535–44.

41. Groucutt HS, Petraglia MD, Bailey G, Scerri EML, Parton A, Clark-Balzan L, et al. Rethinking the dispersal of Homo sapiens out of Africa. Evol Anthropol. 2015;24(4):149–64.

42. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. United States; 2015 Aug;349(6250):aab3884.

43. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S. Massive migration from the steppe is a source for Indo-European languages in Europe. :1–20.

44. Pajnič I, Pogorelc B, Balažic J. Molecular genetic identification of skeletal remains from the Second World War Konfin i mass grave in Slovenia. Int J Legal Med. 2010;124(4):307–17.

45. Baeta M, Nuñez C, Cardoso S, Palencia-Madrid L, Herrasti L, Etxeberria F, et al. Digging up the recent Spanish memory: Genetic identification of human remains from mass graves of the Spanish Civil War and posterior dictatorship. Forensic Sci Int Genet. Elsevier Ireland Ltd; 2015;19:272–9.

46. Haas C, Shved N, Rühli FJ, Papageorgopoulou C, Purps J, Geppert M, et al. Y-chromosomal analysis identifies the skeletal remains of Swiss national hero Jörg Jenatsch Jenatsch (1596-1639). Forensic Sci Int Genet. 2013;7(6):610–7.

47. King TE, Fortes GG, Balaresque P, Thomas MG, Balding D, Maisano Delser P, et al. Identification of the remains of King Richard III. Nat Commun. 2014;5(5631):1–8.

48. Charlier P, Olalde I, Solé N, Ramírez O, Babelon JP, Galland B, et al. Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI (both Kings of France). Forensic Sci Int. 2013;226(1–3):38–40.

49. Lalueza-Fox C, Gigli E, Bini C, Calafell F, Luiselli D, Pelotti S, et al. Genetic analysis of the presumptive blood from Louis XVI, king of France. Forensic Sci Int Genet. Elsevier Ireland Ltd; 2011;5(5):459–63.

50. Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, et al. Jefferson fathered slave's last child. Nature. 1998;396(6706):27–8.

51. King TE, Bowden GR, Balaresque PL, Adams SM, Shanks

ME, Jobling MA. Thomas Jefferson's Y Chromosome Belongs to a Rare European Lineage. Am J Phys Anthropol. 2007;1(May):80–3.

52. Gordon-Reed A. Thomas Jefferson and Sally Hemings: an American Controversy. Virginia: University Press of Virginia; 1997.

53. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The Genetic Legacy of the Mongols. Am J Hum Genet. 2003;72(3):717–21.

54. Larmuseau MHD, Vanoverbeke J, Van Geystelen A, Defraene G, Vanderheyden N, Matthys K, et al. Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. Proc R Soc. 2013 Dec;280(1772):20132400.

55. Colantonio SE, Lasker GW, Kaplan BA, Fuster V. Use of surname models in human population biology: a review of recent developments. Hum Biol. 2003;75(6):785–807.

56. Sibille I, Duverneuil C, Lorin de la Grandmaison G, Guerrouache K, Teissière F, Durigon M, et al. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Sci Int. 2002;125(2–3):212–6.

57. Rolf B, Keil W, Brinkmann B, Roewer L, Fimmers R. Paternity testing using Y-STR haplotypes: Assigning a probability for paternity in cases of mutations. Int J Legal Med. 2001;115(1):12–5.

58. Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, et al. A Comprehensive Survey of Human Y-Chromosomal Microsatellites. Am J Hum Genet. American Society of Human Genetics; 2004 Jun 7;74(6):1183–97.

59. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-chromosomal

microsatellites: Rates, characteristics, molecular bases, and rorensic implications. Am J Hum Genet. 2010;87(3):341–53.

60. Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso C a, Alvarez-Fernández F, et al. Mutation rates at Y chromosome specific microsatellites. Hum Mutat. 2005 Dec;26(6):520–8.

61. Moll F de B, FdB M, Moll F de B. Els llinatges catalans. Mallorca: Ed. Moll; 1982.

62. Albaiges JM. Enciclopedia de los nombres propios. Encicloped. Barcelona, Spain; 1995.

63. Faure R, Ribes MA, García A. Diccionario de apellidos españoles. Madrid, Spain: Espasa-Calpe; 2001.

64. Hanks P, Hodges F. A dictionary of Surnames. Oxford, UK: Oxford University Press; 1988.

65. Feder EK, Karkazis K. What's in a Name? 2008;9(October):33–6.

66. Rodriguez-Larralde a, Gonzales-Martin a, Scapoli C, Barrai I. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol. 2003 Jul;121(3):280–92.

67. Manni F, Toupance B, Sabbagh A, Heyer E. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. Am J Phys Anthropol. 2005 Feb;126(2):214–28.

68. Barrai I, Rodriguez-Larralde A, Manni F, Ruggiero V, Tartari D, Scapoli C. Isolation by language and distance in Belgium. Ann Hum Genet. 2004;68(1):1–16.

69. Jobling MA. In the name of the father: Surnames and genetics. Trends Genet. 2001;17(6):353–7.

70. McKinley R. A history of British surnames. London: Longman; 1990.

71. King TE, Jobling MA. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol. 2009 May;26(5):1093–102.

72. Redmonds G. Names and History: People, Places and Things. Hambledon and London: Continuum; 2004.

73. Sykes B, Irven C. Surnames and the Y chromosome. Am J Hum Genet. 2000 Apr;66(4):1417–9.

74. Martínez-González LJ, Martínez-Espín E, Álvarez JC, Albardaner F, Rickards O, Martínez-Labarga C, et al. Surname and Y chromosome in Southern Europe: a case study with Colom/Colombo. Eur J Hum Genet. 2012 Feb;20(2):211–6.

75. King TE, Ballereau SJ, Schurer KE, Jobling MA, Schürer KE, Jobling MA. Genetic signatures of coancestry within surnames. Curr Biol. 2006 Feb 21;16(4):384–8.

76. McEvoy B, Bradley DG. Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. Hum Genet. 2006 Mar;119(1–2):212–9.

77. Larmuseau MHD, Vanderheyden N, Van Geystelen A, van Oven M, de Knijff P, Decorte R. Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann Hum Genet. 2014 Mar;78(2):92–103.

78. Solé-Morata N, Bertranpetit J, Comas D, Calafell F. Recent Radiation of R-M269 and High Y-STR Haplotype Resemblance Confirmed. Ann Hum Genet. 2014 Jul 13;78(4):253–4.

79. Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, Herrera RJ. Y-chromosome diversity characterizes the Gulf of Oman. Eur J Hum Genet. 2008;16(3):374–86.

80. Rogers AR. Doubts about Isonymy Author. Hum Biol.

1991;63(5):663–8.

81. Hill EW, Jobling M a, Bradley DG. Y-chromosome variation and Irish origins. Nature. 2000;404(6776):351–2.

82. Moore LT, McEvoy B, Cape E, Simms K, Bradley DG. A Y-Chromosome Signature of Hegemony in Gaelic Ireland. Am J Hum Genet. 2006;78(2):334–8.

83. McEvoy B, Simms K, Bradley DG. Genetic investigation of the patrilineal kinship structure of early medieval Ireland. Am J Phys Anthropol. 2008;136(4):415–22.

84. King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, et al. Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. Eur J Hum Genet. 2007;15(3):288–93.

85. Bedoya G, Montoya P, García J, Soto I, Bourgeois S, Carvajal L, et al. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci U S A. 2006;103(19):7234–9.

86. Bowden GR, Balaresque P, King TE, Hansen Z, Lee AC, Pergl-Wilson G, et al. Excavating past population structures by surname-based sampling: The genetic legacy of the vikings in Northwest England. Mol Biol Evol. 2008;25(2):301–9.

87. Gitschier J. Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet. The American Society of Human Genetics; 2008;84(2):251–8.

88. Motluk A. Anonymous sperm donor traced on internet. New Sci Mag. 2005;(2524):6.

89. Trumme T, Herrmann B, Hummel S. Genetics in Genealogical Research--Reconstruction of a Family Tree by Means of Y-Haplotyping. Anthr Anz. 2004;62(4):379–86.

90. Kayser M, Vermeulen M, Knoblauch H, Schuster H, Krawczak M, Roewer L. Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. Forensic Sci Int Genet. 2007;1(2):125–8.

91. Walsh B. Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics. 2001 Jun;158(2):897–912.

92. Larmuseau MHD, Van Geystelen A, Van Oven M, Decorte R. Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics. Am J Phys Anthropol. 2013;150(4):505–11.

93. Dibble HL, Aldeias V, Jacobs Z, Olszewski DI, Rezek Z, Lin SC, et al. On the industrial attributions of the Aterian and Mousterian of the Maghreb. J Hum Evol. Elsevier Ltd; 2013;64(3):194–210.

94. Hublin J-J. Northwestern African Middle Pleistocene hominids and their bearing on the emergence of Homo sapiens. In: Human Roots: Africa and Asia in the Middle Pleistocene. Bristol: Western Academic and Specialist; 2001. p. 99–121.

95. Smith TM, Tafforeau P, Reid DJ, Grun R, Eggins S, Boutakiout M, et al. Earliest evidence of modern human life history in North African early Homo sapiens. Proc Natl Acad Sci. 2007;104(15):6128–33.

96. McDougall I, Brown FH, Fleagle JG. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature. 2005;433(7027):733–6.

97. Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature. Nature Publishing Group; 2017;546(7657):289–92.

98. Balter M. Was North Africa the launch pad for modern

human migrations? Science. 2011;331(6013):20–3.

99. Jacobs Z, Meyer MC, Roberts RG, Aldeias V, Dibble H, El Hajraoui MA. Single-grain OSL dating at La Grotte des Contrebandiers ('Smugglers' Cave'), Morocco: Improved age constraints for the Middle Paleolithic levels. J Archaeol Sci. Elsevier Ltd; 2011;38(12):3631–43.

100. Bouzouggar A, Barton N, Vanhaeren M, D'Errico F, Collcutt S, Higham T, et al. 82,000-year-old shell beads from North Africa and implications for the origins of modern human behavior. Proc Natl Acad Sci. 2007;104(24):9964–9.

101. d'Errico F, Vanhaeren M, Barton N, Bouzouggar A, Mienis H, Richter D, et al. Out of Africa: modern human origins special feature: additional evidence on the use of personal ornaments in the Middle Paleolithic of North Africa. Proc Natl Acad Sci U S A. 2009;106(38):16051–6.

102. Nespoulet R, El Hajraoui MA, Amani F, Ben Ncer A, Debénath A, Idrissi A, et al. Palaeolithic and neolithic occupations in the Témara region (Rabat, Morocco): Recent data on hominin contexts and behavior. African Archaeol Rev. 2008;25(1–2):21–39.

103. Garcea EAA, Giraudi C. Late Quaternary human settlement patterning in the Jebel Gharbi. J Hum Evol. 2006;51(4):411–21.

104. Jacobs Z, Roberts RG. Advances in optically stimulated luminescence dating of individual grains of quartz from archeological deposits. Evol Anthropol. 2007;16(6):210–23.

105. Pagonis V, Chen R, Kitis G. On the intrinsic accuracy and precision of luminescence dating techniques for fired ceramics. J Archaeol Sci. Elsevier Ltd; 2011;38(7):1591–602.

106. Van Peer P, Vermeersch PM. The Nubian complex and the dispersal of modern humans in North Africa. Recent Res into

Stone Age Northeast Africa. 2000;(May 2014):47–60.

107. Jacobs Z, Roberts RG, Nespoulet R, El Hajraoui MA, Debénath A. Single-grain OSL chronologies for Middle Palaeolithic deposits at El Mnasra and El Harhoura 2, Morocco: Implications for Late Pleistocene human-environment interactions along the Atlantic coast of northwest Africa. J Hum Evol. Elsevier Ltd; 2012;62(3):377–94.

108. Close A, Wendorf F. North Africa at 18,000 BP: Low latitudes. London: Unwin Hyman; 1990. 41-57 p.

109. Humphrey L, Bello SM, Turner E, Bouzouggar A, Barton N. Iberomaurusian funerary behaviour: Evidence from Grotte des Pigeons, Taforalt, Morocco. J Hum Evol. Elsevier Ltd; 2012;62(2):261–73.

110. Debenath A. Le peuplement prehistorique du Maroc : donnees recentes et problemes. Anthropologie. 2000;104:131–45.

111. Jackes M, Lubell D. Early and middle holocene environments and Capsian cultural change: Evidence from the Télidjène Basin Basin, Eastern Algeria. African Archaeol Rev. 2008;25(1–2):41–55.

112. Rahmani N. Technological and Cultural Change Among the Last Hunther-Gatherers of the Maghreb: The Caspian (10,000-6000 B.P.). J World Prehistory. 2004;18(1):57–105.

113. Camps G. Les Berbères: Mémoire et identité. Paris: Actes Sud; 1997.

114. Morales J, Pérez-Jordà G, Peña-Chocarro L, Zapata L, Ruíz-Alonso M, López-Sáez JA, et al. The origins of agriculture in North-West Africa: Macro-botanical remains from Epipalaeolithic and Early Neolithic levels of Ifri Oudadane (Morocco). J Archaeol Sci. Elsevier Ltd; 2013;40(6):2659–69.

115. Garcea E a. a. Semi-permanent foragers in semi-arid environments of North Africa. World Archaeol. 2006;38(911087503):197–219.

116. Haaland R. Africa and the Near East: pot and porridge, bread and oven - two food systems maintained over 10,000 years. 12th Congr Panafrican Archaeol Assoc Prehistory Relat Stud. Gaborone: University of Botswana; 2005;

117. Marston E. The Phoenicians. New York: Benchmark Books; 2002.

118. Harden D. The Phoenicians. London: Penguin Books; 1971.

119. Zalloua PA, Platt DE, El Sibai M, Khalife J, Makhoul N, Haber M, et al. Identifying Genetic Traces of Historical Expansions: Phoenician Footprints in the Mediterranean. Am J Hum Genet. 2008;83(5):633–42.

120. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill P a, Bertranpetit J. High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet. 2001 Apr;68(4):1019–29.

121. Hitti P. The Arabs: a short history. Washington, DC: Gateway Editions; 1990.

122. Newman J. The peopling of Africa. New Haven: Yale University Press; 1995.

123. Bosch E, Calafell F, Perez-lezaun A, Comas D, Mateu E, Bertranpetit J. Population History of North Africa: Evidence from Classical Genetic Markers. Hum Biol. 1997;3:295.

124. Cavalli-Sforza L, Menozzi P, Piazza A. The History and Geography of Human Genes. Princeton: Princeton University Press; 1994.

125. Mourant A, Kopec A, Domaniewska-Sobczak K. The distribution of the human blood groups and other polymorphisms. London: Oxford University Press; 1976.

126. Roychoudhury A, Nei M. Human polymorphic genes: world distribution. New York: Oxford University Press; 1988.

127. Tills D, Kopeć A, Tills R, Mourant A. The distribution of the human blood groups, and other polymorphisms. London: Oxford University Press; 1983.

128. Bosch E, Calafell F, Pérez-Lezaun A, Clarimón J, Comas D, Mateu E, et al. Genetic structure of north-west Africa revealed by STR analysis. Eur J Hum Genet. 2000;8(5):360–6.

129. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science. 2008;319(5866):1100–5.

130. Tishkoff SA, Reed FA, Friedlaender FR, Ranciaro A, Froment A, Hirbo JB, et al. The Genetic Structure and History of Africans and African Americans. 2009;324(5930):1035–44.

131. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 2012 Jan;8(1):e1002397.

132. Arauna LR, Mendoza-Revilla J, Mas-Sandoval A, Izaabel H, Bekada A, Benhamamouch S, et al. Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. Mol Biol Evol. 2017;34(2):318–29.

133. Fadhlaoui-Zid K, Rodríguez-Botigu´r L, Naoui N, Benammar-Elgaaied A, Calafell F, Comas D. Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. Am J Phys Anthropol. 2011;145(1):107–

17.

134. González AM, Larruga JM, Abu-Amero KK, Shi Y, Pestano J, Cabrera VM. Mitochondrial lineage M1 traces an early human backflow to Africa. BMC Genomics. 2007;8:223.

135. Hervella M, Svensson EM, Alberdi A, Günther T, Izagirre N, Munters AR, et al. The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. Sci Rep. 2016;6(14):25501.

136. Burton ML, Moore C., Whiting JW, Romney AK. Regions based on social structure. Curr Anthropol. 1996;37(1):87–123.

137. Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, et al. Tracing past human male movements in northern/eastern Africa and western Eurasia: New clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol Biol Evol. 2007;24(6):1300–11.

138. Pereira L, Cerný V, Cerezo M, Silva NM, Hájek M, Vasíková A, 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. 2010;18(8):915–23.

139. Zalloua PA, Xue Y, Khalife J, Makhoul N, Debiane L, Platt DE, et al. Y-Chromosomal Diversity in Lebanon Is Structured by Recent Historical Events. Am J Hum Genet. 2008;82(4):873–82.

140. Haber M, Platt DE, Badro DA, Xue Y, El-Sibai M, Bonab MA, et al. Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. Eur J Hum Genet. Nature Publishing Group; 2011;19:334–40.

141. Fadhlaoui-Zid K, Martinez-Cruz B, Khodjet-el-khil H, Mendizabal I, Benammar-Elgaaied A, Comas D. Genetic structure of Tunisian ethnic groups revealed by paternal

lineages. Am J Phys Anthropol. 2011 Oct;146(2):271–80.

142. Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, et al. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. Am J Hum Genet. 2004;75(2):338–45.

143. Fadhlaoui-Zid K, Haber M, Martínez-Cruz B, Zalloua P, Benammar Elgaaied A, Comas D. Genome-wide and paternal diversity reveal a recent origin of human populations in North Africa. PLoS One. 2013 Jan;8(11):e80293.

144. Goode MR, Cheong SY, Li N, Ray WC, Bartlett CW. Collection and Extraction of Saliva DNA for Next Generation Sequencing Video Link. J Vis Exp. 2014;(10):516973791–51697.

145. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, et al. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci. 2006 Jan;51(1):64–75.

146. Martínez-Cruz B, Ziegle J, Sanz P, Sotelo G, Anglada R, Plaza S, et al. Multiplex single-nucleotide polymorphism typing of the human Y chromosome using TaqMan probes. Investig Genet. BioMed Central Ltd; 2011 Jan;2(1):13.

147. McGuigan FEA, Ralston SH. Single nucleotide polymorphism detection: allelic discrimination using TaqMan. Psychiatr Genet. England; 2002 Sep;12(3):133–6.

148. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.

149. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.

150. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten

years of next-generation sequencing technology. Trends in Genetics. 2014;

151. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012.

152. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;5(6).

153. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012;40(1):1–8.

154. Craig DW, Pearson J V, Szelinger S, Sekar A, Margot R, Corneveaux JJ, et al. Identification of Genetic Variants Using Barcoded Multiplexed Sequencing. Nat Methods. 2008;5(10):887–93.

155. Buermans HPJ, Den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. Elsevier B.V.; 2014;1842(10):1932–41.

156. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. Nature Publishing Group; 2010;11(1):31–46.

157. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

158. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7.

159. Lunter G, Goodson M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936–9.

160. Langmead B, Salzberg SL. Fast gapped-read alignment with

Bowtie 2. Nat Methods. 2012;9(4):357–9.

161. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–9.

162. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.

163. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013. 1-33 p.

164. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics. 1992;131(2):479–91.

165. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10:564–7.

166. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16(1):37–48.

167. Cox MP. Accuracy of Molecular Dating with the Rho Statistic: Deviations from Coalescent Expectations Under a Range of Demographic Models. Hum Biol. Wayne State University Press ; 2009 Dec;81(5–6):911–33.

168. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. Nature Publishing Group; 2017;

169. Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, et al. Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun Commun. 2015;(6):7152.

170. Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, et al. The Effective Mutation Rate at Y Chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. Am J Hum Genet. The American Society of Human Genetics; 2004 Jan 19;74(1):50–61.

171. Busby GBJ, Brisighelli F, Sanchez-Diz P, Ramos-Luis E, Martinez-Cadenas C, Thomas MG, et al. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. Proc R Soc B Biol Sci. 2012;279(1730):884–92.

172. Wei W, Ayub Q, Xue Y, Tyler-Smith C. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci Int Genet. Elsevier; 2013 Dec;7(6):568–72.

173. Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, et al. Inferring continental ancestry of argentineans from Autosomal, Y-chromosomal and mitochondrial DNA. Ann Hum Genet. England; 2010 Jan;74(1):65–76.

174. Rojas W, Parra MV, Campo O, Caro MA, Lopera JG, Arias W, et al. Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. Am J Phys Anthropol. United States; 2010 Sep;143(1):13–20.

175. Alves-Silva J, da Silva Santos M, Guimarães PEM, Ferreira ACS, Bandelt H-J, Pena SDJ, et al. The Ancestry of Brazilian mtDNA Lineages. American Journal of Human Genetics. 2000. p. 444–61.

176. Henn BM, Cavalli-Sforza LL, Feldman MW. The great

human expansion. Proc Natl Acad Sci U S A. 2012 Oct 30;109(44):17758–64.

177. Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, Op den Velde Boots PM, et al. Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: an online community approach. PLoS One. 2012 Jan;7(7):e41634.

178. Vogt PH. AZF deletions and Y chromosomal haplogroups: history and update based on sequence. Hum Reprod Update. England; 2005;11(4):319–36.

179. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. Springer Berlin Heidelberg; 2017;1–15.

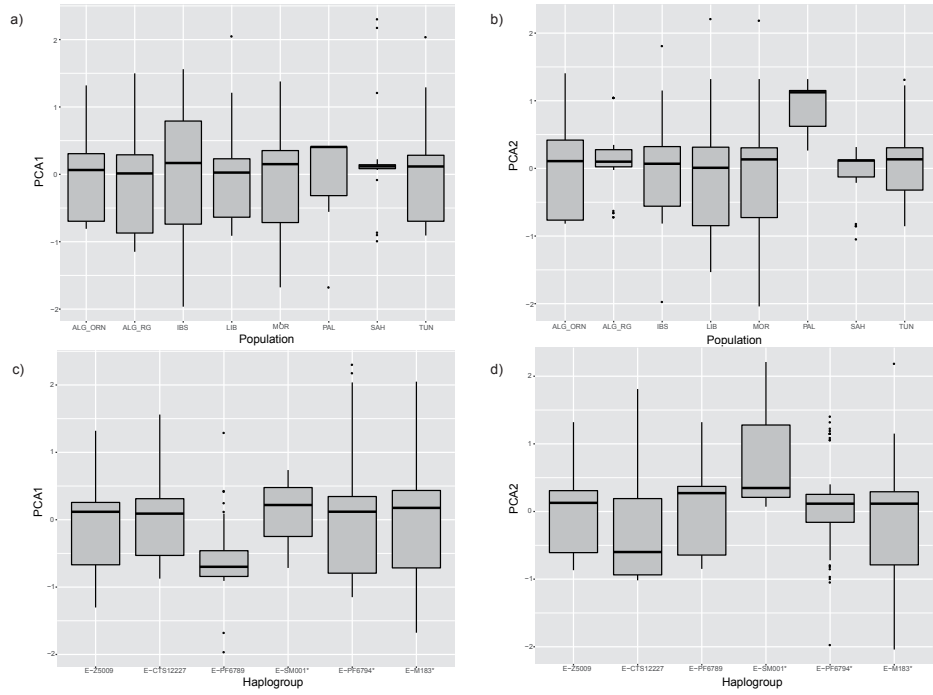
180. Martinez-Cadenas C, Blanco-Verea A, Hernando B, Busby GBJ, Brion M, Carracedo A, et al. The relationship between surname frequency and Y chromosome variation in Spain. Eur J Hum Genet. Nature Publishing Group; 2016;24(1):120–8.

181. Boattini A, Sarno S, Pedrini P, Medoro C, Carta M, Tucci S, et al. Traces of medieval migrations in a socially stratified population from Northern Italy. Evidence from uniparental markers and deep-rooted pedigrees. Heredity (Edinb). 2014 Sep;

182. Gerstenberger J, Hummel S, Schultes T, Häck B, Herrmann B. Reconstruction of a Historical Genealogy by means of STR Analysis and Y-haplotyping of Ancient DNA. Eur J Hum Genet. 1999;7(4):469–77.

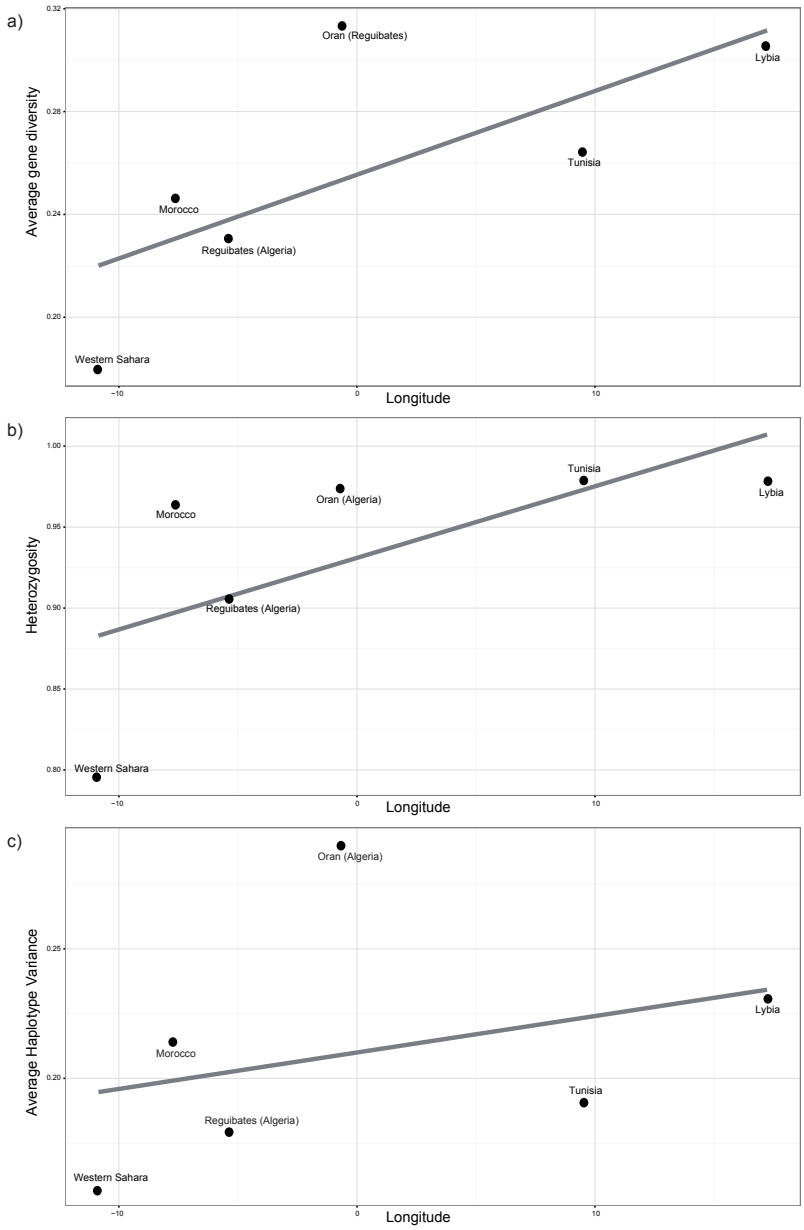
8. ELECTRONIC APPENDIX

8.1. Supplementary information for section 4.3

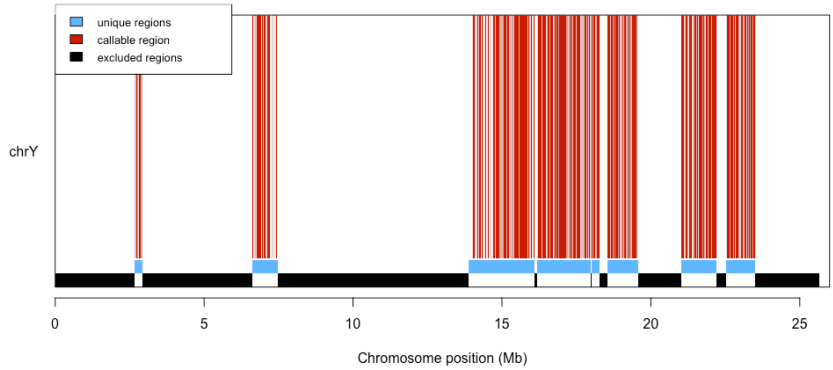


Supplementary Figure S1. Box plots of the mean values of PCA1 (a and c) and PCA2 (b and d). Populations are shown in a) and b), and haplogrups in c) and d).

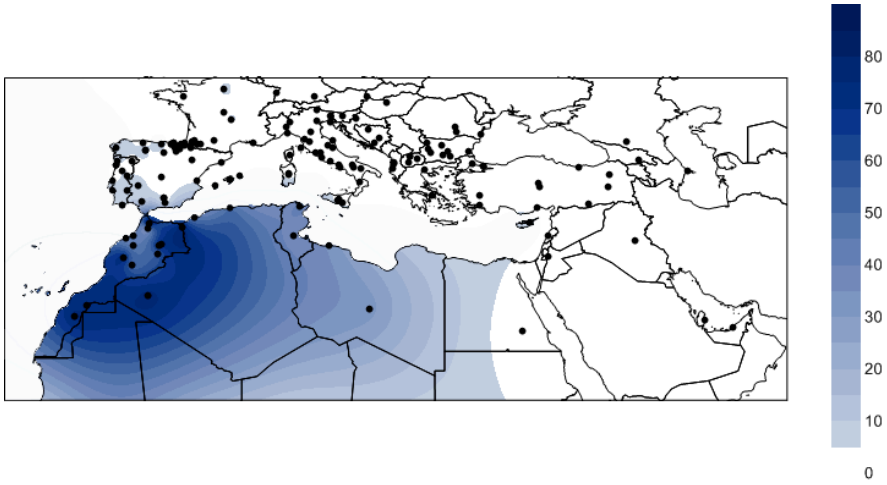
.



Supplementary Figure S2. Diversity indices as a function of the longitude. Population a) gene diversity, b) heterozygosity and c) Y-STR haplotype variance in North Africa. Black lines are showing linear regressions of each diversity index onto longitude.



Supplementary Figure S3. Callability mask for the Y chromosome. Blue lines define the so-called unique regions defined by Wei *et al.* (8), vertical red lines define the callable region, and black lines indicate the regions that have been excluded for the analysis. Each black dot shows geographic data points where data was available.



Supplementary Figure S4. Contour maps of the derived allele frequencies of EM81 constructed using the Surfer Golden software v14 (Golden Software, Golden, CO, USA) (http://www.goldensoftware.com/products/surfer).

Supplementary Table S1. Description of the samples included in the analysis.

SampleID	Origin	References	Mean depth	Haplogroup	Haplogroup (mutation	% missing data	Source
HAN	China (Han)	Reich et al. 2010	14.2	O2a2c2	O-AM01744	0.16	Lymphoblastoid cells
MAN01	Senegal (Mandenka)	Reich et al. 2010	12.51	E1a1	E-M44	0.07	Lymphoblastoid cells
MBU01	Democratic Republic of Congo (Mbuti Pygmy)	Reich et al. 2010	12.61	E1b1a1a1f1a1	E-U174	0.28	Lymphoblastoid cells
SAN01	Namibia (JuhoaniSan)	Reich et al. 2010	16.36	A2a1a	A-P28	0.34	Lymphoblastoid cells
YOR01	Nigeria (Yoruba)	Reich et al. 2010	16.57	E1b1a1a1f1a1	E-U174	0.07	Lymphoblastoid cells
DAI	China (Dai)	Reich et al. 2010	13.04	O1b	O-M268	0.02	Lymphoblastoid cells
SAR	Italy	Reich et al. 2010	12.92	I2a1	I-AM00517	0.09	Lymphoblastoid cells
DNK02	South Sudan (Dinka)	Reich et al. 2010	14.81	E2	E-M75	0.48	Saliva
DNK07	South Sudan (Dinka)	Prüfer et al. 2013	18.02	A3b2f	A-AMM1045	0.11	Saliva
FRE01	France	Reich et al. 2010	13.91	I1	I-M253	0.07	Lymphoblastoid cells
MAN02	Senegal (Mandenka)	Prüfer et al. 2013	18.33	E1b1b1a1a	E-V12	0.02	Lymphoblastoid cells
MBU02	Democratic Republic of Congo (Mbuti Pygmy)	Prüfer et al. 2013	18.22	E2b1a	E-M200	0.05	Lymphoblastoid cells
SAN02	Namibia (JuhoaniSan)	Prüfer et al. 2013	20.05	A3b1	A-M51	0.39	Lymphoblastoid cells
YOR02	Nigeria (Yoruba)	Prüfer et al. 2013	19.45	E1b1a1a1f1a1	E-U174	0.05	Lymphoblastoid cells
YOR03	Nigeria (Yoruba)	Schuster et al. 2010	22.47	E1b1a1a1f1a1	E-U174	0.02	Blood
BAN03	WestAfricanBantu	Lorente-Galdós et al. unpublished work	14.43	E1b1a1a1f1a1	E-U174	0.14	Blood
LIB02	Libya	Lorente-Galdós et al. unpublished work	12.89	E1b1b1b1b	E-M183	0.09	Blood
SAH01	Western Sahara	Lorente-Galdós et al. unpublished work	12.01	E1b1b1b1b	E-M183	0.25	Blood
SAN04	Sub-Saharan Africa (Khomani San)	Lorente-Galdós et al. unpublished work	11.27	A3b1	A-M51	0.64	Blood
TOU	Chad (Toubou)	Lorente-Galdós et al. unpublished work	11.96	T1	T-M70	0.05	Blood
ALG03	Algeria	YFULL	108.66	E1b1b1b1b	E-M183	0.05	Saliva

ALG04	Algeria	YFULL	74.29	E1b1b1b1b	E-M183	0.16	Saliva
ALG05	Algeria	YFULL	58.7	E1b1b1b1b	E-M183	0.05	Saliva
BEL01	Belarus	YFULL	74.37	E1b1b1a2c	E-AMM1902	0.02	Saliva
BUL01	Bulgaria	YFULL	62.73	E1b1b1a2c	E-AMM1902	0.85	Saliva
BUL02	Bulgaria	YFULL	49.99	E1b1b1a2c	E-AMM1902	0.02	Saliva
BUL03	Bulgaria	YFULL	78.09	E1b1b1a2c	E-AMM1902	0.87	Saliva
FRE02	France	YFULL	126.92	E1b1b1b1b	E-M183	0.05	Saliva
FRE03	France	YFULL	71.52	E1b1b1b1b	E-M183	0.16	Saliva
LEB01	Lebanon	YFULL	56.72	E1b1b1b1b	E-M183	0.07	Saliva
LIB03	Libya	YFULL	81.61	E1b1b1b1b	E-M183	0.21	Saliva
LIB04	Libya	YFULL	61.6	E1b1b1b1b	E-M183	0	Saliva
LIB05	Libya	YFULL	68.59	E1b1b1b1b	E-M183	0.21	Saliva
MOR03	Morocco	YFULL	62.8	E1b1b1b1b	E-M183	0.02	Saliva
MOR04	Morocco	YFULL	59.8	E1b1b1b1b	E-M183	0.11	Saliva
NOR01	Norway	YFULL	88.16	E1b1b1a2c	E-AMM1902	0.34	Saliva
NOR02	Norway	YFULL	46.22	E1b1b1a2c	E-AMM1902	1.72	Saliva
PAL01	Palestina	YFULL	80.81	E1b1b1b1b	E-M183	0.05	Saliva
SAH02	Western Sahara	YFULL	72.83	E1b1b1b1b	E-M183	0.07	Saliva
SAH03	Western Sahara	YFULL	35.53	E1b1b1b1b	E-M183	0.02	Saliva
SAH04	Western Sahara	YFULL	94.7	E1b1b1b1b	E-M183	0.05	Saliva
SAU01	Saudi Arabia	YFULL	65.48	E1b1b1b1b	E-M183	0.05	Saliva
TUN02	Tunisia	YFULL	69.26	E1b1b1b1b	E-M183	0	Saliva
UK01	United Kingdom	YFULL	75.9	E1b1b1b1b	E-M183	0	Saliva
UK02	United Kingdom	YFULL	142.12	E1b1b1b1b	E-M183	0	Saliva
USA01	United States of America	YFULL	110.28	E1b1b1b1b	E-M183	0	Saliva
USA02	United States of America	YFULL	70.86	E1b1b1b1b	E-M183	0.02	Saliva
ALG01	Algeria	This study	15.52	E1b1b1b1b	E-M183	0.05	Blood
ALG02	Algeria	This study	15.28	E1b1b1b1b	E-M183	0.02	Blood
BAS01	Spain	This study	14.74	R1b1a2a1b1a1a1a	R-Z214	0	Blood
BTUN01	Tunisia (Berber)	This study	15.4	E1b1b1b1b	E-M183	0	Blood

BTUN02	Tunisia (Berber)	This study	15.01	E1b1b1b1b	E-M183	0	Blood
BZEN01	Algeria (Berber)	This study	14.88	E1b1b1b1b	E-M183	0.02	Blood
BZEN02	Algeria (Berber)	This study	15.32	E1b1b1b1b	E-M183	0.02	Blood
EGY01	Egypt	This study	16.72	T1	T-M70	0.14	Blood
EGY02	Egypt	This study	14.03	E1b1a1a1f1a1	E-U174	0.07	Blood
IRQ01	Iraq	This study	14.8	J1	J-M267	0.07	Blood
IRQ02	Iraq	This study	14.85	L1a	L-M27	0.07	Blood
LIB01	Libya	This study	12.29	T1	T-M70	0.11	Blood
MOR01	Morocco	This study	16.13	E1b1b1b1b	E-M183	0	Blood
MOR02	Morocco	This study	15.7	E1b1b1b1b	E-M183	0	Blood
TUN01	Tunisia	This study	15.06	E1b1b1b1b	E-M183	0.05	Blood

Supplementary Table S2. ANOVA test of principal component mean values. At a a) population level and b) haplogroup level.

a)

		Df	Sum Sq	Mean Sq	F value	Pr (>F)
PCA1	Population	7	1.3282	0.1897	0.2972	0.9545
	Residuals	227	144.9286	0.6385		
PCA2	Population	7	6.5503	0.9358	2.1255	0.0419
	Residuals	227	99.9368	0.4403		

b)

		Df	Sum Sq	Mean Sq	F value	Pr (>F)
PCA1	Haplogroups	5	10.2759	2.0552	3.4611	0.0049
	Residuals	229	135.9809	0.5938		
PCA2	Haplogroups	5	4.2159	0.8432	1.8880	0.0973
	Residuals	229	102.2712	0.4466		

Supplementary Table S3. Diversity indices. At a a) population level and b) haplogroup level.

a)

Population	Het	D	Variance
Western Sahara	0.795 (0.088)	0.18 (0.118)	0.157 (0.187)
Morocco	0.964 (0.011)	0.246 (0.142)	0.214 (0.172)
Algeria (Oran)	0.974 (0.025)	0.314 (0.203)	0.29 (0.32)
Algeria (Reguibates)	0.906 (0.032)	0.231 (0.14)	0.18 (0.22)
Tunisia	0.98 (0.013)	0.264 (0.158)	0.191 (0.166)
Libya	0.978 (0.019)	0.305 (0.186)	0.231 (0.258)
Middle East	1 (0.096)	0.179 (0.13)	0.121 (0.204)
Iberian Peninsula	0.961 (0.024)	0.275 (0.168)	0.227 (0.3)

b)

Haplogroup	Het	D	Variance
E-CTS12227	0.975 (0.013)	0.252 (0.149)	0.233 (0.184)
E-PF6789	0.924 (0.038)	0.243 (0.154)	0.155 (0.136)
E-SM001*	0.926 (0.017)	0.226 (0.135)	0.184 (0.185)
E-Z5009	0.952 (0.013)	0.287 (0.271)	0.2 (0.182)
E-M183*	1 (0.045)	0.303 (0.201)	0.226 (0.273)

Supplementary Table S4. List of sampling location of the volunteers genotyped for the present analysis.

Population	Latitude	Longitude	Sample size	Study	Source
Western Sahara	27.65	-10.86	26	Plaza et al. 2003	Blood
Morocco	33.57	-7.59	141	Plaza et al. 2003	Blood
Algeria (Oran)	35.70	-0.63	51	Bekada et al. 2015	Blood
Algeria (Reguibates)	27.70	-5.33	39	Bekada et al. 2015	Blood
Tunisia	33.89	9.54	91	Plaza et al. 2003	Blood
Lybia	26.34	17.23	77	Fadhlaoui-Zid et al. 2013	Blood
Middle East	32	35.25	374[1]	Zalloua et al. 2008; El-Sibai et al. 2009	Blood & buccal swabs
Iberian Peninsula	41.39	2.17	1085[1]	Solé-Morata et al. 2015	Saliva

[1]From these datasets we genotyped only those samples that were derived for M183 (Iberian=23 and Middle East=6)

Supplementary Table S5. Genotypes of the M183 derived samples.

Sample	Population	HG	M183	SM001	PF6794	PF6789	CTS12227	Z5009
LIB06	Libya	E-PF6789	C	T	C	A	A	G
LIB07	Libya	E-PF6789	C	T	C	A	A	G
LIB08	Libya	E-CTS12227	C	T	T	G	G	G
LIB10	Libya	E-PF6789	C	T	C	A	A	G
LIB11	Libya	E-PF6789	C	T	C	A	A	G
LIB13	Libya	E-PF6789	C	T	C	A	A	G
LIB14	Libya	E-PF6789	C	T	C	A	A	G
LIB16	Libya	E-PF6789	C	T	C	A	A	G
LIB17	Libya	E-PF6789	C	T	C	A	A	G
LIB18	Libya	E-Z5009	C	T	T	G	A	A
LIB19	Libya	E-Z5009	C	T	T	G	A	A
LIB20	Libya	E-SM001	C	T	T	G	A	G
LIB21	Libya	E-Z5009	C	T	T	G	A	A
LIB22	Libya	E-PF6789	C	T	C	A	A	G
LIB23	Libya	E-M183*	C	G	T	G	A	G
LIB24	Libya	UN[1]	C	T	C	UN	A	G
LIB25	Libya	E-PF6789	C	T	C	A	A	G
LIB26	Libya	E-Z5009	C	T	T	G	A	A
LIB27	Libya	E-CTS12227	C	T	T	G	G	G
LIB28	Libya	E-PF6789	C	T	C	A	A	G
LIB29	Libya	E-PF6789	C	T	C	A	A	G
LIB30	Libya	E-CTS12227	C	T	T	G	G	G
LIB31	Libya	E-PF6794*	C	T	C	G	A	G
LIB32	Libya	E-PF6789	C	T	C	A	A	G
LIB33	Libya	E-Z5009	C	T	T	G	A	A
LIB34	Libya	E-M183*	C	G	T	G	A	G
TUN03	Tunisia	E-CTS12227	C	T	T	G	G	G
TUN04	Tunisia	E-Z5009	C	T	T	G	A	A
PAL03	Middle East	E-SM001	C	T	T	G	A	G
TUN06	Tunisia	E-Z5009	C	T	T	G	A	A

TUN01	Tunisia	E-PF6789	C	T	C	A	A	G
TUN08	Tunisia	E-Z5009	C	T	T	G	A	A
PAL04	Middle East	E-PF6789	C	T	C	A	A	G
PAL07	Middle East	E-SM001	C	T	T	G	A	G
TUN09	Tunisia	E-SM001	C	T	T	G	A	G
TUN10	Tunisia	E-Z5009	C	T	T	G	A	A
TUN11	Tunisia	E-PF6789	C	T	C	A	A	G
PAL05	Middle East	E-Z5009	C	T	T	G	A	A
TUN12	Tunisia	E-Z5009	C	T	T	G	A	A
TUN13	Tunisia	E-Z5009	C	T	T	G	A	A
TUN14	Tunisia	E-PF6789	C	T	C	A	A	G
TUN15	Tunisia	E-Z5009	C	T	T	G	A	A
TUN16	Tunisia	E-PF6789	C	T	C	A	A	G
TUN17	Tunisia	E-Z5009	C	T	T	G	A	A
TUN18	Tunisia	E-SM001	C	T	T	G	A	G
TUN19	Tunisia	E-SM001	C	T	T	G	A	G
TUN20	Tunisia	E-SM001	C	T	T	G	A	G
TUN21	Tunisia	E-Z5009	C	T	T	G	A	A
TUN22	Tunisia	E-PF6789	C	T	C	A	A	G
TUN23	Tunisia	E-CTS12227	C	T	T	G	G	G
TUN24	Tunisia	E-SM001	C	T	T	G	A	G
TUN25	Tunisia	E-SM001	C	T	T	G	A	G
PAL02	Middle East	E-SM001	C	T	T	G	A	G
PAL06	Middle East	E-Z5009	C	T	T	G	A	A
TUN26	Tunisia	E-SM001	C	T	T	G	A	G
TUN27	Tunisia	E-SM001	C	T	T	G	A	G
TUN28	Tunisia	E-PF6789	C	T	C	A	A	G
TUN29	Tunisia	E-Z5009	C	T	T	G	A	A
TUN30	Tunisia	E-PF6789	C	T	C	A	A	G
TUN31	Tunisia	E-PF6789	C	T	C	A	A	G
MOR05	Morocco	E-CTS12227	C	T	T	G	G	G
MOR06	Morocco	E-Z5009	C	T	T	G	A	A
MOR07	Morocco	E-M183*	C	G	T	G	A	G
MOR08	Morocco	E-Z5009	C	T	T	G	A	A

MOR09	Morocco	E-SM001	C	T	T	G	A	G
TUN32	Tunisia	E-CTS12227	C	T	T	G	G	G
MOR10	Morocco	E-SM001	C	T	T	G	A	G
CAT01	Iberian Peninsula	E-SM001	C	T	T	G	A	G
TUN33	Tunisia	E-Z5009	C	T	T	G	A	A
CAT02	Iberian Peninsula	E-PF6789	C	T	C	A	A	G
CAT03	Iberian Peninsula	E-CTS12227	C	T	T	G	G	G
CAT04	Iberian Peninsula	E-Z5009	C	T	T	G	A	A
CAT05	Iberian Peninsula	E-CTS12227	C	T	T	G	G	G
CAT06	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT07	Iberian Peninsula	E-Z5009	C	T	T	G	A	A
MOR11	Morocco	E-Z5009	C	T	T	G	A	A
MOR12	Morocco	E-SM001	C	T	T	G	A	G
MOR13	Morocco	E-SM001	C	T	T	G	A	G
CAT08	Iberian Peninsula	E-SM001	C	T	T	G	A	G
MOR14	Morocco	E-Z5009	C	T	T	G	A	A
MOR15	Morocco	E-SM001	C	T	T	G	A	G
MOR16	Morocco	E-Z5009	C	T	T	G	A	A
MOR17	Morocco	E-Z5009	C	T	T	G	A	A
MOR18	Morocco	E-CTS12227	C	T	T	G	G	G
MOR20	Morocco	E-CTS12227	C	T	T	G	G	G
MOR21	Morocco	E-Z5009	C	T	T	G	A	A
MOR22	Morocco	E-Z5009	C	T	T	G	A	A
MOR23	Morocco	E-Z5009	C	T	T	G	A	A
MOR24	Morocco	E-Z5009	C	T	T	G	A	A
MOR01	Morocco	E-PF6789	C	T	C	A	A	G
MOR25	Morocco	E-CTS12227	C	T	T	G	G	G
MOR26	Morocco	E-Z5009	C	T	T	G	A	A
MOR27	Morocco	E-CTS12227	C	T	T	G	G	G
MOR28	Morocco	E-Z5009	C	T	T	G	A	A
MOR29	Morocco	E-Z5009	C	T	T	G	A	A
MOR30	Morocco	E-SM001	C	T	T	G	A	G
MOR31	Morocco	E-Z5009	C	T	T	G	A	A
MOR32	Morocco	E-Z5009	C	T	T	G	A	A

MOR33	Morocco	E-SM001	C	T	T	G	A	G
MOR35	Morocco	E-Z5009	C	T	T	G	G	A
MOR36	Morocco	UN[1]	C	G	UN	UN	UN	UN
MOR37	Morocco	E-Z5009	C	T	T	G	A	A
MOR38	Morocco	E-Z5009	C	T	T	G	A	A
MOR39	Morocco	E-Z5009	C	T	T	G	A	A
MOR40	Morocco	E-Z5009	C	T	T	G	A	A
MOR41	Morocco	E-CTS12227	C	T	T	G	G	G
MOR42	Morocco	E-Z5009	C	T	T	G	A	A
MOR43	Morocco	E-Z5009	C	T	T	G	A	A
MOR44	Morocco	E-CTS12227	C	T	T	G	G	G
MOR45	Morocco	E-Z5009	C	T	T	G	A	A
MOR46	Morocco	E-Z5009	C	T	T	G	A	A
MOR47	Morocco	E-Z5009	C	T	T	G	G	A
MOR48	Morocco	E-CTS12227	C	T	T	G	G	G
MOR49	Morocco	E-CTS12227	C	T	T	G	G	G
MOR50	Morocco	E-M183*	C	G	T	G	A	G
MOR51	Morocco	E-CTS12227	C	T	T	G	G	G
MOR52	Morocco	E-CTS12227	C	T	T	G	G	G
MOR02	Morocco	E-Z5009	C	T	T	G	A	A
MOR53	Morocco	E-SM001	C	T	T	G	A	G
MOR54	Morocco	E-Z5009	C	T	T	G	A	A
MOR55	Morocco	E-Z5009	C	T	T	G	A	A
MOR56	Morocco	E-Z5009	C	T	T	G	A	A
MOR57	Morocco	E-CTS12227	C	T	T	G	G	G
MOR58	Morocco	E-Z5009	C	T	T	G	A	A
MOR59	Morocco	E-SM001	C	T	T	G	A	G
MOR60	Morocco	E-CTS12227	C	T	T	G	G	G
MOR61	Morocco	E-CTS12227	C	T	T	G	G	G
MOR62	Morocco	E-Z5009	C	T	T	G	A	A
MOR63	Morocco	E-Z5009	C	T	T	G	A	A
MOR64	Morocco	E-SM001	C	T	T	G	A	G
MOR65	Morocco	E-Z5009	C	T	T	G	A	A
MOR66	Morocco	E-Z5009	C	T	T	G	A	A

MOR67	Morocco	E-PF6789	C	T	C	A	A	G
MOR68	Morocco	E-SM001	C	T	T	G	A	G
MOR69	Morocco	E-SM001	C	T	T	G	A	G
MOR70	Morocco	E-Z5009	C	T	T	G	A	A
MOR72	Morocco	E-Z5009	C	T	T	G	A	A
MOR73	Morocco	E-Z5009	C	T	T	G	A	A
CAT09	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT10	Iberian Peninsula	E-PF6789	C	T	C	A	A	G
ALGO01	Algeria (ORN)	E-Z5009	C	T	T	G	A	A
ALGO02	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO03	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO04	Algeria (ORN)	E-PF6789	C	T	C	A	A	G
ALGO05	Algeria (ORN)	E-Z5009	C	T	T	G	A	A
ALGO06	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO07	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO08	Algeria (ORN)	E-PF6789	C	T	C	A	A	G
ALGO09	Algeria (ORN)	E-SM001	C	T	T	G	A	G
ALGO10	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO11	Algeria (ORN)	E-PF6789	C	T	C	A	A	G
ALGO12	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO13	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO14	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO15	Algeria (ORN)	E-Z5009	C	T	T	G	A	A
ALGO16	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO17	Algeria (ORN)	E-SM001	C	T	T	G	A	G
ALGO18	Algeria (ORN)	E-SM001	C	T	T	G	A	G
ALGO19	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
ALGO20	Algeria (ORN)	E-SM001	C	T	T	G	A	G
ALGO21	Algeria (ORN)	E-Z5009	C	T	T	G	A	A
ALGO22	Algeria (ORN)	E-CTS12227	C	T	T	G	G	G
TUN35	Tunisia	E-CTS12227	C	T	T	G	G	G
TUN36	Tunisia	E-CTS12227	C	T	T	G	G	G
TUN37	Tunisia	E-PF6789	C	T	C	A	A	G
MOR74	Morocco	E-Z5009	C	T	T	G	A	A

MOR75	Morocco	E-Z5009	C	T	T	G	A	A
MOR76	Morocco	E-Z5009	C	T	T	G	A	A
MOR77	Morocco	E-SM001	C	T	T	G	A	G
MOR78	Morocco	E-Z5009	C	T	T	G	A	A
MOR79	Morocco	E-M183*	C	G	T	G	A	G
MOR80	Morocco	E-CTS12227	C	T	T	G	G	G
MOR81	Morocco	E-SM001	C	T	T	G	A	G
MOR82	Morocco	E-Z5009	C	T	T	G	A	A
MOR83	Morocco	E-CTS12227	C	T	T	G	G	G
MOR84	Morocco	E-PF6789	C	T	C	A	A	G
MOR85	Morocco	E-CTS12227	C	T	T	G	G	G
MOR86	Morocco	E-CTS12227	C	T	T	G	G	G
MOR87	Morocco	E-CTS12227	C	T	T	G	G	G
MOR88	Morocco	E-Z5009	C	T	T	G	A	A
MOR89	Morocco	E-Z5009	C	T	T	G	A	A
MOR90	Morocco	E-Z5009	C	T	T	G	A	A
MOR91	Morocco	E-Z5009	C	T	T	G	A	A
MOR92	Morocco	E-Z5009	C	T	T	G	A	A
MOR93	Morocco	E-Z5009	C	T	T	G	A	A
CAT11	Iberian Peninsula	UN[1]	C	UN	T	G	A	G
ALGR01	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR02	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR03	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR04	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR05	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR06	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR07	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR08	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR09	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR10	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR11	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR12	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR13	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR14	Algeria (RG)	E-SM001	C	T	T	G	A	G

ALGR15	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR16	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR17	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR18	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR19	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR20	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR21	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR22	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR23	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR24	Algeria (RG)	E-Z5009	C	T	T	G	A	A
ALGR25	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR26	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR27	Algeria (RG)	E-PF6794*	C	T	C	G	A	G
ALGR28	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR29	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR30	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR31	Algeria (RG)	E-SM001	C	T	T	G	A	G
ALGR32	Algeria (RG)	E-SM001	C	T	T	G	A	G
CAT12	Iberian Peninsula	E-Z5009	C	T	T	G	A	A
CAT13	Iberian Peninsula	E-PF6794*	C	T	C	G	A	G
SAH05	Western Sahara	E-SM001	C	T	T	G	A	G
SAH06	Western Sahara	E-SM001	C	T	T	G	A	G
SAH07	Western Sahara	E-SM001	C	T	T	G	A	G
SAH08	Western Sahara	E-SM001	C	T	T	G	A	G
SAH09	Western Sahara	E-SM001	C	T	T	G	A	G
SAH10	Western Sahara	E-SM001	C	T	T	G	A	G
SAH11	Western Sahara	E-SM001	C	T	T	G	A	G
SAH12	Western Sahara	E-SM001	C	T	T	G	A	G
SAH13	Western Sahara	E-Z5009	C	T	T	G	A	A
SAH14	Western Sahara	E-SM001	C	T	T	G	A	G
SAH15	Western Sahara	E-SM001	C	T	T	G	A	G
SAH16	Western Sahara	E-Z5009	C	T	T	G	A	A
SAH17	Western Sahara	E-Z5009	C	T	T	G	A	A
SAH18	Western Sahara	E-SM001	C	T	T	G	A	G

SAH19	Western Sahara	E-SM001	C	T	T	G	A	G
SAH20	Western Sahara	E-SM001	C	T	T	G	A	G
SAH21	Western Sahara	E-Z5009	C	T	T	G	A	A
SAH22	Western Sahara	E-SM001	C	T	T	G	A	G
SAH23	Western Sahara	E-SM001	C	T	T	G	A	G
SAH24	Western Sahara	E-SM001	C	T	T	G	A	G
CAT14	Iberian Peninsula	E-M183*	C	G	T	G	A	G
CAT15	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT16	Iberian Peninsula	E-M183*	C	G	T	G	A	G
CAT17	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT18	Iberian Peninsula	E-M183*	C	G	T	G	A	G
CAT19	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT20	Iberian Peninsula	E-M183*	C	G	T	G	A	G
CAT21	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT22	Iberian Peninsula	E-SM001	C	T	T	G	A	G
CAT23	Iberian Peninsula	E-M183*	C	G	T	G	A	G

[1]Samples excluded from the downstream analysis

Supplementary Table S6. Taqman assays information.

Assay Name	Product Type	Catalog numer	Assay ID	dbSNP	Forward Primer Name	Forward Primer Seq.	Forward Primer Conc.
M183	Functionally Tested	4351379	C___1083235_20	rs2032600			
E3_Z5009	Custom Taqman(R) SNP Genotyping Assay Service		AHLJ2I2		E3_Z5009_F	GAAGGACCATGCG TTCAGTTG	36
E2_CTS12227	Custom Taqman(R) SNP Genotyping Assay Service		AHKA4CU		E2_CTS12227_F	AACAGATCTGCTCA ATCTTTGTCTGATT	36
E1a_PF6789	Custom Taqman(R) SNP Genotyping Assay Service		AHI156M		E1a_PF6789_F	ACCTCAGTGCTTGG CCCATATATA	36
E1_PF6794	Custom Taqman(R) SNP Genotyping Assay Service		AHHS70E		E1_PF6794_F	AATGTCCGTCAACT GATGAATGGAT	36
E0_SM001	Custom Taqman(R) SNP Genotyping Assay Service		AHGJ9T6		E0_SM001_F	ATTGGGCCCTCACC ATACAC	36

Reverse Primer Name	Reverse Primer Seq.	Reverse Primer Conc.	Reporter 1 Name	Reporter 1 Dye	Reporter 1 Sequence
E3_Z5009_R	GCATCTATGCCAC TTGCTTAGTTTG	36	E3_Z5009_V	VIC	TAGGTTAGAGGCTTTCTG
E2_CTS12227_R	GCAAAATATCAAG GCCCAAGTAACC	36	E2_CTS12227_V	VIC	TGCATGATCAACTTCA
E1a_PF6789_R	ATGACTCATATTT GCACATCAACACA AC	36	E1a_PF6789_V	VIC	CTCTCATTGGCATTATG
E1_PF6794_R	GTTCATTCATGTTG CAGCATTTATTAG TG	36	E1_PF6794_V	VIC	TGAATAATGTTCCATTGTATGTAT
E0_SM001_R	CATTGCTCACAGT TCTGGAGACT	36	E0_SM001_V	VIC	ATCAAGATATGGTAGATTCA

Reporter 1 Conc.	Reporter 1 Quencher	Reporter 2 Name	Reporter 2 Dye	Reporter 2 Sequence	Reporter 2 Conc.	Reporter 2 Quencher
8	NFQ	E3_Z5009_M	FAM	TAGGTTAGAGACTTTCTG	8	NFQ
8	NFQ	E2_CTS12227_M	FAM	TGCATGATCAGCTTCA	8	NFQ
8	NFQ	E1a_PF6789_M	FAM	CTCTCATTGACATTATG	8	NFQ
8	NFQ	E1_PF6794_M	FAM	ATGTTCCATTGCATGTAT	8	NFQ
8	NFQ	E0_SM001_M	FAM	ATCAAGATATGTTAGATTCA	8	NFQ

Context Sequence [VIC/FAM]	Gene Symbol	Gene Name	Chromosome	Species	SNP Type
ATTATCACAAGGAAAGAATGATTCT[A/C]CCTATGTCGAGGTTTGTGTGAAGTT	USP9Y	ubiquitin specific peptidase 9; Y-linked	Y	Human	Transversion Substitution;Silent Mutation;Intragenic
			Y	Human	
			Y	Human	
			Y	Human	
			Y	Human	
			Y	Human	

Minor Allele Freq - Caucasian	NCBI Assembly Build Number	Location on NCBI Assembly
	37	chr Y: 14888783
	37	chr Y: 16695746
	37	chr Y: 28532290
	37	chr Y: 23125341
	37	chr Y: 17345606
	37	chr Y: 16716119

Reference
Karafet et al. Genome Res. 2008 May; 18(5): 830–838.
Gregory Magoon, Ph.D., Richard Rocca, Vince Tilroe, David F. Reynolds, Bonnie Schrack, Peter M. Op den Velde Boots, Ray H. Banks, Roman Sychev, Victar Mas, Steve Fix, Christian Rottensteiner, Alexander R. Williamson, Ph.D., John Sloan and an anonymous individual, independent researchers of publicly available whole genome sequence datasets, and Thomas Krahn, MSc (Dipl.-Ing.), with support from the genetic genealogy community.
Wei et al. Genome Res. 2013 Feb;23(2):388-95.
Francalacci et al. Science. 2013 Aug 2;341(6145):565-9.
Francalacci et al. Science. 2013 Aug 2;341(6145):565-9.
This study

Supplementary Table S7. Y-STRs haplotypes of each individual.

Code	Population	HG	456	389I	390	458	19	393	391	439	635	H4	437	438	448
ALGO01	Algeria (ORN)	E-Z5009	15	14	24	18	13	13	9			12	14		20
ALGO02	Algeria (ORN)	E-CTS12227	15	14	24	18	13	13	10	11	21	12	14	10	21
ALGO04	Algeria (ORN)	E-PF6789	16	14	24	17	13	13	9		20	12	14	10	20
ALGO05	Algeria (ORN)	E-Z5009	16	13	24	18	13	13	9	10	22	12	14	10	20
ALGO06	Algeria (ORN)	E-CTS12227	16	14	24	19		13	10		22	12	14	11	20
ALGO07	Algeria (ORN)	E-CTS12227	16	14	23	18		13	9			12	14	9	19
ALGO08	Algeria (ORN)	E-PF6789	15	14	24	17	13	13	9		22	12	14	10	20
ALGO09	Algeria (ORN)	E-SM001	17	14	24	17	14	13	9	10	21	12	14	10	21
ALGO10	Algeria (ORN)	E-CTS12227	16	14	24	19	13	13	10	11	21	12	14	10	20
ALGO11	Algeria (ORN)	E-PF6789	16	14	24	17	13	13	9		21	12	14	10	20
ALGO12	Algeria (ORN)	E-CTS12227	15	13	24	18	13	13	9	8	21	12	14	10	
ALGO13	Algeria (ORN)	E-CTS12227	15	14	24	18	13	13	10		21	12	14	10	20
ALGO14	Algeria (ORN)	E-CTS12227	16	14	24	17	13	13	10		21	12	14	10	20
ALGO16	Algeria (ORN)	E-CTS12227	15	14	24	18	13	11	9		21	12	14	10	20
ALGO18	Algeria (ORN)	E-SM001	17	14		18	13	13	9	10	22	12	14	10	20
ALGO19	Algeria (ORN)	E-CTS12227	16	14	24	17	14,3	14	10		20	12	14	10	20
ALGO20	Algeria (ORN)	E-SM001	17	14	24	18		13	9			12	14	10	20
ALGO22	Algeria (ORN)	E-CTS12227	16	14		17	13	13	10			12	14	10	20
ALGR01	Algeria (RG)	E-SM001	16	14	25	17	13	13	9	10	21	12	14	10	19
ALGR02	Algeria (RG)	E-Z5009	17	14	23	19	13	13	9		21	12	14		
ALGR03	Algeria (RG)	E-Z5009	16	14	24	19	13	13	9	10	21	12	15	10	19
ALGR04	Algeria (RG)	E-SM001	15	14	24	17	13	13	9		21	12	14	10	19
ALGR05	Algeria (RG)	E-SM001	15	14	24	17	13	13	9	10	21	12	14	10	19
ALGR06	Algeria (RG)	E-SM001	16	14	25	17	13	13	9	10	21	12	14	10	20
ALGR07	Algeria (RG)	E-SM001	16	13	25	18	13	13	9	10	21	12	14	10	19
ALGR08	Algeria (RG)	E-SM001	16	13	25	18	13	13	9	10	21	12	14	10	19
ALGR09	Algeria (RG)	E-Z5009	17	14	23	19	13	13	9	10	21	12	14	10	20

ALGR10	Algeria (RG)	E-SM001	16	14	25	17	13	13	9	10	21	12	14	10	19
ALGR11	Algeria (RG)	E-SM001	16	14	25	18	13	13	9		21	12	14		20
ALGR12	Algeria (RG)	E-SM001	16	14	25	18	13	13	9		21	12	14	10	20
ALGR13	Algeria (RG)	E-Z5009	16	15	24	19	13	13	9	10	21	12	15	10	
ALGR14	Algeria (RG)	E-SM001	16	14	25	18	13	13	9	10	22	12	14	10	20
ALGR15	Algeria (RG)	E-Z5009	17	14	23	19	13	13	9	10	21	12	14	10	20
ALGR16	Algeria (RG)	E-SM001	15	14	25	17	13	13	9	10	21	12	14	10	19
ALGR17	Algeria (RG)	E-Z5009	16	14	24	17	13	13	9		21	12	14	10	20
ALGR18	Algeria (RG)	E-Z5009	16	14	23	19	13	13	9	10	21	12	14	10	20
ALGR19	Algeria (RG)	E-SM001	16	14	25	18	12	13	9	10	21	12	14	9	19
ALGR20	Algeria (RG)	E-SM001	16	14	25	18	13	13	9		20	12	14		20
ALGR22	Algeria (RG)	E-SM001	16	14		17	13	13	9	10	21	12	14	10	19
ALGR23	Algeria (RG)	E-SM001	16	14	25	18	13	13	9	10	22	13	14	10	20
ALGR24	Algeria (RG)	E-Z5009	17	14	23	19	13	13	9	10	21	12	14	10	20
ALGR25	Algeria (RG)	E-SM001	15	14	24	17	13	13	9	10	21	12	14	10	
ALGR26	Algeria (RG)	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
ALGR27	Algeria (RG)	E-PF6794*	16	14	25	18		13	9	10	22	13	14	10	20
ALGR28	Algeria (RG)	E-SM001	15	14	24	17	13	13	9	10	22	12	14	10	20
ALGR30	Algeria (RG)	E-SM001	16	14	25	17	13	13	9	10	21	12	14	10	19
ALGR31	Algeria (RG)	E-SM001	16	15	25	17	13	13	9	10	21	12	14	10	19
ALGR32	Algeria (RG)	E-SM001	16	14		18	13	13	9		22	12	14		20
CAT01	Iberian Peninsula	E-SM001	16	13	25	18	13	13	9	10	21	12	14	10	20
CAT02	Iberian Peninsula	E-PF6789	15	14	25	16	13	13	9	10	21	12	15	10	20
CAT03	Iberian Peninsula	E-CTS12227	15	14	24	17	13	13	9	10	21	12	14	11	20
CAT04	Iberian Peninsula	E-Z5009	16	14	25	17	13	13	9	10	21	12	14	10	20
CAT05	Iberian Peninsula	E-CTS12227	16	13	24	18	13	13	9	10	21	11	14	10	20
CAT06	Iberian Peninsula	E-SM001	17	14	24	19	15	13	9	10	21	12	14	10	20
CAT07	Iberian Peninsula	E-Z5009	16	14	24	17	13	12	10	10	21	12	14	10	20
CAT08	Iberian Peninsula	E-SM001	17	14	24	18	13	13	9		21	12	14	10	
CAT09	Iberian Peninsula	E-SM001	16	14	24	19	13	13	9	10	23	12	14	10	20
CAT10	Iberian Peninsula	E-PF6789	16	14	24	18	13	13	9	10	21	12	14	10	20
CAT11	Iberian Peninsula	UN[1]	16	14	24	20	13	13	9	10	21	12	14	10	21
CAT12	Iberian Peninsula	E-Z5009	15	14	24	18	13	13	9	10	21	12	14	10	20
CAT13	Iberian Peninsula	E-PF6794*	16	14	24	17	13	13	9	10	21	12	14	10	20

CAT14	Iberian Peninsula	E-M183*	15	14	24	17	13	13	9	10	21	12	14	10	20
CAT15	Iberian Peninsula	E-SM001	16	14	24	19	13	13	9	10	22	12		10	20
CAT16	Iberian Peninsula	E-M183*	18	14	24	19	13	13	9	10	22	11	14	10	20
CAT17	Iberian Peninsula	E-SM001	14	14	24	19	13	13	9	11	21	12	14	10	20
CAT18	Iberian Peninsula	E-M183*	15	14	24	17	14	13	9	10	21	12	14	10	20
CAT19	Iberian Peninsula	E-SM001	15	14	24	17	13	13	9	10	23	12	14	10	20
CAT20	Iberian Peninsula	E-M183*	16	14	24	17		13	9		21	12	14	10	
CAT21	Iberian Peninsula	E-SM001	16	14	24	19		13	9	10	22	12	14	10	
CAT22	Iberian Peninsula	E-SM001	16	14	24	18	13	13	9	10	21	12	14	10	20
CAT23	Iberian Peninsula	E-M183*	16	13	24	18	14	13	9	10	21	12	14	10	20
LIB06	Libya	E-PF6789	15	13	24	17	13	13	9	10	21	12	14	10	20
LIB07	Libya	E-PF6789	16	14	23	17	13	13	9		21	12	14	10	21
LIB08	Libya	E-CTS12227	16	14	24	18	13	13	9		21	11	14	10	20
LIB10	Libya	E-PF6789	16	14	25	18	13	13	9	10	21	12	14	10	20
LIB11	Libya	E-PF6789	15	14	24	18	13	13	9		21	12	14	10	20
LIB13	Libya	E-PF6789	16	14	24	17	13	13	9	10	21	12	14	10	20
LIB14	Libya	E-PF6789	16	14	24	17		13	9		21	12	14		20
LIB16	Libya	E-PF6789	15	13	24	18	13	13	9	10	21	12	14	10	20
LIB18	Libya	E-Z5009	15	14	24	18	13	13	9		21	11	14	10	
LIB19	Libya	E-Z5009	15	14	24	20	13	14	9			12	14		20
LIB21	Libya	E-Z5009	16	15	24	18	13	14	9			12	14		20
LIB22	Libya	E-PF6789	15	13	24	17	13	13	10			12	14		20
LIB23	Libya	E-M183*	15	14	25	18	13	13	9		22	12	14	10	20
LIB24	Libya	UN[1]	16	14	23	19	13	13	9	11	21	12	14	10	20
LIB25	Libya	E-PF6789	17	14	24	17	13	13	9	10	21	12	14	10	20
LIB26	Libya	E-Z5009	15	14	24	19	13	13	9	10	21	12	14	10	20
LIB27	Libya	E-CTS12227	16	14	24	19	13	13	9	10	21	12	14		20
LIB28	Libya	E-PF6789	17	14	24	17	13	13	9	10	21	12	14	10	20
LIB29	Libya	E-PF6789	15	13	24	17	13	13	9		21	12	14	10	20
LIB30	Libya	E-CTS12227	16	14	24	18	13	13	9	10	21	12	14	10	20
LIB31	Libya	E-PF6794*	18	14		18	14	13	9			12	14	10	21
LIB32	Libya	E-PF6789	16	14	23	17	10	13	9			12	14		21
LIB33	Libya	E-Z5009	14	14	23	17	14	13	9	10	21	11	14	10	
LIB34	Libya	E-M183*	15	14	25	18	13	13	9	10	22	12	14	10	20

MOR01	Morocco	E-PF6789	16	13	24	16	13	13	9	10	21	12	14	10	20
MOR02	Morocco	E-Z5009	15	13	24	18	13	13	9	10	21	12	14	10	19
MOR05	Morocco	E-CTS12227	16	14	24	19	13	13	10	11	21	12	14	10	20
MOR06	Morocco	E-Z5009	16	14	24	18	13	13	9	10	22	12	14	10	20
MOR07	Morocco	E-M183*	15	14	24	18	13	13	9	10	22	12	14	10	20
MOR08	Morocco	E-Z5009	16	14	23	18	13	13	9	10	22	12	14	10	20
MOR09	Morocco	E-SM001	16	13	24	18	13	13	9	11	21	12	14	10	20
MOR10	Morocco	E-SM001	16	13	24	18	13	13	9		21	12	14	10	20
MOR11	Morocco	E-Z5009	15	15	24	17	14	13	9	10	22	11	15	10	20
MOR12	Morocco	E-SM001	15	14	24	17	13	13	9	10	22	12	14	10	20
MOR13	Morocco	E-SM001	16	13	23	18	13	13	9	10	21	11	14	10	20
MOR14	Morocco	E-Z5009	16	14	23	19	13	13	9	10	21	12	14	10	20
MOR15	Morocco	E-SM001	15	14	24	19	13	13	9	10	21	12	14	10	19
MOR16	Morocco	E-Z5009	16	13	24	18	13	13	9	10	21	12	14	10	20
MOR17	Morocco	E-Z5009	16	13	24	19	13	14	9	10	23	12	14	10	20
MOR18	Morocco	E-CTS12227	16	14	24	18	13	13	10	11	21	12	14	10	20
MOR20	Morocco	E-CTS12227	15	14	24	18	13	13	9	10	21	12	14	10	20
MOR21	Morocco	E-Z5009	16	14	24	17	13	13	9	10	21	12	14	10	20
MOR22	Morocco	E-Z5009	16	14	24	19	13	13	9	10	23	12	14	10	20
MOR23	Morocco	E-Z5009	16	14	24	17	13	13	9		21	12	14	10	16
MOR24	Morocco	E-Z5009	16	13	24	18	14	13	9	10	21	11	14	10	20
MOR25	Morocco	E-CTS12227	16	15	24	18	13	13	9	10	21	12	14	10	20
MOR26	Morocco	E-Z5009	15	14	24	19	13	13	9	10	21	12	14	10	20
MOR27	Morocco	E-CTS12227	15	14	24	18	13	13	10	11	21	12	14	10	20
MOR28	Morocco	E-Z5009	16	14	24	16	13	13	9	10	22	12	14	10	20
MOR29	Morocco	E-Z5009	15	14	24	18	13	13	9	10	21	13	14	10	20
MOR30	Morocco	E-SM001	16	13	25	17	13	13	9	10	21	11	14	10	20
MOR31	Morocco	E-Z5009	16	14	24	18	13	13	9	10	21	12	14	10	20
MOR32	Morocco	E-Z5009	16	13	24	17	13	13	9	10	21	12	14	10	20
MOR33	Morocco	E-SM001	17	13	24	18	13	13	9	10	20	12	14		20
MOR35	Morocco	E-Z5009	16	14	24	17	13	13	9	10	21	13	14	10	20
MOR36	Morocco	UN[1]		14	24	19	13	13	9		22		14	10	20
MOR37	Morocco	E-Z5009	16	13	24	17	14	13	9	10	21	12	14	10	19
MOR38	Morocco	E-Z5009	16	14	24	17	13	13	9		21	12	14	10	

MOR39	Morocco	E-Z5009	16	13	24	17	13	13	9	10	21	11	14	10	20
MOR40	Morocco	E-Z5009	16	14	24	17	13	13	9	10	21	12	14	10	20
MOR41	Morocco	E-CTS12227	17	14	24	17	13	13	10	11	21	12	14	10	20
MOR42	Morocco	E-Z5009	16	14	24	17	13	13	9	11	21	12	14	10	20
MOR43	Morocco	E-Z5009	17	14	24	18	13	13	9	10	21	12	14	10	20
MOR45	Morocco	E-Z5009	16	14	25	18	13	13	9	10	21	12	14		20
MOR46	Morocco	E-Z5009	16	14	23	18	13	13	9	11	21	12	14	10	20
MOR47	Morocco	E-Z5009	15	14	24	18	14	13	9	10	22	11	14	10	21
MOR48	Morocco	E-CTS12227	16	14	24	18	13	14	9	10	21	12	14	10	20
MOR49	Morocco	E-CTS12227	15	14	24	17	13	13	9	10	21	12	14	10	20
MOR50	Morocco	E-M183*	15	13	24	19	13	13	9	10	21	12	14	10	20
MOR51	Morocco	E-CTS12227	17	14	24	17	13	13	9	10	21	12	14	10	20
MOR52	Morocco	E-CTS12227	16	15	24	17	13	13	9	10	21	11	14	10	20
MOR53	Morocco	E-SM001	16	14	25	18	14	13	9	10	21	12	14	10	20
MOR54	Morocco	E-Z5009	15	13	24	18	13	14	9	11	21	12	14	10	
MOR55	Morocco	E-Z5009	15	14	24	17	13	13	9	10	21	12	14	10	20
MOR56	Morocco	E-Z5009	16	14	23	18	13	13	9	10	21	12	14	10	20
MOR57	Morocco	E-CTS12227	16	14	23	17	13	12	9	10	21	12	14	10	20
MOR58	Morocco	E-Z5009	16	14	24	17	13	13	9	10	21	12	14	10	20
MOR59	Morocco	E-SM001	17	14	24	18		13	9	10	22	12	14	10	
MOR60	Morocco	E-CTS12227	16	14	24	17	13	13	9	10	21	12	13	10	16
MOR62	Morocco	E-Z5009	16	14	25	18	13	13	9	10	21	11	13	10	19
MOR63	Morocco	E-Z5009	16	14	24	19	13	13	9	10	21	13		11	20
MOR64	Morocco	E-SM001	15	14	24	19	13	13	9	10	21	12	14	10	20
MOR65	Morocco	E-Z5009		14	24	18	13	13	9		21		14	10	20
MOR66	Morocco	E-Z5009	16	14	25	17	13	13	9	10	21	12	14	10	20
MOR67	Morocco	E-PF6789	16	14	24	17	13	13	9	10	21	12	14	10	20
MOR68	Morocco	E-SM001	15	14	24	17	13	13	9	10	21	12	14	10	19
MOR69	Morocco	E-SM001	15	14	24	17	14	13	9	10	21	11	14	10	20
MOR70	Morocco	E-Z5009	15	14	24	19	13	13	9	10	22	12	14	10	20
MOR72	Morocco	E-Z5009	16	14	24	18	13	13	9	10	21	12		10	20
MOR73	Morocco	E-Z5009	15	14	24	18	13	13	9	10	21	11	14	10	20
MOR74	Morocco	E-Z5009	15	14	24	17	13	13	9	10	22	11	14	10	20
MOR75	Morocco	E-Z5009	16	14	24	18	13	13	9	10	21	12	14	10	20

MOR76	Morocco	E-Z5009	15	14	24	18	13	13	9	10	21	12	14	10	20
MOR77	Morocco	E-SM001	16	14	24	18	13	13	9	10	21	12	14	10	20
MOR78	Morocco	E-Z5009	15	14	24	19	13	13	9		20	12	14	10	20
MOR79	Morocco	E-M183*	16	14		18	13	13	11	10		13	14	10.2	20
MOR81	Morocco	E-SM001	15	14	24	18	13	13	9	10	21	12	14	10	20
MOR82	Morocco	E-Z5009	16	14		17	13	13	9	10		12	14	9	20
MOR83	Morocco	E-CTS12227	16	14	24	17	13	13	9	10	21	12	14	10	20
MOR84	Morocco	E-PF6789	15	14		17	13	13	9	10		12	14	9	20
MOR85	Morocco	E-CTS12227	16	14		18	13	13	10		20	12	14	12	
MOR86	Morocco	E-CTS12227	15	14	24	18	13	13	11	11	21	12	14	10	20
MOR87	Morocco	E-CTS12227	17	14	24	18	13	13	9	11	21	12	14	10	20
MOR88	Morocco	E-Z5009	14	13	24	19	13	13	9	10	21	11	14	10	20
MOR89	Morocco	E-Z5009	15	13	24	19	13	13	10	11	21	12	14	10	20
MOR90	Morocco	E-Z5009	15	14	24	17	13	13	9	10	21	12	14	10	20
MOR91	Morocco	E-Z5009	15	14		19	13	13	9	10		12	14	10	20
MOR92	Morocco	E-Z5009	16	14	24	18	13	13	9	10	21	12	14	10	20
MOR93	Morocco	E-Z5009	18	14	23	18	13	13	9	10	21	12	14	10	20
JOR03	Middle East	E-SM001	17	14	24	18	13	13	9	10	21	12	14	10	20
PAL02	Middle East	E-SM001	17	14	24	17	13	13	9	10	21	12	14	10	20
JOR01	Middle East	E-PF6789	16	14	24	18	14	13	9	10	21	12	14	10	21
PAL03	Middle East	E-Z5009	17	13	24	18	13	13	9	10	21	12	14	10	20
JOR04	Middle East	E-Z5009	16	13	24	16	13	13	9	10	21	12	14	10	20
JOR02	Middle East	E-SM001	17	14	24	18	13	13	9	10	21	12	14	10	20
TUN01	Tunisia	E-PF6789	15	14	24	17	13	13	9	10	21	11	14	10	20
TUN03	Tunisia	E-CTS12227	16	14	24	18	13	13	10		21	12	14	10	19
TUN04	Tunisia	E-Z5009	15	13	24	18	13	14	9	10	21	12	14	10	20
TUN06	Tunisia	E-Z5009	16	14	24	18	13	13	9	10	22	12	14	10	20
TUN08	Tunisia	E-Z5009	16	14	24	18	14	13	9	10	21	12	14	10	20
TUN09	Tunisia	E-SM001	16	14	24	19	13	13	9	10	21	12	14	10	20
TUN10	Tunisia	E-Z5009	16	14	24	17	13	13	9	10	21	12	14	10	20
TUN11	Tunisia	E-PF6789	15	14	24	17	13	13	9	10	21	12	14	10	20
TUN12	Tunisia	E-Z5009	16	14	25	18	13	13	9	10	21	12	14	10	20
TUN13	Tunisia	E-Z5009	15	14	24	17	13	13	9	11	21	12	14	10	20
TUN14	Tunisia	E-PF6789	16	14	24	17	13	13	9	10	21	12	14	10	21

TUN15	Tunisia	E-Z5009	16	14	24	18	13	13	9	10	22	13	14	10	19
TUN16	Tunisia	E-PF6789		14	24	17	13	13	9		21	13	14	10	20
TUN18	Tunisia	E-SM001	17	14	24	18	15	13	9	10	21	12	14	10	19
TUN19	Tunisia	E-SM001	16	14	24	17	14	13	9	10	21	12	14	10	20
TUN20	Tunisia	E-SM001	16	14	25	20	14	13	9	10	21	11	14	10	20
TUN21	Tunisia	E-Z5009	16	14	24	18	13	13	9	10	22	13	14	10	20
TUN22	Tunisia	E-PF6789	16	14	24	17	13	13	9	10	21	11	14	10	21
TUN23	Tunisia	E-CTS12227	15	14	24	18	13	13	9	11	21	12	14	10	
TUN24	Tunisia	E-SM001	15	14	24	17	13	13	9	10	21	11	14	10	20
TUN25	Tunisia	E-SM001	16	14	24	18	15	13	9	10	21	11	14	10	20
TUN26	Tunisia	E-SM001	16	14	24	17	13	13	9	10	21	11	14	10	20
TUN27	Tunisia	E-SM001	16	14	24	17	13	13	9	10	21	11	14	10	20
TUN28	Tunisia	E-PF6789	16	14	24	17	13	13	9	10	21	12	14	10	20
TUN29	Tunisia	E-Z5009	15	14	24	18	13	13	9			13	14		20
TUN30	Tunisia	E-PF6789	16	14	24	19	13	13	9	10	20	12	14	10	21
TUN31	Tunisia	E-PF6789	17	14	24	17	13	14	9		21	12	15	10	
TUN32	Tunisia	E-CTS12227	16	14	24	19	13	13	10	11	21	12	14	10	20
TUN33	Tunisia	E-Z5009	15	14	24	18	13	14	9	10	21	12	14	10	20
TUN35	Tunisia	E-CTS12227	16	13	24	17	13	13	10	11	21	12	14	10	20
TUN36	Tunisia	E-CTS12227	16	13	24	17	13	13	10	11	21	12	14	10	20
TUN37	Tunisia	E-PF6789	16	14	24	18	13	14	9	10	21	12	14	10	21
SAH05	Western Sahara	E-SM001	16	14		18	13	13	9	10		12	14	9	20
SAH06	Western Sahara	E-SM001	16	14	25	17	13	13	9	11	22	12	14	10	19
SAH07	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH08	Western Sahara	E-SM001	16	14	24	17	13	13	9	10	21	12	14	10	19
SAH09	Western Sahara	E-SM001	16	14	25	18	13	13	9		21	12	14	10	20
SAH10	Western Sahara	E-SM001	16	14	24	19	13	13	9	10	21	12	14	10	20
SAH11	Western Sahara	E-SM001	16	13	24	20	13	13	9	10	21	12	14	10	20
SAH12	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH13	Western Sahara	E-Z5009	16	14	24	18	13	13	9	10	21	11	14	10	20
SAH14	Western Sahara	E-SM001	16	14	23	20	13	13	9	10	21	12	14	10	20
SAH15	Western Sahara	E-SM001	15	14		18	13	13	9	10		12	14	9	19
SAH16	Western Sahara	E-Z5009	15	14		18	13	13	9	10		12	14	9	20
SAH17	Western Sahara	E-Z5009	15	14		18	13	13	9	11		12	14	9	20

SAH18	Western Sahara	E-SM001	16	14	25	17	13	13	9	10	21	12	14	10	19
SAH19	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH20	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH21	Western Sahara	E-Z5009	16	14	24	18	13	13	9	10	21	12		10	19
SAH22	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH23	Western Sahara	E-SM001	16	14	25	18	13	13	9	10	21	12	14	10	20
SAH24	Western Sahara	E-SM001	15	14	25	18	13	13	9	10	23	12	14	10	20

[1]Samples excluded from the downstream analysis

8.2. Supplementary information for section 6.2

Supplementary note. Let *a* be the absolute frequency of haplogroup R1b-M269(xP312) in a sample of *n* Y chromosomes; similarly, let *b*: R1b-P312(xDF27), *c*: R1b-DF27(xZ195), *d*: R1b-Z195(xL176,Z220), *e*: R1b-L176(xM176), *f*:R1b-M167, *g*: R1b-Z220(xM278), *h*: R1b-Z278(xM153), and *i*: R1b-M153. Let *s=a+b+c+...+i*. We have three types of samples with partial information: R1b-M269 without further subtyping (let its frequency be *j*), R1b-P312 (xU152, L21, Z195), but not typed for DF27 (call it *k*), and R1b-Z195(xZ220), not typed for L176 (*l*). *j* individuals may belong to any of the *a*,...,*i* subhaplogroups with probability *a/s*,...,*i/s*; *k* can be R1b-DF27(xZ195) with probability *c/(b+c)*, and R1b-Z195(xZ220,M167) can be either R1b-Z195(xL176,Z220) with probability *d/(d+e)* or R1b-L176(xM167) with probability *e/(d+e)*. Combining these probabilities and turning them into estimated relative frequencies (which we denote with a circumflex over each letter), we have

$$\hat{c} = \frac{c\left(1 + \frac{j}{s} + \frac{k}{b+c}\right)}{n}$$

$$\hat{d} = \frac{d\left(1 + \frac{j}{s} + \frac{l}{d+e}\right)}{n}$$

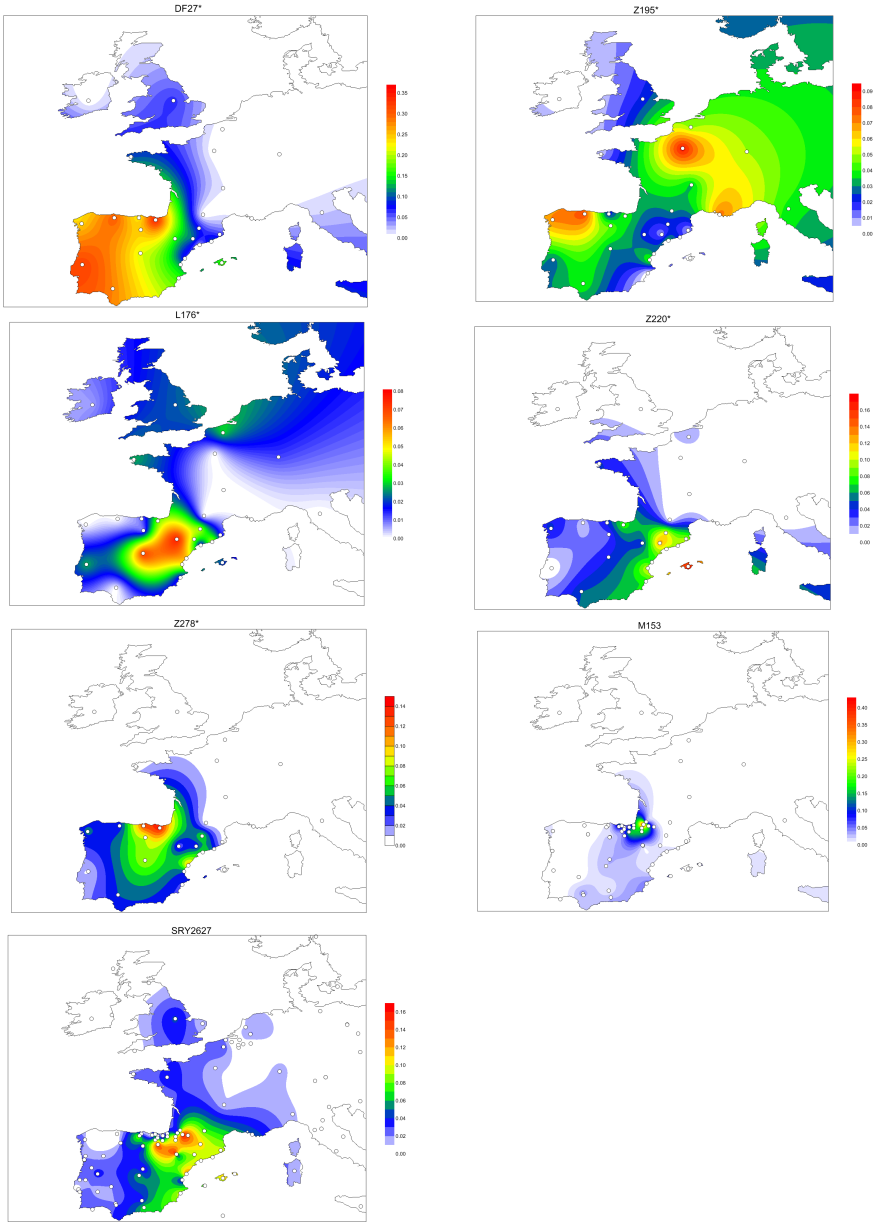
$$\hat{e} = \frac{e\left(1 + \frac{j}{s} + \frac{l}{d+e}\right)}{n}$$

$$\hat{f} = \frac{f\left(1 + \frac{j}{s}\right)}{n}$$

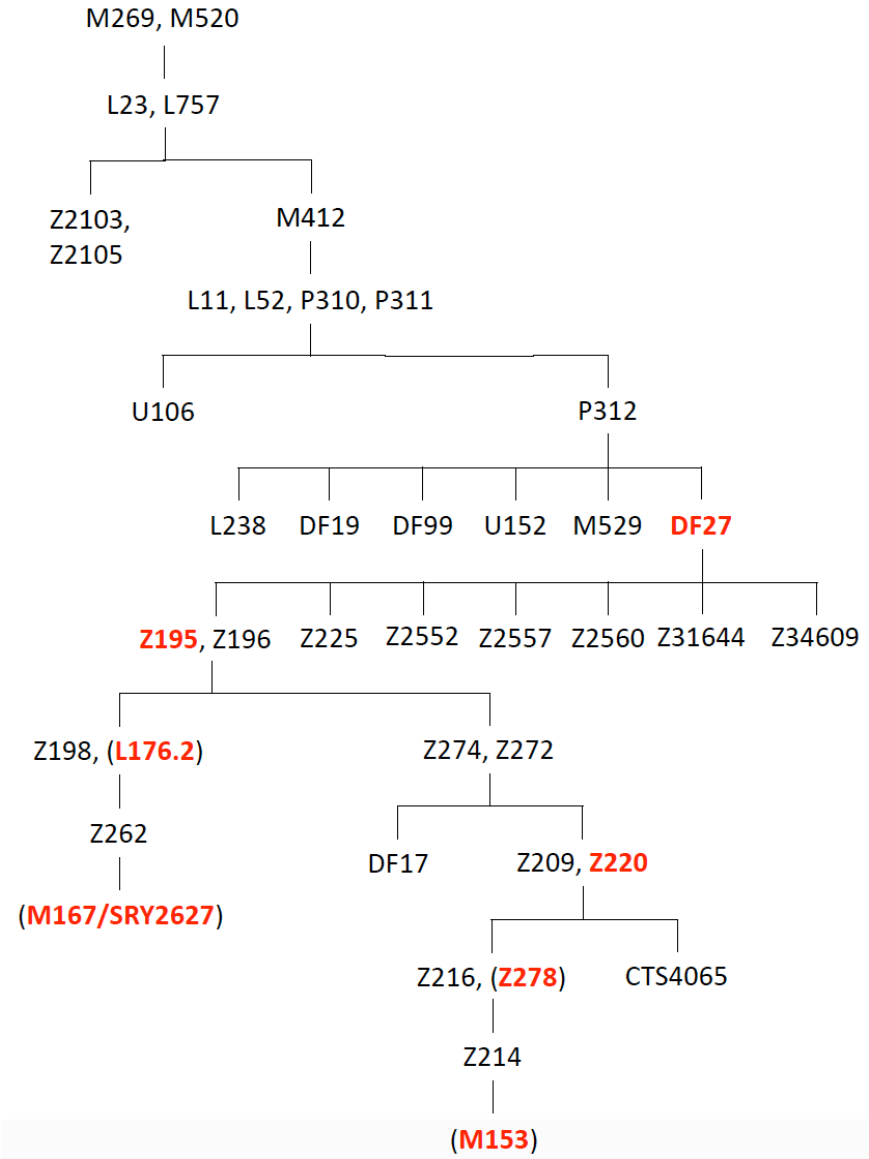
$$\hat{g} = \frac{g\left(1 + \frac{j}{s}\right)}{n}$$

$$\hat{h} = \frac{h\left(1 + \frac{j}{s}\right)}{n}$$

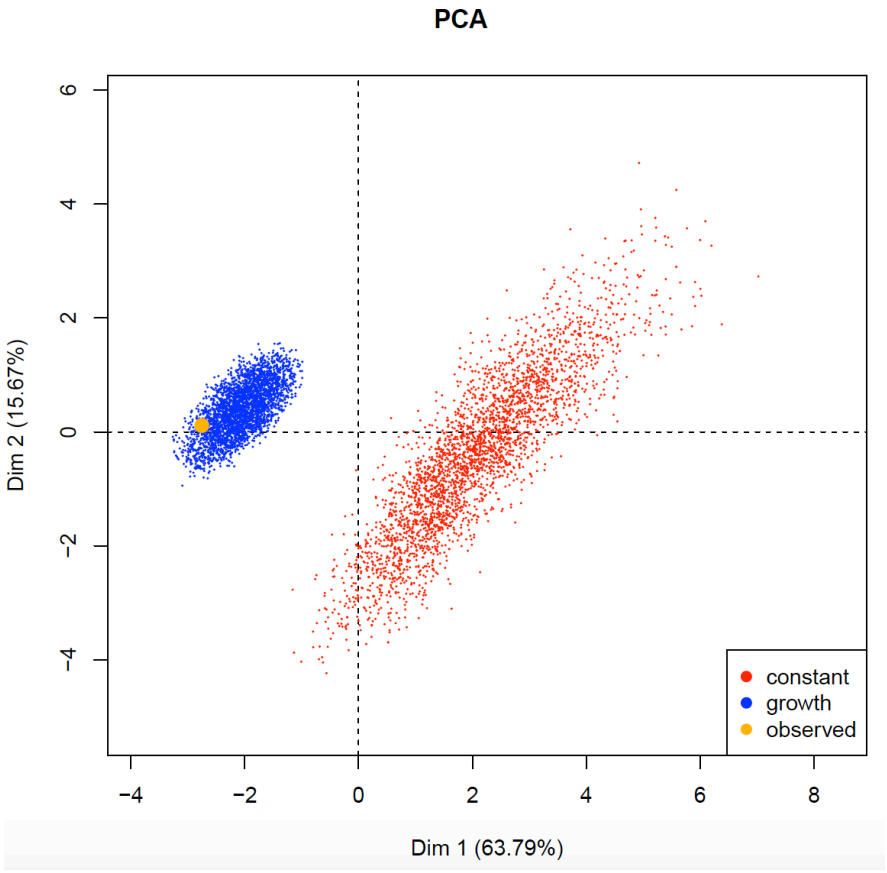
$$\hat{\imath} = \frac{i\left(1 + \frac{j}{s}\right)}{n}$$



Supplementary Figure 1. R1b-DF27 in the context of the Y-SNP tree compiled in ref (32) and available from http://www.phylotree.org/Y/tree/index.htm. In red, SNPs typed in this work. In parentheses, SNPs absent from phylotree-Y. Comas separate different SNPs falling the same phylogenetic branch, while slashes indicate alternate names for the same SNP.



Supplementary Figure 2. Additional frequency contour maps of paragroups, and of SRY2627 and of M153 with additional data from the literature. Maps were drawn with SURFER v. 12 (Golden Software, Golden CO, USA).



Supplementary Figure 3. Principal component analysis of summary statistics in stationary and growth ABC simulations; the observed value falls clearly within the cloud of growth simulations.

Supplementary table S1

population	N	other	R1b-M269	R1b-P312	R1b-DF27	R1b-DF27*	R1b-Z195	R1b-Z195*	R1b-L176.2	R1b-L176.2*	R1b-M167	R1b-Z220	R1b-Z220*	R1b-Z278	R1b-Z278*	R1b-M153
Alacant	142	0.3451	0.6549	0.5845	0.4225	0.1620	0.2606	0.0000	0.1127	0.0141	0.0986	0.1479	0.0704	0.0775	0.0282	0.0493
Alsace	80	0.4125	0.5875	0.3875	0.0750	0.0000	0.0750	0.0500	0.0250	0.0125	0.0125	0.0000	0.0000	0.0000	0.0000	0.0000
Andalucía	100	0.3700	0.6300	0.6000	0.4700	0.2800	0.1900	0.0400	0.0200	0.0000	0.0200	0.1300	0.0500	0.0800	0.0400	0.0400
Aragón	92	0.3370	0.6630	0.6087	0.3696	0.1522	0.2174	0.0217	0.1196	0.0761	0.0435	0.0761	0.0435	0.0326	0.0217	0.0109
Asturias	63	0.4286	0.5714	0.5714	0.4286	0.3016	0.1270	0.0794	0.0000	0.0000	0.0000	0.0476	0.0159	0.0317	0.0317	0.0000
Auvergne	89	0.4719	0.5281	0.4944	0.0562	0.0112	0.0449	0.0337	0.0112	0.0000	0.0112	0.0000	0.0000	0.0000	0.0000	0.0000
Barcelona	571	0.3047	0.6953	0.6162	0.3979	0.0987	0.2992	0.0202	0.1364	0.0444	0.0920	0.1426	0.0957	0.0468	0.0360	0.0108
Brittany	145	0.1310	0.8690	0.8345	0.1931	0.0966	0.0966	0.0069	0.0483	0.0276	0.0207	0.0414	0.0414	0.0000	0.0000	0.0000
Cantabria	96	0.2813	0.7188	0.6250	0.4479	0.2292	0.2188	0.0208	0.0313	0.0104	0.0208	0.1667	0.0417	0.1250	0.1250	0.0000
Castelló	49	0.3265	0.6735	0.6283	0.4717	0.0660	0.4058	0.0292	0.1839	0.0558	0.1282	0.1926	0.0674	0.1252	0.1042	0.0210
Galicia	70	0.3857	0.6143	0.5571	0.4000	0.2429	0.1571	0.0714	0.0000	0.0000	0.0000	0.0857	0.0429	0.0429	0.0429	0.0000
GBR (1000 genomes)	46	0.2609	0.7391	0.5217	0.1522	0.0652	0.0870	0.0217	0.0652	0.0217	0.0435	0.0000	0.0000	0.0000	0.0000	0.0000
Girona	131	0.3969	0.6031	0.5022	0.2874	0.0593	0.2281	0.0081	0.1098	0.0161	0.0937	0.1102	0.0939	0.0163	0.0084	0.0079
Île-de-France	91	0.4396	0.5604	0.4693	0.1026	0.0000	0.1025	0.0900	0.0124	0.0000	0.0121	0.0000	0.0000	0.0000	0.0000	0.0000
Ireland	146	0.1849	0.8151	0.7466	0.0068	0.0000	0.0068	0.0000	0.0068	0.0068	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Lleida	104	0.2788	0.7212	0.6518	0.4146	0.1080	0.3066	0.0102	0.1287	0.0395	0.0892	0.1677	0.1279	0.0398	0.0299	0.0099
Madrid	99	0.3131	0.6869	0.6162	0.4949	0.2121	0.2828	0.0303	0.1010	0.0707	0.0303	0.1515	0.0404	0.1111	0.0808	0.0303
Mallorca	48	0.3125	0.6875	0.6875	0.4587	0.1671	0.2917	0.0000	0.1250	0.0208	0.1042	0.1667	0.1667	0.0000	0.0000	0.0000
Midi-Pyrénées	67	0.4030	0.5970	0.5351	0.1070	0.0000	0.1069	0.0313	0.0756	0.0000	0.0754	0.0000	0.0000	0.0000	0.0000	0.0000
Native Basques	229	0.0786	0.9214	0.9214	0.7380	0.3450	0.3930	0.0393	0.0699	0.0393	0.0306	0.2838	0.0786	0.2052	0.1397	0.0655
Nord-Pas-de-Calais	68	0.3824	0.6176	0.4935	0.1251	0.0000	0.1249	0.0484	0.0616	0.0300	0.0316	0.0149	0.0149	0.0000	0.0000	0.0000
Portugal	109	0.3761	0.6239	0.5046	0.4037	0.3211	0.0826	0.0275	0.0459	0.0275	0.0183	0.0092	0.0000	0.0092	0.0092	0.0000
Provence–Alpes-Côte d'Azur	45	0.4444	0.5556	0.5235	0.1223	0.0000	0.1220	0.0731	0.0486	0.0000	0.0478	0.0000	0.0000	0.0000	0.0000	0.0000
Pyrenees	46	0.3043	0.6957	0.6957	0.4241	0.0763	0.3478	0.0217	0.1522	0.0435	0.1087	0.1739	0.1087	0.0652	0.0652	0.0000
Resident Basques	111	0.3784	0.6216	0.5766	0.4775	0.2432	0.2342	0.0450	0.0721	0.0090	0.0631	0.1171	0.0270	0.0901	0.0811	0.0090
Tarragona	120	0.3583	0.6417	0.5750	0.3459	0.0292	0.3167	0.0364	0.1052	0.0136	0.0917	0.1750	0.0917	0.0833	0.0667	0.0167
TSI (1000 genomes)	53	0.5472	0.4528	0.3774	0.0755	0.0189	0.0566	0.0377	0.0000	0.0000	0.0000	0.0189	0.0000	0.0000	0.0000	0.0000
València	79	0.2911	0.7089	0.6669	0.4076	0.1251	0.2826	0.0438	0.0785	0.0479	0.0306	0.1602	0.1070	0.0532	0.0400	0.0133

Supplementary Table S2

	DF27			Z195			L176.2			M167			Z220			Z278			M153		
	N	Var	sd	N	Var	sd	N	Var	sd	N	Var	sd	N	Var	sd	N	Var	sd	N	Var	sd
All	758	0.330	0.215	510	0.326	0.198	189	0.287	0.213	137	0.245	0.190	267	0.293	0.211	130	0.225	0.141	34	0.146	0.115
Aragón	29	0.372	0.218	18	0.314	0.140	10	0.304	0.225	4	0.239	0.297	7	0.200	0.161	3	0.222	0.272	1		
Basques	154	0.282	0.174	83	0.263	0.179	14	0.300	0.315	7	0.254	0.421	62	0.216	0.209	46	0.177	0.132	15	0.107	0.106
Catalonia	311	0.346	0.218	246	0.343	0.208	108	0.285	0.209	81	0.241	0.172	123	0.319	0.245	41	0.231	0.182	9	0.122	0.151
France	35	0.299	0.207	34	0.302	0.212	15	0.363	0.345	12	0.336	0.330	1								
Mallorca	21	0.341	0.280	14	0.354	0.321	7	0.292	0.325	6	0.311	0.360	7	0.295	0.408						
North Central Spain	105	0.319	0.249	44	0.306	0.262	8	0.248	0.271	6	0.220	0.254	27	0.312	0.325	20	0.279	0.245	1		
València	103	0.349	0.280	71	0.324	0.252	27	0.246	0.233	21	0.192	0.192	40	0.318	0.246	20	0.271	0.180	8	0.188	0.172