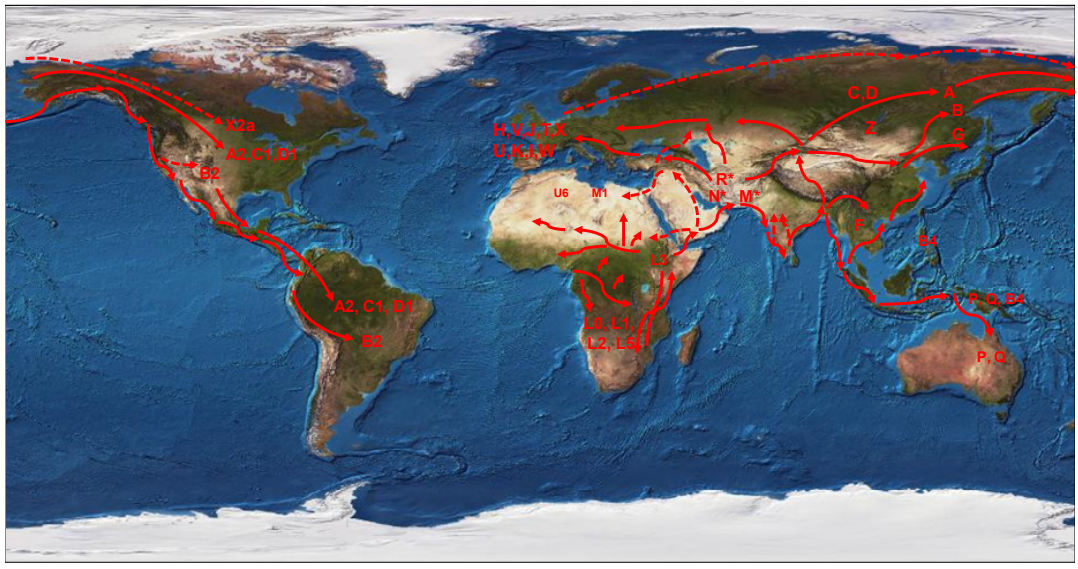
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

FACULTADE DE MEDICINA

INSTITUTO DE CIENCIAS FORENSES “LUIS CONCHEIRO”

HUMAN MITOCHONDRIAL DNA VARIABILITY: MULTIDISCIPLINARY APPLICATIONS IN THE FIELDS OF FORENSIC, MEDICAL AND POPULATION GENETICS



Memoria que presenta para optar al grado de doctor:

María Cerezo Fernández

ISBN 978-84-9887-754-0 (Edición Digital PDF)

El Doctor *D. Ángel Carracedo Álvarez*, Catedrático de Medicina Legal en la Facultad de Medicina de la Universidad de Santiago de Compostela, y el Doctor *D. Antonio Salas Ellacuriaga*, Profesor de la Universidad de Santiago de Compostela,

CERTIFICAN:

Que la presente memoria que lleva por título "***Human mitochondrial DNA variability: Multidisciplinary applications in the fields of forensic, medical and population genetics* "**

de la licenciada en Biología por la Universidad de Santiago de Compostela *María Cerezo Fernández*, ha sido realizada bajo nuestra dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el Tribunal correspondiente.

Y para que así conste a los efectos oportunos, firmamos la presente en Santiago de Compostela, 08 Abril de 2011

Fdo: Prof. Dr. Ángel Carracedo Álvarez

Fdo. Prof.Dr. Antonio Salas Ellacuriaga

María Cerezo Fernández

Este trabajo ha sido financiado en parte, con becas con cargo a los proyectos de investigación (2004/AX268; 2005/AX311), un contrato predoctoral de la Universidad de Santiago de Compostela (convocatoria 2007), contratos con cargo a proyectos (2009/AX615), ayudas para movilidad del CIBERER (convocatoria 2009); los proyectos de investigación de la Xunta de Galicia (PGIDIT06PXIB208079PR y Grupos Emerxentes; 2008/XA122); de la Fundación Mutua Madrileña (2006/ CL370 y 2008/CL444) y del Ministerio de Ciencia e Innovación (SAF2008-02971)

ACKNOWLEDGMENTS

Casi me parece mentira que este momento haya llegado, "sólo" me falta recordar a todos los que habéis estado ahí, compartiendo algún momento de esta etapa que ya se acaba, ayudándome en el esfuerzo que ha requerido llegar a realizar todo este trabajo (perdón de antemano si alguien se me queda en el "tintero").

Creo que me haría falta un "network" para saber cómo poner todo lo que os quiero decir de una manera ordenada, muchos deberíais estar en párrafos compartidos, para mí tenéis "cualidades recurrentes" pero también otras que os hacen únicos. Muchas veces este apartado es utilizado para quedar bien, pero si en este tiempo habéis llegado a conocerme, sabéis que nunca digo cosas que no siento. Y dado que esto no tiene límite de extensión, no lo tengo que hacer en inglés y es lo único que muchos (bueno, más bien la mayoría) os vais a leer…aquí va!!!

Quisiera empezar agradeciéndoos a muchos de los gallegos que aquí aparecéis, he pasado una gran etapa de mi vida en una tierra acostumbrada a ver partir a la gente en vez de recibirla, y debido en gran parte a todo este tiempo con vosotros… ahora muchas veces contesto con otra pregunta; hace tiempo que no siento nostalgia sino que tengo *morriña*; las cosas que antes no me importaban, ahora *"me la traen al pairo"*; desde hace mucho para mí dejó de chispear, ahora *orballa*; cuando algo me hace mucha gracia ya no me rio sino que me *escarallo*; ahora cuando algo me sorprende lo primero que sale de mi boca es un *"¡imimá!!"*; la mayor parte de las veces tengo que decidir los sitios porque vosotr@s siempre respondéis: "*A mí me da igual*", pero lo siento… creo que siempre seguiré *"sin dar hecho"* decir *"no doy hecho"!!*. Por todo esto y por más, GRACIAS por mi enriquecimiento cultural.

Aunque no voy a ser nada innovadora, gracias **Ángel** por la oportunidad. Porque gracias a tí he podido cumplir mi deseo de estar en un laboratorio formando parte de este grupo. Gracias enormes también por asignarme como doctoranda de Toño. Esta tesis nunca podría haber sido así sino hubiese sido por los medios que ambos habéis puesto a mi alcance. No querría dejar de agradecerte tus esfuerzos por fomentar la investigación (a costa de muchas veces no saber ni dónde te levantas y pasar media vida entre aeropuertos, aviones, reuniones, congresos y conferencias), también por ayudar a que muchos biólogos no nos sintamos en tierra de nadie y por tener siempre esa sonrisa para tod@s.

Siguiendo con la especie de plantilla que parece que todo el mundo tiene… gracias sinceras y enormes **Maviki**, por todo tu apoyo y tu cercanía. Gracias, por preocuparte de que esté a gusto. Gracias, por tomarte tantas molestias por mí cuando eso podía suponerte un quebradero de cabeza más que un beneficio, porque siempre te he tenido y te he sentido ahí, eso realmente ha sido muy importante para mí.

Gracias a "mi pequeño gran jefe", **Toño**, por ser tan genial. Porque aunque sinceramente has sido capaz de sacarme de mis casillas (dejémoslo ahí) con tus continuos despistes, también has sabido estimular esa chispa que hace que quiera saber y que quiera hacer más (espero habértelo demostrado), imprescindible para aguantar la maratón que se supone que comencé al entrar al laboratorio. Gracias por haber conseguido que cada día me guste más en lo que trabajo, por permitirme conocer a los "capos" del mitocondrial a nivel mundial, aunque esto se deba a que tú eres uno de ellos. No todo el mundo puede conocer y llegar a trabajar con los mejores en su campo y eso es un lujo que te debo sólo a ti. Sé que en momentos no te lo he puesto nada fácil y lo siento, aunque no te consuele tampoco lo ha sido para mí. Pero sobre todo muchísimas gracias por tantísimas conversaciones y consejos (aunque no siempre los entendí y el tiempo tuvo que ayudarme a hacerlo).

Thanks Chris for your invitation to Cambridge just when I was feeling completely lost in Oxford and an enormous thanks you for your contagious laughter during all these years (always managed to make me smile!). And obviously thanks for each of the funny moments we´ve shared talking about "football".

Thanks to Cristian and the group for your invitation to the laboratory and for those days in Oxford. Thank you for your efforts in making me feel at ease, and particularly for the bike and the guide! Many thanks to George and Sarah also; I wish you luck with your PhDs although I think you are very lucky to have Cristian as your supervisor.

Aunque por haber estado en este grupo, me puedo considerar una privilegiada no puedo decir que haya sido fácil, sobre todo por estar lejos de los míos. Por ello me gustaría agradecer a todos aquellos compañeros de laboratorio que habéis ido apareciendo a lo largo de todos estos años y que tanto me habéis alegrado fuera de el.

Gracias por la ayuda a todos los que estabais cuando llegué; **María B** (por esas sobremesas que ya quedan tan lejanas en el tiempo), **Paula** (por tenerme en cuenta siempre que viajas hacia la meseta), **Vane**, **Meli** (por ser la que domina todo y no perder la paciencia siempre que se acaba algo); **Alex** y **Alejandro B.** (porque con vosotros fue divertido hacer de McGiver y por las largas charlas), **Fonde** (por enseñarme "PhDcomics", con ellos aprendí a reírme de esta etapa"), **Raquel** (por tantos buenos momentos, esas cañas y esos partidos), **Eva** (por alegrarnos a todos tantos días con "movidas" y "paranoias" varias), **Rosa** (por no perder nunca la paciencia con mis preguntas para la fotocopiadora y el fax) y **Jose Manuel** (por tener siempre una sonrisa y ayudarme a que el papeleo se quede en (casi)nada).

Gracias también a todos los que llegasteis después por todo lo que nos habéis ido aportando: **Ana F** (también por las sobremesas con sus charlas y la ayuda con el secuenciador las últimas semanas), **Ana M.**, **Yarimar** (mi compi de puntas), **Carla** (por mi "sprint" final con

el 34-plex y su análisis), **Liliana** (por ampliar mi vocabulario con palabras como "chocolatinosa"), **Miguel; Ana IV, Pedro, Paula, Esther, Jens, Danel, Francesca S, Ana Paz**…. Y a todos los que ya os fuisteis: **Gloria, María M., Ricardo, Ángela** …

Puede que sea una coincidencia que os quiera agradecer tanto a los que empezasteis a la vez…

Gracias infinitas por todos los buenos momentos dentro y fuera del laboratorio en Santiago, gracias enormes por todos los momentos geniales e inolvidables de mi estancia en Oxford, sin ti no habría sido ni parecida junto con todos los "spaniards" que conocí gracias a ti. (Gracias a **Juan, Estrella, Israel, Jose, Ana y Vito** por abrirme los ojos a lo que son los "post-docs" y por hacerme tan agradable la experiencia de irme de "pre-doc"). Eres la responsable de que ya no quiera que siempre pierda Italia y de que a veces entienda mejor el "itañolo" que el inglés, gracias por todo, **Franchi**.

Gracias enormes, **Cata**, por ser capaz de contagiar la alegría allá por donde vas, gracias por tirar de mí para ir al gimnasio lo justo para que acabase siendo yo la que tiró de ti y de **Noa** (gracias por esos buenos momentos de las tres).

Manu, no sé cómo agradecerte todos los momentos de risa que has intentado que tuviese desde que empezamos. Gracias por intentar integrarte en ese grupo mayoritario de biólogos, haciéndonos "coñas" con los reactivos que utilizamos aunque no supieses ni para qué servían. Gracias por estar ahí cuando te necesité (aunque no siempre lo supiste y eso lo hace aún mejor) gracias a mi "tercer hermano".

Aparecisteis en el momento justo, gracias por preocuparos de verme siempre animada y con una sonrisa (aunque fuese a costa de ponerme roja), pero siempre tratándome tan bien, **Luis** (por todas esas horas de música gracias a spotify) y **Alberto** (compañero fatigas con los genomas completos del ADNmt). Pero sobretodo quiero daros las gracias a ambos por intentar que aprenda a tomarme las cosas menos en serio (aunque ya sabéis que aún estoy en ello!!).

Siendo de las últimas en llegar sois muy responsables de ayudarme a llevar los momentos de la cuesta arriba que ha supuesto la última parte de la maratón. Gracias por las risas, las confidencias, los apoyos y los momentos de desconexión, **Ana P., Olalla, Montse** (por acercarme un poco más a mi casa cada vez que me soltabas alguna expresión más castellana) y **Rocío** (por ser siempre tan cariñosa conmigo, llenarme de piropos y hacerme sentir siempre mejor después de verte).

Sinceramente… me has vuelto loca casi desde que llegaste, pero como se trata de agradecer… gracias por todo! Por tantas comidas y sobremesas juntas, por quedarte tantos días hasta "las tantas" conmigo en el depar, por todo lo que me has hecho reír, por estar siempre ahí … pero sobretodo, gracias **Dosil** por considerarme tu amiga.

Gracias además a toda la gente que pasa por el departamento, unas semanas, unos meses o unos años… Por todo lo que nos enseñáis de tantos sitios diferentes, dando un gran colorido al laboratorio. De manera especial quiero agradecer a todos aquellos que aportaron unas muestras que han servido como material imprescindible de esta tesis.

Gracias a los que también considero mis compañeros, pero en el hospital. Gracias a **"Pancho"** y **"Luli"** por haberme permitido trabajar en la Fundación como si siempre hubiese estado allí. A **Ana Vega**, **Clara** y **Celsa** por tener siempre ratillos para hablar conmigo siempre que he ido. A **Jorge**, por las charlas y en especial por acordarte de mí siempre que vas a ir a Burgos. A toda la parte de CEGEN, en especial a **María Torres, Rosana, y Juan** por todo el trabajo con Sequenom.

Gracias chicos por integrarme en el "grupo del Cola Cao", rebautizado por Pablo en "cuchipandi del Nesquik" (aunque yo siempre tomase café) cuando empecé a secuenciar "a granel", gracias por esos inicios y por todo lo que ha venido después a **Ana B., Pablo**, **Alejandro G.** y **Ceres** (mi casi tocaya y compañera de experiencias en Oxford, por ser siempre tan positiva y alegre lo cual hace que siempre me apetezca un rato contigo y con **Juan**, sois tal para cual, gracias a ambos por todo el apoyo, las risas, desconexiones y buenos momentos). Pronto ese grupo creció y pasó a ser el "grupo de las tortillas" gracias enormes a **Laura** (aunque me quieras disputar el título de horas en el laboratorio), **William** (Walter-Wallace), **Marta C., Marta S., Andrés, Lucia, Susana, "Marigli", Marcos, Ramón**… por todo lo que me habéis hecho reír (aunque no vaya a ser vuestro "case-report").

Gracias a mis "Doctoras Nurias" Nunca pensé que os iba a echar tanto de menos, me habéis faltado cuando más me habrían ayudado esos ánimos y esas risas, pero no me olvido de los momentos a mitad de camino que tanto han hecho que os recordase. A **Nuria Naverán**, por ser la principal responsable de que empezase a "engancharme" a mil y un *TV shows*, pero sobretodo porque contigo los días siempre tenían momentos de alegría. Gracias por ser tan positiva y contagiosa, si hubiese más gente como tú, el mundo (no sólo el científico) sería maravilloso. A **Nuria Gómez**, porque me convenciste para hablar con Ángel y sin eso tampoco habría llegado hasta aquí. Gracias por que junto con **Ana Blanco** (y después **Ceres y Laura**), estuvisteis dispuesta a mandarme una por una las secuencias, cargarme las placas, mandarme los datos, y resolver mis dudas. Gracias por ser una gran amiga y por todos esos buenos momentos juntas en el piso que quedarán para nuestro recuerdo.

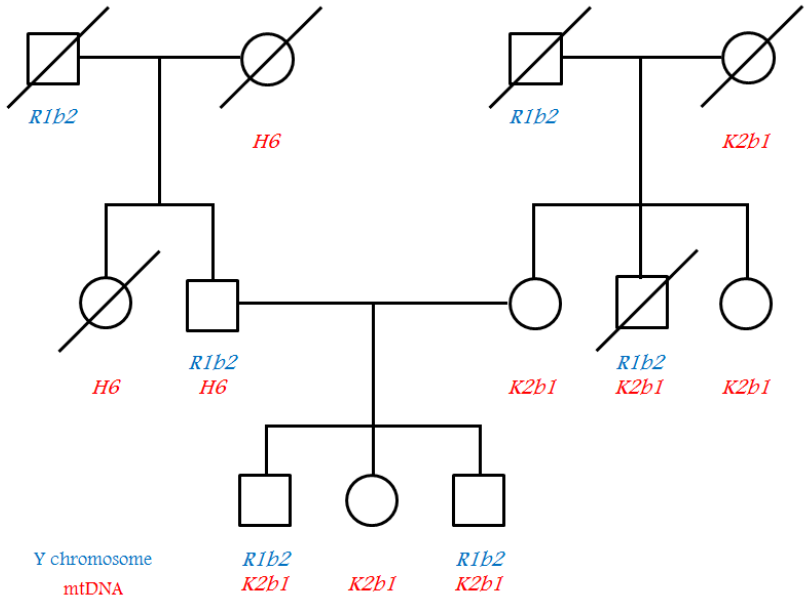
Gracias enormes a mis amig@s de siempre por todos estos años, por los ánimos y los buenos momentos, por perdonarme que haya estado desconectada de todo los últimos meses y seguir contando conmigo, gracias por no tener ni idea de genética y mucho menos del ADN mitocondrial, así con vosotros la desconexión siempre fue total

Y ya en el plano "mucho" más personal quiero agradecer a **mi familia**, en especial a **mis padres**, verdaderos responsables de que haya podido llegar hasta aquí. Como bien dijo Ángel:"La investigación es esfuerzo y dedicación que siempre repercute en tiempo y sacrificio para los más queridos".

Es fácil apoyar a alguien cuando hace lo que tú quieres, sé que hubieseis preferido tenerme más cerca y por eso tiene aún más valor vuestro apoyo junto con el esfuerzo y los sacrificios que han supuesto que llegase hasta aquí. Gracias por ser, junto con mis hermanos, **Leonardo** y **Juanjo,** la constante en mi vida, mi punto de referencia, independientemente del sitio en el que esté viviendo. También gracias a los que ya no están, por todo su apoyo hasta que se fueron, **mi abuela** (la única que pude conocer) y más reciente a mi **tío Paco** (lo más cercano a un abuelo que pude tener). A todos vosotros, gracias por los consejos dados cuando lo necesitaba, porque siempre me habéis hecho sentir mejor con vuestros ánimos, en fin, gracias por estar siempre ahí.

Finalmente, gracias **Gonzalo**, por aguantar esta distancia, sólo porque es lo que quiero hacer. Gracias por conseguir que los kilómetros desapareciesen cada noche por teléfono, y más aún cuando el día no era el mejor. En fin gracias porque no sabes lo importante que ha sido poder contar contigo a pesar del sacrificio que ha supuesto el estar separados, por quererme y apoyarme, gracias.

TO MY FAMILY



A MIS PADRES

A MIS HERMANOS

A GONZALO

Let us suppose the members of a tribe; practising some form of marriage, to spread over an unoccupied continent, they would soon split up into distinct hordes, separated from each other by various barriers, and still more effectually by the incessant wars between all barbarous nations. The hordes would thus be exposed to slightly different conditions and habits of life, and would sooner or later come to differ in some small degree. As soon as this occurred, each isolated tribe would form for itself a slightly different standard of beauty and then unconscious selection would come into action through the more powerful and leading men preferring certain women to others. Thus the differences between the tribes, at first very slight, would gradually and inevitably be more or less increased.

CHARLES DARWIN, *THE DESCENT OF MAN*

CONTENTS

ABBREVIATIONS

AB	Applied Biosystems
BER	Base Excision Repair
Bp	Base pair
BSA	Bovine Serum Albumine
DHPLC	Denaturing High-Performance Liquid Chromatography
DLB	Domain Lysis Buffer
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DTT	dithiothreitol
EDTA	Ethylene Diamine Tetraacetic Acid
HG/hg	Haplogroup
Hg	mercury
HVS	Hipervariable segment
k.a.	Kilo-annum or 1000 years
kya	Kilo-years ago
LGM	Last Glacial Maximum
MALDI	Matrix-Assisted Laser Desorption/Ionization
MJ	Median Joining (networks)
ML	Maximum likehood
MP	Maximum parsimony
mTERF	mitochondrial transcription termination factor
MRCA	Most recent common ancestor
MRO	Multiregional Origin model
MS	Mass Spectrometry
Mya	million years ago
nDNA	nuclear DNA

NJ	Neighbour-joining
nps	nucleotide position
nt	nucleotide
OOA	Out of Africa
OXPHOS	Oxidative Phosphorylation
PCR	Polymerase Chain Reaction
QSN	Quasi-median spanning (networks)
RFLPs	Restriction Fragment Length Polymorphisms
RM	Reduced median (network)
rpm.	Revolutions per minute
SBE	Single Base Extension
SCCP	Single Strand Conformation Polymorphism
SWGDAM	Scientific Working Group on DNA Analysis Methods
Taq	Termus aquaticus
TMRCA	Time to the most recent common ancestor
TOF	Time of flight
UPGMA	Unweighted Paired Group Method with Arithmetic mean
y.b.p.	years before present

LIST OF FIGURES

All the figures were carried out by the author of this document unless otherwise indicated

LIST OF TABLES

I BACKGROUND

I.1. MITOCHONDRIAL DNA

I.1.1 Localization, organization and function of mtDNA

Mitochondrial DNA (mtDNA) is an extra-nuclear genome which is inside the matrix of the mitochondrion, a double membrane-enclosed organelle found in most eukaryotic cells. Mitochondria have several features that resemble those of prokaryotes: somehow the mtDNA functionally behaves as the DNA of a prokaryote, the mtDNA is circular, the mitochondria divides independently of the cell through binary fission, the method of cell division is the most typical in prokaryotes. Nowadays, the most accepted hypothesis about the origin of this organelle in the cell is the endosymbiotic hypothesis formulated by Lynn Margulis (Sagan 1967). The number of mitochondria in a cell varies widely by organism and tissue type and depends on the cell's function. The chief function of the mitochondria is to create energy (ATP) for cellular activity by the process of aerobic respiration. Other functions are the regulation of the membrane potential, the cellular metabolism or the apoptosis programmed cell death.

Mitochondrial DNA is not naked; it is packaged into a nucleoid structure with a group of protein factors and associated with the mitochondrial inner membrane (Holt et al. 2007). For some authors, the nucleoid may serve as a mitochondrial genetic unit as it has been shown that there is very little exchange between nucleoids (Gilkerson et al. 2008). The average number of mtDNA molecules per nucleoid is around six in cultured human cells (Iborra et al. 2004; Legros et al. 2004), but can vary from 2 to 10 mtDNA copies per nucleoid (Legros et al. 2004; Malka et al. 2006; Satoh and Kuroiwa 1991).

The mitochondrial genetic code differs from the universal genetic code in several aspects. In human mtDNA, UGA codes for tryptophan and it is not a termination codon; AUG codes for methionine and it does not code for isoleucine; AGA and AGG are termination codons, not arginine codons; and AUA and AUG are initiation codons not isoleucine codon (the first one) neither methionine codon (the last). The molecule shows asymmetric distribution of functions in both strands and nucleotide compositions; one of the strands is rich in purine nucleotides and for this reason is called heavy-strand (*H-strand*), the other strand is rich in pyrimidine nucleotides and is called light strand (*L-strand*).

In 1981, a research group from Cambridge sequenced the first human mtDNA (Anderson et al. 1981); since them, the human mtDNA is numbered with reference to the L-strand of this *Cambridge reference sequence (CRS)*.

In 1999, the resequencing of the placental mtDNA used by the Cambridge group (Andrews et al. 1999) detected a number of errors (previously called discrepancies (Howell et al. 1992)). The errors were presumable due to sequencing artefacts or because the presence of HeLa and bovine mtDNA in the sample. This study confirmed ten substitution errors and only one cytosine residue between positions 3106 and 3107 (the previous result showed two). In order to prevent the confusion with previous studies the authors suggested to keep the original light strand numbering. This revised sequence was baptized as the *revised Cambridge reference sequence (rCRS).*

Mitochondrial DNA is a circle double stranded DNA of 16569 bp which encodes 37 genes, 13 code polypeptides, 2 rRNAs and 22 tRNAs (see *Figure 1*). *NADH1–NADH6*, and *NADH4L* are seven subunits of complex I (NADH–ubiquinone oxidoreductase). *Cyt b* is the only mtDNA encoded which is a part of complex III subunit (ubiquinolcytochrome *c* oxidase reductase). COXI–COXIII consist of three catalytic subunits for the complex IV (cytochrome *c* oxidase, or COX), and the *ATP 6* and *ATP 8* genes encode for two subunits of complex V (ATP synthase). These genes are asymmetrically distributed in the mtDNA molecule. The H-strand encodes most of the information: the two rRNAs, 14 tRNAs and the mRNAs for 12 out of the 13 polypeptides. The L-strand encodes the remaining eight tRNAs and only one mRNA, corresponding to ND6 subunit (Anderson et al. 1981; Attardi and Schatz 1988).

Although nuclear genes code for the majority of mitochondrial respiratory chain, a disruption of either genes can cause a mitochondrial dysfunction. The tRNAs and rRNAs are required to synthesize the proteins which are then incorporated into four of the five multiprotein enzyme complexes of the mitochondrial oxidative phosphorylation (OXPHOS) system. Carbohydrates and fatty acids are imported into the mitochondria where are converted by ß-oxidation and the tricarboxylic acid (TCA) cycle into the molecules used for the OXPHOS system to generate ATP. The OXPHOS system receives electrons from these molecules and transfers them through the complexes, with the aid of two electron shuttle molecules Coenzyme Q (CoQ) and Cytochrome c (Cyt c), to Complex IV where they are accepted by oxygen. This process generates free energy which is used to pump protons (H+) from the mitochondrial matrix to the intermembrane space, creating an electrochemical gradient across the inner mitochondrial membrane, which is used by Complex V to drive the formation of ATP.

One of the characteristics of the mtDNA molecule is its compact gene organization, with all the coding sequences contiguous to each other or separated by a few bases and without introns (*see Figure 1*). Some of the protein genes overlap, namely, 46 nucleotides between ATPase 6 and 8, and 7 nucleotides between ND4 and ND4L subunits.

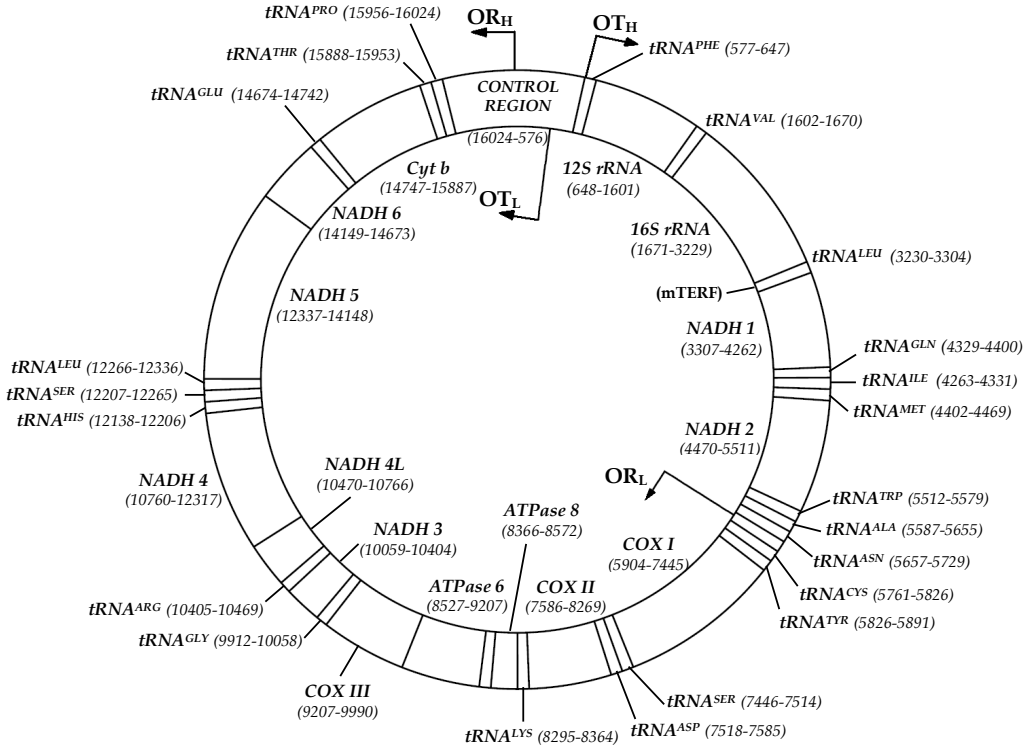


Figure 1. Schematic representation of mitochondrial genome. The position and the length of the different genes and points of origin of replication and transcription are indicated. *OR* are the origins of heavy-strand (OR_H) and light-strand (OR_L) replication. *OT* are the origins of heavy-strand (OT_H), light-strand (OT_L) transcription. *mTERF* is the mitochondrial transcription termination factor. (Positions according to New NCBI Reference Sequence NC_012920.1; http://tinyurl.com/yhzurcf)

The mtDNA displacement loop (D-loop, also called control region) is an ~1122 bp non-coding region (carries on from position 16024 to position 576) which is involved in the regulation of replication and transcription of the molecule (*see Figure 2)*. The D-loop contains three regions which have a highly variable sequence at the population level, popularly called *hypervariable sequences (HVS) HVS-I, HVS-II and HVS-III.*

The precise length of these fragments varies between studies (*see Figure 2*). For example, in the forensic field, the region known as *HVI* ranges from position 16024 to

position 16365, *HVS-II* from 73 to 340, and HVS-III from 438 to 576. In a number of population genetics studies, *HVI* is considered from 16024 to 16400, *HVS-II* from 44 to 340 and *HVS-III* from 438 to 576. The function of these regions is not known but can be important for molecule replication and transcription, because they contain or are near the origin of heavy and light strands.

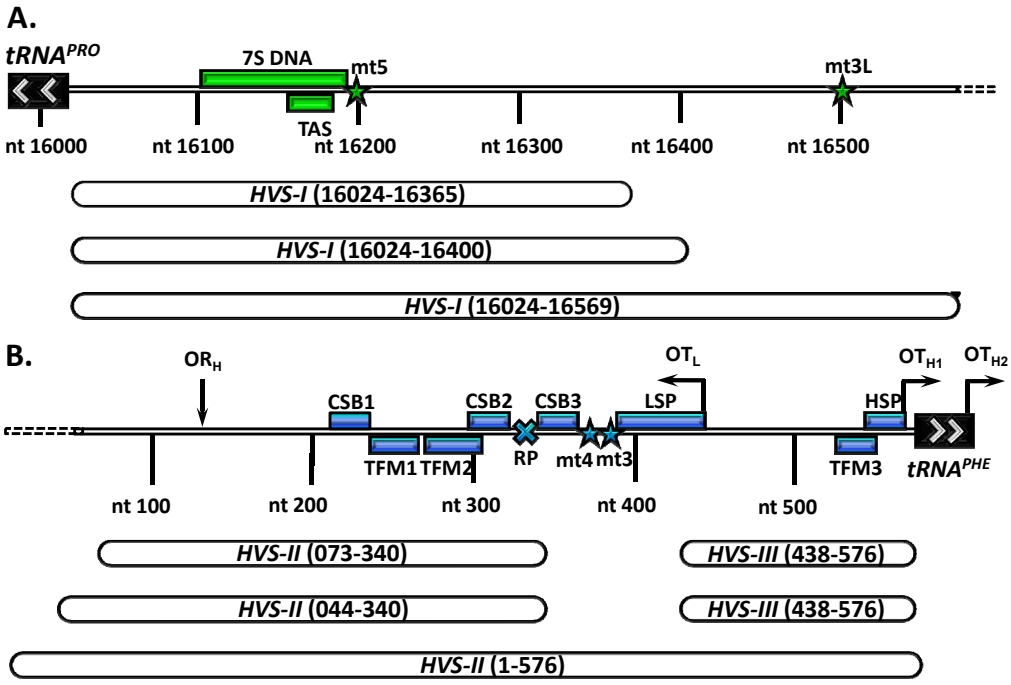


Figure 2 Mitochondrial DNA control region. (A) The first part of the D-loop from 16024 position to 16569, contains according with different works *7S DNA* (Anderson et al. 1981; Brown et al. 1978; Crews et al. 1979; Gillum and Clayton 1978, 1979); *TAS* (termination-associated sequence) (Doda et al. 1981; Roberti et al. 1998); *mt5* (control element) (Ohno et al. 1991) and *mt3L* (L-strand control element)(Suzuki et al. 1991). This first segment also contains the HVI (the different fragments correspond to the length in forensic, population genetics studies and the present study) (B) The second part of the D-loop, which extends from position 1 to position 576, contains according to different studies the elements *CBS1*, *CBS2* and *CBS3* (conserved sequence blocks) (Chang and Clayton 1985); *TFM1*, *TFM2* and *TFM3* (mtDNA transcription factors) (Fisher et al. 1987);*RP* (replication primer) (Chang and Clayton 1987a, b); *mt4* and *mt3* (H-strand control elements) (Suzuki et al. 1991); *LSP* (light-strand promoter)(Clayton 1984; Chang and Clayton 1984; Montoya et al. 1982; Walberg and Clayton 1983); *HSP* (heavy-strand promoter)(Bogenhagen et al. 1984; Chang and Clayton 1984; Hixson et al. 1986; Montoya et al. 1982; Yoza and Bogenhagen 1984); OR_H (origin of heavy-strand replication) OT_L (origin of light-strand transcription) OT_{H1} (origin of mayor heavy-strand transcription) and OT_{H2} (origin of minor heavy-strand transcription). This second part of the D-loop also contains hypervariable sequences (the blocks represent the length of the hypervariable sequences according to forensic, population genetics studies and present work).

There are other non coding short segments and single base pairs along the molecule:

- The nucleotides 3305 and 3306 between *tRNA^{LEU}* and *NADH1* genes.
- The single base 4401 between *tRNA^{GLN}* and *tRNA^{MET}* genes.
- The segment from 5580 position to 5586 between *tRNA^{TRP}* and *tRNA^{ALA}* genes.
- The single base 5656 between *tRNA^{ALA}* and *tRNA^{ASN}* genes.
- The segment from 5730 position to 5760 between *tRNA^{ASN}* and *tRNA^{CYS}* genes, which contains the origin of replication of L-strand (OR_L)
- The short fragment from 5892 position to 5903 between *tRNA^{TYR}* and *COX I* genes.
- The nucleotides 7515, 7516 and 7517 between *tRNA^{SER}* and *tRNA^{ASP}* genes.
- The segment from 8270 position to 8294 between *COX II* and *tRNA^{LYS}* genes.
- The single base 8365 between *tRNA^{LYS}* and *ATPase 8* genes.
- The segment from 14743 position to 14746 between *tRNA^{GLU}* and *Cytb* genes.

I.1.2 Transcription and replication of mtDNA

The current model of mtDNA transcription describes two origins for the H-strand, the first one within the control region (OT_{H1}) and a second (OT_{H2}) close to the 5' end of the 12S rRNA gene (*see Figure 1 and Figure 2*). Transcription from OT_{H1} ends at the 3' end of the 16S rRNA gene, and result in the synthesis of the two rRNAs (12S and 16 S) and two tRNAs ($tRNA^{PHE}$ and $tRNA^{VAL}$). This transcription unit operates with a frequency 20 times greater than the second which begins at IT_{H2} (Fernandez-Silva et al. 2003) which results in a polycistronic mRNA molecule for the other 12 proteins and 12 tRNAs encoded for the H-strand. The activity of the first unit is linked to a transcription termination factor (mTERF) taking place immediately downstream from 16S rRNA, inside the gene for $tRNA^{LEU}$ (*see Figure 1)*

The L-strand transcription gives rise to a single polycistron starting in the control region (OT_L) at the 5´end of 7S RNA, about 150 bp away from the H1 initiation point, from which the eight tRNAs and the ND6 mRNA are derived. The precise locations and mechanisms for L-strand and second H-strand transcription termination are unknown. Only seem to be some relation between mutations in CBS II and shorter prematurely terminated transcripts (Asari et al. 2007; Pham et al. 2006).

OT_{H1} and OT_L are situated within regions highly conserved; consisting of a promoter element and the binding site for mitochondrial transcription factor A (the mitochondrial RNA polymerase requires TFAM and one of the two other mitochondrial transcription factors B1 o B2 to initiate transcription), however OT_{H2} does not have the same promoter features (Gaspari et al. 2004; Taylor and Turnbull 2005).

Mitochondrial DNA is replicated by DNA polymerase γ, a heterodimeric enzyme (with proof-reading activity) which is mitochondria-specific, but also are involved other proteins like Twinkle (which has 5'-3' helicase activity) (Taylor and Turnbull 2005). The replication takes place in the mitochondrial matrix, independently from cell cycle phase or nDNA replication(Bogenhagen and Clayton 1977).

The generally accepted model, describes replication as an asynchronous displacement mechanism involving two unidirectional, independent origins. The synthesis starts at OR_H, (located downstream of the LSP in the D-loop region), and proceeds along the parental L-strand to produce a daughter H-strand circle. When H-strand replication reaches OR_L, the parental H-strand is displaced, the initiation site for L-strand synthesis is exposed and its replication starts and proceeds in the opposite direction producing a daughter L-strand. A second model of replication was proposed for the results with two-dimensional gel electrophoresis, supporting a bidirectional strand-couple mechanism (Holt et al. 2000; Yang et al. 2002). Both modes of replication have been discussed (Bogenhagen and Clayton 2003a, b; Holt and Jacobs 2003)and the description of another D-loop origin of replication (Fish et al. 2004) would add more debate(Holt 2009).

I.1.3 mtDNA characteristics

Mitochondrial DNA shows several features which make it suitable for forensic, evolutionary and clinical studies. These features include high copy number, maternal inheritance, lack of recombination and higher average mutation rate than found in nDNA (nuclear DNA).

I.1.3.1 High copy number, heteroplasmy and threshold effect

Mitochondrial DNA is present in multiple copies per cell, from hundreds to thousands depending on the cell (Bogenhagen and Clayton 1974; Cree et al. 2008; Legros et al. 2004; Piko and Taylor 1987). This characteristic is advantageous for forensic studies or even for ancient DNA studies, where the material recovered is often degraded or in low amounts. Moreover, the change of mtDNA copy number can result in a disease; mtDNA depletion (decrease of mtDNA copy number) has been associated with renal cell carcinoma (Xing et al. 2008), liver disease (Morten et al. 2007) cardiomyopathy (Bai and Wong 2005), and breast cancer (Yu et al. 2007); disease caused by excess mtDNA proliferation is less common, although it was reported as a compensatory effect of mtDNA deletions (Bai and Wong 2005)

For this polyploidy aspect, different levels of *heteroplasmy* (when there is an admixture of two or more mtDNA types) can be present in an individual. There are different types of heteroplasmy in an individual, within tissue, within cells of the same tissue, within mitochondria inside a cell or within a single mitochondrion. The opposite condition is referred to as *homoplasmy* (when all the copies of the mitochondrial genome are identical). When one of the mitochondrial types has a mutation which is responsible for a disease, the effect of this mutation will depend on the level of this mutant versus wild-type mtDNA, this is the *threshold effect.*

Heteroplasmy is an unstable (mutational) condition in individuals and therefore a feature of little interest in populations. It is however important to note that every fixed mutation/polymorphism observed in nature passes through an initial heteroplasmic state. The transition from heteroplasmy to homoplasmy can be driven by a reduction of mtDNA molecules, called *mitochondrial bottleneck,* which is produced during the embryogenesis (Cree et al. 2008) or postnatal folliculogenesis (Wai et al. 2008).

This decreasing in mtDNA content is followed by a huge increase in oogenesis, which can lead to strong founder effects (Bergstrom and Pritchard 1998).

Last year, a study focussed in heteroplasmic mutations in several individuals as well as different tissues was published (He et al. 2010). They proposed three potential ways to explain the heteroplasmic variants. The first of these is paternal: heteroplasmic variants might represent mtDNA inherited from the father. The second is maternal:

heteroplasmic variants might be inherited from the mother, with bottlenecks occurring during embryonic development resulting in tissue-specific variations. The third is de novo generation: heteroplasmic variants might represent new mutations that occurred during embryonic development. Finally they discarded the first option, the second was confirmed in one pedigree and the third was the most frequent (He et al. 2010).

As in clinical genetics, heteroplasmy is important in forensic casework, since different tissues of the same individuals could contain different heteroplasmic levels for a given mutation.

I.1.3.2 Maternal inheritance

From traditional view, mtDNA is maternally inherited (Giles et al. 1980), meaning that it is passed from a mother to all of her children, but only her daughters will pass it on to subsequent generations. The first direct evidence for uniparental, and maternal, mitochondrial inheritance was available in *Xenopus* in 1972 (Dawid and Blackler 1972) and since that time the concept of strict maternal inheritance has been supported from numerous studies (Giles et al. 1980; Gyllensten et al. 1985; Hayashi et al. 1978; Hutchison et al. 1974; Reilly and Thomas 1980).

Although paternal inheritance of mtDNA seems to occur in several organisms like mussels (Zouros et al. 1992), mice (Gyllensten et al. 1991; Kaneda et al. 1995) *Drosophila* (Kondo et al. 1990), the maternal inheritance of human mtDNA was regarded "as an unshakable dogma of the field"(Stoneking and Soodyall 1996; Wallace et al. 1999).

There are different explanations that would explain matrilineal inheritance of the mtDNA genome; for example relatively low proportion of paternal mtDNA to maternal DNA (10^3-10^4 times less) makes the first one prone to loss by drift, amplified by the *mitochondrial bottleneck*. Moreover, sperm mitochondria are selectively destroyed in the oocyte (Manfredi et al. 1997) and paternal mtDNA is marked by ubiquitination and eliminated by proteolytic digestion (Sutovsky et al. 1999, 2000; Thompson et al. 2003).

However, over the last 25 years, the discussion has been opened to a hot debate (Bromham et al. 2003a; White et al. 2008). The hypothesis of some authors is around the detection limits of paternal leakage contribution (Wolff and Gemmell 2008). Others recognize that is a very rare phenomenon but it is possible, however, that the mechanism responsible for the elimination of paternal mtDNA may fail occasionally, potentially

leading to maternal/paternal mtDNA mosaicism in an individual. An example is the case report (Schwartz and Vissing 2002), where a 28 years old man with a metabolic disorder that had resulted in a severe exercise intolerance whose muscular mtDNA was of predominantly paternal (Pakendorf and Stoneking 2005), and this mtDNA is called *Mitochondrial Steve* (Bromham et al. 2003b; Slate and Gemmell 2004).

I.1.3.3 Lack of recombination

Some authors suggested that transmission of paternal mtDNA and subsequent recombination with maternal mtDNA might occur. In 1999, three independent studies reported indirect evidences for recombination in human mtDNA (Awadalla et al. 1999; Eyre-Walker et al. 1999; Hagelberg et al. 1999). One year later, one of these studies (Hagelberg et al. 1999) retracted their conclusions: the extremely rare point mutation at high frequency detected in different lineages in a sample population (which contributed to conclude that recombination was the most plausible explanation) was shown to be an artefact caused by misalignment of the DNA sequences (Hagelberg et al. 2000). Awadalla and colleagues argued in favour of recombination because of linkage disequilibrium declined with distance between variable sites (Awadalla et al. 1999). Eyre-Walker and colleagues (Eyre-Walker et al. 1999) argued for recombination because of the detection of high number of *homoplasies* (the same mutation at the same site independently on two or more occasions unrelated groups) across all the mitochondrial phylogeny, which was interpreted like consistent with undetected recombination between haplotypes.

Many studies have criticized this finding arguing that the conclusions depend on the choice of the method to measure linkage disequilibrium (Jorde and Bamshad 2000; Kumar et al. 2000; White and Gemmell 2009) or that the homoplasy test is sensitive to *mutation hotspots sites* (sites where mutations might have occurred multiple times) (Stoneking 2000). Apart from the methodology, the quality of sequence data used was not as good as it should be; moreover, the results could not be replicated when the data was analyzed under other perspective (Jorde and Bamshad 2000; Kivisild and Villems 2000; Kumar et al. 2000; Macaulay et al. 1999; Merriweather and Kaestle 1999; Parsons and Irwin 2000). Some authors have suggested that the results could be produced for other population genetic processes like mutation rate heterogeneity (Innan and Nordborg 2002) or selection (Wallis 2000).

On the other hand, the evidence for recombination in these studies (Awadalla et al. 1999; Eyre-Walker et al. 1999) is indirect, while other authors using similar indirect studies derived conclusions the opposite direction (Elson et al. 2001; Herrnstadt et al. 2002; Ingman et al. 2000; Jorde and Bamshad 2000; Kumar et al. 2000). Nowadays there is only one empirical study showing evidences that human mtDNA recombination can happen (Kraytsberg et al. 2004), although the mtDNA recombinants were only detected in somatic tissue of a patient; curiously this is the same patient analyzed previously (Schwartz and Vissing 2002), and the authors did not mention that the individual had differences between maternal and paternal mtDNA with the presence of a third mtDNA haplotype.

Even if it is accepted that the human mtDNA recombination occurs, there is no evidence that this event may take place in gametic tissues and therefore be a heritable success. The assumption of maternal inheritance together with no recombination for mtDNA allows tracing genetic lineages straightforward (with the only noise generated by hotspots), because of all variation is generated by mutation. Any attempt aimed to demonstrate mtDNA recombination should also explain why the worldwide mtDNA phylogeny faithfully mirrors the non-recombinant nature of this genome (Bandelt et al. 2005).

I.1.3.4 mtDNA repair mechanisms

It was thought that DNA repair mechanisms were either non-existent or very inefficient in mitochondria and that damaged mtDNA molecules were degraded, and undamaged copies served as templates for new mtDNA synthesis. In addition to environmental factors common to nDNA, mtDNA is more susceptible to oxidative damage by reactive oxygen species (ROS), because it is physically close to the oxidative phosphorylation system which generates these highly active molecules and also to the lack of protective histones.

Nowadays, it has been assumed that mtDNA lacks effective DNA repair mechanisms, like pirimidine dimers for UV-B light. However, several studies have indicated that at least some repair activity exists, called *base excision repair* (BER), which repairs certain types of damage resulting from deamination, simple alkylation and oxidation (Croteau et al. 1999; Kang and Hamasaki 2002; LeDoux et al. 1999; Mason et al. 2003). This mechanism of repair is catalyzed by DNA glycosylases, AP endonuclease, DNA

polymerase, and DNA ligase (Maynard et al. 2010). It is necessary consider that, at least some mutations, may be explained by failures of the mtDNA replication system (Malyarchuk et al. 2002), and indeed mutations in any nuclear genes that code for components of replication or repair systems may expose some individuals to the development of mtDNA errors.

I.1.3.5 High mutation rate

The average mutation rate in mtDNA is at least 10-fold relative to that in nDNA, consequently mtDNA evolves rapidly (Brown et al. 1979; Brown et al. 1982; Wallace et al. 1987) and *D-loop* has an even faster evolutionary rate (*see Figure 3)*, making it useful for studies of human population history. No effective DNA repair mechanism and the lack of histones have been proposed as the main causes for the mutation accumulation in mtDNA.

In recent years, a number of studies have aimed to estimate mutation rates using different approaches, basically, pedigree-based or phylogeny-based methods. For the first one, the mutation rates are estimated by comparing parent-offspring pairs of deep-rooted familial lineages at one loci and counting the number of novel mutations per pair, divided by the number of meiosis (Heyer et al. 2001). Phylogeny-based mutation rate are estimated counting the number of mutations between two species lineages (or two sequences with the greatest number of differences in the case of within-humans comparisons) then the number is divided by the externally derived divergence date.

Several factors have been proposed in order to explain the discrepancy between pedigree-based and phylogeny-based mutation rates. Estimates can be affected by rate heterogeneity between sites; pedigree methods capture “fast” substitution rates whereas phylogenetic methods mostly “slow” substitution rates (Heyer et al. 2001). Moreover, *homoplasy* (occurrence of homoplasies) and back mutations (also called reversions) can be considered common events in ancient lineages, disturbing the number of mutations considered in the case of phylogenetic analysis but not in pedigree studies (Henn et al. 2009).

Some of the most commonly used mutation rate latest years have been 1.79×10^{-7} base substitution per nucleotide per year for 276 bp of the HVI obtained by Forster and colleagues (Forster et al. 1996) using a network-method calibrated by the assumption

that the expansion of haplogroup A2 happened 11300 years ago. Another one was presented by Ingman and colleagues (Ingman et al. 2000); it is based on coding region and they assumed a human-chimp species split at five million years ago (Mya), yielding an estimate of 1.7×10^{-8} base substitution per nucleotide per year. Other value is 1.26×10^{-8} base substitution per nucleotide per year, presented by Mishmar and colleagues (Mishmar et al. 2003b) based on complete genomes using the HKY85 model of nucleotide substitution calibrated by assuming that the split between humans and chimpanzees happened 6,5 Mya ; another one is 3.5×10^{-8} base substitution per nucleotide per year estimated by Kivisild and colleagues (Kivisild et al. 2006) using the same calibration than Mishmar and colleagues but for synonymous changes in mitochondrial protein-coding genes.

Two years ago, Soares and colleagues (2009) obtained a new value for the substitution rate for the entire molecule of 1.665×10^{-8} base substitution per nucleotide per year considering the divergence time between human and chimpanzee was 7 Mya. They calibrated the mtDNA clock considering the divergence human-chimpanzee but adding a correction that considers selection and saturation effects. Endicott and colleagues have criticized the results because the method did not consider variation among contemporaneous lineages (Endicott et al. 2009).

Few months later, Loogvali and colleagues published another value for the synonymous substitutions in the mitochondrial genome: 3×10^{-8} synonymous substitutions per site per year (Loogvali et al. 2009). They presented several explanations for the divergence between their value and the value presented by Soares. One of them was the larger proportion of non-neutral variation in their data in compare with Soares data because of the exclusion of the non-coding region. Another one was that Soares data was not corrected by multiple mutation hits and for this reason the distance was erroneously underestimated (Loogvali et al. 2009).

The discrepancy between the values can be due to the effect of the selection on protein coding sites, the saturation in the control region and the election of the external calibrated estimate.

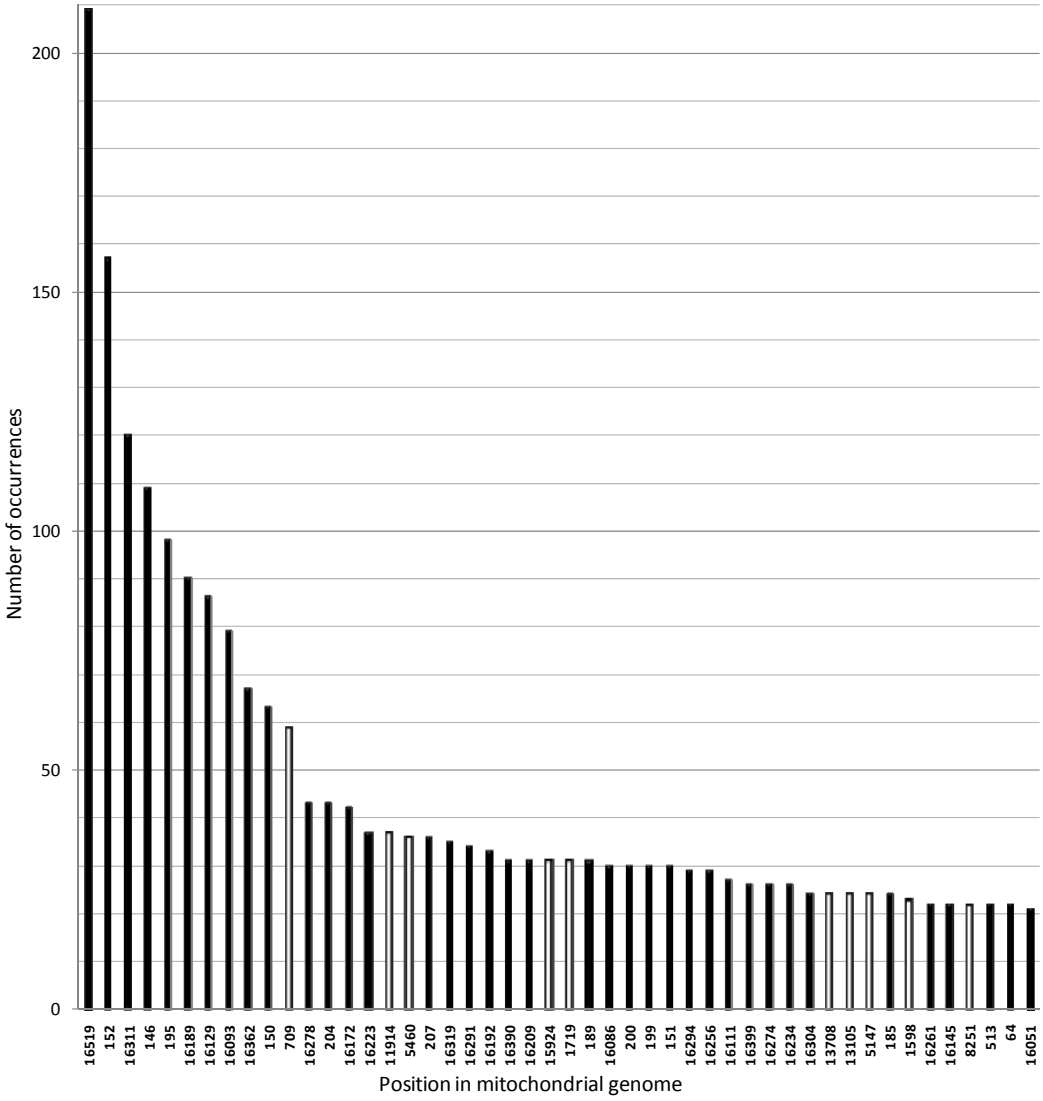


Figure 3 Hot Spots in the mtDNA genome, showing positions that appear > 20 times in the tree. (data from (Soares et al. 2009)). Black blocks are control region positions while white blocks are coding region positions

The mutation rates of mtDNA are used to estimate the timing of several processes due to the consideration of the molecule as a molecular clock. However it has been proposed based on several data that it is not a linear molecular clock (Howell et al. 2004; Loogvali et al. 2009; Soares et al. 2009; Torroni et al. 2001b). Moreover due to the variability of the mutation rate estimates several timescales based on mtDNA variability can be established (*see Figure 4)*.

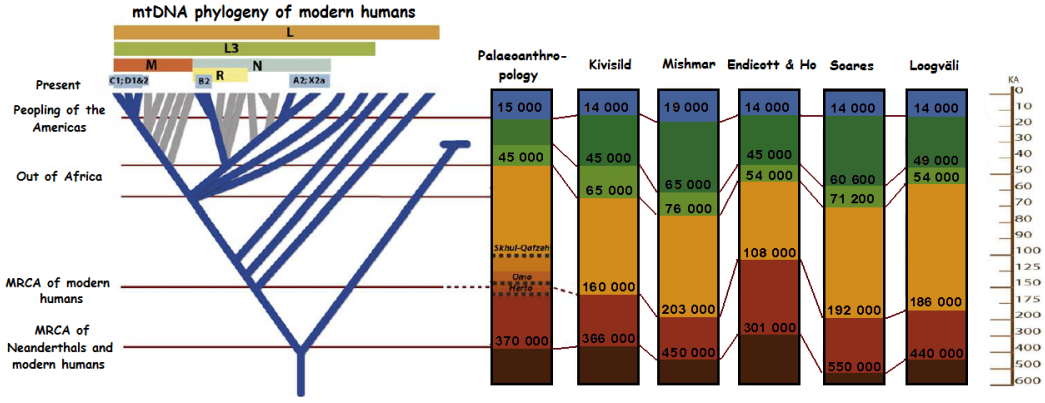


Figure 4 Different timescales of human evolution. Modified from(Endicott et al. 2009)

I.1.3.6 Limitations of mtDNA in human studies

Due to mtDNA is a single locus, it only provides partial information about human evolution, and it only refers to the female population. In forensic casework, the mtDNA test only allows the identification of lineages but not the identification of a single person (like nuclear markers do), although, it could be still useful to exclude suspects.

Due to the maternal inheritance of the mtDNA, it has an effective population size (N_e, which is an estimate of the breeding size of a population) that is about a quarter of the N_e in the autosomal markers. Therefore, the mtDNA is more severely affected by the stochastic effect of genetic drift than autosomal markers. Natural selection can also influence the shape of the phylogenies, patterns of variation, and estimates of coalescence times of mtDNA haplogroups (*see Figure 5*).

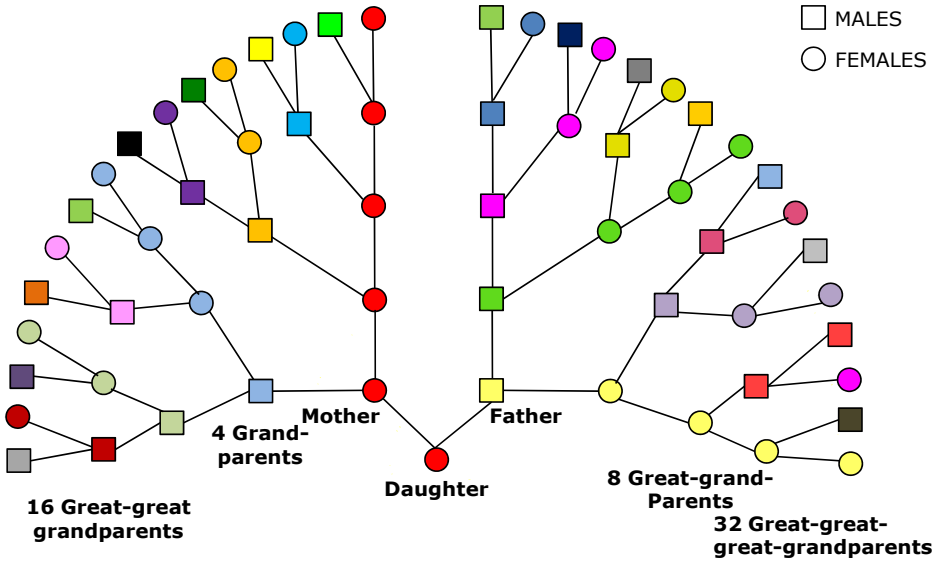


Figure 5 Genealogy tree showing the loss of variation in 5 generations. Although at the beginning there were 32 different mitochondrial haplotypes (each one with a different colour) at the end there is only one type. mtDNA only can shows a part of the history and autosomal markers and Y-chromosome are required in order to know it completetely.

The high level of homoplasies present in mtDNA must be considered in order to prevent the assumption that shared polymorphisms in two branches could be identical by descent. Moreover, it is also necessary to consider the possibility of back mutations to an ancestral state along the tree. Even though the mutation rate estimates are different depending on the method employed, it seems that the mtDNA (*see above*) could not be so useful as other markers (Y-chromosome) to date recent events.

I.2. METHODS TO DETECT MTDNA VARIATION

I.2.1 General technological developments

In the last decades there have been a lot of technological developments, allowing great advances for the detection of DNA variation (see next section for detection of mtDNA variation). The improvements in this line include:

a. Rapid methods for DNA extraction. Classical techniques for extracting DNA require from 3 to 24 hours and the use of toxic chemicals; also in the past, large quantities of biological material were needed to extract reasonable amounts of DNA. Nowadays the DNA extraction can be carried out using automated techniques in as little as 30 minutes and using commercial kits that do not contain toxic reagents.

b. In 1978, Hamilton and Nathan won Nobel Prize in Physiology for their discovery about the effect of the bacterial *RE* (Restriction Enzymes) that cleave DNA at specific sequences. Different patterns of *Restriction Fragment Length Polymorphisms* (RFLPs) were identified on the DNA; during the nineties, several mutations and RE sites were identified by sequencing.

c. The improvement of DNA hybridization techniques allowed comparison of DNA between different species. The method is based on the annealing of two homologous DNA strands, comparing the re-association patterns and their thermal instability. In 1984, this technique allowed to demonstrate that chimpanzees are our closest phylogenetic relative (Sibley and Ahlquist 1984).

d. One of the most important advances in molecular genetics arrived with the *PCR* (Polymerase Chain Reaction) technique, which allows obtaining copies of DNA sequences through a particular region using primers. The method of alternate heating (for denaturing the DNA) and cooling (for annealing the target sequence) allows the synthesis of specific DNA regions in geometric progression. In 1993, Kary Mullis won Nobel Prize in Chemistry for the development of PCR method and thermocycler

e. Automated DNA sequencing, followed by the development of high throughput sequencing machines have allow to obtain rapid characterization of the human genome. The earliest methods were time-consuming and the identification of specific DNA fragments on gels followed by radio-labelling. The high throughput methods using fluorescence terminators and capillary electrophoresis have changed the scenario completely. Nowadays a "new technological revolution" has arisen with the development of new platforms and techniques such as pyrosequencing with 454 Life Sciences (Roche Diagnostics, Indianapolis, USA), ultra-sequencing using oligonucleotide ligation with SOLiD platform (AB, Applied Biosystems, Foster City, CA) and sequencing by synthesis with Solexa (Illumina).

	454 (Roche)	SOLiD (AB)	Solexa (Illumina)
Sequencing method	Pyrosequencing chemistry	Oligo ligation-cleavage	Labeled base addition
Base per run	100 million	1000-1500 million	1000-3000 million
Read length	100-200 bases	35 bases	35-50 bases
Length of run	7,5 hours	2-4 days	2-3 days

Table 1 Comparison of different new sequencing platforms.

I.2.2 The evolution of mtDNA variability detection techniques

One of the pioneering studies on mtDNA variation suggested different specific patterns of cleavage followed by digestion with 18 RE sites in 21 humans of diverse population and geographic origins (Brown 1980). One year later, another study showed similar results using one RE (*Hpa*I) analyzed in five ethnic groups (Denaro et al. 1981). In 1987, Cann and colleagues, based also on the analysis of few RE sites, established the origin of the common ancestor in Africa, popularized in the media as the *Mitochondrial Eve* (Cann et al. 1987).

More recent studies of mtDNA variation were based on the analysis by sequencing of control region segments and/or restriction maps (Vigilant et al. 1991) (Ward et al. 1991). The pioneering studies based on sequencing small mtDNA segments revealed the existence of many haplotypes in populations. Today, sequencing complete genomes has started to be common in population and medical studies (Richards and Macaulay 2001).

Modern technological improvements allow now to sequence samples with tiny amounts of DNA or bearing very degraded DNA; the analysis of mtDNA Neanderthal genomes represent paradigmatic cases (Briggs et al. 2009; Green et al. 2010; Green et al. 2008).

Other techniques to analyze mtDNA variation were development in the last two decades, although most of them are not used anymore. *Single Strand Conformation Polymorphism (*SCCP*)* is based on the different mobility across a poliacrilamide gel of denaturalized DNA strands. RE-SSCP (*Restriction Enzyme-SSCP*) combines a digestion with RE and the SCCP technique, while FSSCP-OF (*Fluorescent-SSCP-Overlapping fragments)* combines PCR amplification of overlapping fragments by PCR and automatic detection using fluorescent reagents. Traditional RFLPs typing is being substituted by new methods such as those based on SBE (Single Base Extension, called also minisequencing) (Chinault et al. 2009; Nishigaki et al. 2010; Thieme et al. 2009; Vallone et al. 2007) and chips (van Eijsden et al. 2006).

I.3. NOMENCLATURE

I.3.1 How to report the differences with respect to reference sequence

In order to standardize the nomenclature, main efforts have been carried out in several fields of research; in the last few years, more intensively in the forensic field for obvious reasons. In 2001 some recommendations were published by forensic geneticists (Tully et al. 2001). Those recommendations are nowadays followed by a wider range of geneticists in other fields of research.

I.3.1.1 Single nucleotide variation

Each mtDNA sequence is described by listing those variants that differs from the rCRS. For example, position 247 is G in the rCRS, if the sample under analysis harbours an A at this site, this difference can be referred as G247A or m.247G>A chiefly in clinical articles, or 247A mainly in forensic studies.

- In a majority of population genetic studies, one usually refers a *transition* (change between purines or between pirimidines) only by using the position number, while a suffix is added in case of a *transversion* (change of a purine base by a pyrimidine). In the example above, we would write 247 for the transition G to A at this site, whereas an A at position 055 (where the rCRS is a T) would be referred as 055A (or 055a in some studies) (*see Figure* 6)

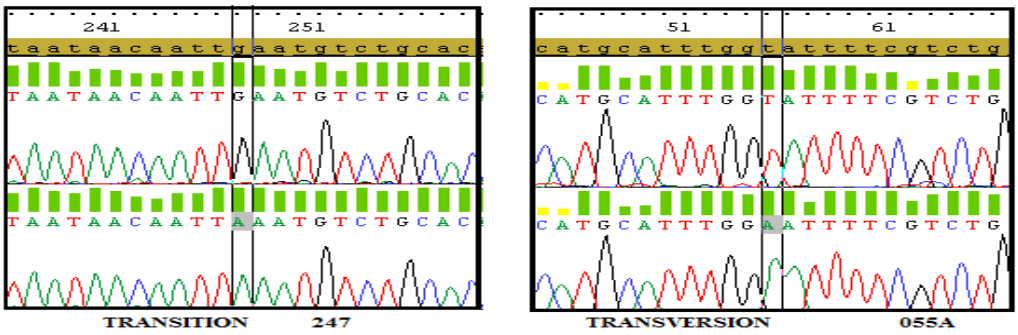


Figure 6 Electropherograms with the examples of transition and tranversion explained on the text. On the top rCRS sequence, at the bottom the sequence with the change

- A deletion is generally referred as “d” or “del”. For example if the deletion occurs between 246 and 248 position, it can be denoted as 247d or 247del. If the deletion is inside a homopolymeric tract, it is not possible to asses which one of the bases was deleted so the recommendation is to assign the deletion at the last position of the tract; for example, rCRS has 5 C’s between positions 494 and 498, if the sample carries only 4

C's it should be indicated as 498d. However, in the literature some authors denotes the deletion at the first np, namely, 494d (Chinnery et al. 2001) (*see Figure 7*).

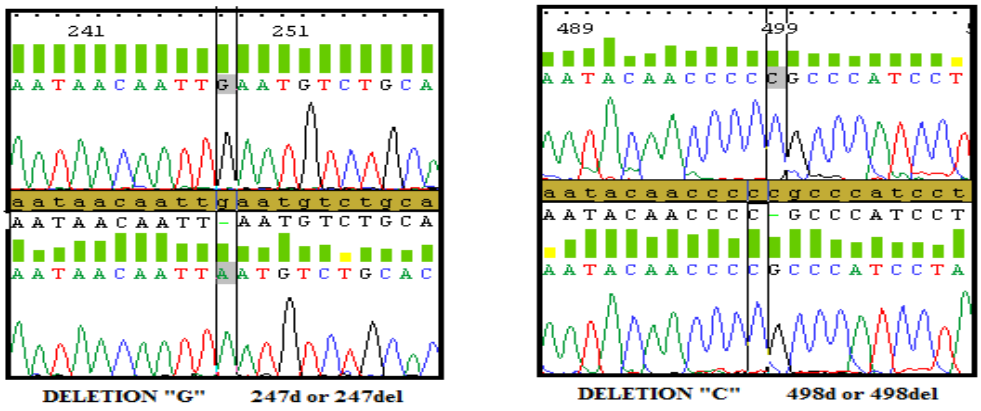


Figure 7 Electropherograms with the examples of deletions explained on the text

I.3.1.2 Length variants

- In the case of the insertions, ".1", "+" or "ins" are used. For example, an insertion of a G after position 042 will be indicated as 043+G. If the insertion is inside a homopolymeric tract the exact position of the insertion is unknown so the recommendation is to assign the insertion to the 3' position. For example the rCRS shows 7 consecutive C's between positions 303 and 309, if a profile carries 8 C's the annotation should be 309.1C or 309+C, with 9 C's it should be denoted as 309.2C or 309+CC or 309+2C (although not all authors follow this rule (Chinnery et al. 2001) (*see Figure 8*)

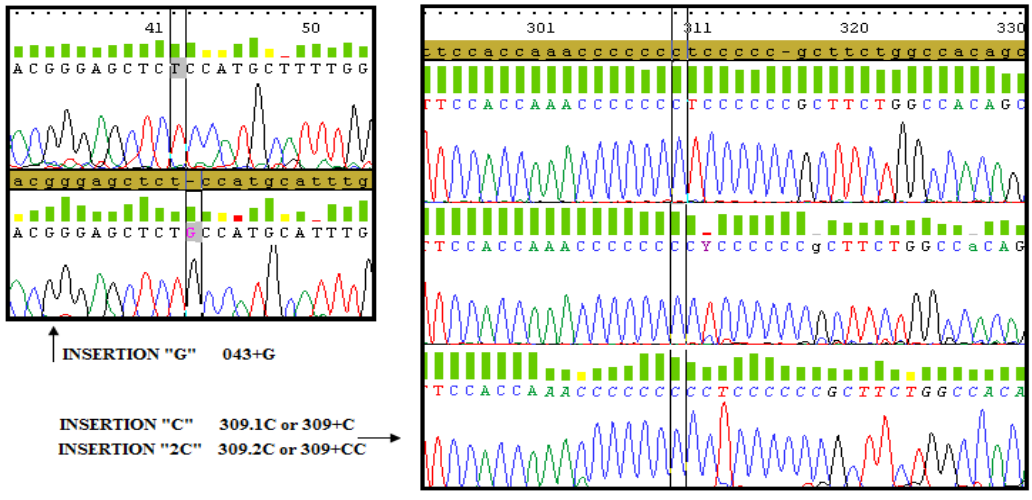


Figure 8 Electropherograms where several insertions appear.

In 2002, some recommendations were published for treatment of length variants in the human mtDNA control region (Wilson et al. 2002). In this study they established three recommendations:

1. Characterize profiles using the least number of differences from the reference sequence

2. If there is more than one way to maintain the same number of differences with respect to the reference sequence, differences should be prioritized as follows: (a) insertions/deletions (indels) (b) transitions (c) transversions.

3. Insertions and deletions should be placed 3' with respect to the L-strand. Insertions and deletions should be combined in situations where the same number of differences to the reference sequence is maintained.

These recommendations were considered to be too naïve by several authors. For instance, Bandelt and Parson (Bandelt and Parson 2008) strongly encourage to use a phylogenetic criteria instead the arbitrary rules proposed by Wilson and coworkers. Budowle and colleagues (Budowle et al. 2008), co-authors in the first Wilson's study, argued that inconsistency in nomenclature due to different approaches do not lead to ''unjustified exclusion of the culprit as the donor of the stain.''. They argued that most comparisons in forensic casework are side-by-side such that the entire sequence information is directly accessible.

Multiple alignments of the mtDNA sequence with the rCRS may be possible not only for length variants and different nomenclatures could be considered, this phenomena has been called *jumping alignments* (Den Hartog et al. 2009). For instance, the sequence in Figure 9 shows the rCRS (top) and two alternative alignments for a given sequence (bottom). By default, the SeqScape software produces an alignment resulting in the following variants: 16183C 16186 16189d 16192; however, when applying the Wilson's rules, the alignment would result in the following variants 16183C 16186 16189 16191 16193d. When using a phylogenetic approach, the best alignment should better be 16183d 16187 16189 16192 (*see Figure 9*).

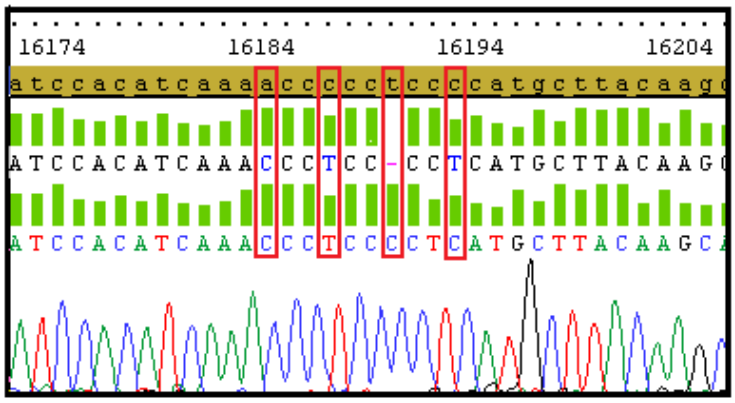


Figure 9 Electropherogram where multiple alignments can lead to different nomenclatures.

Two years ago, the impact of the jumping alignments on mtDNA population analysis and database searching was evaluated in two mitotype databases: Scientific Working Group on DNA Analysis Methods (SWGDAM) and European Mitochondrial DNA Population database (EMPOP). They found 1.86% and 3.14% of the pairwise comparisons in the SWGDAM and EMPOP databases, respectively, experienced incidences of jumping alignments (Den Hartog et al. 2009).

I.3.1.3 Heteroplasmic positions

I.3.1.3.1 Point heteroplasmy

In the case of a clear point heteroplasmy with both variants at same level, IUB designation may be used (*see Figure 10*). Different nomenclatures have been proposed in different fields of research. Usually, the presence of a C and a T at a given position is designated as C~T or C/T.

If the electropherograms shows two nucleotide bases at the same position at different intensities, this could be denoted using symbols “>” or “<”; for instance, C>T would indicate that the signal for C is larger than the signal for T. Generally, it is not easy to discriminate between heteroplasmy and sequencing artefacts or background noise. In order to prevent these errors a confirmation of heteroplasmy by resequencing of both strands is necessary. In case of an ambiguity, N should be used instead.

CODE	TRASLATION
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
B	C, G or T
D	A, G or T
H	A, C or T
V	A, C or G
R	A or G (purRine)
Y	C or T (pYrimidine)
K	G or T (Keto)
M	A or C (aMino)
S	G or C (Strong = 3H bonds)
W	A or T (Weak = 2H bonds)
N	aNy base

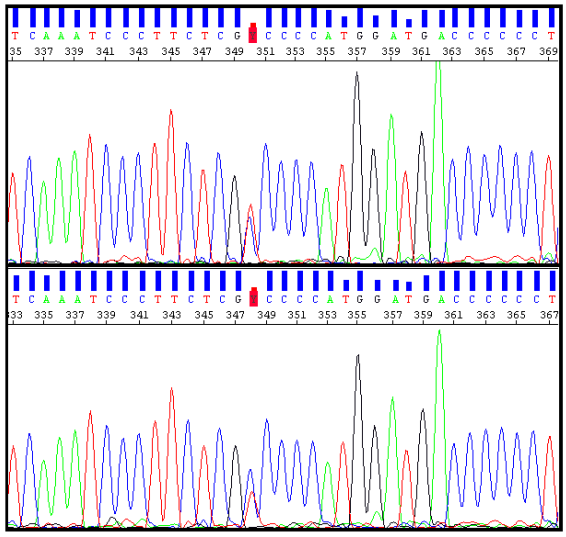


Figure 10 NC-IUB (Nomenclature Committee International Union of Biochemistry) CODE and two examples of point heteroplasmy at position 16362. SeqScape software uses IUPAC code and for this reason appears "Y" in both cases, at the top T>C and at the bottom C>T.

I.3.1.3.2 Length heteroplasmy

Some of the biggest efforts for the detection, confirmation and standardization in the nomenclature of length heteroplasmy have come from forensic genetic research (Brandstatter et al. 2004; Forster et al. 2010; Lutz-Bonengel et al. 2004).

Mononucleotide repeats are known mutation hotspots, which are potentially due to slippage of the DNA polymerase during replication, resulting in DNA length heteroplasmy (Lutz-Bonengel et al. 2004). Human mtDNA has three mutational unstable homopolymeric stretches in the control region that can frequently lead to length heteroplasmy: nps 16184-16189, 303-315 and 568-573; the two former homopolymeric tracks are interrupted by the presence of a thymine residue at np 16189 and 310, respectively. For the first tract, the thymine residue at np 16189 is often replaced by a cytosine, producing an uninterrupted homopolymeric tract with 8–14 cytosine residues. For the second tract the most common phenomenon of HVII length heteroplasmy is the insertion of cytosines in the 303–309 segment as well as replication slippage in the case of a thymine to cytosine transition at np 310. The mechanism that generates heteroplasmy in the track 568-573 is the incorporation of one or more C residues. Some authors consider a fourth tract in the control region, between nps 456-463 which is interrupted by a T residue at np 460. The mechanism to generate the length heteroplasmy due to a homopolimeric tract is a transition at this position (Forster et al. 2010).

Heteroplasmies can be detecting using sequencing only if there minor allele is present at about 10% of the mtDNAs in the sample. When it is not possible to solidly report the exact nature of the heteroplasmy, different amplification and sequencing reactions should be carried out; if the ambiguity persists, the result should be denoted as "N".

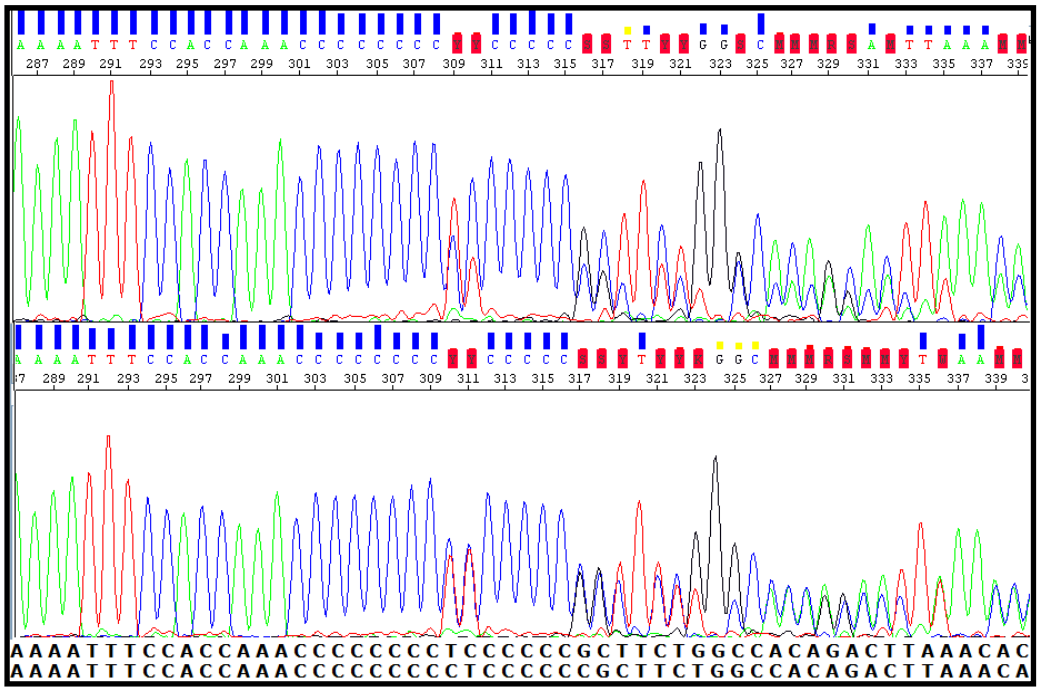


Figure 11 Electropherograms which show different level of length heteroplasmy at the same positions. The sequence at the top should be annotated like 309+C>309+CC while sequence at the bottom should be annotated like 309+C=309+CC.

In order to prevent several notations for the same heteroplasmy some recommendations have been proposed (Wilson et al. 2002), although alternative nomenclatures have been proposed (see above).

I.3.1.4 RFLPs

The nomenclature for RFLPs patterns is straightforward. A plus sign (+) indicates the presence of a restriction site, and a minus sign (-) indicates its absence. Then, the position of the first nucleotide of the recognition sequence is given followed by the RE. For example +12308*Hinf*I indicates that the sample was recognized and cut with the RE *Hinf*I, the position 12308 would be the first one in the recognition site (GANTC). A slash (/) separating states indicates the simultaneous presence or absence of restriction sites that

can be correlated with a single nucleotide substitution. For example, -9052*Hae*II/-9053*Hha*I indicates that the sample was neither recognized at position 9052 with the RE *Hae*II nor at position 9053 with the RE *Hha*I, and the reason is the presence of an "A" at np 9055. For more examples see below (*Table 2*)

RE	*Source*	Recognition Sequence:	Recognition nucleotide (Haplogroup) Position*		
***Acc*I**	*Acinetobacter calcoaceticus*	...GTMKAC... / ...CAKMTG...	+14465 (X) 14470C		
***Alu*I**	*Arthrobacter luteus*	...AGCT... / ...TCGA...	-4310 (L0) 4312T	+4769 (H2a) 4769A	-7025 (H) 7028C
			-7055 (L1) 7055G	+10032 (I) 10034C	+10319 (L1c1'2'4'6) 10321C
			+10397 (M) 10400T	+11329 (U4) 11332T	+13262 (C) 13263G
			+15606 (T, P) 15607G		
***Ava*II**	*Anabaena variabilis*	...GGWCC... / ...CCWGG...	+8249 (L0d, N1e'I, N1b, W) 8251A		-12629 (T1) 12633a
			-13367 (T) 13368A	-16390 (I) 16391A	
***Bam*HI**	*Curtobacterium albidum*	...GGATCC... / ...CCTAGG...	+13666 (T) 13368A	-14258 (L3j, M32c) 14260G	
			+16389 (I) 16391A		
***Bfa*I**	*Bacteroides fragilis*	...CTAG... / ...GATC...	+4914 (T) 4917G	+5004 (H4) 5004C	
***Dde*I**	*Desulfovibrio desulfuricans*	...CTNAG... / ...GANTC...	-679 (L2c) 680C		
			-1715 (B4e, D6, H7a, L3d3, L3h1, M33b, M27b, M28, N1'5, P4, R21, X2) 1719A		
			-5003 (H4) 5004C	-5176 (D) 5178a	-6296 (H7c) 6296a
			+10394 (J, K1, non N) 10398G		+16223 (G2a, H22) 16227G
***Hae*III**	*Haemophilus aegyptius*	...GGCC... / ...CCGG...	-322 (L2c) 325T	+663 (A) 663G	-8250 (N1e'I, N1b, W) 8251A
			-8391 (Y) 8392A	-8994 (W, HV9, H21) 8994A	
			+13803 (L2a) 13803G	-13957 (L2c) 13958c	
			+16398 (F4a, L2d, L2e, M53b, M32c) 16399G		+16517 (H, U*,K) 15519C
***Hae*II**	*Haemophilus aegyptius*	...RGCGCY... / ...YCGCGR...	-4529 (N1e'I) 4529t	+4833 (G) 4833G	-9052 (U8b'K) 9055A
***Hha*I**	*Haemophilus haemolyticus*	...GCGC... / ...CGCG...	+663 (A) 663G	-7598 (E) 7598G	-9053 (U8b'K) 9055A
***Hinc*II**	*Haemophilus influenzae Rc*	...GTYRAC... / ...CARYTG...	-12406 (F1) 12406A	-13259 (C) 13263G	
***Hinf*I**	*Haemophilus influenzae Rf*	...GANTC... / ...CTNAG...	+10806 (L0,L1, L5) 10810	+12308 (U) 12308G	-13103 (U1) 13104G
			+16389 (B4f, E, L2, M48, M52a, M27b, N1b, R8b) 16390A		
			-16065 (J) 16069T		
***Hpa*I**	*Haemophilus parainfluenzae*	...GTTAAC... / ...CAATTG...	+35932 (L non L3) 3594T		
***Mbo*I**	*Moraxella bovis*	...NGATCN... / ...NCTAGN...	-951 (H2a1) 951A	+2349 (L1b, L3e) 2352C	-4990 (U1a) 4991A
			+7933 (Y1a) 7933G	-8616 (L3d) 8618C	+13104 (U1) 13104G
			+13667 (T) 13368A	+16390 (I) 16391A	
***Mse*I**	*Micrococcus species*	...TTAA... / ...AATT...	-14766 (HV, H) 14766T		
***Msp*I**	*Moraxella species*	...CCGG... / ...GGCC...	-8112 (L0d) 8113a	-8150 (L0d) 8152A	+11436(N1c) 11437G
			+13100 (H8) 13101c	-15925 (T) 15928A	
NlaIII	*Neisseria lactamica*	...NCATGN... / ...NGTACN...	+4216 (R2'JT) 4216C	-4577 (V) 4580A	
RsaI	*Rhodobacter sphaeroides*	...GTAC... / ...CATG...	-2758 (L0, L1) 2758A	+4643 (U4) 4646C	+7702 (U4b) 7705C
			-8012 (HV1) 8014t	+12345 (M1a1) 12346T	+12806 (L1c) 12810G
			+15907 (U2e) 15907G	-16049 (U2) 16051G	-16208 (H1a1) 16209C
			-16303 (H5) 16304 C	-16310 (K) 16311C	
TaqI	*Thermus aquaticus YTI*	...TCGA... / ...AGCT...	+5419 (X1b1) 5420C	+9070 (L1c) 9072G	+10084 (L3b) 10086G
			+14068 (U1) 14070G		

Table 2 Different RE used to the genotyping of different positions along the mtDNA molecule, their origin, recognition sequence and some positions which are genotyping with them. * Appears the "*" or "-" and the position which are recognized, between "()" some of the haplogroup or sub-haplogroup which are genotyped by the RE and finally the nucleotide present at this position.

I.3.1.5 Phylogenetic tree nomenclature

Usually mutations are transitions unless a suffix A, G, C, T or d indicates a transversion or a delection, respectively; insertions are indicated by + followed by the inserted nucleotides.

An alternative nomenclature for mutations was used for Behar and colleagues (Behar et al. 2008b) and it is also followed for the trees in the present dissertation: all mutations are indicated with the np plus a capital letter indicating transitions, but tranversions are specified using lowercase letter (*see Figure 12*).

Usually "@" as prefix before the position or "!" as a suffix indicate a back mutation; while underlying a np or round brackets indicate the presence of recurrent mutations in the tree. Mutations in *italics* are disregarded for time estimations; note however that in Phylotree (www.phylotree.org (van Oven and Kayser 2009)), italic nps indicate preliminary diagnostic mutations that need confirmation, whereas round brackets are used to indicate recurrent mutations or uncertainty.

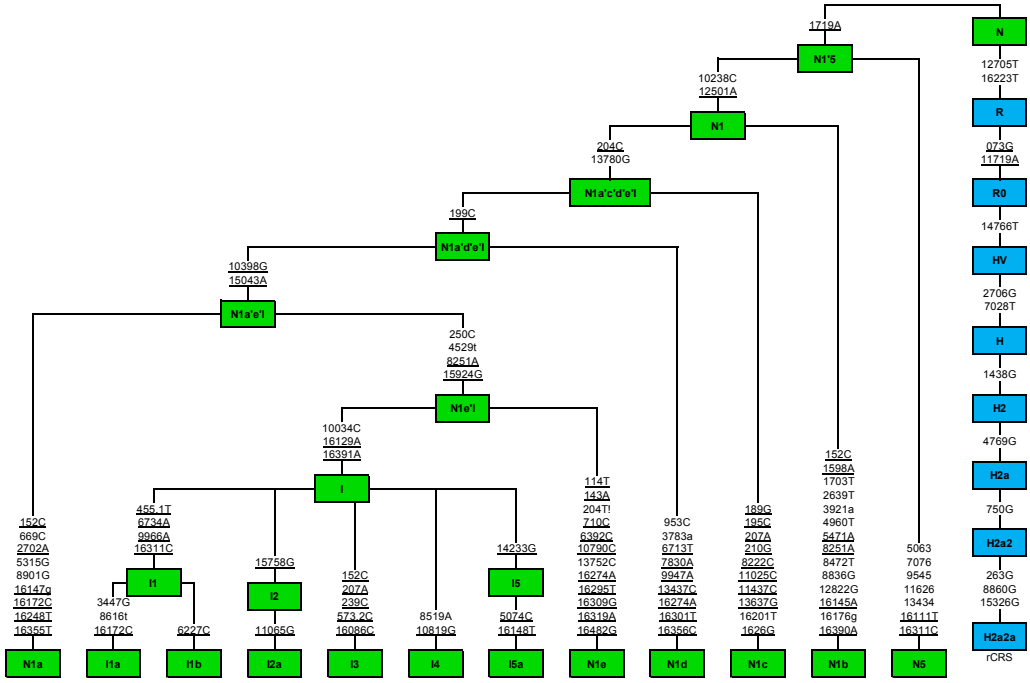


Figure 12 Phylogenetic mitochondrial tree of the haplogroups N1'5 with some of their sub-haplogroups.

I.3.2 How to name haplotypes and haplogroups

The *haplotype* can be defined by a set of variants that are inherited together throughout the maternal line and are usually reported as a list of changes with respect to the rCRS. *Haplogroup* is a group of mtDNA haplotypes derived by descent from the same ancestral mtDNA molecule as revealed by the sharing of a distinguishing *mutational motif* (Torroni et al. 2006). Haplogroup was coined for the first time in 1993 (Torroni et al. 1993) although the concept was introduced before in other studies, e.g. "*clans of types*" (Stoneking et al. 1990), "haplotype cluster" (Torroni et al. 1992), etc..

HAPLOGROUP	HVS-I MOTIF
L0a	129 148 172 187 188G 189 223 230 311 320
L0b	051 129 148 164 172 182C 183C 189 209 230 278 287 291 311 519
L0f	129 169 172 187 189 223 230 278 311 327 354 368 519
L0k	129 166C 172 187 189 214 223 230 278 291G 311 519
L0d	129 187 189 223 230 243 278 311 519
L1b	126 187 189 223 264 270 278 293 311 519
L1c	129 187 189 223 278 294 311 360 519
L5c	111 129 148 166 187 189 223 254 278 311 360
L5a	129 148 166 187 189 223 278 311 355 362
L2a	223 278 294 390
L2b	114A 129 213 223 278 390
L2c	223 278 390
L2d	093 129 189 278 300 354 390 399 519
L2e	111A 145 184 223 239 278 292 355 390 399 400 519
L6a	048 223 224 278 311 519
L6b	048 223 224 278 311 519
L4a	223 260 311 362 519
L4b	223 311 362
L3a	223 316
L3b	124 223 278 362
L3c	223 311 362
L3d	124 223
L3j	223
L3f	209 223 519
L3h	223
L3e	223
L3i	223
L3k	223
L3x	189 223

Table 3 HVS-I motif for mitochondrial L-sub-haplogroups. All the variant position from rCRS minus 16,000 are indicated.

The rules to designate the name to a haplogroup or sub-haplogroup were clearly established by Richards and colleagues (Richards et al. 1998). Main haplogroups (called *clusters* in the Richard's study) should be indicated by capital letters, then a hierarchical notation should be used alternating numbers and small letters, for example L1c1c1, is a sub-lineage of L1c1c, and so on. Each cluster should made up for a *monophyletic* clade in the phylogeny (all of the members of a group have a common ancestor). This straightforward nomenclature needs continuous updating as more resolution with the arrival of new mtDNA data (preferably from complete genomes). Sometimes studies of mitochondrial variation have shown that several of the monophyletic clades were in fact *paraphyletic* (when all the members of a group which have common ancestor are not included). A paradigmatic case is haplogroup K, initially thought to be a different clade outide haplogroup U, but it is now recognized to be one of the many sub-lineages of U. Haplogroup JT includes all the members of haplogroup J and haplogroup T; pre-HV (nowadays called R0) indicates a bigger clade where haplogroups H, V and HV, the haplogroup (pre-HV)1 (nowadays called R0a) are nested. Finally "*" following the name of the haplogroup, indicates the members of a haplogroup that are not included in a known sub-haplogroup; for example K* indicates all sequences belonging to haplogroup K that are not K1 nor K2. There are not enough capital letters for all the haplogroups and some of them were named before having a clear idea of their correct position into the human mitochondrial phylogenetic tree. Some examples are M13'46'61, M29'Q or N1e'I (*see Figure 12*).

This nomenclature has been followed by those interested in Y-chromosomal variation (2002).

Nowadays the standard for haplogroup nomenclature is being compiled in Phylotree (van Oven and Kayser 2009), and can be freely accessed via web in www.phylotree.org.

I.4. PROBLEMS WITH THE SEQUENCE DATA INTERPRETATION

A huge amount of data on mtDNA variation has been generated in the last two decades; in population genetics, forensic genetics and clinical genetics. Unfortunately this data is not always correct, leading to misinterpretations of different nature of results and conclusions.

I.4.1 Laboratory errors

There are several potential artefacts in a laboratory that can lead to problems of different nature. For decades, these errors have left their imprint in different fields of research. In aDNA studies, erroneous theories about evolution of modern humans or past population histories have been suggested; in forensic casework, genotyping errors could have lead to erroneous inclusions or exclusions and have severely affected well known forensic databases (Bandelt et al. 2004a; Bandelt et al. 2004b); and in clinical studies different variants can appear as associated to a given disease when in reality are mere artefacts.

Once the sample "arrives" to the laboratory, it needs to be processed through several steps for DNA typing. The phylogenetic approach has demonstrated to be very useful to detect several kinds of errors arising at different steps of the genotyping process (Bandelt et al. 2001b; Salas et al. 2005a).

FINAL ERROR PHENOTYPE	LABORATORY STEP	CAUSE OF ERROR
Type I or Base shift	5	L, B, N, T
Type II or Reference bias	5	L, T
Type III or Phantom mutation	3, 4, 5	A, L,
Type IV or Base misreporting	5	N, T
Type V or Artificial recombination	1, 2, 3, 4, 5	C, M, S, A, B

Table 4 Classification of the different sources of laboratory errors. The laboratory steps are: (1) DNA extraction, (2) PCR, (3) Sequencing reaction, (4) Electrophoresis, (5) Interpretation and documentation. For the causes of error (C) indicates contamination, (M) sample mix-up, (S) sample manipulation and bias, (A) sequencing artefact, (L) misalignment or incorrect reference sequence, (B) base-call misinterpretation, (N) nomenclature violation, (T) transcription violation.

The different types of errors are as follow (*see also Table 4*):

Type I or base shift, can appear when a position or several positions are miss scored due to an incorrect alignment or reading shift or by a column shift during the preparation of a table.

Type II or reference bias, can be caused by the mistake of nucleotide variants relative to the reference sequence, like positions omitted by overlooking or missing the variation in a stretch because of difficulties in the reading.

Type III or phantom mutations, are due to chemical problems, can appear like sequencing errors during electrophoresis. This kind of error has been intensively studied in the literature (Bandelt et al. 2002; Brandstatter et al. 2005; Herrnstadt et al. 2003).

Type IV or base misreporting, when the nucleotide change is incorrectly written. For example when a transversion is scoring like a transition or the base for a deletion has changed.

Type V or artifactual recombination, this error can be due to sample contamination or mix-up of several fragments from different samples

The errors in mtDNA data have been the main topic of many scientific research and debates (Bandelt and Kivisild 2006; Parson 2007; Stoneking and Nasidze 2006; Yao et al. 2009).

I.4.2 NUMTs

Mitochondria have co-evolved with their host organisms, and during this process many fragments of DNA that have presumably been present in the original endosymbiont have been transferred to the nDNA of the eukaryotic cell and have also been identified in humans. These transferred fragments from mtDNA are called *nuclear inserts of mtDNA* or NUMTs (Lopez et al. 1994); . They are usually non-functional pseudogenes, but it has been suggested that some NUMTs might even be functional (Tourmen et al. 2002; Woischnik and Moraes 2002). The transfer of mtDNA to nDNA seems to be an ongoing process that shapes nuclear genomes. Indeed, *de novo* transfer of mtDNA to nDNA has been shown to cause a genetic disease by introducing a premature stop codon within a functional nuclear gene (Turner et al. 2003).

Several mechanisms of NUMTs insertion into the nucleus have been proposed. The most supported pathway involves the degradation of abnormal mitochondria although others have been suggested, such as lysis of mitochondrial compartment, encapsulation of mtDNA inside the nucleus, direct physical association between the mitochondria and the nucleus membranes, or mtDNA that enters into the nucleus and integrates into nuclear chromosomes (Hazkani-Covo et al. 2010). To explain the integration into the nuclear genome it has been proposed a mechanism where NUMTs show a short homology with respect to nuclear genome and they are inserted into double-strand breaks by the non-homologous end joining machinery (Hazkani-Covo et al. 2010). Insertion of NUMT can also occur without this micro-homology and the process is called blunt-end repair(Hazkani-Covo et al. 2010).

In 1983, Tsuzuki and colleagues were the first authors that found a NUMTs in the human nuclear genome (Tsuzuki et al. 1983). Since this pioneering study, using different methods (hybridization, sequencing, and similarity searches in DNA databases), several NUMTs have been identified. The study of Mourier et al. compiled an inventory of 280 human NUMTs using a BLAST alignment approach (Mourier et al. 2001). In another study a total of 1105 nDNA sequences homologous to mtDNA were found in the August 2001 Goldenpath human genome database with 286 pseudogenes (Tourmen et al. 2002). Mishmar and colleagues analyzed 247 NUMTs using a BLAST searching approach from the Celera and NCBI GenBank human genome sequences using the rCRS as query (Mishmar et al. 2004), Ricchetti and colleagues found 211 NUMTs from a blastn search on the database of *H. sapiens* published by the public consortium in 2001, using as query the CRS (Ricchetti et al. 2004). Based on a comparative study between human and chimpanzees, Hazkani-Covo and Graur (Hazkani-Covo and Graur 2007) have estimated in 452 the number of NUMTs in the human genome.

NUMTs can be a problem when the sample has small amounts or degraded DNA (as in forensic cases) or even when the sample has a great amount of DNA (like from fresh blood). NUMTs can generate artefactual patterns in sequencing analysis such as unspecific bands in acrilamide or agarose gels, sequence ambiguities (e.g. heteroplasmic-like patterns) in polymorphic sites, etc (*see Figure 13)*. Different strategies can be used in order to avoid potential problems with NUMTS: (i) in cases where the amount of DNA is not a limitation, one could carry out a PCR using different pair of primers to obtain larger amplicons (most of the NUMTs are smaller than 1000bp), (ii) performed different PCR using less number of cycles, (iii) isolate each PCR product from a gel and analyzed them independently, or (iv) cloning and subsequent sequencing of the PCR products, etc.

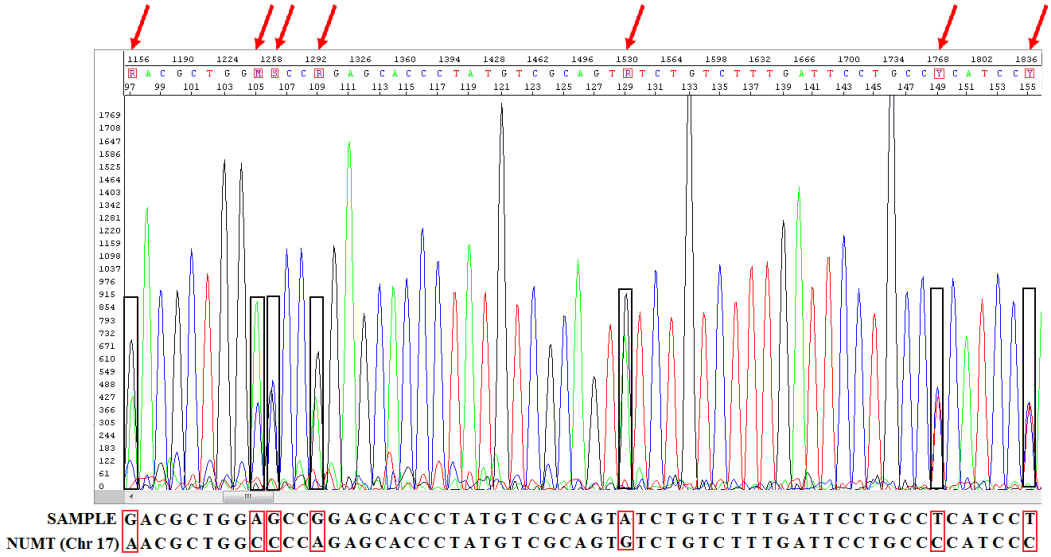


Figure 13 Example of electropherogram with heteroplasmy due to an admixture with a NUMT. The positions which present the heteroplasmy are 094, 102, 103, 106, 126, 146, and 152. These heteroplasmic positions (and also the rest of the non mtDNA sequence) disappeared when the amplicon was larger.

For some authors NUMTs can represent an opportunity to study an ancient mtDNA (Kivisild et al. 2006); such that NUMTs could show previous variation states before the mutations started to accumulate in modern human mtDNAs. Usually these fragments seem to have less mutation rate because of NUMTs usually are inside intronic regions (Ricchetti et al. 2004).

In 2008 a study focused on the compilation of NUMTs using bioinformatic tools (Lascaro et al. 2008). The Yao's et al study focused on artefacts generated by NUMTs in the disease studies (Yao et al. 2008). The study of Goios et al. conducted some experiments to determine the co-amplification of NUMTs, concluding that employing routine techniques, there is no risk of co-amplification.(Goios et al. 2008). Yao and colleagues argued that it was not clear whether in the field of medical genetics routine techniques were always executed correctly.

I.5. ANALISES OF THE GENOTYPING DATA

I.5.1 Methods to analyze mtDNA variation

The study of mtDNA variation can be used to make inferences about different aspects of human evolution. Genetic variation can be analyzed at an intra-population (variation within a population or subdivisions of a population) and/or inter-population

(variation between populations) level. Several analytical methods are available to explore the effects of different evolutionary forces on population subdivision and patterns of variation, such as nucleotide and sequence diversity, tests of neutral evolution, mismatch distribution or analysis of molecular variance (AMOVA), etc. The relationship between haplotypes can be analyzed using phylogenetic trees, or phylogenetic networks. Also, frequencies and geographical distribution of haplotypes or haplogroups can be interpolated in geographic maps.

I.5.1.1 Nucleotide diversity statistics

According to the neutral theory of molecular evolution, the majority of nucleotide substitutions have no effect on fitness of an individual (i.e. are neutral) and most polymorphisms are transient, waiting fixation due to drift. Therefore, assuming mutation-drift equilibrium it is possible to determine the expected level of diversity (θ) in a population or its subdivision using the mutation rate (μ) and effective population size (N_e) for mtDNA, with the equation:

$$\theta = 2N_e\mu$$

The parameter θ is an important factor in several different molecular statistical techniques and is often compared to the nucleotide diversity measure (π) which was introduced by (Nei and Tajima 1981). The π statistic is a measure independent of sample size describes the probability that two copies of the same nucleotide drawn at random from the same set of sequences will differ and is represented using the equation:

$$\pi = n(x_i x_j \pi_{ij})/(n-1)$$

where n equals the number of sampled sequences, x_i and x_j are the frequencies of $_ith$ and $_jth$ sequences and π_{ij} is the proportion of nucleotide differences between them.

I.5.1.2 Measures of neutral evolution

In order to determine whether or not populations are being influenced by evolutionary forces other than selection it is important to ascertain whether the amount of genetic diversity exhibited by these populations deviates from neutrality. Several different neutrality tests exist, including Tajima's D, HKA, McDonald-Kreitman, Fu's Fs, Fu and Li's D as well as other.

These indices have been created to detect signatures of natural selection in DNA sequences. However, all of them are somehow sensitive to demographic changes, so, in mtDNA studies, one cannot be sure to differentiate between the effect of natural selection and neutral evolution. However, under certain assumptions, these indices can still be useful. For example, Tajima's D and Fu's Fs, applied under the assumptions of neutral evolution, could allow the identification of a population expansion or a population under constant population size. Population growth generates an excess of mutations in the external branches of the genealogy and therefore an excess of substitutions are present in only one sampled sequence. This leads to a star-like phylogeny characterizes by a large ancestral central node with several radiating derivative branches. Tajima's D uses information from the sample mutation frequency and is based on the infinite-sites model without recombination.

Fu's Fs is also based on the infinite-site model without recombination but utilizes data from haplotype distributions. Statistically significant negative scores indicate an excess of alleles, a signature of population expansion. This test is considered less conservative than Tajima's D and is more sensitive to large population expansions expressed as large negative numbers whereas positive numbers indicate populations impacted by genetic drift

I.5.1.3 Mismatch distribution

Another popular method in the past for valuating amount of variation at molecular data is the distribution of pairwise differences, also known as mismatch distribution. This method is applicable to molecular data where differences between alleles can be counted and includes nucleotide substitutions, RFLPs, VNTRs, or STRs. The mismatch distribution is constructed by counting the number of differences between each pair of mtDNA sequences and then using histograms or scatter plots to display the frequencies of sites that differ. This measure of diversity summarizes the discernible amount of genetic variation within a population. It was claimed that the shape of the mismatch distribution could be informative regarding the historical demography of a populations. Thus, a unimodal distribution is indicative of population expansion whereas a multimodal distribution indicates constant population size over a long time period. Mismatch distributions can however lead to naïve interpretations of human population history and have been criticized by several authors (Bandelt and Forster 1997)

I.5.1.4 AMOVA

Analysis of Molecular Variance (AMOVA), as applied to mtDNA sequences, takes into account the molecular relationship of alleles. AMOVA is analogous to a nested analysis of variance (ANOVA) derived from a matrix of squared distances among all pairs of haplotypes. This in turn produces variance estimates and F-statistic analogues designated as -statistics that reflect the correlation of haplotypic diversity at different hierarchical levels of population subdivision.

This allows for the hierarchical partition of the haplotypes into sum of squared deviations (SSD) within populations, SSD within regional groups, and SSD among populations within regional groups. The mean squared deviation (MSD) is obtained by dividing the corresponding SSD by the appropriate degrees of freedom. This method is appropriate for any data where genetic distances between alleles can be calculated. AMOVA can be used on mtDNA control region data to determine whether population structure was present

I.5.1.5 Interpolation maps

Interpolation maps just represent frequencies of (typically) haplogroups in geographic maps using different interpolation approaches. In the past, these maps were recurrently used by Luca Cavalli-Sforza to represent the main components of a principal component analysis (PCA). Nowadays, the interpolation maps usually are the representation of F_{ST} values (Brucato et al. 2010) or the frequency of specific haplogroups (Pala et al. 2009; Perego et al. 2009; Perego et al. 2010).

I.5.1.6 Phylogenetic trees

Phylogenetic trees not only contain information about the relationships between populations but also provide information regarding the time of divergence. Briefly, a tree consists of branches, nodes, and tips. The tips are extant representatives which are DNA sequences with different nucleotide substitutions. A node is the point at which one sequence diverged from another to form two branches (lineages). Generally, with molecular data, a mutational event creates a divergence. A branch represents a lineage between divergences, from the last divergence to the present. Several different methods of tree building using different assumptions exist and include: maximum likelihood (ML), maximum parsimony (MP), neighbour-joining (NJ), and unweighted paired group method (UPGMA). Statistical error in tree building can be high but can be somehow monitored by using bootstrapping methods, which may allow for an estimation of confidence limits.

I.5.1.7 Phylogenetic networks

An alternative to phylogenetic trees for certain types of molecular data (mtDNA RFLP, mtDNA control region, and male specific Y-chromosome STRs) are phylogenetic networks. Networks offer an advantage over traditional tree building methods that utilize maximum parsimony or maximum likelihood, because networks can distinguish between irresolvable and resolvable character conflict that may occur due to homoplasy. Networks represent 'all most parsimonious trees' by highlighting conflicts in the form of reticulations (equally possible mutation routes between nodes in the network) and interpreted as homoplasy, (artefactual) recombination, or sequence errors (Bandelt et al. 1995). The network is sequentially built by the addition of consensus points (median vectors) to three mutually close sequences at a time. These median vectors are then inferred as either extinct sequences or extant unsampled sequences within the population. Four different types of networks exist and include: minimum spanning networks (MSN), reduced median networks (RM), median joining networks (MJ), and quasi-median spanning networks (QSN).

I.5.2 Coalescence time

The coalescent theory aims to estimate the number of generations from a set of contemporary haplotypes to the most recent common ancestor (MRCA). The coalescent represents the state when all the lineages of a gene tree have been traced back to the point where only one lineage remains.

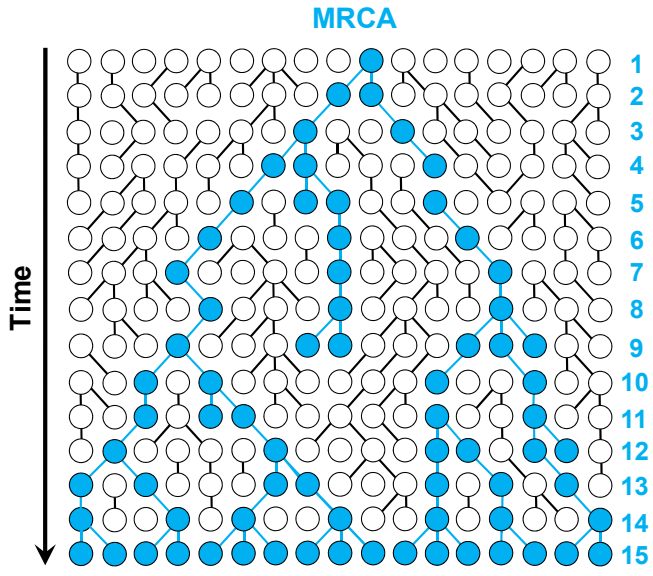


Figure 14 Schematic diagram showing coalescence process for mitochondrial DNA.

In coalescence, a gene tree is constructed by deducing the most likely pattern of sequential nucleotide substitutions that accounts for the differences among the DNA sequences. In this way two haplotypes can differ by one (or more) substitutions and have a common ancestor at time t_1. Thus, in tracing the branches backward, the coalescence time occurs when the number of branches in the tree is reduced to 1.

For mtDNA variation, the probability that two haplotypes are derived from the same ancestor is $1/N_e$ so the probability that they have different origins is $1\text{-}1/N_e$, where N_e is the effective population size. Calculation of the coalescent in a gene tree depends on the number and evolutionary relationships between contemporary haplotypes. Formally, the expected time in generations for a coalescence is $2N_e/(\kappa/2)$ where κ is the number of branches preceding the coalescence time multiplied by those at the coalescence time. Overall, the time to the coalescent in a gene tree is $2N_e(1\text{-}1/n)$, where n is the number of contemporary haplotypes. In practice, the determination of the coalescent requires an estimate of the population-mutation parameter (θ) (*see above*).

The data are analyzed by computer simulations using numerous models that take into consideration assumptions about gene flow, recombination, differential mutation rates at different DNA regions, bottlenecks, expansions, and varying growth rates.

The best fit between the data set and a particular model provides an estimate of the most likely time measured in N_e generations from the present to the MRCA. Finally, N_e generations can be converted in years by using an intergeneration time of e.g. 25.

I.5.3 Phylogeography

The term phylogeography, which means the phylogenetic analysis of biological data in the context of its geographic distribution, was first coined by (Avise et al. 1987). The original definition was “The field of study concerned with the principles and processes governing the geographical distribution of genealogical lineages, especially those at the intraspecific level”. This study aimed to unite evolutionary biologists in the disparate fields of phylogenetics and population genetics.

MtDNA is particularly suitable for phylogeographic analysis specially due to its lack of recombination, but also because its smaller effective population size (compared to nDNA). Phylogeography has been deeply applied to the analysis of mtDNA variation (Bandelt et al. 2001a; Bandelt et al. 2001b; Batini et al. 2007; Hickerson et al. 2010; Richards et al. 1998; Topf et al. 2006).

I.6. mtDNA VARIABILITY STUDY IN HUMAN POPULATION GENETICS

I.6.1 The role of climate and technological development in human dispersion

The climate in the Earth has been a constant switch from warm interglacial to cold glacial conditions, and back again. Actual distribution of modern non admixed human population is, at least in part, a consequence of those conditions since the last 200,000 years. Main migration events all over the world and also growth of the populations have taken place when the climate conditions were warmer.

A *stadial* is a period of colder temperatures during an interglacial (warm period) separating the glacial periods of an ice age. These periods are of insufficient duration or intensity to be considered glacial periods. Notable stadials include the *Oldest Dryas*, *Older Dryas* and *Younger Dryas* stadials and the *Little Ice Age*.

An *interstadial* is a warm period during a glacial period of an ice age that is of insufficient duration or intensity to be considered an interglacial. Generally, interstadials endure for less than ten thousand years and interglacials for more than ten thousand. The *Bølling Oscillation* and the *Allerød Oscillation* were the last interstadial periods before the present.

The names that receive the different glacial and interglacial periods as well as the dating vary depending on the sub-continental region (*see Table 5*)

REGION	GLACIAL 3 130-200 k.a.	INTERGLACIAL 3 110-130 k.a.	GLACIAL 4 110-10 k.a	INTERGLACIAL 4 10 k.a-
Alps	Riss	Riss-Würm	Würm	*Holocene**
North Europe	Saalian	Eemian	Weichselian	*Holocene**
British Isles	Wolstonian	Ipswichian	Devensian	*Flandrian*
Midwest U.S.	Illinoian	Sangamonian	Wisconsinan	*Holocene**

Table 5 Names of the two last glacial periods and the last interglacial period depending of the region. * Holocene is an epoch which started at the end of the last glacial period.

The *Eemian Stage*, which lasted from about 130,000 to 114,000 years ago, was the last interglacial prior to the present *Holocene epoch*. The end of *Eemian Stage* was a relatively sudden event and not a gradual slide into colder conditions taking many thousands of years. The warmest peak of the *Eemian Stage* was about 125,000 years ago and could be the period of a first "*Out of Africa*" of modern humans, via the *northern route* (*see below*), which failed maybe due to the climate conditions in the Middle East area change to a drier conditions too quickly.

Last Glacial Period began after the *Eemian Stage* from about 100,000 to 10,000 years ago. During the glacial periods, due to the volume of ice on land, sea level could be approximately from 80 to 130 meters lower than present (Lambeck et al. 2002), allowing the spread for more regions which were connected or nearer to land than actually, for example Sahul continent (see below). Near to the end of this period was the called *Last Glacial Maximum* (LGM) which is defined as the period when ice sheets had the maximum extension and was the responsible of important human migrations, especially in northern hemisphere, who retreated into refugial areas (see below). Also after LGM, took place multiple human migrations in several regions around the world. One of them was the re-expansion from the refugia in south of Europe to the rest of the mainland when the conditions were warmer. The phenomenon homologous in Africa was the *Last Glacial Aridity Maximum* (LGAM)

As well as climate conditions the development of lithic tools had played an important role in human modern dispersal and therefore the spatial patterns of genetic variation. These technological prehistoric periods are:

- *Paleolithic* or *Old Stone Age* (2,5 m.a-12 k.a), is divided into:
    - Lower Paleolithic (2,5 m.a.-300 k.a) that entails the evolution of several species of genus *Homo*, control of fire and first stone tools
    - Middle Paleolithic (300-40 k.a.), comprising the evolution of *Homo neanderthalensis* and the "Out of Africa" of modern humans (*Homo sapiens)*
    - Upper Paleolithic (40-10 k.a.) which involves abundant artwork and fully developed language.
- *Mesolithic* or Middle Stone Age (12-7 k.a), involves the development of microliths and bows
- *Neolithic* or New Stone Age (7-3 k.a.), involves the agriculture and farming in Europe
- *Chalcolithic, Eneolithic,* Bronze Age or Iron Age (3-2 k.a.), involves the first metal tools

I.6.2 Global mtDNA variability in modern humans

The global patterns of human mtDNA diversity suggest that modern human variation is broadly structured at the continental level, with South Asia and East Asia (and probably also Southeast Asia) forming genetic clusters distinct both from each other and from (Native) America, Australasia, West Eurasia and sub-Saharan Africa. This is the result of sequential colonisation and expansion from very small founder groups who dispersed from an East African homeland within the last 70 kya.

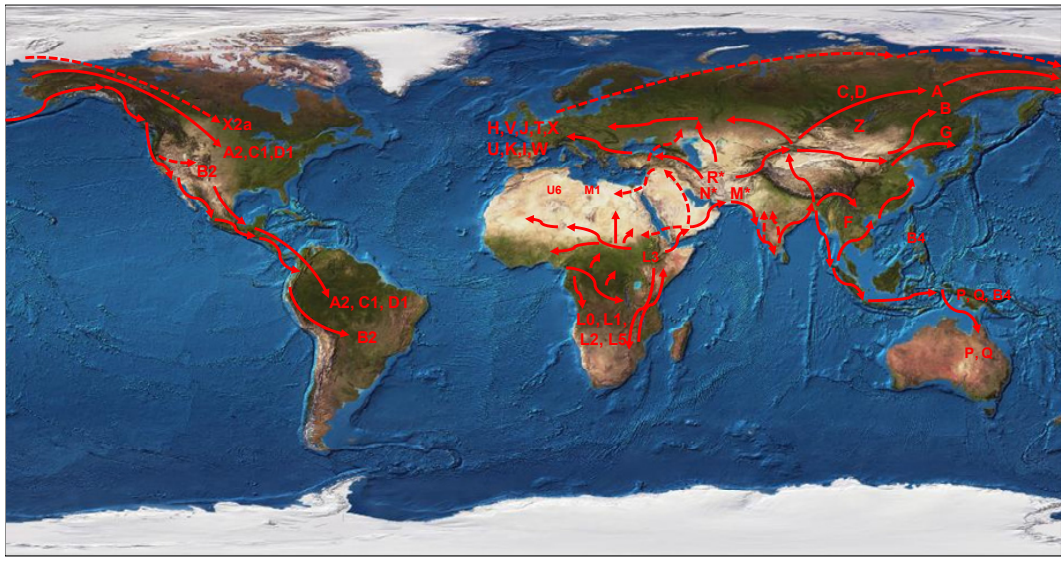


Figure 15 World human migrations with the main haplogroups. Arrows and haplogroup labels aim a rough indication of main migration routes and distribution of main continental haplogroups.

Briefly, haplogroup L, which is predominant in Africa, is the oldest of the human mtDNA lineages and contains the MRCA of all human populations. Haplogroup L can be divided into several macro-lineages (L0, L1, L2, and L3). Two other macro-lineages (M and N) diverged from L3, presumably in the Middle East or Southern Asia. The European haplogroups H, I, J, K, T, U, V, W, and X were subsequently derived only from N, whereas M and N contributed equally to the radiation of mtDNA into the Asian specific haplogroups A, C, D, G, Z, and Y. The Americas was populated from North-eastern Asia through the Bering land bridge mainly by humans belonging to haplogroups A2, C1, and D1, whereas haplogroup B2 may have arrived later and via a coastal route (*see Figure 15*).

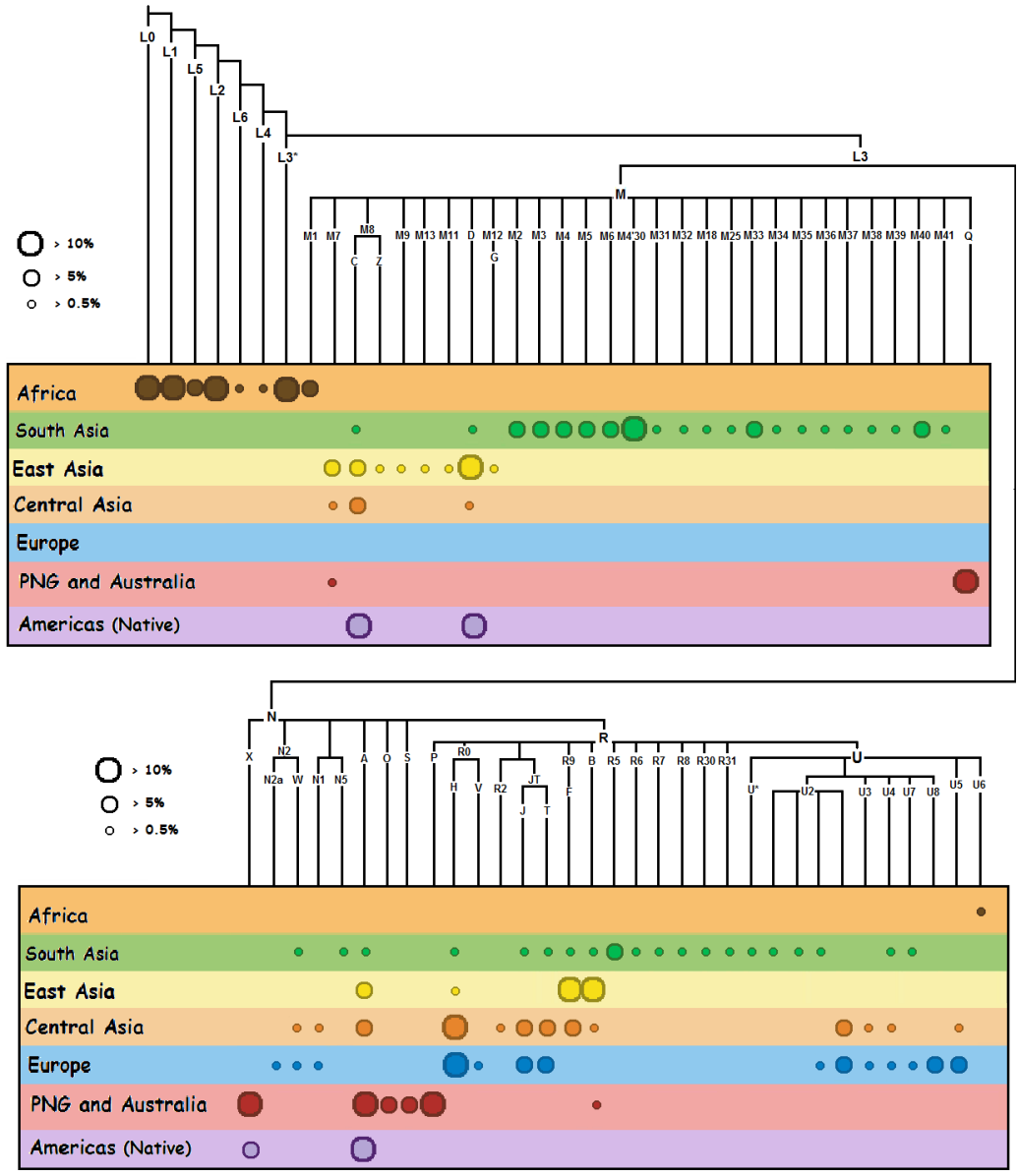


Figure 16 Distribution of global mtDNA. (Modified from (Chaubey et al. 2007))

I.6.2.1 Evolutionary history and mtDNA variability in African populations

Africa is the birthplace of *Anatomically Modern Humans* (AMH), who have lived continuously on the African continent longer than in any other geographic region. The high genetic diversity observed in sub-Saharan Africans is one of the most important arguments supporting the *Recent African Origin* model (also commonly referred to as the '*Out of Africa*' (*OOA)* model) that explains the origin and evolution of AMH. According to

the OOA model, our specie appeared somewhere in sub-Saharan Africa and colonized the Earth, displacing archaic 'human' populations in the course (*see Figure 17*). A modified version of the OOA model, the 'Weak Garden of Eden Hypothesis' suggests that populations may have remained small and subdivided for some time after the initial migration of modern humans out of Africa, followed by recent and rapid population expansion within the past 50,000 years (Harpending et al. 1998). This theory is against the Multiregional Origin (MRO) model of human origins which suggests that there was no single geographic origin for all modern humans. Advocates to the MRO model argue that after the expansion of *Homo erectus* from Africa into Europe and Asia 0.8 to 1.8 Mya, there has been a continuous transition among regional populations from *Homo erectus* to *Homo sapiens*. The MRO model was supported by the observation of regional continuity of certain morphological traits in the fossil record, which suggested that, they should have evolved over very long periods of time in the regions where they are found today.

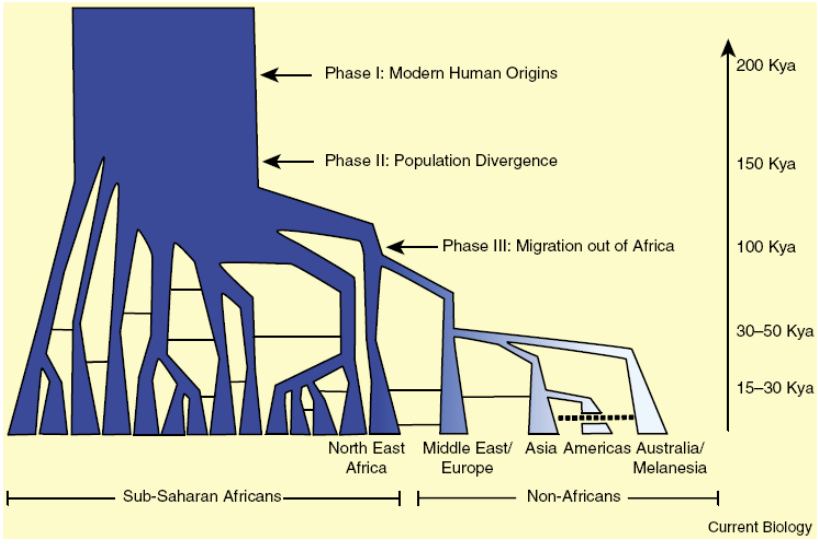


Figure 17 The Recent African Origin model of modern humans and population substructure in Africa (Taken from (Campbell and Tishkoff 2010))

An early support to the OOA model came from mtDNA studies indicating a relatively recent coalescence of human mtDNA, the topology of the mtDNA tree (all the non-African mtDNAs is a sub-tree that spread from sub-Saharan Africa), and also a seeming star-like shape of the non-African mtDNA tree, indicating a major pre-historical population expansion (Cann et al. 1987) (*see above Mitochondrial Eve*).

The OOA model is also supported by the fossil record, in which the earliest AMH fossils are from Africa and the Middle East. In the Levant, the earliest dated AMH were found at Qafzeh and Skhul in the Middle East at 90-115 kilo years ago (kya) and in Ethiopia at 195-154 kya. For these fossil records two major routes of dispersal have been hypothesized: one through North Africa into the Levant, called *northern route*, and one through Ethiopia along the tropical southern Asian coastline, called *southern route.* The last one is supported by the high number of basal mtDNA haplogroup R and M lineages in India (Sun et al. 2006) and by similarities between industries associated with modern humans in South Africa ~60 kya and South Asia at least 35 kya (Mellars 2006). Recently a study based on autosomal SNP has confirmed this hypothesis (Abdulla et al. 2009).

The variation we observed today in the African continent is the consequence of a number of pre-historical fluctuations in population size, migrations, admixture, etc, fuelled by technological innovations and different lifestyles such as e.g. hunter-gatherers vs agriculturalists (*see Figure 18*). East Africa is characterized by high levels of population differentiation and has been proposed as the birthplace of AMH. After that there was an early spread across Africa (with a coalescence date of 130 kya or more), followed by a major re-expansion (60-80 kya) which repopulated Africa and led to the out of Africa of modern humans (Forster 2004). Afterwards there have been several migrations within the continent.

One of the most important and recent events occurred in sub-Saharan Africa, has been the *Bantu expansion* which involved a great migration of Bantu speakers about 5 kya from Nigeria and Cameroon first to the rainforest of equatorial Africa and then into eastern and southern Africa (Forster 2004). There were other internal migrations like the *Nilo-Saharan expansion* around 8-10kya from Middle Nile Basin to Lake Chad or gene-flow between Nilo-Saharan speakers (predominantly central and eastern African pastoralist) and Afroasiatic speakers (predominantly northern and eastern African pastoralist and agropastoralist). Finally, other important demographic events in Africa were the various back-migrations coming from Middle and Near East into Africa which explain for instance the presence of well known African haplogroups of non-African origin (*see below, in Haplogroup M1 and Haplogroup U6 paragraphs*).

African mtDNA haplogroups are also present in America originally due to the Trans-Atlantic slave trade occurred only few hundred years ago, and this has been the main focus of several studies (Mendizabal et al. 2008; Salas et al. 2005b).

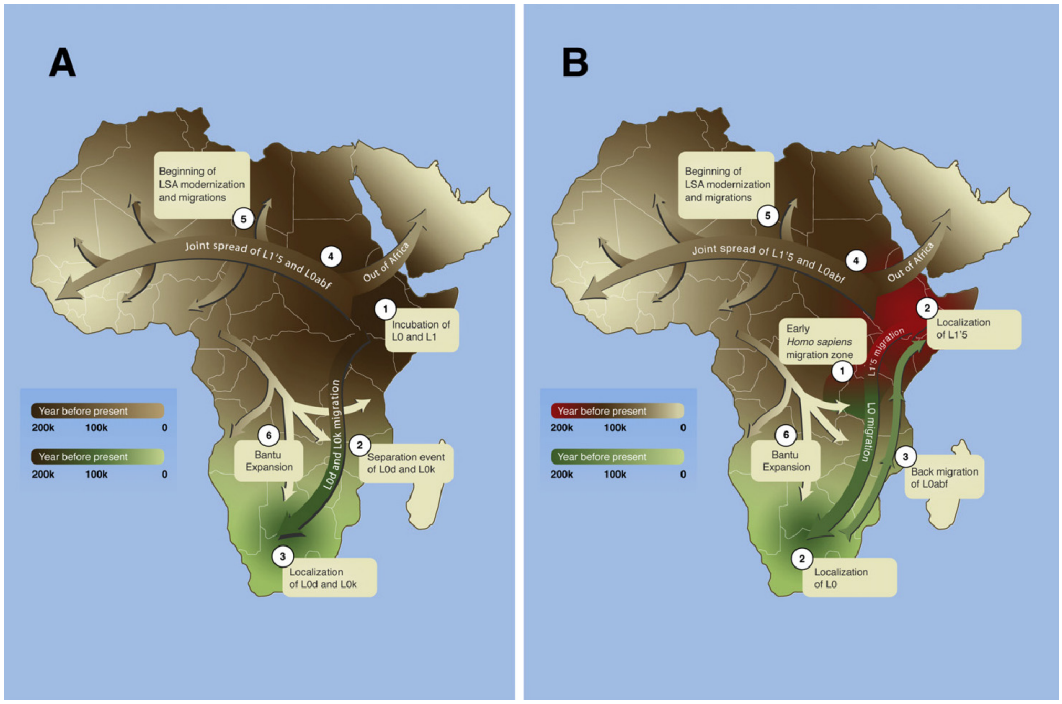


Figure 18 Two hypotheses of maternal gene flow between Africa. (A) An initial prolonged colonization (brown) by AMH (1) is followed by a dispersal wave (green) of a fracture of the population (2) and the localization of L0d and L0k to southern Africa (3) Later dispersal waves from the eastern African population parallels the beginning of African LSA approximately 70,000 ybp (4). Rapid migrations during the LSA (5) brought descendants of the eastern African population into repeated contact with the southern population, peaking during the Bantu expansion (6). (B) An early Homo sapiens division in a hypothetical migration zone (1) resulted in two separately evolving populations (2) and the localization of L0 (green) in southern Africa and L105 (red) in eastern Africa. A subsequent dispersal event of the L0abf subset from the southern population and its mergence with the eastern population (brown) is suggested (3), resulting in the former population composed only of L0d and L0k and the latter composed of L1'5 and L0abf. (Taken from (Behar et al. 2008b)).

Most of the present mtDNA haplogroups observed in sub-Saharan Africa are designated with "L" capital letter. The change of nucleotide at position 3594 (with respect to rCRS) separates only L-African specific haplogroups from all non African mtDNAs. This variation was early detected using *Hpa*I (RE from *Haemophilus parainfluenzae)* that has a restriction site at np 3592; all the L-African lineages that lack this polymorphism were designated as L3. Early studies had shown differences between African populations (Scozzari et al. 1988). However, the first attempt to classify African mtDNAs into haplogroups was carried out several years later (Chen et al. 1995). Although the tree in the Chen et al. study was rooted with an Asian sample as outgroup, the conclusion was that African mtDNA variation was ancestral to that present in Eurasia.

Since the haplogroup classification provided by those pioneering studies, several updates have been made to the labelling of the major and minor African haplogroups (*see Figure 20* and *Figure 22*). For example, nowadays the human mtDNA tree splits into two branches, one defined as haplogroup L0 and the other holding all of the rest of extant African and non-African mtDNA haplogroups.

I.6.2.1.1 Haplogroup L0

Haplogroup L0 is the earliest branch of the mtDNA tree in Africa. This appears as a *sister group* to the branch that holds all the other haplogroups found in the extant humans. It usually includes four sub-clades called L0a, L0d, L0f, and L0k (appears a fifth clade called L0b in (Behar et al. 2008b)) (*see Figure 20*). Depending of the authors the coalescence age of L0 and its sub-haplogroups can vary:

HG	(Salas et al. 2002)	(Gonder et al. 2007)	(Behar et al. 2008b)	(Soares et al. 2009)	(Schuster et al. 2010)
L0	150,000-170,000**	146,400 +/− 25,100	152,384 +/- 12,698	149,700	158,700-187,900
L0d	49,600 +/− 13,450	106,000 +/− 20200	101,589 +/- 10,318	—	107,200-129,800
L0k	—	70,900 +/−19,700	39,683 +/- 8,730	11,200	78,600-94,500
L0f	—	94,900 +/− 9,400	88,097 +/- 9,524	108,000	93,800-117,200
L0a	40,350 +/− 16,250	54,600 +/− 5700	53,176 +/- 7,143	44,800	61,100-71,400

Table 6 Coalescence ages for L0 haplogroup and sub-haplogroups depending on several studies. **The haplogroup L0 did not exit, this estimate includes L0 and L1 haplogroups.

L0a is common in East, Central, and Southeast Africa (~20%-25%) and is almost absent in North, West, and Southern Africa (Salas et al. 2002). L0a1 is the main sub-clade, it has a quite star-like phylogeny and a predominantly East/Southeast African distribution with their root type common in East Africa. L0a seems likely to have been brought to Southeast Africa into the Bantu community by the eastern stream of the Bantu expansion itself, having been picked up in East African non-Bantu speakers (Salas et al. 2002).

L0b is a "sister" clade of L0a, but without the transversion at position 16188. Some authors have designated it as L0g (Poloni et al. 2009).

L0d is one of the most ancient mtDNA lineages; it is common in click-speaking Khoisan populations (Bandelt and Forster 1997), and was identified at low frequency (5%) in Southeastern African samples (Pereira et al. 2001) and also in the click-speaking Sandawe population from Tanzania (Gonder et al. 2007; Tishkoff et al. 2007). L0d was subdivided into two reciprocally monophyletic clades: one clade composed of Southern African Khoisan and the other composed of Tanzanians (Gonder et al. 2007). Recently it has been found at high frequencies in South African Bantu speakers, suggesting more gene flow between Khoe-San and South African Bantu speakers than between Khoe-San and populations from Mozambique and Zimbabwe (Schlebusch et al. 2009).

As L0d, haplogroup **L0k** has been found, nearly exclusively among southern African "click" speakers (Quintana-Murci et al. 2010), although it can be detected at very low levels in other populations (Castri et al. 2009) likely due to gene flow between pre-existing Khoisan populations and Bantu populations (Salas et al. 2002).

Distribution and dating of L0d and L0k suggest their origin amongst ancestors of Khoisan (Salas et al. 2002). In 2008, an study concluded that modern distribution is due to a existence of a single AMH population which probably existed in eastern or southern Africa, so the localization of L0d and L0k to the southern part of Africa must be considered as the result of a population split followed by drift (Behar et al. 2008b). Another very recent study based on self-designed "coloured" people (people with a recent admixture) from South Africa showed that 60% of their mtDNA variability belonged to L0d haplogroup.

L0f is also among one of the most ancient haplogroups although it can only be found at low frequencies in East Africa (Gonder et al. 2007; Salas et al. 2002).

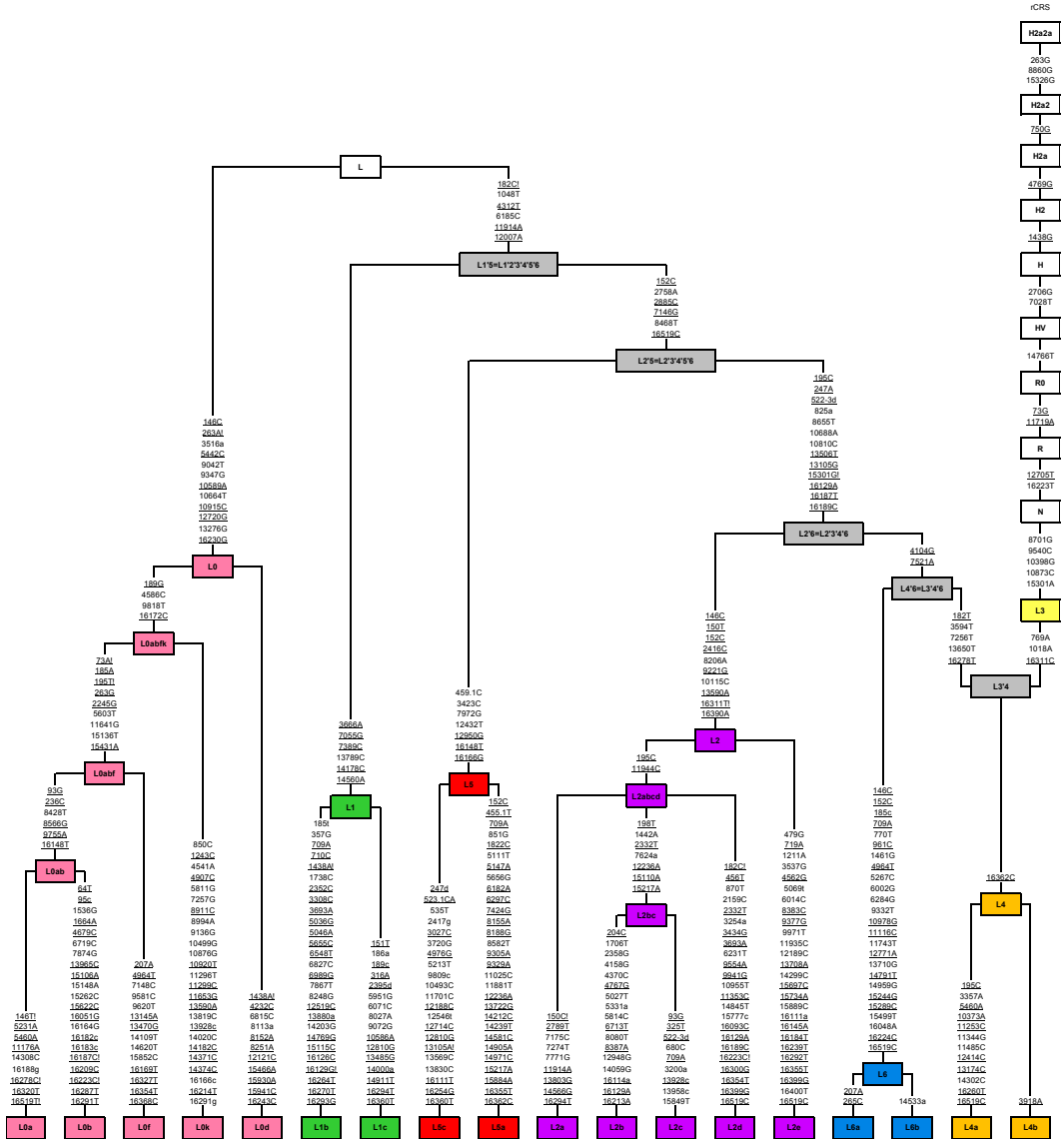


Figure 19 Human mitochondrial L African haplogroups phylogeny (except L3 and its derivates)

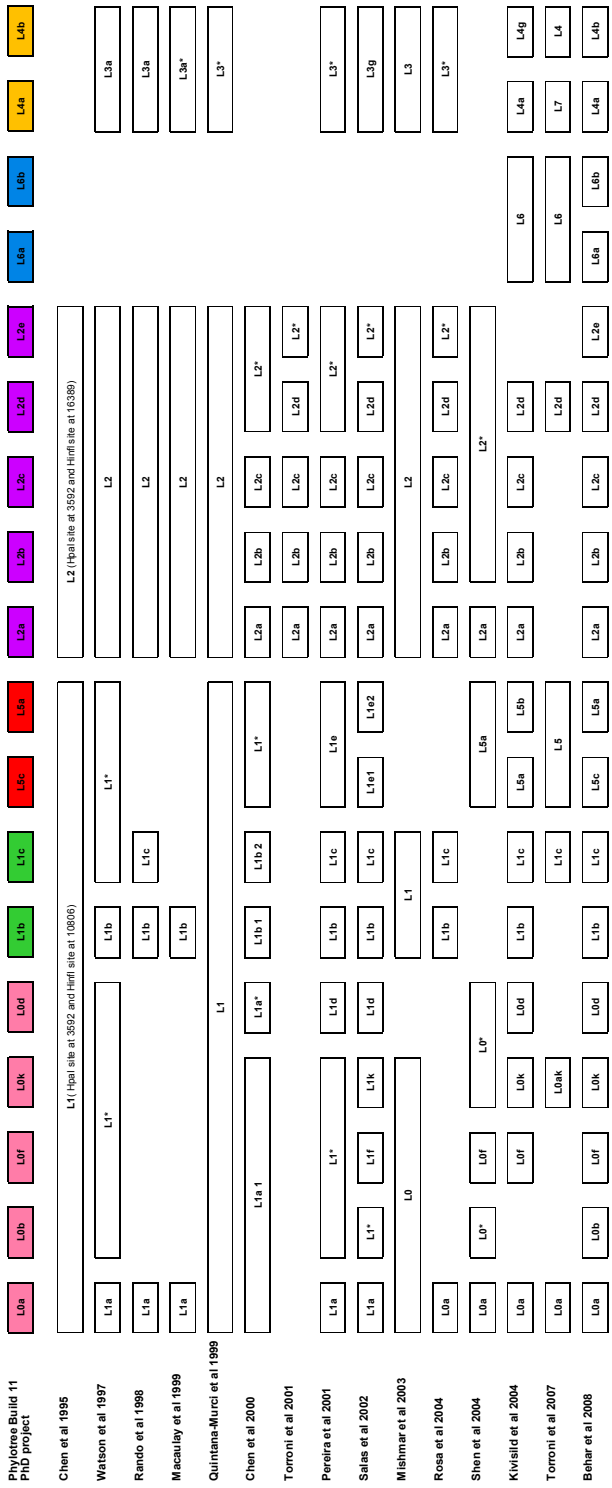


Figure 20 Evolution of human mitochondrial L haplogroups phylogeny (excluding L3). Taken and modified from (Bandelt et al. 2006)

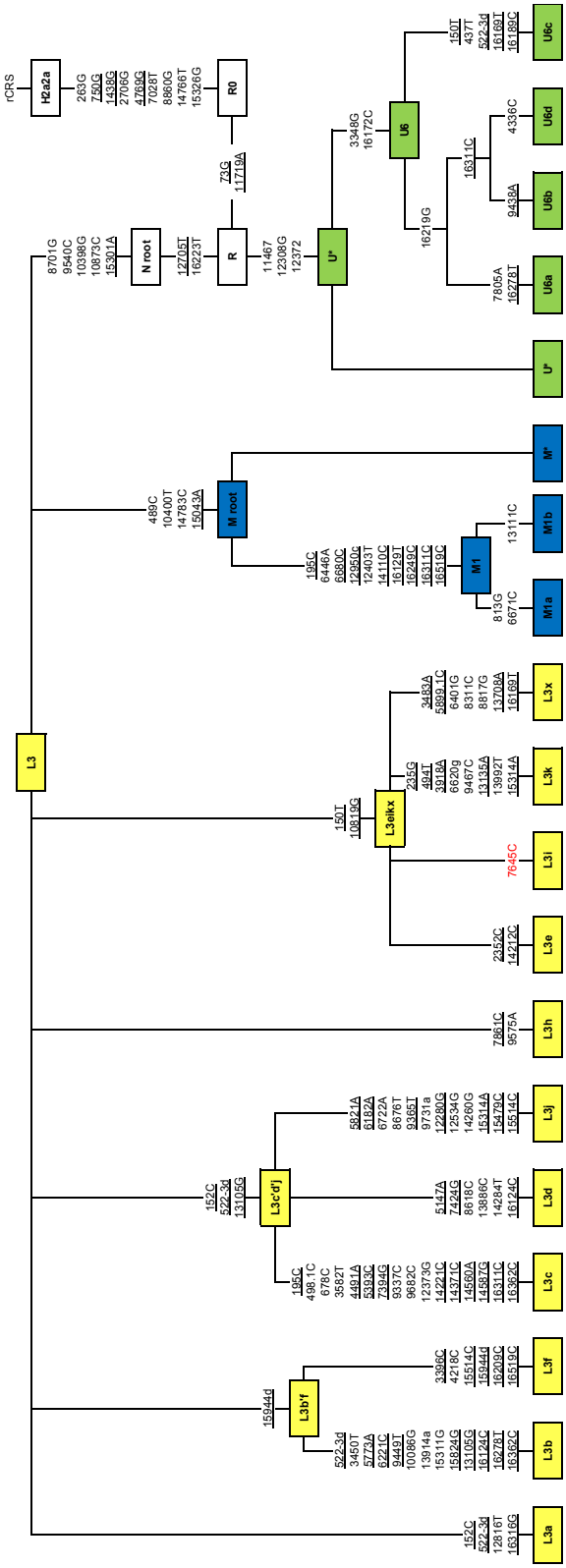


Figure 21 Human mitochondrial L3 and no-L African haplogroups phylogeny

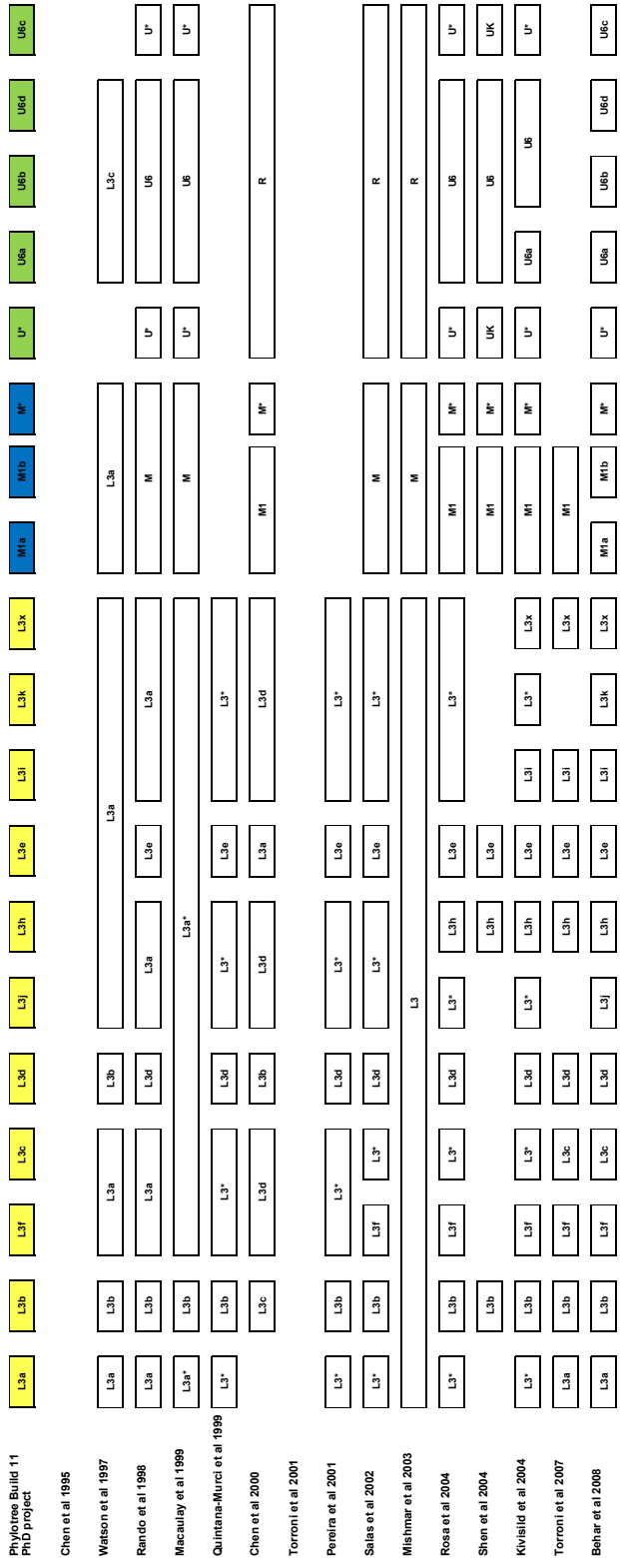


Figure 22 Evolution of human mitochondrial L3 and no-L African haplogroups phylogeny. Taken and modified from (Bandelt et al. 2006)

I.6.2.1.2 Haplogroup L1

Haplogroup L1 is more frequent and diverse in West and Central Africa than in East Africa (Salas et al. 2002). Together with L0, L1 mtDNA haplotypes represent the most basal lineages of the human mtDNA gene tree. L1 appears to be slightly more recent than L0.

HG	(Salas et al. 2002)	(Gonder et al. 2007)	(Behar et al. 2008b)	(Soares et al. 2009)
L1	150,000-170,000**	140,400 +/− 32,900	146,828 +/- 11,905	140,600
L1b	30,550 +/− 16,250	30,550 +/− 16,250	29,366 +/- 7,937	9,700
L1c	59,650 +/− 11,800	59,650 +/− 11,800	102,383 +/- 7,937	85,400

Table 7 Coalescence ages for L1 haplogroup and sub-haplogroups depending on several studies. **The haplogroup L0 did not exit, this estimate includes L0 and L1 haplogroups.

Sub-haplogroup **L1b** is concentrated in West Africa, with some introgression into Central and North Africa, particularly in geographically adjacent areas connected by the West African coastal pathway. It is also common in African-Americans, as a result of the Atlantic slave trade (Salas et al. 2008a). A western origin of L1b, with significant diffusion into North and Central Africa, has been hypothesized. Nevertheless, it is worth noting that their coalescence time is significantly lower than its sister clade L1c which likely originated in Central Africa. Consequently, the hypotheses of a more recent demographic scenario that shaped L1b phylogeny (e.g. bottleneck an re-expansion in West Africa), and an older Central African origin cannot be ruled out (Salas et al. 2002).

The origin of **L1c** haplogroup is still uncertain, although Central Africa is the most likely candidate source region (Salas et al. 2005b; Salas et al. 2002; Salas et al. 2004). It is however more difficult to account for the origin of the large proportion of L1c American lineages that do not find matches in present day Africa (Salas et al. 2008a). In 2007 (Batini et al. 2007) found distinct lineages within L1c, each with different evolutionary histories and proposed that L1c1b, L1c1c and L1c2 originated in Bantu ancestors whereas L1c1a, L1c4 and L1c5 (recalled L1c1a1a1 by (Quintana-Murci et al. 2008)) evolved among Western Pygmies. In order to avoid confusion L1c5 is skipped from the phylogeny. One year later another study focused also in the origin of Pygmies and neighbouring Bantu agriculturalists (Quintana-Murci et al. 2008). They also analyzed complete mitochondrial genomes belonging to L1c in order to improve the phylogeny and find more insights into their origin. Although the origin of L1c cannot be proven, they assign a Central African origin to L1c1a because of its distribution restricted to this region.

More recently, same author has published another study which comprises 205 complete mtDNA from ten central African populations (Batini et al. 2010). They did not find share maternal lineages between Western and Eastern Pygmy groups, and haplogroup L1c remaining as characteristic of the Western group.

I.6.2.1.3 Haplogroup L5

Haplogroup L5 (Kivisild et al. 2004), had been previously referred to as L1e (see Figure 20) occupies an intermediate phylogenetic position between L1 and L2. It has been observed at low frequency only in eastern Africa (Kivisild et al. 2004; Salas et al. 2002), the Sukuma of Tanzania (Knight et al. 2003), the Gurnas of Egypt (Stevanovitch et al. 2004) and among the Mbuti Pygmies in Congo (Kivisild et al. 2006). Haplogroup L5c (which sometimes is called L5b) has a spread more southern (Knight et al. 2003).

HG	(Salas et al. 2002)	(Kivisild et al. 2004)	(Gonder et al. 2007)	(Behar et al. 2008b)
L5	82,950 +/- 24,900	95,900 +/- 26,200	129,400 +/- 22,100	138,098 +/- 11,905
L5a	—	—	—	37,302 +/- 7,143
L5c	—	—	—	31,747 +/- 8,730

Table 8 Coalescence ages for L5 haplogroup and sub-haplogroups depending on several studies.

I.6.2.1.4 Haplogroup L2

L2 is common in western and South-eastern sub-Saharan Africa and it is the most numerous and widespread of the four major L-haplogroups; it encompasses one-quarter (Salas et al. 2002) to one- third of the sub-Saharan African mtDNAs (Torroni et al. 2001b). In contrast to L0 and L1, haplogroup L2 was a well-defined monophyletic haplotype group and traditionally has been divided in four sub-haplogroups called L2a, L2b, L2c and L2d; the fifth sub-haplogroup L2e appeared recently in (Behar et al. 2008b)) (*see Figure 20*).

HG	(Salas et al. 2002)	(Gonder et al. 2007)	(Behar et al. 2008b)	(Soares et al. 2009)
L2	70,100 +/- 15,300	94,500 +/- 4,500	104,764 +/- 8,730	89,300
L2a	55,150 +/- 19,350		46,033 +/- 7,937	48,300
L2b	31,600 +/- 11,200		26,985 +/- 3,968	14,300
L2c	27,500 +/- 7,250		23,810 +/- 3,175	25,200
L2d	121,900+/-34,200		23,810 +/- 7,143	—
L2e	—		46,826 +/- 7,143	—

Table 9 Coalescence ages for L2 haplogroup and sub-haplogroups depending on several studies

Haplogroup **L2a** is the most common and widely distributed South Saharan African haplogroup and is also frequent in the Americas (due to the slave trade). The wide distribution of L2a in Africa makes identification of its geographical origins more complicate. Moreover, some analyses indicated the occurrence of marked homoplasy at multiple sites in the control region (e.g. 16189, 16192, 16309) (Howell et al. 2004), which confound the phylogeny of L2a. Furthermore, only few complete genomes have been sequenced so far (Howell et al. 2004; Ingman et al. 2000; Torroni et al. 2001b).

Haplogroup **L2b** appears to be largely confined, together with the **L2c** and **L2d**, to West and West-Central Africa, where they possibly originated (Salas et al. 2002). Some derived types are present in Southeast Africa too. In the context of L2 haplogroup, clusters L2b and L2c are the most recent, while L2d is the oldest.

I.6.2.1.5 Haplogroup L6

This haplogroup appears for the first time in the literature in 2004 (Kivisild et al. 2004); it is the most frequent haplogroup in Yemenis with a coalescence age of 36,600 ± 23,400 years (*see Figure 20*). This haplogroup derives from the phylogenetic tree of sub-Saharan African mtDNA haplogroups but shows a very low frequency in Ethiopians and is completely absent elsewhere in Africa. Its high frequency in Yemen, together with low haplotype diversity, probably reflects the effect of genetic drift in a small founding population. Kivisild and colleagues proposed that L6 haplogroup could have been originated from the same out-of-Africa migration that carried haplogroups M and N. This would be accompanied by isolation of a southern Arabian population from the others in that region that would explain the absence of L6 types in other populations of the Near East, Arabia, and elsewhere in the world.

I.6.2.1.6 Haplogroup L4

These haplogroup has changed its nomenclature several times. Initially, it was included within L3 haplogroup (*see Figure 21*) and sub-haplogroup L4a was called L7 (Torroni et al. 2006) although it was reverted back to the original label as suggested in (Kivisild et al. 2004) and (Behar et al. 2008b). As explained by Behar and colleagues, both clades shared positions and similar distribution in East Africa and in southern West Eurasia and presented similar coalescence ages (Behar et al. 2008b), namely, 95,240+/-7,143 years.

I.6.2.1.7 Haplogroup L3

The youngest of the major haplogroups, L3, is most common in western and eastern/Southeastern sub-Saharan Africa, particularly among speakers of the Bantu language family, and is thought to have originated in eastern Africa, where it accounts for half of all types (Salas et al. 2002).

HG	(Salas et al. 2002)	(Gonder et al. 2007)	(Behar et al. 2008b)	(Soares et al. 2009)
L3	61,300 +/- 11,650		76,192 +/ -4,762	
L3a			65,081 +/- 9,524	
L3b	21,600 +/- 6,850	19,700 +/- 1,119	28,572 +/- 4,762	16,400
L3c		9,246 +/- 3,444	11,905 +/- 5,556	
L3d	30,250 +/- 8,450	27,109 +/- 1,850	38,096 +/- 5,556	31,000
L3f	36,300 +/- 12,800		60,319 +/- 6,349	53,100
L3h	—		70,636 +/- 5,556	66,700
L3e	49,250 +/- 11,750		44,445 +/- 5,556	39,000
L3i	—		45,239 +/- 9,524	34,800
L3k	—		38,096 +/- 7,143	—
L3x	—		38,096 +/- 5,556	36,100

Table 10 Coalescence ages for L3 haplogroup and sub-haplogroups depending on several studies.

Its most frequent sub-haplogroup is **L3b**, which appears at high frequency in West African, and in consequence also among African Americans (Salas et al. 2004). It has spread over into North Africa and into the Near East. There is very little dispersal into either East Africa or even Central Africa, but several derived types are present in Southeastern Africa (Salas et al. 2002).

Its sister clade haplogroup **L3d**, is also mainly in West African and African-Americans. A number of types are found in Southeastern Africa, including one type (in L3d1), matching a Fulbe lineage, at considerably elevated frequency. A second type (in L3d3) is not seen in our Southeastern African sample but occurs at high frequency in the south, in both Khwe and !Kung, and matches a type apparently found at high frequency in the Herero (Vigilant et al. 1991; not included in the network here because of sequence ambiguities). This likely arose in the Bantu population and spread later into the Khoisan speakers, since a single one-step derivative is present in the Southeast.

Both **L3f** and **L3g** are rare and also appear to have an East African origin. L3f* and L3g are virtually restricted to East Africa (with some dispersal into Central Africa, Southeastern Africa, and the Near East). Sub-clade L3f1, defined by transition at position

16293, hypothetically spread into West Africa at an early date, and is correspondingly well-represented in African Americans. The time to MRCA (TMRCA) for L3f1 has been estimated to be 28,650 +/− 8,650 (Salas et al. 2002).

Haplogroup **L3e** is the most widespread, and frequent of the African L3 clades. It hypothetically originated in Central Africa/Sudan region, about 45,000 years ago (Bandelt et al. 2001a). Bandelt et al. defined four main sub-clades (L3e1-L3e4); later, complete genome data allowed to identify a fifth sub-clade, called L3e5 (Behar et al. 2008b). Sub-haplogroup **L3e1** (32,150 +/− 11,450 years) is distributed throughout South-Saharan Africa, but is especially common in Southeast Africa, among Bantu speakers, while is rare in West Africa. L3e1 is of a likely West-Central African origin and it spread from there to Kenya through the eastern Bantu stream, and successively reached the Southeastern regions of Africa (Salas et al. 2002). Sub-haplogroup **L3e2,** is mostly distributed in Central Africa and West Africa with a coalescence age calculated to be 37,400 +/− 18,350 years (Salas et al. 2002). Sub-haplogroup **L3e3** is a small clade that is mainly distributed in Central, Eastern and Southeast Africa. Its root type is spread at elevated frequencies in Southeast Africa, together with some derivatives. A possible connection with the eastern Bantu stream has been hypothesized. The TMRCA has been estimated to be 14,150 +/− 4,500 years (Salas et al. 2002).

Haplogroup L3w was defined by (Kivisild et al. 2004) although it appears in (Behar et al. 2008b) under the denomination of L3i2. It was characterised by the HVS-I motif 16223-16260, and the transition 15388.

I.6.2.1.8 Haplogroup M1

At first, haplogroup M was regarded as an ancient marker of East Asian origin (Torroni et al. 1994) due to several expansion of the clade were found in India and Eastern Asia although was also found in Ethiopia later(Kivisild et al. 2004) . For its variation and geographical distribution an split between an Asian M haplogroups and eastern-African M haplogroups more than 50,000 years ago was suggested (Quintana-Murci et al. 1999). Because of the absence of haplogroup M in the Levant and its high frequency in the South-Arabian peninsula, M was considered to be the first genetic indicator for the hypothesized *southern route* (see above) of modern humans out of Africa (Quintana-Murci et al. 1999).

HG	(Olivieri et al. 2006)	(Gonzalez et al. 2007) (a)	(Gonzalez et al. 2007) (b)
M1	36,800 +/− 7,100	26,071 +/− 5,297	35,175 +/− 7,147
M1a	28,800 +/− 4,900	>16,756 +/− 5,997**	>22,607 +/− 8,901**
M1b*	23,400 +/− 5,600	19,040 +/− 4,917	25,689 +/− 6,633

Table 11 Coalescence ages for M1 haplogroup and sub-haplogroups depending on several studies. *Nomenclature is according (Olivieri et al. 2006) For example M1b subgroup was called M1c subgroup in (Gonzalez et al. 2007). **M1a in (Gonzalez et al. 2007) correspond to M1a1, for this reason the coalescence age must be ">". (a) coalescence with mutation rate from (Ingman et al. 2000) (b) coalescence with mutation rate from (Mishmar et al. 2003a)

M1 is present at high frequencies in the Horn of Africa and appears to be predominantly African specific. Complete mtDNA sequencing demonstrated that the presence of this lineage in Eastern Africa is the result of a back migration from Southwest Asia. In the same direction their presence also in Northwest Africa cannot be interpreted as a derivative of the East African mtDNA pool (Gonzalez et al. 2007; Olivieri et al. 2006).

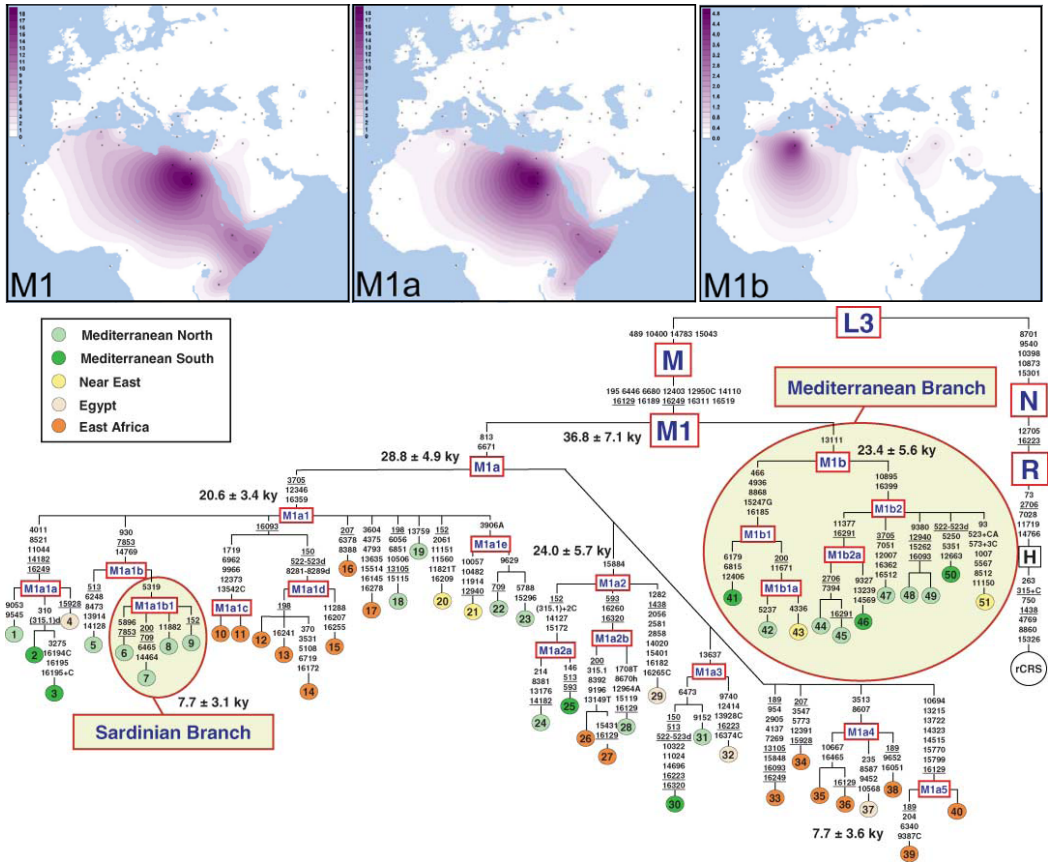


Figure 23 Spatial frequency distributions of haplogroup M1 and mtDNA tree of 51 samples belong to M1 haplogroup or sub-haplogroups (modified from (Olivieri et al. 2006))

I.6.2.1.9 Haplogroup U6

The samples which belong to this haplogroup were hypothesised like the result of a recent admixture from Africa and Asia (Richards et al. 1998). Following this and the previous study in Mozabites from Algeria (Corte-Real et al. 1996) all the samples which belong to U6 called the "*berber motif*"(Rando et al. 1998) due to the higher frequency between this population.

HG	(Olivieri et al. 2006)	(Maca-Meyer et al. 2003) (a)	(Maca-Meyer et al. 2003) (b)	(Maca-Meyer et al. 2003) (c)
U6	45,100 +/− 6,900	25,297 +/− 4,543	34,130 +/− 6,130	66,222 +/− 25,269
U6a	37,500 +/− 4,300	24,277 +/− 4,831	32,754 +/− 6,518	27,590 +/− 13,576
U6b	−	8,568 +/− 2,856	11,560 +/− 3,853	24,411 +/− 15,200
U6c	−	5,712 +/− 3,298	7,707 +/− 4,450	17,658 +/− 12,862

Table 12 Coalescence ages for U6 haplogroup and sub-haplogroups depending on several studies. (a) coalescence for coding region data with mutation rate from (Ingman et al. 2000) (b) coalescence for coding region data with mutation rate from (Mishmar et al. 2003a) (c) coalescence for HVS-I data

It has been proposed that U6 lineages, mainly found in North Africa although also in eastern Africa, are the signatures of a return to Africa around 39–52 kya, like M1 haplogroup (see above) despite the fact that M1 is more common in East Africa than in North Africa(Maca-Meyer et al. 2003). Several studies which carry out complete genomes sequencing are helping to determine their phylogeny and the possible migration events (Olivieri et al. 2006; Pereira et al. 2010).

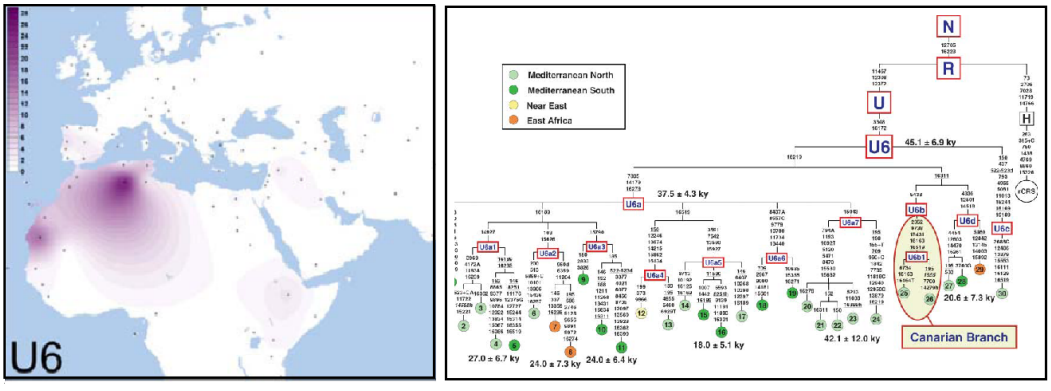


Figure 24 Spatial frequency distributions of haplogroup U6 and mtDNA tree of 30 sequences belong to U6 haplogroup or sub-haplogroups (modified from (Olivieri et al. 2006)).

I.6.2.2 Evolutionary history mtDNA variability in South Asia

South Asia usually comprises India, Pakistan, countries in the sub-Himalayan region and Myanmar although Iran and Afghanistan can also be included. This area was one of the first regions with modern human settlements out of Africa and served as major route of dispersal to other geographical regions along the *southern coastal route*, most likely after 50 kya (Macaulay et al. 2005) or according to a more recent study around 55-70kya (Soares et al. 2009), when much of Island Southeast Asia formed part of the mainland as the Sunda continent (*see Figure 25*). These dates are earlier than the earliest widely accepted archaeological evidence and seem to exclude the possibility that the dispersal into South and Southeast Asia took place before the volcanic eruption of Toba in Sumatra ~74 kya. However this question is under debate (Balter 2010).

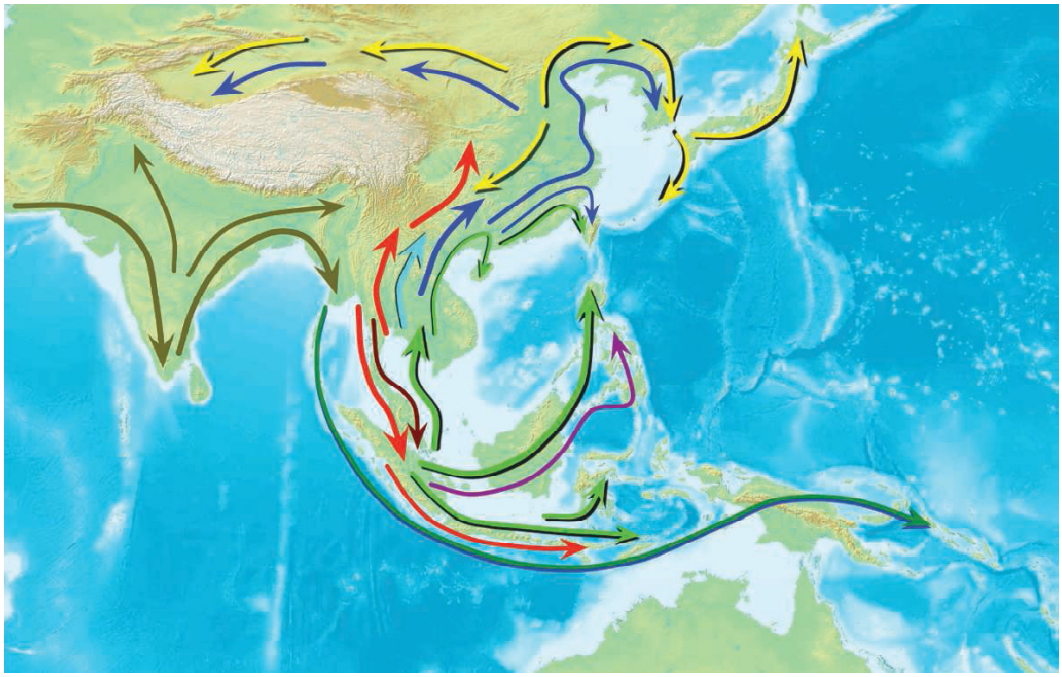


Figure 25 Genetic diversification of humans after migration from South Asia coast and split into numerous genetically distinct groups that moved across Southeast Asia (Sunda continent) and Sahul and also migrated north into East Asia. (Modified from (Normile 2009)).

According the post-Toba expansion theory, the dispersal occurred extremely rapid, within the range of few thousand years, because there is a divergence with different basal mtDNA lineages in each region, rather than a nesting-like phylogeographic pattern. mtDNA diversity suggests that East Asia could be initially settled from South to North (Macaulay et al. 2005), this point has been also confirmed with nDNA (Abdulla et al. 2009; Normile 2009).

There are diverse language and cultural groups within India. According to the language structure, *Dravidian* speaking groups populated southern India, *Indo-European* speakers settled northern India, *Tibeto-Burman* speakers were confined to Northeastern India, while *Austro-Asiatic* speakers moved to fragmented areas of eastern and Central India. On the other hand, culturally most of the inhabitants of India belong to either *tribal* or *caste* societies. There are studies indicating that there is virtually no exchange of genes between tribal populations or between caste and tribal population and there is little exchange between castes (Majumder 2010).

All the mtDNA haplogroups present in Indian or other Asian populations are M or N lineages which are derived from L3 lineage now found only in Africa (Majumder 2010). The sub-branching of haplogroup M in South Asia (mainly in India) is deeply different from any other Asian locality. More than 60% of Indians have their maternal roots in Indian-specific branches of haplogroup M while typical Asian sub-clusters C, D, E and G are found at extremely low frequencies (Kivisild et al. 1999b).

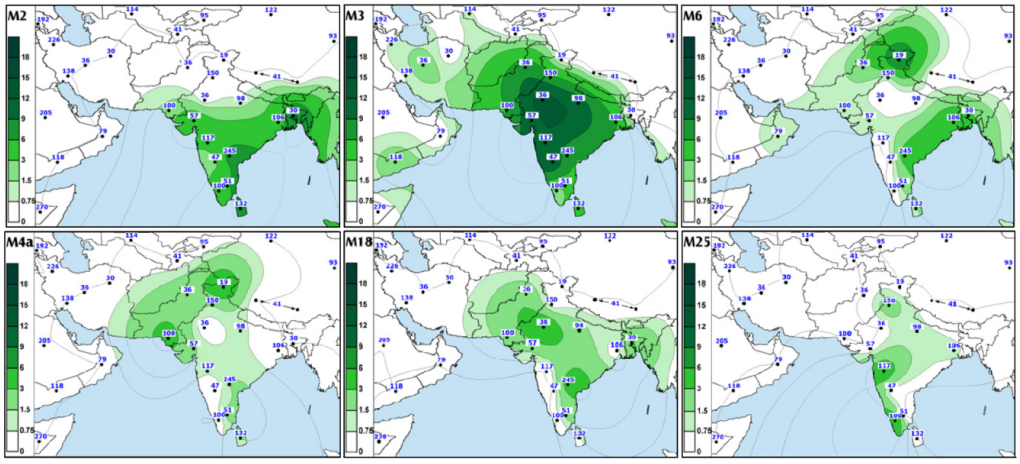


Figure 26 Distribution of M2 M3 M4a M6 M18 and M25 mtDNA lineages in South Asia. Taken from (Metspalu et al. 2004)

Indian specific M sub-clades identified so far are: M2, M3a, M4a, M5, M6, M18, and M25 (Metspalu et al. 2004). The Austro-Asiatic speakers (who are exclusively tribal) show the highest frequencies of M lineages and also the highest levels of M2 with the highest nucleotide diversity. This data seems to indicate that this population could be the modern representatives of earliest settlers of India (Basu et al. 2003; Kumar et al. 2008). There are others M-lineages close to the M-basal node; some of them have been defined with complete genome data and called M30 and M31 (Rajkumar et al. 2005) M53-M64

(Chandrasekar et al. 2009). Based on this phylogenetic distribution, it has been hypothesized that some M lineages originated in southwest Asia but probably arose not long from the time of first settlement (Chandrasekar et al. 2009; Kivisild et al. 2003; Kumar et al. 2008).

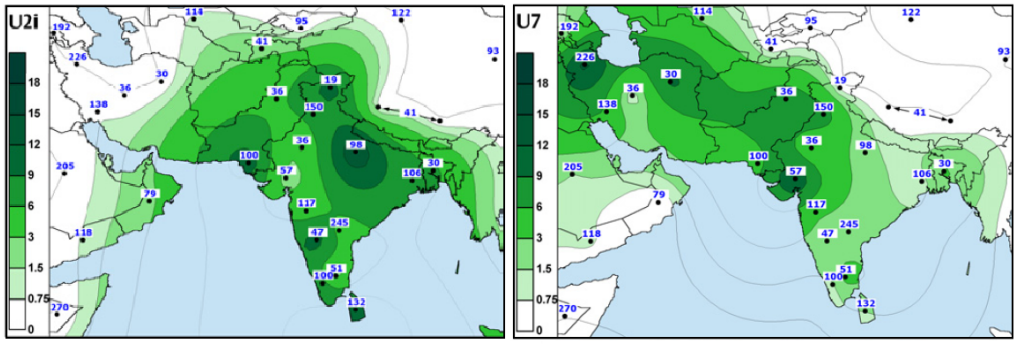


Figure 27 Distribution of U2i and U7 mtDNA lineages in South Asia. Taken from (Metspalu et al. 2004)

Among N lineages, more than 20% of the South Asian mtDNA lineages belong to western-Eurasian specific haplogroups H, I, J, K, U and W. The most frequent sub-clade of R in India is haplogroup U, which reaches 15% among the caste and 8% among the tribal population. There are Indian-specific variants of U, such as U2i (*see Figure 27*) which can reach more than 75% among all U mtDNA Indian lineages (Kivisild et al. 1999a). Another subset of U, haplogroup U7, is also present in India (Kivisild et al. 1999a). It has been found also in Iran (*see Figure 27*) and in some European and Siberian populations (Derbeneva et al. 2002; Richards et al. 2000).

The remaining 19% of South Asian lineages belong to N derivates, most of them to haplogroup R. It was not until 2004 that several of these Indian lineages could not be distinguished from the ancestral node R (Palanichamy et al. 2004). Using complete genome data, Palanichamy and colleagues defined sub-haplogroup R7, R8, R30, and R31 while the definition of R5 and R6 was improved.

I.6.2.3 Evolutionary history and mtDNA variability in East Asia

East Asia has been defined as the geographic region bordered by the Ural Mountains in the West, by the Himalayan Plateau in the southwest, by the Bering Strait in the Northeast, and extending into island Southeast Asia (Stoneking and Delfin 2010).

Genetic diversity suggests that East Asia could be initially settled from the South, and also it indicates discontinuities in the prehistory of East Asia. These discontinuities could be due to subsequent re-dispersals, which may be in part due to Neolithic expansions, but also reflect the more recent expansion of Han Chinese people (1.5 kya). The impact of the LGM is also likely to have been severe in continental East Asia, whereas several refugia existed within Southeast Asia. Sea-level rises beginning ~19 kya had their maximal impact, however, in Southeast Asia, Sunda continent was inundated leading to wide scale dispersals of lineages across actual islands in this area. This event could have had a greater demographic impact than the subsequent Holocene spread of the Neolithic across Southeast Asia and into the Pacific islands (Soares et al. 2008).

The two East Asian haplogroups derivates from super-haplogroup N (A and N9) and are predominantly of northern Asian provenance, while those of the super-haplogroup R (B, R9) are more frequent in southern East Asia (*more details see Figure 28*).

The super-haplogroup M is also more frequent in northern than in southern East Asia, but at the sub-haplogroup level some clades have opposite distributions. For example, M7 and E, are largely specific to mainland and island Southeast Asia (Ballinger et al. 1992) while others such as G, M8 (including C and Z), and the most frequent M subhaplogroup D, are much more frequent in northern East Asia. Haplogroups C and D are co-dominant in southern Siberia (Derenko et al. 2003) while C and G are more frequent in Northeastern Siberia (Tanaka et al. 2004). However, the mtDNA pools of northern and southern East Asia overlap, and the haplogroups that are most frequent among the Siberian populations also amount to one-quarter of the Southeast Asian mtDNA pool. In turn, the Southeastern haplogroups (excluding E, which is absent) take a notable share of the East Asian-specific mtDNAs in Central Asia and southern Siberia (*more details see Figure 28*).

Complete and partial mtDNA coding region sequences have been used to determine the fine-structure of the mtDNA lineages present in Asia (Kivisild et al. 2002; Kong et al. 2003; Yao et al. 2002a; Yao et al. 2002b). In 2004 the analysis of complete mtDNA sequences from 672 Japanese individuals provided a significant refinement of the East Asian mtDNA phylogeny (Tanaka et al. 2004). In 2007, to elucidate the human colonization process of northern Asia and human dispersals to the Americas complete genome sequencing of 71 individuals was published (Derenko et al. 2007). In this study, were identified new haplogroups such as I4, N1e, G1c, M7d, M7e, and J1b2a and the phylogeny of A, D2, G1, M7, I, N1a and G1b was improved.

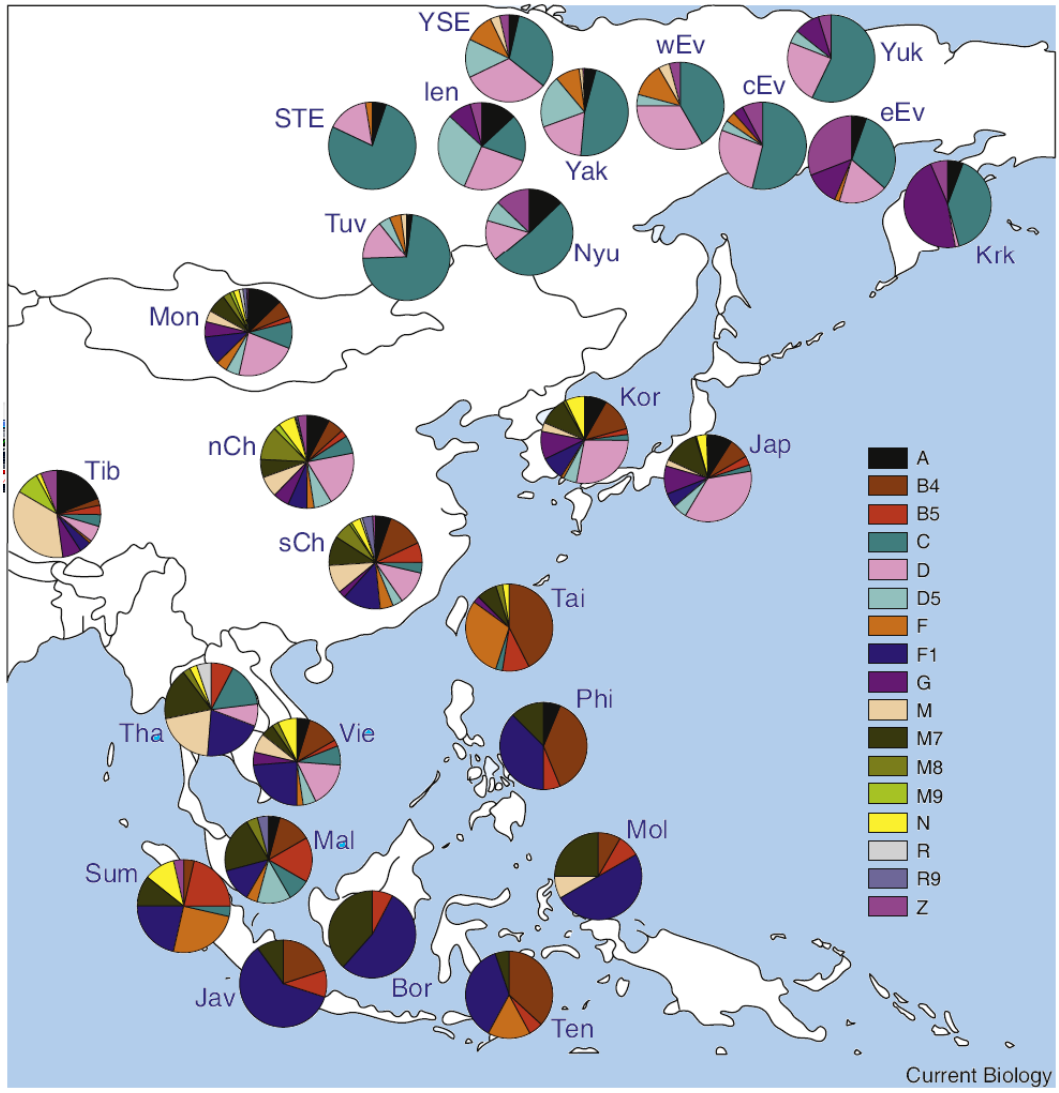


Figure 28 Distribution of major mtDNA haplogroups in East Asia. Abbreviations are: sCh, southern China (Xue et al. 2008); nCh, northern China (Xue et al. 2008); Tib, Tibet (Jin et al. 2009; Qian et al. 2001; Wen et al. 2004); Mon, Mongolia (Cheng et al. 2008; Jin et al. 2009); wEv, western Evens (Pakendorf et al. 2007); cEv, central Evens (Pakendorf et al. 2007); eEv, eastern Evens (Pakendorf et al. 2007); STE, Stony Tunguska Evenk (Pakendorf et al. 2007); Ien, Iengra (Pakendorf et al. 2007); Nyu, Nyukzha (Pakendorf et al. 2007); YSE, Yakutspeaking Evenks (Pakendorf et al. 2007); Yak, Yakuts (Pakendorf et al. 2007); Yuk, Yukaghirs (Pakendorf et al. 2007); Krk, Koryaks (Pakendorf et al. 2007); Tuv, Tuvans (Pakendorf et al. 2007); Kor, Korea (Jin et al. 2009); Jap, Japan (Jin et al. 2009); Tai, Taiwan (Kayser et al. 2008); Vie, Vietnam (Jin et al. 2009); Tha, Thailand (Jin et al. 2009); Mal, Malaysia (Trejaut et al. 2005); Sum, Sumatra (Kayser et al. 2008); Jav, Java (Kayser et al. 2008); Bor, Borneo (Kayser et al. 2008); Ten, Tengarras (Kayser et al. 2008); Mol, Moluccas (Kayser et al. 2008); Phi, Philippines (Kayser et al. 2008). Modified from (Stoneking and Delfin 2010).

It is necessary to take account the fact that several hypothesized aspects of human evolution in Asia based on mtDNA data (also Y-chromosome data) are recently confirmed by autosomal analysis (Abdulla et al. 2009).

I.6.2.4 Evolutionary history and mtDNA variability in West Eurasian population

In contrast to other main continental populations, West Eurasians have a moderate amount of haplogroup diversity within mtDNA haplogroups N, and lack haplogroup M, which is otherwise dominant in Asia. The colonization of West Eurasia was the result of a branch of the colonization along the *southern route*, followed by a lengthy pause, perhaps at the Persian/Arabian Gulf, until the climate improved and the area changed to the called *"Fertile Crescent"*. Under these better environmental conditions the ancestors of West Eurasians were able to enter first into the Levant and then into Europe. Paleo-environmental evidence is crucial to this argument, suggesting that an earlier migration from sub-Saharan Africa toward North Africa >50 kya and then to Europe would have been impossible given the climate of the time, with desert conditions extending from North Africa to Central Asia, rejecting the *northern route* (*see above*).

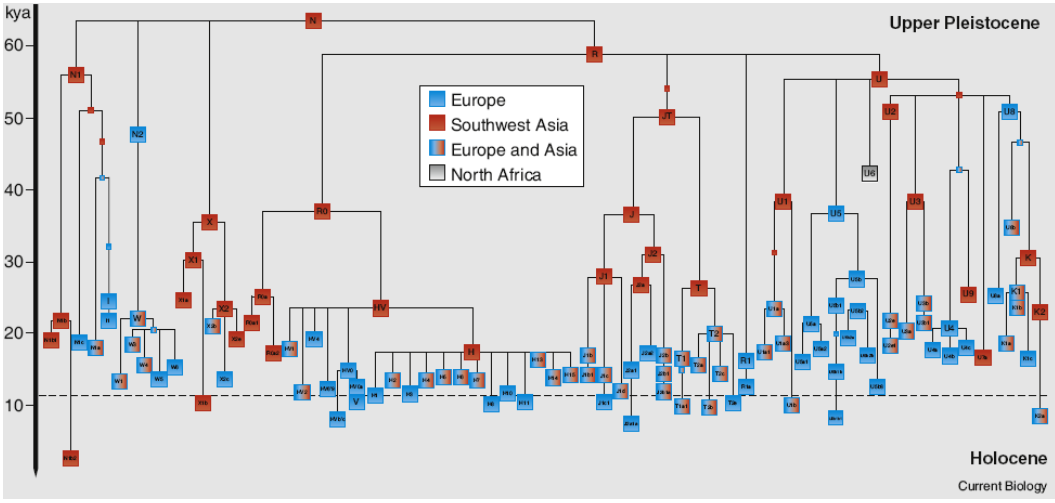


Figure 29 Phylogenetic tree of most common human mtDNA haplogroups in West Eurasian and North African populations. (Modified from (Soares et al. 2010)

A great number of studies have been focused in several aspects of peopling of West Eurasia, such as the pioneer colonization during Palaeolithic with their specific mtDNA haplogroups (Olivieri et al. 2006; Torroni et al. 2001a), the spread of several mtDNA haplogroups from the refugia areas after the LGM (Achilli et al. 2004; Pereira et al. 2005; Torroni et al. 2001a), and the demographic impact of agriculture in Europe.

The most ancient mtDNA lineages which have originated in Europe belong to haplogroup U5 (30 kya) and U8 (50 kya), which could represent initial settlements of Europe, while the other R derivates basal clades R0 and JT appeared in Europe more recently.

During LGM, human populations moved from North and Central Europe to refugial areas mainly located in south Europe. After this period there were several re-expansions and re-settlements of North and Central Europe. Distribution of several extant subhaplogroups like H1, H3 and H5 could reflect some of those dispersal routes back after the Ice Age (Achilli et al. 2004; Alvarez-Iglesias et al. 2009; Pereira et al. 2005; Torroni et al. 2001a)(*see Figure 30*), although some authors suggest than "no experimental evidence was found to support the human refuge-expansion theory" (Garcia et al. 2010). Another evidence of those events could be the un-expected links (especially within haplogroups V and U5b1b1) between several populations like Saami and Berbers (Achilli et al. 2005; Soares et al. 2010). It is important to take into account that although most of mtDNA variability involves nine main haplogroups (H, I, J, K, T, U, V, W, and X), more than 40% belong only to H haplogroup. For this reason, several studies have been focussed in the analysis of evolution and dissection of H haplogroup (Alvarez-Iglesias et al. 2009; Brandstatter et al. 2006; Brandstatter et al. 2008; Loogvali et al. 2004).

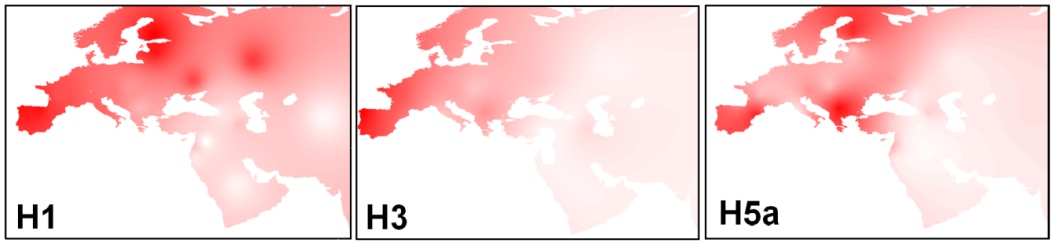


Figure 30 Distribution of H1, H3 and H5a lineages in western Eurasia. Modified from (Alvarez-Iglesias et al. 2009).

After a warmer climate during Mesolithic the style of life of European humans could start to change; thus, gathering and fishing became more important and coastal communities became more sedentary. Under similar conditions communities from Near East also became more sedentary but adopted cereal agriculture. Researchers have long debated whether this agricultural explosion was a demic diffusion which sparked by massive migrations of farmers themselves, or by the spread of farming ideas, known as cultural diffusion (Balter 2009). For this reason understanding how agriculture came to Europe has been the aim of several studies which mtDNA was evolved.

Founder analyses of mtDNA in Europe suggest that less than 15% of European lineages were contributed from the Near Eastern Neolithic component (Richards et al. 2000; Torroni et al. 1998) although Y-chromosomal seem to indicate the opposite trend (Balaresque et al. 2010). The explanation proposed for these results is that although demic diffusion involves both females and males, the disparity between mtDNA and Y-chromosomal patterns could arise from an increased and transmitted reproductive success for male farmers compared to indigenous hunter-gatherers, without a corresponding difference between females from the two groups. Possible mitochondrial candidate lineages for Neolithic dispersal could be J2a1a and K2a (Soares et al. 2010).

Latest year a study focused on N1a haplogroup in Europe as result of spread of agriculture was published (Palanichamy et al. 2010) Under their results, they suggested that the Neolithic transition process was more complex in central Europe and possibly the farmer N1a lineages were a result of a 'leapfrog' colonization process (Palanichamy et al. 2010).

I.6.2.5 Evolutionary history and mtDNA variability in Native Americans

The colonization of the Americas represents the most recent major human occupation of an uninhabited land mass on the planet. This process created a strong population bottleneck, and it leads to significant reductions in genetic variation in Native Americans relative to other global populations. Subsequent strong events of genetic drift followed the initial colonization all the way through the double continent. The (native) mtDNA variability in the Americas actually encompassed five major founder haplogroups A2, B2, C1, D1 and X2a and some minor ones (Achilli et al. 2008; Perego et al. 2009).

Several aspects of the colonization process, such as number of migration events, size of founding population or timing and routes of colonization, are still unclear. A Northeast Asian source population for the Americas, most likely around the Lake Baikal region, is widely accepted based on mtDNA and Y chromosome data. The number of migrations has been under debate since 1991, when four founder lineages were interpreted based on 63 sequences from an Amerindian tribe (Ward et al. 1991). In 1993, those four founder lineages were interpreted as four migrations (Horai et al. 1993). In another study these founder lineages were the first to be designated according the present rules of haplogroup nomenclature as A, B, C, and D (Torroni et al. 1993). Most recent studies indicate a single migration (Fagundes et al. 2008; Mulligan et al. 2004; Wang et al. 2007), although a recent one indicates that initial settlement was mediated by at least a dual migration (Perego et al. 2009).

It is still unclear if the colonization of *Beringia* (land mass which connected present day Alaska and eastern Siberia during several Pleistocene ages due to the sea level) and the first settlements into the Americas were prior or immediately after to LGM. Current coalescence estimates based on variation in extant mtDNA lineages set the event at 25 to 20 kya (Forster 2004) or less than 20 kya (Perego et al. 2009; Perego et al. 2010; Schurr and Sherry 2004).

Another important question for the early colonization of America concerns the availability of routes that would be open for the southern migration. Two such routes have been considered, the first giving an interior access to North America (the so-called 'ice-free corridor'), and the other an access through the coast. In relation to the former route, the corridor probably remained closed by the Laurentide and Cordilleran ice sheets until about the end of the Younger Dryas; while the Northwest coast route could have been opened earlier, by 14.5 kya (*see Figure 31*). Another issue opened to debate regards to the relationship of the first settlements and the Clovis technology.

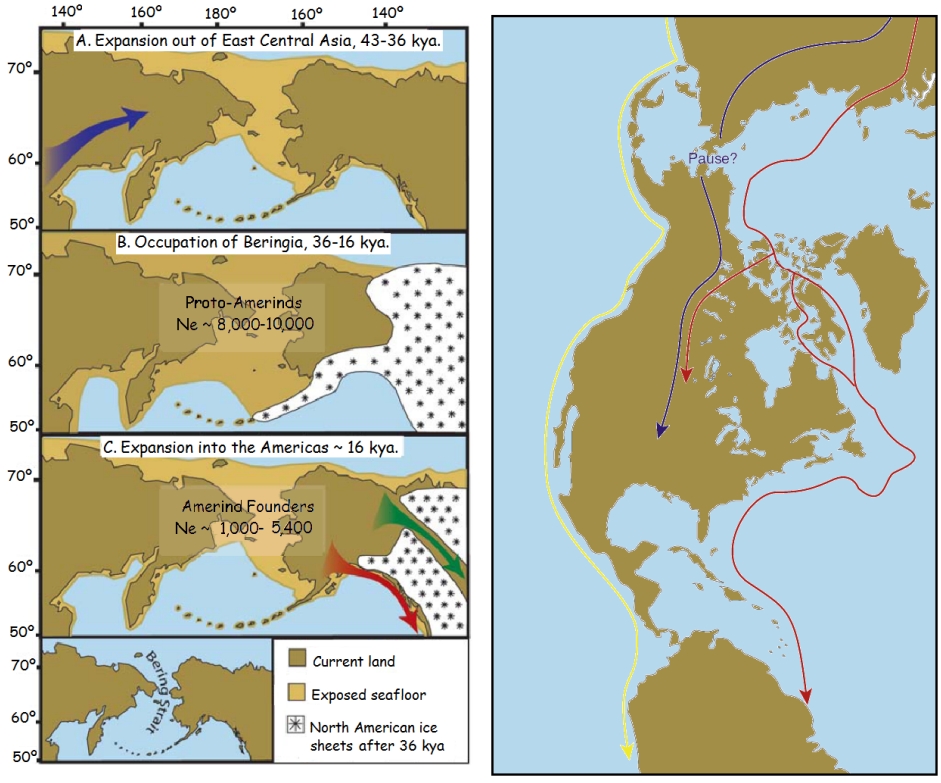


Figure 31 Hypothesised routes of migrations into the Americas. To the left, there is a representation of three stage colonization model proposed by (Kitchen et al. 2008). Rightwards appears the routes hypothesised by (O'Rourke and Raff 2010).(Modified from (Kitchen et al. 2008; O'Rourke and Raff 2010)).

Last year, another study concerning to the initial peopling of the Americas was published (Perego et al. 2010). The authors found that one branch of C1 (C1d), which has been dating as more recent than C1b and C1c, in fact was underestimated. The study is a great example of the application of complete genome sequencing in order to reach the maximum level of resolution and for the establishment of the different mtDNA sequences that might participate in the migration process.

The mtDNA pool of the Americas has completely changed in Colonial and post-Colonial times with the arrival of Europeans and African slaved that were forced immigrants to the continents. There are differences on the proportion of these lineages between American countries due to differences in their own history and the differential gender contribution has also been analyzed in detailed (Alvarez-Iglesias et al. 2007; Mendizabal et al. 2008; Salas et al. 2008a; Salas et al. 2005b; Salas et al. 2008b; Salas et al. 2004; Salas et al. 2005d).

I.6.2.6 Evolutionary history and mtDNA variability in Australia and Oceania

The peopling of the Pacific is particular as for that involves one of the earliest migrations of modern humans, the settlement of Australia and New Guinea, and the last major colonization event, the settlement of Polynesia. Oceania can be divided into *Melanesia*, which includes several islands such as New Guinea; *Polynesia* which include islands such as New Zealand; and *Micronesia* which include north of northern Island Melanesia and Northwest of Polynesia. Another division could consider *Near Oceania*, which comprises mainland New Guinea with surrounding islands, and *Remote Oceania* which would include all islands further eastward, as well as Micronesia and Polynesia (Kayser 2010).

Previous to the human colonization process of this area, sea levels were significantly lower than they are today, exposing two major continental shelves: *Sunda*, the greater Asian landmass, and *Sahul,* the southern continent made up of present-day Australia, New Guinea and Tasmania. Although there are no consensus about timing of the occupation of Sahul region from Sunda, this pioneer colonization process has been interpreted as the result of the first exodus of AMH out of Africa (*see Figure 32*).

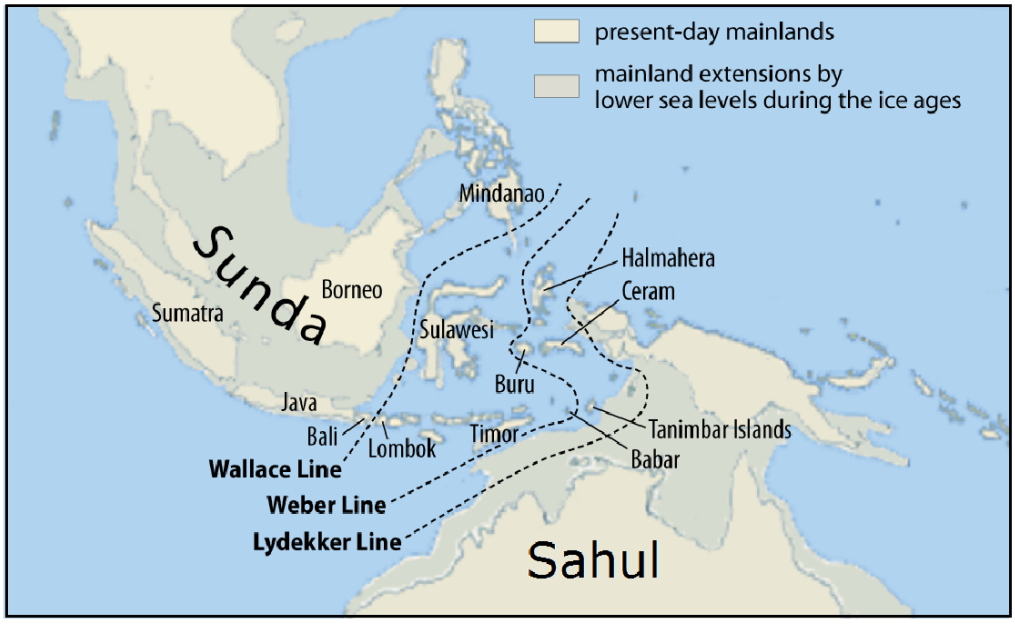


Figure 32 Representation of present day mainland extensions and their extension during the ice ages.

Several approaches have been followed to the analysis of mtDNA variation in order to investigate the origins of Polynesians. One of the most widespread and studied marker has been the 9bp deletion between nps 8281 and 8989 called *Polynesian motif* (Melton et al. 1998; Redd et al. 1995) in combination with three transition substitutions, at positions 16217, 16261, and 16247, defining mtDNA haplogroup B4a1a1 (Trejaut et al. 2005). This haplogroup reaches the highest frequencies in East Polynesian populations. Complete mitochondrial genome studies have provided a more refined knowledge of the evolution of the Polynesian motif (Friedlaender et al. 2005; Ingman and Gyllensten 2003; Pierson et al. 2006; Trejaut et al. 2005) (Soares et al. 2011)

The earliest settlement of Near Oceania appears to be most closely associated with two ancient mtDNA haplogroups, P and Q, which belong to the deep non-African N and M clades, respectively. It has been suggested that several of these lineages could have originated in Near Oceania (Friedlaender et al. 2005; Lum and Cann 2000).

Various dates have been calculated for the origins of these lineages, with P in the range of 50-60 kya and Q being slightly younger at about 45kya (Forster et al. 2001; Friedlaender et al. 2005; Ingman and Gyllensten 2003). Forster and colleagues identified the structure within the P and Q haplogroups, with seven branches now identified within P (with internal branching in P1) and three major branches within Q (Forster et al. 2001). Ingman and Gyllesten reported data that clarified those previous definitions of P and Q (Ingman and Gyllensten 2003).

Haplogroups P and Q make up about 90% of mtDNA lineages in New Guinea. They are not found on the Asian mainland and are rare in Southeast Asia Islands. The distribution of P includes Australia, but generally Australia and New Guinea have separate sub-haplogroups of P, with only the sub-haplogroup P3 found in both regions. P1 is the most common sub-haplogroup of P observed in Near Oceania, where it is found throughout the New Guinea mainland, while it is rare in the other islands of Near Oceania (Friedlaender et al. 2005). Haplogroup Q is much more common in the islands of Near Oceania although Q1 and Q2 are also found at low frequencies in Remote Oceania extending into Polynesia where Q (undefined) is found at a very low frequency (Friedlaender et al. 2005) (*see Figure 33*).

One major branch of haplogroup P is common in Papuans (languages of the western Pacific which are neither Austronesian nor Australia), while four major branches survived among Aboriginal Australians. This suggests that the early Australasians may have entered Sahul as a single group (Ingman and Gyllensten 2003). The hypothesis of a single wave was recently confirmed in a study where more than 50,000 autosomal SNPs were analyzed (Abdulla et al. 2009).

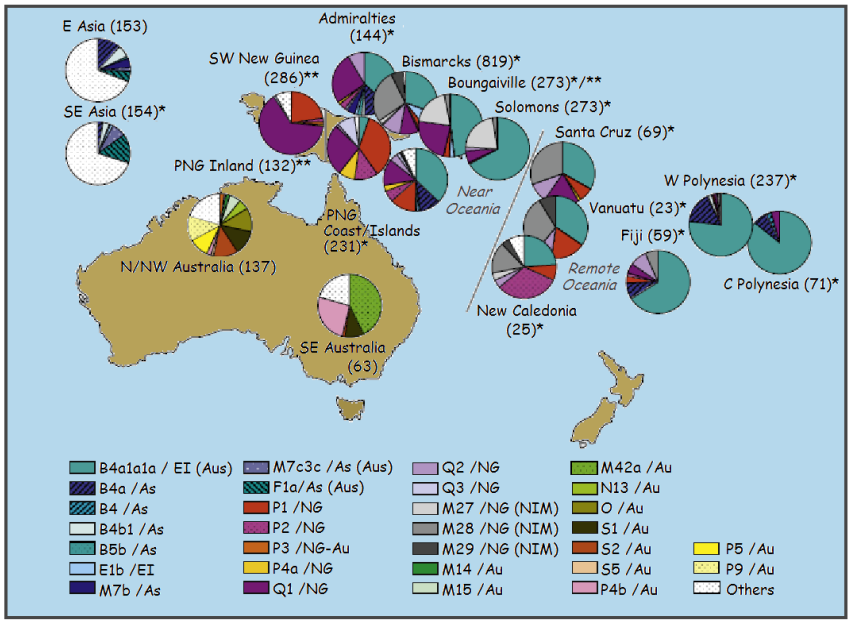


Figure 33 Distribution of mtDNA lineages across Oceania. AS indicates Asian origin; EI eastern Indonesian origin; Aus Austronesian speakers; NG New Guinean origin; NIM northern island Melanesian; Au Australian origin. Sample size is provided in brackets and language group affiliation is indicated by one asterisk for only or mostly Austronesian speakers and two asterisks for only or mostly non-Austronesian speakers. Modified from (Kayser 2010)

In addition to the P, Q and B haplogroups found in Near and Remote Oceania, complete mitochondrial genome sequencing in the region has revealed a number of new and deep lineages belonging to macro-haplogroup M (Merriwether et al. 2005). Many branches of M are found throughout mainland and Southeast Asia and it is generally accepted that M was brought to Asia and ultimately Australasia as part of the first modern human migrations out of Africa, which proceeded along the southern coastal route (Macaulay et al. 2005) These new ancient M lineages identified in Near Oceania, M27, M28 and M29 are not closely related to one another or to M lineages identified in Australian aboriginal, Indian or other Southeast Asian populations (Merriwether et al. 2005).

I.7. mtDNA VARIABILITY STUDY IN FORENSIC GENETICS

I.7.1 General considerations

While anthropological genetics usually looks into populations and demography, forensic casework focuses in the evaluation of the DNA evidence in criminalistic and affiliation casework. The interplay between both disciplines is important because forensic casework needs a population framework.

The efforts in standardization of the genotyping protocols, along with the development of standard nomenclature and the establishment of quality controls have been important contributions of forensic genetics to human mtDNA studies. However, the improvements reached in other fields of research (especially population genetics and phylogenetics) have not always been followed by forensic analysts (Salas et al. 2007).

The development of new genotyping techniques along with ancient DNA studies have contributed to the improvement of forensic casework especially in those cases where there are low amounts of DNA or the DNA is highly degraded (Alonso et al. 2006; Capelli et al. 2003).

I.7.2 Applications and limitations

mtDNA can be analyzed in whatever kind of biological evidence, including saliva and blood. Nevertheless, the mtDNA test is particularly useful for the analysis of samples containing low amounts of DNA and/or degraded DNA such as bones or teeth, or samples that do not contain (or contain very minor amounts) of nDNA such as hairs or hair shafts.

The maternal inheritance of mtDNA allows comparing a mtDNA profile to reference samples from that person or any other maternally related individuals, therefore, an individual identification is not possible.

I.7.2.1 Identification and relationship

mtDNA is mainly used in criminalistics for both associating crime scene samples against victims and suspects but also for the identification of human remains. In both cases the profile that has been generated from unknown sample has to be compared to a reference profile. For instance, in crime scene investigations, the reference sample generally comes from a suspect while in human identification, a sample is compared with a maternal relative.

Several improvements in the analysis of mtDNA evidences have come from ancient DNA studies (Capelli et al. 2003). Usually ancient DNA sources and forensic evidences have been exposed to environmental damage due to variation in temperature, humidity and ultraviolet light, favoring DNA degradation; at extreme conditions, only few molecules of mtDNA survive but not nDNA. The techniques to extract and test ancient DNA and forensic evidences are very similar in many cases (Fondevila et al. 2008); indeed, the development of new protocols or the improvement of old ones increase the success probability (Berger and Parson 2009). There is a shared need in both fields of research to avoid contamination with exogenous DNA and to eliminate PCR inhibitors (see below).

I.7.2.2 Quality controls

Some of the greater contributions of forensic casework at mtDNA human studies have been the organization by International consortiums of quality controls and standardization techniques for its genotyping as well as interpretation. Most of the forensic laboratories (at least in America and Europe) belong to *International Society of Forensic Genetics* (ISFG), where several sub-organizations are nested, such as the Grupo Español y Portugués de la ISFG (GEP-ISFG). Every year there several quality controls are organized all around the world by these organizations with the main aim of checking the ability of laboratories for genotyping and interpretation, including e.g. the use of different statistical procedures (Alonso et al. 2002; Crespillo et al. 2006; Prieto et al. 2003; Salas et al. 2005c). Indeed several recommendations for these aspects have been regularly published (Carracedo et al. 2000). In the same direction are the efforts of *European DNA profiling group* (EDNAP) (Tully et al. 2001).

Forensic casework makes strong efforts in a correct genotyping and monitorization and control of foreign DNA (contamination). Several recommendations have been published focused into control and surveillance of contamination (Bar et al. 2000; Carracedo et al. 2000). The protocols usually include use of dedicated laboratory areas, laboratory coats and disposable laboratory material. Pre- and post-amplification areas should be physically separated, work surface areas should be thoroughly cleaned before and after use, and workspaces under dedicated hoods should be employed when possible. In addition, exposing all appropriate materials and reagents to UV light should be carried out (Bar et al. 2000; Carracedo et al. 2000). In order to monitor a possible contamination, reagent blanks and negative controls should be used, and all laboratory personnel involved in the mtDNA analysis should be typed.

Some authors have proposed that these efforts in improving genotyping to control for quality should be complemented with phylogenetic control procedures (Yao et al. 2004). This approach could be very useful to reveal errors of different nature in forensic casework as it was also demonstrated in publications regarding population studies and databases (Bandelt et al. 2004a).

I.7.2.3 Criteria for inclusion and exclusion

Both, declaring an inclusion or exclusion can be problematic. A match between two samples could be indicative of an inclusion, but weighting the evidence can be problematic. A mismatch as synonymous of exclusion can be also problematic due to the possibility of mutation rate, either, intergeneration mutation rate or somatic mutation. For instance, when a questioned sample and a reference sample differ at only one position, ideally the likelihood of that one base difference occurring though a mutation should be estimated.

The following guidelines have been suggested or discussed in the literature regarding inclusion or exclusion criteria (Carracedo et al. 2000; Egeland and Salas 2008; Tully et al. 2001) (each single case has however to be examined and evaluated according to its particularities):

1. When two mtDNA sequences from separate sources match, the two sources cannot be excluded as originating from the same person or from persons with the same maternal origin.

2. When two mtDNA sequences from separate sources do not match, a number of different possible conclusions can be drawn.

a. In most cases, if there are three or more sequence differences between the two sources, the sources can be excluded as originating from the same person or from persons with the same maternal origin. Ideally, site specific mutation rates should be considered (Soares et al. 2009).

b. If there are fewer than three sequence differences, cases can be evaluated individually to determine if exclusion can be reported. The type of differences (hotspots, C-stretch regions; sites at which heteroplasmy is commonly observed) and the relationship of the reference source to the sample should be evaluated.

3. If the number or types of sequence differences are insufficient to render a conclusion, the resulting comparison should be considered to be inconclusive.

I.7.2.4 The weight of evidence

Interpretation guidelines to evaluate sequencing results from evidence and reference samples are necessary in order to prevent a type 1 error (a false inclusion) or a type 2 error (a false exclusion)(Parson and Bandelt 2007). The mtDNA genome is inherited as a single locus and this limits the evidential value of the marker in forensic cases. Haplotype frequencies have to be measured directly by counting the occurrence of a particular haplotype in a database and reporting the size of the database. A database is absolutely needed in order to interpret the weight of the DNA evidence in a criminal or identification context. There are three main problems concerning mtDNA databases:

- Firstly, the weight of evidence depends on the frequency of the profile in the reference population but, since sample sizes of most current databases are very low in relation to the large amount of variability in populations and most mitochondrial haplotypes are rare, there is a large degree of uncertainty concerning the estimation of profile frequencies. (Egeland and Salas 2008).
- Secondly, on several occasions the literature has alerted users to the high prevalence of errors in datasets produced in forensic labs, as well as laboratories from other disciplines. A notable example is one of the most widely used databases employed by forensic practitioners: SWGDAM (Bandelt et al. 2004a). In contrast to SWGDAM, the EMPOP database (www.empop.org) has been designed to avoid the addition of erroneous profiles as much as possible with the main aim of covering samples from worldwide populations (Parson and Dur 2007)

- Thirdly, the use of a database in a particular criminal case implicitly assumes that the database is representative of the populations found in the region.

There are mechanisms that compensate for the limitations of reference databases, such as minimum haplotype frequencies, and employing standard error calculations and correction factors to allow for subpopulation.

The progress made by population geneticists in the study of the phylogeny and phylogeography of mtDNA variants in human populations is now of special interest to forensic geneticists in order to prevent errors in mtDNA databases (Salas et al. 2007).

I.8. mtDNA VARIABILITY STUDY IN CLINICAL GENETICS

I.8.1 General considerations

Study of mtDNA variability can be applied in clinical genetics at several levels. The first pathogenic mtDNA mutations were identified in the late 1980s (see below). Since then, there has been a rapid rise in the number of mtDNA mutations identified in association with clinical disorders, although in some cases, the pathogenicity of the mutations has not been well established.

Mitochondrial DNA mutations have been considered to play an important role in common diseases where a particular mtDNA mutation or several mutations acting together (haplogroup) could predispose or protect to the disease. In particular, special emphasis has been dedicated to the analysis of mtDNA mutation in aging, as well as in common age-related neurodegenerative disorders such as Parkinson´s disease. The third application could be the study of the mtDNA instability and its relation with several processes like apoptosis or cancer development or ageing.

I.8.2 Applications and limitations

Mitochondria are important integrators of cellular function and therefore affect the homeostatic balance of the cell. Besides their important role in producing ATP through OXPHOS, mitochondria are involved in the control of cytosolic calcium concentration, metabolism of key cellular intermediates and contributed to a programmed cell death. Due to all those roles, mitochondrial genome variation usually appears as a good candidate to link with several pathologies.

The majority of pathogenic mtDNA mutations are heteroplasmic in affected individuals with varying proportions of mutated mtDNA within the same individual, between different tissues even between several cells in the same tissue. The presence of effects depends on the proportion (see above the threshold effect) and also the detection by genotyping.

One of the major problems of human mtDNA variability study in clinical genetic research is the absence of phylogenetic knowledge. For example, names such as UKJT, can be found in the medical literature, that are supposed to designate haplogroup clusters constitute polyphyletic entities and are therefore meaningless (Pyle et al. 2005).

Other kind of error usually appears as a mix-up of several fragments which belong to different haplogroups and obviously there are not from the same sample (Tanaka et al. 2004)(see above in the errors paragraph). A third kind of problem can be NUMTs, which are responsible of misinterpretations and due this the variants detected are implicated in causing diseases (Hazkani-Covo et al. 2010).

Moreover, in many cases the conclusions of these studies are based on poor sample sizes, the statistical analysis is deficient, there are not replication in other laboratories or even the definition of the disease is ambiguous or can involve several kind of similar diseases.

I.8.2.1 Mitochondrial disorders

In 1988, a pioneering study reported the role of a pathogenic deletion in the mtDNA molecule as responsible for a characteristic muscle pathology involving ragged red muscle fibers and abnormal mitochondria (Holt et al. 1988). Then, a second article appeared showing the presence of a missense mutation at np 11,778 located at the mtDNA ND4 gene as the cause of maternally inherited Leber Hereditary Optic Neuropathy (LHON) (Wallace et al. 1988).

Mitochondrial disorders are generally associated with defects of the mitochondrial respiratory chain, and fall into four major categories:

- Defects due to mutations in respiratory chain subunits.
- Defects due to mutations that affect respiratory chain assembly.

- Defects due to mutations that affect respiratory chain function indirectly, either via alterations in the translation of mtDNA-encoded polypeptides or via alterations in mtDNA integrity.
- Defects due to mutations in nDNA that affect organellar morphology and mobility, in which defects in respiratory chain function can be considered to be 'collateral damage'.

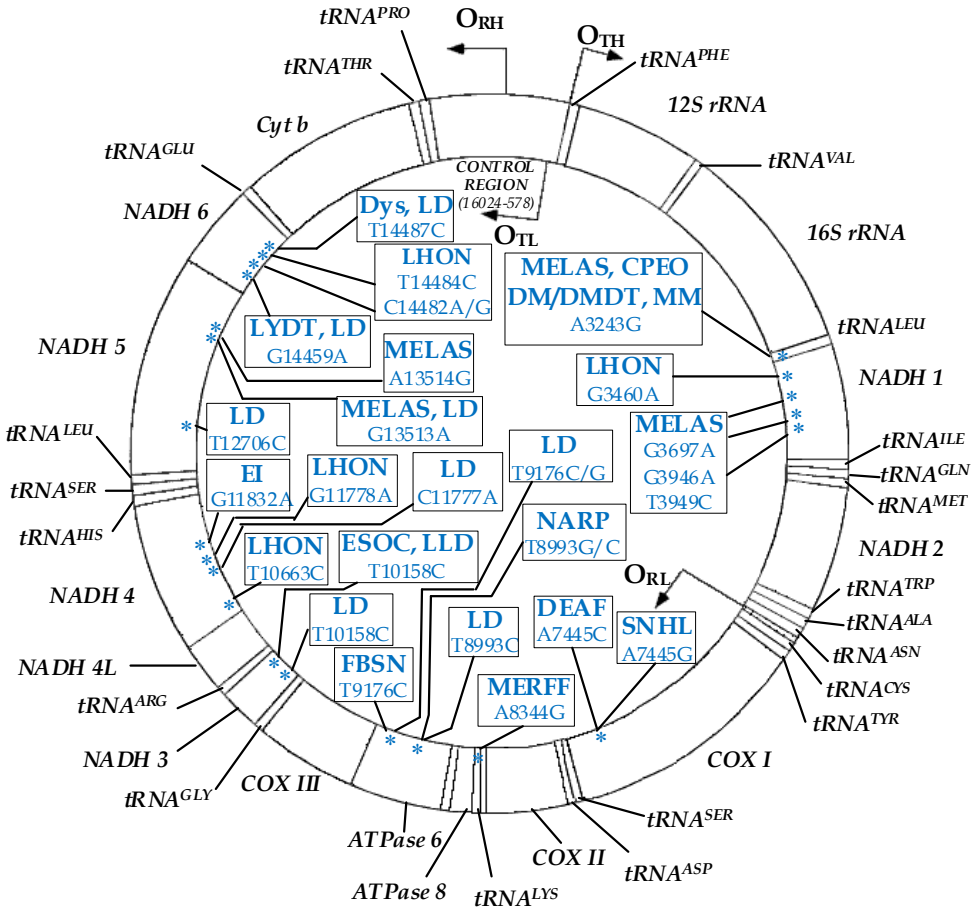


Figure 34 Approximate locations and nps of the mtDNA mutations most commonly associated with selected mitochondrial disorders. Chronic Progressive External Ophthalmoplegia (CPEO); Maternally inherited DEAFness or aminoglycoside-induced DEAFness (DEAF); DysTonia (Dys); Diabetes Mellitus (DM); Exercise Intolerance (EI); Epilepsy, Strokes, Optic atrophy, & Cognitive decline (ESOC); Familial Bilateral Striatal Necrosis (FBSN); Leigh Disease (LD); Leber's hereditary optic neuropathy and DysTonia (LDYT); Leber Hereditary Optic Neuropathy (LHON); Leigh-like Disease (LLD); Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes (MELAS); Mitochondrial Myopathy (MM); Neurogenic muscle weakness, Ataxia, and Retinitis Pigmentosa (NARP); SensoriNeural Hearing Loss (SNHL)

Some of the most frequent mitochondrial disorders are represented in Figure 34 and the most common ones, which are due to punctual mutations, are briefly commented below:

Mitochondrial encephalomyopathy with lactic acidosis and strokelike episodes (MELAS) is the most common maternally inherited mitochondrial disease. The clinical features include recurrent strokes that begin before 40 years of age, myopathy, ataxia, muscle twitching (myoclonus), dementia, and deafness. The transition from A to G at nps 3243 accounts for more than 80% of the cases of MELAS. This base substitution occurs in the tRNALeu(UUR) gene (*see Figure 34*).

Myclonus epilepsy and ragged red fibers (MERRF) is a rare, maternally inherited, heteroplasmic, debilitating multisystem disorder that includes transient seizures (myoclonus epilepsy), failure to coordinate muscle movement (ataxia), loss of muscle cells (myopathy), and, to a lesser extent, dementia, deafness, and degeneration of spinal nerves. The term "ragged red fibers" refers to large clumps of abnormal mitochondria that accumulate mostly in muscle cells and are stained red by a dye that is specific for complex II of the electron transport chain. In general terms, MERRF is a member of a group of disorders called mitochondrial encephalomyopathies that feature mitochondrial defects with altered brain and muscle functions. The majority of MERRF cases are the result of a mutation at nps 8344 in the transfer RNALys gene of the mitochondrial genome (*see Figure 34*).

Leber hereditary optic neuropathy (LHON) is a rare mitochondrial disorder of the eye. Degeneration of the optic nerve and neurons of the retina are the principal pathological features of LHON. Additional abnormalities, such as deterioration of peripheral nerves, tremors, heart conduction defects, and diminished muscle tonicity (dystonia) sometimes accompany the bilateral blindness. The onset of LHON is usually when patients are in their mid-20s, but can range from childhood to adults older than 70 years. There is generally a sex bias, with approximately five times more males than females showing the disorder. The reason for this difference has not been determined. At least 18 different missense mutations in nine genes (*NADH1*, *NADH2*, *COX I*, *ATP6*, *COX III*, *NADH4*, *NADH5*, *NADH6*, *Cyt b*) that encode mitochondrial proteins are responsible directly or indirectly for the LHON phenotype. Five of these mutations are sufficient, each on their own, to cause LHON. However, with primary LHON, three mutations at nps 3460, 11778 and 14484 account for more than 90% of all cases and the 11778A mutation is

found in 50% to 70% of these patients. Homoplasmy is a common occurrence among many primary LHON families. The threshold level for mutant mitochondrial DNA in the heteroplasmic LHON families is ≥70% (*see Figure 34*).

Neuropathy, ataxia, and retinitis pigmentosa (NARP) is a rare heteroplasmic mitochondrial disorder characterized by delayed development, muscle weakness, dementia, seizures, retinitis pigmentosa, and diminished sensory functions. Mutations at nps 8993 from T to G or C are associated with NARP when the level of heteroplasmy of the mutant gene ranges from 70% to 90%. When heteroplasmy for either 8993 mutation is greater than 90%, a fatal infancy disorder called Leigh disease (LD) occurs.

Apart from mitochondrial diseases caused by punctual mutations, there are also several mitochondrial deletion syndromes. These syndromes comprise three overlapping phenotypes that may be observed in different members of the same family or may evolve in a given individual over time: *Kearns-Sayre syndrome* (KSS), Pearson syndrome, and *Progressive External Ophthalmoplegia* (PEO). One third of *PEO/KSS/PS* patients have the so-called common deletion that removes 4977 bp of the mtDNA molecule (from nps 8469 to nps 13447)(Mao and Holt 2009).

A third group of mitochondrial disorders are those caused by a decrease of mtDNA copy number in the cell, technically called *depletion*. The density of mitochondria is different in various tissues dependent upon the demands of oxidative phosphorylation. mtDNA depletion is caused by defects in the nuclear genes that are responsible for maintenance of integrity of mtDNA or mtDNA biogenesis. The mtDNA depletion syndrome (MDS) includes: predominant myopathy, PEO, *Mitochondrial Neurogastrointestinal Encephalomyopathy* (MNGIE), Sensory-Ataxic Neuropathy, Dysarthria, and Ophthalmoplegia (SANDO) and hepato-encephalopathy. The most common tissues or organs involved in MDS and related disorders include the brain, liver and muscles.

There are several studies focussed on the detection of pathogenic mutations (Alvarez-Iglesias et al. 2008) as well as the development of techniques to prevent transmission of a given mtDNA disease (Brown et al. 2006; Craven et al. 2010). Most of the effort in the literature is however devoted to the analysis of new sequence variants as cause of one of the disorders in patient or a familial pedigree. Many of these studies have been however questioned (Bandelt et al. 2009).

I.8.2.2 Instabilities

The stability of mitochondrial genome is essential for the maintenance of energy requirements. Although there is no consensus (Bianchi 2010; Maximo et al. 2005), it has been suggested that instability can play an important role in several diseases (Bianchi et al. 2001; Liu et al. 2003; Richard et al. 2000)

Due to the features of the mtDNA concerning to its location and protection as well as its function, some authors have try to look for a causal link between the instability of mtDNA and a disease. The proper functions of mtDNA depend almost totally on specific proteins that are encoded by the nucleus and then imported into mitochondria. Thus, mtDNA instabilities can be categorized in accordance with their molecular etiologies, those that are caused by primary defects of mtDNA, and those by secondary effects from abnormalities in nDNA.

Most of the literature dealing with mtDNA instability focused on *DNA polymerase γ* (POLG), *DNA helicase Twinkle* (Twinkle), and *mitochondrial transcription factor A* (TFAM). Mutations at these genes generally lead to mtDNA deletions and/or errors in transcription of mtDNA genes frequently observed as length heteroplasmies at homopolimeric tracts.

I.8.2.3 Case-control studies

Several human common and multifactorial diseases, including neurodegenerative diseases and cancer, are associated with mitochondrial dysfunction and increased ROS damage. In 1956, Warburg proposed that alterations to respiratory capacity generated by mitochondrial impairment could be the origin of cancer. This theory was based on the observation that there are a higher glucose consumption and higher lactate production by tumor cells (in compare with normal cells) even in the presence of sufficient oxygen, suggesting that these cells preferentially use glycolysis to produce ATP. This phenomenon is known as *"aerobic glycolysis"* or the *"Warburg hypothesis".* Although it is clear that tumor cells have altered metabolism when compared with normal cells, it is difficult to understand why tumors cells prefer to generate ATP through glycolysis, even though oxygen is still present. One reason could be that glycolysis produces ATP more rapidly than OXPHOS and it is required for higher rates of cell growth and proliferation in tumour.

There are a huge amount of studies which have tried to find a causal link between mtDNA common variation and several diseases, such as several kinds of cancer (Akouchekian et al. 2009; Arnold et al. 2009; Burgart et al. 1995; Namslauer and Brzezinski 2009), autism(Graf et al. 2000; Lombard 1998; Palmieri and Persico 2010), Parkinson disease (Ko et al. 2001; van der Walt et al. 2004; Wang et al. 2008), etc. Most of these studies are based on poor sample sizes (which generally involve a low statistical power), lack of control for population stratification, deficient statistical analysis (e.g. no adjustment for multiple test), lack of a replication of the findings using independent cohorts, etc; and therefore, most of these findings (if not all) could be just type I errors and should remain on quarantine awaiting for proper replication.

II AIMS OF THE PRESENT STUDY

Analysis of mtDNA variation has demonstrated to be very useful for the reconstruction of past and modern migration events in human populations. Variation at the mtDNA molecule is also of interest in forensic genetics, and it is in fact the only genetic variation that can be explored in certain circumstances regarding forensic casework (e.g. hair shafts, degraded DNA, etc). Mutations in the mtDNA can be also responsible for different mtDNA diseases, and some polymorphisms could also contribute additively or epistatically to different complex diseases. Interplay between the different disciplines is necessary in order to comprehensively understand patterns of mtDNA variation and the different implications that these patterns could have in different biomedical areas of research. Several studies have been carried out as part of the present project with this multi-disciplinary view in mind.

From a population genetic point of view, we will go through different human populations in order to improve our knowledge of the mtDNA phylogeny or understand more about their demographic histories. Thus, for instance, we aimed to analyze in deep detail variation within the main European macro-haplogroup, R0. This clade embraces the most common mtDNA lineage in West Eurasia, namely, haplogroup H (~40%). The control region variability is not enough to discriminate between different R0 sub-lineages, so the genotyping of coding region variation is necessary in order to gain resolution and improve our understanding of European mtDNA variation.

Several aspects of the African variation are programmed to be analyzed in the present project. Although several studies have been already carried out in different African populations, there are still many questions that remain unresolved. We will firstly develop a new technique that allows targeting several mtDNA coding region SNPs to a population scale. This would allow to improve our knowledge of the African variation, this time, to a fine level of phylogenetic resolution. Some populations in Africa are particularly interesting. Thus, the *Tuareg* are a nomadic population that nowadays live along Saharan desert with several origins depending on the source. Analysis of several Tuareg populations would allow to understand the origin of this nomad groups and their relationships with other populations that have been interacting with them along their nomadic history. We are also interesting in digging into the near 1% of Sub-Saharan mtDNA lineages that are present in European population. It is belief that most of these sequences arrived to Europe very recently, although the evidences are very weak. The genotyping of complete mitochondrial genomes of several individuals from Europe

belonging to different African haplogroups can lead to the improvement of knowledge of mtDNA variability as well as to date the timing of the possible migration events from African into Europe. A deep knowledge of the African variation can also help to understand the patterns of variation observed in African-Americans. We will go through this issue by way of comparing large databases in both continents and carry out studies of particular populations in America that have been important harbours for African slaves. For instance, Colombian populations are the fruit of several complex processes of admixture between Europeans, Africans, and Native Americans to different degrees depending on the region. Although census data do not include ethnicity, more than 50 different indigenous ethnic groups have been described in the country. One of the aims of this project will be to investigate the mtDNA ancestry of different Colombian admixed group, and explore if their self-reported ancestry is supported by genetic mtDNA variation.

Among the aims of the present study, we are also interested in analyzing contrasting patterns of variation between highly admixed populations such as Colombia with other American populations that have mainly preserved their Native American original legacy. This is also interesting if we consider that there exists human populations in America that have never been analyzed to date. In this line of research, we will focus on the study of the mtDNA variation in a population sample from El Salvador; a country that has a predominant Native American nature, although no ethnic groups survived Colonial times.

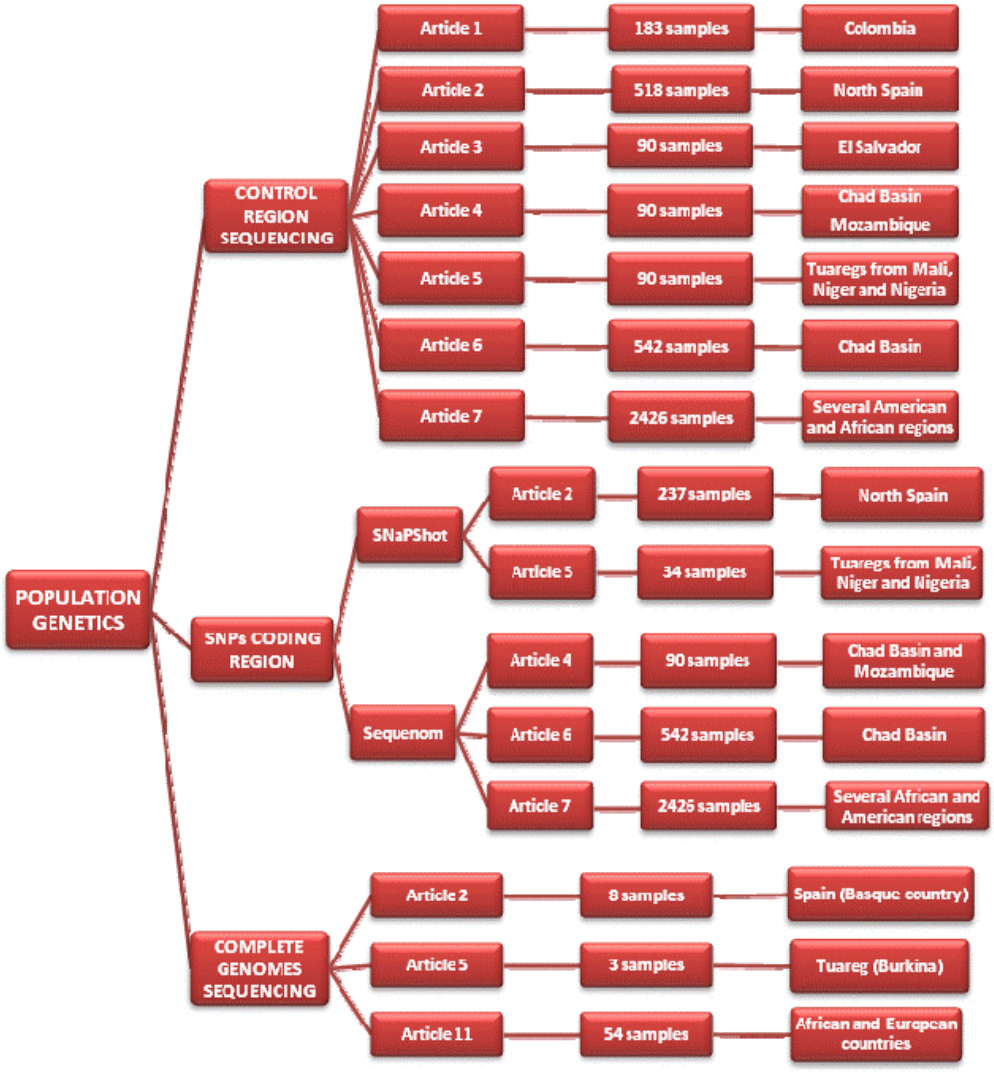
Some effort in the present project has been devoted to the field of forensic genetics. Thus, it is of interest of the present project to see the reliability and efficiency of minisequencing techniques for the analysis of complex forensic samples; all in context with standard sequencing procedures.

Finally, an issue of interest in medical genetics is the presumable role that mtDNA mutations could have in tumorigenesis. We will apply forensic standards to the analysis of mtDNA instability in tumours, in order to determine if previous findings considering an active role of mtDNA mutations in tumorigenesis are real or just false positives of instability.

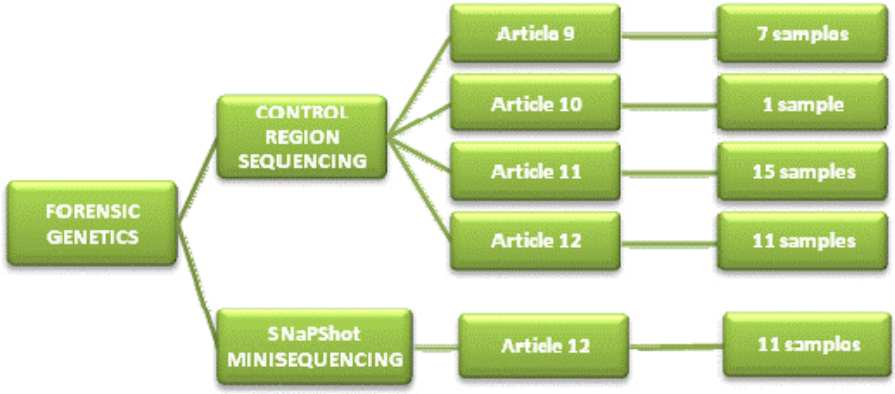
IIIMATERIALS AND METHODS

III.1 SAMPLES

III.1.1 SAMPLES FOR POPULATION STUDIES



III.1.2 SAMPLES FOR FORENSIC STUDIES



III.1.3 SAMPLES FOR CLINICAL STUDIES

III.2 DNA EXTRACTION

III.2.1 From blood stains with phenol-clorophorm-isoamilic alcohol protocol

For the extraction of DNA from blood stains, semen and saliva has been used a method of extraction with phenol-chloroform specially designed for biological fluid stains. Phenol denatures proteins and removes solutes, but it does better in the presence of a second organic solvent such as chloroform. Chloroform also stabilizes the phenol/aqueous phase boundary and improves yield by reducing the amount of aqueous phase retained by phenol. Isoamyl alcohol is used to enhance the separation of organic and aqueous phases and to reduce foaming. The next extraction with chloroform/isoamyl alcohol is used to remove residual phenol. Ethanol precipitation is used to concentrate DNA and remove residual organic solvent.

Phenol is mixed with the supernatant resulting from cell lysis (aqueous phase) and denature the sample proteins, which pass into the phenolic phase while nucleic acids remain soluble in the aqueous phase. The chloroform dissolves the fat and improves the efficiency of extraction because of its ability to denature proteins. The last part of the extraction is performed only with chloroform in order to remove any traces of phenol in the aqueous phase. The different density of the aqueous and organic (phenol or chloroform) phases, allows their separation by centrifugation in an easy and fast protocol.

- Cut 1cm^2 blood stain and add 500µL DLB (Tris-ClH 1M, ClNa 5M, EDTA 0.5M pH 8.0).
- Add 50µL SDS al 10% and 5µl proteinase K (20 mg/ml).
- Heating all night to 56 ºC with gentle agitation
- Add 20µl NaCl 5M and 575µl phenol:chloroform:isoamyl alcohol (25:24:1) and mix to form an emulsion
- Centrifugate 3 minutes at 12000rpm. and transfer aqueous phase to a fresh tube
- Add 575 µL phenol:cloroform (24:1) and mix by hand
- Centrifugate 3 minutes at 12000rpm
- Recover aqueous phase and transfer it to another 1.5mL tube, then add 1mL absolute ethanol
- Incubate 15 minutes at -80ºC and centrifugate 15 minutes at 12000rpm.
- Remove most of the ethanol from the tubes pouring off and the rest by evaporation
- Add 50-100 µl H_2O and incubate at 56ºC with a gentle rotating between 2 and 16 hours.

III.2.2 From hair without bulb samples with Chelex ®

For the extraction of DNA from hair without bulb using Chelex ® (BioRad, CA, USA) is a chelating ion exchange resin with high affinity for polyvalent metal ions and for monovalent cations such as sodium and potassium. This extraction protocol presents some modifications in compare with the first described by (Singer-Sam et al. 1989) and the described by (Sweet et al. 1996).

- 200 µl of 5% Chelex ®solution were added to at least 2-3cm of hair shaft in an 1,5 mL tube
- Then 20 ng proteinase K and 7µl DTT 1M were added and the tube were agitated.
- After that the tubes were incubated at 56°C for 60 minutes
- Then, the tubes were agitated in the vortex for 5-10 seconds
- The tubes were centrifugated 10-20 seconds at 10000rpm
- Then the tubes were incubated at 100°C for 8 minutes
- Then, the tubes were agitated in the vortex for 5-10 seconds
- Just before to do the PCR, the tubes were centrifugated 3 minutes at 10000-15000 rpm

III.3 PREVIOUS WHOLE GENOME AMPLIFICATION

Sometimes, when there is not enough amount of DNA needed for genotyping, a preliminary amplification of the entire genome present in the sample has to be carried out using commercial kit *GenomiPhi DNA Amplification Kit* (GE Healthcare Life Sciences; Uppsala, Sweden). During the making of this thesis, there was an optimization of the product, carrying out the protocols for both versions, although the quantities of reagents were the same, the optimization occurred in the time required to produce the reaction. This kit achieves a random amplification of whole genome due to different primers used have only 6 pairs of bases which are not specific to a particular region of the genome and will anneal to the template DNA at multiple sites.

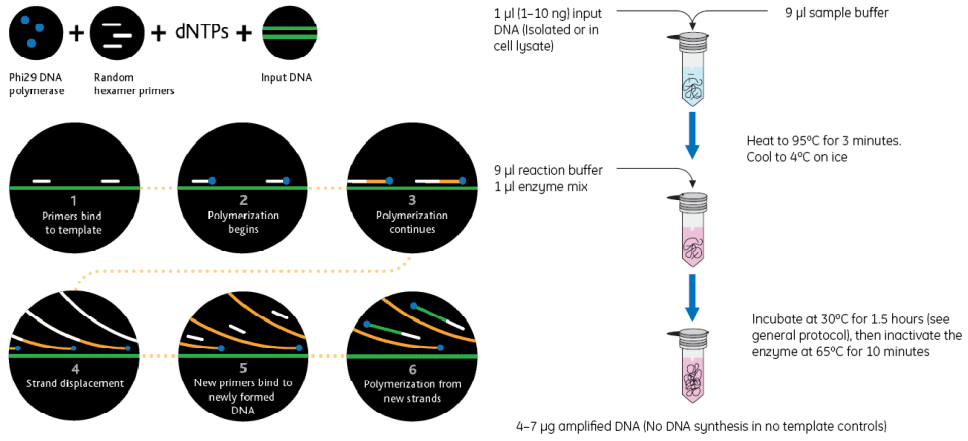


Figure 35 Different steps of the amplification with Genomiphi v2 kit (Taken from Genomiphi v2 kit Product Web Protocol)

The kit includes three components:

- *Sample buffer* which contains the random hexamer primers that nonspecifically prime polymerization catalyzed by Phi29 DNA polymerase.
- *Reaction buffer*, contains salts and deoxynucleotides, and is adjusted to a pH that is optimal for Phi29 DNA polymerase catalyzed synthesis.
- *Enzyme mix*, contains Phi29 DNA polymerase.

The steps of the protocol for amplifying template DNA are:

1. Add 9 µl *Sample Buffer* to 1 µl of (at least) 10 ng template DNA and heat the samples to 95ºC for 3 minutes then cool to 4ºC on ice
2. Prepare the master mix for each amplification reaction, on ice combine 9 µl of *Reaction Buffer* with 1 µl of *Enzyme Mix*.

3. Transfer 10 µl of prepared master mix to 10 µl of prepared cooled sample, on ice.
4. Incubate the samples for DNA amplification. at 30ºC for 1.5 hours.(16-18hours in the first version)
5. Inactivate the Phi29 DNA polymerase enzyme heating the samples to 65ºC for 10 minutes and then cool to 4ºC.
6. Then, you can use the product directly or dilute for PCR.

III.4 SEQUENCING CONTROL REGION

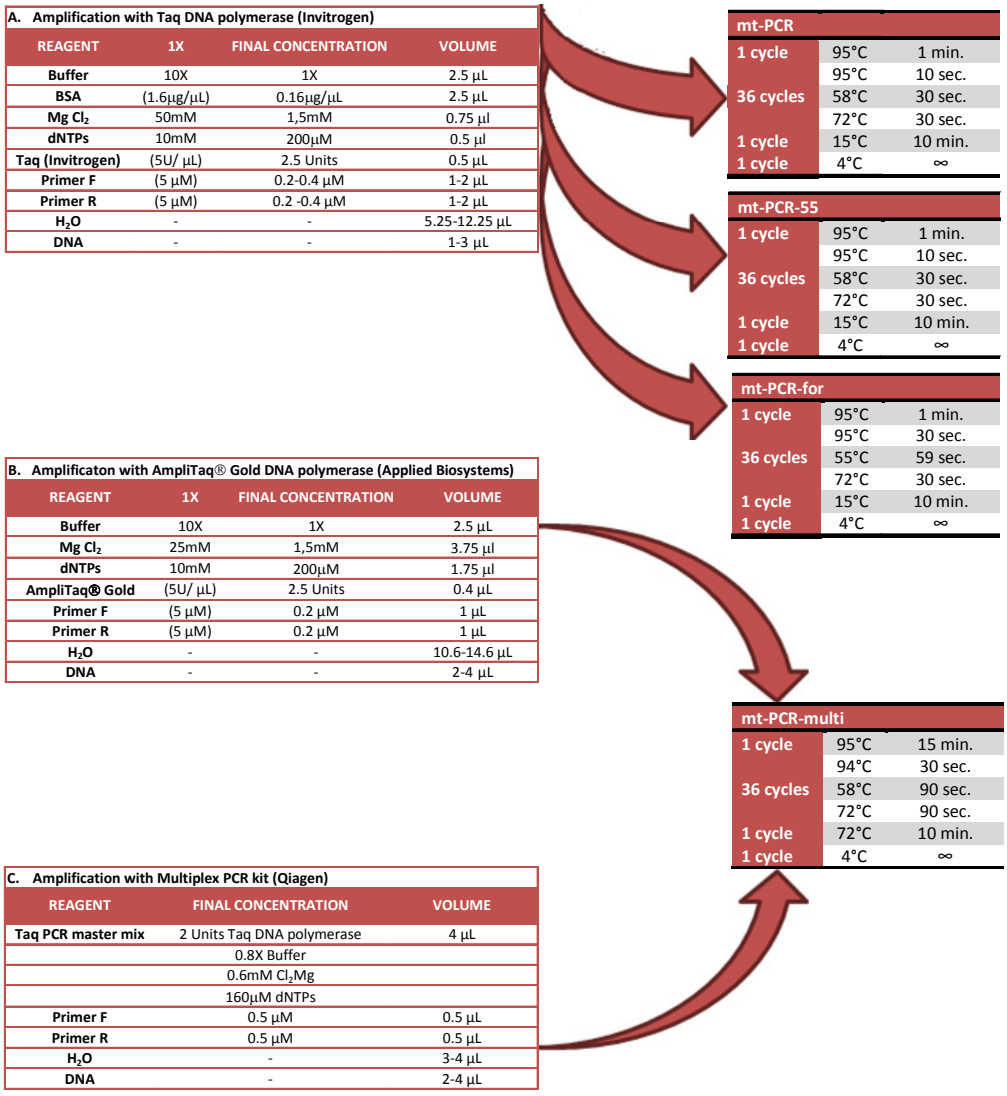
III.4.1 PCR amplification

In most of the samples the whole control region analysis was carried out (from nps 16,024 to nps 16,569 and from nps 1 to nps 576) or at least from nps 16,024 to nps 16,569. In the rest of the cases the fragment analyzed was from nps 16,024 to np 16,400.

Depending on quality of the samples, several PCR protocols were developed. In the same way various commercial kits were used for (*see Table 13*). Thus, like the first option, *Taq DNA Polymerase* (Invitrogen) was chosen (*see Table 13 A*), but when the results were not as good *AmpliTaq® Gold DNA* polymerase (AB) (*see Table 13 B*) was used and at the end *Multiplex PCR kit* (Qiagen) (*see Table 13 C*) was chosen and being the best option with the best results.

Each unit of DNA polymerase can be defined as the amount which is able to incorporate 10nmol of deoxyribonucleotide in to acid-precipitable material in 30 minutes at 74ºC. *AmpliTaq® Gold* DNA polymerase needs a previous activation step by heating which consists in 5-10 minutes to 95°C. Although this step is not necessary for the rest of the enzymes, also works suitably for the *Multiplex PCR kit*. A possible explanation for this could be that during the step the proteases present in the sample could be inactivated and in this way could not reduce amplification efficiency. In addition to the previous step to 95°C, in sample with degraded DNA or with a lot of PCR inhibitors several reagent, which improve the amplification efficiency because of their neutralizing effect against inhibitory substances were used like BSA (for most of the samples) DMSO, or glycerol.

Taq DNA Polymerase and *AmpliTaq® Gold DNA* Polymerase are commercialized with Cl2Mg and a buffer while *Multiplex PCR kit* include them and also the dNTPs leading to a reduction into pipetting errors and the possibilities of a incorrect mix of reagents.



A. Amplification with Taq DNA polymerase (Invitrogen)

REAGENT	1X	FINAL CONCENTRATION	VOLUME
Buffer	10X	1X	2.5 µL
BSA	(1.6µg/µL)	0.16µg/µL	2.5 µL
Mg Cl_2	50mM	1,5mM	0.75 µl
dNTPs	10mM	200µM	0.5 µl
Taq (Invitrogen)	(5U/ µL)	2.5 Units	0.5 µL
Primer F	(5 µM)	0.2-0.4 µM	1-2 µL
Primer R	(5 µM)	0.2 -0.4 µM	1-2 µL
H_2O	-	-	5.25-12.25 µL
DNA	-	-	1-3 µL

mt-PCR		
1 cycle	95°C	1 min.
	95°C	10 sec.
36 cycles	58°C	30 sec.
	72°C	30 sec.
1 cycle	15°C	10 min.
1 cycle	4°C	∞

mt-PCR-55		
1 cycle	95°C	1 min.
	95°C	10 sec.
36 cycles	58°C	30 sec.
	72°C	30 sec.
1 cycle	15°C	10 min.
1 cycle	4°C	∞

mt-PCR-for		
1 cycle	95°C	1 min.
	95°C	30 sec.
36 cycles	55°C	59 sec.
	72°C	30 sec.
1 cycle	15°C	10 min.
1 cycle	4°C	∞

B. Amplificaton with AmpliTaq® Gold DNA polymerase (Applied Biosystems)

REAGENT	1X	FINAL CONCENTRATION	VOLUME
Buffer	10X	1X	2.5 µL
Mg Cl_2	25mM	1,5mM	3.75 µl
dNTPs	10mM	200µM	1.75 µl
AmpliTaq® Gold	(5U/ µL)	2.5 Units	0.4 µL
Primer F	(5 µM)	0.2 µM	1 µL
Primer R	(5 µM)	0.2 µM	1 µL
H_2O	-	-	10.6-14.6 µL
DNA	-	-	2-4 µL

mt-PCR-multi		
1 cycle	95°C	15 min.
	94°C	30 sec.
36 cycles	58°C	90 sec.
	72°C	90 sec.
1 cycle	72°C	10 min.
1 cycle	4°C	∞

C. Amplification with Multiplex PCR kit (Qiagen)

REAGENT	FINAL CONCENTRATION	VOLUME
Taq PCR master mix	2 Units Taq DNA polymerase	4 µL
	0.8X Buffer	
	0.6mM Cl_2Mg	
	160µM dNTPs	
Primer F	0.5 µM	0.5 µL
Primer R	0.5 µM	0.5 µL
H_2O	-	3-4 µL
DNA	-	2-4 µL

Table 13 Control region amplification conditions with the different polymerases (A) Invitrogen, (B) Applied Biosystems y (C) Qiagen

The control region amplification was carried out in two overlapping fragments of 600bp for most of the samples, named HVI (for the PCR product which include from nps 16,024 to nps 16,569) and HVII (from nps 1 to nps 576) (*see Figure 2*). Depending on the quality of the DNA present in the sample the size of the PCR needed to be changed. When there was not amplification of these two fragments, three PCR for overlapping fragments of ~420bp (called A, B y C), or six PCR for overlapping fragments of ~240bp (called 1a, 1b, 2a, 2b, 1c y 2c) were carried out (*see Figure 36*)

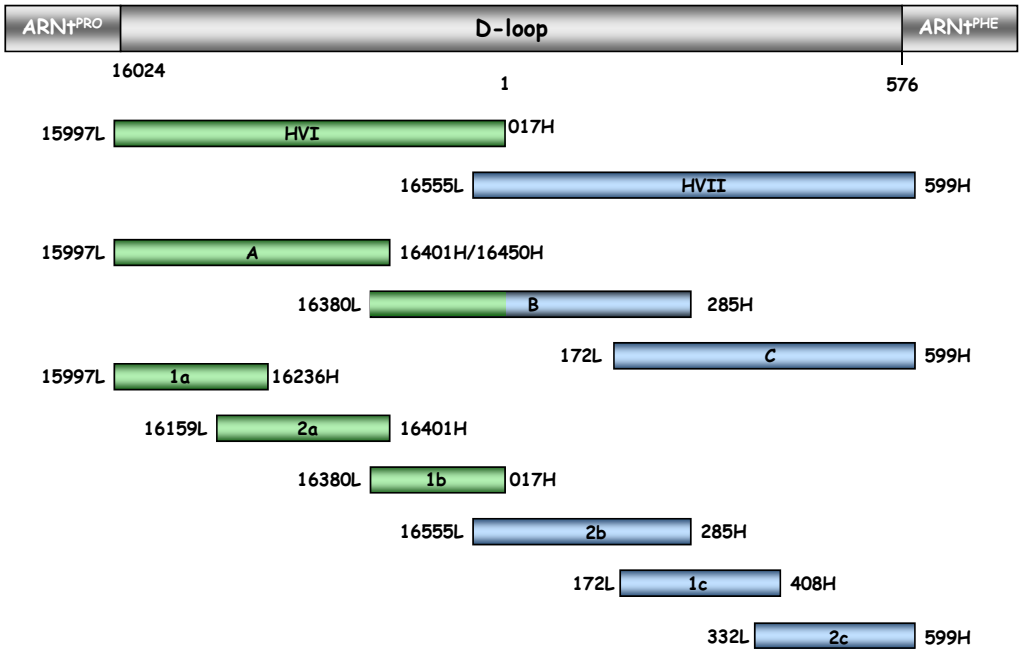


Figure 36 Different PCR products needed to obtain control region depending on the quality of the DNA

When the sample had a homopolimeric tract more primers were necessary, for example primer 16,254L (Alvarez-Iglesias et al. 2007) was used for the poli-C tract between positions 16,184-16,193; primer 370L (Alvarez-Iglesias et al. 2007) was used at the beginning to resolve the poli-C tract between 303-315 positions but then primer 332L (Yao et al. 2007) was used with better results (*for more details see Figure 37*).

All the PCR were carried out in different models of thermocyclers like GeneAmp PCR® System 2700, 2720 Thermal Cycler, GeneAmp PCR® System 9700, 9800 Fast Thermal Cycler and Veriti™ Thermal Cycler (AB).

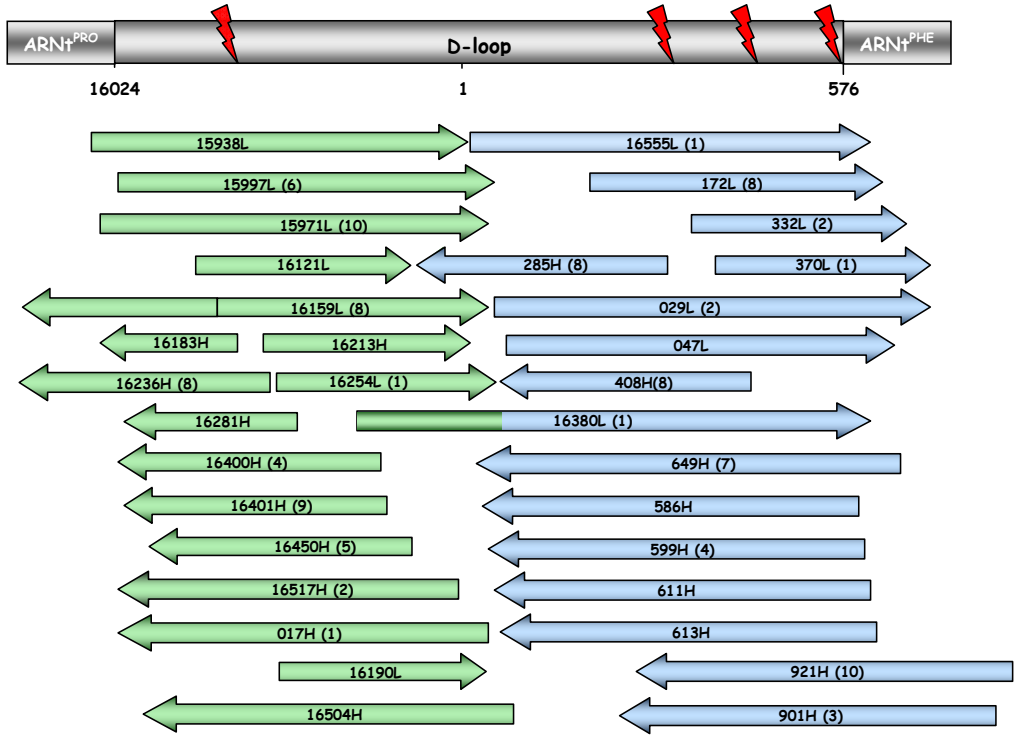


Figure 37 Primers used in order to obtain the control region sequencing. Moreover the hompolimeric tracts are indicated at the top of the figure. The numbers next to the primers indicate authors who designed them. [1] Alvarez-Iglesias *et al* 2007; [2] Yao *et al* 2007; [3] Kivisild *et al* 2006; [4] Brandstatter *et al* 2004;[5] Kao *et al* 1998; [6] Ward *et al* 1991;[7] Levin *et al* 2003; [8] Wilson *et al* 1995; [9] Vigilant *et al* 1989; [10] Levin *et al* 1999; no number is used to indicate present study.

III.4.2 PCR checking and PCR product purification

Purification of PCR product was the next step each time the checking in polyacrilamide gel with visualization by silver staining confirmed PCR was correct. There are several kinds of purification, like an enzymatic purification (see bellow in SNaPShot protocol) or physical purification. With the exception of forensic samples, all the PCR products which were performed to obtain sequencing genotypes were purified with physical purification in a vacuum manifold with MultiScreen®PCR$_{\mu 96}$ Plate (Millipore)(*see Figure 38*). In case of forensic samples the PCR product purification was carried out with Exo-SAP (*see below*).

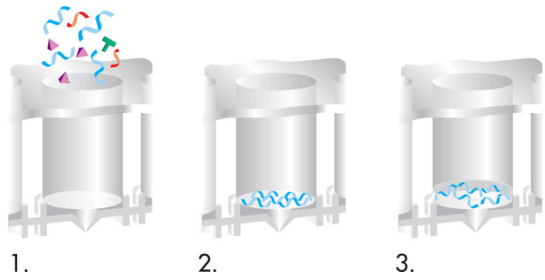


Figure 38 Steps which purification in MultiScreen®PCR$_{\mu 96}$ Plate is based on. 1. PCR product is added to the well. 2. After the vacuum all the reagents except the DNA are removed. 3. DNA can be retrieved mixing it with buffer or water and pipetting. (Taken from Millipore website)

In order to obtain the purified PCR product the steps are:

1. Adjust the volume of the PCR reactions to 100 µL with Milli-Q® grade water.
2. Transfer PCR reactions to the MultiScreen PCRµ96 plate and place the plate in the vacuum manifold
3. Apply vacuum at 18-20 inches Hg until the well are completely empty.
4. An optional wash step may be added, using 40 µL of Milli-Q® grade water, followed by vacuum filtration like previous step.
5. Dissolve the samples in 20-30 µL of Milli-Q® grade water by mixing.
6. Retrieve the purified PCR products from each well by pipetting.

III.4.3 Sequencing reaction with Rhodamina or BigDye terminators

During the performance of the present dissertation there was a change in the chemical used in the sequencing. However, both chemicals are based on fluorescent dye-labelled terminators which fluoresce at a different wavelength. In this way, at first the method used was *ABI PRISM® dRhodamine Terminator Cycle Sequencing Ready Reaction Kit* (AB) and subsequently went on to use *BigDye® Terminator v. 3.1 Cycle Sequencing Kit* (AB) according to the conditions indicated in Table 14. The signal emitted by the BigDye terminators is greater and more closely than that provided by the Rhodamine terminators which translates into a better signal, for more time in the electropherograms and a minor overlap between the bases.

A. Sequencing with ABI PRISM® dRhodamine Terminator Cycle Sequencing Ready Reaction Kit	
REAGENT	VOLUME
Sequencing kit	2 µL
Primer	1 µL
Buffer	-
H_2O	-
PCR product	5 µL

mt-Seq		
1 cycle	96°C	4 min.
36 cycles	96°C	15 sec.
	50°C	10 sec.
	60°C	2 min.
1 cycle	60°C	10 min.
1 cycle	4°C	∞

B. Sequencing with BigDye® Terminator v3.1 Cycle Sequencing Kit	
REAGENT	VOLUME
Sequencing kit	0,5 µL
Primer	1 µL
Buffer	2.5 µL
H_2O	4.5-5.5 µL
PCR product	3-4 µL

mt-Seq-bigdye		
1 cycle	96°C	3 min.
25 cycles	96°C	30 sec.
	50°C	15 sec.
	60°C	4 min.
1 cycle	60°C	10 min.
1 cycle	4°C	∞

mt-Seq-bigdye FAST		
1 cycle	96°C	1 min.
25 cycles	96°C	10 sec.
	50°C	5 sec.
	60°C	1 min.
1 cycle	60°C	1 min.
1 cycle	4°C	∞

Table 14 Sequencing conditions (A) dRhodamina terminators or (B) BigDye terminators

All the sequencing reactions were carried out in different models of thermocyclers like GeneAmp PCR® System 2700, 2720 Thermal Cycler, GeneAmp PCR® System 9700, 9800 Fast Thermal Cycler and Veriti™ Thermal Cycler (AB).

III.4.4 Sequencing product purification

After the sequencing reaction is necessary to remove all the reagents tan could cover the signal in the posterior electrophoresis. Depending on the chemical used different kinds of purification were carried out.

III.4.4.1 Ethanol precipitation of DNA

When sequencing products were obtained with Rhodamine terminators then a physical purification with ethanol and Cl_2Mg was carried out. The procedure was as follows:

1. Add to sequencing product 4.75µL 2mM Cl_2Mg and 13.75µL 96% ethanol.
2. Centrifuge sample for 20 minutes at 12,000 rpm
3. Remove as much supernatant as possible with a micropipette.
4. Add 62.5µl 70% ethanol
5. Centrifuge sample for 10 minutes at 12,000 rpm.

6. Remove as much supernatant as possible and evaporate remaining ethanol in a 37°C thermal reactor.
7. Resuspend pellet in 10 µl HiDi™ formamide(AB) and finally centrifuge sample 1 minute at 1,000 rpm in order to remove air bubbles.

III.4.4.2 Purification with MontageSEQ96 Clean Up Kit and Sephadex

When sequencing products were obtained with Big Dye terminators, several purification can be was carried out. One of this purification protocols is carry out with MontageSEQ$_{96}$ Clean Up Kit (Millipore) followed by another physical purification with Sephadex^{TM} G-50 (GE Healthcare, Bio-Sciences, Upsala).

The steps to carry out purification with MontageSEQ$_{96}$ Clean Up Kit are:

1. Dilute sequence reactions by adding 25 µL of Injection Solution and mix gently by pipetting up and down.
2. Transfer diluted reactions from the thermal cycling plate into the bottom of SEQ$_{96}$ plate wells. Place the SEQ$_{96}$ plate on the vacuum manifold.
3. Set the vacuum to 10–25" Hg and apply vacuum until the solution has been completely removed from all the wells
4. Shut off the vacuum source and remove the SEQ$_{96}$ plate from the manifold and blot the excess liquid from the bottom of the SEQ$_{96}$ plate by briefly pressing the plate on an absorbent material such as paper towels.
5. Add 25 µL of Injection Solution into the bottom of each well and repeat 3rd and 4th steps
6. Add 40 µL of Injection Solution into the bottom of each well and repeat 3rd and 4th steps
7. Add 25 µL of Injection Solution into the bottom of each well and resuspend the purified sequencing products in the Injection Solution by pipetting up and down. Alternatively, the DNA can be resuspended by shaking for 10 minutes on a microplate shaker.
8. Transfer the purified sequencing products to a plate with Sephadex (which have to been prepared previously according to the manufacturer protocol).
9. Centrifugate the Sephadex plate with MicroAmp™ Optical 96-Well Reaction Plate (AB) below it in order to collect the purified sequencing product.
10. The plate can be transfer to the ABI sequencer.

III.4.4.3 Purification with SAP and MontageSEQ96 Clean Up Kit

The other possibility when sequencing products were obtained with Big Dye terminators is an enzymatic purification with shrimp alkaline phosphatase SAP (GE Healthcare) followed by the physical purification with MontageSEQ$_{96}$ Clean Up Kit (Millipore).The SAP cleaves a phosphate from the unincorporated dNTPs, converting them to dNDPs and rendering them unavailable to future reaction.

After enzymatic purification according conditions indicated in Table 15 the sequencing product obtained is purified again with the same condition indicated in previous paragraph although in the 8rd step the purified sequencing products are transferred to a MicroAmp™ Optical 96-Well Reaction Plate (AB), then the plate needs to be centrifugate in order to remove air bubbles and finally the plate can be transfer to the ABI sequencer.

Enzymatic purification with SAP (GE Healthcare)	
REAGENT	VOLUME
Sequencing product	12.5 µL
SAP	1 µL

SAP Clean up		
1 cycle	37°C	80 min.
1 cycle	80°C	15 min.
1 cycle	4°C	∞

Table 15 Enzymatic purification with SAP conditions

III.4.5 Capillary electrophoresis in ABI 3100/3130/3730xl sequencers

Sequencing products which were obtained with Rhodamine terminators were injected in ABI 3100 sequencer (AB) with POP-6 polymer. In case of sequencing products obtained with BigDye terminators were injected in ABI 3130*xl* sequencer (AB) with POP-6 polymer or with POP-7 polymer or were injected ABI 3730*xl* sequencer (AB) only with POP-7 polymer. The best results were obtained with ABI 3730*xl* sequencer. The results were aligned with several version of the software SeqScape (AB).

III.5 CODING REGION SNPs GENOTYPING WITH SNaPShot (AB)

SNaPshot (AB) is a commercial kit which consists on a SBE method of a primer immediately adjacent to the SNP using fluorescently labelled ddNTPs. As four different dyes can be detected simultaneously, any base position can be assessed for the presence of an A, C, G, or T base (*see Figure 39*). Indeed, several fluorescently labelled extension products can be separated and visualized by electrophoresis and fluorescence detection. In this way numerous mtDNA substitution polymorphisms within a single SNaPshot reaction can be genotyped.

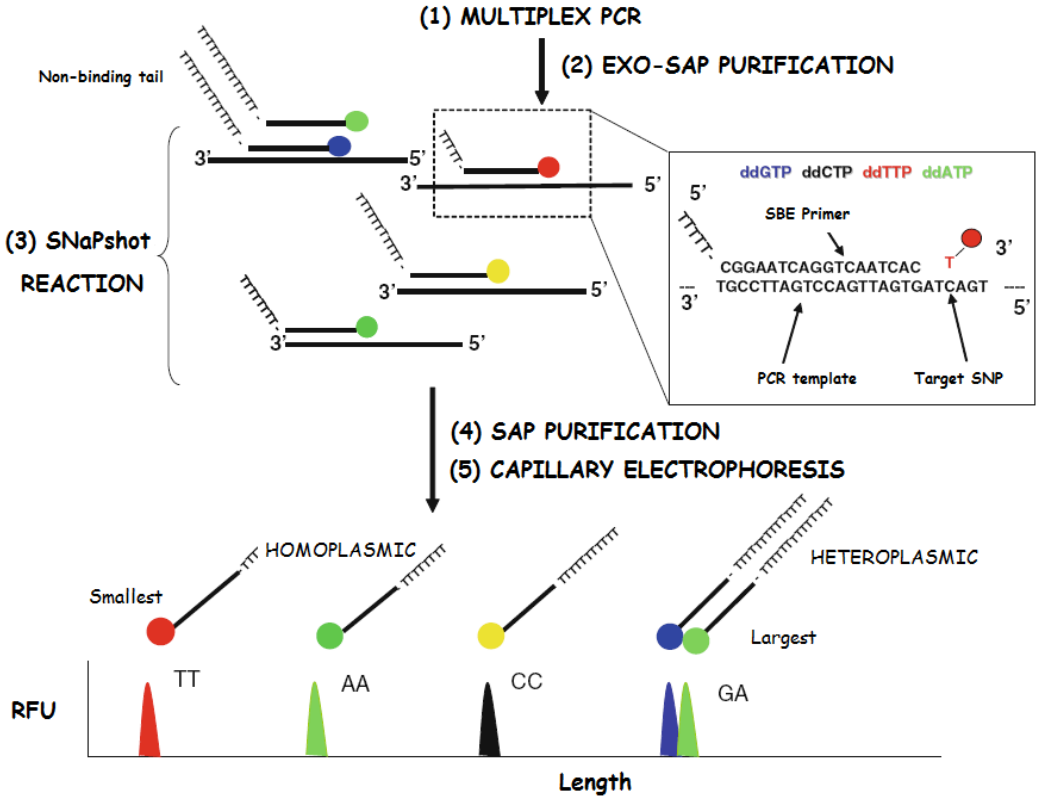


Figure 39 SNaPshot procedure. The steps are the same for all the SBE assays. RFU relative fluorescent units). Modified from (Podini and Vallone 2009).

III.5.1 Assay design

The first step to carry out this procedure is the design for the primers for PCR and the primer for mini-sequencing reaction. For this purpose the free software Primer3 was chosen, it can be downloaded from website frodo.wi.mit.edu/primer3/. The 71 SNPs chosen for a better resolution of the typical European macro-haplogroup R0 were recovered from the literature (Abu-Amero et al. 2007; Achilli et al. 2004; Behar et al. 2008a; Brandstatter et al. 2006; Brandstatter et al. 2008; Loogvali et al. 2004; Quintans et al. 2004; Roostalu et al. 2007; van Oven and Kayser 2009).

Then it is necessary a BLAST search to check if the primers chosen are only humans and exclusive of the region of interest. After this step all the primers need to be test in silico in order to check if can appear some problems when all of them work together. For this purpose AutoDimer software was chosen, it can be downloaded from www.cstl.nist.gov/strbase/AutoDimerHomepage/AutoDimerProgramHomepage.htm.

Finally, multiplexing of a SBE assay is achieved by adding a nonbinding tail sequence to the 5' end of the SBE primer. The tail is typically a poly-T or a repeating AGCT sequence. Each SBE primer is usually separated in size by three to four nucleotides in order to ensure resolution for a correct discrimination between primers.

III.5.2 PCR amplification

The multiplex PCR amplification was carried out with the same conditions as were explained in III.4.1 paragraph. Only the *Taq DNA Polymerase* (Invitrogen) was not suitable for multiplex PCR, so the other polymerases were chosen to carry out this procedure. The main change is the amount of each one of the pair of primers which is necessary to obtain enough PCR product for each one of the SNPs. Depending on the volume of the primers, the amount of water can change in order to obtain the final volume of the reaction. The 71 SNPs were genotyping in three multiplex PCR. The first one amplified 14 PCR products which contain 24 SNPs, the second 17 PCR products with 22 SNPs and the third 12 PCR products with 25 SNPs. The final concentrations as well as the sequence of the primers, the SNP and the length of the PCR products are indicated Supplementary data.

III.5.3 PCR checking and PCR product purification

The correct PCR amplification was checked in poliacrilamide gels and visualized by silver staining. In case of genotyping with SNaPShot as well as in case of sequencing of forensic samples due to their features (tiny amount of DNA and usually degraded) and to avoid possible contaminations the enzymatic purification with ExoSAP-IT was chosen. This reagent is a mixture of Exonuclease I and Shrimp Alkaline Phosphatase (SAP). Typically, the excess primers and any other foreign single-stranded DNA present in PCR products will interfere with subsequent enzymatic reactions involving DNA synthesis. The hydrolytic properties of Exonuclease I degrade all single-stranded DNA present in the PCR mixture allowing the product to be used more efficiently in other applications. SAP removes phosphatise groups from the 5'-ends of DNA, RNA and nucleotides. Working together, PCR products are free of excess of primers and nucleotides. The procedure was carried out according to the condition explained in Table 16.

Enzymatic purification with Exo-SapIT (GE Healthcare)		
REAGENT	VOLUME (SNaPShot procedure)	VOLUME (Sequencing procedure)
PCR product	1 µL	2.15 µL
Exo-SapIT	0.5 µL	0.85 µL

Exo SapIT Clean up		
1 cycle	37°C	20 min.
1 cycle	80°C	15 min.
1 cycle	4°C	∞

Table 16 Enzymatic purification with Exo-SAP conditions for SNaPShot and sequencing procedures.

III.5.4 Minisequencing reaction

Once the purification PCR products can be used as template, the minisequencing reaction or SBE can be carried out in one of the thermocyclers previously named. The SNaPshot® kit contains buffer, polymerase, and fluorescently labelled dideoxynucleotides (ddNTP) (one dye for each nucleotide).

Minisequencing with SNaPshot® kit	
REAGENT	VOLUME
SNaPshot® kit	2-4 µL
Mix SBE primers	1 µL
PCR purified product	1.5 µL
H_2O	3.5-5.5 µL

SNaPshot		
25 cycles	96°C	10 sec.
	50°C	5 sec.
	60°C	30 sec.
1 cycle	4°C	∞

Table 17 SNaPshot procedure conditions.

Each one of the three multiplex PCR was used as template for their minisequencing reaction, according to the conditions explained in Table 17. The final concentrations as well as SBE primer sequences, the strand of the primer, the alleles and the length of the primer with the tail are indicated in Supplementary data.

III.5.5 Minisequencing product purification

The enzymatic purification of minisequencing product with SAP was previously described for sequencing products in III.4.4.3 paragraph. The new conditions are explained in Table 18.

Enzymatic purification with SAP (GE Healthcare)	
REAGENT	VOLUME
SNaPshot product	10 µL
SAP	1 µL

SAP Clean up		
1 cycle	37°C	60 min.
1 cycle	80°C	15 min.
1 cycle	4°C	∞

Table 18 Enzymatic purification with SAP for SNaPshot procedure conditions.

III.5.6 Capillary electrophoresis in ABI 3100/3130

The electrophoresis must be carried out in presence of the internal ladder GeneScan-120LIZ™ (AB) to indicate the length of each peak. This ladder consists in 9 fluorescently labelled fragments with length between 15 and 120 nucleotides.

For this procedure, 1.5µL of minisequencing product is mixed with 0.25µL of GeneScan and 10 µL of formamide HiDi™(AB). Then the admixtures for each sample were injected in ABI 3130*xl* sequencer (AB) with POP-6 polymer or with POP-7 polymer (AB). The results were analyzed with the software GeneMapper™ v.3.2 (AB).

III.6 CODING REGION SNPs GENOTYPING WITH iPLEX (SEQUENOM)

It is another SBE method which resolves, via matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS), differences in primer masses due to changes in sequence due to incorporation of different terminator nucleotides at the 3'-end of a primer bound adjacent to a variant site. No indirect detection techniques such as mass tags, fluorescent groups, or radioactive labels are required for accurate resolution of single base differences such as an A–G SNP (16-Da difference in mass). This is because the analytical accuracy of MALDI-TOF MS is quite high, 0.1–0.01% of the determined mass. Given that all of the primers and extension products are between 15 and 30 bp in length (approximately 4,500– 9,000 Da) this puts the smallest separation of 9 Da (A to T) within the resolving capability of the instrument.

III.6.1 Assay design

A single Mass EXTEND assay analyzes a single variation (SNP, insertion or deletion) and is defined by three oligo primer sequences. Two oligos are PCR primers that are designed to amplify a small region (about 100 bases) of DNA containing the targeted variation site. The third oligo is the Mass EXTEND probe primer that is designed to extend into the variation site in a subsequent and distinct polymerase reaction to produce an allele-specific sequence, which may be later identified in a mass spectrum. Although PCR and probe primers are used for quite different purposes, their design is mostly restricted by the same criteria.

The software MassARRAY™ Assay Design (SEQUENOM) was chosen for the assay design. The 230 SNPs were chosen based on the main research article of African complete genomes (Behar et al. 2008b). The design tried to cover all the branches and sub-branches for each one of the L macro-haplogroups as well as M1 and U6 haplogroups in order to include all Sub-Saharan mtDNA variation.

III.6.2 PCR amplification

In general, 37 assays are usually multiplexed per reaction although vary between 3 and 37 (*see Figure 40*). PCR primers for each assay within the multiplex are mixed together with a hot start Taq master mix containing buffers, $MgCl_2$, and dNTPs, which in turn is mixed with the DNA, the conditions are explained in Table 19. The DNA was previously amplified with Genomiphi v.2 kit (*GE Healthcare*).

MULTIPLEX	1	2	3	4	5	6	7	8	9
NUMBER OF SNPs	37	37	37	37	32	28	22	8	3

Figure 40 Number of SNPs for each one of multiplexes for MALDI-TOF assay.

Each PCR primer contains a 10-bp, non-templated tail which adds mass to the PCR primer so that unextended primer has a mass larger than that of the smallest primer extension product and does not, therefore, interfere with automated genotype interpretation

Multiplex PCR Amplification with HotstarTaq (Qiagen)			
REAGENT	1X	FINAL CONCENTRATION	VOLUME
Water (HPLC grade)		-	1.850 µL
PCR Buffer (15 mM MgCl2)*	10X	1,25X	0.625 µL
MgCl2 **	25 mM	1.625 mM	0.325 µL
dNTP mix	25 mM each	500 µM	0.100 µL
Primer mix	500 nM each	100 nM	1.000 µL
HotstarTaq®	5 U/µL	0.5 U/rxn	0.100 µL
DNA			1.000 µL
*The final $MgCl_2$ concentration is 3.5 mM, 1.875 mM from the PCR buffer and 1.625 mM from the $MgCl_2$			
**For plexes >27, increase HotstarTaq® to 1U/rxn			

PCR-SEQUENOM		
1 cycle	94°C	15 min.
	94°C	20 sec.
45 cycles	56°C	30 sec.
	72°C	1 min.
1 cycle	72°C	3 min.
1 cycle	4°C	∞

Table 19 Multiplex PCR conditions for genotyping by MassARRAY®.

III.6.3 SAP treatment

The enzymatic purification of PCR product with SAP was previously described for sequencing products in III.4.4.3 paragraph and for minisequencing products in III.5.5. The new conditions are explained in Table 20.

SAP MIX	
REAGENT	VOLUME
Nanopure H2O	1.530 µL
10x SAP buffer	0.170 µL
SAP enzyme (1U/µL)	0.300 µL

Purification with SAP	
REAGENT	VOLUME
SAP mix	2 µL
PCR product	5 µL

SAP Clean up		
1 cycle	37°C	40 min.
1 cycle	85°C	5 min.
1 cycle	4°C	∞

Table 20 SAP treatment conditions for genotyping by MassARRAY®.

III.6.4 iPLEX reaction

Next to the SAP treatment, iPLEX Gold reaction cocktail (primer, enzyme, buffer, mass-modified nucleotides) is added to the amplification products. The amplification products and iPLEX Gold reaction cocktail are thermocycled to process the iPLEX Gold reaction, which involves the enzymatic addition of mass modified nucleotides into the diagnostic. During the iPLEX Gold reaction, the primer is extended by one of the nucleotides, which terminates the extension of the primer. Using a DNA polymerase that incorporates nucleotides, the iPLEX Gold reaction produces allele-specific extension products of different masses depending on the sequence analyzed. The conditions are explained in Table 21.

iPLEX Gold Reaction		
REAGENT	FINAL CONCENTRATION	VOLUME
iPLEX Buffer Plus (10x)	0.222X	0.200 µL
iPLEX Termination mix	0.5x	0.100 µL
iPLEX enzyme	0.5x	0.0205 µL
Primer mix (7 µM: 14 µM)	0.625 µM: 1.25 µM	0.94 µL
Water (HPLC grade)	-	0.7395 µL
PCR purified product		7 µL

iPLEX Gold Reaction			
1 cycle		94°C	30 sec.
40 cycles		94°C	5 sec.
	5 cycles	50°C	15 sec.
		60°C	4 min.
1 cycle		60°C	10 min.
1 cycle		4°C	∞

Table 21 iPLEX Gold reaction conditions

III.6.5 Desalting iPLEX reaction products and dispensing to SpectroCHIP® Bioarrays

This cleanup step is important to optimize mass spectrometry analysis of the iPLEX Gold reaction products. The clean resin is a cationic resin pretreated with acid reagents. The resin is added directly to primer extension reaction products to remove salts such as Na+, K+, and Mg2+ ions. If not removed, these ions can result in high background noise in the mass spectra. Due to the detection is inside a machine of mass spectrometry is necessary transfer the samples from 384- well plates to a chip with a nanodispenser.

III.6.6 MALDI-TOF MS analysis

The chip is placed into the mass spectrometer and each spot is then shot with a laser under vacuum by the MALDI-TOF method. A laser beam serves as desorption and ionization source in MALDI mass spectrometry. Once the sample molecules are vaporized and ionized, they are transferred electrostatically into a time-of-flight mass spectrometer (TOF-MS), where they are separated from the matrix ions, individually detected based on their mass-to-charge (*m*/*z*) ratios, and analyzed. High transmission and sensitivity, along with theoretically unlimited mass range, are among the inherent advantages of TOF instruments. Detection of an ion at the end of the tube is based on its flight time, which is proportional to the square root of its *m*/*z*. iPLEX SpectroCHIP bioarrays were processed and analyzed with MassARRAY Workstation version 3.3 software.

III.7 COMPLETE GENOMES SEQUENCING

The genotyping of complete genomes was carried out in Instituto de Ciencias Forense in Santiago de Compostela as well as in Zoology Department in Oxford during the predoctoral stay.

III.7.1 Previous studies

In order to obtain a good protocol for complete genomes, instead of the design of new primers for amplification and sequencing, previous studies were analysed to choose between them. There are multiple protocols in the literature although a lot of authors have chosen between only a few of them (*see* Figure 41).

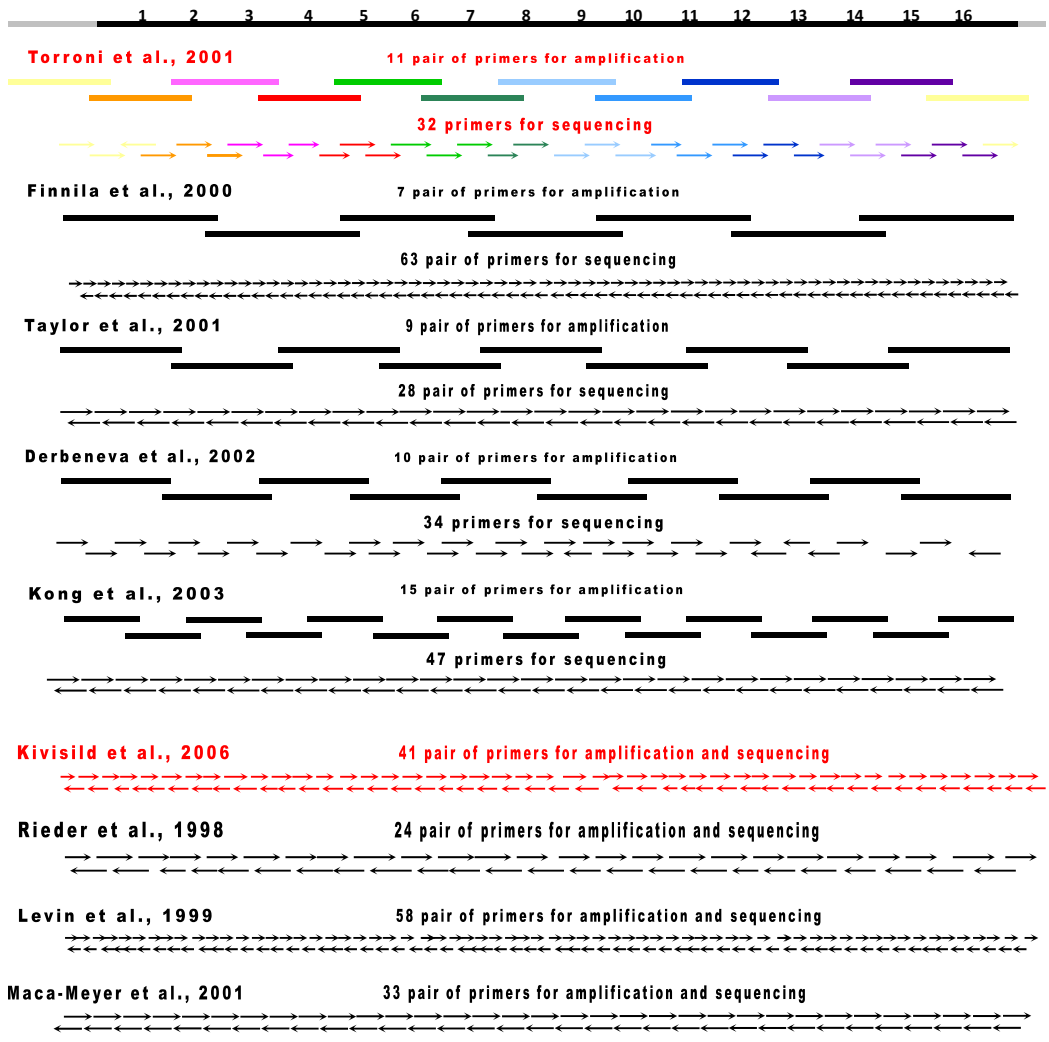


Figure 41 Several of the complete mitochondrial genomes strategies more used in literature. The set of primers chosen for carried out the complete genome sequencing in the present PhD project appear in colours.

III.7.2 PCR amplification

Depending on the quality of the DNA, samples for complete sequencing were amplified with different protocols:

1. Amplification of three overlapping long PCR fragments following by 34 sequencing reactions.
2. Amplification of eleven overlapping PCR fragments following by 34 sequencing reactions.

3. Amplification of 44 overlapping PCR fragments following 44 by sequencing reactions.

The first protocol was based on long PCRs which allowed the complete mitochondrial genome to be amplified in three overlapping fragments, named fragment 1, fragment 2 and fragment 3 (see below). The majority of the long PCR reactions were carried out using the Expand Long Template PCR System (Roche Applied Science, Mannheim, Germany) following the conditions indicated in Table 22

Amplificaton with Expand Long Rang dNTPack (Roche)			
REAGENT	**1X**	**FINAL CONCENTRATION**	**VOLUME**
Buffer (with 12.5mM Mg Cl_2)	5X	1X	5 µL
dNTPs	10mM	500µM	1.25 µl
Enzyme mix	(5U/ µL)	3.5 Units	0.35 µL
DMSO	100%	8%	2 µL
Primer F	(10 µM)	0.3 µM	0.75 µL
Primer R	(10 µM)	0.3 µM	0.75 µL
H_2O	-	-	11.9 µL
DNA	-	-	3 µL

LONG-PCR		
1 cycle	92°C	2 min.
10 cycles	92°C	10 sec.
	55°C	15 sec.
	68°C	2 min.
25 cycles	92°C	10 sec.
	55°C	15 sec.
	68°C	2 min. +20sec/cycle
1 cycle	68°C	7 min.
1 cycle	4°C	∞

Table 22 Amplification conditions for long PCR with Expand Long Range dNTPack (Roche)

The primer combinations used to amplify the entire mitochondrial genomes in three overlapping products were 1F and 4R to amplify 'Fragment 1'; 4F and 7R for 'Fragment 2', and finally 7F and 11R for 'Fragment 3' from the set of 11 primer pairs (Table 2 in Appendix) described by Torroni (Torroni et al. 2001b). These combinations generated products of length 6355bp, 6318bp, and 8110bp respectively, according to the rCRS. Other primer combinations were tried, but resulted in multiple products worse amplification or no amplification.

Sometimes, the quality of DNA did not allow the amplification in only three fragments. In those cases, *Multiplex PCR kit* (Qiagen) was chosen to carry out the PCR and the conditions were the same explained previously for the control region sequencing.

Along with the thermal cycles indicated in previous paragraphs MJ Thermal Cycler PTC-225 was used for the amplification reactions as well as for the sequencing reactions.

III.7.3 PCR checking and PCR products purification

Most of the PCR fragments were checked in agarose gels (1%) and visualised by stained with ethidium bromide. After this step the fragments were recovered with a

Pasteur pipette instead of a sharp scalpel from the agarose. The extraction and purification of the DNA from the agarose was carried out with the *QIAquick Gel Extraction Kit* (Qiagen).

The steps to obtain the purified DNA from agarose are:

1. Transfer the portion of gel with the fragment of interest from Pasteur pipette to a 1.5mL microcentrifuge tube.
2. Add 100µL of buffer QG and incubate until the gel slice has completely dissolved.
3. Add 50µL of isopropanol and mix.
4. Transfer the sample to the QIAquick column with a collection tube and centrifuge for 1 minute at 10,000 x *g.*
5. Discard flow-through and place QIAquick column back in the same collection tube.
6. Add 750µL of buffer PE to the QIAquick column with a collection tube and centrifuge for 1 minute at 10,000 x *g.*
7. Discard the flow-through and centrifuge the QIAquick column for an additional 1 min at 10,000 x *g.*
8. Place QIAquick column into a clean 1.5 ml microcentrifuge tube.
9. Add 30 µl of Buffer EB to the center of the QIAquick membrane, let the column stand for 1 min, and then centrifuge for 1 min at 10,000 x *g.*

III.7.4 Sequencing reaction

The sequencing reaction protocol was explained previously for the sequencing of control region.

III.7.5 Sequencing product purification

Moreover the combination of enzymatic purification with SAP and physical purification with MontageSEQ$_{96}$ Clean Up Kit, the purification of sequencing products was carried out with SAP followed a purification with Sephadex with the same conditions indicated in previous paragraphs.

III.7.6 Capillary electrophoresis in ABI 3730xl

In this case all the sequencing products were obtained with BigDye terminators and injected in ABI 3730*xl* sequencer (AB) with with POP-7 polymer, and then results were aligned with version 2.5 of the software SeqScape (AB).

III.8 ANALYSIS OF AUTOSOMAL MARKERS

Apart from all the genotyping analysis carried out for mtDNA, autosomal makers were analysed in two of the studies of the present PhD project

The first of this analysis was carried out in the study concerning to Sub-saharan African mtDNA lineages in European samples. In this study we carried out an analysis of 34 autosomal SNPs which are classified as *Ancestral Informative Markers* (AIMs)(Phillips et al. 2007). The objective of this analysis was discard that the samples which belong to Sub-saharan African mtDNA lineages belonged to recent migrants.

The other study, where autosomal markers were analysed, is about mtDNA instability. In this case an analysis of *Short Tandem Repeats* (STRs) was carried out with PowerPlex® 16 System (Promega) in order to discard that the mtDNA instability detected were due to an admixture from two samples.

III.9 ANALYSIS OF THE GENOTYPING DATA

The analysis carried out for the present PhD project were explained more extensively previously (*see I.5 paragraph*). Here we only list the package software used to carry out all the statistical analyses procedures used in the present project.

METHOD/ANALYSIS	SOFTWARE	URL
Intrapopulation variation parameters Genetic distances AMOVA	Arlequin v. 3.11	http://cmpg.unibe.ch/software/arlequin3/
Principal Component Analysis	STATA v.9 R	http://www.stata.com/ http://www.r-project.org/
Median-Joining Network	Network 4.5	www.fluxus-engineering.com
Diversity indices	DNAsp 5.0	http://www.ub.es/dnasp/
Model of nucleotide substitution Base frequencies Substitution rates	jMoldelTest v.0.1.1	http://darwin.uvigo.es/software/jmodeltest.html
Population size Population growth Migration rates	LAMARC v.2.1.6 MIGRATE v.3.2.6	http://evolution.genetics.washington.edu/lamarc http://popgen.sc.fsu.edu/Migrate/Migrate-n.html
Interpolation maps	Surfer v.8	http://www.goldensoftware.com
Population structure	STRUCTURE v.2.3.3	http://pritch.bsd.uchicago.edu/structure.html
Ancestral origin	SNIPPER	http://mathgene.usc.es/snipper/

IV RESULTS

IV.1 POPULATION GENETICS

IV.1.1 Article 1: The mtDNA ancestry of admixed Colombian populations ***American Journal of Human Biology***

IV.1.2. Article 2: New population and phylogenetic features on the internal variation of the mtDNA macro-haplogroup R0 ***PLoS ONE***

IV.1.3. Article 3: Mitochondrial echoes of first settlement and genetic continuity in El Salvador ***PLoS ONE***

IV.1.4. Article 4: Applications of MALDI-TOF MS to large scale human mtDNA population-based studies ***Electrophoresis***

IV.1.5. Article 5: Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel ***European Journal of Human Genetics***

IV.1.6. Article 6: New insights into the Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs. ***PLoS ONE (accepted)***

IV.1.7. Article 7: Manuscript under preparation concerning to phylogeny and demography of African mtDNA variability in Africa and The Americas due to Slave Trade

IV.1.8. Article 8: Manuscript under preparation concerning to the spread of African mtDNA lineages in Europe

AMERICAN JOURNAL OF HUMAN BIOLOGY 20:584–591 (2008)

Original Research Article

The mtDNA Ancestry of Admixed Colombian Populations

A. SALAS,[1,2*] A. ACOSTA,[1] V. ÁLVAREZ-IGLESIAS,[1] M. CEREZO,[1] C. PHILLIPS,[1] M. V. LAREU,[1] AND Á. CARRACEDO[1,2]
[1]*Unidade de Xenética, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, 15782 Galicia, Spain*
[2]*Grupo de Medicina Xenómica, CIBERER, Hospital Clínico Universitario, 15706 Santiago de Compostela, Galicia, Spain*

ABSTRACT A total of 185 individuals from Colombia were sequenced for the first hypervariable region (HVS-I) of the mitochondrial DNA (mtDNA) genome, and a subset of these individuals were additionally genotyped for the second hypervariable segment (HVS-II). These individuals were collected according to their "self-reported ethnicity" in Colombia, comprising "Mestizos," "Mulatos," and "Afro-Colombians." We used databases containing more than 4,300 Native American lineages, 6,800 Africans, and 15,600 Europeans for population comparisons and phylogeographic inferences. We observe that Mulatos and Afro-Colombians have a dominant African mtDNA component, whereas Mestizos carry predominantly Native American haplotypes. All the populations analyzed have high diversity indices and there are no signatures of dramatic genetic drift episodes. Central and South America are the main candidate source populations of the Colombian Native American lineages, whereas west-central, southwest, and southeast Africa are the main original mtDNA sources for the African Colombian mtDNAs. We found that our results differ from those obtained in other studies for the same "population groups" in terms of haplogroup frequencies. This observation leads us to conclude that (i) self-reported ancestry is not a reliable proxy to indicate an individual's "ethinicity" in Colombia, (ii) our results do not support the use of outmoded race descriptions (Mestizos, Mulatos, etc.) mainly because these labels do not correspond to any genetically homogeneous population group, and (iii) studies relying on these terms to describe the population group of the individual, which then treat them as genetically homogeneous, carry a high risk of type I error (false positives) in medical studies in this country and of misinterpretation of the frequency of observed variation in forensic casework. Am. J. Hum. Biol. 20:584–591, 2008.

The two main linguistic groups that dominated the territory of Colombia during pre-Colombian times were the Chibcha and the Carib. Examples of these Native American groups were the Tayronas living in the Caribbean region and the Muiscas in the highlands near Bogotá, both of which belonged to the Chibcha language family. The Muisca people had one of the most developed political systems in South America, surpassed only by the Incas. These populations constituted mainly hunter-gatherer societies and they traded with one another (and with other cultures living in the Magdalena River valley) by exchanging salt, emeralds, beans, maize, other crops, etc.

When the first Spanish arrived to what is now Colombia, the largest and most widespread culture was the Chibchas; these population groups were concentrated mainly in the highland basins and valleys of the Cordillera Oriental. The colonial period led to a dramatic change in the political and socioeconomic regimes of the indigenous Colombian people. The Spanish settled along the north coast of modern Colombia as early as 1500, but their first permanent settlement at Santa Marta and Cartagena was not established until 1525 and 1533, respectively. The city of Bogotá, founded in 1538, became one of the principal administrative centers of the Spanish possessions in the New World (along with Lima and Mexico City).

The arrival of African slaves for forced labor to New Granada (the old name of Colombia used 1533–1858) would dramatically change the demographic landscape of the region. It is estimated that the city of Cartagena alone received more than 200,000 African slaves destined for the Viceroyalty of Perú (Curtin, 1979). The late 18th century saw progress toward the abolition of slavery as part of a movement for independence from Spain.

Today, more than 50 different indigenous ethnic groups have been described in Colombia, and most of them have preserved their original languages (belonging to the Chibchan and Caribean linguistic families). The census data in the country do not record ethnicity, and so percentages are based on estimates from other sources and can vary from one region to another. In general, it is thought that the Colombian population is the result of a complex process of admixture between Europeans, Africans, and Native Americans to different degrees depending on the region.

In the last 20 years, the analysis of molecular DNA markers has contributed significantly to an understanding of the prehistory and history of human populations. Colombia has been the target of a large number of genetic studies, most of them, however, comprise forensic genetics analysis, e.g. short tandem repeat (STR) databases of forensic use (Bravo et al., 2001; Paredes et al., 2003), but with little focus on the anthropological and demographic issues. One of the first attempts to unravel the ancestry of "Mestizo" and "Afro-Colombians" was performed by Rodas et al. (2003) by analyzing mitochondrial DNA (mtDNA) restriction fragment length polymorphisms (RFLP) variation and characterizing the main Native American, West

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/1042-0533/suppmat.

Contract grant sponsor: Ministerio de Educación y Ciencia; Contract grant number: RYC2005-3; Contract grant sponsor: Xunta de Galicia; Contract grant number: PGIDIT06PXIB208079PR; Contract grant sponsor: Fundación de Investigación Médica Mutua Madrileña.

*Correspondence to: Antonio Salas, Unidade de Xenética, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, 15782 Galicia, Spain. E-mail: apimlase@usc.es

Received 23 November 2007; Revision received 25 January 2008; Accepted 19 February 2008

DOI 10.1002/ajhb.20783

Published online 28 April 2008 in Wiley InterScience (www.interscience.wiley.com).

European, and African haplogroups; whereas, Torres et al. (2006) studied the mtDNA RFLP variation in Colombian Native Americans alone. Recently, Bedoya et al. (2006) studied the genetic composition of a sample described as "Hispanics" from Antioquia (Colombia). Previous work from the same group characterized the admixture Amerind/"white" sex bias in a sample from Colombia (Carvajal-Carmona et al., 2000, 2003). Lastly, a recent article by Melton et al. (2007) examined the mtDNA variation of three Chibchan and one Arawak population from north-east Colombia mainly focusing on the demographic expansion of the Chibchan-speakers into South America.

The present study aimed to investigate the mtDNA ancestry of admixed Colombian groups of individuals, including "Mestizos" (the term widely used in America to designate individuals of European and Native American co-ancestry), "Mulatos" (the term used in Colombia, to designate individuals of African and European coancestry), and "Afro-Colombians," by analyzing a sample of 185 individuals and focusing on a phylogeographic approach that complements the previous studies by other authors. One of the main aims of this study is to demonstrate that the use of the ancestry descriptors in Colombia (and in many other American countries), whether self declared or assigned during sample collection, is not supported by the genetic variation we have characterized. It should be noted that, just for clarity, we report the ancestry descriptors used by other studies or those based on self declaration in our own study. Furthermore, we advocate the use of simple continental descriptors of ancestry or coancestry, namely European, African, American, European-American, although the term Native American avoids confusion with African American and European American of near universal use when applied to U.S. population studies.

MATERIALS AND METHODS

Samples

We have analyzed 185 individuals from Colombia, most of them from the South (departments of Cauca and Valle del Cauca). These samples consist of individuals belonging to different "self reported ethnicities": 67 Mestizos, 11 Mulatos, and 95 Afro-Colombians. We also included 10 Native Americans (Páez Indians). Informed consent was given by all participants. The uneven nature of the sample sizes of the population groups studied was related to the logistics of sample collection, which is particularly complex given the size, linguistic diversity and difficulty of travel in the Colombian regions analyzed. The protocol and procedures employed were approved by the review committee of the University of Santiago de Compostela (Spain) where the study was performed.

PCR and sequencing

All samples were amplified and double-strand sequenced for the HVS-I; however, because of limitations related to the amount of DNA available, we could not analyze all samples for the same sequence range; however, most were read for the sequence segment 16024–16569. For the same reason, only a subset of samples could be sequenced for the HVS-II segment. PCR amplification was carried out using the GenAmp PCR sequencing system, as described in the work of Álvarez-Iglesias et al. (2007).

Mutations are referenced with respect to the revised Cambridge Reference Sequence (Anderson et al., 1981; Andrews et al., 1999). We followed a standardized forensic nomenclature system as indicated in the work of Carracedo et al. (2000), but with slight modifications considered in that of Salas et al. (2005a). The data were checked following the phylogenetic principles described in previous works (Bandelt et al., 2002, 2004a,b; Salas et al., 2005a,d, 2007) in order to reduce to an absolute minimum possible sequence artifacts. The final results are presented in Table S1.

Statistical approach and mtDNA nomenclature

For phylogeographic purposes we have collected different available population datasets from the literature. Thus, for the African dataset, we have used nearly the same data reported in the work of Černý et al. (2007), which consists of 6,856 profiles from different regions in the African continent.

For Native American lineages, an HVS-I database that consists of 4,394 profiles was employed. Finally, the European database consists of more than 15,600 HVS-I profiles. More information concerning the literature source of these large datasets will be provided under request to the corresponding author.

Only the HVS-I sequence range from position 16090 to 16365 was used for population comparisons, because this is the common segment available for the different population datasets used. Nomenclature of African haplotypes follows (Salas et al., 2002, 2004) with the updates in the works of Kivisild et al. (2004) and Torroni et al. (2006), and slight modifications reported in that of Černý et al. (2007). For Native American haplogroups, we use the most updated nomenclature reported by Kong et al. (2006), Tamm et al. (2007), and Achilli et al. (2008).

Finally, principal component analysis was carried out using the software Stata v.9 based on haplogroup frequency profiles, and including the main Colombian population groups (Mestizos, Mulatos, and Afro-Colombians) and few other populations representing the three main Colombian source populations (Native Americans, Africans, and Europeans).

RESULTS

Variability of the Native American and African Colombian lineages

For the Native American haplogroups (A2, B2, C1, and D1), 45 haplotypes appeared only once in the whole Colombian sample, five appeared twice, five were found three times, three were found six times, and one was found 11 times (Table S1). A total of 33 haplotype sequences were not found in our American database. Among the observed sequences, 15 match in North America (n = 2,005), nine in Central America (n = 485), and 16 in South America (n = 1,657); but many of them are present in the three main regions or at least in two of them (Table 1).

For the African lineages (56 sequences belonging to haplogroup L), 42 sequences appeared once in the whole Colombian sample, nine sequences were found twice, one haplotype was found three times, and three sequences were found five times (Table S1). Forty-seven out of 56 (~84%) mtDNAs had been previously observed in other studies on other American locations. Colombia shares 18

TABLE 1. Shared haplotypes between Colombian Native American lineages and the main American regions

Colombian Native American lineages (minus 16000)	*n*	AM-C	AM-N	AM-S	Total	HG
111 223 290 319 362	6	45	161	50	262	A2
093 111 223 290 311 319 362	1	–	–	–	1	A2
111 129 192 223 290 319 362	1	–	–	–	1	A2
111 183C 189 223 287 290 319 362	3	–	–	3	6	A2
111 187 223 290 319	1	–	–	–	1	A2
111 187 223 290 319 362	1	25	2	–	28	A2
111 189 193+C 223 290 319 362	1	–	1	–	2	A2
111 192 223 274 290 311 319 362	1	–	–	–	1	A2
111 213 223 290 319	1	–	–	–	1	A2
111 213 223 290 319 356 362	1	–	–	–	1	A2
111 213 223 290 319 362	11	–	–	3	14	A2
111 213 223 319 362	1	–	–	–	1	A2
111 223 247T 290 295 319 362	1	–	–	–	1	A2
111 223 274 290 311 319 362	1	–	–	–	1	A2
111 223 290 300 319 362	1	1	–	–	2	A2
111 223 290 311 319 362	1	2	2	2	7	A2
111 223 290 319	1	3	11	1	16	A2
111 223 290 319 356 362	3	–	4	3	10	A2
093 129 183C 189 217	1	–	–	–	1	B2
175 183 189 217	1	–	–	–	1	B2
182C 183C 189 217 300	1	–	–	–	1	B2
182C 183C 189 217 301 304	1	–	–	–	1	B2
183C 189 193+C 217	1	–	1	4	6	B2
183C 189 217	6	1	14	86	107	B2
183C 189 217 235 319C	1	–	–	–	1	B2
223 295 298 325 327	1	–	6	–	7	C1
223 298 325	1	–	–	2	3	C1
092 176 218 223 298 325 327	1	3	1	–	5	C1
104 174 223 298 325 327	1	–	–	–	1	C1
127 223 295 298 325 327	1	–	–	–	1	C1
128 209 223 298 325 327	1	–	–	–	1	C1
129 223 234 298 325 327	3	–	–	–	3	C1
129 223 274 298 325 327	1	–	–	–	1	C1
169 209 223 298 325 327	1	–	–	–	1	C1
169 216 223 298 325 327	1	–	–	–	1	C1
169 223 298 325 327	3	–	–	–	3	C1
172 223 298 325 327	2	–	–	6	8	C1
185 223 239 298 311 325 327	1	–	–	–	1	C1
185 223 298 311 325 327	1	–	–	–	1	C1
185 223 325 327	2	–	–	–	2	C1
209 223 289 298 325 327	1	–	–	–	1	C1
209 223 298 300 325 327	1	–	–	–	1	C1
209 223 298 325 327	1	–	–	1	2	C1
209 223 298 325 327 357	1	–	–	–	1	C1
209 223 325 327	1	–	–	1	2	C1
209 298 325 327	1	–	–	–	1	C1
216 223 298 325 327 356 362	1	–	–	–	1	C1
223 274 298 325 327	3	–	5	–	8	C1
223 298 325 327	6	19	111	195	331	C1
223 298 325 327 335	2	–	1	1	4	C1
223 325 362	2	5	34	35	76	C1
092 223 311 325 362	1	–	–	–	1	D1
142 188 223 325 362	1	–	–	–	1	D1
142 207 223 325 362	2	–	–	–	2	D1
172 185 192 223 301 342 362	1	–	–	–	1	D1
183C 189 223 270 325 362	1	–	–	–	1	D1
223 232A 325 362	1	–	–	–	1	D1
223 311 325 362	1	–	–	–	1	D1
Total	*99*	*485*	*2005*	*1657*	*949*	

AM-C, Central America; AM-N, North America; AM-S, South America; HG, haplogroup.

L-haplotypes each with west and west-central Africa, 12 haplotypes with southeast, and eight with southwest; the latter two are more surprising considering the relatively lower sample sizes of studies on these regions (Table 2) in comparison with the Atlantic African coast.

Haplogroup composition of the different admixed Colombian groups

The Mestizo population consists of 67 individuals. The majority (97%) carry Native American haplogroups: 40% belong to haplogroup A2, 13% to B2, 34% to C1, and 7% to D1. Two haplotypes are of European origin while only one belongs to the typical sub-Saharan haplogroup L2a. In contrast, the Mulato sample is dominated by L haplotypes (~81%); however, the sample size is low (n = 11) and therefore reliable frequency estimation of minor haplogroups is not feasible. The Afro-Colombians (n = 95) also carry predominantly L-haplotypes (72.6%), but the Native American component is also very significant (23.2%), comprising mtDNAs belonging to haplogroup C1, the main indigenous component (11%).

TABLE 2. Shared haplotypes between Colombian African lineages the main American and African regions

Colombian African lineages (minus 16000)	*n*	AM-C	AM-N	AM-S	AF-E	AF-N	AF-S	AF-SE	AF-SW	AF-W	AF-WC	Total	HG
129 148 168 172 187 223 230 278 293 311 320	1	–	–	–	–	–	–	–	–	–	–	1	L01
129 148 168 172 187 188G 189 223 230 278 293 311 320	2	1	2	2	5	1	1	22	4	–	11	51	L0a1
126 187 189 223 264 270 278 311	2	1	24	2	–	6	1	2	3	17	6	64	L1b
111 126 187 189 223 239 270 278 293 311	1	–	3	–	–	–	–	–	–	3	5	12	L1b1
126 187 189 223 248 264 270 278 293 311	1	–	–	–	–	–	–	–	–	–	–	1	L1b1
126 187 189 223 264 270 278 284 311	1	–	–	–	–	–	–	–	–	–	–	1	L1b1
126 187 189 223 264 270 278 293 311 355	1	–	–	–	–	–	–	–	–	–	–	1	L1b1
126 187 189 223 264 270 278 293 311 362	1	–	–	–	–	2	–	–	–	7	–	10	L1b1
092 126 184 187 189 223 278 294 301 311 360	1	–	–	–	–	–	–	–	–	–	–	1	L1c
093 129 140 184 187 189 223 278 294 301 311 360	2	–	1	–	–	–	–	–	–	–	–	3	L1c
104 129 163 187 189 223 278 293 294 311 360	1	–	–	–	–	–	–	–	–	–	–	1	L1c
129 163 187 189 223 278 293 294 304 311 360	1	–	–	–	–	–	–	–	–	16	2	19	L1c
129 169 172 187 189 192 223 261 278 293 311 360	1	–	–	–	–	–	–	–	–	–	–	1	L1c
129 183C 189 215 223 278 294 311 360	2	–	2	–	–	–	–	2	–	–	2	8	L1c
129 184 187 189 270 278 293 301 311	1	–	–	–	–	–	–	–	–	–	–	1	L1c
129 187 189 214 234 249 258 274 278 293 294 311 360	1	–	1	–	–	–	–	–	–	–	1–	12	L1c
129 187 189 223 265C 278 286G 294 311 359 360	1	–	1	1	–	–	–	–	2	–	–	5	L1c
129 187 189 223 278 284 293 294 311 360	2	–	5	–	–	–	–	–	–	–	–	7	L1c
129 187 189 223 274 278 293 294 311 360	1	–	1	1	–	–	–	2	–	–	24	29	L1c1a
187 189 223 278 284 293 294 311 360	1	–	–	–	–	–	–	–	–	–	–	1	L1c1a
093 213 223 290 294	1	–	–	–	–	–	–	–	–	–	–	1	L2a
183C 189 192 223 294 309 311	1	–	–	–	–	–	–	–	–	–	–	1	L2a
184 223 278 294	1	–	–	–	–	–	–	–	–	–	–	1	L2a
189 192 223 278 294	2	–	5	–	4	–	–	3	–	2	3	19	L2a
189 223 278 294 362	1	–	–	–	–	–	–	–	–	7	2	10	L2a
192 223 278 294	1	–	–	–	–	–	–	–	–	–	–	1	L2a
223 278 290 294 309	1	–	1	–	–	1	–	–	–	3	–	6	L2a
223 278 294 309	5	3	48	8	4	12	–	1–	1	23	2–	134	L2a
093 189 192 223 278 294 309	1	–	2	–	–	–	–	–	–	2	–	5	L2a1
126 223 278 309	3	–	1	–	–	–	–	–	–	–	–	4	L2a1
189 223 278 294 309	1	2	12	1	26	3	–	2	–	13	6	66	L2a1
193 213 223 239 278 294 309	1	–	2	–	–	–	–	–	–	–	1	4	L2a1
223 264 278	1	–	5	–	–	–	–	–	–	5	3	14	L2a1
215 223 278 286 294 309	1	–	–	–	–	–	–	–	–	–	–	1	L2a1a
114A 129 142 213 223 325 362	1	–	–	–	–	–	–	–	–	–	–	1	L2b1
129 213 223 278 354	1	–	–	–	–	–	–	–	–	–	–	1	L2b1
169 223 264 278	2	–	–	–	–	–	–	–	–	–	–	2	L2c2
093 223 278	1	–	–	–	–	–	–	–	–	5	–	6	L3*?
223 319	1	–	–	–	1	1	–	–	–	–	–	3	L3*?
124 183C 187A 189 223 278 362	1	–	–	–	–	–	–	–	–	–	–	1	L3b
124 223 240 278 362	1	–	–	–	–	–	–	–	–	–	–	1	L3b
124 223 362	1	–	2	–	–	–	–	–	–	–	2	5	L3b
223 240 278 362	1	–	–	–	–	–	–	–	–	–	–	1	L3b
124 166 223	1	–	2	–	–	1	–	–	–	3	7	14	L3d
124 223	5	2	8	1	2	3	–	3	–	2–	12	56	L3d
124 223 291	1	1	3	–	–	–	–	–	–	1	–	6	L3d
093 148 223 265T	1	–	3	1	–	–	–	–	–	–	–	5	L3e
185 223 327	1	–	4	2	–	–	1	7	3	–	–	18	L3e1a
213 223 265 320	1	–	–	–	–	–	–	–	–	–	–	1	L3e3
223 265T	1	–	15	–	5	1	1	1–	2	–	7	42	L3e3
223 264 319	1	–	–	–	–	–	–	–	–	–	–	1	L3e4
209 223 311	2	–	3	1	1	5	–	3	8	2	13	38	L3f
209 218 223 256 292 311	2	–	2	–	–	–	–	2	2	–	–	8	L3f1
209 223 235 292 311	1	–	–	–	–	–	–	–	–	2	–	3	L3f1
209 223 292 295 311	5	1	4	–	–	–	–	–	–	13	–	23	L3f1
Total	*78*	*83*	*1148*	*143*	*835*	*1312*	*264*	*416*	*157*	*1184*	*1202*	*732*	

AM-C, Central America; AM-N, North America; AM-S, South America; AF-E, East Africa; AF-N, North Africa; AF-S, South Africa; AF-SE, Southeast Africa; AF-SW, Southwest Africa; AF-W, West Africa; AF-WC, West-Central Africa.

The African component of the Afro-Colombian sample is less diverse in terms of haplotype diversity than the African component of Mulatos (Table 2), but this difference is not statistically significant because of the large standard error affecting the Mulato's sample (Table 3). Nucleotide diversity is slightly higher, but significantly so, in the African mtDNA component of Afro-Colombians than in the L- lineages observed in Mulatos.

On the other hand, the Native American component of Afro-Colombians is more diverse in terms of haplotype diversity than its counterpart in Mestizo, but again, this difference is not statistically significant (Table 3). The opposite pattern occurs for the nucleotide diversity values.

As expected the nucleotide diversity of the African lineages in Colombians is higher than the values for the Native American mtDNAs. The same occurs for the haplotype diversity, although the differences are not statistically significant (Table 3).

The phylogeographic features of admixed Colombians

The most common Native American haplotypes in Colombians match some of the most common haplotypes in America; and this is particularly true for the root haplotypes of the main Native American haplogroups. There is no evidence of important founder effects in our sample;

TABLE 3. Diversity indices for HVS-I (sequence range 16024–16365) in Colombian populations

	n	K (K/n)	π (SE)	H (SE)	M
Afro-Colombians	95	71 (74.7)	0.025 (0.001)	0.991 (0.003)	8.4
African component	69	48 (70.3)	0.024 (0.001)	0.985 (0.006)	8.3
Native American component	22	19 (82.6)	0.018 (0.001)	0.984 (0.017)	6.0
Mestizo	67	43 (64.2)	0.021 (0.001)	0.974 (0.009)	7.0
Native American component	64	40 (62.5)	0.020 (0.001)	0.971 (0.010)	6.9
Mulato	11	11 (100.0)	0.021 (0.003)	1.000 (0.039)	7.1
African component	9	9 (100.0)	0.020 (0.003)	1.000 (0.052)	6.9

K = number of different sequences found and percentage of sample size in brackets; π = nucleotide diversity; H = haplotype diversity; M = average number of pairwise difference.

however, some haplotypes appear at relatively high frequencies in Colombia. For instance, A2 haplotype: C16111T G16213A C16223T C16290T G16319A T16362C occurs 10 times in this sample of Colombia, but only five times in previous studies, three of them in an independent Colombian sample (Torres et al., 2006) and two in the Genographic dataset (https://www3.nationalgeographic.com/genographic/resources.html). The one-step mutation haplotypes carrying transition: T16209C on top of the later haplotype is highly prevalent in Venezuela; of the 12 haplotypes observed in the whole database, 11 are found in the indigenous Guahibo people (Vona et al., 2005). Haplotype C16111T C16223T C16290T G16319A T16362C is observed at high frequency in the Haida from Queen Charlotte Islands (North America), but it is also present in the Xavante Brazilians (Ward et al., 1996) or in the Kuna from Panama (Batista et al., 1995). In the neighboring country of Ecuador (Rickards et al., 1999) we also find the only three matching sequences of the Colombian type: C16111T A16183C T16189C C16223T C16287T C16290T G16319A T16362C (either including or excluding the highly unstable variant A16183C). In the Ngöbe of Panama (Kolman et al., 1995) we detect the majority of matching sequences for haplotype: C16111T C16187T C16223T C16290T G16319A T16362C. Also particularly noticeable is the fact that perfect matches of the A2 haplotype: C16111T C16223T C16290T G16319A are only found in the Inuit from Canada. Finally, we observed identical matches of the haplotype C16111T C16223T C16290T G16319A T16356C T16362C in another independent Colombian sample (Torres et al., 2006) as well as in some regions neighboring Colombia.

It can be said that the Native American component of admixed Colombians is more closely phylogenetically related to northern-central South America (specifically the northern regions that encompass northern Brazil, Ecuador, and Venezuela) than to any other American region.

Most of the African lineages found in Colombia are present in Africa. Only four of them were not previously observed in this continent or in Afro-American samples; however, some of these mtDNAs have well-known one-step mutation representatives in Africa. The most frequent L-haplotypes in Colombia are also common in Africa; for instance, the L1b haplotype characterized by the following substitutions: T16126C C16187T T16189C C16223T C16264T C16270T C16278T T16311C is the root type of haplogroup L1b and was observed 16 times in our Colombian sample and 62 times in Africa, specially in West but also in North, Central, and East at lower frequencies; it is also common in other parts of America (see Fig. 4 in Salas et al., 2002). The L1b1 lineage: T16126C C16187T T16189C C16223T C16264T C16270T C16278T A16293G T16311C T16362C was only observed in three different Atlantic Islands: Azores, Cape Verde, and Canary Islands; note that the two latter locations were important ports for slave transport from Africa to America. Another common haplotype in Colombia is the root type of L2a: C16223T C16278T C16294T A16309G (n = 13 in Colombia) which is also widespread in Africa and in other parts of America.

The origin of L1c haplogroup is still uncertain, although Central Africa is the most likely candidate source region (Quintana-Murci et al., 2007; Salas et al., 2002, 2004, 2005b); it is however more difficult to account for the origin of the large proportion of L1c American lineages that do not find matches in present day Africa. Three different L1c haplotypes were observed in Colombia. Haplotype G16129A C16187T T16189C C16223T A16265C C16278T C16286G C16294T T16311C T16359C C16360T is observed in, e.g., Brazil (Silva et al., 2006) and could have originated from Angola, where it has been observed at least twice (Plaza et al., 2004), whereas the one-step mutation derivative: G16129A C16187T T16189C C16223T A16265C C16278T C16286G C16294T T16311C T16360C was observed once in the neighboring African population of Cabinda (Beleza et al., 2005). Haplotype: L1c2 G16129A A16183C T16189C A16215G C16223T C16278T C16294T T16311C T16360C could readily originate from Equatorial Guinea (Mateu et al., 1997), where it was found at high frequency (eight matches). Other continental locations are also good candidates, notably Central Africa is a likely source region because there are few one-step mutation haplotypes located in this area. The origin of the common Colombian L1c mtDNA characterized by substitutions: T16093C G16129A T16140C C16184T C16187T T16189C C16223T C16278T C16294T 16301 T16311C T16360C (n = 11) is more enigmatic; it was observed only once in America (Monson et al., 2002).

The L3e haplotype: T16093C C16148T C16223T A16265T was only observed in America in the SWGDAM database (Monson et al., 2002) and in another Colombian population with an important recent African ancestry from the department of Chocó (Bravo et al., 2001; Paredes et al., 2003; Salas et al., 2005c), while there are also three matches in the Genographic database for this haplotype. The L3e1a Colombian profile: C16185T C16223T C16327T was detected several times in Mozambique (Pereira et al., 2001; Salas et al., 2002) but also in Cabinda (Beleza et al., 2005) and Angola (Plaza et al., 2004); these two southwest African locations are good candidate source regions considering their important role in the past as source populations for slavery to America.

The 10 Páez Indian samples analyzed in this study carry Native American mtDNAs; half of them belong to haplogroup A2, four individuals bear haplogroup C1 haplotypes, and one individual carries a B2 mtDNA.

Principal component analysis

The two first principal components (PC1 and PC2) account for 61% of the observed variability (see Fig. 1).

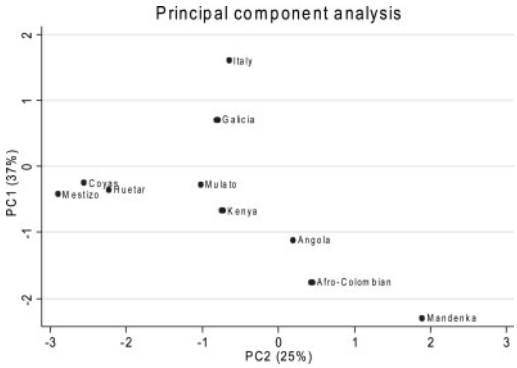


Fig. 1. Scatter plot showing the first two principal component analysis of haplogroup frequency profiles for several populations, including the Colombian samples (Mestizo, Mulato, and Afro-Colombians).

PC1 (37% of the genetic variance) clusters in one pole the sub-Saharan populations of Angola (Plaza et al., 2004) and Mandenka (Graven et al., 1995) together with the Afro-Colombians; on the opposite side of the scatter plot are the European populations of Italy (Tagliabracci et al., 2001) and Galicia, southwest Iberia (Salas et al., 1998); the rest of the populations stand in between (see Fig. 1). The PC2 (25% of the genetic variance) clearly separates the Native American populations of Coyas (Álvarez-Iglesias et al., 2007) and Huetar (Santos and Barrantes, 1994) together with the Colombian Mestizos, from the rest of the populations (see Fig. 1). Finally, the small Mulato sample is closely located to the Kikuyu (Watson et al., 1996), showing mtDNA affinities to sub-Saharan Africa (see Fig. 1). As expected, the main determinants of the PC1 pattern are in general the whole set of African (L-haplogroups) and European lineages, while the main determinants of PC2 are the Native American haplogroups A2, B2, C1, and D1.

DISCUSSION

We have dissected the different mtDNA ancestries of various Colombian admixed "self-reported" population groups from a predominantly phylogeographic perspective.

The Native American component of Colombians is more closely related to Central America and northern South America. Some lineages have high frequencies in Colombia, but we did not observe evidence for the action of genetic drift in these populations. This is also manifested by the high values of the diversity indices computed (Table 3), demonstrating that an important amount of autochthonous Native variation in many admixed populations has survived the demographic impact of the western European colonization (notably in the case of Colombia from Spain).

Several past studies have attempted to disentangle the complex African contribution to America arising from the slave trade, either on a continental scale (Salas et al., 2002, 2004, 2005b), or focusing on contributions on more regionalized variation (Alves-Silva et al., 2000; Bravi et al., 1997; Carvajal-Carmona et al., 2000; Ely et al., 2006; Parra et al., 1998; Salas et al., 2005c). We show that the main African component in Colombia appears to be West/west-central, southeast, and southwest Africa, in agreement with previous findings (Salas et al., 2004, 2005b) and historical records. The amount of variation in the African component is also comparable with that found in the original African continent; this reflects the large number of slaves forcibly moved to Colombia and other American regions which constitutes a large effective population size.

The PC analysis accounts for a substantial percentage of the variability (61%) and helps to summarize the global haplogroup composition of the admixed Colombian populations here analyzed: Mulatos and Afro-Colombians cluster together with other sub-Saharan samples, whereas the Mestizo sample locates close to other Native American groups reflecting the high percentage of L-lineages in the two former samples and the large amount of Native American haplogroups in the latter.

An important aspect of this study is the possibility of detailed evaluation of self-reported ethnicity, used during the sampling process, as a proxy for genetic ancestry. In a previous study (Rodas et al., 2003) it was reported that the frequency of L lineages in five different Afro-Colombian population samples ranged from 21.4% (Quibdó) to 52.5% (Providencia). In our Colombian sample, we found that 72.6% of the lineages belong to some L-haplogroup; this pattern differs significantly both from the Quibdó (Fisher's exact test; $P < 0.0000$) and from the Providencia sample (Fisher's exact test; $P = 0.0203$). The frequency of the Native American component in the Mestizo also differs significantly in both studies, 78% in Rodas' study (2003) *versus* ~97% in our Mestizo sample (Fisher's exact test; $P = 0.0004$). We therefore observe that independently collected samples of the same self-reported population affiliations from Colombia can lead to quite different haplogroup spectra. This has been observed to occur in the study of other admixed populations from other regions or from the use of population descriptors that have a larger cultural than geographical point of reference such as "Hispanic" (Salas et al., 2007). Our results therefore do not support the use of the terminology still found in many population studies that are outmoded and inaccurate (e.g. Mestizo), lack a correctly defined geographic description (e.g. Caucasian rather than European), or have a predominantly cultural or linguistic definition (e.g. Hispanic). Our study and numerous others clearly illustrate that, when the genetic markers used have high resolution for characterizing the ancestry of an individual, these terms incorrectly denote membership of a homogenous group. Individuals belonging to different and variable mtDNA genetic ancestries may use self-descriptive terms that fail to reflect these differences, and studies that rely on such descriptions run the risk of spurious conclusions based on presumed genetically homogeneous groups.

Assumptions about the homogeneity of a group of individuals that are grouped into improperly defined populations can have consequences for forensic genetics and association studies. Undetected population stratification has an important bearing in forensic casework if an expert consulting a database wrongly assumes that the populations (lacking correctly defined genetic ancestry) are genetically homogenous. Such a consideration was highlighted by a recent review of the widely used SWGDAM forensic database (Salas et al., 2007). In case–control association studies, undetected population substructure is a well-known cause of type I error (Mosquera-Miguel et al.,

2008). Often cases and control subjects are matched by recombining individuals ethnically (e.g. both groups are Hispanics) in order to prevent the inflation of association values caused by differences in ancestry between each study group (Salas and Carracedo, 2007). However, we wish to conclude by emphasizing that the only way to prevent the risk of spurious findings in the above situations is to properly define the population in terms of genetic ancestry and levels of admixture using the most informative markers and appropriate sampling designs.

LITERATURE CITED

Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, Woodward SR, Salas A, Torroni A, Bandelt HJ. 2008. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3(3):e1764.

Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A. 2007. Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int: Genet 1:44–55.

Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt H-J, Pena SD, Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67:444–461.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge Reference Sequence for human mitochondrial DNA. Nat Genet 23:147.

Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V. 2002. The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150–1160.

Bandelt H-J, Salas A, Bravi CM. 2004a. Problems in FBI mtDNA database. Science 305:1402–1404.

Bandelt H-J, Salas A, Lutz-Bonengel S. 2004b. Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118:267–273.

Batista O, Kolman CJ, Bermingham E. 1995. Mitochondrial DNA diversity in the Kuna Amerinds of Panama. Hum Mol Genet 4:921–929.

Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S, Carvajal L, Labuda D, Alvarez V, Ospina J, Hedrick PW, et al. 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc Natl Acad Sci USA 103:7234–7239.

Beleza S, Gusmão L, Amorim A, Carracedo Á, Salas A. 2005. The genetic legacy of western Bantu migrations. Hum Genet 117:366–375.

Bravi CM, Sans M, Bailliet G, Martinez-Marignac VL, Portas M, Barreto I, Bonilla C, Bianchi NO. 1997. Characterization of mitochondrial DNA and Y-chromosome haplotypes in a Uruguayan population of African ancestry. Hum Biol 69:641–652.

Bravo ML, Moreno MA, Builes JJ, Salas A, Lareu MV, Carracedo A. 2001. Autosomal STR genetic variation in negroid Choco and Bogota populations. Int J Legal Med 115:102–104.

Carracedo A, Bar W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, et al. 2000. DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 110:79–85.

Carvajal-Carmona LG, Ophoff R, Service S, Hartiala J, Molina J, Leon P, Ospina J, Bedoya G, Freimer N, Ruiz-Linares A. 2003. Genetic demography of Antioquia (Colombia) and the central valley of Costa Rica. Hum Genet. 112:534–541.

Carvajal-Carmona LG, Soto ID, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, McCarthy M, Montoya P, Alvarez VM, Bedoya G, et al. 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am J Hum Genet 67:1287–1295.

Černý V, Salas A, Hájek M, Žaloudková M, Brdička R. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71:433–452.

Curtin D. 1979. The Atlantic slave trade 1600-1800. In: Ajayi JFA, Crowder M, editors. New York: Columbia University Press.

Ely B, Wilson JL, Jackson F, Jackson BA. 2006. African-American mitochondrial DNAs often match mtDNAs found in multiple African ethnic groups. BMC Biol 4:34.

Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L. 1995. Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol Biol Evol 12:334–345.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752–770.

Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F. 1995. Reduced mtDNA diversity in the Ngöbé Amerinds of Panamá. Genetics 140:275–283.

Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, Achilli A, Wang CY, Zhong L, Zhu CL, Wu SF, et al. 2006. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15:2076–2086.

Mateu E, Comas D, Calafell F, Pérez-Lezaun A, Abade A, Bertranpetit J. 1997. A tale of two islands: population history and mitochondrial sequence variation of Bioko and São Tomé, Gulf of Guinea. Annals of Hum Genet 61:507–518.

Melton PE, Briceno I, Gomez A, Devor EJ, Bernal JE, Crawford MH. 2007. Biological relationship between Central and South American Chibchan speaking populations: evidence from mtDNA. Am J Phys Anthropol 133:753–770.

Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B. 2002. The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Sci Commun4.

Mosquera-Miguel A, Álvarez-Iglesias V, Vega A, Milne R, Cabrera de León A, Benitez J, Carracedo Á, Salas A. 2008. Is mitochondrial DNA variation associated with sporadic breast cancer risk? Cancer Res 68:623–625.

Paredes M, Galindo A, Bernal M, Ávila S, Andrade D, Vergara C, Rincón M, Romero RE, Navarrete M, Cardenas M, et al. 2003. Analysis of the CODIS autosomal STR loci in four main Colombian regions. Forensic Sci Int 137:67–73.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, et al. 1998. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63:1839–1851.

Pereira L, Macaulay V, Torroni A, Scozzari R, Prata M-J, Amorim A. 2001. Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Annals Hum Genet 65:439–458.

Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo Á, Comas D. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet 115:439–447.

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, et al. 2007. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci USA 105: 1596–1601.

Rickards O, Martinez-Labarga C, Lum JK, De Stefano GF, Cann RL. 1999. mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations. Am J Hum Genet 65:519–530.

Rodas C, Gelvez N, Keyeux G. 2003. Mitochondrial DNA studies show asymmetrical Amerindian admixture in Afro-Colombian and Mestizo populations. Hum Biol 75:13–30.

Salas A, Bandelt HJ, Macaulay V, Richards MB. 2007. Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168:1–13.

Salas A, Carracedo Á. 2007. Studies of association in complex diseases: Statistical problems related to the analysis of genetic polymorphisms. Rev Clin Esp 207:563–565.

Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J. 2005a. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335:891–899.

Salas A, Carracedo Á, Richards M, Macaulay V. 2005b. Charting the ancestry of African Americans. Am J Hum Genet 77:676–680.

Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo Á. 1998. mtDNA analysis of the Galician population: a genetic edge of European variation. Eur J Hum Genet 6:365–375.

Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo Á. 2002. The making of the African mtDNA landscape. Am J Hum Genet 71:1082–1111.

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo Á. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74:454–465.

Salas A, Richards M, Lareu MV, Sobrino B, Silva S, Matamoros M, Macaulay V, Carracedo Á. 2005c. Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. Am J Phys Anthropol 128:855–860.

Salas A, Yao Y-G, Macaulay V, Vega A, Carracedo Á, Bandelt H-J. 2005d. A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2:e296.

Santos M, Barrantes R. 1994. D-loop mtDNA deletion as a unique marker of Chibchan Amerindians. Am J Hum Genet 55:413–414.

Silva WA, Bortolini MC, Schneider MP, Marrero A, Elion J, Krishnamoorthy R, Zago MA. 2006. MtDNA haplogroup analysis of black Brazilian and sub-Saharan populations: implications for the Atlantic slave trade. Hum Biol 78:29–41.

Tagliabracci C, Turchi C, Buscemi L, Sassaroli C. 2001. Polymorphism of the mitochondrial DNA control region in Italians. Int J Legal Med 114:224–228.

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, et al. 2007. Beringian standstill and spread of Native American founders. PLoS ONE 2:e829.

Torres MM, Bravi CM, Bortolini MC, Duque C, Callegari-Jacques S, Ortiz D, Bedoya G, Groot de Restrepo H, Ruiz-Linares A. 2006. A revertant of the major founder Native American haplogroup C common in populations from northern South America. Am J Hum Biol 18:59–65.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. 2006. Harvesting the fruit of the human mtDNA tree. Trends Genet 22:339–345.

Vona G, Falchi A, Moral P, Calo CM, Varesi L. 2005. Mitochondrial sequence variation in the Guahibo Amerindian population from Venezuela. Am J Phys Anthropol 127:361–369.

Ward RH, Salzano FM, Bonatto SL, Hutz MH, Coimbra CEA, Santos RV. 1996. Mitochondrial DNA polymorphism in 3 Brazilian Indian tribes. Am J Hum Biol 8:317–323.

Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Pääbo S. 1996. mtDNA sequence diversity in Africa. Am J Hum Genet 59:437–444.

OPEN ACCESS Freely available online

PLoS one

New Population and Phylogenetic Features of the Internal Variation within Mitochondrial DNA Macro-Haplogroup R0

Vanesa Álvarez-Iglesias[1☯], Ana Mosquera-Miguel[1☯], Maria Cerezo[1], Beatriz Quintáns[2], Maria Teresa Zarrabeitia[3], Ivon Cuscó[4], Maria Victoria Lareu[1], Óscar García[5], Luis Pérez-Jurado[4,6], Ángel Carracedo[1,2], Antonio Salas[1]*

1 Unidade de Xenética, Instituto de Medicina Legal and Departamento de Anatomía Patolóxica y Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain, **2** Fundación Pública Galega de Medicina Xenómica (FPGMX), and Ciber de enfermedades raras (CIBERER), Hospital Clínico Universitario, Universidade de Santiago de Compostela, Galicia, Spain, **3** Medicina Legal, Universidad de Cantabria, Santander, Spain, **4** Unidad de Genética, Universitat Pompeu Fabra, and U735 CIBER de enfermedades raras (CIBERER), Barcelona, Spain, **5** Laboratorio de la Ertzaintza, Bilbao, Spain, **6** Programa de Medicina Molecular y Genética, Hospital Universitari Vall d'Hebron, Barcelona, Spain

Abstract

Background: R0 embraces the most common mitochondrial DNA (mtDNA) lineage in West Eurasia, namely, haplogroup H (~40%). R0 sub-lineages are badly defined in the control region and therefore, the analysis of diagnostic coding region polymorphisms is needed in order to gain resolution in population and medical studies.

Methodology/Principal Findings: We sequenced the first hypervariable segment (HVS-I) of 518 individuals from different North Iberian regions. The mtDNAs belonging to R0 (~57%) were further genotyped for a set of 71 coding region SNPs characterizing major and minor branches of R0. We found that the North Iberian Peninsula shows moderate levels of population stratification; for instance, haplogroup V reaches the highest frequency in Cantabria (north-central Iberia), but lower in Galicia (northwest Iberia) and Catalonia (northeast Iberia). When compared to other European and Middle East populations, haplogroups H1, H3 and H5a show frequency peaks in the Franco-Cantabrian region, declining from West towards the East and South Europe. In addition, we have characterized, by way of complete genome sequencing, a new autochthonous clade of haplogroup H in the Basque country, named H2a5. Its coalescence age, 15.6±8 thousand years ago (kya), dates to the period immediately after the Last Glacial Maximum (LGM).

Conclusions/Significance: In contrast to other H lineages that experienced re-expansion outside the Franco-Cantabrian refuge after the LGM (e.g. H1 and H3), H2a5 most likely remained confined to this area till present days.

Citation: Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, et al. (2009) New Population and Phylogenetic Features of the Internal Variation within Mitochondrial DNA Macro-Haplogroup R0. PLoS ONE 4(4): e5112. doi:10.1371/journal.pone.0005112

Editor: Vincent Macaulay, University of Glasgow, United Kingdom

Received December 8, 2008; **Accepted** March 9, 2009; **Published** April 2, 2009

Funding: This research received support from Xunta de Galicia (Grupos Emerxentes; 2008/XA122), two grants from the Fundación de Investigación Médica Mutua Madrileña, and a grant from the Ministerio de Ciencia e Innovación (SAF2008-02971) given to AS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: antonio.salas@usc.es

☯ These authors contributed equally to this work.

Introduction

Haplogroup R0, formerly known as pre-HV [1], is defined by the absence of transitions A73G and G11719A relative to haplogroup R. There is one main sub-branch of R0 defined by the lack of C14766T (haplogroup HV) and a minor branch known as R0a [2]. HV embraces the most frequent haplogroup in Europe (~40%), namely, haplogroup H, which is defined by the lack of the characteristic transitions A2706G and C7028T. HV also contains some other less frequent clades, such as HV1, HV2, and specially HV0, where haplogroup V is nested. Most of the haplogroup H sub-lineages are most likely of Middle Eastern origin (as it is the case of the majority of the typical West European clades). Overall, R0 shows frequency patterns declining from West towards East and South Europe and Middle East [1,3,4].

By way of complete genome sequencing, Achilli et al. [1] identified numerous sub-branches of haplogroup H. These authors demonstrated for the first time that, although haplogroup H overall in Europe is rather uniform, the sub-clades H1 and H3 show frequency peaks centered in Iberia and surrounding areas. The phylogeographic distribution of these lineages and their coalescence ages (~11 kya) lead these authors to conclude that H1 and H3 represent a signal of late-glacial expansion of hunter-gatherers that repopulated Central and Northern Europe from about 15,000 years ago, after the LGM. These patterns mirror those previously observed by the same authors for haplogroup V [5], which also shows a clear-cut clinal geographical distribution in Europe, with a peak in the

Franco-Cantabrian area and coalescence ages ranging from 16.3±4.8 kya in West Europe to 8.5±2.3 kya in the East Europe. Therefore, the geographic and diversity patterns of these three lineages pointed to a re-colonization period of Europe from western refuge locations after the LGM period. Apart from adding substantially to the resolution of the haplogroup H phylogeny, another contemporary study [3] also showed that some lineages such as H1*, H1b, H1f, H2a, H3, H6a, H6b, and H8, have distinct phylogeographic patterns within Europe. The study by Brandstätter et al. [6] further contributed to the dissection of the phylogeny of haplogroup H, although with a more technical perspective. More recently, Roostalu et al. [4] studied the population stratification of haplogroup H sub-lineages in West Eurasia, with special focus to Near Eastern populations and the Caucasus. Again, the authors demonstrated that most of the haplogroup H lineages of present-day Near Eastern-Caucasus area expanded after the LGM and presumably before the Holocene. The study of Abu-Amero et al. [2] was also very useful in providing further resolution at the level of complete genome sequencing within R0a. The refined knowledge of the mtDNA phylogeny to the level of complete genomes opened the doors to a wide spectrum of different applications of medical and forensic interest; see also [7,8,9,10].

On the other hand, the mtDNA phylogeny needs continuous updating in order to ease future population and phylogenetic studies (e.g. [11,12]). Due to the growing interest of geneticists in unraveling the internal variation within haplogroup H, several conflicts have arisen in the phylogeny and nomenclature of R0 sub-clades. For example, the recent publication of Roostalu et al. [4] added new branches to the phylogeny of haplogroup H, but for instance, their label H19 was used to name a different branch in the contemporary study of Achilli et al. [1]. To give another example, based on complete genome sequencing, Behar et al. [13] referred to a new clade, R0a1, with three minor sub-clades (R0a1a, R0a1b, and R0a1c); however, they did not notice the contribution by Abu-Amero et al. [2], where new complete genomes and new sub-branches of R0a had been reported; thus, for instance, the R0a1 in Behar et al. [13] matched a branch previously coined (preHV)1b by Abu-Amero et al. [2] (therefore using also the old nomenclature; see [14]). Some of the problems related to R0a were mentioned in Brandstätter et al. [6] and Brandstätter et al. [15]; although many problems still remain (see below).

The goals of the present study are: i) provide new insights into the distribution and population variability of haplogroup H sub-lineages in North Iberia to a high level of phylogenetic resolution; ii) resolve the many existing conflicts in the nomenclature and phylogeny of R0 that nowadays represent a challenge for future inter-population studies; iii) refine the phylogeny of R0 by way of inspecting the existing mtDNA complete genomes (plus coding region segments) available in the literature and GenBank (>1,100); and iv) contribute to enrich the known phylogeny of haplogroup H at the level of complete genome sequencing, by characterizing a new autochthonous clade observed in the Basque Country, namely H2a5.

Methods

Samples

We have collected samples from three main North Iberia regions. A total of 282 healthy unrelated individuals were obtained from Galicia (northwest Iberia) (which is an independent sample to the one reported in [16,17]). Three different locations were sampled in Cantabria (N= 135; North-Central Iberia), including 39 healthy unrelated individuals from Valle del Pas, 45 from Valle del Liébana, and 51 from Santander. Several individuals from Valle del Pas were previously reported for the HVS-I segment in Maca-Meyer et al. [18] . For most of the analysis, these three locations were lumped in a single group (Cantabria). A total of 101 individuals suffering autism were collected from Catalonia (northeast Iberia). Since mtDNA lineages do not play a role as medium to high penetrance factors in autism (which is likely to be a polygenic and multifactorial disease), this sample can be considered to represent (from a mtDNA point of view) a random sample from the region (a case-control association study performed by the authors adds support to this statement). Finally, eight samples carrying substitution C4592T in the sample set of 75 individuals from the Basque Country (bordering East Cantabria) screened in [19], and presumably belonging to a new minor clade of haplogroup H (here baptized as H2a5), were selected for complete genome sequencing.

Oral informed consent was required for the samples collected in Galicia and Cantabria, and all of them were anonymized. Written informed consent was required for the samples collected in Catalonia and were also anonymized; then, DNA extracts were submitted to the laboratory in Santiago de Compostela were the genotyping was carried out. In addition, the study was approved by the Ethical committee of the University of Santiago de Compostela. The study conforms to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July)."

Genotyping protocols and nomenclature

All the samples from Galicia, Cantabria, and Catalonia were sequenced for the HVS-I region (N= 518). Those samples belonging to R0 or with a dubious adscription to other non-R0 haplogroups (N= 293; ~57%) were further genotyped for a set of 71 coding region SNPs mainly defining different branches within R0 (more information below). A total of 283 samples (~55%) were finally classified as belonging to some R0 sub-branch.

The protocol for PCR amplification and automatic minisequencing is fully described in Text S1. Protocols for automatic sequencing of control region mtDNAs and complete genome sequencing are also shown in Text S1.

MtDNA variation is referred to the revised Cambridge Reference Sequence [20]. Haplogroup nomenclature is based on previous studies [1,2,3,4,5,6,7,13,15,21,22,23]. As introduced above, an important number of conflicts have arisen among past studies, most of them because of the neglect of already existing nomenclature, or the delay of updating results from information available in the literature, or simply because overlapping of different publications. In order to reconcile the nomenclature conflicts between different studies, we have followed a chronological criterion when possible but only in case this nomenclature was harmonious with the almost worldwide accepted nomenclature rules and phylogenetic features [11].

Monitoring genotyping errors

We have used the mtDNA tree as a reference to avoid as much as possible artefactual profiles and documentation errors in mtDNA sequences and in SNP genotypes [24,25,26,27,28,29]. When detecting some unexpected SNP pattern, we confirmed the genotypes by repeating the SNP genotyping using single-plex minisequencing and automatic sequencing, as performed in Álvarez-Iglesias et al. [9].

Genetic diversity estimates and analysis of geographical patterns

DnaSP 4.10.3 software [30] was used for the computation of different diversity indices, including haplotype and nucleotide

diversities and mean number of pairwise differences [31,32,33]. Departure from normal distribution of pairwise differences was checked using the Harpending's r (raggedness) index [34]. Selective neutrality was tested using the Tajima [35] and Fu and Li tests [36].

The geographical representations of haplogroup frequencies were obtained using Surfer 8.0 (http://www.goldensoftware.com). The data used was collected from previous studies [3,4,6,14] and the present one, representing a total of 24 different population samples. We used the inverse-squared distance method. Haplogroup frequencies are presented in a regular grid covering part of Eurasia (including Europe), Middle East and the Arabian Peninsula. Only data points within the same landmass, either island or continent, were considered for interpolation. In addition, we carried out analysis of spatial autocorrelation using the Spatial Autocorrelation Analysis Program (SAAP; http://www.exetersoftware.com/cat/saap.html) in order to detect and evaluate statistically signals of gradients (clines), gradients irradiating from the center of a particular area (depressions) or isolation by distance models; see for instance Barbujani [37].

Phylogeographic analysis

R0 and its different sub-lineages are the main focus of the present article; however, there are only few studies focusing on the internal variability of R0 suitable for population comparisons [1,3,4,14,38,39]. In addition, different haplotype searches were carried out using literature mtDNA datasets, most of them containing just HVS-I data; thus, in the literature there are more than 30,000 West Eurasian mtDNA profiles available that can be used for phylogeographic purposes.

Estimation of the time to the most recent common ancestor of each cluster and SDs were carried out according to Saillard et al. [40] and employing an evolutionary rate estimate which is intermediate between the one provided in [41] and [42], according to the procedures followed in [43]. Thus, the calibration corresponds now to 4,610 years per substitutions considering all the substitutions in the entire coding-region.

Results

The rationale for SNP selection and the R0 phylogeny

R0 differs from R* by lacking A73G and G11719A. R0 contains haplogroup HV which likewise embraces the most common haplogroup in Europe, H, but also haplogroup HV0a (where haplogroup V is nested) and some other minor branches such as HV1 and HV2. Within haplogroup H, there are at least 25 sub-haplogroups; many of them can be further sub-divided into minor branches.

MtDNA coding region SNP genotyping has been designed here with the aim of covering as much as possible the R0 phylogeny; given priority to those SNPs representing the most frequent sub-lineages, and also those characterizing branches that do not have any known diagnostic polymorphism in the control region. SNP selection in the present study considers the full set of SNPs reported in Brandstätter et al. [6] (with the exception of SNP A14552G which is replace here by C3936T; both leading to haplogroup H12) plus a selection of additional variants that define further sub-branches of R0 within Europe; see e.g. [4]. In addition, the analysis of the literature and complete genomes sequences available in GenBank has allowed us to infer new minor sub-lineages of R0 (see e.g. Text S2 and Table S1).

When selecting mtDNA SNPs, it called our attention the many inconsistencies existing in the nomenclature of haplogroup H and its sub-lineages. One of the aims of the present study was therefore to resolve these nomenclature conflicts in order to ease inter-population genetic studies. These problems and the rationale to determine new sub-branches of R0 are shown in Text S2 and Table S1. The updated classification tree of haplogroup R0 and its sub-clades is shown in Figure 1 and Figure 2. These figures also indicate the SNPs selected and genotyped in the present study. We also incorporated in the minisequencing assays various diagnostic sites for haplogroups HV1 and HV2 (sister clades of H and HV0), and other polymorphisms covering several major branches of haplogroup R, namely, haplogroup U (A12308G) and JT

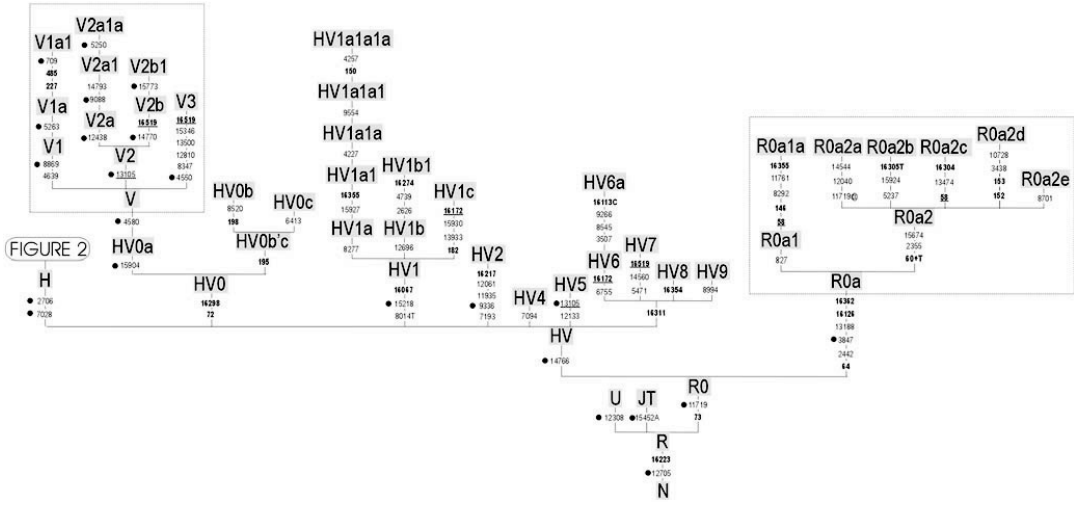


Figure 1. Phylogeny of haplogroup R0. An expanded view of the haplogroup H phylogeny is shown in Figure 2. Underlined positions signal parallel mutations, while @ indicates a back mutation. In bold are the control region variants, whereas dots indicate the SNPs selected and genotyped in the present study.
doi:10.1371/journal.pone.0005112.g001

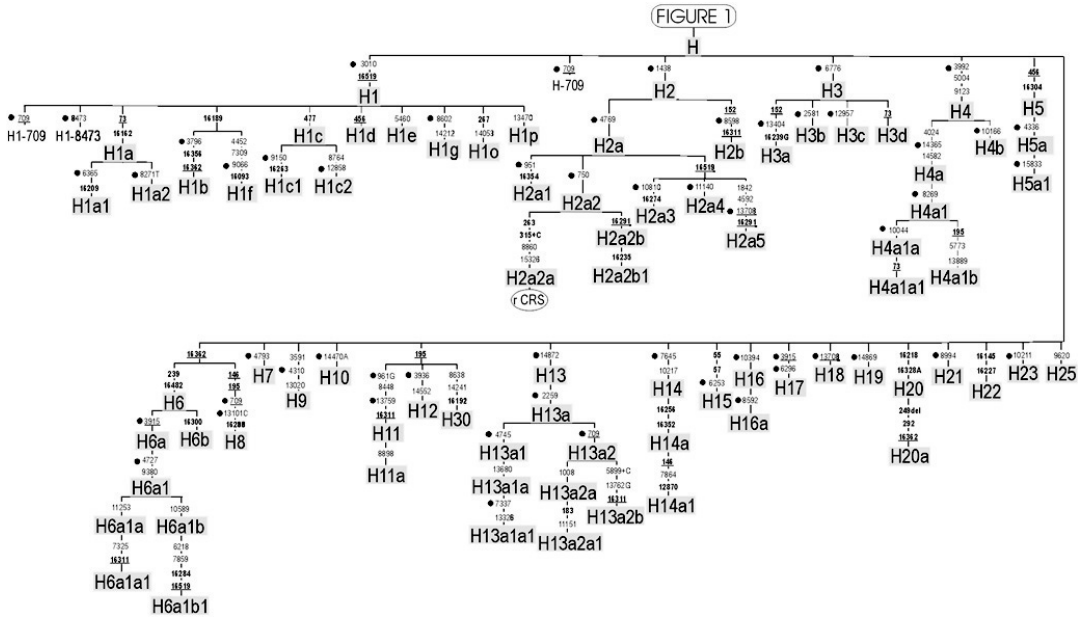


Figure 2. Phylogeny of haplogroup H. See legend of Figure 1 for more details.
doi:10.1371/journal.pone.0005112.g002

(C15452A). The transition C12705T defining macro-haplogroup N was also added.

The advantages of using a minisequencing multiplex genotyping procedure *versus* other mtDNA SNP genotyping methods are reported and explained in Text S3. Some phylogenetic inconsistencies have been observed in our data, but all of them were confirmed by sequencing (see M&M); the most relevant ones are also described in Supplementary Data S2.

Global mtDNA patterns in North Iberia

The three North Iberian samples analyzed in the present study show a typical West European mtDNA haplogroup composition (Table S2). Some haplogroups show slight differences in frequency. For example, while haplogroup H sums ~39% of the total mtDNAs in Catalonia and Cantabria, it makes-up ~44% of the mtDNA pool in Galicia. Haplogroup V reaches its highest frequency in Cantabria, ~9%, and decrease substantially in Galicia (~4%), and in Catalonia (~3%). These differences in frequencies are not statistically significant (under a Chi-square test) but the patterns observed are in agreement with previous findings [5,16].

All the HVS-I profiles obtained were searched among datasets compiled from the literature (more than 83.000 profiles) but only considering the common sequence range from position 16090 to 16365. A total of ~5%, ~10%, and ~14% of the mtDNAs from Cantabria, Galicia and Catalonia, respectively, were still not observed in the literature. Catalonia shows the highest levels of sequence diversity, followed by Galicia and Cantabria (see also below and Table 1 for variability within haplogroup H). As expected, the most common haplotype was the one that matches the rCRS sequence, being very common in Galicia (~20%; range 16090–16365), but decreasing in frequency towards Cantabria (~13%) and Catalonia (~12%).

A small percentage of the total mtDNAs analyzed belonged to non-Eurasian lineages. Thus, several sub-Saharan mtDNA profiles were detected in Galicia (~2.5%) and Catalonia (~3%); none in Cantabria. Curiously, six out of the ten sub-Saharan haplotypes observed belong to haplogroup L1b; this clade originated in western Africa but it was also carried to America during the period of the Atlantic slave trade [44,45,46]. For instance, the L1b1 profile found in Galicia, T16126C C16187T T16189C C16223T C16264T C16270T C16278T A16293G T16311C (note also the presence in Galicia of another derivative haplotype with A16317G on top), is found in many sub-Saharan regions [47,48], but also in America [49,50,51]. The rare haplotype, C16169T C16193T T16195C C16223T T16243C C16261T, observed in Galicia belongs to the uncommon sub-Saharan haplogroup L3x2 typically found in Ethiopia and Yemen [52]; peculiarly, this haplotype was also detected in an independent sample collected from the same region some years ago [16,17].

Some typical Native American profiles were also observed in the Catalonian sample. For instance, the haplogroup D1 profile T16189C C16223T T16325C T16362C (excluding the 'speedy' transversion A16183C) is commonly found in South America [53,54], or the A2 haplotype C16111T T16189C C16223T C16290T G16319A T16362C (excluding highly mutable transition T16519C), which is also common in all around America [49,50].

In Catalonia we have also observed one rare East Asian profile, C16104T C16111T T16140C A16162G 16169+C A16182C A16183C T16189C C16228T C16234T T16243C, belonging to B5b. Members of this haplogroup appear frequently in Japan, Taiwan, Korea, etc. [9,55].

The presence at low frequency of non-western European lineages in Catalonia could be explained by recent gene flow because it is well-known that this region has received important

Table 1. Summary statistics of HVS-I sequences in the North Iberian populations analyzed in the present article and other European populations.

HG	Population	*N*	*k*	*S*	N_{mut}	*H*±SE	π±SE	*M*	$V_O(M)$	*r*	*D*	*FL*
All the sample												
	Galicia[*1]	282	150 (0.53)	93	102	0.952±0.010	0.0138±0.001	3.76	5.06	0.012	−2.328**	−4.727**
	Catalonia[*1]	101	79 (0.78)	71	73	0.984±0.007	0.0166±0.001	4.59	6.29	0.014	−2.187**	−3.557**
	Cantabria[*1]	135	61 (0.45)	60	62	0.971±0.007	0.0135±0.001	3.72	3.85	0.018	−2.099*	−2.596*
HG-H												
	Galicia[*1]	124	51 (0.41)	49	50	0.800±0.038	0.006±0.001	1.73	2.08	0.035	−2.528***	−4.447**
	Catalonia[*1]	44	30 (0.68)	33	33	0.937±0.030	0.009±0.001	2.48	1.93	0.043	−2.300**	−3.836**
	Cantabria[*1]	52	26 (0.50)	25	26	0.875±0.042	0.006±0.001	1.78	1.33	0.067	−2.251**	−2.480**
	Volga-Ural[*2]	50	18 (0.36)	17	18	0.819±0.049	0.006±0.001	1.61	1.39	0.050	−1.884*	−1.966
	Finland[*2]	31	16 (0.52)	15	16	0.908±0.035	0.009±0.001	2.42	1.53	0.092	−1.338	−1.083
	Estonia[*2]	50	31 (0.62)	30	31	0.936±0.026	0.009±0.001	2.54	2.31	0.035	−2.114*	−3.113*
	Slovakia[*2]	50	30 (0.60)	31	30	0.939±0.027	0.009±0.001	2.49	2.23	0.045	−2.090*	−2.455*
	France[*2]	50	19 (0.38)	17	19	0.762±0.063	0.005±0.001	1.31	1.33	0.097	−2.187**	−2.569*
	Balkans[*2]	50	31 (0.62)	30	31	0.953±0.018	0.009±0.001	2.52	1.77	0.053	−2.120*	−2.852*
	Turkey[*2]	50	31 (0.62)	27	31	0.914±0.032	0.008±0.001	2.13	1.59	0.055	−2.311**	−3.113*
	Near East[*2]	50	36 (0.72)	30	36	0.943±0-023	0.009±0.001	2.56	2.08	0.040	−2.301**	−4.097**
	Asia[*2]	48	29 (0.60)	26	29	0.947±0.019	0.010±0.001	2.89	2.37	0.029	−1.962*	−2.261
	Eastern Slavs[2]	50	30 (0.60)	31	30	0.944±0.023	0.009±0.001	2.35	1.67	0.057	−2.162*	−3.280*
	Arabian Peninsula[*3]	52	29 (0.56)	30	30	0.947±0.017	0.008±0.001	2.32	1.34	0.074	−2.153*	−3.050*
	Armenia[*3]	54	27 (0.50)	33	33	0.914±0.031	0.009±0.001	2.53	2.35	0.030	−2.158*	−1.685
	Daghestan[*3]	60	26 (0.43)	33	33	0.859±0.042	0.008±0.001	2.17	2.28	0.023	−2.268**	−2.323
	Georgia[*3]	30	15 (0.50)	16	16	0.874±0.050	0.008±0.001	2.11	2.12	0.031	−1.617	−0.682
	Jordan[*3]	33	18 (0.55)	25	25	0.847±0.062	0.008±0.001	2.24	2.30	0.024	−2.227**	−2.586*
	Karatchaians-Balkanians[*3]	50	21 (0.42)	23	23	0.943±0.017	0.012±0.001	3.23	2.00	0.059	−1.202	0.411
	Lebanon[*3]	34	20 (0.59)	23	23	0.907±0.041	0.008±0.001	2.09	1.88	0.061	−2.171*	−3.548**
	Northwest Caucasus[*3]	69	35 (0.51)	38	38	0.895±0.034	0.009±0.001	2.42	2.70	0.026	−2.256**	−2.953*
	Ossetia[*3]	45	22 (0.49)	26	27	0.883±0.002	0.009±0.001	2.58	2.84	0.029	−1.950*	−2.445
	Syria[*3]	28	19 (0.68)	23	23	0.966±0.019	0.009±0.001	2.38	1.38	0.098	−2.139*	−2.667*
	Turkey[*3]	90	46 (0.51)	44	46	0.898±0.029	0.008±0.001	2.24	2.10	0.037	−2.408**	−2.957*
	Austria[*4]	964	116 (0.12)	75	81	0.683±0.017	0.005±0.001	1.15	1.07	0.041	−2.468***	−5.322
	Germany[*4]	28	20 (0.71)	20	20	0.952±0.030	0.010±0.001	2.73	1.88	0.042	−1.657	−1.116
	Hungary[*4]	55	15 (0.27)	22	22	0.677±0.070	0.006±0.001	1.64	2.61	0.073	−2.059*	−2.160
	Macedonia[*4]	88	30 (0.34)	28	29	0.892±0.025	0.007±0.001	2.01	1.84	0.058	−2.000*	−1.707
	Romania[*4]	100	29 (0.29)	29	29	0.917±0.017	0.009±0.001	2.48	2.04	0.034	−1.690	−2.160

N = sample size, *k* = number of different haplotypes (divided by *N* in brackets).
S = Number of polymorphic (segregating) sites.
N_{mut} = total number of mutations.
H = haplotype diversity and standard error.
π = nucleotide diversity and standard error.
M = average number of nucleotide differences.
$V_O(M)$ = observed variance of *M*.
r = Harpending's (raggedness) index.
D = Tajima's test of selective neutrality.
FL = Fu and Li's D* statistics.
Statistical significance: *, *P*-value<0.05.
**, *P*-value<0.02.
[*1]Present study.
[*2]Loogväli et al. [3].
[*3]Roostalu et al. [4].
[*4]Brandstätter et al. [15].
doi:10.1371/journal.pone.0005112.t001

mtDNA Macro-Haplogroup R0

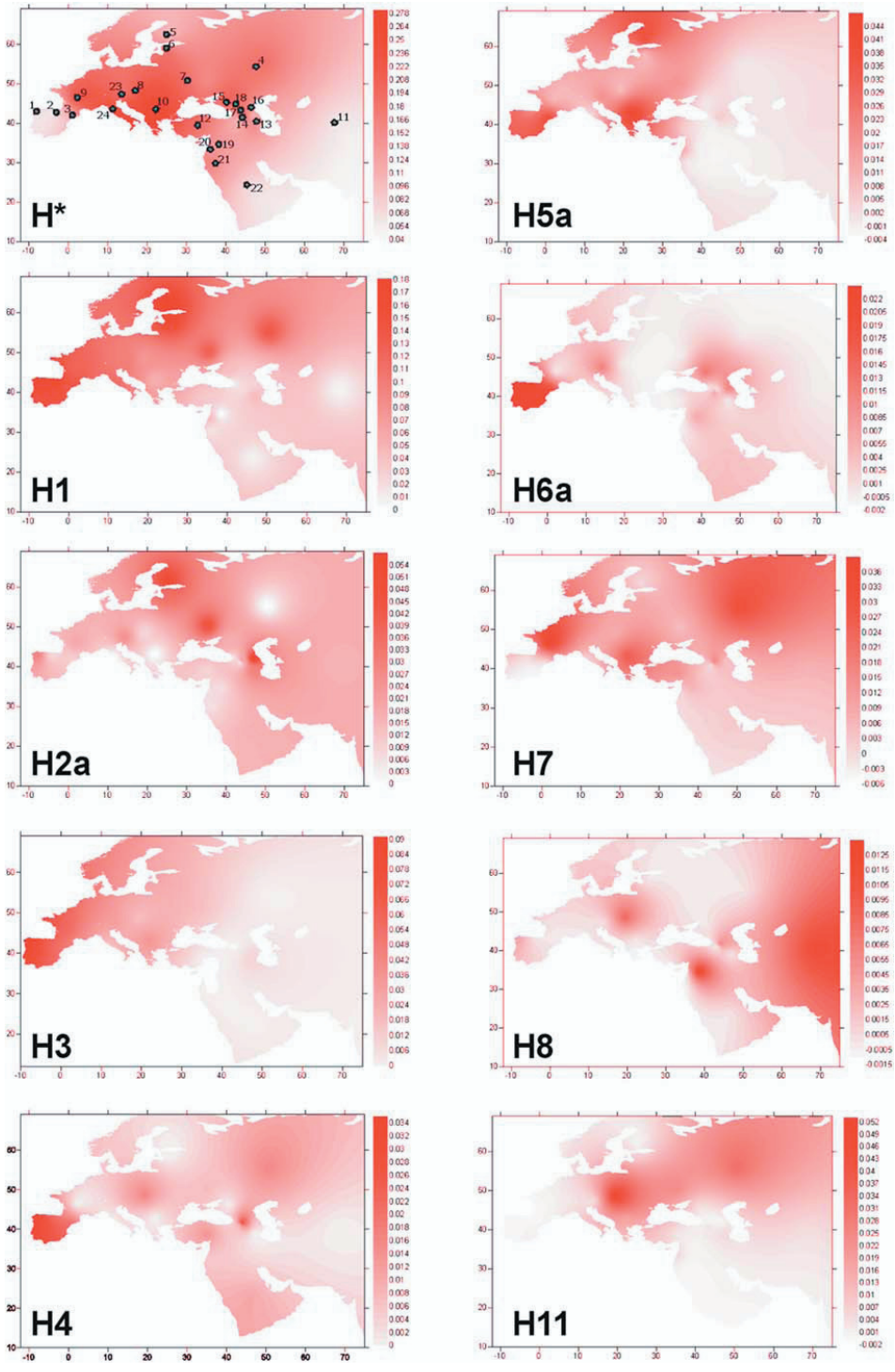


Figure 3. Geographic maps of haplogroup frequencies for haplogroups H*, H1, H2a, H3, H4, H5a, H6a, H7, H8, H11. Dots in the map of H* indicate the location of the populations used. Codes for populations are: [1] Galicia, [2] Cantabria, [3] Catalonia (present study); [4] Volga-Ural, [5] Finland, [6] Estonia, [7] Eastern Slavs, [8] Slovakia, [9] France, [10] Balkans, [11] Central Asia [3]; [12] Turkey, [13] Armenia, [14] Georgia, [15] Northwest Caucasus, [16] Daghestan, [17] Ossetia, [18] Karatchians-Balkarians, [19] Syria, [20] Lebanon, [21] Jordan, [22] Arabian Peninsula [4]; [23] Austria [6]; [24] Tuscani [14].
doi:10.1371/journal.pone.0005112.g003

flow of immigrants in the last decades; more than Galicia and Cantabria.

Diversity patterns of R0 in North Iberia

Several diversity indices were computed for the three North Iberian samples analyzed in the present study (Table 1). Overall, the Catalonian sample shows the highest values of sequence and nucleotide diversity (however with overlapping ranges considering a confidence interval of two standard errors) and also for the average number of nucleotide differences. The Cantabrian region shows the lowest values again for the three mentioned indices.

The patterns of variability within haplogroup H are quite different around Europe and Middle East. For instance, Galicia shows one of the lowest sequence diversity values within Eurasia (Table 1), in agreement with a previous independent study from the region [16]); and it is also among the regions with lowest values of nucleotide diversities (together with Cantabria).

Both the Tajima's *D* and the Fu and Li's tests show significantly negative values in almost all the populations (Table 1), suggesting that all of them have passed through population expansions. The mismatch distributions (data not shown) also support this hypothesis as well as the raggedness *r* index (Table 1) indicating that the mismatch distributions are unimodal and then compatible with population expansion.

Phylogeographical patterns of R0 sub-lineages

Using the SNP genotyping strategy described above, less than 10% of the lineages within haplogroup H could not be allocated to some of the already known H sub-branches (Table S3).

The distribution of haplogroup frequencies along the North Iberian fringe shows patterns moderately stratified.

On average, ~42% of the mtDNAs in the total sample belongs to haplogroup H; the Galician sample reaches the highest frequency (~44%), and it is slightly lower (~39%) in Cantabria and Catalonia.

H* represents 15% and 10% of the total haplogroup H lineages in Catalonia and Galicia, respectively, but only 4% in Cantabria. H1a (without counting H1a1 and H1a2) represents 8% of the haplogroup H mtDNAs in Cantabria, but it makes-up only 2% in Galicia and 0% in Catalonia. Again with respect to haplogroup H, H1 is more frequent also in Cantabria (46%) than in Galicia (38%) and Catalonia (36%); whereas haplogroup V has a clear peak in Cantabria, ~16% of the total R0 haplotypes (but only ~9% in Galicia and ~6% in Catalonia).

The maps of Figure 3 show the spatial frequency distribution of different sub-lineages of haplogroup H. Some clades get the highest frequency in Iberia, such as H1, H3, and H5a or are only observed in this region (H4); while others are virtually absent in Iberia but are significantly more prevalent in Central Europe (e.g. H11).

In addition, haplogroups H1, H3, and H5a display clinal patterns as determined by their spatial correlograms (Figure S1). The frequency of these three haplogroups has a peak in the Franco-Cantabrian refuge area and declines towards East Europe.

The autochthonous nature of the H2a5 clade in the Basque Country

It was first notice in a study by Pereira et al. [19] by way of sequencing several small coding region mtDNA segments, the presence of the coding region variant C4952T in ~6% of their samples from the Basque Country.

A scrutiny of more than 5,500 coding region segments (most of them available in GenBank and some only in the literature) and in Google searches (*sensu* [56,57]) revealed that this variant was only reported twice, curiously in two medical studies [58,59] where no detailed information on the geographical origin of the carriers was provided. Therefore, the multiple occurrence of this transition in the Basque Country could point to a diagnostic site for an autochthonous lineage in this region. These features lead us to further investigate these mtDNAs by way of complete genome sequencing the eight available Basque samples carrying transition C4592T.

This analysis revealed a new sub-clade of haplogroup H, baptized here as H2a5. All these sequences share the following diagnostic variants: A1842G C4592T G13708A C16291T T16519C (Figure 4). Six out of the eight complete genomes are identical while the other two show one private variant each. The coalescence age for this sub-lineage is 15.7±9 kya.

Discussion

Analysis of mtDNA variation based exclusively on few RFLP markers and/or the HVS-I region have lead in the past to simplistic perceptions of Europe as a uniform population. The results presented in previous studies [1,3,4,6,15,60] and those shown here, demonstrate that population stratification in European population can only be revealed when using higher phylogenetic resolution. Analysis of complete genomes ideally provides the maximum level of resolution; however, genotyping

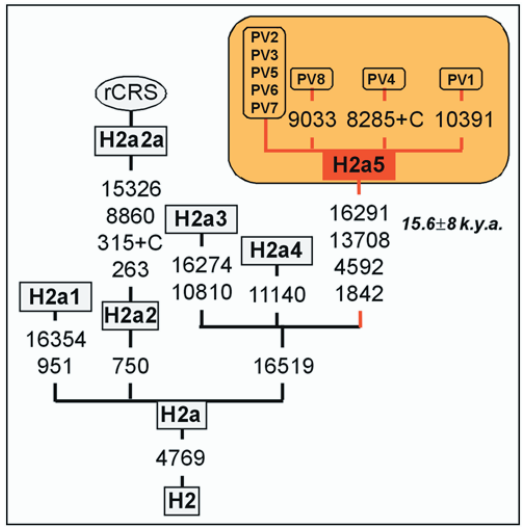


Figure 4. Phylogeny of haplogroup H2a5.
doi:10.1371/journal.pone.0005112.g004

complete mtDNA molecules demands great economical and personal effort in large-scale population projects. Complete genome sequencing is generally carried out when the analysis focuses on a particular group of mtDNAs presenting some interesting phylogenetic or population feature [1,3], such as performed here for the analysis of members belonging to haplogroup H2a5. At a population level, the coding region genotyping strategy presented here represents a way to overcome the drawback of whole genome genotyping and allow at the same time obtaining high resolution information from the mtDNA genome.

A total of 518 samples from three main locations in North Iberia were sequenced for the HVS-I segment. About 55% of them could be ascribed to R0. All these samples were further screened for a set of 71 coding region SNPs in order to sub-classify them into different R0 sub-clades. As indicated by the various diversity indices computed, Galicia and Cantabria show low diversity values, especially for the overall haplogroup H. The present study also revealed moderate levels of stratification in North Iberia, which could be relevant in other fields of research, such as in forensic casework [61] or in medical studies, where population sub-structure could explain most of the false positives of association in case-control studies [62].

When compared to other European and Middle East populations, we observed geographical patterns for H1, H3 and H5a that are statistically clinal, with frequency peaks in the Franco-Cantabrian region decreasing towards East Europe. This is compatible with a process of demographic repopulation of Europe after the LGM period centered in this climatic and geographic refuge, as it was previously demonstrated by Torroni et al. [5] and Achilli et al. [1].

We have also described a new minor autochthonous clade in Basques, H2a5. This lineage has been dated in 15.6±8 kya; this age fits also with the period of population expansion that followed the LGM (although with a large standard error). However, this branch was exclusively found in the Basque country at a significant frequency (~6%). The absence of this clade in other parts of Europe could be due to the limited sample size still available in the literature; however, we can speculate with the fact that all the evidences taken together resemble the findings of Torroni et al. [5] and Achilli et al. [1] regarding the 'imprint' of post-LGM human population re-expansions centered in the Franco-Cantabrian refuge on the mtDNA variability.

Supporting Information

Text S1 Genotyping protocols.
Found at: doi:10.1371/journal.pone.0005112.s001 (0.27 MB DOC)

Text S2 Reconciliation of the nomenclature conflicts in haplogroup R0.
Found at: doi:10.1371/journal.pone.0005112.s002 (0.14 MB DOC)

Text S3 Note about the advantages of using minisequencing high throughput SNP genotyping and report of the phylogenetic inconsistencies observed in the data from North Iberia.
Found at: doi:10.1371/journal.pone.0005112.s003 (0.04 MB DOC)

Table S1 Compendium of the problems related to the nomenclature of the R0 phylogeny and update of the nomenclature.
Found at: doi:10.1371/journal.pone.0005112.s004 (0.05 MB XLS)

Table S2 HVS-I and coding region SNP variation for the Iberian samples analyzed in the present study.
Found at: doi:10.1371/journal.pone.0005112.s005 (0.71 MB XLS)

Table S3 Comparative population frequencies of different haplogroup H (sub)lineages. In bold we collapse frequencies into higher hierarchical phylogenetic clades as a function of the SNPs genotyped in the referred studies, such that only these 'bolded' categories are fully comparable between the different studies considered. This is because haplogroup categories are not fully comparable among populations when the samples have undetermined (nd) SNPs; for instance, H* embraces different lineages in our study because we genotyped to a higher level of resolution than previous attempts (where different lineages were already collapsed into H*). For nomenclature we follow the scheme of Figure 1 and Figure 2.
Found at: doi:10.1371/journal.pone.0005112.s006 (0.08 MB XLS)

Figure S1 Autocorrelograms for the most frequent R0 subclades observed in North Iberia.
Found at: doi:10.1371/journal.pone.0005112.s007 (0.12 MB PPT)

Acknowledgments

We would like to thank Francesc Calafell and Oscar Lao for their help with the Surfer software and the spatial representations. The eight complete mtDNA genomes analyzed in the present study are available in GenBank under accession numbers FJ527772–FJ527779.

Author Contributions

Conceived and designed the experiments: VAI AMM MC AS. Performed the experiments: VAI AMM MC BQ. Analyzed the data: VAI AMM MC AS. Contributed reagents/materials/analysis tools: MTZ IC MVL OG LPJ AC AS. Wrote the paper: VAI AMM MC AS.

References

1. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.
2. Abu-Amero KK, González AM, Larruga JM, Bosley TM, Cabrera VM (2007) Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC Evol Biol 7: 32.
3. Loogväli E-L, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, et al. (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21: 2012–2021.
4. Roostalu U, Kutuev I, Loogväli E-L, Metspalu E, Tambets K, et al. (2007) Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol 24: 436–448.
5. Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, et al. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69: 844–852.
6. Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo Á, et al. (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27: 2541–2550.
7. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, et al. (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251–257.
8. Álvarez-Iglesias V, Barros F, Carracedo Á, Salas A (2008) Minisequencing mitochondrial DNA pathogenic mutations. BMC Med Genet 9: 26.
9. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int: Genet 1: 44–55.

10. Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, et al. (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 118: 137–146.
11. Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, et al. (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15: 2076–2086.
12. Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, et al. (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3: e1764.
13. Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, et al. (2008) Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS ONE 3: e2062.
14. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, et al. (2007) Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet 80: 759–768.
15. Brandstätter A, Zimmermann B, Wagner J, Göbel T, Röck A, et al. (2008) Timing and deciphering mitochondrial DNA macro-haplogroup R0 variability in Central Europe and Middle East. BMC Evol Biol 4: 191.
16. Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo Á (1998) mtDNA analysis of the Galician population: a genetic edge of European variation. Eur J Hum Genet 6: 365–375.
17. Salas A, Lareu MV, Sánchez-Diz P, Calafell F, Carracedo A (2000) mtDNA hypervariable region II (HVII) sequences in human evolution studies: impact of mutation rate heterogeneity. Progress in Forensic Genetics 8: 329–331.
18. Maca-Meyer N, Sánchez-Velasco P, Flores C, Larruga JM, González AM, et al. (2003) Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain). Ann Hum Genet 67: 329–339.
19. Pereira L, Richards M, Alonso A, Albarrán C, Garcia O, et al. (2004) Subdividing mtDNA haplogroup H based on coding-region polymorphisms—a study in Iberia. Int Congress Series 1261: 416–418.
20. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.
21. van Oven M, Kayser M (2008) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat in press.
22. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.
23. Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, et al. (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137–1152.
24. Bandelt H-J, Kivisild T, Parik J, Villems R, Bravi CM, et al. (2006) Lab-specific mutation processes. In: Bandelt H-J, Richards M, Macaulay V, eds. Human mitochondrial DNA and the evolution of Homo sapiens. Berlin: Springer-Verlag.
25. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891–899.
26. Salas A, Prieto L, Montesino M, Albarrán C, Arroyo E, et al. (2005) Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02-03 proficiency testing trial. Forensic Sci Int 148: 191–198.
27. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150–1160.
28. Bandelt H-J, Salas A, Bravi CM (2004) Problems in FBI mtDNA database. Science 305: 1402–1404.
29. Bandelt H-J, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267–273.
30. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496–2497.
31. Nei N (1987) Molecular evolutionary genetics. New York: Columbi University Press.
32. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.
33. Tajima F (1993) Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10: 677–688.
34. Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66: 591–600.
35. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–589.
36. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.
37. Barbujani G (2000) Geographic patterns: how to identify them and why. Hum Biol 72: 133–153.
38. Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, et al. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64: 232–249.
39. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, et al. (2005) Saami and Berbers–an unexpected mitochondrial DNA link. Am J Hum Genet 76: 883–886.
40. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718–726.
41. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, et al. (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171–176.
42. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373–387.
43. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, et al. (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1–8.
44. Salas A, Carracedo Á, Richards M, Macaulay V (2005) Charting the Ancestry of African Americans. Am J Hum Genet 77: 676–680.
45. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, et al. (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082–1111.
46. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, et al. (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454–465.
47. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, et al. (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596–1601.
48. Černý V, Salas A, Hájek M, Zaloudkova M, Brdička R (2007) A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71: 433–452.
49. Salas A, Acosta A, Álvarez-Iglesias V, Cerezo M, Phillips C, et al. (2008) The mtDNA ancestry of admixed Colombian populations. Am J Hum Biol 20: 584–591.
50. Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, et al. (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol 8: 213.
51. Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt H-J, et al. (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67: 444–461.
52. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, et al. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752–770.
53. Santos M, Ward RH, Barrantes R (1994) mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Hum Biol 66: 963–977.
54. Salas A, Richards M, Lareu MV, Sobrino B, Silva S, et al. (2005) Shipwrecks and founder effects: Divergent demographic histories reflected in Caribbean mtDNA. Am J Phys Anthropol 128: 855–860.
55. Hill C, Soares P, Mormina M, Macaulay V, Clarke D, et al. (2007) A mitochondrial stratigraphy for island southeast Asia. Am J Hum Genet 80: 29–43.
56. Bandelt H-J, Salas A, Bravi CM (2006) What is a 'novel' mtDNA mutation – and does 'novelty' really matter? J Hum Genet 51: 1073–1082.
57. Bandelt H-J, Yao Y-G, Salas A (2008) The search of 'novel' mtDNA mutations in hypertrophic cardiomyopathy: MITOMAPping as a risk factor. Int J Cardiol 126: 439–442.
58. Schwartz M, Vissing J (2002) Paternal inheritance of mitochondrial DNA. N Engl J Med 347: 576–580.
59. Pulkes T, Liolitsa D, Nelson IP, Hanna MG (2003) Classical mitochondrial phenotypes without mtDNA mutations: the possible role of nuclear genes. Neurology 61: 1144–1147.
60. Richards M, Macaulay V, Torroni A, Bandelt H-J (2002) In search of geographical patterns in European mitochondrial DNA. Am J Hum Genet 71: 1168–1174.
61. Egeland T, Salas A (2008) Estimating haplotype frequency and coverage of databases. PLoS ONE 3: e3988.
62. Mosquera-Miguel A, Álvarez-Iglesias V, Vega A, Milne R, Cabrera de León A, et al. (2008) Is mitochondrial DNA variation associated with sporadic breast cancer risk? Cancer Res 68: 623–625.

OPEN ACCESS Freely available online

PLoS one

Mitochondrial Echoes of First Settlement and Genetic Continuity in El Salvador

Antonio Salas[1]*, José Lovo-Gómez[1,2], Vanesa Álvarez-Iglesias[1], María Cerezo[1], María Victoria Lareu[1], Vincent Macaulay[3], Martin B. Richards[4], Ángel Carracedo[1]

1 Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain, **2** Laboratorio de Genética Forense, Instituto de Medicina Legal, Dr. Roberto Masferrer, Corte Suprema de Justicia, San Salvador, El Salvador, **3** Department of Statistics, University of Glasgow, Glasgow, United Kingdom, **4** Institute of Integrative and Comparative Biology, Faculty of Biological Sciences, University of Leeds, Leeds, United Kingdom

Abstract

Background: From Paleo-Indian times to recent historical episodes, the Mesoamerican isthmus played an important role in the distribution and patterns of variability all around the double American continent. However, the amount of genetic information currently available on Central American continental populations is very scarce. In order to shed light on the role of Mesoamerica in the peopling of the New World, the present study focuses on the analysis of the mtDNA variation in a population sample from El Salvador.

Methodology/Principal Findings: We have carried out DNA sequencing of the entire control region of the mitochondrial DNA (mtDNA) genome in 90 individuals from El Salvador. We have also compiled more than 3,985 control region profiles from the public domain and the literature in order to carry out inter-population comparisons. The results reveal a predominant Native American component in this region: by far, the most prevalent mtDNA haplogroup in this country (at ~90%) is A2, in contrast with other North, Meso- and South American populations. Haplogroup A2 shows a star-like phylogeny and is very diverse with a substantial proportion of mtDNAs (45%; sequence range 16090–16365) still unobserved in other American populations. Two different Bayesian approaches used to estimate admixture proportions in El Salvador shows that the majority of the mtDNAs observed come from North America. A preliminary founder analysis indicates that the settlement of El Salvador occurred about 13,400±5,200 Y.B.P.. The founder age of A2 in El Salvador is close to the overall age of A2 in America, which suggests that the colonization of this region occurred within a few thousand years of the initial expansion into the Americas.

Conclusions/Significance: As a whole, the results are compatible with the hypothesis that today's A2 variability in El Salvador represents to a large extent the indigenous component of the region. Concordant with this hypothesis is also the observation of a very limited contribution from European and African women (~5%). This implies that the Atlantic slave trade had a very small demographic impact in El Salvador in contrast to its transformation of the gene pool in neighbouring populations from the Caribbean facade.

Citation: Salas A, Lovo-Gómez J, Álvarez-Iglesias V, Cerezo M, Lareu MV, et al. (2009) Mitochondrial Echoes of First Settlement and Genetic Continuity in El Salvador. PLoS ONE 4(9): e6882. doi:10.1371/journal.pone.0006882

Editor: Manfred Kayser, Erasmus University Medical Center Rotterdam, Netherlands

Received May 5, 2009; **Accepted** July 29, 2009; **Published** September 2, 2009

Funding: This work was partially supported by grants from the Xunta de Galicia (Grupos Emerxentes; 2008/037), Ministerio de Ciencia e Innovacion (SAF2008-02971), and Fundacion de Investigacion Medica Mutua Madrilena (2008/CL444) given to AS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: antonio.salas@usc.es

Introduction

El Salvador lies on the Pacific coast (without an Atlantic seaboard) and it is the smallest of the Central American countries. Most of the country rests on a fertile volcanic plateau. It is segmented by two volcanic ranges running roughly west to east, separated by broad, fertile valleys, such as that of the river Lempa. El Salvador was inhabited by Native American groups who were in part descendants of the Aztecs and Toltec of Mexico, such as the Pipil (a Nahua tribe) and the Lenca. These two Native American communities inhabited mainly the western regions, constituting about 60% of the population throughout the colonial era and into the early decades of independence [1,2].

The development of coffee estates led to the slow but continuous dissolution of most of the communal lands of Native villages [1,2]. Thus, the 1930 census, the last to contain the category, designated only 5.6% of the population as "Indian" – although it is not clear what criteria were used in arriving at this figure. Other independent estimates (considering religious activities, distinctive women's dress, language, and involvement in various handicrafts) placed the mid-twentieth-century Indian population at 20% (~400,000 persons). The abandonment of Indian language and customs was hastened by political repression; most natives stopped wearing traditional dress, abandoned the Pipil language, and adopted ladino customs. By 1975 no more than ~1% of the population wore distinctive Indian clothing or followed Indian

customs. Nowadays, the official language in El Salvador is Spanish, although Nahua is still spoken among some natives.

Although the American continent has been the target of many forensic and population genetic studies, there are nevertheless many American regions, such as El Salvador, that remain genetically uncharacterized. The mtDNA molecule is commonly used in anthropological contexts because of particular features (maternal inheritance, lack of recombination and high average mutation rate) that confer great power for phylogenetic and phylogeographic inferences. Many mtDNA studies of Native Americans have, however, been limited to genotyping a handful of mtDNA coding region sites that simply distinguish the four major Native American mtDNA haplogroups, A2, B2, C1 and D1 (generally using RFLP typing); unfortunately, the information provided by these few SNPs is of limited value in forensic and population genetics.

Here we have sequenced the mtDNA control region in a sample from El Salvador in order to investigate to what extent the Native American component has survived the impact of European colonialism and the concomitant influx of African slaves to the Caribbean and Meso-America.

Materials and Methods

Sample collection and DNA extraction

A total of 90 saliva samples were collected from healthy unrelated individuals from El Salvador. DNA extraction was undertaken following standard phenol-chloroform protocol. DNA quantification was carried out using DyNA Quant 200 Fluorometer, Hoefer (APB, Uppsala, Sweden).

All the samples were collected anonymously by the Laboratorio de Genética Forense from the Instituto de Medicina Legal that belongs to the Corte Suprema de Justicia from San Salvador (El Salvador). Oral informed consent was required in all the cases. The study, including the oral informed consent protocol, was approved by the Ethical committee of the University of Santiago de Compostela, and it conforms to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July).

PCR and sequence analysis

We analyzed the first and second hypervariable segments (HVS-I and HVS-II) of the mtDNA genome. We performed PCR amplifications using a 2700 Thermocycler (Applied Biosystems), using PCR and sequencing primers as reported in [3]. Cycling parameters were 95°C for 1 min, followed by 36 cycles of 95°C for 10 sec, 55°C for 30 sec and 72°C for 30 sec, and followed by 10 min at 15°C. We checked amplification products on a polyacrylamide gel visualized by silver staining and purified with *Montage* (Multiscreen PCR, Millipore Corporation, USA). We performed sequence reaction products on each strand by means of the ABI Prism dRhodamine Terminator cycle sequencing reaction kit (Applied Biosystems). DNA products were then purified by ethanol precipitation and sequence reaction products analyzed on the ABI Prism 3100 automatic sequencer (Applied Biosystems). We omitted population variation at the hypervariable sites (mainly related to the cytosine homopolymeric track around 310 and the CA-dinucleotide repeat around positions 522) from inter-population comparisons and phylogeographic analyses. We have used the same primers for amplification and sequencing described in [4].

Sequences were edited using the numbering system of the revised Cambridge Reference Sequence [5]. Most of the sequences could be read from np 16042–16569 and 21–550; for convenience, we will refer to these as HVS-I and HVS-II, although these sequence ranges encompass more than the canonical ranges of these control-region segments.

Quality checking

Problems with the quality of mtDNA data in forensic, clinical, and population genetic studies are unfortunately rather common; see, for instance, [6,7,8,9,10,11]. In order to minimize the effects of potential laboratory and documentation errors, the data were read separately by two independent persons in the light of the known phylogeny. We checked phylogenetic inconsistencies by hand with special attention to private or unusual variants (e.g. rare transitions or indels). In some cases, we confirmed the sequences by repeated extraction and sequencing. In addition, to detect potential "phantom mutations" [7], we also checked the data using the computer program SPECTRA ([7], available at http://www.stats.gla.ac.uk/~vincent/fingerprint/index.html).

Statistical analysis and population comparison

Haplogroup nomenclature follows the most recently updated versions of the Native American phylogeny given in [12,13,14]. Diversity indices of HVS-I sequences (haplotype diversity, nucleotide diversity, average number of pairwise differences) were calculated using Arlequin 3.0 software [15]. Nucleotide and sequence diversity was computed as in the manner of Nei [16].

We estimated median-joining networks of HVS-I sequences using the Network 4.1.1.2 software [17,18]. Coalescence times were calculated using the ρ statistic [19,20] with an HVS-I mutation rate of one transition per 18,845 years applied for the sequence range 16090-16365 using the most recent estimates provided by Soares et al. [21].

An mtDNA database of Native American populations was compiled for population comparisons: (i) from North America: Aleuts [22], Athapaskans, Inupiaq, Yakima [23], Chukchi and Siberian Eskimos [24], Bella Coola and Haida [25], Nuu-Chah-Nulth [26], Cheyenne [27], North Native Americans [various ethnic groups; 28,29,30,31,32,33], Apache and Navajo [34], (ii) from Meso-America: Pima [27], Maya [33], Huetar [35], Kuna [36], Ngöbe [37], Quiché [38], Emberá and Wounan [39], Mexico [40], Central Native Americans [various ethnic groups] [28], El Salvador (present study), and (iii) from South America: Native Brazilians and Araucanians or Chileans [33], Equador [41], Embera and Gavião [42], Amazonas [43], Ayoreo [44], Chilean Mapuche and Pehueche, Yaghan [45], Argentinian Mapuche [46], Cayapas [41], Xavante, Zoró, and Gavião [42], Yanomami [47,48], South Native Americans [various ethinic groups] [29], Tuacuarembó [49], Uruguay [50], Guahibo [51], Colombia [33,52], Yuracaré, Trinitario, Movima, and Ignaciano [53], and 105 from Arequipa, Tayacaja and San Martin in Peru [54]. We also included the data collected from several studies of ancient DNA [55,56,57,58,59,60,61]. In addition, other datasets were additionally used for haplotype matching comparisons [34,62,63,64,65,66]. In total, 3,843 mtDNAs profiles (mainly HVS-I segments) were compiled for comparison with our sample from El Salvador. Those population samples consisting of less than 15 individuals were only used for haplotype matching between populations. For comparison purposes the common reading frame 16090–16365 of the HVS-I was used.

Admixture analysis

We took two different approaches to carry out an admixture analysis.

The first model was applied as described in [67] although, instead of using haplogroup frequencies as variables, we used the frequencies of the shared haplotypes (matching haplotypes) between the source populations (North and South America) and El Salvador. The number of mtDNAs within each matching haplotype in El Salvador (n_i: $1 \leq i \leq C$, the number of different

matching haplotypes in the sample) was assumed to be a draw from a multinomial distribution with parameters $N=\sum_{i=1}^{c} n_i$ and $p_i=\sum_{j=1}^{R} a_j f_{ji}$ $(1\leq i\leq C)$, where R is the number of source regions in America, f_{ji} is the frequency of the ith cluster in the jth source region (assumed to be known), and α_j are the admixture coefficients. This model describes samples from an urn containing C different kinds of ball, where the urn has been created by mixing together R other urns in proportions given by the admixture coefficients. We chose to analyze this model in a Bayesian framework, which meant that we had to explore the distribution of the admixture coefficients, given the data. The prior distribution of the admixture coefficients was taken to be uninformative—namely, uniform on $a_j\geq 0, \sum_{j=1}^{R} a_j=1$. The posterior distribution of the $\{\alpha_j\}$ was explored with the Metropolis-Hastings algorithm, using a simple proposal, and was summarized by the posterior mean of each a_j and its root-mean-square deviation about the mean. To assess model fit, we examined plots of standardized residuals.

The second admixed model was applied as described [68]. The probability of origin of each of the sub-continental region was computed as $p_{o,}=\frac{1}{n}\sum_{i=1}^{n} k_i \frac{p_{ic}}{p_{iC}}$ where, n is the number of El Salvador sequences with matches (≥ 1) in the whole continental dataset; k_i, the number of times the sequence i is found in El Salvador; p_{ic}, the frequency of the sequence i in the sub-continental region dataset; and p_{ic}, the frequency of the sequence i in whole continental dataset.

Founder analysis

The time to the most recent common ancestor (TMRCA) of haplogroup A2 in the phylogeny was estimated as described [19,20].

In order to carry out a founder analysis [19,69], we made some simplifying assumptions about the founding of El Salvador. We assumed (i) a single migration to El Salvador and (ii) that North America was the unique source population. Founder sequences were inferred as matches with samples from North America. An estimate of the time of the migration event was determined by averaging diversity over the clusters derived from each founder in El Salvador, as follows. Suppose there are r founder clusters. Let, ρ_i be the ρ value (average distance of the haplotypes of a clade from the respective root [19,20]) for the i^{th} founder cluster, σ_i be its estimated standard error [20] and n_i be the number of sampled individuals in that cluster. Define,

$$n=\sum_{i=1}^{r} n_i \text{ and } w_i = n_i/n.$$

Then,

$$\text{overall } \rho = \sum_{i=1}^{r} w_i\rho_i \text{ and overall } \sigma=\sqrt{\sum_{i-1}^{r} w_i\rho_i}.$$

Values of ρ and σ were converted to age using the most recent mutation rate available for the HVS-I segment of 1 transition per 18,845 years (in the sequence range 16090–16365).

Results

Summary statistics

We observed a total of 55 different HVS-I, 40 different HVS-II, and 76 different combined HVSI/II mtDNA haplotypes in El Salvador ($N=90$). Some HVS-I profiles are quite common, such as C16111T–T16223C–C16290T–G16319A–T16362T, appearing in 12 mtDNAs, and its one step-mutation 'neighbour' haplotype (16519 on top) appearing 12 times.

As shown in Table 1, El Salvador shows haplotype and nucleotide diversity values slightly lower than those observed in the continental North, South, and other Meso American populations, which is in part due to the fact that there is virtually only one Native American haplogroup (A2) represented in El Salvador sample. Note that these comparisons have to beviewed with care because the terms "North", "South" and "Meso-American" refer to groups of population samples of different nature; some are Native American groups that have passed through severe prehistoric bottlenecks while others are at different levels of admixture with *e.g.* Europeans and Africans.

Phylogeography of Salvadorian Native American mtDNAs

Table 2 shows the full list of control region profiles from El Salvador and their haplogroup allocation. Frequencies of the typical Native American haplogroups A2, B2, and C1 are ~91%, ~2%, and ~2%, respectively.

Figure 1 shows the frequency distribution of the main mtDNA American haplogroups in Native American populations. Although haplogroup A2 is at high frequencies in Meso America, El Salvador is particularly distinct from the other populations by its extremely high A2 haplogroup frequency. Note also that there exists substantial heterogeneity of haplogroup frequency patterns in America (even between neighbouring populations).

The phylogeny of A2 in El Salvador is clearly star-like (Figure 2); its root is, identified by the diagnostic sites C16111T–T16223C–C16290T–G16319A–T16362C in HVS-I, and C64T–A73G–T146C–A153G –A235G–A263G–315+C in HVS-II. There are no very solid diagnostic sites in the control region that would allow us to classify A2 sub-lineages from El Salvador [12,14]. Moreover, several control-region variants regarded as haplogroup diagnostic, such as C64T and A153G, show reversions: complete genome sequence data confirm the existence of multiple back and parallel mutations within haplogroup A2 [12,14]. Although many of them are well-known hotspots (e.g. T146C), others such as position 64, seem to behave as hotspots only within A2 (see e.g. [12]). Other Native American lineages, like D1, D4h3 and X2a [13] are absent from our sample from El Salvador.

The sub-clade of A2 carrying C16360T is particularly prevalent in Meso America, especially in the Huetar (12 matches; ~44% of the Huetar sample) from Costa Rica [35] and the Ngöbé (three matches; ~7% of the Ngöbe sample) from Panama [37]; in El Salvador this variant was also present in two individuals. The haplotype C16111T–C16187T–T16223C–C16290T–G16319A–T16362C is virtually only shared with the Ngöbe (19 matches that make up ~41% of the Ngöbe sample) but was also detected in one

Table 1. Native American population mtDNA database considering the sequence range 16090–16365.

	n	H	D	π	M
El Salvador	90	49	0.917±0.025	0.013±0.002	3.5
North America	1215	243	0.950±0.003	0.020±0.000	5.1
Meso America	394	142	0.968±0.004	0.023±0.012	6.2
South America	1144	265	0.956±0.003	0.019±0.000	5.3

n = sample size; H = number of different haplotypes; D = haplotype diversity; π = nucleotide diversity; M = average number of nucleotide differences.
doi:10.1371/journal.pone.0006882.t001

Table 2. MtDNA haplotypes in El Salvador.

ID#	HVS-I (minus 16000)	HVS-I reading range (minus 16000)	HVS-II	HVS-II reading range	HG
1	42 111 223 244 290 319 362 519	024-569	64 73 146 152 153 154 235 263 309+C 315+C 523-524del	021-540	A2
2	51 111 223 290 299 319 362	024-569	64 73 146 153 235 263 315+C 523-524del	021-589	A2
90	111 136 153 223 290 311 319 362	024-560	64 73 146 153 235 263 309+CC 315+C 523-524del	021-595	A2
10	111 136 153 223 290 319 362	024-520	64 73 146 153 235 263 309+CC 315+C 523-524del	021-590	A2
8	111 136 223 290 311 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-560	A2
43	111 136 223 290 311 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
55	111 136 223 290 319 362	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
62	111 136 223 290 319 362	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-540	A2
93	111 136 223 290 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
61	111 172 223 290 319 362 519	024-589	64 73 146 153 195 235 263 309+CC 315+C	021-590	A2
83	111 175 223 290 300 319 362	024-569	64 73 153 235 263 309+C 315+C	021-440	A2
77	111 175 290 300 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-560	A2
35	111 181 187 223 290 304 319 362	024-564	64 73 146 153 207 235 263 309+C 315+C	021-320	A2
64	111 182C 183C 189 223 290 319 362	024-549	64 73 146 153 235 263 309+CC 315+C	021-320	A2
67	111 182C 183C 189 223 290 319 362	024-560	64 73 146 153 235 263 315+C 523-524del	021-600	A2
48	111 183C 189 223 290 319 362 381 519	024-560	64 73 146 153 235 263 309+C 315+C	021-310	A2
46	111 187 209 223 290 319 362 371 519	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-550	A2
40	111 187 223 234 290 319 362 390 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-570	A2
54	111 187 223 290 319 362	024-589	64 73 146 153 235 263 309+C 315+C	021-410	A2
6	111 187 223 290 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-560	A2
13	111 187 223 290 319 362	024-520	64 73 146 153 235 263 309+C 315+C 523-524del	021-535	A2
17	111 187 223 290 319 362	024-530	64 73 146 153 235 263 315+C 523-524del	021-569	A2
47	111 189 223 290 319 362	024-530	64 73 146 153 235 263 309+C 315+C 523-524del	021-550	A2
89	111 189 223 274 290 319 362	024-540	64 73 146 153 235 263 309+CC 315+C	021-510	A2
87	111 189 223 290 311 319 362	024-550	64 73 146 153 235 263 292 309+C 315+C 523-524del	021-600	A2
24	111 189 223 290 319 324 362	034-569	64 73 146 153 235 263 309+CC 315+C 523-524del	021-535	A2
32	111 189 223 290 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-535	A2
57	111 189 223 290 319 362	024-560	73 146 152 153 235 263 309+CC 315+C	021-320	A2
75	111 209 223 290 291 319 362 477	024-569	64 73 146 152 153 235 263 315+C 523-524del	021-600	A2
74	111 209 223 290 293C 319 362 519	024-569	64 73 146 153 235 263 309+CC 315+C4 523-524del	021-550	A2
28	111 209 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
31	111 209 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-595	A2
99	111 223 234 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 356+C 523-524del	021-600	A2
21	111 223 243 290 299 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
7	111 223 290 299 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-560	A2
37	111 223 290 299 319 362	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-535	A2
44	111 223 290 299 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
49	111 223 290 299 319 362	024-520	64 73 146 153 235 263 309+C 315+C 523-524del	021-570	A2
82	111 223 290 300 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
23	111 223 290 311 319 360 362	024-569	146 153 235 263 309+CC 315+C 523-524del	021-539	A2
26	111 223 290 311 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
4	111 223 290 311 319 362	024-569	64 73 146 153 235 263 315+C 523-524del	021-600	A2
3	111 223 290 319 360 362	024-569	143 146 152 153 204 235 263 309+CC 315+C 523-524del	021-570	A2
59	111 223 290 319 362	024-560	64 73 146 152 153 215 235 263 309+C 315+C 523-524del	021-600	A2
92	111 223 290 319 362	024-569	64 73 146 152 153 235 263 315+C	021-320	A2
38	111 223 290 319 362	024-560	64 73 146 153 235 263 309+C 315+C	021-535	A2
78	111 223 290 319 362	024-520	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
79	111 223 290 319 362	024-530	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
84	111 223 290 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-530	A2

Table 2. Cont.

ID#	HVS-I (minus 16000)	HVS-I reading range (minus 16000)	HVS-II	HVS-II reading range	HG
20	111 223 290 319 362	024-549	64 73 146 153 235 263 309+CC 315+C	021-560	A2
29	111 223 290 319 362	024-500	64 73 146 153 235 263 309+CC 315+C 523-524del	021-570	A2
36	111 223 290 319 362	024-560	64 73 153 214 235 263 315+C 523-524del	021-600	A2
63	111 223 290 319 362	024-525	73 146 150 153 235 263 315+C 523-524del	021-560	A2
68	111 223 290 319 362	024-550	n.d	–	A2
42	111 223 290 319 362 391	024-569	64 73 146 153 235 263 309+CC 315+C 523-524del	021-600	A2
103	111 223 290 319 362 518 519	024-569	64 73 146 153 235 263 315+C 523-524del	021-550	A2
91	111 223 290 319 362 519	024-560	64 73 146 150 153 174+C 235 263 309+CC 315+C 523-524del	021-320	A2
53	111 223 290 319 362 519	024-569	64 73 146 150 153 235 263 315+C 523-524del	021-600	A2
34	111 223 290 319 362 519	024-569	64 73 146 150 153 235 263 315+C 523-524del	021-600	A2
76	111 223 290 319 362 519	024-550	64 73 146 152 153 235 263 309+C 315+C	021-320	A2
27	111 223 290 319 362 519	024-569	64 73 146 152 153 235 263 309+C 315+C 523-524del	021-589	A2
14	111 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
51	111 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
86	111 223 290 319 362 519	021-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
88	111 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
11	111 223 290 319 362 519	024-569	64 73 146 153 235 263 315+C 523-524del	021-570	A2
9	111 223 290 319 362 519	024-569	73 146 152 153 197 235 263 309+C 315+C 523-524del	021-590	A2
15	111 223 290 319 362 519	024-569	73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
25	111 223 290 319 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-544	A2
97	111 261 290 319 362 519	024-550	73 146 152 153 235 263 315+C 523-524del	021-600	A2
45	111 290 311 319 362 391	024-569	64 73 146 153 235 263 315+C 523-524del	021-590	A2
50	111 290 311 319 362 391	024-569	64 73 146 153 235 263 315+C 523-524del	021-589	A2
19	111 290 319 362 391	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
58	153 223 240 290 319 362	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-590	A2
18	189 223 290 319 362	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-585	A2
56	223 290 311 319 362	024-560	64 73 146 153 235 263 309+C 315+C 523-524del	021-530	A2
5	223 290 316 319 362	024-569	64 73 146 153 182 235 263 309+C 315+C 523-524del	021-560	A2
60	223 290 319 352 362	024-560	64 73 146 153 182 235 263 309+C 315+C 523-524del	021-589	A2
70	223 290 319 362	024-545	64 73 146 153 235 263 315+C 523-524del	021-600	A2
100	223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-600	A2
16	92 111 189 223 290 319 362 519	024-569	64 73 146 153 235 263 309+C 315+C 523-524del	021-570	A2
81	93 111 136 223 290 319 324 362	024-569	73 146 235 263 309+C 315+C 523-524del	021-580	A2
33	129 183C 189 217 269 519	024-569	73 146 152 195 234 263 309+CC 315+C 499	021-570	B2
22	183C 189 217 519	024-569	73 263 309+C 315+C 499	021-570	B2
41	183C 189 223 256 298 325 327	024-560	73 249del 263 290-291del 309+C 315+C 489 523-524del	021-555	C1
39	189 223 298 325 327 362 519	024-560	73 195 249del 263 290-291del 315+C 489	021-580	C1
12	519	024-569	153 263 315+C 523-524del	021-600	H?
52	129 148 168 172 187 188G 189 223 230 278 293 311 320	024-560	93 95C 185 189 236 247 263 315+C 523-524del	021-560	L0a1a
30	126 271 294 296 304 519	024-569	73 263 315+C	021-580	T2
96	51 129C 189 362	024-540	73 152 217 263 309+C 315+C 340 508	021-600	U2e

HG = haplogroup.
n.d. = non determined.
doi:10.1371/journal.pone.0006882.t002

Uruguayan [50]. El Salvador shares a higher number of haplotypes with North America (19), followed by Meso-America (10), and then South America (8); note however that the database for Meso-America ($n = 395$) is of a much lower sample size than the one from North ($n = 2,010$) and South ($n = 1,596$). These results roughly indicate a clear imprint of North in Meso-America and also the existence of lineages that are mainly concentrated in Meso-America (probably due to the fact that these were founders in the region and experienced posterior expansion); in some instances, some of these South mtDNAs could have been carried

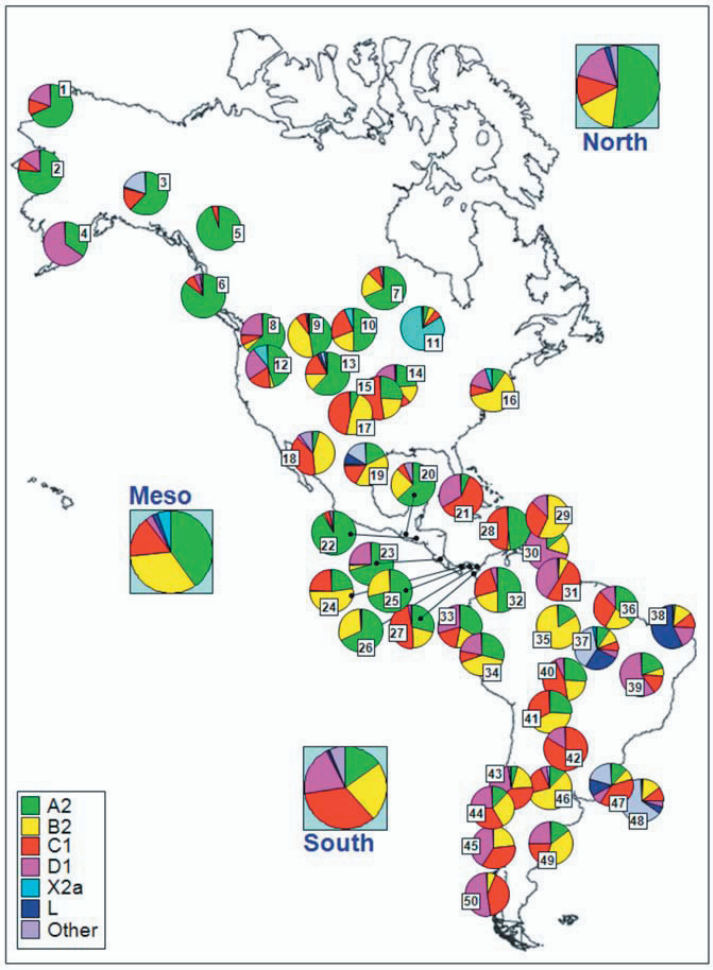


Figure 1. Haplogroup patterns in America. Codes for populations are as follow: North America: 1 = Chukchy, 2 = Eskimos [24]; 3 = Inuit (collected from the HvrBase database [84]; 4 = Aleuts [22]; 5 = Athapaskan [23]; 6 = Haida [25]; 7 = Apache [85], 8 = Bella Coola [25]; 9 = Navajo [85]; 10 = Sioux, 11 = Chippewa [66], 12 = Nuu-Chah-Nult [26]; 13 = Cheyenne [66]; 14 = Muskogean populations [31]; 15 = Cheyenne-Arapaho [30]; 16 = Yakima [23]; 17 = Stillwell Cherokee [30]; Meso-America: 18 = Pima [27]; 19 = Mexico [40]; 20 = Quiche [38]; 21 = Cuba [38]; 22 = El Salvador (present study); 23 = Huetar [35]; 24 = Emberá [39]; 25 = Kuna [36]; 26 = Ngöbé [37]; 27 = Wounan [37]; South America: 28 = Guahibo [51]; 29 = Yanomamo from Venezuela [48]; 30 = Gaviao [42]; 31 = Yanomamo from Venezuela and Brazil [86]; 32 = Colombia [52]; 33 = Ecuador (general population), 34 = Cayapa [41]; 35 = Xavante [42]; 36 = North Brazil [43]; 37 = Brazil [64]; 38 = Curiau [60]; 39 = Zoró [42]; 40 = Ignaciano, 41 = Yuracare [53]; 42 = Ayoreo [44]; 43 = Araucarians [33]; 44 = Pehuenche, 45 = Mapuche from Chile [45]; 46 = Coyas [4]; 47 = Tacuarembó [49]; 48 = Uruguay [50]; 49 = Mapuches from Argentina [46]; 50 = Yaghan [59].
doi:10.1371/journal.pone.0006882.g001

from Meso-America in some wave of migration towards the South (such as the one indicated above observed in Uruguay, or *e.g.* C16111T–T16189C–T16223C–C16290T–16311–G16319A–T16362C, which was also found in three Brazilians [64]).

We found only four Native American mtDNAs not belonging to haplogroup A2: two haplogroup B2 and two haplogroup C1 mtDNAs. We did not find any exact match amongst published data for the B2 sequence #33 that carries the distinctive variant A16269G. The haplogroup C1 sequence #41 carries C16256T; this uncommon variant within haplogroup C1 has been also observed in the Yanomama from Venezuela [47] and the Zoró from Brazil [42]. Haplotype C1 #39 was only observed in one Brazilian [43] and one Guahibo from Venezuela [51], but also in two ancient Taino samples from the Caribbean [57].

Non-Native American haplotypes in El Salvador

Signals of a European contribution to our sample from El Salvador are limited to three haplotypes (see Table 1): haplotype #96 belongs to haplogroup U2e, with exact matches in several West European locations (e.g. Northwest Spain and Portugal [70]); in Madeira [71], etc. Haplotype #12 can be assigned most plausibly to haplogroup H, while #30 probably belongs to haplogroup T2; this sequence curiously matches published sequences only observed in Portugal and Brazil [64,72] but also a single hit in Poland [73].

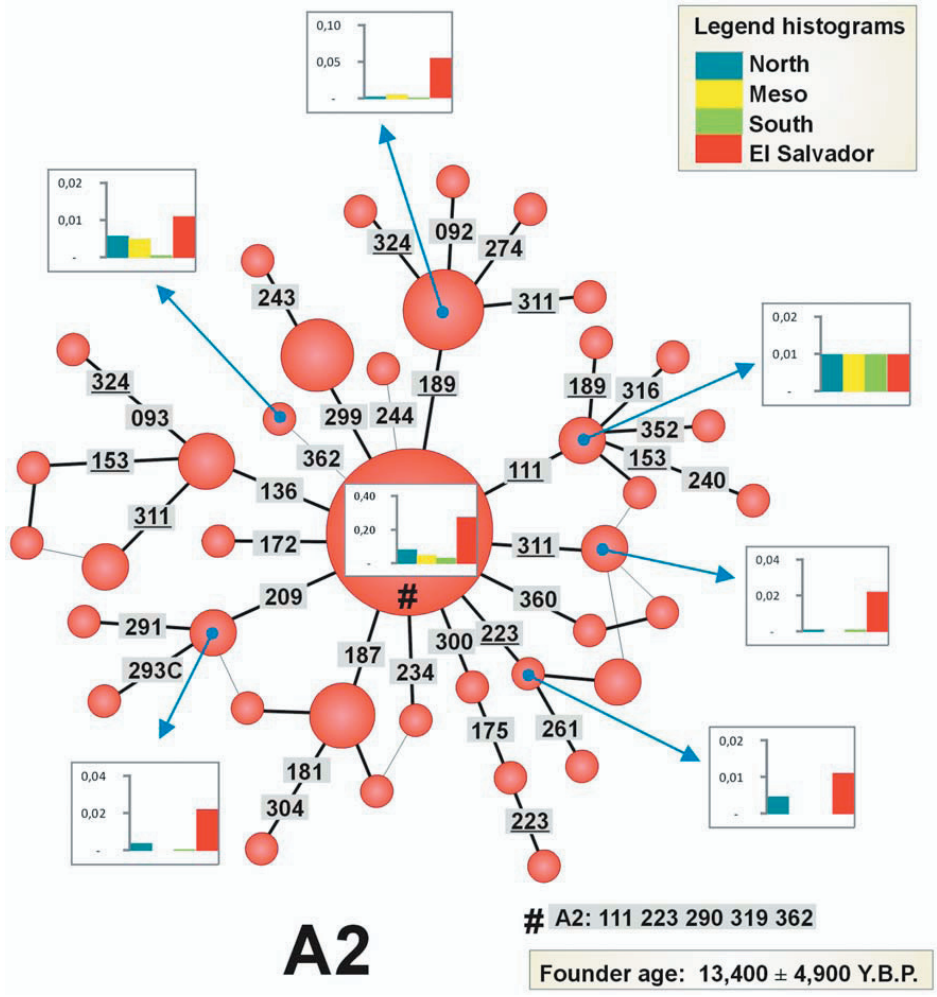


Figure 2. Phylogenetic network of haplogroup A2 mtDNA sequences from El Salvador. Only the variation contained in the HVS-I segment (range 16090–16365) was used. Positions are referred to the rCRS minus 16000; a transversion is specifically indicated as a suffix. The areas of the circles are proportional to the number of individuals bearing the corresponding haplotype in the El Salvador sample. The histograms indicate the frequency of the El Salvadorian mtDNA founder clusters in other continental regions (North, Meso, and South America). To compute the TMRCA and founder ages, we resolved the network to a tree (shown by the bold connections between haplotypes), using the positional mutation rates reported in [21]. Resolved parallel mutations are shown underlined. The ancestral haplotype is indicated by a hash.
doi:10.1371/journal.pone.0006882.g002

We detected only one sequence belonging to a typical sub-Saharan haplogroup in El Salvador. It belongs to L0a1a, a sub-clade highly prevalent in southeast Africa [67,74,75], where we find exact matches in HVS-I and HVS-II. Exact matches are also found in, for example, the Atlantic African southwest coast, in Cabinda [76], and in the Tongas [77]. Although it is not possible to determine with precision the African origin of this haplotype, southeast Africa (Mozambique) is probably the best candidate population source.

Admixture analysis

The admixture analysis carried out as in [67] indicated that North America accounts for ~92% of the lineages in El Salvador, the remaining ~8% coming from South America. The method described in [68] indicated that North America contributed to El Salvador ~76% of the mtDNA lineages, in contrast to the ~24% coming from South America.

Founder analysis

We inferred seven founders in our Salvadorian sample, all present in North American populations. Some sequence matches were not considered founders because they were detected only in Mexico and not in North American populations; they are more likely the result of recent gene flow between El Salvador and neighbouring populations. Some other potential founders were also rejected because they were present mostly as singletons analyzed in North American laboratories but belonging to e.g.

'Hispanic' or other sample populations without information about their ethnic affiliation. All the founders are indicated in Figure 1. The founder age of haplogroup A2 in El Salvador was estimated as 13,400±5,200 Y.B.P.

Discussion

El Salvador is the smallest Latin American republic and also the most densely populated. Although historically El Salvador has been home to a culturally diverse mix of peoples, including Native Americans, Africans, and west Europeans, by the 1980s the population of the country was essentially considered to be homogeneous in terms of ethnicity and basic cultural identity. Virtually all Salvadorans speak Spanish, the official language, as their mother tongue, and the vast majority are generally characterized as "mestizos" (or "ladinos", a term more commonly used in Central America), popularly used to refer to those persons of mosaic geographic ancestry who follow a wide variety of indigenous and "hispanic" customs and habits that over the centuries have come to constitute Spanish-American cultural patterns. In the late 1980s, the ethnic composition of the population was estimated as 89% "mestizo", 10% Native American, and 1% "white" [78]. Therefore, in contrast to most other Central American countries, El Salvador no longer possessed an ethnically or linguistically distinct Native American population, although persons of native-like ethnicity or cultural heritage still lived in the western parts of the country. Similarly, there was no ethnically or culturally distinct African-American population as there is in neighbouring populations [79]. However, there is a general belief that much of the Salvadorian population in the 1980s had a predominantly Native American ancestry[1].

The results of the present study have shown that, in contrast to the cultural patterns observed in the today's El Salvador population, most of the mtDNA profiles found are typically Native American; haplogroup A2 account for ~90% of the Salvadorian sample. Correspondingly, the impact of Europeans on the mtDNA pool of El Salvador is very low (~2%). It seems that the Spanish conquerors and more recent European demographic influences did not contribute significantly to the today's genetic composition of El Salvador in the maternal side. This contrasts with the European Y-chromosome contribution to the El Salvador gene pool. According to [80] about one half in metropolitan areas and two thirds in rural populations of El Salvador belong to non-Native American haplogroups; for instance, the most common Y-chromosome haplogroup in Europe (namely, R1b) is present in El Salvador at 24% in metropolitan areas and 43% in rural regions. Concomitantly, the Native American Y-chromosome proportion in El Salvador (represented by haplogroup Q3) is about 31–49%.

Therefore, the mtDNA and Y-chromosome variation in El Salvador displays an extreme version of a pattern that was also observed in other American populations [81,82]: the indigenous female contribution is much higher than the indigenous male contribution.

Our results show that the impact of African-American lineages on the mtDNA pool of El Salvador was very low, as indicated by the presence of only one mtDNA of sub-Saharan origin in our sample. The scarcity of the sub-Saharan component strikingly contrasts with the situation on the Caribbean coast, where (as a consequence of the Atlantic slave trade) it is clearly predominant [67,74,75,79]. The Y-chromosome variation shows a similar pattern: no lineages of African ancestry have been detected in El Salvador [80].

There are no clear signals of recent genetic drift events in the general population from El Salvador, as observed in, for instance, neighbouring but isolated Native American populations such as the Ngöbé from Panamá [37] which shows extremely reduced levels of mtDNA diversity (reflecting passage through post-conquest population bottlenecks). Haplogroup A2 is at high frequency in El Salvador (~90% of the sample) and a high percent of the lineages (45%; computed using the sequence range 16090–16365) remain unobserved in other American populations. Admixture analysis indicates that the main mtDNA influence in El Salvador can be attributed to North America. The phylogeny of A2 is rather star-like and the founder age was 12,600±4,900 years. The shape of this phylogeny points to the existence of a prehistoric demographic expansion. Considering the most recently estimated age of A2 in the American continent as a whole of 13,400±5,200 [21] (largely determined from North American samples) as a proxy for the time of the expansion into the Americas, it can be tentatively suggested that the initial settlement of El Salvador occurred rather soon after the initial colonization of the American continent, and that El Salvador largely contains the descendants of the mtDNAs in that original pool with scarce subsequent demographic influence from other American or non-American populations. Indeed, since we have genotyped samples collected in urban areas we would expect to have an even higher prevalence of the Native American component in more isolated groups from the country, as is in fact observed on the Y-chromosome side where the Native American component is higher in rural than in metropolitan areas [80].

In contrast to the high impact of the Atlantic slave trade on the Central American Caribbean coast [79], the Pacific side (at least for El Salvador) appears to have preserved its Native American mtDNA heritage intact to the present day. At the same time, this study has also shown that El Salvador harbours haplogroup frequency patterns quite different from other modern Native American communities. At the individual haplotype level, El Salvador shows numerous mtDNAs that have never been observed in other American regions, even within Central America. These features provide little support to those that assume (or claim) that "Hispanics" or Native American communities are sufficiently homogeneous to justify the portability of forensic databases from one country to another (e.g. SWGDAM; [34]); see [83] for a discussion.

Acknowledgments

We would like to thank Vilma de Aguilar (Ministerio de Salud Pública de El Salvador) for helping with the sample collection.

Author Contributions

Conceived and designed the experiments: AS. Performed the experiments: JLG VAI MC. Analyzed the data: AS VM MR. Contributed reagents/materials/analysis tools: AS MVL AC. Wrote the paper: AS VM MR.

References

1. Ministerio de Educación ES (1994) Historia de El Salvador.
2. Lardé y Larín J (2000) El Salvador: descubrimiento, conquista y colonización: Dirección de Publicaciones e Impresos, El Salvador.
3. Álvarez-Iglesias V, Salas A, Cerezo M, Ramos-Luis E, Jaime JC, et al. (2006) Genotyping coding region mtDNA SNPs for Asian and Native American haplogroup assignation. Int Congress Series 11: 4–6.
4. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int: Genet 1: 44–55.
5. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.

6. Bandelt H-J, Salas A, Bravi CM (2004) Problems in FBI mtDNA database. Science 305: 1402–1404.
7. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150–1160.
8. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891–899.
9. Salas A, Yao Y-G, Macaulay V, Vega A, Carracedo Á, et al. (2005) A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2: e296.
10. Yao Y-G, Salas A, Bravi CM, Bandelt H-J (2006) A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum Genet 119: 505–515.
11. Bandelt HJ, Salas A (2009) Contamination and sample mix-up can best explain some patterns of mtDNA instabilities in buccal cells and oral squamous cell carcinoma. BMC Cancer 16: 113.
12. Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, et al. (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3: e1764.
13. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, et al. (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1–8.
14. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, et al. (2007) Beringian standstill and spread of Native American founders. PLoS ONE 2: e829.
15. Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinformatics Online 1: 47–50.
16. Nei N (1987) Molecular evolutionary genetics: New York: Columbi University Press.
17. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37–48.
18. Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141: 743–753.
19. Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59: 935–945.
20. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718–726.
21. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet in press.
22. Rubicz R, Schurr TG, Babb PL, Crawford MH (2003) Mitochondrial DNA variation and the origins of the Aleuts. Hum Biol 75: 809–835.
23. Shields GF, Schmiechen AM, Frazier BL, Redd A, Voevoda MI, et al. (1993) mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. Am J Hum Genet 53: 549–562.
24. Starikovskaya YB, Sukernik RI, Schurr TG, Kogelnik AM, Wallace DC (1998) mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of ancient Beringia and the peopling of the New World. Am J Hum Genet 63: 1473–1491.
25. Ward RH, Redd A, Valencia D, Frazier B, Pääbo S (1993) Genetic and linguistic differentiation in the Americas. Proc Natl Acad Sci U S A 90: 10663–10667.
26. Ward RH, Frazier BL, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci U S A 88: 8720–8724.
27. Kittles RA, Bergen AW, Urbanek M, Virkkunen M, Linnoila M, et al. (1999) Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. Am J Phys Anthropol 108: 381–399.
28. Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, et al. (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53: 563–590.
29. Torroni A, Sukernik RI, Schurr TG, Starikovskaya YB, Cabell MF, et al. (1993) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet 53: 591–608.
30. Malhi RS, Schultz BA, Smith DG (2001) Distribution of mitochondrial DNA lineages among Native American tribes of Northeastern North America. Hum Biol 73: 17–55.
31. Bolnick DA, Smith DG (2003) Unexpected patterns of mitochondrial DNA variation among Native Americans from the southeastern United States. Am J Phys Anthropol 122: 336–354.
32. Lorenz JG, Smith DG (1994) Distribution of the 9-bp mitochondrial DNA region V deletion among North American Indians. Hum Biol 66: 777–788.
33. Horai S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, et al. (1993) Peopling of the Americas, founded by four major lineages of mitochondrial DNA. Mol Biol Evol 10: 23–47.
34. Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B (2002) The mtDNA Population Database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 4: no 2.
35. Santos M, Ward RH, Barrantes R (1994) mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Hum Biol 66: 963–977.
36. Batista O, Kolman CJ, Bermingham E (1995) Mitochondrial DNA diversity in the Kuna Amerinds of Panama. Hum Mol Genet 4: 921–929.
37. Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, et al. (1995) Reduced mtDNA diversity in the Ngöbé Amerinds of Panamá. Genetics 140: 275–283.
38. Boles TC, Snow CC, Stover E (1995) Forensic DNA testing on skeletal remains from mass graves: a pilot project in Guatemala. J Forensic Sci 40: 349–355.
39. Kolman CJ, Bermingham E (1997) Mitochondrial and nuclear DNA diversity in the Choco and Chibcha Amerinds of Panama. Genetics 147: 1289–1302.
40. Green LD, Derr JN, Knight A (2000) mtDNA affinities of the peoples of North-Central Mexico. Am J Hum Genet 66: 989–998.
41. Rickards O, Martinez-Labarga C, Lum JK, De Stefano GF, Cann RL (1999) mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations. Am J Hum Genet 65: 519–530.
42. Ward RH, Salzano FM, Bonatto SL, Hutz MH, Coimbra CEA, et al. (1996) Mitochondrial DNA polymorphism in 3 Brazilian Indian tribes. Am J Hum Biol 8: 317–323.
43. Santos SE, Ribeiro-Dos-Santos AK, Meyer D, Zago MA (1996) Multiple founder haplotypes of mitochondrial DNA in Amerindians revealed by RFLP and sequencing. Ann Hum Genet 60 (Pt 4): 305–319.
44. Dornelles CL, Battilana J, Fagundes NJ, Freitas LB, Bonatto SL, et al. (2004) Mitochondrial DNA and Alu insertions in a genetically peculiar population: the Ayoreo Indians of Bolivia and Paraguay. Am J Hum Biol 16: 479–488.
45. Moraga ML, Rocco P, Miquel JF, Nervi F, Llop E, et al. (2000) Mitochondrial DNA polymorphisms in Chilean aboriginal populations: implications for the peopling of the southern cone of the continent. Am J Phys Anthropol 113: 19–29.
46. Ginther C, Corach D, Penacino GA, Rey JA, Carnese FR, et al. (1993) Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. Exs 67: 211–219.
47. Merriwether DA, Kemp BM, Crews DE, Neel JV (2000) Gene flow and genetic variation in the Yanomama as revealed by mitochondrial DNA. In: Renfrew C, ed. America Past, America Present: Genes and Languages in the Americas and Beyond. Cambridge: McDonald Institute for Archaeological Research. pp 89–124.
48. Williams SR, Chagnon NA, Spielman RS (2002) Nuclear and mitochondrial genetic variation in the Yanomamo: a test case for ancient DNA studies of prehistoric populations. Am J Phys Anthropol 117: 246–259.
49. Bonilla C, Bertoni B, Gonzalez S, Cardoso H, Brum-Zorrilla N, et al. (2004) Substantial Native American female contribution to the population of Tacuarembo, Uruguay, reveals past episodes of sex-biased gene flow. Am J Hum Biol 16: 289–297.
50. Pagano S, Sans M, Pimenoff V, Cantera AM, Alvarez JC, et al. (2005) Assessment of HV1 and HV2 mtDNA variation for forensic purposes in an Uruguayan population sample. J Forensic Sci 50: 1239–1242.
51. Vona G, Falchi A, Moral P, Calo CM, Varesi L (2005) Mitochondrial sequence variation in the Guahibo Amerindian population from Venezuela. Am J Phys Anthropol 127: 361–369.
52. Torres MM, Bravi CM, Bortolini MC, Duque C, Callegari-Jacques S, et al. (2006) A revertant of the major founder Native American haplogroup C common in populations from northern South America. Am J Hum Biol 18: 59–65.
53. Bert F, Corella A, Gene M, Perez-Perez A, Turbon D (2004) Mitochondrial DNA diversity in the Llanos de Moxos: Moxo, Movima and Yuracare Amerindian populations from Bolivia lowlands. Ann Hum Biol 31: 9–28.
54. Vernesi C, Fuselli S, Castri L, Bertorelle G, Barbujani G (2002) Mitochondrial diversity in linguistic isolates of the Alps: a reappraisal. Hum Biol 74: 725–730.
55. Monsalve MV, Stone AC, Lewis CM, Rempel A, Richards M, et al. (2002) Brief communication: molecular analysis of the Kwaday Dan Ts'finchi ancient remains found in a glacier in Canada. Am J Phys Anthropol 119: 288–291.
56. Stone AC, Stoneking M (1993) Ancient DNA from a pre-Columbian Amerindian population. Am J Phys Anthropol 92: 463–471.
57. Lalueza-Fox C, Calderon FL, Calafell F, Morera B, Bertranpetit J (2001) MtDNA from extinct Tainos and the peopling of the Caribbean. Ann Hum Genet 65: 137–151.
58. Lalueza-Fox C, Gilbert MT, Martinez-Fuentes AJ, Calafell F, Bertranpetit J (2003) Mitochondrial DNA from pre-Columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean. Am J Phys Anthropol 121: 97–108.
59. Moraga M, Santoro CM, Standen VG, Carvallo P, Rothhammer F (2005) Microevolution in prehistoric Andean populations: chronologic mtDNA variation in the desert valleys of northern Chile. Am J Phys Anthropol 127: 170–181.
60. Ribeiro-Dos-Santos AK, Santos SE, Machado AL, Guapindaia V, Zago MA (1996) Heterogeneity of mitochondrial DNA haplotypes in Pre-Columbian Natives of the Amazon region. Am J Phys Anthropol 101: 29–37.
61. García-Bour J, Pérez-Pérez A, Álvarez S, Fernández E, López-Parra AM, et al. (2004) Early population differentiation in extinct aborigines from Tierra del Fuego-Patagonia: ancient mtDNA sequences and Y-chromosome STR characterization. Am J Phys Anthropol 123: 361–370.
62. Brown MD, Hosseini SH, Torroni A, Bandelt H-J, Allen JC, et al. (1998) mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? Am J Hum Genet 63: 1852–1861.
63. Dornelles CL, Bonatto SL, De Freitas LB, Salzano FM (2005) Is haplogroup X present in extant South American Indians? Am J Phys Anthropol 127: 439–448.

64. Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt H-J, et al. (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67: 444–461.
65. Ribeiro-Dos-Santos AK, Carvalho BM, Feio-Dos-Santos AC, Santos SE (2006) Nucleotide variability of HV-I in Afro-descendents populations of the Brazilian Amazon Region. Forensic Sci Int.
66. Smith DG, Malhi RS, Eshleman J, Lorenz JG, Kaestle FA (1999) Distribution of mtDNA haplogroup X among Native North Americans. Am J Phys Anthropol 110: 271–284.
67. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, et al. (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454–465.
68. Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, et al. (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol 8: 213.
69. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276.
70. González AM, Brehm A, Pérez JA, Maca-Meyer N, Flores C, et al. (2003) Mitochondrial DNA affinities at the Atlantic fringe of Europe. Am J Phys Anthropol 120: 391–404.
71. Brehm A, Pereira L, Kivisild T, Amorim A (2003) Mitochondrial portraits of the Madeira and Acores archipelagos witness different genetic pools of its settlers. Hum Genet 114: 77–86.
72. Pereira L, Prata MJ, Amorim A (2000) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann Hum Genet 64: 491–506.
73. Malyarchuk BA, Grzybowski T, Derenko MV, Czarny J, Wozniak M, et al. (2002) Mitochondrial DNA variability in Poles and Russians. Ann Hum Genet 66: 261–283.
74. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, et al. (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082–1111.
75. Salas A, Carracedo Á, Richards M, Macaulay V (2005) Charting the Ancestry of African Americans. Am J Hum Genet 77: 676–680.
76. Beleza S, Gusmão L, Amorim A, Carracedo Á, Salas A (2005) The genetic legacy of western Bantu migrations. Hum Genet 117: 366–375.
77. Trovoada MJ, Pereira L, Gusmao L, Abade A, Amorim A, et al. (2004) Pattern of mtDNA variation in three populations from Sao Tome e Principe. Ann Hum Genet 68: 40–54.
78. CONCULTURA (2000) Biblioteca de Historia Salvadoreña, El Salvador, Historia de sus pueblos, villas y ciudades. Consejo Nacional para la Cultura y el Arte.
79. Salas A, Richards M, Lareu MV, Sobrino B, Silva S, et al. (2005) Shipwrecks and founder effects: Divergent demographic histories reflected in Caribbean mtDNA. Am J Phys Anthropol 128: 855–860.
80. Lovo-Gómez J, Blanco-Verea A, Lareu MV, Brion M, Carracedo Á (2007) The genetic male legacy from El Salvador. Forensic Sci Int 171: 198–203.
81. Dipierri JE, Alfaro E, Martínez-Marignac VL, Bailliet G, Bravi CM, et al. (1998) Paternal directional mating in two Amerindian subpopulations located at different altitudes in northwestern Argentina. Hum Biol 70: 1001–1010.
82. Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo Á (2008) Gender bias in the multi-ethnic genetic composition of Central Argentina. J Hum Genet 53: 662–674.
83. Salas A, Bandelt H-J, Macaulay V, Richards MB (2007) Phylogeographic investigations: The role of trees in forensic genetics. Forensic Sci Int 168: 1–13.
84. Burckhardt F, von Haeseler A, Meyer S (1999) HvrBase: compilation of mtDNA control region sequences from primates. Nucleic Acids Res 27: 138–142.
85. Budowle B, Allard MW, Fisher CL, Isenberg AR, Monson KL, et al. (2002) HVI and HVII mitochondrial DNA data in Apaches and Navajos. Int J Legal Med 116: 212–215.
86. Merriwether DA, Kemp BM, Crews DE, Neel JV (2002) Gene flow and genetic variation in the Yanomama as revealed by mitochondrial DNA. Cambridge: McDonald Institute for Archaeological Research. pp 89–124.

Electrophoresis 2009, *30*, 3665–3673

María Cerezo[1]
Viktor Černý[2]
Ángel Carracedo[1]
Antonio Salas[1]

[1]Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain
[2]Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, The Czech Republic

Received May 6, 2009
Revised July 2, 2009
Accepted July 9, 2009

Research Article

Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies

Analysis of the mitochondrial DNA variation in populations is commonly carried out in many fields of biomedical research. We propose the analysis of mitochondrial DNA coding region SNP (mtSNP) variation to a high level of phylogenetic resolution based on MALDI-TOF MS. The African phylogeny has been chosen to test the applicability of the technique but any other part of the worldwide phylogeny (or any other mtSNP panel) could be equally suitable for MALDI-TOF MS genotyping. SNP selection thus aimed to fully cover all the mtSNPs defining major and minor branches of the known African tree, including, macro-haplogroup L, and haplogroups M1, and U6. A total of 230 mtSNPs were finally selected. We used tests samples collected from two different African locations, namely, Mozambique and Chad Basin. Different internal genotyping controls and other indirect approaches (*e.g.* phylogenetic checking coupled with automatic sequencing) were used in order to evaluate the reproducibility of the technique, which resulted to be 100% using samples previously subjected to whole genome amplification. The advantages of the MALDI-TOF MS are also discussed in comparison with other popular methods such as minisequencing, highlighting its high-throughput nature, which is particularly suitable for case–control medical studies, forensic databasing or population and anthropological studies.

Keywords:
Haplogroup / High-throughput SNP genotyping / MALDI-TOF MS / Mitochondrial DNA / Multiplex assay DOI 10.1002/elps.200900294

1 Introduction

The analysis of mitochondrial DNA (mtDNA) variation in human populations has captivated many laboratories interested in different forensic, medical, and anthropological applications. Molecular anthropology is devoted to the analysis of genetic variation in human population groups generally pre-selected on the basis of geographic and/or ethnical criteria. Most of the contemporary studies in this field of research commonly analyze the first (and sometimes the second) hypervariable segment (HVS-I/II) of the mtDNA control region. In the best of the cases, these data are usually complemented with the genotyping of few selected mitochondrial DNA coding region SNPs (mtSNPs) that allow a better classification of samples into haplogroups (group of phylogenetically related haplotypes). Genotyping of mtSNPs is frequently performed using RFLP analysis. Ideally, whole genome sequencing would provide the maximum level of resolution of the mtDNA molecule. In practice, analysis of complete genome sequencing is only performed nowadays when there is a particular interest in resolving some specific branch of the mtDNA phylogeny to the highest level of resolution, or when there is some interest to search for potential pathogenic mutations (*e.g.* [1–9]). Population studies based on complete genome sequencing are still limited because they demand a very intense economical, technical, and personal effort. The same reasons preclude the analysis of complete genomes in forensic applications, with the aggravated fact that usually, forensic casework samples contain degraded and/or very limited amounts of DNA.

Many techniques have been applied for mtSNP genotyping. The ones used in the past have been designed primarily for screening purposes (*e.g.* SSCP or heteroduplex analyses) or for the genotyping of pre-selected SNPs in some specific application (*e.g.* RFLPs classification of mtDNA haplotypes into haplogroups) (see, *e.g.* [10–13]). Most recently, minisequencing has gained more popularity among geneticists due to the fact that the technique is

Correspondence: Dr. Antonio Salas, Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, 15782, Universidade de Santiago de Compostela, Galicia, Spain
E-mail: antonio.salas@usc.es
Fax: +34-981-580-336

Abbreviations: mtDNA, mitochondrial DNA; **mtSNP,** mitochondrial DNA coding region; **WGA,** whole genome amplification

relatively simple and is more cost effective than other alternative ones (*e.g.* [14–17]).

Any of the above-mentioned techniques allow the genotyping of large amount of SNPs in large collections of samples as demanded in current molecular anthropological studies or even in many forensic applications (*e.g.* creation of SNP forensic database). However, some more suitable techniques are available for these purposes, such as those based on MS. In fact, there are many medical applications that have benefitted from the use of MALDI-TOF MS to the genotyping of medium to large number of SNPs in disease studies (*e.g.* [18–22]), although the technique has been more popularly used for protein characterization [23, 24].

To our knowledge, there was only one attempt to apply MALDI-TOF MS for the genotyping of mtDNA variation. Xiu-Cheng Fan *et al.* [25] have developed a Sequenom assay to genotype 18 control region SNPs. However, the potential applications of MS for high-throughput mtSNP genotyping and the rationale for SNP selection played a secondary role in this pioneering study.

In brief, the present study has been motivated by the unfeasibility of today's techniques for complete genome sequencing applied to large population scale studies, and the low-throughput capacity of today's mtSNP genotyping techniques. In consequence, we here aimed to evaluate the ability of MALDI-TOF MS to genotype large collection of mtSNPs (pre-selected on the basis of phylogenetic criteria) and samples that generally characterize population-based studies.

2 Materials and methods

2.1 Population samples

A total of 90 blood stain and saliva samples were used to test the MALDI-TOF MS assay; 30 were collected in Mozambique and 60 in the Chad Basin region comprising northern Cameroon, neighboring part of Chad, southeastern Niger and northeastern Nigeria [26] (in what follows, Chad Basin region). The ones from Mozambique are different to those genotyped in [27], while the ones in the Chad Basin overlapped with the ones genotyped for the HVS-I region and some RFLPs in [26].

DNA extraction of the Mozambique samples was carried out using standard phenol–chloroform methods. The samples from Chad were extracted as indicated in [26]. Theoretically, the DNA required for MALDI-TOF MS genotyping considering the seven multiplexes developed in the present study would be ~500 ng, an amount of DNA that is rarely available in population and anthropological projects. Therefore, all the samples analyzed in the present study were previously subjected to whole genome amplification (WGA) using Genomiphi (GE Healthcare Life Sciences, Uppsala, Sweden). Only 1 μL of the original extracted DNA was used for WGA (without undertaken previous DNA quantification) using the manufactured protocol. The WGA product was subsequently water diluted 1:16 and then directly used for MALDI-TOF MS genotyping.

2.2 Rationale for SNP selection

The rationale of SNP design was based on the following criteria: (i) we have considered diagnostic SNPs defining main and minor branches of the mtDNA African phylogeny, including not only the autochthonous sub-Saharan macrohaplogroup L, but also, haplogroups M1 and U6, the two southwestern Asian clades that have also spread into North and sub-Saharan Africa; (ii) mtSNPs were selected with the primary aim of optimizing the level of haplogroup resolution, given the fact that the control region (or primarily the HVS-I) was already sequenced; (iii) recurrent mutations and known hotspots in the phylogeny [28–30] were avoided if other candidate variants were available for the same phylogenetic branch; (iv) transversions were preferred over transitions; and (v) indels were selected if any other variant was available for the targeted phylogenetic branch. The final list of mtSNP candidates aimed to cover almost all the phylogenetic branches of the known African mtDNA phylogeny.

The African phylogeny and nomenclature are very complex and was elaborated during the last few years based on several control region and complete genome sequencing efforts (*e.g.* [3, 5, 10, 27, 31–37]). For the sake of clarity, we provide two phylogenetic trees (Figs. 1 and 2) mainly inspired and adapted from the Behar *et al.*'s study [5], indicating all the mtSNPs genotyped in the present study plus diagnostic control region variants.

2.3 MALDI-TOF MS genotyping

Genotyping was performed using the MassARRAY SNP genotyping system (Sequenom, San Diego, CA, USA) located at the Universidad de Santiago de Compostela (Galicia, Spain), following the manufacturer's instructions. The technology was applied as explained in [19, 20]. In brief, this typing assay uses the extension of a single primer that binds to the sequence flanking the mutation site. Base-specific primer extension products are created one to four bases long depending on the substitution present. The different primer extension products are then differentiated by mass. Multiple mtSNPs can be typed simultaneously by multiplexing the extension reaction. We used the iPLEX^{TM} assay to increase plexing efficiency and flexibility with Mass ARRAY platform. Detection uses MALDI-TOF MS with samples automatically genotyped from each mass spectrum produced.

The MALDI-TOF MS assays were designed using Assay Design 3.1 software. All the samples were genotyped in 384-well plates and following automated protocols. Spectro TYPER-RT software was used for allele-calling of all possible SNPs in each DNA sample. Amplification and extension probes are reported in Supporting Information Table 1. Figure 3 shows an example of a Sequenom mass spectrum, genotyping calls for variants at np 5096, and a Sequenom mass spectrum for the mtSNP 3420.

Genotyped error rate was 0% as assessed by the use of several positive and negative controls as well as inferred from other indirect approaches (see below).

Electrophoresis 2009, *30*, 3665–3673

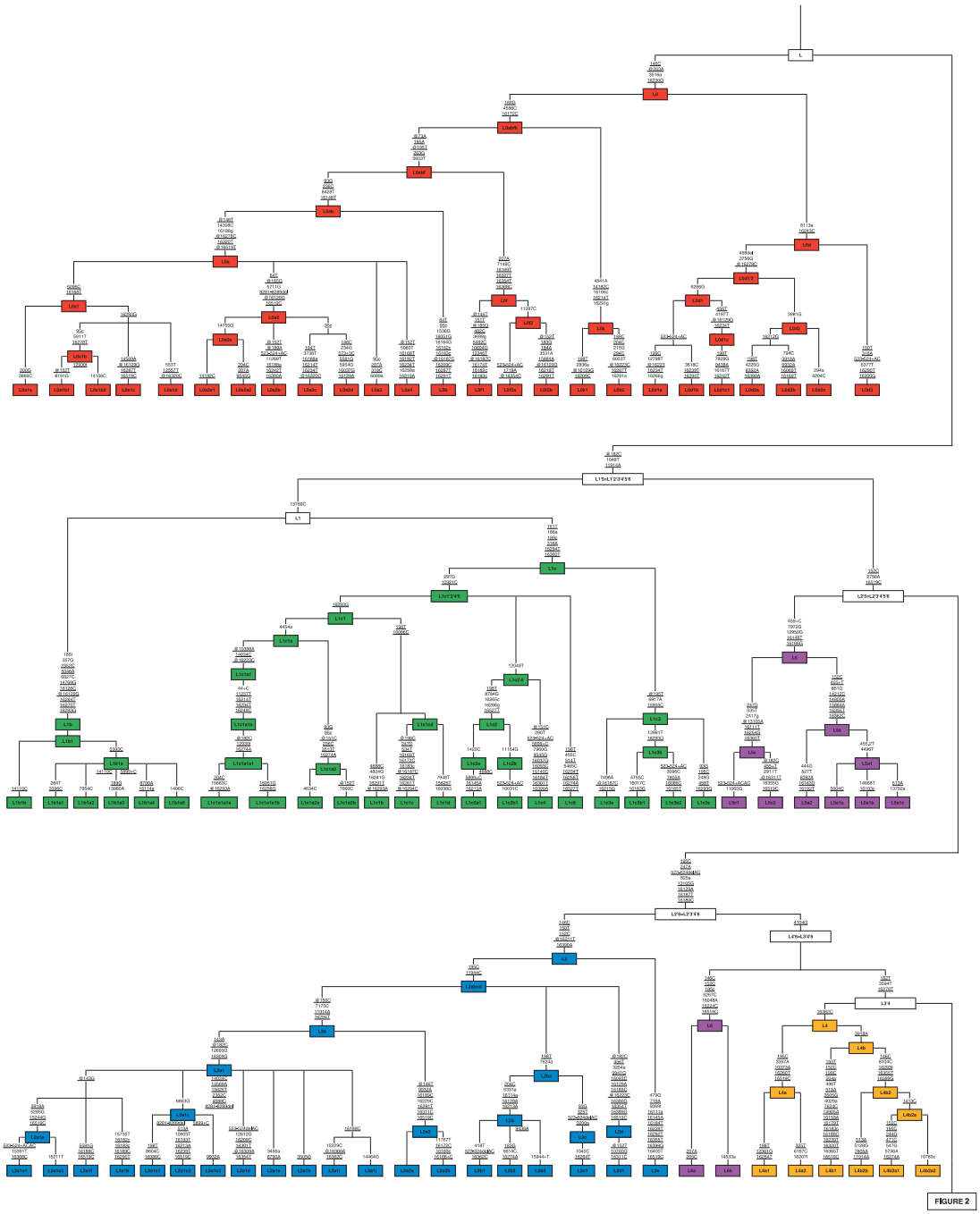


Figure 1. Phylogeny of L haplogroups (excluding L3) as defined by the mtSNPs selected in the present study. The diagnostic control region variants for each branch are also indicated in the tree. Transitions are indicated using suffixes in upper case, while transversions are indicated using suffixes in lower case; parallel mutations are underlined; ''@'' indicates a reversion, ''del'' a deletion, and ''+'' an insertion.

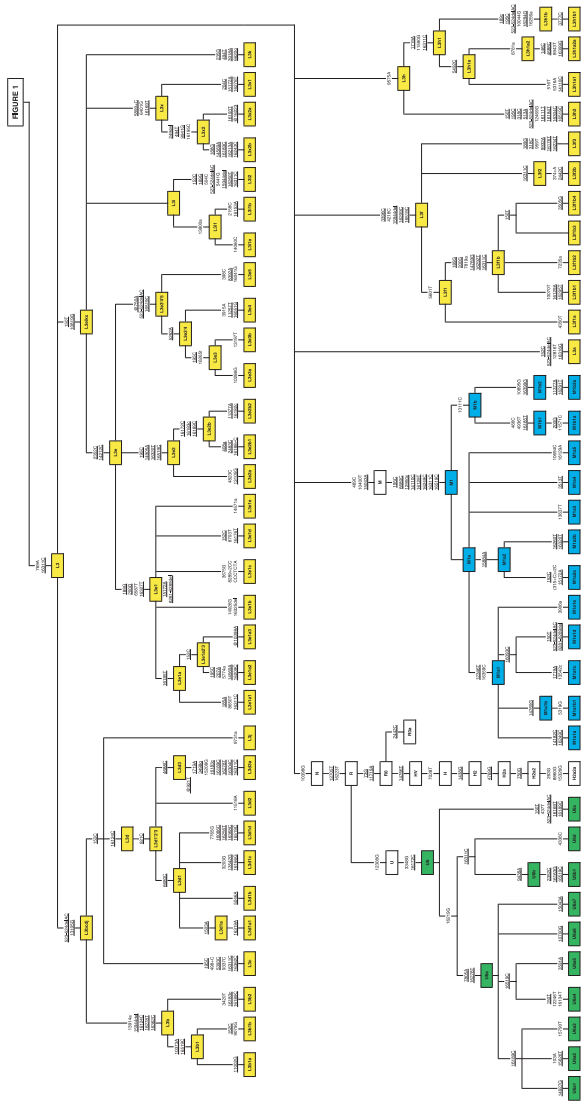


Figure 2. Phylogeny of L3 and non-L haplogroups as defined by the mtSNPs selected in the present study. Codes are as in Fig. 1.

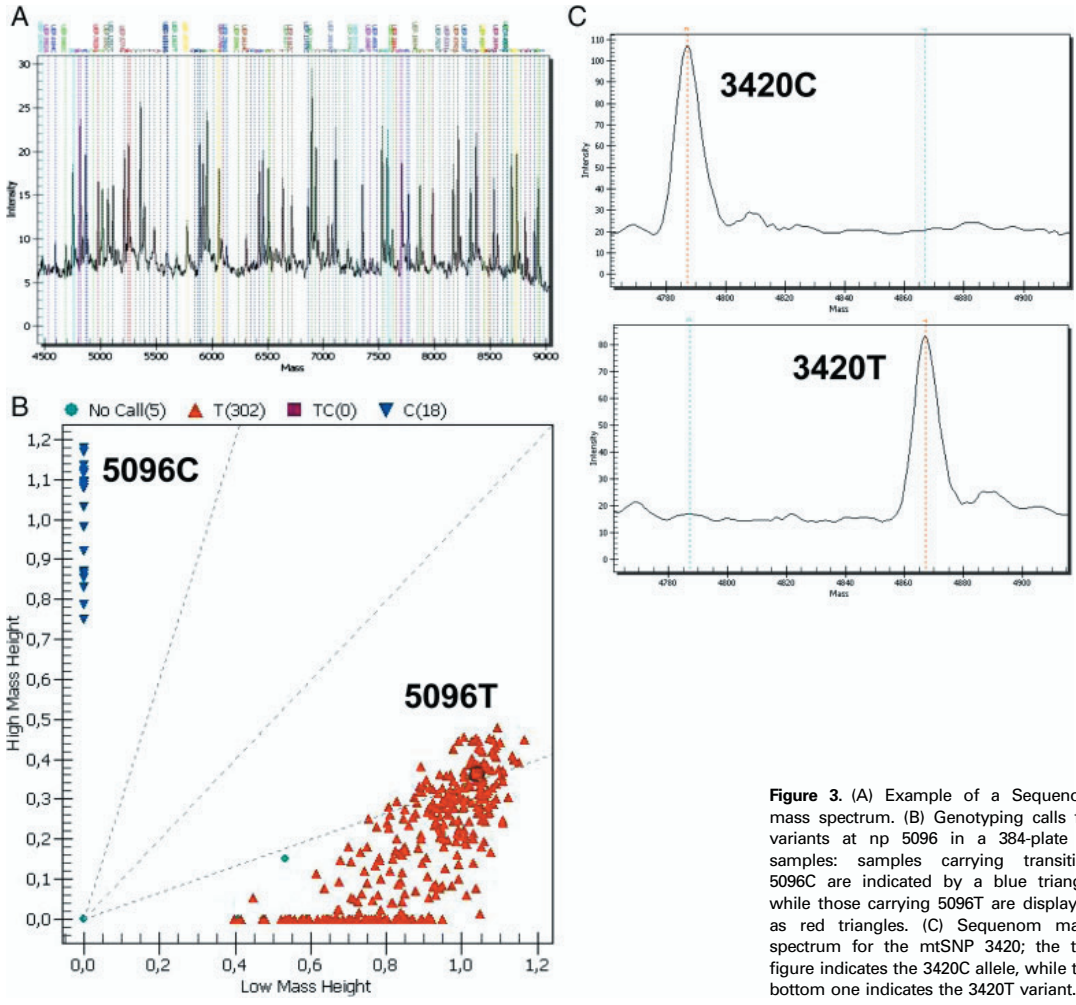


Figure 3. (A) Example of a Sequenom mass spectrum. (B) Genotyping calls for variants at np 5096 in a 384-plate of samples: samples carrying transition 5096C are indicated by a blue triangle while those carrying 5096T are displayed as red triangles. (C) Sequenom mass spectrum for the mtSNP 3420; the top figure indicates the 3420C allele, while the bottom one indicates the 3420T variant.

2.4 Standard sequencing analysis

The mtSNP variation observed in the genotyped samples was contrasted with the mtDNA variation expected according to the known mtDNA phylogeny. Some phylogenetic inconsistencies were observed (sensus, *e.g.* [38, 39, 49]). When testing a genotyping platform for a new application, it is convenient to disregard other potential sources of inconsistencies [16], as it could be the case of genotyping errors or mix-up and contamination [40–45]. Therefore, to confirm that these inconsistencies correspond to natural reversions (commonly observed along the mtDNA phylogeny, specially at mutational hotspots [28, 30, 38]) and not to artifactual errors introduced by the MALDI-TOF MS technique, we double checked these variants using standard automatic sequencing procedures, as described in [46].

3 Results and discussion

3.1 Monitoring quality control

The full set of sequencing results and mtSNP data for the samples analyzed in the present project are reported in Supporting Information Table 2. We have used the known phylogeny to identify all potential mtSNP inconsistencies (indicated in Supporting Information Table 2) according to the procedures employed in previous studies [38, 42, 47–49]. All the observed phylogenetic inconsistencies were sequenced using standard protocols and all of them could be fully corroborated. In addition, an internal control to evaluate the reproducibility of the genotyping was also carried out by replicating a total of 3330 genotypes. All the genotypes (excluding failed ones that could not be compared, 0.14%) were 100% replicated.

The reproducibility of the MS results could also be confirmed by direct comparison with complete genome sequencing data. Thus, a set of complete genomes belonging to L3f from the Chad Basin was independently obtained in [50]; some of them overlapping with the samples analyzed in the present study. All the diagnostic coding region variants captured with the MS assays developed in the present study were observed by complete genome sequencing, including T959C and A9932G defining L3f3, or T4350C defining L3f1a, *etc*. Apart from the L3f branches identified in our samples, the set of mtSNP genotyped using MALDI-TOF MS are prepared to discriminate among many other L3f branches, as for instance, L3f2, L3f1b, *etc*. (see Fig. 2); these branches were however not observed in our set of samples.

Call rate of our MALDI-TOF MS assay was above 90% for 75% of the mtSNPs. We are aware that other publications (*e.g.* [51]) support a much better average calling rate for the technique (~95%). This parameter is however strongly dependent on the quality of the samples. In fact, the ones used in the present project are not the paradigm of best quality DNA since all of them were DNA extracts from blood stain or saliva samples that have been stored for more than a decade at room temperature (blood stains) or frozen (saliva hyssops). The fact that mtDNA is inherited as an haplotype block coupled with the fact that the today's mtDNA phylogeny is reasonably well known, allowed the classification of any mtDNA into the corresponding sub-haplogroups even in cases where the calling rate was not very favorable; note that the most informative SNPs for haplogroup classification are generally those located at the tip branches of the phylogeny.

3.2 Variation captured by control region sequencing and the MS assay

It is worth mentioning that the mtSNPs considered in the present study were selected with the main aim of allocating mtDNAs of African ancestry to their corresponding (sub)haplogroups, in part motivated by the relatively low phylogenetic resolution provided by previous studies dealing with mtDNA variation in African populations (*e.g.* [31, 52–55]) or African descends in America (*e.g.* [27, 33, 56–59]), with the exception of those projects based on complete genome sequencing that are however mainly focused on the African phylogeny instead of population variability (*e.g.* [5, 32, 34, 66]). The success of the present design cannot be evaluated by their discrimination power (a parameter of more interest among, *e.g.* forensic than population genetics), for which a selection of hotspots and/or in combination with diagnostic sites could be more appropriate. It can be said however that the assays provided in the present study are actually a good discrimination tool if we consider that the whole SNP panel coupled with the HVS-I information allows to discriminate 77 haplotypes classified into 48 different haplogroups among the 90 samples available.

Theoretically, the whole mtSNP panel genotyped using MALDI-TOF MS would allow the classification of African mtDNAs into at least 236 different haplogroups (mtSNP-haplotypes), while the HVS-I segment, for example, could only discriminate among 132 haplogroups, with the additional drawback of (generally) substantiating the phylogenetic classification on more unstable variants.

3.3 Advantages and disadvantages of MALDI-TOF MS

The main advantages of high-throughput genotyping using MALDI-TOF MS with respect to other techniques can be summarized as follows: (i) the technique is suitable for the genotyping of both large collections of samples and mtSNPs, in contrast to other more popular genotyping methodologies that are more effective for low-scale genotyping projects (*e.g.* minisequencing); (ii) almost all the steps involved in the study design and genotyping can be monitorized by using *ad hoc* software (*e.g.* primer design and automatic reading of the genotyping spectrum); (iii) MALDI-TOF is also suitable for robot monitoring of pre-genotyping tasks (*e.g.* sample managing, sample dilutions, *etc*.), which constitutes an important advantage to prevent sample mix-up and contamination, namely, one of the main source of error in mtDNA studies [49–51, 61–63]; (iv) MALDI-TOF MS is much faster than other common techniques – genotyping of few thousand samples (including also the pre-genotyping tasks involved in the preparation of samples) for a set of 230 SNPs can be executed in less than 2 wk, in comparison to the several months needed to carry out the same genotyping effort using, *e.g.* minisequencing or RFLPs-based procedures; (v) when genotyping large collection of samples the cost-benefit of MALDI-TOF MS improves substantially with respect to other techniques; for instance, we estimate the cost *per* genotype using MS to be at least half the price (~0.1 Euros/genotype) of the cost using minisequencing-SNaPshot (averaging the costs of previous assays designed by the authors; see [14–16]); (vi) the genotyping is highly reproducible; this is in part because allele assignment is deduced from the intrinsic molecular mass of the products and, therefore, the results are not influenced by external reaction conditions [64]; (vii) the technique allows a great flexibility in terms of plexing design, and therefore it can be used for many other biomedical, forensic, or anthropological applications; note however that each application would demand different mtSNP panels; and (viii) the small sizes of the amplicons makes the technique suitable for the analysis of highly degraded samples, a fact that is of special interest for forensic and other medical applications; MALDI-TOF MS seems to be even more sensitive than sequencing procedures for the detection of heteroplasmy status, as previously described in [25]. Note that the present study does not focus in the latter applications of the technique because

heteroplasmic status is irrelevant in population-based studies (this condition is an individual attribute that is not stable in populations). A careful evaluation of the MALDI-TOF MS applicability in forensic casework of molecular diagnosis would require more specific designs in order to properly evaluate its ability for the detection of heteroplasmy. Another issue that would remain to be evaluated is the ability of the MALDI-TOF MS to deal with the genotyping of repetitive regions in the mtDNA molecule (*e.g.* homopolymeric tracks in the control region). This variation was not targeted in the present study because it is – as heteroplasmies – virtually uninformative from the point of view of a population geneticist (due to their high mutation rate). Apart from the mentioned population and forensic applications mentioned above, we foresee other implementations that could benefit from the MALDI-TOF MS technique. For instance, genotyping of large collections of mtSNPs in population samples could be a good approach for the estimation of positional mutation rates [28]. Case–control association studies aimed to disentangle the complex nature of multi-factorial diseases are very popular since the mtDNA is assumed to contribute to the disease phenotype of a wide spectrum of common diseases (*e.g.* [65, 66]), usually, these studies generally focus in large collections of samples and a number of SNPs representing the main branches of the phylogeny of interest (*e.g.* European, Asian, *etc.*).

Apart from the high prices of the instruments, the main disadvantage of the MALDI-TOF MS approach is the need of relatively large amounts of DNA, which however is comparable to other techniques aimed to genotype the same amount of SNPs or much lower than the amount of DNA needed to carry out complete genome sequencing. Note however that WGA could be an optimal solution to overcome the limitation of having low DNA amount (provided the DNA is not degraded). As demonstrated in the present study, WGA is fully reliable as testified by the robustness of the genotyping results using MALDI-TOF MS and the several checking procedures used to monitor genotyping error. WGA would enormously benefit those laboratories handling samples containing low DNA quantity (*e.g.* hair shafts, saliva, blood stains, *etc.*), which is a common scenario in population-based studies.

In contrast to the whole genome sequencing approach, MALDI-TOF MS cannot detect new mutations, but as new branches of mtDNA phylogeny are already discovered, the technique can be very useful to allocate mtDNA into the corresponding haplogroups and the detection of phylogenetically related samples.

4 Concluding remarks

The MALDI-TOF MS assays proposed in the present study have been designed in order to provide a useful and efficient tool for population-based mtDNA studies, involving large amount of samples and mtSNPs. The technique is highly reproducible and cost-effective in this high-throughput context. The careful selection of mtSNPs sets carried out in the present study allows capturing the haplogroup status of any African mtDNA considering the extant African mtDNA phylogeny. The technique is highly flexible and would allow to genotype whatever mtSNP combination of interest in other human population backgrounds (European, worldwide, admixed, *etc.*) or other medical applications related to disease studies or forensic casework.

We would like to thank María Torres for technical assistance. This work was supported by grants from the Xunta de Galicia (Grupos Emerxentes; 2008/037), Ministerio de Ciencia e Innovación (SAF2008-02971), and Fundación de Investigación Médica Mutua Madrileña (2008/CL444) given to A. S., and Grant Agency of the Czech Republic (grant no. 206/08/1587) given to V. č.

The authors have declared no conflict of interest.

5 References

[1] Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J. *et al.*, *Science* 2005, *308*, 1034–1036.

[2] Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q. P., Woodward, S. R., Salas, A. *et al.*, *PLoS ONE* 2008, *3*, e1764.

[3] Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R. *et al.*, *Science* 2006, *314*, 1767–1770.

[4] Brisighelli, F., Capelli, C., Alvarez-Iglesias, V., Onofri, V., Paoli, G., Tofanelli, S., Carracedo, A. *et al.*, *Eur. J. Hum. Genet.* 2009, *17*, 693–696.

[5] Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzar, R. *et al.*, *Am. J. Hum. Genet.* 2008, *82*, 1130–1140.

[6] Palanichamy, M. g., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y. *et al.*, *Am. J. Hum. Genet.* 2004, *75*, 966–978.

[7] Coble, M. D., Just, R. S., O'Callaghan, J. E., Letmanyi, I. H., Peterson, C. T., Irwin, J. A., Parsons, T. J. *et al.*, *Int. J. Legal. Med.* 2004, *118*, 137–146.

[8] Ingman, M., Gyllensten, U., *Nucleic Acids Res.* 2006, *34*, D749–D751.

[9] Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 171–176.

[10] Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J. *et al.*, *Am. J. Hum. Genet.* 2004, *75*, 752–770.

[11] Salas, A., Rasmussen, E. M., Lareu, M. V., Morling, N., Carracedo, Á., *Forensic Sci. Int.* 2001, *124*, 97–103.

[12] Barros, F., Lareu, M. V., Salas, A., Carracedo, A., *Electrophoresis* 1997, *18*, 52–54.

[13] Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, E., Scozzari, R. *et al., Am. J. Hum. Genet.* 1999, *64*, 232–249.

[14] Quintáns, B., Álvarez-Iglesias, V., Salas, A., Phillips, C., Lareu, M. V., Carracedo, Á., *Forensic Sci. Int.* 2004, *140*, 251–257.

[15] Álvarez-Iglesias, V., Barros, F., Carracedo, Á., Salas, A., *BMC Med. Genet.* 2008, *9*, 26.

[16] Álvarez-Iglesias, V., Jaime, J. C., Carracedo, Á., Salas, A., *Forensic Sci. Int. Genet.* 2007, *1*, 44–55.

[17] Brandstätter, A., Salas, A., Niederstätter, H., Gassner, C., Carracedo, Á., Parson, W., *Electrophoresis* 2006, *27*, 2541–2550.

[18] Ribas, G., González-Neira, A., Salas, A., Milne, R. L., Vega, A., Carracedo, B., Gonzélez, E. *et al., Hum. Genet.* 2006, *118*, 669–679.

[19] Vega, A., Salas, A., Milne, R. L., Carracedo, B., Ribas, G., Ruibal, A., de Leon, A. C. *et al., Gynecol. Oncol.* 2009, *112*, 210–214.

[20] Salas, A., Vega, A., Milne, R. L., García-Magariños, M., Ruibal, Á., Benítez, J., Carracedo, Á. *et al., Clin. Med. Oncol.* 2008, *2*, 357–362.

[21] Huang, H. L., Stasyk, T., Morandell, S., Dieplinger, H., Falkensammer, G., Griesmacher, A., Mogg, M. *et al., Electrophoresis* 2006, *27*, 1641–1650.

[22] Li, Y., Wenzel, F., Holzgreve, W., Hahn, S., *Electrophoresis* 2006, *27*, 3889–3896.

[23] Fernández, J., Gharahdaghi, F., Mische, S. M., *Electrophoresis* 1998, *19*, 1036–1045.

[24] Silvertand, L. H., Torano, J. S., de Jong, G. J., van Bennekom, W. P., *Electrophoresis* 2009, *30*, 1828–1835.

[25] Xiu-Cheng Fan, A., Garritsen, H. S., Tarhouny, S. E., Morris, M., Hahn, S., Holzgreve, W., Zhong, X. Y., *Clin. Chem. Lab. Med.* 2008, *46*, 299–305.

[26] Černý, V., Salas, A., Hájek, M., Zaloudkova, M., Brdička, R., *Ann. Hum. Genet.* 2007, *71*, 433–452.

[27] Salas, A., Richards, M., De la Fé, T., Lareu, M. V., Sobrino, B., Sánchez-Diz, P., Macaulay, V. *et al., Am. J. Hum. Genet.* 2002, *71*, 1082–1111.

[28] Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A. *et al., Am. J. Hum. Genet.* 2009, *84*, 740–759.

[29] Bandelt, H.-J., Quintana-Murci, L., Salas, A., Macaulay, V., *Am. J. Hum. Genet.* 2002, *71*, 1150–1160.

[30] Malyarchuk, B. A., Rogozin, I. B., *Ann. Hum. Genet.* 2004, *68*, 324–339.

[31] Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L. *et al., Proc. Natl. Acad. Sci. USA* 2008, *105*, 1596–1601.

[32] Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A., Tishkoff, S. A., *Mol. Biol. Evol.* 2007, *24*, 757–768.

[33] Salas, A., Richards, M., Lareu, M. V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V. *et al., Am. J. Hum. Genet.* 2004, *74*, 454–465.

[34] Torroni, A., Achilli, A., Macaulay, V., Richards, M., Bandelt, H.-J., *Trends Genet.* 2006, *22*, 339–345.

[35] Beleza, S., Gusmão, L., Amorim, A., Carracedo, Á., Salas, A., *Hum. Genet.* 2005, *117*, 366–375.

[36] Plaza, S., Salas, A., Calafell, F., Corte-Real, F., Bertranpetit, J., Carracedo, Á., Comas, D., *Hum. Genet.* 2004, *115*, 439–447.

[37] van Oven, M., Kayser, M., *Hum. Mutat.* 2009, *30*, E386–E394.

[38] Bandelt, H.-J., Quintana-Murci, L., Salas, A., Macaulay, V., *Am. J. Hum. Genet.* 2002, *71*, 1150–1160.

[39] Kong, Q. P., Bandelt, H.-J., Sun, C., Yong, Y. G., Salas, A., Achilli, A., Wang, C. Y. *et al., Hum. Mol. Genet.* 2006, *15*, 2076–2086.

[40] Bandelt, H.-J., Salas, A., *BMC Cancer* 2009, *9*, 113.

[41] Bandelt, H.-J., Olivieri, A., Bravi, C., Yong, Y. G., Torroni, A., Salas, A., *Eur. J. Hum. Genet.* 2007, *15*, 402–404.

[42] Salas, A., Yong, Y. G., Macaulay, V., Vega, A., Carracedo, A., Bandelt, H.-J., *PLoS Med.* 2005, *2*, e296.

[43] Yong, Y. G., Kong, Q. P., Salas, A., Bandelt, H.-J., *J. Med. Genet.* 2008, *45*, 769–772.

[44] Yong, Y. G., Salas, A., Bravi, C. M., Bandelt, H.-J., *Hum. Genet.* 2006, *119*, 505–515.

[45] Salas, A., Bandelt, H.-J., Macaulay, V., Richards, M. B., *Forensic Sci. Int.* 2007, *168*, 1–13.

[46] Álvarez-Iglesias, V., Mosquera-Miguel, A., Cerezo, M., Quintáns, B., Zarrabeitia, M. T., Cuscó, I., Lareu, M. V., *et al., PLoS ONE* 2009, *4*, e5112.

[47] Bandelt, H.-J., Salas, A., Bravi, C. M., *Science* 2004, *305*, 1402–1404.

[48] Bandelt, H.-J., Salas, A., Lutz-Bonengel, S., *Int. J. Legal Med.* 2004, *118*, 267–273.

[49] Salas, A., Carracedo, Á., Macaulay, V., Richards, M., Bandelt, H.-J., *Biochem. Biophys. Res. Commun.* 2005, *335*, 891–899.

[50] Černý, V., Fernandes, V., Costa, M. D., Hájek, M., Mulligan, C. J., Pereira, L., *BMC Evol. Biol.* 2009, *9*, 63.

[51] Oeth, P., Beaulieu, M., Park, C., Kosman, D., del Mistro, G., van den Boom, D., Jurinke, C., *Sequenom Application Note* 2007, *Doc. No. 8876-006*, R05.

[52] Beleza, S., Gusmao, L., Amorim, A., Carracedo, A., Salas, A., *Hum. Genet.* 2005, *117*, 366–375.

[53] Plaza, S., Salas, A., Calafell, F., Corte-Real, F., Bertranpetit, J., Carracedo, A., Comas, D., *Hum. Genet.* 2004, *115*, 439–447.

[54] Rando, J. C., Cabrera, V. M., Larruga, J. M., Hernández, M., González, A. M., Pinto, F., Bandelt, H.-J., *Ann. Hum. Genet.* 1999, *63*, 413–428.

[55] Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., Pääbo, S., *Am. J. Hum. Genet.* 1996, *59*, 437–444.

[56] Salas, A., Carracedo, Á., Richards, M., Macaulay, V., *Am. J. Hum. Genet.* 2005, *77*, 676–680.

[57] Salas, A., Richards, M., Lareu, M. V., Sobrino, B., Silva, S., Matamoros, M., Macaulay, V. *et al., Am. J. Phys. Anthropol.* 2005, *128*, 855–860.

[58] Silva, W. A., Bortolini, M. C., Schneider, M. P., Marrero, A., Elion, J., Krishnamoorthy, R., Zago, M. A., *Hum. Biol.* 2006, *78*, 29–41.

[59] Ely, B., Wilson, J. L., Jackson, F., Jackson, B. A., *BMC Biol.* 2006, *4*, 34.

[60] Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K., Passarino, G. *et al.*, *Genetics* 2006, *172*, 373–387.

[61] Salas, A., Bandelt, H.-J., Macaulay, V., Richards, M. B., *Forensic Sci. Int.* 2007, *168*, 1–13.

[62] Kong, Q. P., Salas, A., Sun, C., Fuku, N., Tanaka, M., Zhong, L., Wang, C. Y. *et al.*, *PLoS ONE* 2008, *3*, e3016.

[63] Bandelt, H.-J., Kong, Q. P., Parson, W., Salas, A., *J. Med. Genet.* 2005, *42*, 957–960.

[64] Petkovski, E., Keyser-Tracqui, C., Hienne, R., Ludes, B., *J. Forensic Sci.* 2005, *50*, 535–541.

[65] Mosquera-Miguel, A., Álvarez-Iglesias, V., Vega, A., Milne, R., Cabrera de León, A., Benitez, J., Carracedo, Á. *et al.*, *Cancer Res.* 2008, *68*, 623–625.

[66] Baudouin, S. V., Saunders, D., Tiangyou, W., Elson, J. L., Poynter, J., Pyle, A., Keers, S. *et al.*, *Lancet* 2005, *366*, 2118–2121.

European Journal of Human Genetics (2010), 1–9

www.nature.com/ejhg

ARTICLE

Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel

Luísa Pereira[1,2], Viktor Černý*[,3], María Cerezo[4], Nuno M Silva[1], Martin Hájek[3], Alžběta Vašíková[5], Martina Kujanová[6], Radim Brdička[5] and Antonio Salas[4]

The Tuareg presently live in the Sahara and the Sahel. Their ancestors are commonly believed to be the Garamantes of the Libyan Fezzan, ever since it was suggested by authors of antiquity. Biological evidence, based on classical genetic markers, however, indicates kinship with the Beja of Eastern Sudan. Our study of mitochondrial DNA (mtDNA) sequences and Y chromosome SNPs of three different southern Tuareg groups from Mali, Burkina Faso and the Republic of Niger reveals a West Eurasian-North African composition of their gene pool. The data show that certain genetic lineages could not have been introduced into this population earlier than ~9000 years ago whereas local expansions establish a minimal date at around 3000 years ago. Some of the mtDNA haplogroups observed in the Tuareg population were involved in the post-Last Glacial Maximum human expansion from Iberian refugia towards both Europe and North Africa. Interestingly, no Near Eastern mtDNA lineages connected with the Neolithic expansion have been observed in our population sample. On the other hand, the Y chromosome SNPs data show that the paternal lineages can very probably be traced to the Near Eastern Neolithic demic expansion towards North Africa, a period that is otherwise concordant with the above-mentioned mtDNA expansion. The time frame for the migration of the Tuareg towards the African Sahel belt overlaps that of early Holocene climatic changes across the Sahara (from the optimal greening ~10 000 YBP to the extant aridity beginning at ~6000 YBP) and the migrations of other African nomadic peoples in the area.
European Journal of Human Genetics advance online publication, 17 March 2010; doi:10.1038/ejhg.2010.21

Keywords: Tuareg; genetic diversity; phylogeography

INTRODUCTION

The Tuareg call themselves *Kel Tamasheq* (people of the *Tamasheq* language) or *Imashaghen* (free people). The *Tamasheq* tongue is a Berber language belonging to the Afro-Asiatic phylum. The Tuareg maintain a nomadic and/or semi-nomadic lifestyle in the Central Sahara and adjacent regions of the African Sahel, where they number about 1 262 000 in total. Their contemporary geographic distribution is shown in the upper map in Figure 1.

The 5th century BC Greek historian Herodotus suggested that the ancestral homeland of the ancient Tuareg (ie Garamantes) was the Libyan Fezzan.[1] It has been suggested that subsequent to the camel being adopted for Saharan trade in the 1st or 2nd century AD, the area of Tuareg influence expanded further to the south. Oral accounts and sparse written records in *Tifinagh* (the script of the *Tamasheq* language) date back to 14th century AD when the first caravan traders were documented in the Air Mountains. Hereafter, it seems that from the 17th century onwards the increasingly frequent invasions of North Africa by various Arabic tribes drove the Tuareg yet further southward to the African Sahel.

Since the beginning of European explorations of the Sahara and the Sahel, the Tuareg have been known mainly as caravan traders linking the sub-Saharan and Mediterranean cultures. Their contact with the various sub-Saharan peoples was not always peaceful and they were known to take war captives. Centuries of mutual contact led to substantial assimilation of others into the Tuareg population.

Carrying out biological or genetic investigations of the Tuareg has not always been easy because of their demanding lifestyle and their often negative attitude to the European colonists. Cavalli-Sforza *et al*,[2] whose synthesized study of classical protein and serological markers is well known, noticed a genetic link between the Tuareg and Beja from Eastern Sudan. The fact that the genetic distances between the Tuareg and Berber/North-western Africans were larger than that between the Tuareg and Beja, provides a picture of a common origin and population separation at some point more than 5000 years ago. Interestingly, both people are also pastoralist and speak Afro-Asiatic languages, even if the Beja language (*Bedawi*), with its four dialects, belongs to the Cushitic branch, whereas *Tamasheq* belongs to the Berber branch. The fact that these two peoples today speak different languages might be

[1]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal; [2]Faculdade de Medicina da Universidade do Porto, Porto, Portugal; [3]Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, The Czech Republic; [4]Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Facultade de Medicina, Instituto de Medicina Legal, Universidade de Santiago de Compostela, Galicia, Spain; [5]Institute of Hematology and Blood Transfusion, Prague, The Czech Republic; [6]Department of Anthropology and Human Genetics, Faculty of Science, Charles University, Prague, The Czech Republic
*Correspondence: Dr V Černý, Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Letenska 4, 118 01, Prague 1, The Czech Republic.
Tel: +420 2570 14304; Fax: +420 2575 32288; E-mail: cerny@aup.cas.cz
Received 21 July 2009; revised 15 January 2010; accepted 20 January 2010

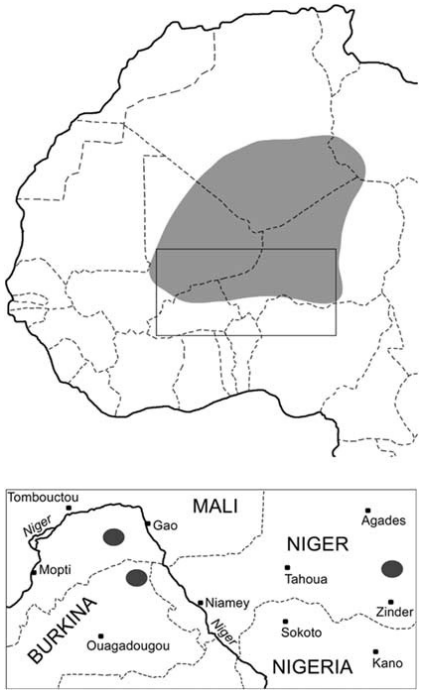


Figure 1 The geographical location of southern Tuareg populations, including the ones studied here: TTan in the Republic of Niger, TGor in Burkina Faso and TGos in Mali.

explained either by the Tuareg having acquired the Berber language during their westwards migration, or possibly by the Beja coming under the influence of some Eastern African peoples as language shift is a relatively common phenomenon.

Among the first African mitochondrial DNA (mtDNA) sequences were those from data sets[3,4] obtained mostly from Tuareg living in Niger and Nigeria, and which revealed a rather sub-Saharan affinity of their population. More recently, however, a study based on 129 Tuareg samples from two villages of the Libyan Fezzan, stressed a high frequency but concomitant low diversity of the West Eurasian component, bearing only haplogroups H1, V and M1. The sub-Saharan component of the Libyan Tuareg was more diversified but predominantly represented by only two haplogroups (L2a1 and L0a1a). The Tuareg population from Libya was homogenous with very low estimates of haplotype diversity suggesting high genetic drift.[5]

The above-mentioned studies have thus revealed a dual influence in the genetic make-up of this African people. In this study, we provide new mtDNA and Y chromosome data sets of three unrelated Tuareg groups from three different countries (Niger, Mali and Burkina Faso). At the same time, we try to unravel the questions of their genetic origin, the mutual relationships among their sub-populations as well as possible links to neighbouring populations. The genetic heritage of the Tuareg population is analysed within the context of the West Eurasian *versus* sub-Saharan contributions to their gene pool.

MATERIALS AND METHODS

Subjects

The biological samples (buccal swabs) were obtained from three different groups of self-identified Tuareg (90 unrelated individuals in total). One population sample (n=38) was secured in Burkina Faso around the village Gorom-Gorom (further referred to as TGor). The second sample (n=31) was taken in the Republic of Niger in the vicinity of Tanut (TTan). The third sample (n=21) was collected in Mali near Gossi (TGos). The samples from Mali and Burkina Faso are geographically relatively close to each other as they are located within the bend of the Niger River, whereas the sample collected in the central part of the Republic of Niger is located some 1500 km eastward (Figure 1). The field sampling was undertaken with the collaboration of local Tuareg assistants. Of these 90 healthy and unrelated individuals, 47 were male and 43 were female. Oral informed consent was obtained from all participants in the study and research permits were obtained from the Ministries of Education and/or Health in all the three countries.

Laboratory analyses

DNA extractions and PCR amplifications of mtDNA hypervariable segments I and II (HVS-I and HVS-II) were carried out as in earlier studies.[6,7] Amplicons were purified and sequenced using forward PCR primers. In some cases the reverse primer was also used for sequencing (due to for example, the presence of poly-C stretches between nt 16184 and 16193). In some samples, SNP testing was analysed through matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and minisequencing.[7,8] Whole genome sequencing of mtDNAs affiliated by D-loop to the haplogroup M1 was undertaken following the protocols reported elsewhere.[7,9–11]

For the detection of Y chromosome SNP polymorphisms, the Signet Y-SNP Identification System v 2.0 (Marligen, Rockville, MD, USA) was used. All the samples were first analysed by A-R multiplex (polymorphisms M122, M168, M175, M207, M304, M343, M45, M89, M96 and M9), differentiating main evolutionary branches of Y chromosome phylogeny.[12] Subsequently, samples belonging to haplogroup E were further analysed for polymorphisms DYS391, M2, M33, M35, M58, M75, M78, M81 and M123.

Haplogroup nomenclature

mtDNA classification into haplogroups L, H, U and M was carried out in accordance with the most recent phylogenetic studies of Salas *et al*,[13] Kivisild *et al*[14] and Behar *et al*[15] for L; Olivieri *et al*[16] for U and M; and Achilli *et al*[17] for H (see also van Oven and Kayser, 2009).[18] The numbering is consistent with the revised Cambridge Reference Sequence.[19] The three complete mtDNA sequences have been deposited in GenBank (accession numbers GQ377749; GQ377750; and GQ377751). For the Y chromosome SNP affiliation, the nomenclature of Karafet *et al*[12] was followed.

Statistical and phylogenetic analyses

Analysis of population structure and molecular diversity measures was calculated by using Arlequin software version 3.0.[20] Two-tailed Fisher's exact test P-values of 2×2 contingency tables were calculated in DnaSP.[21] F_{ST} genetic distances calculated by Arlequin were subsequently visualized in multidimensional scaling (MDS) by means of SPSS 10.0 software (SPSS Inc, Chicago, IL, USA). Extensive data sets of both mtDNA sequences and Y-SNPs were used to characterize the Tuareg diversity within Mediterranean and sub-Saharan population contexts (see Supplementary Material SM1 and SM2).

Phylogenetic reconstruction of mtDNA diversity was based on both HVS-I and complete sequences. The dates of the most recent common ancestor of specific subclusters in the phylogeny were estimated using ρ.[22] The average number of transitions from the ancestral haplotype to all haplotypes in the cluster, for both coding (between positions 577 and 16 023) and HVS-I (between positions 16 090 and 16 365) regions, was considered with respect to mutation rate estimates of 5138 years[23] and 20 180 years per transition[24] within the region, respectively. Standard errors were calculated as in Saillard *et al*.[25] Recently, updated mutation rates published by Soares *et al*[26] were also used for the entire molecule (1 mutation every 3624 years) and synonymous substitutions (1 mutation every 7884 years). As the mutation rate determined by these authors for the HVS-I (between positions 16 090 and 16 365) region,

however, is very similar to the one used above (1 transition every 20 129 years *versus* 20 180 years) the calculations overlapped and we present only those estimated with the original ρ mutation rate of 1 substitution every 20 180 years.

Interpolation maps

To determine and visualize the geographical distribution of haplogroups H1, H3 and V, we drew interpolation maps using the 'Spatial Analyst Extension' of ArcView version 3.2 (http://www.esri.com/software/arcview/). Inverse distance weighted option that we used assumes that each input point has a local influence that diminishes with distance. The geographic location used is the centre of the distribution area, from where the individual samples of each population were collected. Comparative data for H1 and H3 were taken from Finnila *et al*,[27] Herrnstadt *et al*,[28] Pereira *et al*,[29] Cherni *et al*[30] and Ennafaa *et al*;[31] and those for haplogroup V from Torroni *et al*,[10] Pereira *et al*,[32] Behar *et al*[33] and Cherni *et al*.[30]

RESULTS

The mtDNA pool of the Tuareg

The polymorphisms present in the 90 Tuareg individuals led to the identification of 53 different D-loop haplotypes (Table 1). As can be seen in the network based only on HVS-I diversity (Supplementary Material SM3), for which only 33 different haplotypes are observed, there are varying degrees of sharing of haplotypes among the analysed groups: only one belonging to haplogroup H was shared by all three groups; two haplotypes were shared by TGos–TGor and TGos–TTan; and three haplotypes were shared by TGor–TTan. Only 18 of the 33 haplotypes are unique, what is a rather low proportion when compared to most African samples.[32] This is further corroborated by haplotype diversities in the three Tuareg samples, which are lower as compared with other populations – especially in the two groups of the Niger bend (0.861 ± 0.027 in TGor; 0.910 ± 0.037 in TGos; and 0.963 ± 0.020 in TTan; see Supplementary Material SM4).

A total of 48% of the mtDNA haplotypes observed in the Tuareg populations could be ascribed to sub-Saharan haplogroups. Another 39%, however, were of West Eurasian ancestry (non-L types in Table 1), which is a substantial proportion considering the sub-Saharan geographical location. In fact, it has been observed that in typical North African populations there is a gradient of increasing frequency of West Eurasian lineages ranging from around 50–75% in the northernmost locations.[34] The Tuareg's neighbours, however, have a markedly smaller proportion of West Eurasian haplotypes (22% in Western Chad Arabs, 8% in Shuwa Arabs from North-eastern Nigeria, 7% in the Buduma from South-eastern Niger and 6% in the Kanuri from North-eastern Nigeria).[35] The remaining 13% of Tuareg haplotypes belong to the typical East African haplogroup M1.

Furthermore, we noticed some differences in the distribution of West Eurasian mtDNA haplogroups between Tuareg groups. Most of the West Eurasian haplogroups (30 out of 35 sequences, amounting to 6 out of 9 HVS-I haplotypes) and the East African M1 (11 out of 12 sequences but amounting to only 2 out of 3 HVS-I haplotypes) are observed in the two Tuareg populations – TGos and TGor – located within the bend of the Niger. Tuareg from the Republic of Niger, TTan, have much higher proportion of sub-Saharan (81%) haplogroups than of West Eurasian (16%) and East African (3%) ones. These differences in haplogroup distribution led to statistically significant genetic distances when comparing HVS-I haplotypes between Tuareg from Mali (TGos) with those from the Republic of Niger (TTan) (F_{ST}=0.048; unadjusted *P*-value=0.009), as well as Tuareg from Burkina Faso (TGor) with those from the Republic of Niger (TTan) (F_{ST}=0.064; unadjusted *P*-value=0.000), whereas Tuareg from Mali (TGos) and from Burkina Faso (TGor) are not statistically different (F_{ST}=0.012; unadjusted *P*-value=0.234). Similarly, analysis of MDS based on F_{ST} distances and using a large database of West Eurasian and African mtDNA sequences has shown a very good separation of the sub-Saharan and West Eurasian-North African gene pools (Figure 2). Only some East African populations are closer to the West Eurasian samples, respectively, to the North African populations analysed here. This picture is a good representation of F_{ST} values as the normalized raw stress is very low (0.01165). However, the analysed Tuareg populations are divided between two gene pools: like the sample from Libya,[5] the groups located within the bend of Niger (TGor and TGos) fall into the West Eurasian gene pool, whereas the Tuareg from the Republic of Niger (TTan) and the Tuareg sample from the Watson's data set[3,4] are permeated by the sub-Saharan mtDNA gene pool.

The West Eurasian component observed in the Tuareg is highly interesting. A major proportion (94%) could be allocated to haplogroups H1, H3 and V, West Eurasian lineages of Iberian origin that spread to Europe[7,10,17,26,29,36] and most probably North Africa[30,31] with the improvement of the climatic conditions after the retreat of the ice sheets 15 000–13 000 years ago. The interpolation maps of these lineages across North Africa and Europe (Supplementary Material SM5) clearly place the Tuareg population in the path of the southern African edge of post-Last Glacial Maximum expansions. The H1 haplogroup (Supplementary Material SM5A and SM5B, with and without the outlier Norway, respectively) is as frequent in our southern Tuareg groups as in Libya and the centre of the dispersion within the Iberian Peninsula. The H3 haplogroup is almost vestigial in Tuareg (Supplementary Material SM5C), having the highest observed frequencies outside of Iberia in Algeria and Tunisia. Again for haplogroup V, Tuareg present frequencies as high as in the Basque country (Supplementary Material SM5D).

Both H1 and H3 commonly display rather low diversity in the D-loop region, but the Tuareg haplotypes belonging to haplogroup V have a specific diagnostic mutation – the transition at position 16 234. All the Tuareg V haplotype samples collected in Burkina Faso and the Republic of Niger (three haplotypes observed in 11 individuals) bear this mutation together with the defining substitution at position 16 298, This polymorphism is present in two of the five V haplotypes observed in the recently published Libyan Tuaregs.[5] This fact seems to point to a founder effect in haplogroup V occurring in our southern Tuareg population; the further presence of two other polymorphisms in two V samples (substitutions at positions 16 189 and 16 293) allows a very preliminary estimation for the Time to the Most Recent Common Ancestor (TMRCA) of this Tuareg V sub-lineage at around 3600 ± 2600 years ago or at a maximum of 8800 years ago if using a 95% confidence interval (see Supplementary Material SM6 for the network).

Another very interesting characteristic of the West Eurasian mtDNA pool in the Tuareg population as a whole (including Watson's and Ottoni's data sets) is the total absence of so-called Neolithic haplogroups derived from the branch JT, which are otherwise common in Near Eastern, North African, Mediterranean and even some East African populations. The virtual absence of these lineages in the Tuareg is statistically significant when comparing the frequency of these lineages in Morocco[34] (18%; unadjusted *P*-value=0.000), Tunisia[30] (12%; unadjusted *P*-value=0.000) and Egypt[37] (29%; unadjusted *P*-value=0.000). Notice also the absence of the haplogroup U6, which is present mainly in Berbers but also in several others North African groups.[13,16,38]

The sub-Saharan mtDNA pool of the Tuareg is composed of various lineages from the major L-type haplogroups including: 2.3% of L0; 14.0% of L1; 58.1% of L2; 23.3% L3; and 2.3% of L4. We assayed to

Table 1 HVS-I, HVS-II and other polymorphisms observed in the 90 Tuareg samples from the three populations (TGos, TGor and TTan), as well as its haplogroup (HG) affiliation

Haplotype ID	*TGos*	*TGor*	*TTan*	*HV-I*	*HV-II*	*Other polymorphisms*	*HG*
T1	1			129 183C 189 519	263 315.1C	750 3010 4769 8602	H1*
T2		1		189 519	073 263 315.1C	750 3010 4769 14552	H1a?
T3	2			189 519	073 263 315.1C	750 3010 4769 8602	H1
T4	1			189 519	073 263 315.1C 328	750 3010 4769 8602	H1
T5	1			189 192C>T 519	073 263 315.1C	750 3010 4769 8602	H1
T6	1			311 519	152 263 315.1C	750 3010 4769 8602	H1
T7		1		311 519	263 315.1C	750	U3/H?
T8	1			311 519	263 315.1C	750 3010 4769 8602	H1
T9		1		519	073 309.1 315.1C	750 3010 4769 12308 14552	H1a
T10		1		519	152 263 315.1C	750 3010 4769 14552	H1
T11		2		519	263 309.1 315.1C	750 3010 4769 14552T	H1
T12	1			519	263 309.1C 315.1C	750 4769 8602	H3*
T13		2	1	519	263 315.1C	750 3010 4769 14552	H1
T14		3		519	263 315.1C	750 3010 4769 14552T	H1
T15	3			519	263 315.1C	750 3010 4769 8602	H1
T16			1	093 260 343 390 519	073 150 263 315.1C		U3a
T17		1		234 298	064 072 239 263 309.1 315.1C	750 2706 4580 4769 7028 14552	V
T18			1	234 298	072 263 309.1 315.1C	750 2706 4580 4769 7028 12308 14552	V
T19			1	189 234 298	072 263 309.1 315.1C	750 2706 4580 4769 7028 14552T	V
T20			1	234 293 298	072 263 309.1 315.1C	750 2706 4580 4769 7028 12308 14552	V
T21		5		234 298	072 239 263 309.1 315.1C	750 2706 4580 4769 7028 14552	V
T22		1		234 298	072 239 263 309.1 315.1C	750 2706 4580 4769 7028 12308 14552	V
T23		1		234 298	072 239 263 309.1 315.1C	750 2706 4580 4769 7028 14552T	V
T24			1	129 148 168 172 187 188G 189 193 223 230 278 293 311 320	093 095C 152 185 189 236 247 263 309.1C 315.1 523–524delAC		L0a1a
T25	1			093 126 145 187 189 223 264 270 278 293 311 519	073 152 182 185T 195 247 263 315.1C 357 523–524 delAC		L1b
T26			1	126 172 187 189 223 264 270 278 293 311 519	073 146 152 182 185T 195 247 263 315.1C 357 523–524delAC		L1b1
T27			1	126 187 189 223 264 270 278 293 311	073 146 152 263 315.1C 357 523–524delAC 573insCCCC		L1b1
T28			1	126 187 189 223 264 270 278 293 311 519	073 152 182 185T 189 195 247 263 315.1C 357 523–524delAC		L1b1
T29	1			126 187 189 223 264 270 278 293 311 519	073 152 182 185T 189 195 247 263 309.1 315.1C 357 523–524delAC		L1b1
T30			1	129 183C 189 215 223 261 278 294 311 360 519	073 151 152 182 186A 189C 247 263 315.1C 316 523–524delAC		L1c
T31	1			172 213 223 261 278 318 390	073 146 150 152 182 195 198 263 315.1C 523–524delAC		L2*
T32	1	2		093 189 192 223 278 294 311 390 519	073 143 146 152 195 263 309.1 315.1C		L2a
T33		2	1	093 189 193 223 278 294 311 390 519	073 143 146 152 195 263 315.1C		L2a
T34	1			169 223 239 278 294 309 390 526	073 143 146 152 189 195 263 309.1C 315.1C 523–524delAC		L2a

Table 1 (Continued)

Haplotype ID	*TGos*	*TGor*	*TTan*	*HV-I*	*HV-II*	*Other polymorphisms*	*HG*
T35			2	183C 189 223 278 294 390	073 143 146 152 189 195 263 315.1C 517		L2a
T36			1	188 189 223 278 294 311 320 390	073 146 152 263 309.1 315.1C		L2a
T37		3		189 223 278 294 390	073 143 146 152 195 263 309.1 315.1C		L2a
T38		2		189 223 278 294 390	073 143 146 152 195 263 315.1C		L2a
T39			5	189 192 223 278 294 309 390 519	073 146 152 195 263 315.1C		L2a1
T40			2	223 278 294 309 390 519	073 143 146 152 195 263 315.1 523–524delAC		L2a1
T41			1	223 278 294 309 390 519	073 143 146 152 189 195 203 204 263 309.1C 315.1C		L2a1
T42	1			223 278 294 309 390 519	073 146 152 195 263 315.1C 398		L2a1
T43			1	223 355	073 150 152 235 263 309.1C 315.1C 494		L3
T44		2		124 223 278 311 362 519	073 152 263 315.1 523–524delAC		L3b
T45			2	124 223 278 362 519	073 152 200 263 309.1C 315.1C 523–524delAC		L3b
T46			2	124 166 223 309	073 152 263 315.1C 523–524delAC		L3d
T47			1	209 223 261G 292 311 519	073 152 189 194 200 263 315.1C		L3f1
T48			1	209 223 292 311 519	073 189 200 214 263 309.1C 315.1C		L3f1
T49			1	129 174 192 218 223 256 311 362	073 151 152 189C 195 263 294 315.1C 523–524delAC		L3h
T50		1		093G 223 287A 293T 301 311 355 362 399	073 146 150 152 192 200 244 263 315.1C 513		L4g
T51		7		129 183C 189 223 249 311 519	073 152 195 263 (315.1)+2C 489 513		M1a2
T52			1	129 182C 183C 189 223 249 311 362 399 519	073 195 263 315.1 489		M1b2
T53	4			129 185 189 223 249 311 519	073 195 198 263 309.1 315.1 466 489		M1b1
Total	21	38	31				

Variant positions from the rCRS are shown at least for the minimum segment between nucleotide position 16017 and 16399 in HVS-I and 064 and 340 for HVS-II. Number of variants in the HVS-I region are referred with respect to the rCRS minus 16000. Substitutions are transitions unless the base change or a deletion is explicitly indicated, while insertions in poly-C tract are indicated by '.1'. An heteroplasmy for position 16192 is indicated by 'C>T'.

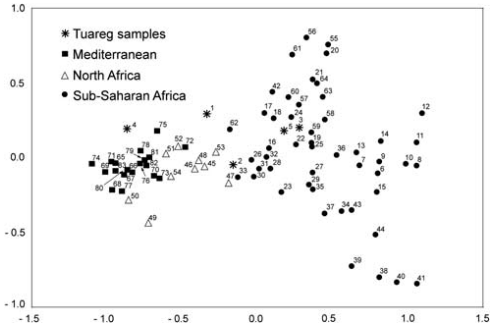


Figure 2 MDS plot of F_{ST} genetic distances calculated from HVS-I mtDNA sequences. For numbers see Supplementary Material SM1.

search for haplotype matches in an extensive database of 7211 individuals from all over Africa (Table 2). The most ancient lineages L0a1a and L1c, characteristic of east/southeast Africa[13] and the Pygmies,[39] respectively, were each observed in only one individual. The highly frequent African haplogroup L2, and specifically its dominant clade L2a, is also dominant in Tuareg – it is probable that some branches of L2a were involved in the Bantu expansion towards the African south[13,40] and many matches are observed for these haplotypes all over the continent. Curiously, the two L2a lineages having substitutions at positions 16 192 and 16 193, respectively, have no match in Africa. As far as the L3 macrohaplogroup is concerned, the two L3b haplotypes observed in the Tuareg are widespread throughout the continent, but one of the L3f1 haplotypes (T47 in Table 1) has no matches. Both are included in the L3f1 sub-haplogroup, which is quite frequent and widespread, and which very probably originated in East Africa. No L3f3, a typical marker of the Chadic migration,[41] has been observed in the Tuareg.

In summary, the matches between Tuareg sub-Saharan haplotypes and the diverse African regions were, after correcting for the size for each region, 5.6% with Africa-Central; 4.3% with Africa-East; 3.4% with Africa-North; 1.1% with Africa-South; 4.3% with Africa-southeast; 4.5% with Africa-southwest; 12.7% with Africa-West; and 13.4% with Africa-westcentral. The West Africa or West-Central African lineages thus are clearly dominant in the extant Tuareg.

The influence of East Africa in the Tuareg can be investigated more directly through haplogroup M1.[16] As concerns the finer classification of Tuareg M1 haplotypes, two of them (5 sequences out of 12) belong to M1b, which has a clear Mediterranean distribution, pointing to North Africa as its most probable gateway to the Tuareg. This finding is inconsistent with the absence of U6, which is believed to have entered Africa together with M1 in a back migration from western Eurasia around 45 000 years ago. The time estimate for M1b, based on the coding region, is 23 400 ± 5600 years,[16] placing its origin in the Early Upper Palaeolithic. More promising in ascertaining Eastern African origin is another haplotype observed in seven Tuareg individuals from Burkina Faso belonging to haplogroup M1a, which, though being considered dominant in East Africa[42] also spread to the Mediterranean, and which has a total age of 28 800 ± 4900 years.[16] We performed the complete sequencing of three individuals, which despite not displaying any difference at HVS-I and HVS-II, might present some substitutions in the coding region, allowing for a better estimate of a TMRCA. These three samples, however, did not bear any difference even when sequencing the complete genome. Nonetheless, when taken together with the other M1a2a individuals (Figure 3) reported in Olivieri *et al*[16] (sample 1 in Figure 3, accession number EF060335; sample 2, accession number EF060336), González *et al*[43] (sample 3; accession number DQ779927) and Maca-Meyer *et al*[44] (sample 4; accession number AF381984) allowed an age estimation for this sub-haplogroup at 8000 ± 2400-years old based on diversity in the coding region. We checked the TMRCA using Soares *et al*[26] mutations rates for the entire molecule and for the synonymous substitutions, obtaining, respectively, the following concordant dates: 10 400 ± 2300 and 10 200 ± 3400. Notice, however, that all the other four M1a2a complete sequences were observed in the Mediterranean region and in Table 2 the HVS-I motif observed in Tuareg has 10 perfect matches in the Africa-North data set and one in Africa-westcentral.

Y chromosome pool in Tuareg

From the 20 branches of the Y chromosome tree, which could be discriminated by the analyses performed, only 7 were observed in our Tuareg population sample (Supplementary Material SM7). Again, from this perspective of Y chromosome diversity, TTan is closer to sub-Saharan populations than the other two Tuareg populations, presenting 5.6% of the old AB lineages and 44.4% of E1b1a, whereas TGor and TGos have, respectively, 16.7 and 9.1% of E1b1a. Curiously, TTan also presents the highest frequency (33.3%) of West Eurasian R1b lineages whereas TGor presents only 5.6% of lineage K* (xO,P), and TGos presents none. There were no instances of the Eurasian J haplogroup in the Tuareg, which is otherwise frequent in North Africa (an average of 20%; see Arredi *et al*[45]), and attains the highest frequency in the Middle East (around 50%; see Semino *et al*)[46].

The dominant haplogroup in TGor (77.8%) and TGos (81.8%) is E1b1b1b, which has a much lower frequency in TTan (11.1%). This haplogroup reaches a mean frequency of 42% in North Africa, decreasing in frequency from 76% in Morocco to ~10% in Egypt.[45] Arredi *et al*[45] dated this haplogroup in North Africa from 2800 to 9800 YBP, associating its expansion with the Neolithic demic diffusion of Afro-Asiatic-speaking pastoralists from the Middle East.

The low level of diversity attained in the Tuareg populations (see Supplementary Material SM8) is consistent with a model of population constancy, although it can also be due in part to the ascertainment bias in the selection of a few Y-SNPs. Haplotype diversities and mean number of pairwise differences were very low in TGor and TGos, being among the lowest values observed in many populations, but TTan showed much higher levels of diversity.

MDS of F_{ST} distances based on available Y-SNP West Eurasian and African population data sets shows, as in the case of mtDNA, separation of the West Eurasian-North African and sub-Saharan populations (Figure 4). A certain separation between the Iberian and Near Eastern groups can be explained by the absence of samples from the Central Mediterranean for the Y-NRY data set. However, though the Tuareg groups from the Niger bend (TGor and TGos) belong clearly on the West Eurasian side, the Tuareg from central Niger lean towards sub-Saharan variability.

DISCUSSION

The Tuareg have a nomadic lifestyle and according to some demographic reports they show reduced fertility in comparison with their neighbours.[47–49] The data observed here for mtDNA and Y-SNP diversities are concordant with those independent reports, especially for the Tuareg living within the bend of the Niger.

Table 2 HVS-I haplotype match for the haplotypes observed in Tuaregs samples and an extensive African database (composed of the following geographical regions: Central (AF-C), East (AF-E), North (AF-N), South (AF-S), Southeast (AF-SE), Southwest (AF-SW), West (AF-W) and West-Central (AF-WC)). Number of variants in the HVS-I region are referred with respect to the rCRS minus 16000

Haplotypes	*Tuareg*	*AF-C*	*AF-E*	*AF-N*	*AF-S*	*AF-SE*	*AF-SW*	*AF-W*	*AF-WC*	*Totals*	*Haplogroup*
rCRS	14	0	4	193	2	0	1	7	5	226	H?
093 126 145 187 189 223 264 270 278 293 311	1	0	0	0	0	0	0	1	0	2	L1b
093 189 192 223 278 294 311	3	0	0	0	0	0	0	0	0	3	L2a
093 189 193 223 278 294 311	3	0	0	0	0	0	0	0	0	3	L2a
093 260 343	1	0	0	0	0	0	0	0	0	1	U3a
093G 223 287A 293T 301 311 355 362	1	2	0	0	0		0	0	9	12	L4g
124 166 223 309	2	0	0	0	0	0	0	0	0	2	L3d
124 223 278 311 362	2	11	3	0	0	3	1	2	14	36	L3b
124 223 278 362	2	5	3	4	0	1	2	67	34	118	L3b
126 172 187 189 223 264 270 278 293 311	1	2	0	1	0	0	0	0	3	7	L1b1
126 187 189 223 264 270 278 293 311	3	26	2	15	0	1	0	47	37	131	L1b1
129 148 168 172 187 188G 189 193 223 230 278 293 311 320	1	0	0	0	0	0	0	0	0	1	L0a1a
129 174 192 218 223 256 311 362	1	0	0	0	0	0	0	0	0	1	L3h
129 182C 1831 189 223 249 311 362	1	0	0	0	0	0	0	0	0	1	M1b2
129 183C 189	1	0	0	0	0	0	0	0	0	1	H1
129 183C 189 215 223 261 278 294 311 360	1	0	0	0	0	0	0	0	0	1	L1c
129 183C 189 223 249 311	7	0	0	10	0	0	0	0	1	18	M1a2
129 185 189 223 249 311	4	0	0	5	0	0	0	0	0	9	M1b1
169 223 239 278 294 309	1	0	0	0	0	0	0	0	0	1	L2a
172 213 223 261 278 318	1	0	0	0	0	0	0	0	0	1	L2*
183C 189 223 278 294	2	0	0	0	0	0	0	0	1	3	L2a
188 189 223 278 294 311 320	1	0	0	1	0	0	0	0	0	2	L2a
189	5	0	0	7	0	0	0	0	0	12	H
189 192 223 278 294 309	5	14	9	8	3	3	3	19	24	88	L2a1
189 223 278 294	5	0	4	1	0	0	0	3	0	13	L2a
189 234 298	1	0	0	0	0	0	0	0	0	1	V
209 223 261G 292 311	1	0	0	0	0	0	0	0	0	1	L3f1
209 223 292 311	1	1	12	4	0	0	0	4	18	40	L3f1
223 278 294 309	4	18	4	12	0	10	1	38	20	107	L2a1
223 355	1	0	0	0	0	0	0	3	1	5	L3
234 293 298	1	0	0	0	0	0	0	0	0	1	V
234 298	9	0	0	0	0	0	0	0	0	9	V
311	3	0	4	19	1	0	0	0	0	27	U3?/H
Partial totals	90	79	45	280	6	18	8	191	167	884	
Totals (sample sizes per continental region)	90	1404	862	1360	266	416	157	1454	1202	7211	

The overall West Eurasian mtDNA gene pool in the Tuareg population as a whole (H1, H3 and V) seems to favour a North African heritage.[50] The only exception is the absence of the otherwise rare U5b that might have rather come to Africa through the Near East, and then drifted to higher frequencies only in some isolated populations such as in the Egyptian oasis Siwa.[51] The absence of U6 can further be explained by genetic drift during the expansion of this haplogroup within North Africa.[51] Note that U6 was observed at low frequencies in several population groups from the Chad Basin, such as in the Nilo-Saharan Kanuri and the Afro-Asiatic Masa.[35]

Relationships with the peoples of Eastern Sudan (the Beja) as pointed to by the study of classical genetic markers[2] cannot yet be disregarded here as there is still no mtDNA of the Beja people available for study. However, according to historical reports, the origin of the Beja is more likely to be traceable to the Arabian Peninsula[52] and the West Eurasian mtDNA lineages seen in the Tuareg have a rather Iberian affiliation in the post-LGM, and probably expanded to North Africa first.[30,31] The weak Eastern African influence in Tuareg is further supported by the M1 haplotypes belonging to the lineages characteristic of the later Mediterranean expansion (M1b and M1a2a) and the presence of very few matches for sub-Saharan L haplotypes with East Africa. The main post-LGM Eurasian and M1a2a lineages found in the Tuareg favour North African origin with migration to its southern location in the Sahel between ~9000 and ~3000 years ago. The upper time limit is defined by the age of the M1a2a, (estimated here from the coding region diversity observed in the three Tuareg, two North and two south Mediterranean individuals at 8000 ± 2400), and by the upper 95% confidence interval for the Tuareg V lineages having polymorphism 16 234 (8800 years ago); the lower limit is defined by the age of the Tuareg V lineages having polymorphism 16 234 (3600 years ago).

The dates obtained from the genetic data coincide well with climatic changes in the Sahara, which resulted in repopulation during the first half of the Holocene when by ~10 000 YBP (the Holocene climatic optimum) humid conditions and greening were established. The climatic optimum lasted until ~6000 YBP, when the shift towards more permanent aridity occurred, culminating with the formation of the current Sahara desert. This desertification could have entrapped Tuareg populations coming from North Africa to the Sahel belt together with other pastoralists such as the Chadic speaking peoples[41]

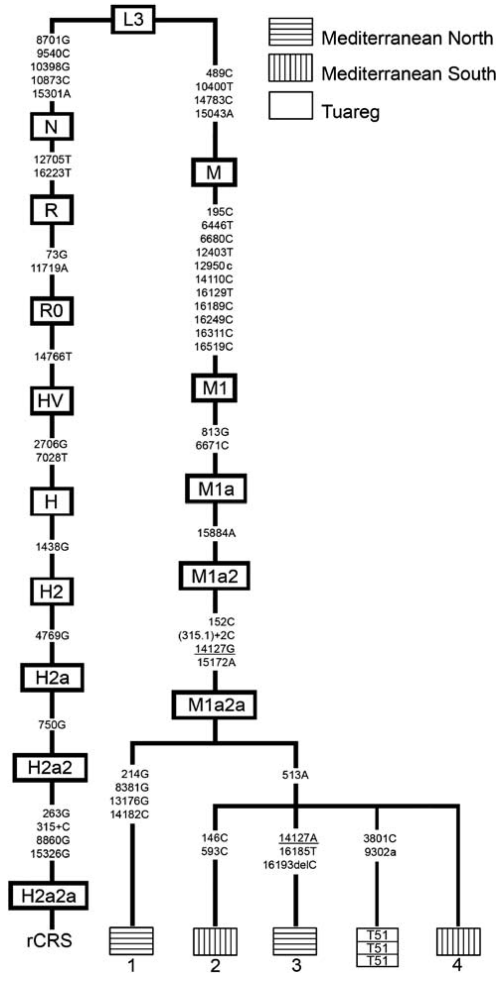


Figure 3 Phylogeny of the complete M1a2a mtDNA sequences, including the ones from Tuaregs and the published so far. Integers represent transitions, and an upper case suffix indicates a transition while a lower case suffix indicates a transversion. Deletions are indicated by a 'del' following the deleted nucleotide position. Underlined nucleotide positions appear more than once in the tree.

coming from East Africa and Fulani nomads[6] coming from West Africa. In fact, by performing complete mtDNA sequencing of the L3f3 lineage, specific for Chadic-speaking groups of the Chad Basin, Černý *et al*[41] estimated a local demographic expansion during the Holocene period at about 8000 ± 2500 YBP. No doubt all populations arriving to the Sahel were further enriched by various admixtures of many other sub-Saharan lineages, an effect even more pronounced in the Chadic groups who adopted a sedentary lifestyle soon after their arrival to the fertile Chad Basin than in the Tuareg who remain nomadic until present.

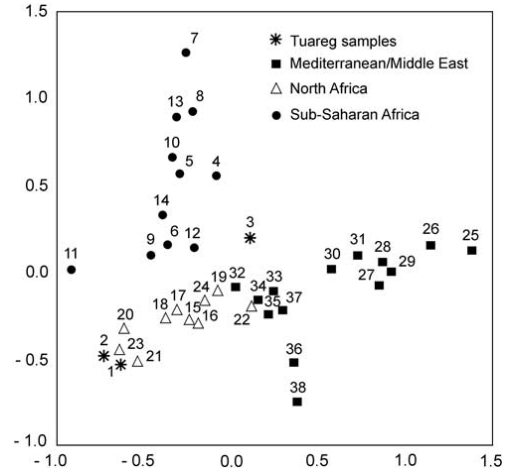


Figure 4 MDS plot of F_{ST} distances calculated from NRY haplogroup frequencies. Codes for numbers are as in Supplementary Material SM2.

It is curious that, at least for the Tuareg maternal gene pool, there are no mtDNA lineages connected with the Neolithic expansion from the Near East despite being present in considerable frequencies in other North African populations. For example, the conservation of the high frequency and remarkable internal variability of T1 haplotypes within the distant and relatively isolated Egyptian oasis of el-Hayez led to an estimation of local expansion at around 5138 ± 3633 YBP.[37] There are no indications yet of the ages of local expansions in the more central and western regions of North Africa, which could contribute further insights for its absence in the Tuareg population as a whole.

Interestingly, for the Y chromosome, the dominant haplogroup in North Africa as well as the Tuareg is E1b1b1b. This haplogroup was associated with Neolithic diffusion in North Africa, with an age estimation of 2800–9800 YBP,[45] but the lower resolution of the Y chromosome tree did not allow us to investigate this issue further. Nonetheless, disregarding whether they are in fact Neolithic, the ages for the mtDNA and Y chromosome lineages of North African origin observed in southern Tuareg are consistent with the same period, between 9000 and 3000 years ago.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The project was supported by the Grant Agency of the Czech Republic (Grant no. 206/08/1587), the Portuguese Fundação para a Ciência e a Tecnologia (PTDC/ANT/66275/2006 and Programa Operacional Ciência, Tecnologia e Inovação – Quadro Comunitario de Apoio III) (LP), Xunta de Galicia (Grupos Emerxentes; 2008/XA122), Fundación de Investigación Médica Mutua Madrileña (2006/CL370 and 2008/CL444) and Ministerio de Ciencia e Innovación (SAF2008-02971)(AS).

1 Lhote H: *Les Touaregs du Hoggar (Ahaggar).* Paris: Payot, 1955.
2 Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press, 1994.

3 Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Paabo S: mtDNA sequence diversity in Africa. *Am J Hum Genet* 1996; **59**: 437–444.
4 Watson E, Forster P, Richards M, Bandelt HJ: Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* 1997; **61**: 691–704.
5 Ottoni C, Martinez-Labarga C, Loogväli EL *et al*: First genetic insight into Libyan Tuaregs: a maternal perspective. *Ann Hum Genet* 2009; **73**: 438–448.
6 Černý V, Hájek M, Bromová M, Čmejla R, Diallo I, Brdička R: MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. *Hum Biol* 2006; **78**: 9–27.
7 Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M *et al*: New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS One* 2009; **4**: e5112.
8 Cerezo M, Cerny V, Carracedo A, Salas A: Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. *Electrophoresis* 2009; **30**: 3665–3673.
9 Brisighelli F, Capelli C, Álvarez-Iglesias V *et al*: The Etruscan timeline: a recent Anatolian connection. *Eur J Hum Genet* 2009; **17**: 693–696.
10 Torroni A, Bandelt HJ, Macaulay V *et al*: A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 2001; **69**: 844–852.
11 Torroni A, Rengo C, Guida V *et al*: Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* 2001; **69**: 1348–1356.
12 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**: 830–838.
13 Salas A, Richards M, De la Fé T *et al*: The making of the African mtDNA landscape. *Am J Hum Genet* 2002; **71**: 1082–1111.
14 Kivisild T, Reidla M, Metspalu E *et al*: Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am J Hum Genet* 2004; **75**: 752–770.
15 Behar DM, Villems R, Soodyall H *et al*: The dawn of human matrilineal diversity. *Am J Hum Genet* 2008; **82**: 1130–1140.
16 Olivieri A, Achilli A, Pala M *et al*: The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* 2006; **314**: 1767–1770.
17 Achilli A, Rengo C, Magri C *et al*: The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am J Hum Genet* 2004; **75**: 910–918.
18 van Oven M, Kayser M: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 2009; **30**: E386–E394.
19 Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N: Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 1999; **23**: 147.
20 Excoffier LGL, Schneider S: Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online* 2005; **1**: 47–50.
21 Rozas J, Rozas R: DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 1999; **15**: 174–175.
22 Bandelt HJ, Forster P, Sykes BC, Richards MB: Mitochondrial portraits of human populations using median networks. *Genetics* 1995; **141**: 743–753.
23 Mishmar D, Ruiz-Pesini E, Golik P *et al*: Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 2003; **100**: 171–176.
24 Forster P, Harding R, Torroni A, Bandelt HJ: Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* 1996; **59**: 935–945.
25 Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S: mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* 2000; **67**: 718–726.
26 Soares P, Ermini L, Thomson N *et al*: Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 2009; **84**: 740–759.
27 Finnila S, Lehtonen MS, Majamaa K: Phylogenetic network for European mtDNA. *Am J Hum Genet* 2001; **68**: 1475–1484.
28 Herrnstadt C, Elson JL, Fahy E *et al*: Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet* 2002; **70**: 1152–1171.
29 Pereira L, Richards M, Goios A *et al*: High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res* 2005; **15**: 19–24.
30 Cherni L, Fernandes V, Pereira JB *et al*: Post-last glacial maximum expansion from Iberia to North Africa revealed by fine characterization of mtDNA H haplogroup in Tunisia. *Am J Phys Anthropol* 2009; **139**: 253–260.
31 Ennafaa H, Cabrera VM, Abu-Amero KK *et al*: Mitochondrial DNA haplogroup H structure in North Africa. *BMC Genet* 2009; **10**: 8.
32 Pereira L, Cunha C, Amorim A: Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *Int J Legal Med* 2004; **118**: 132–136.
33 Behar DM, Garrigan D, Kaplan ME *et al*: Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. *Hum Genet* 2004; **114**: 354–365.
34 Plaza S, Calafell F, Helal A *et al*: Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* 2003; **67**: 312–328.
35 Černý V, Salas A, Hájek M, Žaloudková M, Brdička R: A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. *Ann Hum Genet* 2007; **71**: 433–452.
36 Loogväli EL, Roostalu U, Malyarchuk BA *et al*: Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol Biol Evol* 2004; **21**: 2012–2021.
37 Kujanova M, Pereira L, Fernandes V, Pereira JB, Cerny V: Near eastern neolithic genetic input in a small oasis of the Egyptian Western Desert. *Am J Phys Anthropol* 2009; **140**: 336–346.
38 Pereira L, Cunha C, Alves C, Amorim A: African female heritage in Iberia: a reassessment of mtDNA lineage distribution in present times. *Hum Biol* 2005; **77**: 213–229.
39 Quintana-Murci L, Quach H, Harmant C *et al*: Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci USA* 2008; **105**: 1596–1601.
40 Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A: Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* 2001; **65**: 439–458.
41 Černý V, Fernandes V, Costa MD, Hájek M, Mulligan CJ, Pereira L: Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. *BMC Evol Biol* 2009; **9**: 63.
42 Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS: Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* 1999; **23**: 437–441.
43 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.
44 Maca-Meyer N, González AM, Larruga JM, Flores C, Cabrera VM: Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* 2001; **2**: 13.
45 Arredi B, Poloni ES, Paracchini S *et al*: A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* 2004; **75**: 338–345.
46 Semino O, Magri C, Benuzzi G *et al*: Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* 2004; **74**: 1023–1034.
47 Gallais J: *Pasteurs et paysans du Gourma: la condition sahélienne*. Paris: C.N.R.S.,, 1975.
48 Hill A, Randall S: Différences géographiques et sociales dans la mortalité infantile et juvenile au Mali. *Population* 1984; **39**: 921–946.
49 Randall S, Winter M: *The Reluctant Spouse and the Illegitimate Slave: Marriage, Household Formation, and Demographic Behaviour Amongst Malian Tamesheq from the Niger Delta and the Gourma*. London: Overseas Development Institute, Agricultural Administration Unit,, 1986.
50 Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ: Harvesting the fruit of the human mtDNA tree. *Trends Genet* 2006; **22**: 339–345.
51 Coudray C, Olivieri A, Achilli A *et al*: The complex and diversified mitochondrial gene pool of Berber populations. *Ann Hum Genet* 2009; **73**: 196–214.
52 Paul A: *A History of the Beja Tribes of the Sudan*. London: F. Cass, 1971.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

PLoS ONE (accepted)

New insights into the Lake Chad Basin population structure revealed by high-throughput genotyping of mitochondrial DNA coding SNPs

María Cerezo[1&], Viktor Černý[2], Ángel Carracedo[1], Antonio Salas[1*&]

[1] Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, 15782, Universidade de Santiago de Compostela, CIBERER, Galicia, Spain
[2] Archaeogenetics Laboratory, Institute of Archaeology of the Academy of Sciences of the Czech Republic, Prague, The Czech Republic

*Corresponding author: Antonio Salas; Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, 15782, Universidade de Santiago de Compostela, Galicia, Spain; Email: antonio.salas@usc.es

[&] Both authors contributed equally to this study

Keywords: mtDNA, haplotype, haplogroup, SNP, Chad Basin, Atlantic slave trade, Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry

Abstract

Background: Located in the Sudan belt, the Chad Basin forms a remarkable ecosystem, where several unique agricultural and pastoral techniques have been developed. Both from an archaeological and a genetic point of view, this region has been interpreted to be the center of a bidirectional corridor connecting West and East Africa, as well as a meeting point for populations coming from North Africa through the Saharan desert.

Methodology/Principal Findings: Samples from twelve ethnic groups from the Chad Basin (*n*= 542) have been high-throughput genotyped for 230 coding region mitochondrial DNA (mtDNA) Single Nucleotide Polymorphisms (mtSNPs) using Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI-TOF) mass spectrometry. This set of mtSNPs allowed for much better phylogenetic resolution than previous studies of this geographic region, enabling new insights into its population history. Notable haplogroup (hg) heterogeneity has been observed in the Chad Basin mirroring the different demographic histories of these ethnic groups. As estimated using a Bayesian framework, nomadic populations showed negative growth which was not always correlated to their estimated effective population sizes. Nomads also showed lower diversity values than sedentary groups.

Conclusions/Significance: Compared to sedentary population, nomads showed signals of stronger genetic drift occurring in their ancestral populations. These populations, however, retained more haplotype diversity in their hypervariable segments I (HVS-I), but not their mtSNPs, suggesting a more ancestral ethnogenesis. Whereas the nomadic population showed a higher Mediterranean influence signaled mainly by sub-lineages of M1, R0, U6, and U5, the other populations showed a more consistent sub-Saharan pattern. Although lifestyle may have an influence on diversity patterns and hg composition, analysis of molecular variance has not identified these differences. The present study indicates that analysis of mtSNPs at high resolution could be a fast and extensive approach for screening variation in population studies where labor-intensive techniques such as entire genome sequencing remain unfeasible.

Introduction

The African Sahel together with a more southerly localized zone of savannah forms a clearly distinguishable biome. Also known as the Sudan or Macro-Sudan belt, this region displays some common linguistic features across current linguistic families [1]. The Sudan belt lacks higher mountains or other geographic barriers to migration and in genetics has been interpreted as a bidirectional corridor of human migrations [2,3,4]. From an ecological point of view this zone contains both high grasses in the north, and more or less dispersed shrubs and trees in the south, and comprises the natural surroundings known to humans in Africa from their early beginnings. The Sudan belt also experiences annual cycles of wet and dry seasons allowing for the coexistence of two populations with different lifestyles: nomadic pastoralists and sedentary farmers.

Approximately in the middle of the Sudan belt is the Lake Chad Basin, with Lake Chad in its imaginary centre. The Lake Chad Basin forms a remarkable ecosystem, where several unique agricultural and pastoral techniques have been developed [5]. Due to Pleistocene climatic oscillations Lake Chad has often changed in size and shape. For example, in the early Holocene, Lake Mega-Chad was formed covering a maximum surface area of 350,000 km^2. Such a giant lake, the largest in Africa at the time, with a wealth of food resources undoubtedly attracted long-term human settlements. Lake Chad reached its current size of about 20,000 km^2 approximately 3,000 years ago, around the time that the Sahara grew to its present size and became nearly impenetrable for humans [6]. Historically, inhabitants of the Chad Basin constantly migrated around Lake Chad in synchrony with the receding shorelines of the lake. It seems probable that in its present form Lake Chad acted as a final destination for two population movements in the Sudan belt – one from West represented by the Fulani and the other from East represented by the Arabs.

Variation in mitochondrial DNA (mtDNA) has demonstrated to be useful for the interpretation of historical and contemporary demographic events around the world, and in particular for reconstructing the evolution and origin of human populations. Most population studies carried out in Africa have been based on

analysis of the control region of mtDNA, sometimes complemented with analysis of hg diagnostic coding region sites [7,8,9,10,11,12,13]. Although a number of complete African mtDNA genomes have been obtained and deposited in GenBank or other databases, most of these studies were focused on a phylogenetic rather than a demographic perspective [14,15,16].

On the other hand, variation in mtDNA is commonly analyzed using standard sequencing procedures targeting the first and/or the second hypervariable regions (HVS-I/II) or the whole control region (including HVS-I/II). Analysis of coding region mtDNA SNPs using minisequencing is another common approach [17,18,19,20,21]. However, this technique is inadequate for genotyping large amounts of SNPs. Recently, Cerezo et al. [22] reported a novel MassARRAY SNP genotyping system for genotyping the large number of SNPs located in the coding region of mtDNA using MALDI-TOF. In this work, we present the first population application of this technique, genotyping 542 samples from 12 different ethnic groups of the Lake Chad Basin.

Material and Methods

Ethics Statement

Oral informed consent was required for the samples, and all of them were anonymized. The study was approved by the Ethical committee of the University of Santiago de Compostela. The study also conforms to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July).

Population samples

We analyzed 542 individuals from 12 different ethnic populations sampled around the Lake Chad Basin. Most of these samples (*n* = 441; 80%) were previously reported for the HVS-I segment and selected Restriction Fragment Length Polymorphisms (RPLPs) in Černý et al. [3]. Information about the ethnic adscription and the rationale for sampling collection is provided in [3] (see also Table 1). Briefly, we have analyzed the following population samples: Hide (*n* = 47), Kotoko (*n* = 62), Mafa (*n* = 57), Masa (*n* = 41), Buduma (*n* = 30), Chad Arabs (*n* = 27), Shuwa Arabs (*n* = 39), Fali (*n* = 40), Bongor Fulani (*n* = 50), Tcheboua Fulani (*n* = 40), Kanembu (*n* = 50) and Kanuri (*n* = 59). DNA

extract were then submitted to the laboratory of Santiago de Compostela where the genotyping was carried out.'

DNA extraction of new samples was carried out as described in Černý et al. [3]. All samples analyzed in the present study were previously subjected to whole genome amplification (WGA) using the Genomiphi v.2 kit (GE Healthcare Life Sciences; Uppsala, Sweden) according to the manufacturer's protocol. Only 1 µl of the original extracted DNA (at least >10 ng/µl) was used for WGA. The WGA product was subsequently diluted 1:16 in water and then directly used for MALDI-TOF MS genotyping.

MALDI-TOF MS mtSNP genotyping

All samples were genotyped for a set of 230 mtSNPs using the technology described in Cerezo et al. [22]. Sixty of the samples (11%) were already genotyped for the whole set of mtSNPs [22] (Table 2).

The two phylogenetic trees in Figure 1 and 2 of Cerezo et al. [22] indicate all of the mtSNPs genotyped in the present study, including diagnostic control region variants.

Assessment of the genotyping quality was carried out by replicating some samples from the Chad Basin in different runs plus six good quality DNA samples from the CEPH (Centre d'Etude du Polymorphisme Humain; http://www.cephb.fr/en/cephdb/), namely, NA10830, NA10831, NA10860, NA10861, NA11984, NA12147 (that were used as positive controls). Table S1 summarizes information on call rates per mtSNP. We have not detected genotyping inconsistencies among 7,436 replicated genotypes. For 77% of the mtSNPs the calling rate was above 90%; which can be considered quite acceptable if we take into account that all the samples were collected several years ago (some of the buccal swabs are now more than 10 years old). Some mtSNPs however virtually failed (e.g. 5128G; 5746A) or yielded poor results (650C; 9545G). More information is available in Table S1.

Given that the majority of the failed mtSNPs did not occur at the final branches, there were no problems to assign lineages to their maximum known level of phylogenetic resolution.

Standard sequencing analysis

All phylogenetic inconsistencies observed using MALDI-TOF MS were automatically sequenced using the protocol described in Álvarez-Iglesias et al. [20] and as indicated in Cerezo et al. [22]. New samples were also sequenced for the control region (see Table S2).

Nomenclature

African phylogeny and nomenclature is very complex and has been elaborated during the last decade based on the control region and complete genome sequencing efforts [8,9,13,14,15,23]. All of these phylogenetic efforts have been compiled in the Phylotree project (http://www.phylotree.org/; mtDNA tree Build 11; 7 Feb 2011) [24].

Statistical analysis

DnaSP v.5 [25] was used to compute diversity indices, including nucleotide and haplotype diversity and the average number of nucleotide differences. Arlequin 3.5.1.2 [26] was used to conduct analysis of molecular variance (AMOVA); the significance of the covariance components associated with different levels of genetic structure was tested using a non-parametric permutation procedure [26].

Principal Component Analysis was carried out on population hg frequencies and using R (http://www.r-project.org/).

Lamarc (Likelihood Analysis with Metropolis Algorithm using Random Coalescence) [27] was used to estimate: (i) θ, which for females it is expected to be equal to $2N_e\mu$ for neutral mutations in mtDNA (where μ is the neutral mutation rate per generation and N_e the effective population size of females), (ii) population growth as $g = -(\ln\theta_t / \ln\theta_{present\ day})/t$, where g is expressed as the relationship between θ at a time $t > 0$ in the past and θ at the present ($t = 0$), and (iii) migration rate, defined as $M = m/\mu$, where m is the immigration rate per generation), between the 12 ethnic groups used in the present study and using information from mtSNPs. Estimates were obtained for three independent replicates using a Bayesian framework. The jModelTest v.0.1.1. [28] software (with default heating and burn-in parameters) was used to obtain the base

frequencies, mutation parameters and the best mutation model (according to the Akaike information criteria), which (for our data) was the general time reversible (*GTR*) model. Estimates from iModelTest were obtained using three replicates using 10 initial chains (sampling interval of 20 and burn-in period of 1000) and two final chains (sampling interval of 20 and burn-in period of 1000). Transition:transversion rate was set to 30.

Some caveats should be considered in regards to the demographic estimates obtained. The mtDNA is in reality a single locus and therefore all the values should be taken with caution: "*For estimation of Theta and migration rate it is possible to get results with one region but they will improve markedly with more; doubling the number of regions nearly doubles the available information. Estimation of growth rate is very poor with less than 3 unlinked regions and particularly benefits from having more.*" (http://src.gnu-darwin.org/ports/biology/lamarc/work/lamarc-2.1.2b/doc/html/data_required.html). On the other hand, sample sizes for some of the groups are relatively low and therefore the impact on the different estimates is unpredictable. Last but not least, the neighboring source populations for the Chad Basin are not represented in the mathematical model (e.g. East, North, western Africa, etc); therefore, we are committed to assume a simplistic model were the provenance of the different lineages comes from one of the 12 ethnic groups considered.

For all of these computations and in order to account for missing data, failed SNPs were imputed according to known phylogeny. Given the fact that there exists a robust mtDNA worldwide phylogeny based on entire mtDNA genomes (>8,700), the phylogenetic-based approach for imputation seems more reliable than those based on e.g. metrics for linkage disequilibrium [29].

Results

Genetic diversity in the Chad Basin populations

Several diversity indices have been computed for the 12 ethnic groups analyzed in this study. These indices have been obtained both individually for HVS-I and mtSNPs, and in combination for HVS-I plus mtSNPs (Table 2). With few exceptions, haplotype diversity yielded slightly higher values for HVS-I than

for mtSNPs whereas nucleotide diversity was approximately twice as large for the mtSNPs as for HVS-I. Diversity values are very heterogeneous among the 12 population samples analyzed. The Hide sample shows high values of diversity independently of the mtDNA segment analyzed, whereas the opposite pattern was observed in the two Fulani samples. In general, the four nomadic populations included in this study have lower diversity values than the sedentary populations (Table 2).

Some minor differences may be related to language family, providing an explanation for why the Niger-Congo family has lower diversity values than the other two groups (Table 2), but the latter probably reflects the presence of the low diversity characterizing the two nomadic Fulani groups. No differences were observed in diversity values between populations located in the north (the Shuwa and Chad Arabs, Kanuri, Buduma, Kanembu, and Kotoko) *versus* those located in the south (the Bongor and Tcheboua Fulani, Hide, Mafa, and Masa; see map in Figure 1).

The diversity values obtained for the combined HVS-I plus mtSNPs are approximately an average of the values obtained for the two segments individually for nucleotide diversity and the average number of nucleotide differences, but, as expected, are slightly higher for haplotype diversity in most of the groups.

Phylogeography of the populations in the Chad Basin

The graphs in Figure 1 show the distribution of main African hgs in the Chad Basin. This broad hg classification clearly indicates substantial heterogeneity in the region. For instance, the Kotoko, Masa, Kanuri, and Mafa have frequencies of L3 haplotypes above 60%, in contrast with frequencies of only 30% for the Kanembu, Buduma and Chad Arabs. The Shuwa Arabs and both Fulani populations (Tcheboua and Bongor) do not have L0 haplotypes, which are approximately 11% in the Kotoko and 13% in the Hide. L2 is above 40% in the Buduma and Kanembu, but only 20% or less in the Masa, Kanuri, Fali, and both of the Fulani populations. Haplogroup M1 is present only in both Arab populations and the Buduma. Percentages of non-Sub-Saharan lineages vary also among ethnic groups (included in the category “Others” in Figure 1).

Figure 2 shows a medium resolution phylogeny indicating hg frequencies in the 12 groups genotyped. The most common sub-lineages in all the Chad Basin populations are L2a, L3b, L3f, and L3e.

Haplogroup frequencies measured to the maximum obtainable resolution for the mtDNAs genotyped in this study are shown in Table S3. The data indicate that the sub-hgs, L3b1a and L3e5, are the most common lineages in the Chad Basin (both accounting for >17% of the total sample). These two lineages are present in nearly all of the populations included in this study with the exception of the Chad Arabs (both L3b1a and L3e5) and the Buduma (L3e5). The unusual high frequency of L3e5 in the Chad Basin could be explained by a local expansion; however, because nearly all of the population samples analyzed carry L3e5 mtDNAs, it is more likely that this event occurred before the ethnogenesis of the region. At this level of resolution, it is remarkable that some sub-lineages are frequent in some populations but nearly absent in the rest. For instance, L1b1a appears in the two Fulani groups with frequencies greater than 18%, but with significantly lower frequencies in the other groups (below 6% in the Kanuri, and less than 3% in the other populations). L3b1b is also observed with high frequency in the Fulani Tcheboua (15%), but only appears with low frequency in the Fulani Bongor (4%) and is absent in the rest of the populations. Chad Arabs account for all the R0a mtDNAs in the Chad Basin (19%), in agreement with the high frequency of R0a reported in the Arabian Peninsula [30] and especially in its Southern tip and Socotra island [31]. The whole genome of two of these samples was recently reported [32] and classified within the widespread subclade R0a2f characterized by a substitution at position 8251. No mtDNAs of Eurasian ancestry have been observed in samples from the Fali, Kanembu, Mafa, and Masa. The typical North African lineages (M1 and U6; 2% in the total sample) are mainly observed in the two Arab groups. Most of the M1 lineages likely come from the Mediterranean instead of East Africa. For instance, four Shuwa Arabs belong to M1a1 and another within the sub-branch M1a1a (as indicated by a transition at position 14182 and a reversion at position 16249); the distribution of M1a1 is mainly Mediterranean. The Chad Arab sample #AC92 and the two Buduma #Bu87 and #Bu89 belong to the M1a3 branch, which has a predominant Mediterranean distribution. Finally, two other Buduma samples belong to M1a3, also mainly of

a Mediterranean distribution [23]. The three mtDNAs belonging to U6 were found in one Shuwa Arab, one Kanuri and one Mafa, with the one in the Kanuri belonging to U6a5, again a Mediterranean branch. More interestingly, the U6b lineage found in the Mafa is of the so called “Canarian Branch”, indicating that perhaps the Chad Basin could participate in the demographic wave that originally moved U6 hg towards the Canary Island from East Africa.

Analysis of molecular variance in the Chad Basin

Analyses of molecular variance (AMOVA) were carried out on the 12 Chad Basin populations analyzed in this study using the following grouping schemes: all of the populations individually, populations grouped by language family, and populations grouped according to their locations in the North or the South of the Chad Basin (Table 3). Most of the genetic variation (~96%) was found to occur within populations, whereas variation between populations accounted for only 4%. These values were virtually the same independently of the grouping scheme. Genetic variation among these groups is therefore below the inter-population differentiation reported to exist on the African continent (~12%; see [3]). The level of molecular resolution does not seem to be an influencing factor in the apportioning of genetic variance in the Chad Basin (Table 3), although mtSNPs do seem to contribute a subtle increment to the genetic variation.

Principal Component Analysis of the Chad Basin populations

PCA was carried out based on hg frequencies to the maximum level of resolution. The three first components account for a total of ~40% of the variation, and it shows notable divergence between the different ethnic groups from the Chad Basin. Thus, the first principal component (PC1), which accounts for 15% of the variation, locates Mafa and Kanuri in one side of the plot, and the two Fulani populations in the opposite pole. PC2 (13%) shows also the Mafa in one side of the plot and the Kanembu in the other extreme. PC3 (12%) displays again Mafa in one pole opposed to the Kotoko. There are not unique features in the Mafa that makes this population different to the other ethnic groups, but an accumulative effect of several differences in hg frequencies (some of them are more pronounced than others e.g. high frequency of hgs L2b2, L3d1d, L3e2).

Apart from the two Fulani samples, the two Arab ones are proximal in the plot indicating a close maternal phylogenetic relationship. The most distinctive features of the two Arab populations compared to the other Chad populations are the presence of non sub-Saharan lineages, such as R0a or M1 hgs. In agreement with the analysis of the different genetic diversity metrics, it is interesting to note that the nomadic populations are more tightly grouped in the scatter plot than the sedentary ones (Figure 3), mirroring their more reduced genetic diversity.

Population growth, effective population size and migration rates

We further estimated the population mutation parameter, θ, which in conjunction with the average mtDNA mutation rate was used to infer the effective population sizes of the different Chad Basin populations assuming a neutral model of molecular evolution (Table 4). For an average entire genome mutation rate of 1.655×10^{-8} base substitution per nucleotide per year [33], female effective population size ranges from 359,200 in Masa to 5,423,000 in Buduma (Table 4). Curiously, three out of four populations (Chad Arabs, and Bongor and Tcheboua Fulani) showed negative growth rates, while others have positive values, such as the Buduma and the nomadic Shuwa Arabs (Table 4).

There are not unique features that would explain the observed migration rates values (although not all should be considered as reliable estimates; see M&M; and Table 4). Thus, for instance, the highest migration rate was obtained for Kanuri into Masa (Table 4); from a phylogeographic point of view these two populations share the highest amount of sub-clades (Table S3), and they are displayed together for the PC3 (and to a minor extent for the PC2) in the PCA. Bongor Fulani have the highest frequencies for hg L1b1a and L3b1a, which could explain their influence in the Fali.

Discussion

The values of the diversity indices computed for the HVS-I and the mtSNPs show clear-cut differences, mirroring the fact that haplotype diversity is enriched in the HVS-I segment by the presence of rare (or private) variants, whereas the agglomeration of identical sequences into different hgs (possibly suffering from bias in mtSNPs selection) enriches the nucleotide diversity.

Therefore, values computed using these different mtDNA segments summarize different aspects of the molecular diversity in populations. In fact, diversity values of HVS-I and mtSNPs for the 12 different ethnic groups analyzed moderately correlate for the haplotype diversity (*h*, r^2=0.90), but very poorly correlate for the nucleotide diversity (π, r^2=0.42). Thus, one can speculate different demographic scenarios for each population according to their differential diversity values. For instance, nomadic or semi-nomadic populations tend to experience loss of diversity by genetic drift (assuming moderate admixture with those populations they meet sporadically), reducing both haplotype and nucleotide diversities. Provided that the gene flow is negligible for a given effective population size, the HVS-I region of more ancestral nomadic populations would be expected to retain more haplotype diversity (due to the presence of more rare variants) than younger nomadic groups. Thus, although the (semi-)nomadic populations (the Chad Arabs, Shuwa Arabs, Bongor Fulani, and Tcheboua Fulani) all have very low nucleotide and haplotype diversity values compared with sedentary populations, the two Arab populations retained higher values of haplotype diversity in the HVS-I segment than the Fulani (this signal is not as clear for nucleotide diversity in the HVS-I segment, possibly due to the low inter-population differences observed for all of the population groups analyzed). This hypothesis is compatible with the fact that nomadic populations as a whole have lower diversity than sedentary groups (Table 2). Finally, values for the average number of pairwise differences and nucleotide diversity are highly correlated, as expected given that both indices are based on the same principles.

Demographic inferences carried out only using summarizing indices (such as nucleotide and haplotype diversities) have to be considered with caution because in reality each human population defined only by ethno-linguistic criteria is composed of an amalgamation of genetic lineages of different ages and origins, and therefore, none has a simple past.

Haplogroup patterns vary substantially among the different ethnic groups studied. In some, hg composition seems to correlate well with historical documentation and their known demographic past. Thus, the presence of R0a only in Chad Arabs is expected given the high frequency of this hg in Southern Arabia [32]. In addition, the M1 sub-lineages observed in our samples have a

mainly Mediterranean distribution, and are exclusively found in the two Arab populations and the Buduma (also located in the northern Chad Basin). The spread of this hg to the African Sahel (and possibly further into the Chad Basin) might have been mediated by the Tuareg nomads [34]. Also, the presence of L1c sub-lineages in the Hide (Cameroon; Chad Basin) compared with the rest of the populations indicate narrow contact of this population with Central African populations (including Pygmy populations), where this lineage is found with high frequency [7,8,16].

The few Eurasian profiles observed in the Chad Basin did not cluster in any particular ethnic group. Their control region segments are not informative from a phylogeographic point of view, and these sequences are broadly distributed around Eurasia. The only exception is the U5b1 HVS-I profile T16189C C16192T C16270T C16320T that is detected mainly in Africa [35] and is a perfect match with a sample from Spain (https://www.policia.es/cgpc/index.htm).

With the exception of a few population studies based on complete genomes [36,37] or coding region segments [38,39,40], most of the genotyping studies carried out to date were based on control region sequences [8] and/or mtSNPs at a low to moderate level of hg definition [18,19,20,21,41,42]. Some other studies [15,43,44,45] focused on phylogenetic issues by genotyping selected branches of the mtDNA tree, but did not consider the population as a whole. The mtSNPs genotyped in this study were designed to identify mtDNAs of African ancestry to the maximum level of molecular resolution provided by known phylogeny based on complete genome sequences. In theory, the 230 mtSNPs should be able to discriminate among 147 different terminal branches of the Sub-Saharan phylogeny (L-hgs), along with dozens of intermediate hgs, and other African non-L branches (such as sub-hgs of U6 or M1). Moreover, the mtSNPs allow for a more rigorous classification of mtDNAs into hgs due to the (average) low mutation rate characterizing these SNPs compared with the mutation rate in the control region [33].

Analyses of mtSNPs in combination with sequencing information (control region) has provided new insights regarding population features of the Chad Basin populations [3], and open new perspectives for new pan-African

phylogenetic studies as well as for the reconstruction of the patterns of Trans-Atlantic slave trade into America:

a) Since the mtSNPs used in the present study were designed to detect major and minor branches of the phylogeny, while the control region variation analyzed previously [3] accounts for both (unbiased) common and rare variants, the evolutionary histories told by both sets of markers are different. Thus, as dissected in the present study, different styles of life (nomadic *versus* sedentary populations) can leave different signatures in the two sets of markers.
b) We have estimated for the first time different demographic parameters of the Chad Basin populations, including N_e, population growth, and migration rates. Nomadic populations show signals of negative growth (as also indirectly indicated by diversity metrics), which not always coincide with those that have higher N_e; in fact, both parameters are just moderately correlated ($r^2 = 0.59$). However, larger sample sizes would be required in order to yield solid figures for all these parameters.
c) Analysis of mtSNPs have allowed to reveal new phylogeographic features in the Chad populations, not discussed previously [3]. Given that this is the first study analyzing a pan-African mtSNP hg panel to a population level (with the only exception of a test sample analyzed previously [22]), it is still not possible to make inferences concerning the gene flow of neighboring source populations to the different Chad ethnic groups; however, the present study provides a high resolution hg map for future African studies.
d) It has been previously demonstrated [7,9,46] that only broad patterns of variability can be established in Africa with the current mtDNA data (basically HVS-I segments); therefore, tracking 'African-American' lineages to particular African regions might be fraught with problems due to the low level of genetic resolution. Analysis of mtSNPs, as undertaken in the Chad Basin, could help to achieve new insights into the patterns of Atlantic slave trade.
e) Analyzing mtSNPs to a high level of hg resolution in populations allows a better selection of mtDNAs for further entire genome sequencing or for

the design of multiplex mtSNP panels of interest in population, medical, and forensic genetics [20,21,47,48].

In conclusion, given that genotyping mtSNPs is straightforward compared with the intense effort demanded by sequencing complete genomes, the present study opens the door to more ambitious pan-African studies that would improve our knowledge on the mtDNA phylogeography in this continent.

Acknowledgements

We thank María Torres for his assistance with the genotyping and sample management. MALDI-TOF genotyping was carried out in the CEGEN (Centro Nacional de Genotipado) node from Santiago de Compostela (Spain).

References

1. Güldemann T (2008) The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa; Heine B, Nurse D, editors. Cambridge: Cambridge University Press. .
2. Bereir RE, Hassan HY, Salih NA, Underhill PA, Cavalli-Sforza LL, et al. (2007) Co-introgression of Y-chromosome haplogroups and the sickle cell gene across Africa's Sahel. Eur J Hum Genet 15: 1183-1185.
3. Černý V, Salas A, Hájek M, Žaloudková M, Brdička R (2007) A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71: 433-452.
4. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-1044.
5. Batello C, Marzot M, Touré AH, Kenmore PE (2004) The future is an ancient lake : traditional knowledge, biodiversity, and genetic resources for food and agriculture in Lake Chad Basin ecosystems. : Food and Agriculture Organization of the United Nations., and FAO Inter-Departmental Working Group on Biological Diversity for Food and Agriculture.

6. Kropelin S, Verschuren D, Lezine AM, Eggermont H, Cocquyt C, et al. (2008) Climate-driven ecosystem succession in the Sahara: the past 6000 years. Science 320: 765-768.
7. Salas A, Carracedo Á, Richards M, Macaulay V (2005) Charting the Ancestry of African Americans. Am J Hum Genet 77: 676-680.
8. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, et al. (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082-1111.
9. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, et al. (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454-465.
10. Richards M, Macaulay V, Hill C, Carracedo Á, Salas A (2004) The archaeogenetics of the dispersals of the Bantu-speaking peoples. In: Jones M, editor. Studies in honour of Colin Renfrew. Cambridge: McDonald Institute for Archaeological Research. pp. 1363-1349.
11. Beleza S, Gusmão L, Amorim A, Carracedo Á, Salas A (2005) The genetic legacy of western Bantu migrations. Hum Genet 117: 366-375.
12. Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, et al. (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet 115: 439-447.
13. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, et al. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752-770.
14. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22: 339-345.
15. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, et al. (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82: 1130-1140.
16. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, et al. (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596-1601.
17. Coble M (2004) The identification of single nucleotide polymorphisms in the entire mitochondrial genome to increase the forensic discrimination of

common HV1/HV2 types in the Caucasian population. Washington: The George Washington University. 206 p.

18. Álvarez-Iglesias V, Barros F, Carracedo Á, Salas A (2008) Minisequencing mitochondrial DNA pathogenic mutations. BMC Med Genet 9: 26.
19. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44-55.
20. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, et al. (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE 4: e5112.
21. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, et al. (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251-257.
22. Cerezo M, Černý V, Carracedo Á, Salas A (2009) Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. Electrophoresis 30: 3665-3673.
23. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, et al. (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314: 1767-1770.
24. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: E386-394.
25. Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.
26. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479-491.
27. Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768-770.
28. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25: 1253-1256.

29. Elson JL, Majamaa K, Howell N, Chinnery PF (2007) Associating mitochondrial DNA variation with complex traits. Am J Hum Genet 80: 378-382; author reply 382-373.
30. Černý V, Mulligan CJ, Rídl J, Žaloudková M, Edens CM, et al. (2008) Regional differences in the distribution of the sub-Saharan, West Eurasian, and South Asian mtDNA lineages in Yemen. Am J Phys Anthropol 136: 128-137.
31. Černý V, Pereira L, Kujanová M, Vašíková A, Hájek M, et al. (2009) Out of Arabia-the settlement of island Soqotra as revealed by mitochondrial and Y chromosome genetic diversity. Am J Phys Anthropol 138: 439-447.
32. Černý V, Mulligan CJ, Fernandes V, Silva NM, Alshamali F, et al. (2010) Internal diversification of mitochondrial haplogroup R0a reveals post-Last Glacial Maximum demographic expansions in South Arabia. Mol Biol Evol 28: 71–78.
33. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740-759.
34. Pereira L, Černý V, Cerezo M, Silva NM, Hájek M, et al. (2010) Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. Eur J Hum Genet 18: 915-923.
35. Rando JC, Pinto F, González AM, Hernández M, Larruga JM, et al. (1998) Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann Hum Genet 62 (Pt 6): 531-550.
36. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24: 757-768.
37. Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, et al. (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14: 1832-1850.
38. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA

coding-region sequences from the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152-1171.

39. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373-387.
40. Finnilä S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68: 1475-1484.
41. Coble MD, Vallone PM, Just RS, Diegoli TM, Smith BC, et al. (2006) Effective strategies for forensic analysis in the mitochondrial DNA coding region. Int J Legal Med 120: 27-32.
42. Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo Á, et al. (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27: 2541-2550.
43. Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, et al. (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3: e1764.
44. Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, et al. (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84: 814-821.
45. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, et al. (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1-8.
46. Salas A, Torroni A, Richards M, Quintana-Murci L, Hill C, et al. (2004) The phylogeography of mitochondrial DNA haplogroup L3g in Africa and the Atlantic slave trade. Am J Hum Genet 75: 524-526.
47. Salas A, Amigo J (2010) A reduced number of mtSNPs saturates mitochondrial DNA haplotype diversity of worldwide population groups. PLoS One 5: e10218.
48. Mosquera-Miguel A, Álvarez-Iglesias V, Lareu MV, Carracedo Á, Salas A (2009) Testing the performance of mtSNP minisequencing in forensic samples. Forensic Sci Int Genet 3: 261-264.

49. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.

Figure legends

Figure 1. Map of the Lake Chad Basin showing frequencies of the main African hgs in the different ethnic groups analyzed.

Figure 2. Phylogeny of African hgs at a medium level of phylogenetic resolution and (below branches) counts of these hgs for the different ethnic groups. Bottom of the figure: population labels have in brackets the sample sizes; numbers below branches indicate the hg relative frequencies in each population group and in the total sample size (row "Total"); therefore, each row sums to 1. The counts for the maximum level of resolution are provided in Table S2; the full phylogenetic tree for the SNPs considered in the present study is provided in Cerezo et al. [22]. All positions in the tree refer to the revised Cambridge Reference Sequence (rCRS; [49]); all positions are transitions unless a letter indicates a transversion. Underlined positions are parallel mutations within this tree, while "!" indicates a back mutation. A deletion is indicated as "del", while "+" indicates an insertion.

Figure 3. PC plot of ethnic relationships based on hg frequencies Percentage values in brackets refer to the amount of variation accounted by the first three principal components (PC1, PC2, and PC3). Codes for populations are as indicated in Table 1. Nomadic populations are plotted in blue while sedentary ones are plotted in red.

Table 1. Populations analyzed in the present study. Most of this information is available in Table 1 of Černý et al. [3] and it is replicated here for the sake of clarity.

Population	Code	*n*	Geographical Region	Language branch/Language Family	Lifestyle
Hide	Hi	47	Northern Cameroon	Chadic/AA	Sedentary
Kotoko	Ko	62	Northern Cameroon	Chadic/AA	Sedentary
Mafa	Mf	57	Northern Cameroon	Chadic/AA	Sedentary
Masa	Ms	41	Northern Cameroon	Chadic/AA	Sedentary
Buduma	Bu	30	South-eastern Niger	Chadic/AA	Sedentary
Chad Arabs	CA	27	Southwestern Chad	Semitic/AA	Nomadic
Shuwa Arabs	SA	39	Southwestern Chad	Semitic/AA	Semi-nomadic
Fali	Fa	40	Northern Cameroon	Adamawa-Ubangui/NC	Sedentary
Bongor Fulani	BF	50	Southwest Chad	Atlantic/NC	Nomadic
Tcheboua Fulani	TF	40	Northern Cameroon	Atlantic/NC	Nomadic
Kanembu	Kb	50	South- western Chad	Saharan/NS	Sedentary
Kanuri	Ka	59	North-eastern Nigeria	Saharan/NS	Sedentary

NOTE: AA = Afro-Asiatic; NC = Niger-Congo; NS = Nilo-Saharan

Table 2. mtDNA diversity in the Chad Basin.

		HVS-I						mtSNP						HVS-I plus mtSNP					
	n	*k*	*k/n*	*S*	*h*	*Π*	*M*	*k*	*k/n*	*S*	*h*	*π*	*M*	*k*	*k/n*	*S*	*h*	*π*	*M*
Ethnic group																			
Hide	47	38	0.81	51	0.991±0.006	0.025±0.002	8.60	25	0.53	70	0.963±0.013	0.051±0.003	11.61	38	0.81	121	0.991±0.006	0.036±0.002	20.21
Kotoko	62	33	0.53	51	0.961±0.014	0.020±0.002	6.91	29	0.47	72	0.947±0.016	0.051±0.003	11.61	40	0.65	123	0.977±0.010	0.033±0.002	18.52
Mafa	57	37	0.65	62	0.980±0.008	0.023±0.002	7.80	32	0.56	71	0.961±0.012	0.047±0.012	10.52	42	0.74	133	0.985±0.007	0.032±0.002	18.32
Masa	41	35	0.85	43	0.991±0.008	0.022±0.002	7.34	26	0.63	67	0.971±0.012	0.049±0.003	11.01	36	0.88	110	0.993±0.008	0.032±0.002	18.35
Buduma	30	22	0.73	43	0.968±0.021	0.023±0.002	7.74	18	0.60	57	0.954±0.021	0.041±0.003	9.34	22	0.73	100	0.968±0.021	0.030±0.002	17.09
Chad Arabs	27	20	0.74	36	0.963±0.023	0.020±0.002	6.78	19	0.70	54	0.969±0.018	0.046±0.003	10.52	22	0.81	90	0.980±0.017	0.030±0.001	17.30
Shuwa Arabs	39	29	0.74	44	0.980±0.011	0.018±0.001	6.04	23	0.59	54	0.968±0.012	0.042±0.002	9.57	31	0.79	98	0.987±0.009	0.027±0.001	15.61
Fali	40	23	0.58	44	0.962±0.014	0.022±0.002	7.43	22	0.55	55	0.953±0.017	0.050±0003	11.32	26	0.65	99	0.971±0.013	0.033±0.002	18.75
Bongor Fulani	50	27	0.54	35	0.937±0.023	0.020±0.001	6.87	24	0.48	50	0.937±0.023	0.045±0.002	10.19	32	0.64	85	0.954±0.021	0.030±0.001	17.05
Tcheboua Fulani	40	21	0.53	42	0.953±0.016	0.021±0.002	7.31	19	1.30	52	0.942±0.017	0.043±0.002	9.87	27	0.68	94	0.971±0.014	0.030±0.002	17.19
Kanembu	50	38	0.76	57	0.989±0.006	0.026±0.002	8.80	31	0.62	68	0.978±0.008	0.049±0.003	10.99	42	0.84	125	0.993±0.006	0.035±0.002	19.79
Kanuri	59	47	0.80	55	0.990±0.006	0.022±0.002	7.44	35	0.59	82	0.976±0.008	0.047±0.003	10.74	51	0.86	137	0.994±0.005	0.032±0.002	18.18
TOTAL	542	248	0.46	117	0.991±0.001	0.022±0.001	7.44	143	0.26	136	0.977±0.002	0.049±0.001	11.08	315	0.58	253	0.995±0.001	0.033±0.000	18.53
Lifestyle																			
Sedentary	386	195	0.51	107	0.990±0.001	0.022±0.001	7.57	112	0.29	128	0.975±0.002	0.050±0.001	11.22	237	0.61	235	0.994±0.001	0.033±0.001	18.79
Nomadic	156	83	0.53	73	0.979±0.005	0.021±0.001	7.01	59	0.38	91	0.969±0.0006	0.046±0.001	10.45	101	0.65	164	0.987±0.004	0.031±0.001	17.46
Language Family																			
Niger-Congo	130	60	0.46	61	0.964±0.007	0.021±0.001	7.28	45	0.35	80	0.956±0.009	0.047±0.001	10.56	75	0.58	141	0.980±0.006	0.031±0.001	17.85
Nilo-Saharan	109	80	0.73	74	0.993±0.002	0.024±0.001	8.25	55	0.50	96	0.978±0.005	0.050±0.002	11.25	88	0.81	170	0.995±0.002	0.034±0.001	19.49
Afro-Asiatic	303	151	0.50	105	0.990±0.001	0.021±0.001	7.23	101	0.33	119	0.977±0.003	0.049±0.001	10.98	192	0.63	217	0.995±0.001	0.032±0.001	18.21
Geographical region																			
North	267	143	0.54	98	0.990±0.002	0.021±0.001	7.26	96	0.36	117	0.978±0.003	0.049±0.001	11.00	1789	6.70	215	0.994±0.001	0.032±0.001	18.26
South	275	143	0.52	87	0.987±0.002	0.023±0.001	7.76	83	0.30	112	0.970±0.004	0.049±0.001	10.96	170	0.62	199	0.993±0.001	0.033±0.001	18.72

NOTE: *n* = sample size; *k* = number of different sequences; *S* = number of segregating sites; *h* = haplotype diversity; *π* = nucleotide diversity; *M* = average number of pairwise differences (mismatch observed mean). For all of the samples, the common segment of the HVS-I region analyzed ranges from position 16030 to 16370 (with the exception of samples #Fa108 and #Hi14 that present sequence ranges outside 16030-16370 and were therefore eliminated from the analysis; see Table S2).

Table 3. Apportioning of genetic variance considering different genomic regions (HVS-I, mtSNPs, and both in combination) and groups (populations, language families and geography).

	HVS-I		**mtSNPs**		**HVS-I+ mtSNPs**	
	Among populations[*1]	Within populations[*1]	Among populations[*1]	Within populations[*1]	Among populations[*1]	Within populations[*1]
All populations	3.24	96.76	3.83	96.17	3.59	96.41
Language	3.82[*2]	96.18	4.28[*2]	95.71	4.10[*2]	95.90
Geography	3.50[*2]	96.50	4.03[*2]	95.97	4.39[*2]	95.61

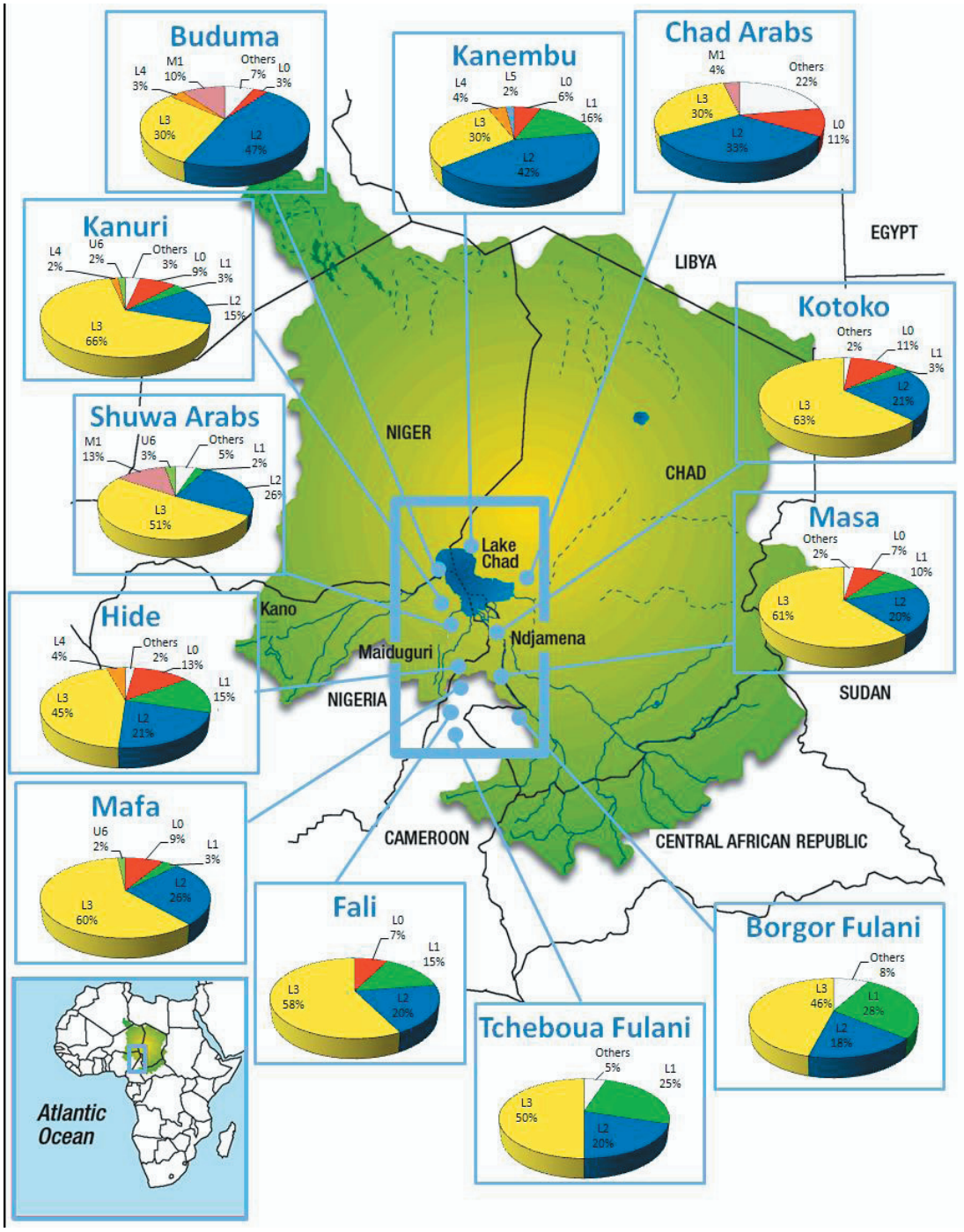
[*1] *P*-values are below 0.0000 using a significance test based on 20.000 permutations

[*2] Among groups + Among populations within groups

Table 4. Inter-population migration rates, population growth, and effective population sizes for the different ethnic groups from the Chad Basin. Migration rates: numbers indicate the gene flow from each population group (as indicated in the first column) into the other populations (as indicated in the first row); for instance, 88.5 would be the migration rate from Shuwa Arabs into Chad Arabs. Population codes are as indicated in Table 1. Stars indicate estimates that have to be taken with care due to limited sample sizes (as inferred by Lamarc).

	CA	SA	Bu	Fa	BF	TF	Hi	Kb	Ka	Ko	Ms	Mf
Migration rates												
CA	–	163.3	110.4	173.3	216.2	119.3*	67.0	166.9*	70.5	112.1	96.1	76.1
SA	88.5	–	41.9	152.0	231.9	83.6	69.1*	28.2	107.1	93.2	109.4	74.7
Bu	135.8	36.1	–	108.8	79.5	66.1	94.1	346.1*	53.1	104.0*	89.2	159.9*
Fa	59.9*	121.2	43.6	–	163.9	122.7	243.9	123.1	87.1	87.1	59.0	86.0
BF	189.7*	270.7	23.7	319.2	–	140.3	19.9	45.3	69.8	51.4	189.9	63.0
TF	60.2	61.9	36.6	162.8	154.0	–	117.9	50.3	95.4	76.7	153.0	50.6
Hi	74.1	143.5*	139.3	55.9	24.9	201.6	–	163.7*	71.0	198.5*	486.9*	152.6*
Kb	404.9*	81.0	575.2*	232.4	25.6	69.4	96.9*	–	109.1	76.7	82.6	112.7*
Ka	69.0	161.4	99.3	156.5	361.1	119.1	66.6	144.1	–	138.5	647.0	63.6
Ko	98.1	97.0	73.9	110.1	53.7	95.0	116.2*	113.7	100.1	–	202.9*	113.1
Ms	140.5	219.9*	91.5	327.5	91.0	174.9	62.9*	74.8	104.0	195.3	–	146.4
Mf	146.4	105.8	136.0	79.5	130.1	115.8	193.8*	144.2	27.5	103.1	90.2	–
Theta	0.017*	0.026*	0.181*	0.014*	0.012*	0.018*	0.025*	0.038*	0.066*	0.023*	0.012*	0.050*
N_e (females)	520,600	769,100	5,423,000	414,700	362,800	545,800	740,400	1,136,000	1,977,900	677,100	359,200	1,498,000
Growth	-138.66	252.15	311.13*	150.23*	-48.97*	-64.94	91.96*	41.77	64.25	20.38	131.07	147.53*

Figure 1


Buduma
L4 3%
M1 10%
Others 7%
L0 3%
L3 30%
L2 47%
Kanembu
L4 4%
L5 2%
L0 6%
L1 16%
L3 30%
L2 42%
Chad Arabs
M1 4%
Others 22%
L3 30%
L0 11%
L2 33%
Kanuri
L4 2%
U6 2%
Others 3%
L0 9%
L1 3%
L2 15%
L3 66%
Shuwa Arabs
M1 13%
U6 3%
Others 5%
L1 2%
L2 26%
L3 51%
Kotoko
Others 2%
L0 11%
L1 3%
L2 21%
L3 63%
Masa
Others 2%
L0 7%
L1 10%
L3 61%
L2 20%
Hide
L4 4%
Others 2%
L0 13%
L1 15%
L3 45%
L2 21%
Mafa
U6 2%
L0 9%
L1 3%
L2 26%
L3 60%
Fali
L0 7%
L1 15%
L3 58%
L2 20%
Tcheboua Fulani
Others 5%
L1 25%
L3 50%
L2 20%
Borgor Fulani
Others 8%
L3 46%
L1 28%
L2 18%
Atlantic Ocean
LIBYA
EGYPT
NIGER
CHAD
Lake Chad
Kano
Maiduguri
Ndjamena
NIGERIA
SUDAN
CAMEROON
CENTRAL AFRICAN REPUBLIC

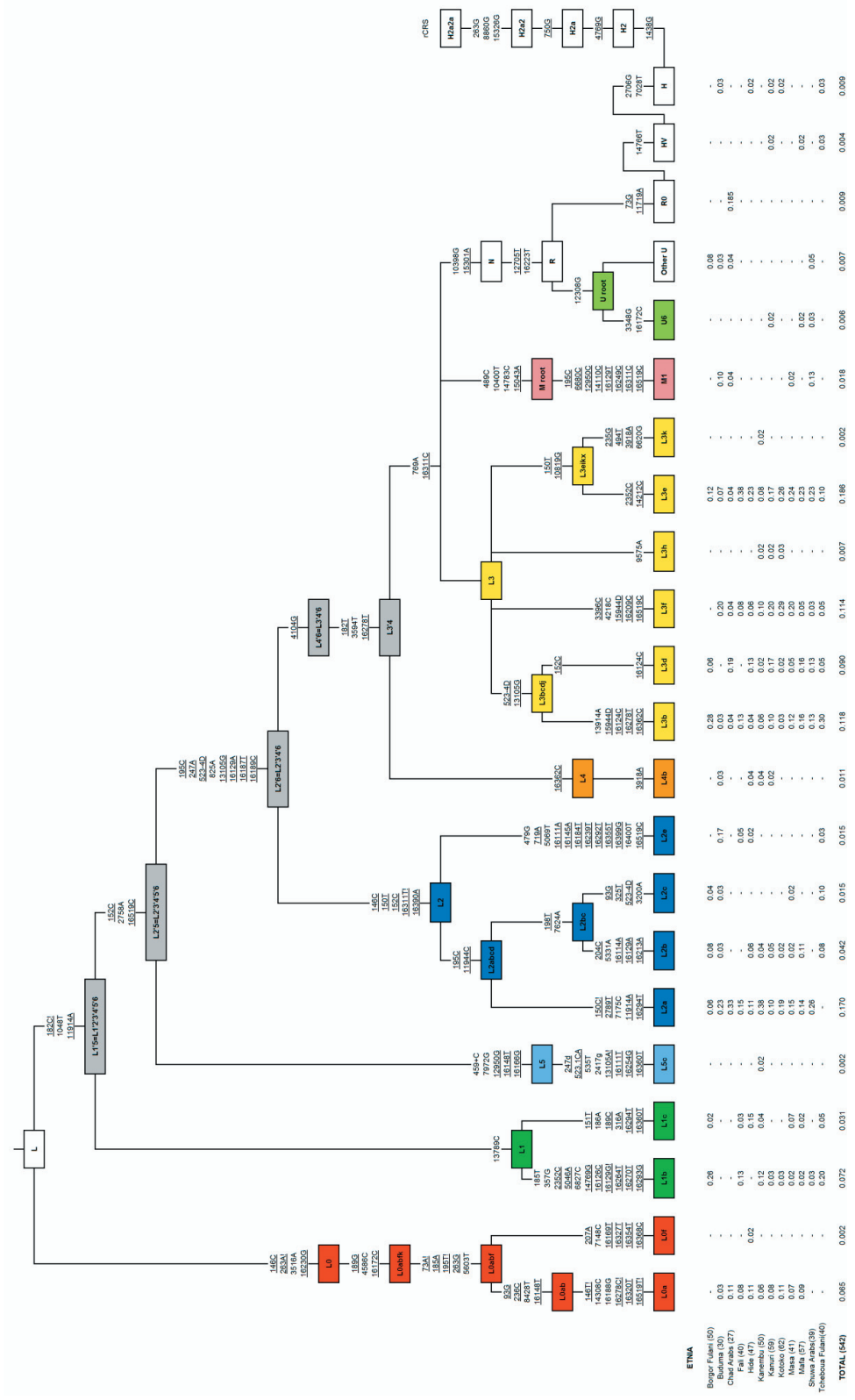
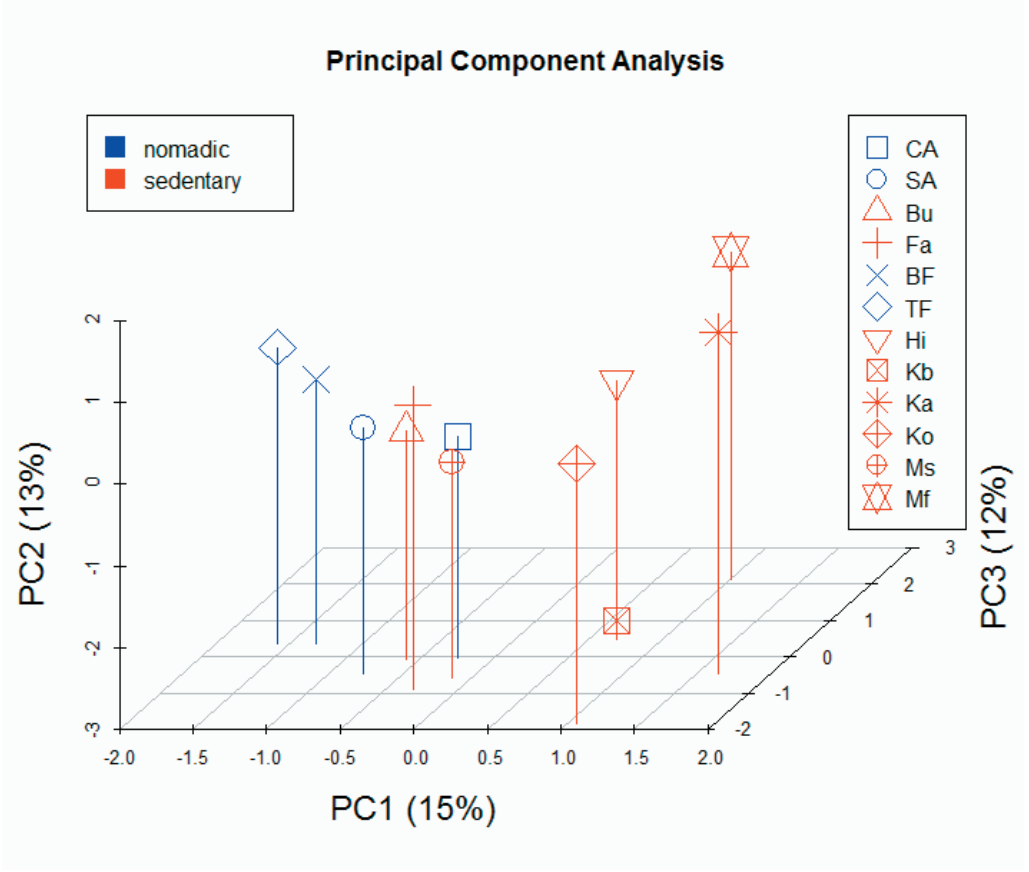


Figure 3


Principal Component Analysis
nomadic
sedentary
CA
SA
Bu
Fa
BF
TF
Hi
Kb
Ka
Ko
Ms
Mf
PC2 (13%)
PC1 (15%)
PC3 (12%)
-3
-2
-1
0
1
2
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
-2
-1
0
1
2
3

Supporting Information Legends

Table S1. mtSNP calling rates and number of phylogenetic inconcistencies in the global dataset. Mutation hits in Soares et al [33] are also given for the phylogenetic inconsistencies.

Table S2. mtSNP genotypes and control region data for the 542 individuals from the Chad Basin region genotyped.

Table S3. Haplogroup frequencies in all the population samples analyzed at the maximum level of resolution provided by the control region and the mtSNPs.

In preparation

Meta-analysis of Pan- African mtDNA variation provides new clues on the past continental demography and the patterns of the Trans-Atlantic slave trade

María Cerezo[1] et al

[1] Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Ciencias Forenses, Facultade de Medicina, Universidad de Santiago de Compostela, Santiago de Compostela, Galicia, Spain.

Keywords: mtDNA, haplotype, haplogroup, SNP, Chad Basin, Atlantic slave trade, Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry

Short title: Haplogroup mtDNA variation in the Chad

Abstract

Although Africa has been recognized to be the cradle of all human beings, the various genetic studies carried out in this continent have left many questions unresolved. A meticulous knowledge of the patterns of variation in Africa is also important when trying to map the African variation observed in America that arrived there as a consequence of the Trans-Atlantic slave trade. In order to tackle the analysis of modern African and African-American patterns, we have carried out one of the largest and comprehensive meta-analysis carried out to date in human populations: (i) new generated high-through put mtDNA SNP data of 2426 African and African-American samples, (ii) XXX control region profiles collected from the literature and representing XXX and XXX populations from Africa and (African)-America respectively, (i) and XXX entire genomes retrieved from the literature… Insights into the African home of humans are also discussed, and in opposition to previous hypothesis suggesting East or South Africa as the cradle of the human beings, the data is more consistent with a West-Central origin. Although, progress have been made in order to allocate the African-American variation into African regions (indicating the important role of the Golden Coast and the Bight of Biafra), the present study shows that mtDNA more ambitious entire genome sequencing projects are needed in order to narrow the most likely geographical origin of individual African-American profiles to particular regions in Africa.

Introduction

A meta-analysis of the African mtDNA variation was carried out in Salas et al. (Salas et al. 2002) in order to reveal the making of the African mtDNA landscape. A phylogeographic analysis of >2800 samples throughout the continent allowed to reveal for the first time a genetic phylogeographic map that correlated well with the southeastern route of Bantu expansion. Since then, several studies has been published in order to reconstruct the south western side of the Bantu expansion (Plaza et al. 2004; Beleza et al. 2005), or the demographic history of particular regions in Africa (Kivisild et al. 2004; Gonder et al. 2007), or e.g. histories of gene flow between pastoralist and hunter-gathers pygmies (Quintana-Murci et al. 2008). Behar

Most of the previous studies carried out in Africa and African-Americans were based on control region data and/or few RFLPs or mtSNPs (Beleza et al. 2005; Černý et al. 2006; Černý et al. 2007; Pereira et al. 2010). Only few of them targeted complete genome sequences but with a focus on the African phylogeny (Behar et al. 2008) or a regional one (Kivisild et al. 2004; Gonder et al. 2007). The present study aims to overcome the limitations of past studies by following a double strategy: (i) the huge amount of data that has accumulated in the literature during the past decade allows now the inference of demographic and diversity parameters that was not possible before (Salas et al. 2002); and (ii) the use of a high-throughput mtSNP genotyping technique allows to gain in phylogeographic resolution with respect to previous studies based exclusively on control region data. This double strategy can now be applied to the reconstruction of the African past but also to the analysis of the mtDNA patterns of the African Diaspora to America.

On the other hand, a detailed map of the African mtDNA variation provides a good opportunity for the reconstruction of the Trans-Atlantic slave trade occurred during the 16th and 19th centuries. Previous studies (Salas et al. 2004; Salas et al. 2005) have tried to reconstruct the global patterns of the Trans-Atlantic slave trade, indicating that with the data available at that time (mainly control region HVS-I segments), mtDNA could be traced to brad geographical regions within Africa, largely coinciding with the historical evidence. The results by Salas et al. indicated that West and West-Central accounted by most of the mtDNA variability in America, with a minor role of South-East. Further studies were carried out during the last few years. Brucato et al. (Brucato et al. 2010) focused on the analysis of the Noir Marron of French Guiana and analyzed patterns of mtDNA, Y-chromosome and HTLV-1 sequences; indicating a major role of the Gold Coast and the Bight of Benin in providing slaves to the island, although with some sex-bias. Most recently, the study by Stefflova et al. (2011) based on the analysis of HVS-I sequences collected from America and Africa concluded that African populations contributed differently to distinct populations of the New World. taimed to dissect the within African ancestry of admixed populations in America and concluding that

Material and Methods

Samples

Apart from the data collected from the literature, new samples have been analyzed here for the first time for a set of 230 mtSNPs. In total, we have genotyped 2024 African and 402 American samples as detailed in Table S2.

Datamining

Control region data has been collected from XXX articles. Most of the data corresponds to the HVS-I segment, and the sequence range 16030 to 16365. A total of XXX profiles have been compiled in Africa as described in Table SXXXX. This data represents quite well the main geographical sub-regions (see below). In America, we focused in lineages of recent African ancestry, namely, all haplogroup L, M1, and U6 mtDNAs.

Rationale of geographical sub-division of Africa and America

The present study focuses on two main topics, one regarding the demography of the African continent, and the other the trans-Atlantic slave trade. Division of populations into main regions is necessary for the purpose of statistical meta-analysis; however, grouping of populations might be somehow arbitrary. We here adopt a criterion that tries to fit the interests of the two mentioned topics of the present study, and it is modified from the subdivision proposed by Brucato et al. (Brucato et al. 2010). (which is mainly inspired in the historical regions of slavery described by Klein (Klein 1999), the genetic criteria and published genetic studies (Salas et al. 2002; Salas et al. 2005; Brucato et al. 2010; Veeramah et al. 2010; Stefflova et al. 2011)): (i) Northwest Africa (Canary Island, Mauritania, Western Sahara, Morocco, and Algeria, and Tunisia), (ii) Northeast Africa (Libya, and Egypt), (iii) Windward Coast, Senegambia and Sierra Leone (Cape Verde, Senegal, Gambia, Guinea-Bissau, Guinea, Mali, Sierra Leone, and Liberia), (iv) Gold Coast and Bight of Benin (Burkina Faso, Ivory Coast, Ghana, Togo and Benin), (v) Bight of Biafra (Republic of Niger, Chad, Nigeria, Cameroon, Central African Republic, Equatorial Guinea, Gabon, and Republic Democratic of the Congo [Zaire], São Tome and Príncipe, and Bioko), (vi) Southwest Africa (Democratic Republic of Congo, Cabinda, and Angola), (vii) South Africa (Namibia, Botswana, South Africa, Lesotho, and Swaziland), (viii) South East (Zambia, Zimbabwe, Mozambique, Malawi, and Tanzania), and (ix) East Africa (Uganda,Rwanda, Kenya, Somalia, Ethiopia, Sudan).

The above subdivision was used for the statistical analysis of control region data given the fact that the data compiled from the literature covers reasonably well all the mentioned sub-regions. However, the mtSNP data generated in the present study does not equally fit this scenario. Therefore, for the analysis of mtSNP we merged several sub-regions as follows: (i) North Africa (including Northwest and Northeast), (ii) West-central (that includes Windward Coast, Senegambia, Sierra Leone, Gold Coast, Bight of Biafra), (iii) Southwest, (iv) Southeast, and (v) East Africa.

America is subdivided in three main regions: (i) North (Canada and USA), (ii) Central or Meso-America (from Mexico to Panama and the Caribbean), and (iii) South (from Colombia to the southern cone). Only L, M1

and U6 lineages have been compiled from the literature and sub-genotyped in the present study. This will probably underestimate the contribution of North Africa to the pool of African-American lineages, but it shouldn't affect the proportions regarding its contribution regarding L, M1 and U6.

Genotyping

Apart from the mtSNPs that were genotyping using MALDITOF, three additional SNPs were genotyping using sequencing procedures. These SNPs were added a posteriori in order to identify three additional clades of the mtDNA phylogeny not considered in the initial MALDITOF design.

Some of the samples were already sequenced for partial or entire control region information in previous studies. See Table 1 for more details.

Statistical analysis

Analysis of several diversity indices, such as DnaSP v.5 (Librado, Rozas 2009) was used to compute diversity indices, including nucleotide and haplotype diversity and the average number of nucleotide differences. Analysis of molecular variance (AMOVA) and the significance of the covariance components associated with different levels of genetic structure was tested applying a non-parametric permutation procedure using Arlequin 3.5.1.2 (Excoffier, Smouse, Quattro 1992).

Principal Component Analysis was carried out on population hg frequencies and using R (http://www.r-project.org/).

jModelTest v.0.1.1. (Posada 2008) was used (using default parameters) to obtain the base frequencies, mutation parameters and the best mutation model (according to the Akaike information criteria), which (for our data) was the general time reversible (*GTR*) model. Estimates from iModelTest were obtained using three replicates using 10 initial chains (sampling interval of 20 and burn-in period of 1000) and two final chains (sampling interval of 20 and burn-in period of 1000). The results from iModelTest were applied in Lamarc (Kuhner 2006) as described in Cerezo et al. (XXX). Briefly, we have estimated several demographic parameters: (i) θ, which for females it is expected to be equal to $2N_e\mu$ for neutral mutations in mtDNA (where μ is the neutral mutation rate per generation and N_e the effective population size of females), (ii) population growth as $g = -(\ln\theta_t / \ln\theta_{present\ day})/t$, where g is expressed as the relationship between θ at a time $t > 0$ in the past and θ at the present ($t = 0$), and (iii) migration rate, defined as $M = m/\mu$, where m is the immigration rate per generation), between the 12 ethnic groups used in the present study and using information from mtSNPs. Estimates were obtained for three independent replicates using a Bayesian framework.

Apart from the migration rates obtained from Lamarc, another independent estimate was obtained using a procedures based on haplotype sharing. Details on these estimates are provided in XXX. XXXbla bla bla

As commented in Cerezo et al. (XXX), these estimates should be taken with care given that the mtDNA is in reality a single locus. In contrast with our previous study (Cerezo et al. XXX), sample sizes are larger now (with only few exceptions), and the models for admixture are more realistic since all the populations that presumable admixed within the African continent and the potential source populations to America are represented all in the model.

As done previously (Cerezo et al. XXX), for all of these computations involving mtSNPs, missing data was imputed according to known phylogeny. Given the fact that there exists a robust mtDNA worldwide phylogeny based on entire mtDNA genomes (>8,700), the phylogenetic-based approach for imputation seems more reliable than those based on e.g. metrics for linkage disequilibrium (Elson et al. 2007).

Results

Genetic diversity

As expected, North Africa shows very low values for all the diversity indices computed in comparison to sub-Saharan Africa, mirroring the more Mediterranean nature of the region (Table 1). Note that only 17% of the lineages belong to L-haplogroup branches. The values are consistent for the mtSNPs and the HVS-I sequences.

Within sub-Saharan Africa, South-East is the region with the lowest values of haplotype diversity, which is compatible with the genetic drift events of the Bantu expansion that end in the region (Salas et al. 2002); other parameters, such as nucleotide diversity (which is correlated with the average number of nucleotide differences) do not show the same signal probably due to the presence of Khoisan lineages in the population (Salas et al. 2002). The same pattern occurs in South Africa for the HVS-I sequences (there are not mtSNP data for the region), another region where Khoisan and Bantu lineages coexist.

Diversity values of 'African-American' lineages are comparable to those observed in sub-Saharan Africa mirroring the fact that no genetic drift occurred during the slave trade, as expected from a process that involved massive forced migrations.

Phylogeography of the main African sub-regions

Analysis of mtDNA phylogeographic patterns in Africa can now be inspected with a much higher phylogeographic resolution than in previous attempts (Salas et al. 2002). More than 2000 mtDNAs could be classified to the maximum known level of phylogenetic resolution.

As expected, the mtDNA patterns of North Africa clearly differentiate from those of sub-Saharan Africa. While L haplogroups represent only 17% of the lineages in the region, L makes-up more than 92% in Sub-Saharan African. Since our samples represent the northwest of the African continent, the patterns of L-lineages resemble quite well the haplogroup frequencies observed in west central Africa. For instance, L3e5, and L3f3 make up in North Africa 33% and 17% of L3 lineages, which are two of the most frequent L3 clades in west-central. The presence of the African M1 and U6 representatives in North African is another distinctive feature of the North; this mtDNAs most likely come from East Africa and Middle East. Most of the non-African lineages in North come most likely from Europe or Middle East.

L0 reaches the highest frequencies in East (22%) and southeast Africa (32%), in contrast to the 6% and 11% of West-Central and South-West, respectively. It is also very diverse in the East, where we observe 11 different sub-clades, in contrast to the XXX sub-clades found in southeast and the four

found in e.g. West-Central (where L0a2a represents by far the most frequent lineage). Moreover, the haplotype diversity is highest in the East (Table 2) and significantly lower in the southeast; thus, L0a1b1 and L0a2a2 are the most frequency sub-clades in South-East by far. When looking at the diversity of HVS-I segments, it is also noticeable that the Gold Coast has the lowest diversity values of continental Africa; curiously, the second most lower value for L0 lineages is found in Central America, perhaps signaling a major impact of the Gold Coast in the slave trade to the Caribbean. The opposite pattern is found in the Bight of Biafra, where the haplotype diversity is the highest followed by North America.

In contrast to L0, haplogroup L1 is at highest frequency (28%) and surprisingly diverse in South-West Africa. There are 17 different lineages in this region while it drops to 11%, 9%, 2%, and 1% in West-Central, South-East, East and North, respectively. Consistent with this fact, is that South-West Africa shows the highest L1 haplotype diversity. However, note that all but one, are L1c lineages. L1c is also the main branch in the other regions with the exception of West-Central, where other clades reach also high frequencies e.g. L1b1a (46% of all L1) and L1b1a3 (14% of all L1). In America, the South show higher haplotype diversity than Central and North, which is also consistent with the highest influence of the South-West African region in South America, which is in agreement with the historical documentation.

L2 is very diverse in West-Central Africa, with 25 lineages, (21 of them with frequencies above 1% within L2). L2 is more frequent in South-East (36%) than in West-Central (29%) but it is less diverse (only two lineages make 74% of the L2 mtDNAs). Within West-Central Africa, the Bight of Biafra reaches the highest haplotype and nucleotide diversity. In America, the South accounts for the highest haplotype diversity. L2a is the most frequent clade in all Africa (above XXX%) with the exception of South-West.

L3 has the highest frequency in South-West Africa (45%), followed by West-Central Africa (43%) and East Africa (34%). L3 is however significantly more diverse in the West-Central than in any other region. Thus, in West-Central we detected 40 different L3 sub-clades, followed by the 27 seen in East Africa. However, this observation is in apparent contradiction with the fact that the haplotype diversity reaches the highest value in East Africa. This could be due to the fact that many L3 lineages are very badly defined at the level of the control region (e.g. L3e is the L3 sub-clade more frequent in the West-Central), with few exceptions such as L3i and L3h that are significantly more frequent in the East.

L4 and L5 are significantly more frequent in East Africa than in the other African regions. L4 is 15% of all the mtDNAs in the East, with four different clades, being the L4b2a2 the most frequent clade. L5 is less frequent (6%) but is slightly more diverse. L5 is also frequent in South-East but at very low frequency (1%) and less diverse.

Apart from the North, M1 appears only in East Africa and West-Central Africa. The ones observed in East Africa came most likely from Middle East while those from West-Central arrived to the Northern groups in the Chad Basin (Shuwa and Chad Arabs and Buduma) (Černý et al. 2007) from Central-North Africa. It can be inferred from control region data that M1 appears only sporadically in more southern/western regions to the Chad Basin.

Phylogeography of lineages of recent African ancestry in America

Figure XXX shows that North America have more different clades than Central and South. Nearly all the African lineages belong to L haplogroups; U6 is only present at 1% in South and North America, respectively, while M1 is only present in South America (2%). The distribution of the main phylogenetic L-branches is similar in North, Central and South America. L2 is the most frequency haplogroup (above 37%) followed by L3, and L1. The patterns of sub-clades vary significantly in the different regions.

The diversity of L0 sub-clades in America is comparable to West-Central and South-West Africa. One of the most outstanding features of L American lineages is that L0a1a is by far the most frequent L0 sub-clade (56% within L0) in North America, while it is completely absent in Central and South.

L1b1a is the most frequent L1 clade in North and Central America, and the second most frequent in South America, again resembling the frequency observed in West-Central Africa.

It is also noticeable that the typical East African clades, L4 and L5 are completely absent in America.

Demographic estimates in Africa

Demographic estimates in regards to the Atlantic Slave Trade

Admixture in African-America

A Bayesian-based model has been used in order to estimate the most likely contribution of the different African locations in present-day African-American locations.

Principal Component Analysis of African and 'African-American' populations

Analysis of Molecular Variance (AMOVA) in Africa and America

AMOVA have been carried out based only on control region data gathered from the literature in order to take advantage of the very large amount of samples available in African (240 ethnic groups) and America (64 sampled groups) compared to the data newly generated here for the mtSNPs.

As expected, most of the variation occurs within populations, but variable values of among population variation were observed depending of the population groups targeted. In Africa, analysis of 240 population samples the among population variation accounts for 14.5%, which is the highest level observed among the different continents in agreement with the demographic ancestral history of all human beings. When the analysis is carried out hierarchically considering the main African regions, 8-8.6% of the variation (depending on the grouping scheme) is accounted by among population within groups, while 8.1-9.2% accounts for among group variation; the remaining correspond to within population variation. When carrying out AMOVA analysis in

the different African regions, by far, the Bight of Biafra is the region that shows the highest levels of among population variation, as high as for the value obtained for the whole continent (14.5%). South East also shows significant levels of population stratification indicated by the 6% of variation among groups.

In America, the among population variation of haplogroup L haplotypes drops to 2.3%, a significant much lower value than the one observed in Africa. Given that West-Central Africa was the main source region for the Atlantic slave trade, it could be contradictory to find so different values of among group variation in America and West-Central Africa. However, the lack of stratification in African-American lineages can be explain if one consider that the different L haplotypes from America arrived there during the slave trade as a consequence of a kind of random sampling of West-Central African individuals. The highest levels of among group variation were found in Central (F_{ST} = 4.5) and South (F_{ST} = 2.9) America; perhaps mirroring the genetic drift that most likely hit the different Caribbean islands (Central America) (compared to continental locations) and the growing contribution of other African sources like South-West and South-East (especially in Central-America) during the final period of the slave trade before abolition.

Discussion

In the present study, we have used a combined approach of (i) high-through put mtSNP data of newly sample locations from Africa and America (ii) a survey of partial control region data (mainly HVS-I segments) from the literature that involves more than XXX profiles from Africa and XXX African-American profiles. We have also explored the patterns of variability of XXX complete genomes in order to make inferences about the potential origin of mtDNA African-American genomes to their original locations in Africa.

The original home of modern humans has been a common topic of debate in the last few decades among geneticists and molecular anthropologists. A phylogenetic study based on entire mtDNA genomes (Behar et al. 2008) and mainly centered in the analysis of Khoisan people proposed two scenarios, one favoring an East African origin, and another one supporting a Central or West-Central African one. However, the phylogenetic signal for any of these two scenarios is hampered by several limitations. First, the root of the African mtDNA tree has not been observed in anywhere in Africa; the deepest phylogenetic mtDNA split leads on one hand to a dichotomic branch (L0) that is found in Khoisan populations (L0d haplogroup) and Khoisan/Bantu populations (L0a'b'f'k), and another branch (L1) from which a plethora of nested sub-branches emerge that are observed in different African locations; the latter contains also the macro-haplogroups that explain the out of African mtDNA variation (M and N that derived from haplogroup L3). Second, sub-Saharan Africa has experienced important demographic movements that have probably blurred the ancestral patterns of mtDNA variability; one of the most important movements occurred only few thousand years ago and it is commonly known as the Bantu expansion (Salas et al. 2002; Beleza et al. 2005); therefore, present patterns perhaps do not capture past events. Third, the Khoisan peoples occupied a wide range of territory in all the South cone of Africa but were probably forced to retract to their modern locations in South Africa as a consequence of the southwards Bantu expansion. Although empirical validation of different scenarios for the origin of human is not possible, it is tempting to

seek our result for an independent perspective to the phylogenetic views proposed by Behar et al., but this time based on patterns of sequence diversity. The values of mtDNA diversity observed in West-Central and East Africa are not very informative in this regard (Table XXX). However, variation accounted among populations is significantly much higher in West-Central than in any other region in Africa. Population stratification would therefore support more strongly a West-Central origin of early humans than other regions in Africa.

It remains to be elucidated how diversity values could have been determined by the complex demographical movements occurred in Africa; but perhaps a phylogenetic view based on a more in-depth sampling of African populations would allow to shed more light on the geographical origin of modern humans. New archaeological findings (Balter 2011) pointed to North Africa as the cradle of human origin; the genetic landscape of this region however has been also enormously altered since ancient times by the complex interactions experienced with peoples coming from the Middle East and the Mediterranean Basin.

The present study has shed new light into the process of the trans-Atlantic slave trade. According to the documentation, West-Central Africa seems to have played the main role in providing slaves into America. Within West-Central Africa, the Bight of Biafra and the Golden Coast contributed on average about XXX of the African-American lineages. The South-West and the South-East African regions gained role in the final days of the slave trade and their contribution was more important in South America, as testified by the model of admixture. It is interesting to note that diversity indices are similar for the three American regions. However, when computed per haplogroup, there are significant differences that seem to be correlated with the values observed in the most likely African source regions. This pattern is clearly visible when observing the frequency distribution of some sub-clades (Figure XXX).

Further studies are needed in order to improved the limitations of the present study. For instance, the database is still very limited in representing African and American regions and populations. For example, while West-Central Africa is quite well sampled, the Bight of Biafra accounts for XXX of the region and XXX of the total African samples. To what extent this could bias some of the analyses carried out in this and other's studies remains to be elucidated. We here provided the view of the mtDNA, which although is the marker for which more data has been accumulated in the literature in terms of sample size and populations represented, it only represents a single marker that is inherited throughout the matriline. Finally, we have seem ideally, entire genome data would be necessary to allow a more detailed analysis of within African genetic patterns and to narrow the most likely African origin of African-American lineages.

Acknowledgements

References

Balter, M. 2011. Was North Africa the launch pad for modern human migrations? Science 331:20-23.

Behar, DM, R Villems, H Soodyall, et al. 2008. The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82:1130-1140.

Beleza, S, L Gusmão, A Amorim, Á Carracedo, A Salas. 2005. The genetic legacy of western Bantu migrations. Hum Genet 117:366-375.

Brucato, N, O Cassar, L Tonasso, P Tortevoye, F Migot-Nabias, S Plancoulaine, E Guitard, G Larrouy, A Gessain, JM Dugoujon. 2010. The imprint of the Slave Trade in an African American population: mitochondrial DNA, Y chromosome and HTLV-1 analysis in the Noir Marron of French Guiana. BMC Evol. Biol. 10:314.

Černý, V, M Hájek, M Bromova, R Cmejla, I Diallo, R Brdička. 2006. MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. Hum Biol 78:9-27.

Černý, V, A Salas, M Hájek, M Žaloudková, R Brdička. 2007. A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71:433-452.

Elson, JL, K Majamaa, N Howell, PF Chinnery. 2007. Associating mitochondrial DNA variation with complex traits. Am J Hum Genet 80:378-382; author reply 382-373.

Excoffier, L, PE Smouse, JM Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Gonder, MK, HM Mortensen, FA Reed, A de Sousa, SA Tishkoff. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. 24:757-768.

Kivisild, T, M Reidla, E Metspalu, A Rosa, A Brehm, E Pennarun, J Parik, T Geberhiwot, E Usanga, R Villems. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am. J. Hum. Genet. 75:752-770.

Klein, HS. 1999. The Atlantic Slave Trade. Cambridge: Cambridge University Press.

Kuhner, MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768-770.

Librado, P, J Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452.

Pereira, L, V Černý, M Cerezo, NM Silva, M Hájek, A Vašíková, M Kujanová, R Brdička, A Salas. 2010. Linking the sub-Saharan and West Eurasian gene pools: maternal and paternal heritage of the Tuareg nomads from the African Sahel. Eur J Hum Genet 18:915-923.

Plaza, S, A Salas, F Calafell, F Corte-Real, J Bertranpetit, Á Carracedo, D Comas. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet 115:439-447.

Posada, D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253-1256.

Quintana-Murci, L, H Quach, C Harmant, et al. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc. Natl. Acad. Sci. U S A. 105:1596-1601.

Salas, A, Á Carracedo, M Richards, V Macaulay. 2005. Charting the Ancestry of African Americans. Am. J. Hum. Genet. 77:676-680.

Salas, A, M Richards, T De la Fé, MV Lareu, B Sobrino, P Sánchez-Diz, V Macaulay, Á Carracedo. 2002. The making of the African mtDNA landscape. Am. J. Hum. Genet. 71:1082-1111.

Salas, A, M Richards, MV Lareu, R Scozzari, A Coppa, A Torroni, V Macaulay, Á Carracedo. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am. J. Hum. Genet. 74:454-465.

Stefflova, K, MC Dulik, JS Barnholtz-Sloan, AA Pai, AH Walker, TR Rebbeck. 2011. Dissecting the within-Africa ancestry of populations of African descent in the Americas. PLoS One 6:e14495.

Veeramah, KR, BA Connell, NA Pour, A Powell, CA Plaster, D Zeitlyn, NR Mendell, ME Weale, N Bradman, MG Thomas. 2010. Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. BMC Evol. Biol. 10:92.

Table 1. Diversity indices in the main African regions; based on mtSNPs (samples genotyped in the present study), and based on the HVS-I (data compiled from the literature; sequence range: 16090-16365); *n* = sample size; *k* = number of different sequences; *S* = number of segregating sites; *h* = haplotype diversity; *π* = nucleotide diversity; *M* = average number of pairwise differences (mismatch observed mean)

Continent	Continental region	n	K	k/n	S	H	π	M
mtSNPs								
Africa	North	107	52		79	.938 (.017)	.032 (.003)	7.5
	West-Central	745	224		168	.984 (.001)	.049 (.001)	11.3
	South-West	106	68		106	.986 (.004)	.055 (.002)	12.7
	South-East	835	148		142	.935 (.004)	.055 (.001)	12.6
	East	231	111		130	.982 (.003)	.050 (.001)	11.7
	Total	2024	471		205	.982 (.001)	.054 (.000)	12.6
America	North America	219	104		128	.985 (.002)	.050 (.002)	11.5
	Central America	56	47		99	.993 (.005)	.052 (.002)	12.1
	South America	107	60		113	.966 (.009)	.050 (.002)	11.5
	Total	382	152		154	.984 (.002)	.050 (.001)	11.6
HVS-I								
Africa	North-East	600	250		102	.963 (.005)	.020 (.001)	5.2
	North-West	2619	764		156	.970 (.002)	.018 (.000)	4.9
	North	3219	876		151	.966 (002)	.018 (.000)	4.7
	Senegambia	1394	397		116	.983 (.001)	.020 (.000)	5.4
	Bight of Biafra	4063	950		155	.989 (.001)	.033 (.000)	8.8
	Gold Coast	713	258		100	.984 (.002)	.025 (.000)	6.9
	West-Central	6170	1093		157	.984 (.001)	.025 (.000)	6.3
	South-West	157	96		77	.991 (.002)	.032 (.001)	8.9
	South-East	587	190		95	.971 (.003)	.030 (.000)	8.2
	East	627	343		123	.995 (.001)	.030 (.000)	8.1
	South	33	12		34	.898 (.027)	.031 (.002)	8.7
	Total	10793	1706		170	.984 (.000)	.021 (.000)	5.1
America	North America	1476	415		107	.978 (.002)	.024 (.000)	5.7
	Central America	451	213		98	.988 (.002)	.028 (.001)	7.6
	South America	1036	358		115	.990 (.001)	.028 (.000)	7.4
	Total	2963	609		125	.978 (.001)	.023 (.000)	5.4

Table 2. Diversity indices in the main African regions and haplogroups based on the HVS-I (data compiled from the literature). Codes are as in Table 1.

HG-Continent	Continental region	*N*	*K*	*k/n*	*S*	*H*	*π*	*M*
L0-Africa	North-East	18	7		17	.752 (.082)	.009 (004)	2.6
	North-West	12	8		16	.894 (.078)	.019 (.002)	5.2
	North	30	14		22	.885 (.038)	.015 (003)	4.0
	Senegambia	39	16		17	.838 (.053)	.008 (001)	2.1
	Bight of Biafra	274	53		42	.922 (.009)	.011 (.000)	3.1
	Gold Coast	15	5		5	.629 (.125)	.003 (.001)	0.9
	West-Central	328	61		48	.903 (.011)	.011 (.000)	2.9
	South-West	20	9		8	.889 (.038)	.010 (.001)	2.7
	South-East	165	48		50	.836 (.025)	.016 (.001)	4.4
	East	89	49		54	.953 (.015)	.023 (.002)	6.2
	South	17	6		16	.765 (.075)	.023 (.003)	6.3
	Total	649	154		88	.932 (.005)	.015 (.000)	4.2
L0-America	North	59	22		22	.904 (.024)	.010 (.001)	2.6
	Central	27	7		6	.752 (.064)	.007 (.001)	2.0
	South	76	19		25	.846 (.027)	.011 (.001)	3.1
	Total	162	37		34	.873 (.014)	.010 (.001)	2.7
L1-Africa	North-East	15	10		23	.943 (.040)	.027 (.003)	7.3
	North-West	134	36		36	.825 (.032)	.010 (.001)	2.7
	North	149	43		41	..853 (.027)	.012 (.001)	3.3
	Senegambia	282	90		63	.932 (.011)	.020 (.001)	5.2
	Bight of Biafra	1320	222		97	.941 (.000)	.025 (..000)	6.9
	Gold Coast	131	47		43	.906 (.017)	.020 (.001)	5.6
	West-Central	1733	285		99	.935 (.003)	.025 (.000)	6.6
	South-West	40	30		43	.985 (.009)	.025 (.002)	7.0
	South-East	37	23		37	.958 (.017)	.026 (.002)	7.1
	East	19	12		24	.936 (.037)	.025 (.004)	6.8
	South	–	–	–	–	–	–	–
	Total	1979	327		104	.935 (.003)	.025 (.000)	6.5
L1-America	North	331	127		68	.938 (.010)	.025 (.001)	6.4
	Central	100	53		58	.965 (.009)	.023 (.001)	6.3
	South	275	120		70	.968 (.005)	.024 (.001)	6.5
	Total	706	222		87	.957 (.005)	.024 (.000)	6.2
L2-Africa	North-East	41	18		22	.804 (.063)	.009 (002)	2.5
	North-West	163	75		58	.960 (.009)	.016 (.001)	4.4
	North	204	78		60	.944 (.009)	.013 (001)	3.6
	Senegambia	555	186		82	.972 (.003)	.017 (.000)	4.5
	Bight of Biafra	891	248		101	.976 (.002)	.019 (.000)	5.2
	Gold Coast	268	98		62	.952 (.008)	.015 (.001)	4.2
	West-Central	1714	425		125	.975 (.002)	.017 (.000)	4.7
	South-West	27	16		27	.946 (.025)	.021 (.002)	5.7
	South-East	185	47		43	.886 (.013)	.011 (.001)	3.1
	East	104	58		53	.973 (.007)	.017 (.001)	4.7
	South	–	–	–	–	–	–	–
	Total	2239	505		138	.967 (.002)	.016 (.000)	4.2
L2-America	North	483	167		88	.954 (.000)	.016 (.000)	4.3
	Central	167	81		59	.960 (.008)	.017 (.001)	4.7
	South	300	107		88	.964 (.005)	.016 (.000)	4.3
	Total	950	276		110	.958 (.003)	..017 (.000)	4.4
L3-Africa	North-East	73	44		42	.956 (.016)	.017 (.001)	4.5
	North-West	228	102		70	.963 (.007)	.015 (.001)	4.0
	North	301	128		77	.972 (.005)	.015 (.001)	4.0
	Senegambia	458	154		78	.971 (.003)	.016 (.000)	4.2
	Bight of Biafra	1453	408		124	.980 (.001)	.019 (.000)	5.3
	Gold Coast	253	97		64	.960 (.007)	.017 (.000)	4.6
	West-Central	2164	550		135	.980 (.001)	.018 (.000)	4.8
	South-West	62	34		31	.967 (.011)	.018 (.001)	5.0
	South-East	143	56		46	.957 (.008)	.015 (.001)	4.1
	East	169	96		71	.988 (.002)	.020 (.001)	5.2
	South	–	–	–	–	–	–	–
	Total	2849	659		145	.977 (.001)	.016 (.000)	4.4
L3-America	North	594	182		86	.970 (.003)	.016 (.000)	4.1
	Central	152	70		55	.967 (.007)	.018 (.001)	4.9
	South	374	122		70	.972 (.003)	.015 (.000)	4.0
	Total	1120	260		110	.965 (.002)	.013 (.000)	3.6

Table 3. AMOVA analysis of African and American populations to different hierarchical levels; *n* = number of populations in each category. Codes for populations: AF-NE: North-East Africa, AF-NW: North-West Africa; Se: Senegambia; BB: Bight of Biafra; GC: Gold Coast; AF-WC = Se + BB + GC; AF-S: South-West Africa; AF-S: South Africa; AF-SE: South East Africa; AM-N: North America; AM-C: Central America; AM-S: South America. AMOVA was not computed for South Africa due to the small sample size of this region, but the data was incorporated when computing AMOVA in other categories.

	N_g	N_p	Among Pop./Groups	Among Pop. within groups	Within Populations
Africa					
All populations	1	240	14.5	–	85.5
AF-NE/AF-NW/Se/BB/GC/ AF-SW/AF-SE/AF-E/AF-S	9	240	8.1	8.0	83.9
AF-N/AF-WC/AF-SW/AF-SE/AF-E/AF-S	6	240	9.2	8.6	82.2
AF-NE	1	3		–	
AF-NW	1	38	2.5	–	97.5
AF-N	1	41	2.6	–	97.4
Senegambia	1	28	2.6	–	97.4
Bight of Biafra	1	96	14.5	–	85.5
Gold Coast	1	38	3.4	–	96.6
AF-WC	1	160	12.7	–	87.3
AF-SW	1	3	1.0	–	99.0
AF-SE	1	23	6.0	–	94.0
AF-E	1	11	2.7	–	97.3
America					
All populations	1	64	2.3	–	97.7
AM-N/AM-C/AM-S					
AM-N	1	10	0.8	–	99.2
AM-C	1	13	4.5	–	95.5
AM-S	1	42	2.9	–	97.1

Legend to the Figures

Figure 1. Haplogroup patterns of mtDNA lineages in African populations. For the sake of clarity, haplogroup categories are not represented at the maximum level of phylogenetic resolution (see Supplementary Data S2).

Figure 2. Haplogroup frequencies of mtDNA lineages in America.

Figure 3. Principal component analysis based on haplogroup frequencies of the main African and American regions.

Figure 1

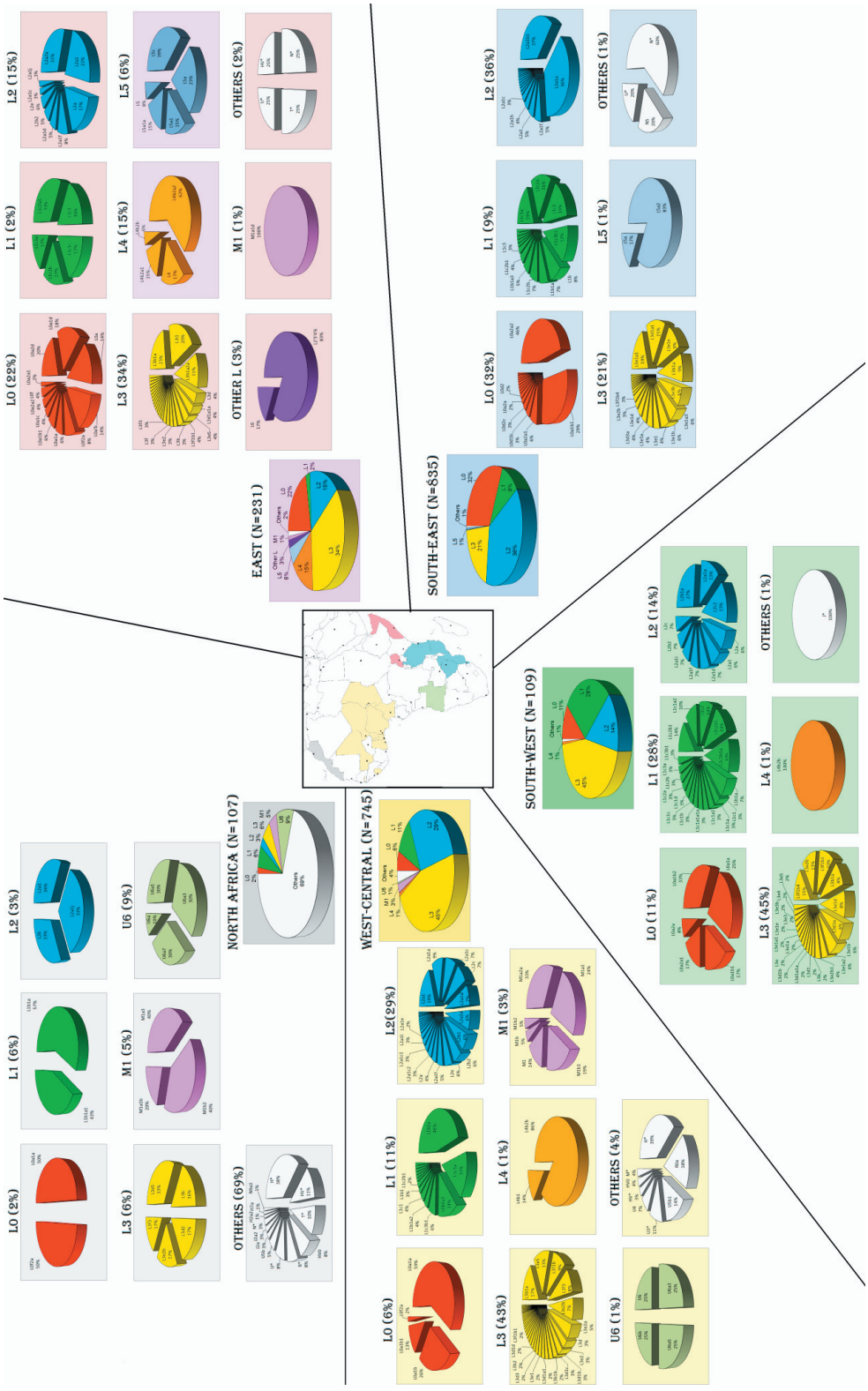
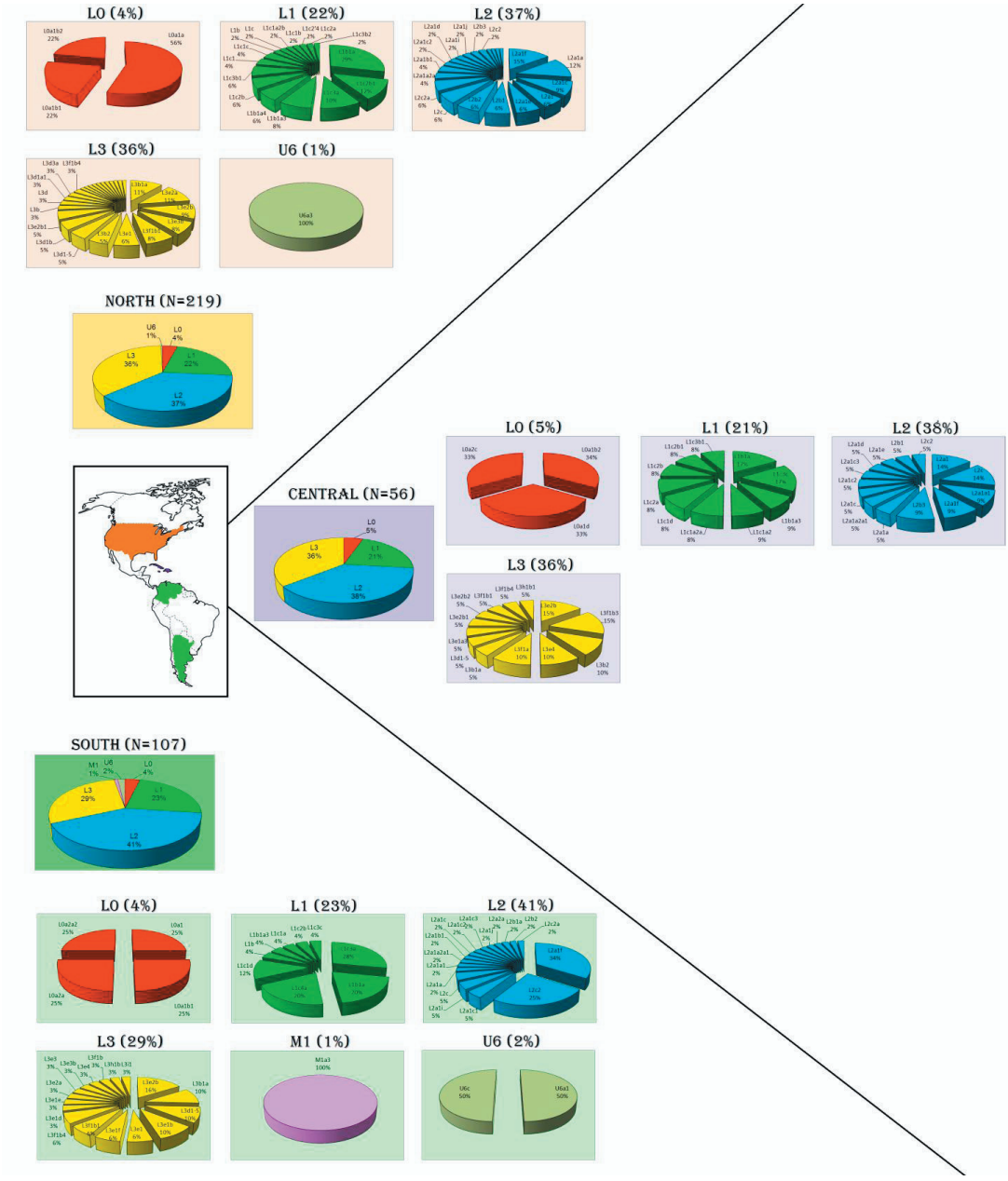

L2 (15%)
L1 (2%)
L5 (6%)
L4 (15%)
OTHERS (2%)
M1 (1%)
L0 (22%)
L3 (34%)
OTHER L (3%)
EAST (N=231)
L2 (36%)
OTHERS (1%)
L1 (9%)
L5 (1%)
L0 (32%)
L3 (21%)
SOUTH-EAST (N=835)
L2 (14%)
L1 (28%)
OTHERS (1%)
L4 (1%)
L0 (11%)
L3 (45%)
SOUTH-WEST (N=109)
L2 (3%)
L1 (6%)
U6 (9%)
M1 (5%)
NORTH AFRICA (N=107)
L0 (2%)
L3 (6%)
OTHERS (69%)
L2(29%)
L1 (11%)
M1 (3%)
L4 (1%)
OTHERS (4%)
WEST-CENTRAL (N=745)
L0 (6%)
L3 (43%)
U6 (1%)

Figure 2


L0 (4%)
L1 (22%)
L2 (37%)
L3 (36%)
U6 (1%)
NORTH (N=219)
CENTRAL (N=56)
L0 (5%)
L1 (21%)
L2 (38%)
L3 (36%)
SOUTH (N=107)
L0 (4%)
L1 (23%)
L2 (41%)
L3 (29%)
M1 (1%)
U6 (2%)

SUPPLEMENTARY MATERIAL

Supplementary Data S1. List of samples analyzed in the present study.

Supplementary Data S2. Mitochondrial DNA SNP genotyping for the XXX samples analyzed in the present study.

Supplementary Data S3. Haplogroup frequencies at the maximum level of resolution in the populations analyzed in the present study.

In preparation

Reconstructing mtDNA bridges used for African Diasporas into Europe

María Cerezo[1] et al

[1] Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Ciencias Forenses, Facultade de Medicina, Universidad de Santiago de Compostela, Santiago de Compostela, Galicia, Spain.

Abstract

Most of the African mitochondrial DNA lineages (mtDNA) belong to the macro-haplogroup L. In Europe, these lineages represent less than 1%. It has been suggested that these mtDNAs most likely arrived to Europe along a recent historical period. However, this hypothesis was based on the observation that mtDNA control region data alone does not show signals of having evolved within the European continent. Analysis of the mtDNA to a much higher resolution could however reveal different patterns and would allow dating specific lineages to different time-scales. Analysis of 69 haplogroup L-European mtDNA entire genomes revealed the existence of a number of lineages that seem to have evolved within Europe long time ago. Their European autochthonous nature find additional support in a surveyed of control region segments that involves few hundred profiles (from Europe, Near East, and Africa). Haplogroup L1b is by far the most common haplogroup in Europe, which has the highest frequent in West-Central African. A Bayesian-based approach indicated that the origin of European haplogroup L lineages can most likely be attributed to West-Central Africa, followed to East-Africa. Coalescent ages of genuine European autochthonous African haplogroups indicated that sub-Saharan Africa and Europe have maintained sporadic contacts since at least 15000 years ago although most of the African lineages have most likely arrived to Europe in more recent times, including the Romanization period, the Arab conquest of the Iberian Peninsula, and the Atlantic slave grade.

Analysis of the mitochondrial DNA (mtDNA) variation has been used for successfully reconstructing ancestral and modern human migrations. Most of the sub-Saharan African mtDNAs fall into one of the multiple sub-clades of macro-haplogroup L. In Europe, L haplotypes account for <1% of the mtDNAs [1]. It has been hypothesized that their arrival to Europe could have occurred most likely in recent times, during the Romanization period, the Arab/Berber conquest of Iberia, and more recently during the Atlantic slave trade. This hypothesis has been postulated under the observation that most of the L-mtDNAs in Europe represent patchy members of African phylogeny as randomly taken from different African source populations, without strong evidences indicating accumulation of divergence with the time within the European continent.

Analysis of the African Trans-Atlantic slave routes have received some attention in the literature [1-4], but its impact on Europe has not been analyzed in detail, mainly due to the scantiness of data. Salas et al. [1] surveyed the mtDNA HVS-I (hypervariable segment I) of ~15.000 individuals from Europe, and found 113 mtDNA of recent African ancestry (including haplogroups U6, M1 and L). According to these authors, African types in Eurasia can be broadly attributed to gene flow from both eastern Africa, western, and southeastern, but the proportions were not estimated nor the time frame for their arrival to Europe. Thus, based on the analysis of control region data compiled from the literature, Malyarchuk & Czarny [5] observed that most of the European L-lineages could have arrived to Europe very recently, with the exception of two monophyletic clusters of L1b and L3b that could be older (no more than 6,500 years). In a follow-up study based on the analysis of African-specific mtDNA entire genomes carried by eight Slav individuals, Malyarchuk et al. [6] could not replicate the previous findings; but instead, they observed a new sub-clade L2a1k (named as L2a1a in their study) that could have been specifically evolved within Europe about 10,000 years ago. Other studies carried out in specific European populations (e.g. from Iberia) have tried to explain the origin of the hg L lineages observed in those populations, in general concluding that virtually all these lineages most likely arrived recently to the European continent [7-9], mostly as a result of the movements of the European Empires in colonial times.

We have carried out the largest study of European haplogroup L entire genomes to date (n = 69) in order to unravel their temporal and geographic origin, under the hypothesis that most of them could have arrived to Europe very recently according to previous reports, but without ruling out the possibility of finding signals of ancestral African migrations into Europe. Information concerning the geographical origin of the grandparents was collected in order to disregard African individuals that could have arrived to Europe in very modern times (Supplemental Data). Moreover, analysis of a panel of ancestry informative markers (AIMs) was used to further corroborate their predominant

European ancestry. Written informed consent was required for all the samples. The study was approved by the Ethical committee of the University of Santiago de Compostela. The study conforms to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July)." More than 200 HVS-I profiles of European, Near East, and African origin were surveyed from the literature; XXX of the European ones (XXX%) belonged to different sub-lineages of macro-haplogroup L. Several sequences with dubious assignment into haplogroup L3 were excluded because they were not characterized by solid diagnostic sites.

DNA extraction was carried out using standard phenol-chloroform methods. Entire genome sequencing was carried out for all the samples. The protocol described in Álvarez-Iglesias et al. [9] was used in those cases involving good quality DNA which allowed large PCR amplicons; the primers were therefore the ones described in Torroni et al. [10]. Some of the samples were sequenced using the set of primers described by Kivisild et al. [11] following the conditions described in [9; 12]. MtDNA variation is referred to the revised Cambridge Reference Sequence [13]. Haplogroup nomenclature is based on previous studies e.g. [10; 11; 14-19]; and updated in Phylotree [20]. A posteriori checking of the sequences was carried out in order to prevent sequencing and documentation errors [21; 22]. Phylogenetic inconsistencies were confirmed by re-amplification and re-sequencing of both strands. DnaSP 4.10.3 software [23] was used for the computation of different diversity indices, including haplotype and nucleotide diversities and mean number of pairwise differences [24-26]. A multiplex of 34 ancestry informative markers (AIMS) have been genotyped for all the African-European samples following [27]. Structure 2.3.3. [28] was used to infer the ancestry proportions of the individuals using training datasets retrieved from HapMap data using SPSmart [29; 30]. Likelihood classification of European haplogroup L carriers were obtained using Snipper (http://mathgene.usc.es/snipper/)[27]. Spatial geographical representations of haplogroup frequencies were obtained using Surfer 8.0 (http://www.goldensoftware.com). The data used was collected from previous studies [31-34]. We used the inverse-squared distance method for interpolating frequency values. Haplogroup frequencies are presented in a regular grid covering Africa and West Eurasia and Middle East. Only data points within the same landmass, either island or continent, were considered for interpolation. Estimation of the time to the most recent common ancestor of each cluster and SDs were carried out according to Saillard et al. [35] and employing an evolutionary rate estimate according to [36].

Analysis of European L-haplogroup entire genomes revealed a number of novel clades of the African mtDNA phylogeny. In several instances, representatives of these clades have not been found in Africa suggesting that these branches could have been evolved within Europe. One of the most singular European haplogroup L sub-clade is L3d1b1a, defined by the stable

diagnostic transversion A8014C [36]. L3d1b1a has been found exclusively in five Italian citizens and seems to have evolved only locally. The age of this clade is about 3632 (95% SD:).

Malyarchuck et al. [6] recently proposed that the sub-clade L2a1k (defined by transition G6722A T12903C C16218T T16519C) could have been originated in Europe about 10,280 ± 5,140 years ago. Our data confirmed this finding as we did not observe representatives of L2a1k in our databank collection of African sequences (n = 2,426). We also searched L2a profiles with C16218T in a large survey of African sequences but any L2a1k candidate was observed. It is however important to note that variant C16218T could be a misleading diagnostic marker for L2a1k; this is because two of our samples from Benin and Cameroon carrying this variant were sequenced for the entire genomes but resulted to belong to a different clade, L2a1c (Figure S3 of Supplemental Data).

L1b is by far the most common L-African lineage in Europe, 49% according complete genome data and XXX% according to control region data. Previous studies mainly based on control region data [14] indicated that L1b hg is more frequent and diverse in West-Central Africa. We have collected from the literature and GenBank 73 L1b entire complete genomes (mainly from Africans and 'African-Americans'), which together with the ones analyzed in the present study from Europe sum-up to 103. From the total, the vast majority of the non European lineages were sampled in North America ('Afro-Americans' and 'Hispanics'; *n* = 40), followed by West-Central Africa (*n* = 13). The European L1b lineages make up 33% of the total (*n* = 34), most of them are newly reported here (*n* = 30).

The phylogeny of entire L1b points to a West-Central African origin of this haplogroup. However, control region data indicates that, although L1b is more frequent in West than in any other region in Africa (Figure 1), this haplogroup is slightly more diverse in East Africa than in the West (Table 1). L1b is also highly prevalent in America indicating that West-Central Africa was by far the main source of Atlantic African slaves [3]. Demographic movements within sub-Saharan Africa, could have carried this lineage to other African regions, including sporadic occurrences in the North (XX% and XX% are the frequencies of L1b in North-East and North-West Africa). L1b is present at low frequency and showing lower levels of diversity in Eastern Africa (Table 1); from here it probably moved to other African and non-African locations. There is in fact a clade of L1b defined by transition T7954C (Figure 2) and named here as L1b1a2, that probably evolved locally in East Africa (represented by three divergent sequences from Ethiopia that define the sub-clade L1b1a2a: EU092952, EU092942, and EU092950); L1b1a2a next could have moved from East Africa to the North along the Nile towards Egypt (represented by the complete genome EU092775). Another entire genome of L1b1a2 was found in Middle East and could have brought here from North or East Africa through the Arabian Peninsula (represented by the Israel Bedouin sequence EU092672)[17]. There is two representatives of L1b1a2a in Spain (one of them in Galicia),

which could have arrived during the period of the Atlantic slave trade or the Arab invasion of the Iberian Peninsula.

We have also identified a new sub-clade of L1b1a, L1b1a9, characterized by transversion G185C and transition T14040C. In contrast to most of the L1b sub-clades, L1b1a9 has a clear North African and Mediterranean distribution. It was perhaps originated in North-West Africa (as represented by the Moroccan Jews sequence EU092667), and afterwards moved to different European Mediterranean locations (mainly Iberia and Italy). Two L1b1a9 sequences were found in Iberia (Galicia and Catalonia), three in the Italian Peninsula, and one in France.

We have also found a new clade of L1b1a that has most likely evolved exclusively within Europe; L1b1a8. This clade is defined by transition A7298G. Figure 1 shows five European members that conform the L1b1a8 clade; three Andalucians, one Galician (present study), and one Russian [6]. The L1b1a8 status was investigated in the high mtDNA throughput data analyzed by the authors (Cerezo et al. unpublished). Here, 43 mtDNAs belonged to hg L1b1a from a total of 2426 (~1.7%) profiles representing different African and 'Afro-American' donors. These L1b1a sequences were observed in America ('Afro-America', 'Afro-Caribbean', Colombia and Argentina) and sub-Sahara Africa (Angola, Ghana, Morocco, Nigeria, Sierra Leone, Ivory Coast, Togo, Tanzania and Mozambique), but any of them carried the transition A7298G, therefore, further supporting an European origin of L1b1a8.

Apart from L1b1a8, there are other minor clades of L1b that could be good candidates as originated within Europe. Members of haplogroup L1b1a11 were only found in North-Central Europe (Ireland, Switzerland, and Slovenia), while L1b1a12 has representatives only in Iberia (Portugal and Catalonia). L1b1a6a is defined by a reversion at position 16093 (T to C, which is likely to be more mutationally stable than the common 16093 C to T transition) and it is present in two Portuguese, one Spaniard, and one individual from Wales. Finally, the immediate ancestor of L1b1a6a seems to have been evolved in West-Central Africa (as represented by two entire genomes from Burkina Faso and Guinea Bissau), and from here, it could have spread into Europe thought the Atlantic facade.

AGE ESTIMATE

Admixture proportions were calculated as done previously (REF). The main source populations for the European L lineages come from Bight of Biafra (32%), Senegambia (17%) and North(18%) (Figure 1).

An important effort in this study has been devoted to the investigation of the origin of haplogroup L lineages of African origin in Europe. Given the criteria of autochthony employed in the present study coupled with the results yielded by the analysis of ancestry based on AIMs, we could rule out (at least in our

samples) the bias that recent immigrants could have in the results. A proportion of the African-European mtDNAs investigated could be attributed to modern and well documented demographic routes that existed during the Romanization period, the Arab conquest and the slave trade (XXX%). However, XXX% of them point to the existence of sporadic population movements between both continents occurring during a large time-frame, even as older as 15,000 years ago. These contacts were not only restricted to North Africa, but to sub-Saharan regions via coastal routes or first crossing North African territories towards the Mediterranean with final destination in Europe.

Supplemental Data

Supplemental Data include XXX tables and can be found with this article online at http://www.ajhg.org/.

Acknowledgements

Mitochondrial DNA genomes analyzed in the present study were submitted to GenBank; accession numbers XXXXXXX- XXXXXXX. This research received support from the Fundación de Investigación Médica Mutua Madrileña, the Ministerio de Ciencia e Innovación (SAF2008-02971) (AS), XXX. We are grateful to all the donors for providing blood samples.

Web Resources

Phylotree: http://www.phylotree.org/
GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

References

1. Salas, A., Richards, M., Lareu, M.V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V., and Carracedo, Á. (2004). The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74, 454-465.
2. Veeramah, K.R., Connell, B.A., Pour, N.A., Powell, A., Plaster, C.A., Zeitlyn, D., Mendell, N.R., Weale, M.E., Bradman, N., and Thomas, M.G. (2010). Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. BMC Evol Biol 10, 92.
3. Salas, A., Carracedo, Á., Richards, M., and Macaulay, V. (2005). Charting the Ancestry of African Americans. Am J Hum Genet 77, 676-680.
4. Stefflova, K., Dulik, M.C., Barnholtz-Sloan, J.S., Pai, A.A., Walker, A.H., and Rebbeck, T.R. (2011). Dissecting the within-Africa ancestry of populations of African descent in the Americas. PLoS One 6, e14495.
5. Malyarchuk, B.A., and Czarny, J. (2005). African DNA lineages in the mitochondrial gene pool of Europeans. Mol Biol 39, 703-709.

6. Malyarchuk, B.A., Derenko, M., Perkova, M., Grzybowski, T., Vanecek, T., and Lazur, J. (2008). Reconstructing the phylogeny of African mitochondrial DNA lineages in Slavs. Eur J Hum Genet 16, 1091-1096.
7. Larruga, J.M., Diez, F., Pinto, F.M., Flores, C., and Gonzalez, A.M. (2001). Mitochondrial DNA characterisation of European isolates: the Maragatos from Spain. Eur J Hum Genet 9, 708-716.
8. Pereira, L., Prata, M.J., and Amorim, A. (2000). Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann Hum Genet 64, 491-506.
9. Álvarez-Iglesias, V., Mosquera-Miguel, A., Cerezo, M., Quintáns, B., Zarrabeitia, M.T., Cuscó, I., Lareu, M.V., García, O., Pérez-Jurado, L., Carracedo, Á., et al. (2009). New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE 4, e5112.
10. Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., Calderon, F.L., Simionati, B., Valle, G., Richards, M., et al. (2001). Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69, 1348-1356.
11. Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. Genetics 172, 373-387.
12. Brisighelli, F., Capelli, C., Álvarez-Iglesias, V., Onofri, V., Paoli, G., Tofanelli, S., Carracedo, Á., Pascali, V.L., and Salas, A. (2009). The Etruscan timeline: a recent Anatolian connection. Eur J Hum Genet 17, 693-696.
13. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23, 147.
14. Salas, A., Richards, M., De la Fé, T., Lareu, M.V., Sobrino, B., Sánchez-Diz, P., Macaulay, V., and Carracedo, Á. (2002). The making of the African mtDNA landscape. Am J Hum Genet 71, 1082-1111.
15. Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., and Villems, R. (2004). Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75, 752-770.
16. Salas, A., Torroni, A., Richards, M., Quintana-Murci, L., Hill, C., Macaulay, V., and Carracedo, Á. (2004). The phylogeography of mitochondrial DNA haplogroup L3g in Africa and the Atlantic slave trade. Am J Hum Genet 75, 524-526.
17. Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. Am J Hum Genet 82, 1130-1140.

18. Černý, V., Fernandes, V., Costa, M.D., Hájek, M., Mulligan, C.J., and Pereira, L. (2009). Migration of Chadic speaking pastoralists within Africa based on population structure of Chad Basin and phylogeography of mitochondrial L3f haplogroup. BMC Evol Biol 9, 63.
19. Černý, V., Salas, A., Hájek, M., Žaloudková, M., and Brdička, R. (2007). A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome. Ann Hum Genet 71, 433-452.
20. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30, E386-394.
21. Salas, A., Carracedo, Á., Macaulay, V., Richards, M., and Bandelt, H.-J. (2005). A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335, 891-899.
22. Bandelt, H.-J., Salas, A., and Lutz-Bonengel, S. (2004). Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118, 267-273.
23. Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496-2497.
24. Nei, N. (1987). Molecular evolutionary genetics.(New York: Columbi University Press).
25. Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437-460.
26. Tajima, F. (1993). Unbiased estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 10, 677-688.
27. Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., et al. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1, 273-280.
28. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945-959.
29. Amigo, J., Phillips, C., Salas, A., and Carracedo, A. (2009). Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. BMC Bioinformatics 10, S5.
30. Amigo, J., Salas, A., Phillips, C., and Carracedo, A. (2008). SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9, 428.
31. Loogväli, E.-L., Roostalu, U., Malyarchuk, B.A., Derenko, M.V., Kivisild, T., Metspalu, E., Tambets, K., Reidla, M., Tolk, H.-V., Parik, J., et al. (2004). Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21, 2012-2021.

32. Roostalu, U., Kutuev, I., Loogväli, E.-L., Metspalu, E., Tambets, K., Reidla, M., Khusnutdinova, E.K., Usanga, E., Kivisild, T., and Villems, R. (2007). Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol 24, 436-448.
33. Brandstätter, A., Salas, A., Niederstätter, H., Gassner, C., Carracedo, Á., and Parson, W. (2006). Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27, 2541-2550.
34. Achilli, A., Olivieri, A., Pala, M., Metspalu, E., Fornarino, S., Battaglia, V., Accetturo, M., Kutuev, I., Khusnutdinova, E., Pennarun, E., et al. (2007). Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. Am J Hum Genet 80, 759-768.
35. Saillard, J., Forster, P., Lynnerup, N., Bandelt, H.-J., and Nørby, S. (2000). mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67, 718-726.
36. Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M.B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84, 740-759.

Table 1. Diversity indices of L1b HVS-I sequences (sequence range 16090 to 16365) from Europe and different African regions

Population	*N*	*K*	*k/N*	*S*	*H*	*π*	*M*
Europe	32	17	0.531	20	0.927± 0.028	0.00999±0.0015	2.7
Africa							
North-East	1	–	–	–	–	–	–
North-West	89	26	0.291	26	0.757±0.048	0.00600±0.0009	1.7
Senegambia	5	–	–	–	–	–	–
Bight of Biafra	151	36	0.238	35	0.815±0.029	0.00698±0.0006	1.9
Gold Coast	37	9	0.243	10	0.805±0.046	0.00583±0.0058	1.6
South-West	21	8	0.381	10	0.724±0.101	0.00583±0.0015	1.6
South-East	6	–		–	–	–	–
East	19	10	0.526	10	0.842 ±0.070	0.00593±0.0011	1.6
South	2	–		–	–	–	–

Legend to the figures

Figure 1. (A) Spatial haplogroup distribution of sub-Saharan African lineages (macro-haplogroup L and haplogroup L1b) in Europe based on control region data. Pink dots indicate the sampled regions (see Supplementary Data XXX). (B) Distribution of African haplogroup frequencies in Europe. (C) Admixture components of L-European lineages in Africa; and (D) Structure triangle and bar plot indicating ancestral components of L-European mtDNA carriers based on profiles derived from a set of 34 AIMs.

Figure 2. Maximum parsimony tree of entire L1b mtDNA genomes.

Figure 1.

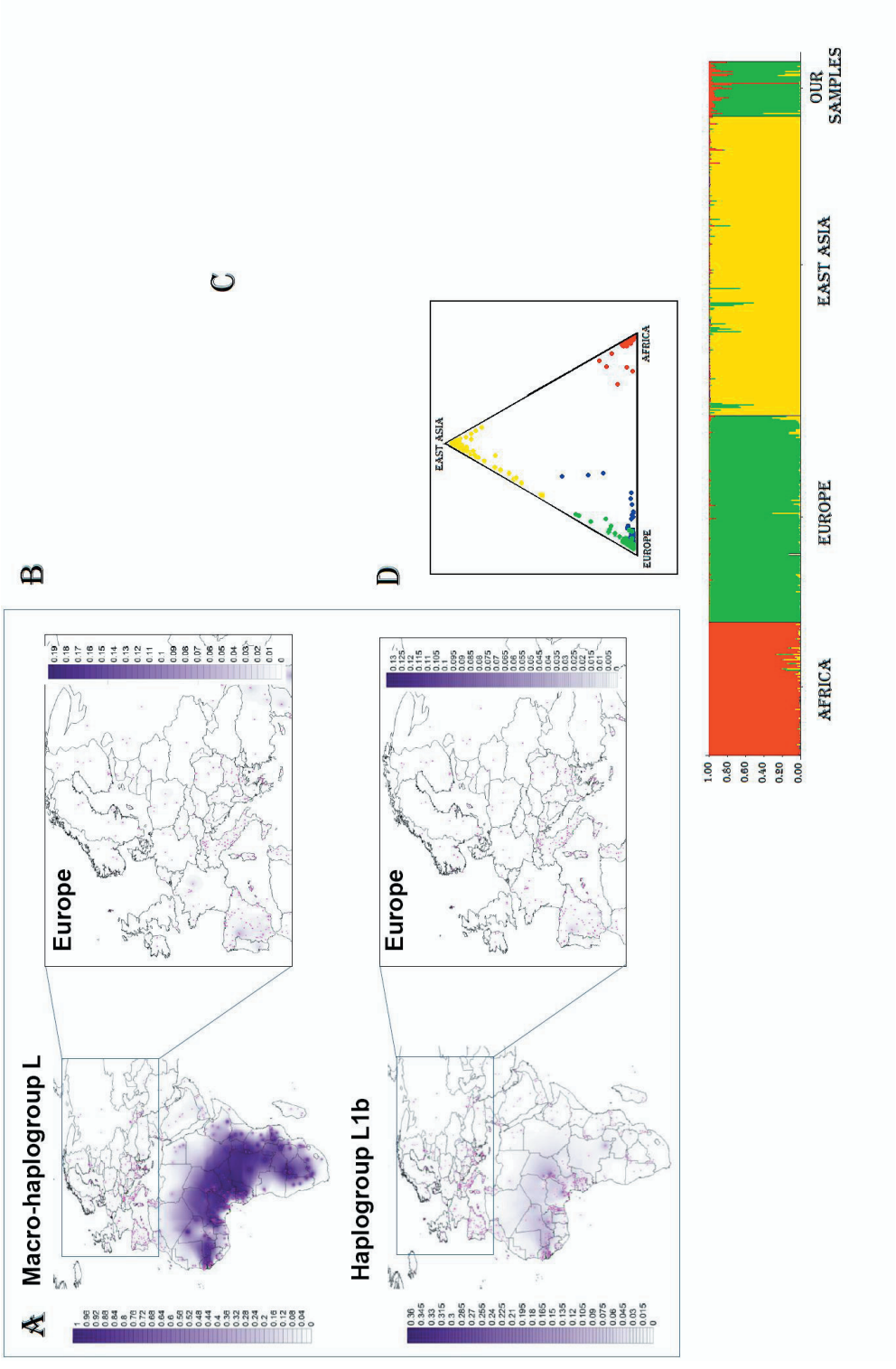
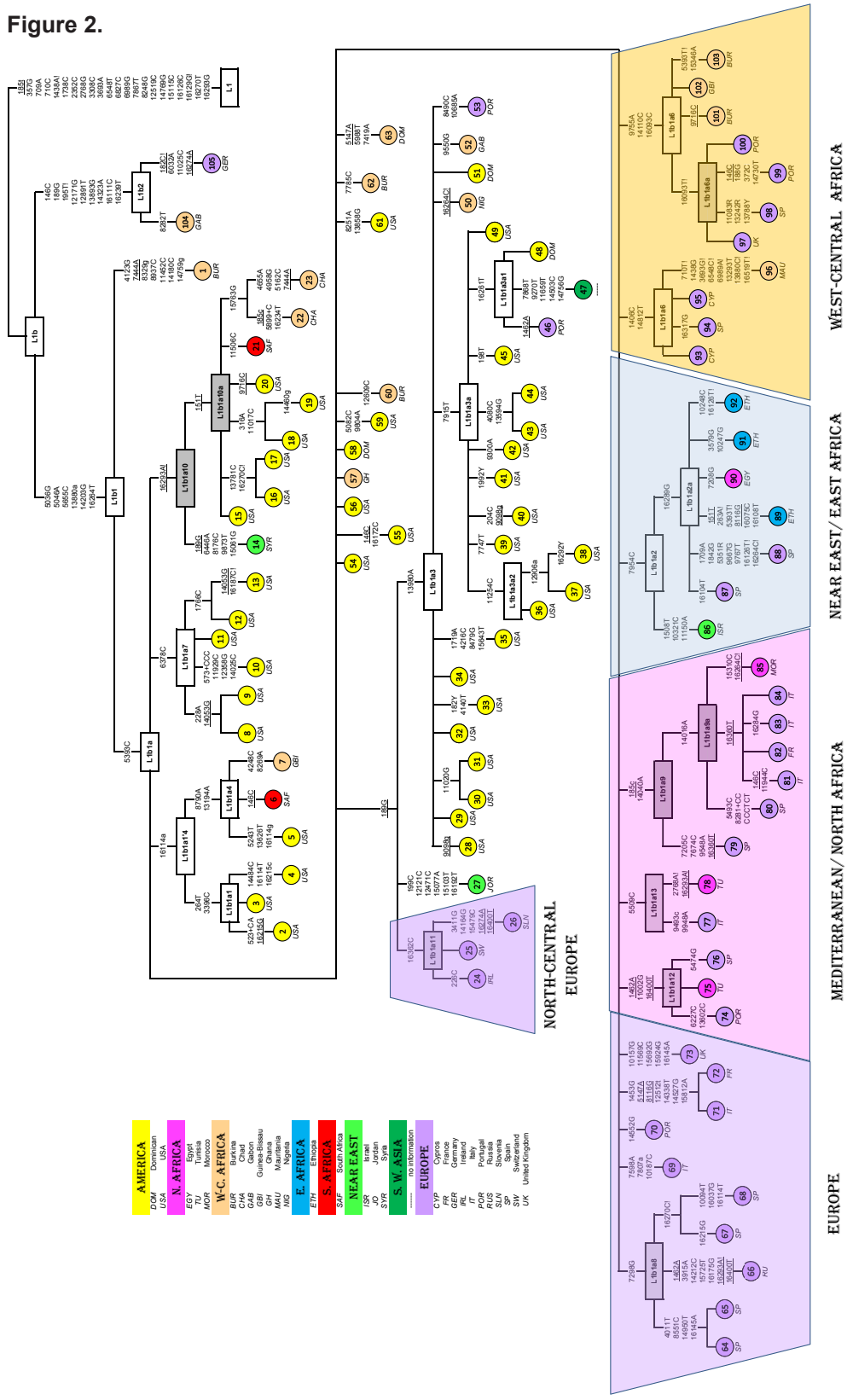

A
Macro-haplogroup L
Europe
Haplogroup L1b
Europe
B
C
D
EAST ASIA
AFRICA
EUROPE
1.00
0.80
0.60
0.40
0.20
0.00
AFRICA
EUROPE
EAST ASIA
OUR SAMPLES

Figure 2.


AMERICA
DOM Dominican
USA USA
N. AFRICA
EGY Egypt
TU Tunisia
MOR Morocco
W-C. AFRICA
BUR Burkina
CHA Chad
GAB Gabon
GNB Guinea-Bissau
GH Ghana
MAU Mauritania
NIG Nigeria
E. AFRICA
ETH Ethiopia
S. AFRICA
SAF South Africa
NEAR EAST
ISR Israel
JO Jordan
SYR Syria
S. W. ASIA
— no information
EUROPE
CYP Cyprus
FR France
GER Germany
IRL Ireland
IT Italy
POR Portugal
RUS Russia
SLV Slovakia
SP Spain
SW Switzerland
UK United Kingdom
L1b
L1b1
L1b1a
L1b1a14
L1b1a1
L1b1a4
L1b1a10
L1b1a10a
L1b1a7
L1b1a3
L1b1a11
L1b1a12
L1b1a13
L1b1a9
L1b1a8a
L1b1a8
L1b1a6
L1b1a6a
L1b1a6a1
L1b1a2
L1b1a2a
L1b1a3
L1b1a3a
L1b1a3a1
L1b1a2a2
L1b2
L1
NORTH-CENTRAL EUROPE
MEDITERRANEAN/ NORTH AFRICA
NEAR EAST/ EAST AFRICA
WEST-CENTRAL AFRICA
EUROPE

Supplemental Data

Title: Reconstructing mtDNA bridges used for African Diasporas into Europe

Authors: Cerezo et al

Table S1. Origin of the entire L-mtDNA genomes from Europe used in the present study

Five genomes were selected from a sample set from Galicia (north-west Spain) [1; 2] and three from a sample set from Catalonia (north-east Spain) [2] for which the HVS-I was already reported). Two additional Galician samples were selected from a cohort of Noonan patients [3], one from a cohort of Galician oligoastenospermic patients (unplublihed), and other two were taken from our local databank of Galician donors. Six DNAs belonged to Andalucian donors (South Spain; unpublished dataset). In addition, seven samples were selected from a dataset representing different regions in Italy, Spain, Portugal, United Kingdom, Ireland, Hungary, Germany, Slovenia, and Switzerland (unpublished data). Information concerning the geographical location and GenBank accession numbers of all these genomes is also provided.

ID Fig. S1	GenBank	Reference	Country	Sub-	HG	ID Fig. 2
1	XXXXXXXX	present study	Spain	–	L1b1a2a	87
2	XXXXXXXX	present study	Spain	Galicia	L1b1a2a	88
3	XXXXXXXX	present study	Portugal	–	L1b1a3a	53
4	EU092713	[4]	Portugal	–	L1b1a3a1	46
5	XXXXXXXX	present study	Ireland	–	L1b1a11	24
6	XXXXXXXX	present study	Switzerland	–	L1b1a11	25
7	XXXXXXXX	present study	Slovenia	–	L1b1a11	26
8	EU092930	[4]	Cypros	–	L1b1a5	93
9	XXXXXXXX	present study	Spain	Galicia	L1b1a5	94
10	EU092931	[4]	Cypros	–	L1b1a5	95
11	XXXXXXXX	present study	Wales	–	L1b1a6a	97
12	XXXXXXXX	present study	Spain	Galicia	L1b1a6a	98
13	XXXXXXXX	present study	Portugal	–	L1b1a6a	99
14	XXXXXXXX	present study	Portugal	–	L1b1a6a	100
15	XXXXXXXX	present study	Spain	Andalucia	L1b1a8	64
16	XXXXXXXX	present study	Spain	Andalucia	L1b1a8	65
17	EU200764	[5]	Russia	–	L1b1a8	66
18	XXXXXXXX	present study	Spain	Andalucia	L1b1a8	67
19	XXXXXXXX	present study	Spain	Galicia	L1b1a8	68
20	XXXXXXXX	present study	Italy	Sardinia	L1b1a	77
21	XXXXXXXX	present study	Italy	Sicilia	L1b1a	69
22	XXXXXXXX	present study	Portugal	–	L1b1a12	74
23	XXXXXXXX	present study	Spain	Catalonia	L1b1a12	76
24	XXXXXXXX	present study	Portugal	–	L1b1a	70

25	XXXXXXXX	present study	Italy	–	L1b1a	71
26	XXXXXXXX	present study	France	–	L1b1a	72
27	XXXXXXXX	present study	England	–	L1b1a	73
28	XXXXXXXX	present study	Spain	Catalonia	L1b1a9	79
29	XXXXXXXX	present study	Spain	Galicia	L1b1a9a	80
30	XXXXXXXX	present study	Italy	Tuscany	L1b1a9a	81
31	XXXXXXXX	present study	France	–	L1b1a9a	82
32	XXXXXXXX	present study	Italy	Marche	L1b1a9a	83
33	XXXXXXXX	present study	Central Italy	–	L1b1a9a	84
34	XXXXXXXX	present study	Germany	–	L1b2	105
35	EU092712	[4]	Portugal	–	L1c2b1	–
36	XXXXXXXX	present study	Central Italy	–	L2a1a2	–
37	XXXXXXXX	present study	Italy	Marche	L2a1a2	–
38	EU092711	[4]	Portugal	–	L2a1a3	–
39	EU200762	[5]	Slovakia	–	L2a1c3	–
40	XXXXXXXX	present study	Spain	Andalucia	L2a1c4	–
41	XXXXXXXX	present study	Spain	Galicia	L2a1c6	–
42	EF177417	[6]	Portugal	–	L2a1c6	–
43	XXXXXXXX	present study	Spain	Galicia	L2a1c	–
44	EU200763	[5]	Slovakia	–	L2a1k	–
45	EU200760	[5]	Czech Republic	–	L2a1k	–
46	HQ384198	[7]	Spain	Galicia	L2a1	–
47	HQ384199	[7]	Spain	Galicia	L2a5	–
48	XXXXXXXX	present study	Italy	Liguria	L2b1a1	–
49	XXXXXXXX	present study	Italy	Liguria	L2b1a1	–
50	XXXXXXXX	present study	Spain	Galicia	L2b3	–
51	XXXXXXXX	present study	Spain	Andalucia	L2c	–
52	EU092710	[4]	Netherlands	–	L2c2	–
53	XXXXXXXX	present study	Spain	Catalonia	L2c3	–
54	EU200761	[5]	Russia	–	L3b1b1	–
55	XXXXXXXX	present study	Italy	Sicilia	L3b	–
56	XXXXXXXX	present study	Spain	Andalucia	L3f1b	–
57	EU200759	[5]	Poland	–	L3d1b1	–
58	XXXXXXXX	present study	Central Italy	–	L3d1b1a	–
59	XXXXXXXX	present study	Italy	Tuscany	L3d1b1a	–
60	XXXXXXXX	present study	Italy	Campania	L3d1b1a	–
61	XXXXXXXX	present study	Italy	Tuscany	L3d1b1a	–
62	XXXXXXXX	present study	Central Italy	–	L3d1b1a	–
63	XXXXXXXX	present study	Italy	Sicilia	L3d	–
64	XXXXXXXX	present study	Italy	Puglia	L3e1f	–
65	XXXXXXXX	present study	Italy	Marche	L3e2b4	–
66	XXXXXXXX	present study	Italy	–	L3e2b4	–
67	XXXXXXXX	present study	Hungary	–	L3e5	–
68	XXXXXXXX	present study	Spain	Galicia	L3x2b	–
69	XXXXXXXX	present study	Spain	Galicia	L3h1b1a1	–

Figure S1. Maximum parsimony tree of entire L-mtDNA genomes from Europe used in the present study

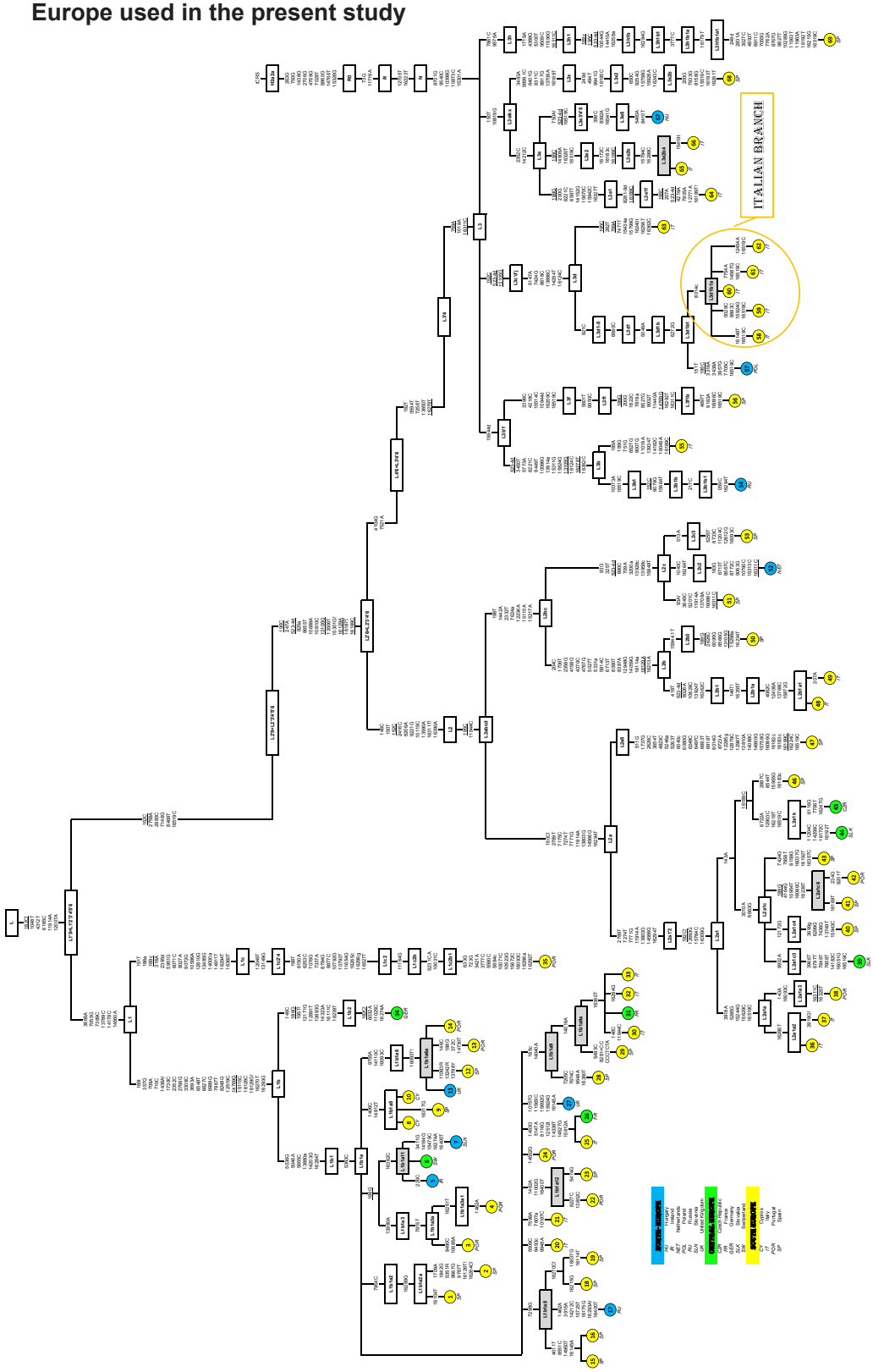


Table S2. Origin of the entire L2a1-mtDNA genomes used in the present study

Figure S2. Maximum parsimony tree of L2a1 entire genomes

The yellow square highlights the L2 sub-clade that is most likely of European origin, L2a1k. Sample codes are as indicated in Table S3.

Table S3. Origin of the entire genomes belonging to haplogroup L1b

ID Fig. 2	GenBank	Continental region	Country	Reference	HG	ID Fig. S1
1	DQ112737	West-Central Africa	Burkina	[8]	L1b1	–
2	–	America	USA	[4; 9]	L1b1a1	–
3	–	America	USA	[10]	L1b1a1	–
4	–	America	USA	[4; 9]	L1b1a1	–
5	–	America	USA	[4; 9]	L1b1a4	–
6	AY195783	South Africa	South Africa	[11]	L1b1a4	–
7	EU092715	West-Central Africa	Guinea-Bissau	[4]	L1b1a4	–
8	DQ282505	America	USA	[12]	L1b1a7	–
9	DQ282506	America	USA	[12]	L1b1a7	–
10	DQ304921	America	USA	[12]	L1b1a7	–
11	DQ304923	America	USA	[12]	L1b1a7	–
12	–	America	USA	[10]	L1b1a7	–
13	–	America	USA	[4; 9]	L1b1a7	–
14	EU092737	Near East	Syria	[4]	L1b1a10	–
15	–	America	USA	[4; 9]	L1b1a10a	–
16	–	America	USA	[4; 9]	L1b1a10a	–
17	–	America	USA	[10]	L1b1a10a	–
18	–	America	USA	[4; 9]	L1b1a10a	–
19	–	America	USA	[10]	L1b1a10a	–
20	–	America	USA	[4; 9]	L1b1a10a	–
21	EU092854	South Africa	South Africa	[4]	L1b1a10a	–
22	EU092886	West-Central Africa	Chad	[4]	L1b1a10a	–
23	EU092895	West-Central Africa	Chad	[4]	L1b1a10a	–
24	XXXXXXXX	Europe	Ireland	present study	L1b1a11	5
25	XXXXXXXX	Europe	Switzerland	present study	L1b1a11	6
26	XXXXXXXX	Europe	Slovenia	present study	L1b1a11	7
27	EU092755	Near East	Jordan	[4]	L1b1a	–
28	–	America	USA	[10]	L1b1a3	–
29	DQ304908	America	USA	[12]	L1b1a3	–
30	DQ304917	America	USA	[12]	L1b1a3	–
31	DQ304916	America	USA	[12]	L1b1a3	–
32	–	America	USA	[10]	L1b1a3	–
33	DQ304909	America	USA	[12]	L1b1a3	–

34	DQ304910	America	USA	[12]	L1b1a3	–
35	DQ304907	America	USA	[12]	L1b1a3	–
36	DQ304906	America	USA	[12]	L1b1a3a2	–
37	DQ304911	America	USA	[12]	L1b1a3a2	–
38	DQ304918	America	USA	[12]	L1b1a3a2	–
39	DQ304915	America	USA	[12]	L1b1a3a	–
40	–	America	USA	[4; 9]	L1b1a3a	–
41	DQ304912	America	USA	[12]	L1b1a3a	–
42	–	America	USA	[4; 9]	L1b1a3a	–
43	DQ304905	America	USA	[12]	L1b1a3a	–
44	DQ304913	America	USA	[12]	L1b1a3a	–
45	–	America	USA	[4; 9]	L1b1a3a	–
46	EU092713	Europe	Portugal	[4]	L1b1a3a1	4
47	DQ112881	South West Asia	no information	[8]	L1b1a3a1	–
48	DQ112690	America	Dominican	[8]	L1b1a3a1	–
49	DQ304914	America	USA	[12]	L1b1a3a	–
50	AF346986	West-Central Africa	Nigeria	Ingman 2000	L1b1a3	–
51	DQ112691	America	Dominican	[8]	L1b1a3	–
52	HM771163	West-Central Africa	Gabon	[13]	L1b1a3	–
53	XXXXXXXX	Europe	Portugal	present study	L1b1a3	3
54	DQ304919	America	USA	[12]	L1b1a	–
55	–	America	USA	[4; 9]	L1b1a	–
56	–	America	USA	[10]	L1b1a	–
57	DQ112829	West-Central Africa	Ghana	[8]	L1b1a	–
58	DQ112693	America	Dominican	[8]	L1b1a	–
59	DQ304920	America	USA	[12]	L1b1a	–
60	DQ112727	West-Central Africa	Burkina	[8]	L1b1a	–
61	DQ304922	America	USA	[12]	L1b1a	–
62	DQ112733	West-Central Africa	Burkina	[8]	L1b1a	–
63	DQ112692	America	Dominican	[8]	L1b1a	–
64	XXXXXXXX	Europe	Spain	present study	L1b1a8	15
65	XXXXXXXX	Europe	Spain	present study	L1b1a8	16
66	EU200764	Europe	Russia	[5]	L1b1a8	17
67	XXXXXXXX	Europe	Spain	present study	L1b1a8	18
68	XXXXXXXX	Europe	Spain	present study	L1b1a8	19
77	XXXXXXXX	Europe	Italy	present study	L1b1a	20
69	XXXXXXXX	Europe	Italy	present study	L1b1a	21
74	XXXXXXXX	Europe	Portugal	present study	L1b1a12	22
75	FJ460537	North Africa	Tunisia	[14]	L1b1a12	–
76	XXXXXXXX	Europe	Spain	present study	L1b1a12	23
70	XXXXXXXX	Europe	Portugal	present study	L1b1a	24
71	XXXXXXXX	Europe	Italy	present study	L1b1a	25
72	XXXXXXXX	Europe	France	present study	L1b1a	26
73	XXXXXXXX	Europe	England	present study	L1b1a	27
78	FJ460522	North Africa	Tunisia	[14]	L1b1a13	–
79	XXXXXXXX	Europe	Spain	present study	L1b1a9	28

80	XXXXXXXX	Europe	Spain	present study	L1b1a9a	29
81	XXXXXXXX	Europe	Italy	present study	L1b1a9a	30
82	XXXXXXXX	Europe	France	present study	L1b1a9a	31
83	XXXXXXXX	Europe	Italy	present study	L1b1a9a	32
84	XXXXXXXX	Europe	Central Italy	present study	L1b1a9a	33
85	EU092667	Near East	Israel	[4]	L1b1a9a	–
86	EU092672	Near East	Israel	[4]	L1b1a2	–
87	XXXXXXXX	Europe	Spain	present study	L1b1a2a	1
88	XXXXXXXX	Europe	Spain	present study	L1b1a2a	2
89	EU092942	East Africa	Ethiopia	[4]	L1b1a2a	–
90	EU092775,	North Africa	Egypt	[4]	L1b1a2a	–
91	EU092952	East Africa	Ethiopia	[4]	L1b1a2a	–
92	EU092950	East Africa	Ethiopia	[4]	L1b1a2a	–
93	EU092930	Europe	Cypros	[4]	L1b1a5	8
95	XXXXXXXX	Europe	Spain	present study	L1b1a5	9
95	EU092931	Europe	Cypros	[4]	L1b1a5	10
96	AF381994	West-Central Africa	Mauritania	[15]	L1b1a5	–
97	XXXXXXXX	Europe	Wales	present study	L1b1a6a	11
98	XXXXXXXX	Europe	Spain	present study	L1b1a6a	12
99	XXXXXXXX	Europe	Portugal	present study	L1b1a6a	13
100	XXXXXXXX	Europe	Portugal	present study	L1b1a6a	14
101	DQ112759	West-Central Africa	Burkina	[8]	L1b1a6	–
101	EU092716	West-Central Africa	Guinea-Bissau	[4]	L1b1a6	–
103	DQ112741	West-Central Africa	Burkina	[8]	L1b1a6	–
104	HM771162	West-Central Africa	Gabon	[13]	L1b2	–
105	XXXXXXXX	Europe	Germany	present study	L1b2	34

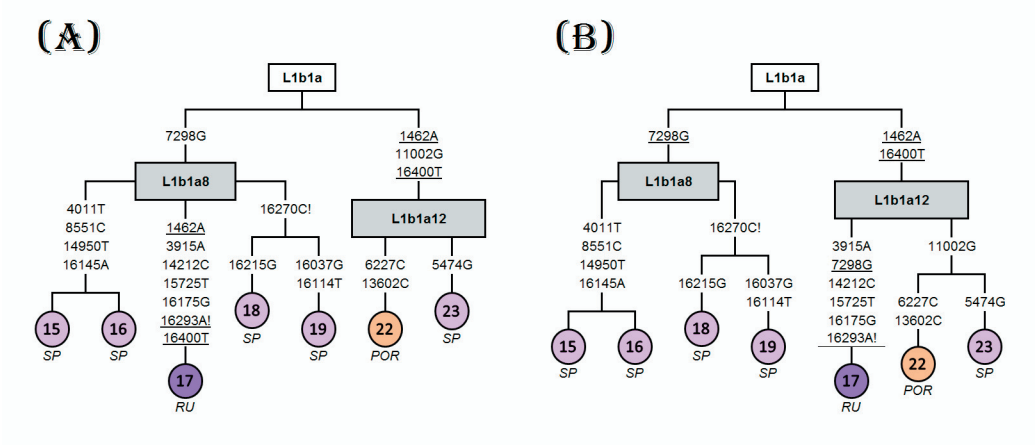
Table S4. List of AIMs genotyped in the present study for the European samples carrying L-haplogroup lineages. Divergence for all the SNPs is also indicated.

	SNP	Alleles	Divergence	Accumulated
1	rs2814778	C/T	0.64	0.64
2	rs1426654	C/T	0.603	1.243
3	rs16891982	C/G	0.5916	1.8346
4	rs773658	C/G	0.3966	2.2312
5	rs881929	G/T	0.3934	2.6246
6	rs239031	C/T	0.3925	3.0171
7	rs1335873	A/T	0.3325	3.3495
8	rs1573020	A/G	0.3251	3.6746
9	rs2065982	A/G	0.2967	3.9713
10	rs4540055	A/C/T	0.287	4.2583
11	rs730570	C/T	0.2867	4.545
12	rs12913832	A/G	0.27	4.815
13	rs2026721	A/G	0.2638	5.0788
14	rs2303798	C/T	0.2557	5.3345
15	rs2040411	A/G	0.2294	5.5639
16	rs1978806	C/T	0.2263	5.7903
17	rs2572307	A/G	0.2161	6.0064
18	rs1321333	C/T	0.2137	6.2201
19	rs3785181	C/T	0.2072	6.4273
20	rs5997008	A/C	0.1956	6.623
21	rs5030240	A/C/G	0.1893	6.8123
22	rs722098	A/G	0.1813	6.9936
23	rs182549	C/T	0.1622	7.1558
24	rs917118	A/G	0.1569	7.3127
25	rs2065160	A/G	0.1192	7.4319
26	rs727811	A/C	0.1181	7.55
27	rs10843344	C/T	0.1103	7.6603
28	rs1886510	A/G	0.1067	7.7669
29	rs10141763	A/T	0.1045	7.8714
30	rs896788	C/T	0.0929	7.9644
31	rs1024116	A/G	0.0885	8.0529
32	rs1498444	A/C	0.0879	8.1408
33	rs7897550	C/T	0.0675	8.2083
34	rs2304925	G/T	0.0259	8.2341

Table S5. Three-way prediction of ancestral origin (Africa, Europe, and East Asia) for the European samples that carry L-mtDNAs. SNP data was from HapMap populations were used as training sets, including Africans (Yoruba), Europeans (CEU), and East Asians (Chinese and Japanese). Prediction was based on maximum likelihood and the genotyping profiles for the AIMs listed in Table S4. Classification was carried out using http://mathgene.usc.es/snipper/ as reported in [16]. Sample ID codes are as in Figure 2.

Sample ID	Europe	Africa	East Asia	Prediction
1	3.08E+01	7.56E+01	6.71E+01	European
2	2.06E+01	7.27E+01	5.41E+01	European
3	4.20E+01	6.93E+01	4.51E+01	European
5	2.16E+01	8.77E+01	5.83E+01	European
6	2.51E+01	7.90E+01	5.49E+01	European
7	2.40E+01	7.68E+01	6.17E+01	European
9	2.33E+01	8.15E+01	5.92E+01	European
11	2.28E+01	8.11E+01	6.34E+01	European
12	2.28E+01	6.61E+01	4.75E+01	European
14	2.31E+01	8.44E+01	5.73E+01	European
15	1.98E+01	6.85E+01	4.90E+01	European
18	2.24E+01	8.33E+01	5.70E+01	European
19	2.87E+01	7.30E+01	6.63E+01	European
20	2.56E+01	7.30E+01	6.36E+01	European
22	3.18E+01	6.37E+01	5.33E+01	European
23	2.48E+01	7.15E+01	5.53E+01	European
24	3.15E+01	8.81E+01	6.59E+01	European
25	2.66E+01	8.79E+01	6.55E+01	European
26	2.08E+01	7.38E+01	5.39E+01	European
27	2.23E+01	7.33E+01	5.19E+01	European
28	1.59E+01	5.68E+01	4.83E+01	European
30	3.15E+01	6.29E+01	5.82E+01	European
31	2.68E+01	7.72E+01	6.53E+01	European
32	2.09E+01	8.15E+01	5.56E+01	European
33	3.15E+01	6.29E+01	5.82E+01	European
34	8.49E+01	2.42E+01	7.34E+01	AFRICA
36	2.07E+01	7.25E+01	4.55E+01	European
40	2.35E+01	8.40E+01	6.38E+01	European
43	2.28E+01	8.50E+01	6.06E+01	European
46	2.35E+01	8.40E+01	6.38E+01	European
16	2.09E+01	7.92E+01	5.86E+01	European
47	4.01E+01	6.88E+01	4.90E+01	European
53	4.05E+01	6.68E+01	5.18E+01	European
50	2.31E+01	8.25E+01	6.14E+01	European
56	4.06E+01	6.68E+01	5.18E+01	European
58	2.21E+01	5.34E+01	4.34E+01	European
60	2.14E+01	6.46E+01	5.31E+01	European
61	1.99E+01	8.69E+01	6.47E+01	European
62	2.40E+01	7.80E+01	4.66E+01	European
66	2.16E+01	9.41E+01	7.19E+01	European
68	2.02E+01	8.70E+01	5.56E+01	European
69	2.32E+01	5.31E+01	4.46E+01	European

Figure S3. Alternative phylogenies for haplogroup L1b1a8 and L1b1a12



References

1. Quintáns, B., Álvarez-Iglesias, V., Salas, A., Phillips, C., Lareu, M.V., and Carracedo, Á. (2004). Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140, 251-257.
2. Álvarez-Iglesias, V., Mosquera-Miguel, A., Cerezo, M., Quintáns, B., Zarrabeitia, M.T., Cuscó, I., Lareu, M.V., García, O., Pérez-Jurado, L., Carracedo, Á., et al. (2009). New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE 4, e5112.
3. Gómez-Carballa, A., Cerezo, M., Balboa, E., Heredia, C., Castro-Feijóo, L., Rica, I., Barreiro, J., Eirís, J., Cabanas, P., Martínez-Soto, I., et al. (2011). Evolutionary analyses of entire genomes does not support the association of mtDNA mutations with Ras/MAPK pathway syndromes. Mol Biol Evol; submitted.
4. Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. Am J Hum Genet 82, 1130-1140.
5. Malyarchuk, B.A., Derenko, M., Perkova, M., Grzybowski, T., Vanecek, T., and Lazur, J. (2008). Reconstructing the phylogeny of African mitochondrial DNA lineages in Slavs. Eur J Hum Genet 16, 1091-1096.
6. Pereira, L., Goncalves, J., Franco-Duarte, R., Silva, J., Rocha, T., Arnold, C., Richards, M., and Macaulay, V. (2007). No evidence for an mtDNA role in sperm motility: data from complete sequencing of asthenozoospermic males. Mol Biol Evol 24, 868-874.
7. Gómez-Carballa, A., Cerezo, M., Balboa, E., Heredia, C., Castro-Feijóo, L., Rica, I., Barreiro, J., Eirís, J., Cabanas, P., Martínez-Soto, I., et al. (2011). Evolutionary analyses of entire genomes does not support the association of mtDNA mutations with Ras/MAPK pathway syndromes. PLoS One, in press.

8. Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. Genetics 172, 373-387.
9. Howell, N., Elson, J.L., Turnbull, D.M., and Herrnstadt, C. (2004). African Haplogroup L mtDNA sequences show violations of clock-like evolution. Mol Biol Evol 21, 1843-1854.
10. Herrnstadt, C., Elson, J.L., Fahy, E., Preston, G., Turnbull, D.M., Anderson, C., Ghosh, S.S., Olefsky, J.M., Beal, M.F., Davis, R.E., et al. (2002). Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences from the major African, Asian, and European haplogroups. Am J Hum Genet 70, 1152-1171.
11. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100, 171-176.
12. Just, R.S., Diegoli, T.M., Saunier, J.L., Irwin, J.A., and Parsons, T.J. (2008). Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. Forensic Sci Int Genet 2, e45-48.
13. Batini, C., Lopes, J., Behar, D.M., Calafell, F., Jorde, L.B., van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D. (2011). Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol Biol Evol 28, 1099-1110.
14. Costa, M.D., Cherni, L., Fernandes, V., Freitas, F., Ammar El Gaaied, A.B., and Pereira, L. (2009). Data from complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association and the "golden mean" to longevity. Mech Ageing Dev 130, 222-226.
15. Maca-Meyer, N., González, A.M., Larruga, J.M., Flores, C., and Cabrera, V.M. (2001). Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2, 13.
16. Phillips, C., Salas, A., Sánchez, J.J., Fondevila, M., Gómez-Tato, A., Álvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., et al. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1, 273-280.

IV.2 FORENSIC GENETICS

IV.2.1. Article 9: 2006 GEP-ISFG collaborative exercise on mtDNA: reflections about interpretation, artefacts, and DNA mixtures *Forensic Science International: Genetics*

IV.2.2. Article 10: Case Report: Identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur *Forensic Science International: Genetics*

IV.2.3. Article 11: Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples F*orensic Science International: Genetics Supplement Series*

IV.2.4. Article 12: Testing the performance of mtSNP minisequencing in forensic samples *Forensic Science International Genetics*

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 2 (2008) 126–133
FSI GENETICS
www.elsevier.com/locate/fsig

2006 GEP-ISFG collaborative exercise on mtDNA: reflections about interpretation, artefacts, and DNA mixtures

L. Prieto [a,*], A. Alonso [b], C. Alves [c], M. Crespillo [d], M. Montesino [a], A. Picornell [e], A. Brehm [f], J.L. Ramírez [g], M.R. Whittle [h], M.J. Anjos [i], I. Boschi [j], J. Buj [k], M. Cerezo [l], S. Cardoso [m], R. Cicarelli [n], D. Comas [o], D. Corach [p], C. Doutremepuich [q], R.M. Espinheira [r], I. Fernández-Fernández [s], S. Filippini [t], Julia Garcia-Hirschfeld [b], A. González [u], B. Heinrichs [v], A. Hernández [w], F.P.N. Leite [x], R.P. Lizarazo [y], A.M. López-Parra [z], M. López-Soto [a1], J.A. Lorente [a2], B. Mechoso [a3], I. Navarro [a4], S. Pagano [a5], J.J. Pestano [a6], J. Puente [a7], E. Raimondi [a8], A. Rodríguez-Quesada [a9], M.F. Terra-Pinheiro [a10], L. Vidal-Rioja [a11], C. Vullo [a12], A. Salas [l]

[a] *Comisaría General de Policía Científica, DNA Laboratory, Madrid, Spain*
[b] *Instituto Nacional de Toxicología y Ciencias Forenses, Departamento de Madrid, Spain*
[c] *Instituto de Patología e Inmunología Molecular IPATIMUP, Universidade do Porto, Portugal*
[d] *Instituto Nacional de Toxicología y Ciencias Forenses, Departamento de Barcelona, Spain*
[e] *Lab. Genètica, Ins. Universitari d'Investigacions en C.C. de la Salut i Dep. de Biol., Universitat Illes Balears, Spain*
[f] *Laboratorio de Genética Humana, Universidade da Madeira, Portugal*
[g] *Fundación Ins. de Estudios Avanzados, C. Biotecnología, U. Pol. Genéticos, Caracas, Venezuela*
[h] *Genomic Engenharia Molecular LTDA, Sao Paulo, Brazil*
[i] *Serviço de Genética e Biologia Forense, Inst. Nacional de Med. Legal, Delegação do Centro, Portugal*
[j] *Istituto di Medicina Legale e delle Assicurazioni, Università Cattolica del S. Cuore, Rome, Italy*
[k] *Neodiagnóstica, Lleida, Spain*
[l] *Unidade de Xenética, Ins. de Medicina Legal, Universidade de Santiago de Compostela, Galicia, Spain*
[m] *Banco de ADN, Universidad del País Vasco, Vitoria-Gasteiz, Spain*
[n] *Laboratorio de Investigaçáo de Paternidade, Universidade Estadual Paulista – UNESP, Sao Paulo, Brazil*
[o] *Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain*
[p] *Servicio de Huellas Digitales Genéticas, Fac. Farmacia y Bioquímica, Universidad de Buenos Aires, Argentina*
[q] *Laboratoire ÐHématologie, Bordeaux, France*
[r] *Serviço de Genética e Biologia Forense, Inst. Nacional de Medicina Legal, Delegação do Sul, Portugal*
[s] *DataGene, Parque Tecnológico de Derio, Bizkaia, Spain*
[t] *Banco Nacional de Datos Genéticos, Buenos Aires, Argentina*
[u] *ADF TecnoGen, Madrid, Spain*
[v] *Dirección General de la Guardia Civil, Madrid, Spain*
[w] *Instituto Nacional de Toxicología y Ciencias Forenses, Delegación de Canarias, La Laguna, Spain*
[x] *Laboratorio de Pericias, Instituto Geral de Pericias, Porto Alegre, Brazil*
[y] *Laboratorio de DNA, Inst. Nacional de Med. Legal y C.C. Forenses, Santafé de Bogotá, Colombia*
[z] *Departamento de Medicina Legal y Forense, Facultad de Medicina, Universidad Complutense, Madrid, Spain*
[a1] *Instituto Nacional de Toxicología y Ciencias Forenses, Departamento de Sevilla, Spain*
[a2] *Laboratorio de Identificación Genética, Dpto. de Medicina Legal, Universidad de Granada, Spain*
[a3] *Instituto de Genética Médica, Hospital Italiano, Montevideo, Uruguay*
[a4] *Centro de Análisis Genéticos C.A.G.T., Zaragoza, Spain*
[a5] *Laboratorio Biológico, Dirección Nacional de Policía Técnica, Montevideo, Uruguay*
[a6] *Laboratorio de Genética, Instituto Anatómico Forense, Las Palmas, Spain*
[a7] *Laboratorio de Genética Clínica LabGenetics, Madrid, Spain*
[a8] *PRICAI -Fundación Favaloro, Buenos Aires, Argentina*
[a9] *Unidad de Genética Forense, Sección de Bioquímica, Dep. de Ciencias Forenses, Heredia, Costa Rica*
[a10] *Serviço de Genética e Biologia Forense, Inst. Nacional de Med. Legal, Delegação do Norte, Portugal*
[a11] *Lab. de Identificación Genética, Ins. Multidisciplinario de Biol. Celular (IMBICE), La Plata, Argentina*
[a12] *Laboratorio de Inmunogenética y Diagnostico Molecular, Córdoba, Argentina*

Received 26 July 2007; received in revised form 11 September 2007; accepted 2 October 2007

* Corresponding author.
E-mail address: lourditasmt@ya.com (L. Prieto).

1872-4973/$ – see front matter
doi:10.1016/j.fsigen.2007.10.010

Abstract

We report the results of the seventh edition of the GEP-ISFG mitochondrial DNA (mtDNA) collaborative exercise. The samples submitted to the participant laboratories were blood stains from a maternity case and simulated forensic samples, including a case of mixture. The success rate for the blood stains was moderate (~77%); even though four inexperienced laboratories concentrated about one-third of the total errors. A similar success was obtained for the analysis of mixed samples (78.8% for a hair–saliva mixture and 69.2% for a saliva–saliva mixture). Two laboratories also dissected the haplotypes contributing to the saliva–saliva mixture. Most of the errors were due to reading problems and misinterpretation of electropherograms, demonstrating once more that the lack of a solid devised experimental approach is the main cause of error in mtDNA testing.

Keywords: mtDNA; Mixture; Recommendations: Phylogeny; GEP-ISFG; Collaborative exercise

1. Introduction

GEP-ISFG mtDNA collaborative exercises have been performed in the last seven years [1–5]. One of the aims of these exercises is to improve the quality and standardization of mtDNA analyses in both technical and interpretation issues. In the 2006 exercise, we analysed seven samples consisting of four blood stains (labelled M1 to M4) from a maternity case, and three simulated forensic samples, namely, a reference blood stain labelled as M5, an unknown colourless stain labelled as M6 and two unknown hair shafts labelled as M7. M6 was composed of a mixture of saliva from the donor of M5 and saliva from an unknown donor, whereas M7 consisted of two hair shafts from the above-mentioned unknown donor, but coated with saliva from the M5 donor. MtDNA haplotypes were unknown before the samples were submitted to the laboratories, so the donors were not deliberately selected.

2. Participating laboratories

The number of laboratories that analysed samples M1 to M7 is indicated in Table 1. Different DNA extraction methods, amplification and sequencing primers, purification strategies, and sequencing equipment were used among these laboratories. Only one laboratory carried out mtDNA quantification using real time-PCR [6]. As in previous exercises, there was not an apparent relation between the different technologies employed by the laboratories and the amount and type of errors.

3. Results regarding the analysis of blood stain samples

A summary of the results obtained for samples M1 to M5, including the consensus haplotypes and the number of laboratories reporting them is shown in Table 2. As a rule in this exercise, an out-of-consensus result in the report of a particular haplotype constitutes an error. For those samples that were analysed by at least five laboratories, 'consensus' means that at least 70% of them (rounding up decimals) report exactly the same result for a given sequence range, but the remaining 30% do not fully coincide and provide a different result to the consensus (the later only applies just in case this 30% includes at least three laboratories).

All laboratories indicated their sequence ranges; only two laboratories reported 'extended' haplotypes that included regions outside of classical HVS-I and HVS-II (16024–16365 and 72–576 in one case and 16024–16569 and 1–576 in the other).

Complete results are shown as supplementary material (Tables S1–S4). Although the global success rate for samples M1–M2–M3–M4–M5 was moderate (77.5%), it is important to highlight that a substantial amount of errors were concentrated in only four inexpert laboratories (laboratories 7, 9, 10 and 12 committed errors in all blood samples; see Tables S1–S3); the success rate excluding these laboratories was 88.2%. The GEP-ISFG group has the policy of allowing any laboratory to participate in its annual quality control (QC) exercise, even those that are still in the process of implementing or have recently implemented the mtDNA analysis technique in their routine work. As a consequence, most of the out-of-consensus results are provided by these inexperienced laboratories, a recurrent situation from previous editions of the GEP-ISFG QC. In contrast, 25 laboratories did not commit any error in samples M1–M2–M3–M4–M5, while seven laboratories reported sequences with only one inconsistency with respect to the consensus haplotypes.

4. Detecting DNA saliva–saliva and hair–saliva mixtures using mtDNA

The GEP-ISFG mtDNA working group was the first consortium analysing and interpreting mtDNA sequencing profiles from sample mixtures [7]. Several parameters influence the detection of nucleotide variants in mtDNA mixtures: (i) the type of tissues contributing to the mixture (related to different mtDNA copy number per cell in different tissues), (ii) differences in mtDNA content among donors, (iii) differences in the amount of fluid from each donor present in a mixture, and (iv) technical factors that can lead to undesirable interpretation artefacts, such as the varying signal throughout the electropherograms when using dye terminator chemistries.

The M6 sample consisted of a 50:50 mixture (80 μL) of saliva from the M5 donor plus saliva from an unknown donor. The presence of a DNA mixture in M6 sample was questioned to the participating laboratories. A total of 18 out of 26 laboratories reported the correct mixture haplotype, and only two laboratories dissected the mixture haplotype into the two potential contributing haplotypes, namely, one sequence belonging to haplogroup U5b (16051G 16189C 16270T 73G 146C 150T 263G 309.1C 315.1C) and another mtDNA probably belonging to haplogroup H (263G 315.1C). Dis-

Table 1
Number of participating laboratories in the mtDNA GEP-ISFG collaborative exercise of 2006–2007

Maternity case				Forensic case		
M1 (blood stain)	M2 (blood stain)	M3 (blood stain)	M4 (blood stain)	M5 (blood stain)	M6 (saliva mixture)	M7 (hair shaft contaminated with saliva)
36	36	40	36	36	26	33

Table 2
Consensus haplotypes and number of laboratories reporting them

Samples	Consensus haplotypes (16024–16365 and 73–340)	No. of laboratories/total laboratories (%)
M1–M2–M4	16188T 16311C 152C 263G 309.1C 315.1C	30/36 (83.3)
M3	93G 151T 263G 315.1C	33/40 (82.5)
M5	16051G 16189C 16270T 73G 146C 150T 263G 309.1C 315.1C	24/36 (66.7)
M6	16051R 16189Y 16270Y 73R 146Y 150Y 263G 309.1C 315.1C	18/26 (69.2)
M7	263G 315.1C	26/33 (78.8)

crepancies were mainly due to poor quality electropherograms, nomenclature problems, clerical errors or misinterpretation/misreading of the electropherograms (see Table 3).

Hair shafts are frequently covered with other fluids (like blood or vaginal fluid in rape cases) coming from a different donor, a fact that is often unperceived by the forensic analyst if a previous morphological study is not carried out. Therefore, it is recommended to wash the hair shaft before carrying out the DNA extraction in order to remove possible contaminant agents and to analyse both the liquid and hair samples separately. This procedure prevents the haplotype from the contaminating fluid to predominate or even mask the signal coming from the hair shaft, thus leading to a false exclusion. The hair shaft can be washed using a cotton swab or carrying out a preferential lysis [8].

The M7 sample consisted of two hair shafts from an unknown donor that were deliberately wet with saliva from the M5 donor, thus emulating a quite typical forensic sample. The aim of the analysis of the M7 sample was to know if the hair shafts could belong to the M5 donor. Twenty-six out of 33 laboratories reported the consensus haplotype (the one from the hair donor) for the M7 sample. One laboratory reported the haplotypes from the saliva and hairs separately; the saliva traces were firstly collected by rubbing the surface of the hair shaft using a cotton swab, and secondly the hair shafts were washed several times before DNA extraction.

Regarding the five laboratories that reported non-consensus sequences, two of them detected either the mixed haplotype (matching both haplotypes coming from the saliva donor and from the hair shaft) or only the saliva haplotype, which would

Table 3
M5, unknown donor and M6 consensus haplotypes and out of consensus haplotypes reported for the M6 mixture

Donor	Haplotype
M5 donor	16051G 16189C 16270T 73G 146C 150T 263G 309.1C 315.1C
Unknown donor	263G 315.1C
Consensus M6	16051R 16189Y 16270Y 73R 146Y 150Y 263G 309.1C 315.1C

Lab. ID	Reported haplotype	Error	Comments concerning electropherogram
1	16051R 16189Y 73R 146Y 150Y 263G 309.1C	16270Y and 315.1C omitted	Only F electropherogram available
2	73R 146Y 150Y 263G C8TC6 and C7TC6	Nomenclature and HVS-I omitted	16051R, 16189Y, and 16270Y in electropherograms
3	263G 315.1C	M5 not detected	Presence of 146Y and other unclear positions
4	16189Y 16270Y 73R 146Y 150Y 263G 309.1Y 310Y 315.1C	16051R omitted	16051R in electropherograms
5	263G	315.1C omitted, M5 not detected	Unavailable
6	16189Y 16270Y 73R 146C 150T 263G 309.1C 315.1C	16051R omitted, 146Y 150Y not detected	Low quality electropherograms
7	16051G 16189Y 16270T 73G 92A 146C 150T 263G	HVS-II-PolyC omitted, 146Y 150Y not detected, 92A detected	Low quality electropherograms (deficient purification?) Phantom mutation
8	51R 16189Y 16270Y 73R 146Y 150Y 263G 309.1C/-315.1C	Clerical error	16051R in electropherograms

clearly lead to a false exclusion in real casework in the case the hair shaft would belong to a suspect.

5. Statistical interpretation of the results: match–mismatch criteria and database searching

The GEP-ISFG mtDNA exercise also collects information regarding interpretation of the mtDNA matching evidence in the simulated forensic cases emulating the work carried out by the laboratories in real forensic situations. A plethora of different interpretations were used in this part of the exercise. This is to some extent expected and reflects the lack of standards and consensus among forensic geneticists. On the other hand, some laboratories only reported the match or mismatch status between reference and unknown samples avoiding any kind of statistical interpretation. On the contrary, other laboratories reported the number of matches of one specific haplotype in a specific database (mainly SWGDAM; www.fbi.gov/hq/lab/fsc/april2002/mtDNA.htm; [9]) or the frequency of the haplotype in that database, and finally, a few laboratories calculated a likelihood ratio value by applying the Balding and Nichols correction method [10] and using either their own or the SWGDAM database. A different and independent exercise carried out by the GEP-ISFG group on interpretation (data not shown) also highlights the lack of consensus among laboratories and the key role of chosen databases (specially the SWGDAM) for haplotype frequency and likelihood ratio estimation.

The interpretation of the number of nucleotide differences among haplotypes also differed substantially between laboratories. Some of the laboratories adopt a simplistic rule that considers two haplotypes as being different if there are two or more different nucleotide positions between them. Others are reluctant to adopt this convention mainly because the mutation rate in the mitochondrial genome dramatically varies among nucleotides; therefore, evaluating an exclusion/inclusion solely by the number of nucleotide differences can be problematic (see [11] for a review). The type of tissues or fluids involved in the haplotype comparison also plays a role since some of them are more prone to mutation than others. In problematic cases where enough information to establish an exclusion or an inclusion is not available, it might be necessary to enlarge the mtDNA fragment under study, in some cases analysing the complete control region sequence (from 16024 to 16569 and from 1 to 576) or looking for particular polymorphic positions in the coding region [12–15].

Concerning databases, some laboratories used their own (generally containing small amounts of profiles) while others used published data or the SWGDAM database. At the time of this collaborative exercise, the EMPOP database (http://www.empop.org/) was not operative; during the last GEP-ISFG meeting however this database was recommended (especially for those laboratories lacking their own databases) instead of the SWGDAM since the former contains a higher number of haplotypes and the data were carefully checked before their inclusion in the database [11,16–18]. However, laboratories need to be aware of potential problems related to the representativeness of external databases in their casework routine (see [11] for some caveat).

6. Conclusions and recommendations in relation to methodology

We have detected several aspects of the methodology that could help improve the results in routine casework and future editions of the GEP-ISFG exercise.

Only one laboratory carried out a specific quantification of the amount of mtDNA contained in the samples considered in this exercise. MtDNA quantification allows us to control the analysis process more accurately. The main advantage of quantifying mtDNA is that the results of this assay can assist in deciding the best strategy for posterior analysis (e.g. PCRs producing short amplicons in critical samples *versus* long amplicon PCRs in good quality samples). It also provides adjustments in the amount of target mtDNA for the PCR (avoiding the unnecessary loss of mtDNA in critical samples) and provides information regarding the convenience of handling low and high mtDNA content samples separately in order to prevent cross-contamination. Additionally, knowing the mtDNA content in forensic samples may help to detect contamination, especially in those samples with low amount of DNA: if an unexpected high amount of mtDNA is observed in a casework sample (a hair shaft, for instance), we may suspect contamination occurred when the sample was originated (e.g. a hair shaft mixed with vaginal fluid) or during collection and/or the analytical process. There are several examples of mtDNA quantification protocols [6,19], some of which can even give information about the presence of PCR inhibitors as well as the degree of mtDNA degradation.

Visualization of PCR products before the extension reactions is also a useful practice; the selection of good amplicons and the adjustment of the PCR product volume for the sequencing reaction help to guaranty a better sequence performance and to reduce the presence of artefacts in the electropherograms.

Concerning the length of mtDNA fragments studied, the analysis, when restricted to the classical HVS-I and HVS-II, may yield limited information in some forensic cases, such as those where the haplotype of unknown and reference samples match one of the most frequent in a reference population (i.e. the M7 haplotype is 263G 315.1C, the most common in Europe). Analysing the complete control region is the common strategy to increase the discrimination power of mtDNA. For example, it can be easily inferred from the data contained in the Human Mitochondrial Genome Database (http://www.genpat.uu.se/mtDB/; [20]) that there are at least 60 polymorphic sites between nucleotides 16365–16569 and 1–72, and 77 additional variants between nucleotides 341 and 574. Some polymorphisms are very informative for haplogroup assignment (e.g. 72C) while others have high mutation rate, and therefore are highly variable among populations (appearing in different haplogroup backgrounds), substantially increasing the discrimination power (e.g. 16519) [11]. Analysing the complete control region does not necessarily involve additional efforts. The

whole control region can be initially amplified in a single PCR reaction by using a single primer pair, following by sequencing shorter fragments using internal primers (i.e. [21]). These strategies do not require new technology, equipment or special training.

Analysis of SNPs located in the coding region also allows increasing the discrimination power of the mtDNA test in a forensic casework. There are several kinds of multiplexes designed to differentiate between H sub-haplogroups [13,14], East-Asian and Native-American haplogroups [15] and West European haplogroups [13,22]. Although no new equipment is usually necessary for these analyses, further technical training might be necessary to read and interpret the results properly.

Concerning the analyses of fluid mixtures, it is worth mentioning that additional non-mitochondrial genetic markers should be studied whenever possible for two main reasons: (a) autosomal markers are generally more informative than mtDNA and (b) the results of the present exercise indicate a level of success lower than desirable (69.2% in M6 sample). MiniSTRs analyses are a good choice in the study of forensic mixtures, and could also be a valuable tool for the genotyping of telogen hairs [23]. It is also recommended to carry out preliminary microscopic examinations of the hair strands aimed to detect possible contaminants and therefore evaluate appropriate protocols for decontamination.

7. Recommendations to avoid errors

The types of errors and their most probable causes are summarized in Table 4. Omission and misdocumentation of nucleotide variants are the most typical discrepancies. Omission of variants is mainly due to clerical errors (HVS-II-polyC stretch forgotten in the report but present in the electropherograms) and poor quality electropherograms (most of them due to the lack of double strand sequencing strategies and the use of additional primers in samples with length heteroplasmy at homopolymeric tracks). Deficient electropherograms were also the cause of a clear example of a phantom mutation: laboratory 9 reported 16469G in all samples.

In order to improve our results, in the present edition of the GEP-ISFG exercise we emphasized several simple recommendations that would have prevented most of the errors [24,25].

7.1. L and H strand sequencing in presence of length heteroplasmy

In order to read both the L and the H strands throughout all the hypervariable segments in presence of length heteroplasmy around position 16189 in HVS-I and around position 310 in HVS-II, it is necessary to use internal sequencing primers (e.g. L16209, H 16164) [26]. The results of the present exercise demonstrate that single strand analysis/reading is a common source of error (Table 2 and Supplementary material): the percentage of consensus results obtained for samples M1, M2, M3 and M4 was clearly superior to that of the M5 sample, despite the fact that all the samples had similar characteristics and contained a sufficient amount of mtDNA. This fact can be attributed to the presence of T16189C in sample M5 which produced an unstable poly-C stretch; thus, for instance, two laboratories reported the profile 16051G 16189C 73G 146C 150T 263G 309.1C 315.1C, but omitted the 16270T variant.

7.2. Reading the electropherograms

A substantial amount of errors was caused by artefacts produced by automatic reading of the electropherograms but without carrying out further visual inspection. The 5′-ends of the sequencing electropherograms are common hotspots for errors and therefore these segments should be systematically reviewed by visual inspection. If for some reason some part of the electropherogram cannot be unambiguously read (and there is no more sample available for further sequencing), it is

Table 4
Types of discrepancies in the present study

Type of error	Cause	Number of times
Position omitted	Clerical error	10
	Poor quality electropherograms/only one electropherogram per region	8
	Undetermined (no electros available)	14
Incorrect position reported	Phantom mutation	6
	Poor quality electropherograms	6
Typing error	Confusion	6
Nomenclature	–	2
Unresolved bases (Ns)	Poor purification	2
	Poor quality electros	2
	Length heteroplasmy and only one strand sequenced	3
Different haplotype	Unwashed hair (not decontaminated)	2
M6 mixture not detected	Poor quality electropherograms	1
	Undetermined (no electros available)	1
Mixed bases not reported	Poor quality electros	4

mandatory to clearly report the real reading range in that specific sample.

7.3. Data entry or editing

This exercise also recorded several examples of documentation errors: (i) reporting the 16189 nucleotide as different from rCRS instead of the 16188T polymorphism, (ii) 051G instead of 16051G, (iii) nomenclature errors such as C8TC6 instead of 309.1C 315.1C, (iv) missing nucleotides (16270 instead of 16270T), (v) reporting the rCRS base instead of the one in the sample (16189T instead of 16189C), etc. Correcting such errors is feasible; for instance, haplotypes can be electronically transferred to the final report and can be double checked by two independent analysts.

7.4. A posteriori QC

It has been demonstrated several times that *a posteriori* inspection of the mtDNA profiles to the light of the phylogeny or simple database searches contributes significantly to prevent a high proportion of the errors in mtDNA reports. Here we just indicate three different related steps that should be followed in this regard:

- Verify if the polymorphisms observed in our haplotypes have been already described in the literature or in databases. It is possible to carry out a quick search in several web sources (e.g. mtDB (www.genpat.uu.se/mtDB), EMPOP (http://www.empop.org/), and SWGDAM (www.fbi.gov/hq/lab/fsc/april2002/mtDNA.htm)). This verification process takes only few minutes and can help to prevent a great amount of errors and misleading interpretations [27]. Nevertheless, the above databases are not fully exhaustive so it is possible that common variants reported in the population or anthropological literature are still not recorded [27,28].
- Checking the haplotype from a phylogenetic perspective helps to detect common errors such as sample mix-up or contamination (see, for instance [18,24,25]). The mtDNA phylogeny is continuously improving and therefore, nomenclature and branching patterns (especially at the tips) are continuously changing. Therefore, it is fairly difficult to be knowledgeable about phylogeny; some publications should be used as references but always bearing in mind the last update; thus for instance, we have good sources for the European [29,30], Asian or Native-American [31,32] and African [33,34] phylogenies.

In the case of samples showing sequence heteroplasmy, it is instructive to check the rate of mutation of the nucleotide position where the heteroplasmy is located, as this helps the interpretation of results. The mutation rate is not uniform throughout the mtDNA molecule and there are some positions that are prone to accumulate changes, i.e. hotspots such as 16189 in HVS-I or 152 in HVS-II. We are aware that a specific mutation rate for each nucleotide position has not been established yet. Some useful information can be gathered in [11,35–37].

The use of phylogenetics as a tool for *a posteriori* checking can help to detect a great number of errors. Unfortunately, a substantial proportion of the laboratories show obvious difficulties in using this approach. In this regard, it is worth mentioning that the EMPOP web resource has a variety of different tools that could assist forensic geneticists in different tasks related to error detection.

8. Final remarks

In general, the outcome of the GEP-ISFG consortium exercise reflects only a modest improvement in its global outcome; the knowledge acquired in previous editions regarding the most common causes of errors helped to prevent the incidence of common mistakes in those laboratories with more experience. Since the electropherograms were also submitted by the majority of the laboratories, most of the errors could be catalogued. The causes of errors were similar to those in previous exercises, mainly comprising edition mistakes, lack of electropherogram quality at the 5′- and 3′-ends of the sequence, as well as nomenclature deficiencies. The exception this time is the apparent lack of contamination problems. We can conclude that in general the laboratories still lack solid devised experimental approaches and protocols, the keystone for preventing errors in mtDNA casework. The results of a QC are not anecdotic since these generally mirror the quality of the casework forensic practice. A great effort is still needed among forensic practitioners regarding the methodological and the theoretical framework in order to improve the health of the (somehow ‘damaged’) mtDNA test.

Acknowledgments

We are very grateful to the samples’ donors and to Milton A. Chin for revising the English style. We also wish to thank the suggestions of two anonymous reviewers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2007.10.010.

References

[1] J. Gómez, Á. Carracedo, The 1999 collaborative exercises and proficiency test program on DNA typing of the Spanish and Portuguese of the International Society for Forensic Genetics (GEP-ISFH), Forensic Sci. Int. 114 (1) (2000) 21–30.

[2] A. Alonso, A. Salas, C. Albarrán, E. Arroyo, A. Castro, M. Crespillo, A.M. Di Lonardo, M.V. Lareu, C.L. Cubría, M.L. Soto, J.A. Lorente, M.M. Semper, A. Palacio, M. Paredes, L. Pereira, A.P. Lezaun, J.P. Brito, A. Sala, M.C. Vide, M.R. Whittle, J.J. Yunis, J. Gómez, Results of the 2000 collaborative exercise and proficiency testing program of mitochondrial DNA of the GEP-ISFG: an inter-laboratory study of the observed variability in the heteroplasmy level of hair from the same donor, Forensic Sci. Int. 125 (1) (2002) 1–7.

[3] L. Prieto, M. Montesino, A. Salas, A. Alonso, C. Albarrán, S. Álvarez, M. Crespillo, A.M. Di Lonardo, C. Doutremepuich, I. Fernández-Fernández, A. González de la Vega, L. Gusmão, C.M. López, M. López-Soto, J.A. Lorente, M. Malaghini, C.A. Martínez, N.M. Modesti, A.M. Palacio, M. Paredes, S.D.J. Pena, A. Pérez-Lezaun, J.J. Pestano, J. Puente, A. Sala, M.C. Vide, M.R. Whittle, J.J. Yunis, J. Gómez, The 2001 GEP–ISFG Collaborative Exercise on mtDNA: assessing the cause of unsuccessful mtDNA PCR amplification of hair shaft samples, Forensic Sci. Int. 134 (1) (2003) 46–53.

[4] A. Salas, L. Prieto, M. Montesino, C. Albarrán, E. Arroyo, M.R. Paredes-Herrera, A.M. Di Lonardo, C. Doutremepuich, I. Fernández-Fernández, A. González de la Vega, C. Alves, C.M. López, M. López-Soto, J.A. Lorente, A. Picornell, R.M. Espinheira, A. Hernández, A.M. Palacio, M. Espinoza, J.J. Yunis, A. Pérez-Lezaun, J.J. Pestano, J.C. Carril, D. Corach, M.C. Vide, V. Álvarez-Iglesias, M.F. Pinheiro, M.R. Whittle, A. Brehm, J. Gómez, Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02–03 proficiency testing trial, Forensic Sci. Int. 148 (2–3) (2005) 191–198.

[5] M. Crespillo, M.R. Paredes, L. Prieto, M. Montesino, C. Albarrán, A. Salas, V. Álvarez-Iglesias, A. Amorin, G. Berniell-Lee, A. Brehm, J.C. Carril, D. Corach, N. Cuevas, A.M. Di Lonardo, C. Doutremepuich, R.M. Espinheira, M. Espinoza, F. Gómez, A. González, A. Hernández, M. Hidalgo, M. Jiménez, F.P. Leite, A.M. López, M. López-Soto, J.A. Lorente, S. Pagano, A.M. Palacio, J.J. Pestano, M.F. Pinheiro, E. Raimondi, M.M. Ramón, F. Tovar, L. Vidal-Rioja, M.C. Vide, M.R. Whittle, J.J. Yunis, J. García-Hirschfeld, Results of the 2003–2004 GEP-ISFG collaborative study on mitochondrial DNA: focus on the mtDNA profile of a mixed semen-saliva stain, Forensic Sci. Int. 169 (2–3) (2006) 157–167.

[6] A. Alonso, P. Martín, C. Albarrán, P. García, O. García, L. Fernández de Simón, J. García-Hirschfeld, M. Sancho, C. de la Rúa, J. Fernández-Piqueras, Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies, Forensic Sci. Int. 139 (2–3) (2004) 141–149.

[7] M. Montesino, A. Salas, M. Crespillo, C. Albarrán, A. Alonso, V. Álvarez-Iglesias, J.A. Cano, M. Carvalho, D. Corach, C. Cruz, A.M. Di Lonardo, R. Espinheira, M.J. Farfán, S. Filippini, J. García-Hirschfeld, A. Hernández, G. Lima, C.M. López-Cubría, M. López-Soto, S. Pagano, M. Paredes, M.F. Pinheiro, A.M. Rodríguez-Monge, A. Sala, S. Sóñora, D.R. Sumita, M.C. Vide, M.R. Whittle, A. Zurita, L. Prieto, Analysis of body fluid mixtures by mtDNA sequencing: an inter-laboratory study of the GEP-ISFG working group, Forensic Sci. Int. 168 (1) (2007) 42–56.

[8] E. Jehaes, A. Gilissen, J.J. Cassiman, R. Decorte, Evaluation of a decontamination protocol for hair shafts before mtDNA sequencing, Forensic Sci. Int. 94 (1–2) (1998) 65–71.

[9] K.L. Monson, K.W.P. Miller, M.R. Wilson, J.A. DiZinno, B. Budowle, The mtDNA Population Database: an integrated software and database resource for forensic comparison, Forensic Sci. Commun. 4 (2002) 2.

[10] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (2–3) (1994) 125–140.

[11] A. Salas, H.-J. Bandelt, V. Macaulay, M.B. Richards, Phylogeographic investigations: the role of trees in forensic genetics, Forensic Sci. Int. 168 (1) (2007) 1–13.

[12] T.J. Parsons, M. Coble, Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome, Croatian Med. J. 42 (3) (2001) 304–309.

[13] B. Quintáns, V. Álvarez-Iglesias, A. Salas, C. Phillips, M.V. Lareu, Á. Carracedo, Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, Forensic Sci. Int. 140 (2–3) (2004) 251–257.

[14] A. Brandstätter, A. Salas, H. Niederstätter, C. Gassner, A. Carracedo, W. Parson, Dissection of mitochondrial super-haplogroup H using coding region SNPs, Electrophoresis 27 (13) (2006) 2541–2550.

[15] V. Álvarez-Iglesias, J.C. Jaime, Á. Carracedo, A. Salas, Coding region mitochondrial DNA SNPs: targeting the East Asian and Native American haplogroups, Forensic Sci. Int.: Genet. 1 (1) (2007) 44–55.

[16] W. Parson, A. Dür, EMPOP—A forensic mtDNA database, Forensic Sci. Int.: Genet. 1 (2) (2007) 88–92.

[17] T.J. Parsons, Mitochondrial DNA Genome Sequencing and SNP Assay Development for Increased Power of Discrimination, 2006, http://www.ncjrs.gov/pdffiles1/nij/grants/213502.pdf.

[18] H.-J. Bandelt, A. Salas, C. Bravi, Problems in FBI mtDNA database, Science 305 (5689) (2004) 1402–1404.

[19] D.M. Timken, K.L. Swango, C. Orrego, M.R. Buoncristiani, A duplex real time qPCR assay for the quantification of human nuclear and mitochondrial DNA in forensic samples: implications for quantifying DNA in degraded samples, J. Forensic Sci. 50 (5) (2005) 1044–1060.

[20] M. Ingman, U. Gyllensten, mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences, Nucleic Acids Res. 34 (Database Issue) (2006) D749–D751.

[21] W. Parson, H.-J. Bandelt, Extended guidelines for mtDNA typing of population data in forensic science, Forensic Sci. Int.: Genet. 1 (1) (2007) 13–19.

[22] M.D. Coble, R.S. Just, J.E. O'Callaghan, I.H. Letmanyi, C.T. Peterson, J.A. Irwin, T.J. Parsons, Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians, Int. J. Legal Med. 118 (3) (2004) 137–146.

[23] A. Hellmann, U. Rohleder, H. Schmitter, M. Wittig, STR typing of human telogen hairs-a new approach, Int. J. Legal Med. 114 (4–5) (2001) 269–273.

[24] A. Salas, Á. Carracedo, V. Macaulay, M. Richards, H.-J. Bandelt, A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics, Biochem. Biophys. Res. Commun. 335 (3) (2005) 891–899.

[25] H.-J. Bandelt, A. Salas, S. Lutz-Bonengel, Artificial recombination in forensic mtDNA population databases, Int. J. Legal Med. 118 (5) (2004) 267–273.

[26] G. Tully, W. Bar, B. Brinkmann, Á. Carracedo, P. Gill, N. Morling, W. Parson, P. Schneider, Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles, Forensic Sci. Int. 124 (1) (2001) 83–91.

[27] H.-J. Bandelt, A. Salas, C.M. Bravi, What is a 'novel' mtDNA mutation—and does 'novelty' really matter? J. Hum. Genet. 51 (12) (2006) 1073–1082.

[28] H.-J. Bandelt, Y.-G. Yao, A. Salas, The search of 'novel' mtDNA mutations in hypertrophic cardiomyopathy: MITOMAPping as a risk factor, Int. J. Cardiol., 2007, doi:10.1016/j.ijcard.2007.02.049.

[29] M. Richards, V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F. Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, G. Stefanescu, J. Hatina, M. Belledi, A. Di Rienzo, A. Novelletto, A. Oppenheim, S. Norby, N. Al-Zaheri, S. Santachiara-Benerecetti, R. Scozari, A. Torroni, H.-J. Bandelt, Tracing European founder lineages in the Near Eastern mtDNA pool, Am. J. Hum. Genet. 67 (5) (2000) 1251–1276.

[30] S. Finnilä, M.S. Lehtonen, K. Majamaa, Phylogenetic network of European mtDNA, Am. J. Hum. Genet. 68 (6) (2001) 1475–1484.

[31] H.-J. Bandelt, C. Herrnstadt, Y.-G. Yao, Q.-P. Kong, T. Kivisild, T. Rengo, R. Scozzari, M. Richards, R. Villems, V. Macaulay, N. Howell, A. Torroni, Y.-P. Zhang, Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats, Ann. Hum. Genet. 67 (Pt 6) (2003) 512–524.

[32] Q.-P. Kong, H.-J. Bandelt, C. Sun, Y.-G. Yao, A. Salas, A. Achilli, C.Y. Wang, L. Zhong, C.L. Zhu, S.F. Wu, A. Torroni, Y.P. Zhang, Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations, Hum. Mol. Genet. 15 (13) (2006) 2076–2086.

[33] A. Salas, M. Richards, T. De la Fé, M.V. Lareu, B. Sobrino, P. Sánchez-Diz, V. Macaulay, Á. Carracedo, The making of the African mtDNA landscape, Am. J. Hum. Genet. 71 (5) (2002) 1082–1111.

[34] A. Torroni, A. Achilli, V. Macaulay, M. Richards, H.-J. Bandelt, Harvesting the fruit of the human mtDNA tree, Trends Genet. 22 (6) (2006) 339–345.

[35] H.-J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, Am. J. Hum. Genet. 71 (5) (2002) 1150–1160.

[36] B.A. Malyarchuk, I.B. Rogozin, V.B. Berikov, M.V. Derenko, Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region, Hum. Genet. 111 (1) (2002) 46–53.

[37] H.-J. Bandelt, Q.P. Kong, M. Richards, V. Macaulay, Estimation of mutation rates and coalescence times: some caveats, in: H.-J. Bandelt, V. Macaulay, M. Richards (Eds.), Human Mitochondrial DNA and the Evolution of *Homo sapiens*, Springer-Verlag Press, Berlin-Heidelberg, 2006, pp. 227–268.

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Forensic Science International: Genetics 2 (2008) 212–218
FSI GENETICS
www.elsevier.com/locate/fsig

Case report: Identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur

M. Fondevila [a], C. Phillips [a,b,*], N. Naveran [a], L. Fernandez [a], M. Cerezo [a], A. Salas [a,b], Á. Carracedo [a,b], M.V. Lareu [a,b]

[a] *Institute of Legal Medicine, Genomic Medicine Group, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain*
[b] *Genomic Medicine Group, CIBERER, University of Santiago de Compostela, Spain*

Received 30 October 2007; received in revised form 22 January 2008; accepted 18 February 2008

Abstract

Applying two extraction protocols to isolate DNA from a charred femur recovered after a major forest fire, a range of established and recently developed forensic marker sets that included mini-STRs and SNPs were used to type the sample and confirm identity by comparison to a claimed daughter of the deceased. Identification of the remains suggested that the individual had been dead for 10 years and the DNA was therefore likely to be severely degraded from the combined effects of decomposition and exposure to very high temperatures. We used new marker sets specifically developed to analyze degraded DNA comprising both reduced-length amplicon STR sets and autosomal SNP multiplexes, giving an opportunity to assess the ability of each approach to successfully type highly degraded material from a challenging case. The results also suggest a modified ancient DNA extraction procedure offers improved typing success from degraded skeletal material.

Keywords: Degraded DNA analysis; Forensic identification; STR; Mini-STR; SNP

1. Introduction

A frequently encountered challenge in forensic casework is the analysis of highly degraded DNA that requires extraction from difficult material such as bones and teeth from long deceased individuals. In such cases standard STR markers often fail, primarily because of their amplicon lengths, with a direct relationship between the length of amplified fragment and the frequency at which the locus fails to amplify completely or shows signal imbalance between short and long tandem repeats [1,2]. Two alternative approaches are possible for analyzing severely degraded material: mitochondrial DNA typing (mtDNA) and autosomal short-amplicon markers. Until recently, mtDNA offered the only realistic alternative to standard STRs for highly degraded DNA, however it has certain acknowledged drawbacks: (i) sequence variation is much less informative than STRs; (ii) matrilineal inheritance rules out some relationship comparisons such as father and daughter and; (iii) proper interpretation of the significance of mtDNA variation requires large-scale haplotype databases. Alternatively reducing the length of marker amplicons can be achieved by re-designing the primers of existing STRs, commonly termed mini-STRs [3,4] or by analyzing single nucleotide polymorphisms (SNPs) [5,6]. However, while bringing the primers of STRs closer to the repeat region reduces overall amplicon size, loci with the largest range of alleles such as FGA and D21S11 will still suffer from big differences in allele size and therefore disparity in PCR performance between the extremes in repeat number. Although with SNPs designing short-amplicon primers around one base is clearly easier, multiplexes of 45–50 loci are required before a sufficient number of binary polymorphisms can match the discriminatory power of STRs [5].

Since successful typing of severely degraded DNA should be the primary characteristic of new marker sets designed for the purpose regardless of their shortcomings, it is important to

* Corresponding author at: Institute of Legal Medicine, Genomic Medicine Group, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain.
E-mail address: c.phillips@mac.com (C. Phillips).

1872-4973/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2008.02.005

properly assess how each of the two short-amplicon approaches perform in real cases compared to conventional typing strategies. This report outlines the analysis of DNA extracted from skeletal remains that, once identified, were shown to have undergone a 10-year period of decomposition followed by exposure to the extreme temperatures associated with forest fires.

1.1. Case report

During the summer of 2006 a large number of forest fires afflicted Galicia, NW Spain. Forest fires in this region are characterized by very high temperatures caused by the resinous nature of the wood in mixed Pine and Eucalyptus plantations. After one major fire was extinguished investigators discovered a set of charred skeletal remains uncovered by the burning of underlying foliage. Analysis of the skeleton was initiated to identify if they were from a fire victim or of a man reported as missing from the area 10 years back. The daughter of the missing man was available for comparison, illustrating here the unsuitability of mtDNA and Y-chromosome loci as uni-parental lineage markers for the identification of fathers and their female descendants. A complete femur was the only bone showing some intact areas with reduced charring and was submitted for analysis, despite other signs of exposure to intense heat, as shown in Fig. 1. When discovered this femur was seen to be half buried in the soil, suggesting a degree of protection for part of the bone, while several other bones had been fragmented, with pathology indicating acts of violence.

2. Materials and methods

DNA was isolated with two different extraction methods and using the product of each in parallel analyses, typed with two standard STR kits: AmplSTR® Identifiler™ (Applied Biosystems: AB) and Powerplex 16™ (Promega), then compared to two SNP assays developed by the SNP*for*ID consortium (http://www.snpforid.org), three mini-STRs multiplexes and, solely in this case as a performance benchmark, HV1 mtDNA sequencing.

Fig. 1. Photograph of the femur as received at the laboratory prior to cleaning and DNA extraction.

2.1. Sample preparation

Prior to DNA extraction the femur was cleaned thoroughly using a scalpel and mild-grade sandpaper, limiting attention to a small, 12 cm length of bone in an area visibly less carbonized by the fire. Once the external layer of bone containing possible contaminants had been fully removed the bone was cut into small portions. The internal face of the bone was further cleaned and portions were pulverized in a liquid N_2 mill yielding a total of 8 ml of fine bone dust.

2.2. DNA extraction and quantitation

Two extraction methods were performed in parallel: (A) a normal phenol–chloroform extraction process adapted for bone material, plus: (B) an ancient DNA extraction protocol [7] we have enhanced for forensic use to improve the quality of extracted DNA. (A) Phenol–chloroform samples were digested as two 2.5 ml aliquots of bone dust plus 2.3 ml of a lysis mix comprising 2 ml of 0.5 M EDTA pH 8 as buffer, 80 μl of DTT 1 M, 140 μl of 10% SDS and 50 μl of protease-K 20 ng/ml each. Overnight lysis at 56 °C was followed by standard phenol–chloroform extraction with DNA purified and concentrated with Centricon centrifugal filter devices (Millipore) following the manufacturers protocol. Each dust aliquot was recovered in TE buffer and combined as a single extract. (B) 1.5 ml of bone dust was decalcified with 10 ml of an EDTA solution (pH 8.8) and left at 37 °C overnight. Digestion involved sample centrifugation for 10 min at 15,000 rpm, discarding of the pellet followed by addition of 1 ml 5% SDS, 500 μl 1 M Tris–HCl, 500 μl 0.5 M NaCl, 500 μl 0.1 M $CaCl_2$, 2 × 50 μl protease-K (20 mg/ml), 75 μl of DTT and 7 ml of H_2O. The mixture was incubated for 24 h at 65 °C, then 24 h at 75 °C. The final extraction step was a phenol–chloroform method as detailed in the original protocol [7]. Extraction procedures (A) and (B) were made on individual occasions separated by 24 h with negative extraction controls made in parallel. All extracts were quantified with the real-time PCR Quantifiler™ human DNA quantification kit (AB) using an AB7300 real-time PCR thermal cycler following manufacturers guidelines. Quantifiler™ includes an internal PCR control (IPC) detecting PCR inhibition. Our approach to minimizing inhibition from severely degraded DNA in standard STR typing is to run five dilutions as tandem PCRs, i.e. neat, 1/2, 1/4, 1/8, and 1/16. All other marker sets used only neat extracts.

2.3. Standard STR and mtDNA typing

Identifiler™ and Powerplex 16™ STR typing followed manufacturers guidelines except use of 12.5 μl reaction volumes. PCR comprised 28 cycles and 32 cycles respectively and we note that although increasing cycle number beyond the above numbers can influence PCR yield and quality, these cycles represent the STR amplification conditions we have optimized from numerous challenging DNA cases analyzed. In addition, the optimization of bone and tooth extract preparations in such cases,

Table 1
Genotyping success using SNP multiplexes (shown as total loci genotyped)

SNP multiplex	Total loci	Classical extraction	Ancient DNA extraction
52 plex PCR + Auto1 extension (23 plex)	23	21	21
52 plex PCR + Auto2 extension (29 plex)	29	26	28
Dedicated 23 plex PCR + SBE	23	20	23
Dedicated 29 plex PCR + SBE	29	26	29
34 plex AIMs	34	31	34

detailed elsewhere [8], gives strong indications that inhibition control by running a dilution series together with assessment of the IPC readings is the most effective way to improve yield and quality with standard STR typing. Capillary electrophoresis was performed throughout using an AB3130 with a 50 cm capillary array and POP-7™ polymer, injecting: 1 μl of PCR product, 15 μl of HI-DI™ formamide and 0.35 μl of LIZ GS500™ size standard. Mitochondrial DNA sequencing analysis followed standard protocols outlined previously [9] with primers: 15997L-16236H, 16159L-16401H and 16380L-017H.

2.4. SNP typing

Two validated SNP multiplexes were genotyped using SNaPshot™ primer extension (AB), comprising a 52 plex forensic identification set [10] and a 34 plex ancestry indicative set [11]. The 52 plex assay uses two parallel primer extension reactions detecting 23 and 29 SNPs (termed Auto1 and Auto2, respectively), the 34 plex assay detects all SNPs in a single extension reaction. Both SNP sets were amplified with a single PCR multiplex although the 52 plex can provide improved sensitivity for degraded DNA using reduced-scale 23 plex and 29 plex PCR. We performed both 52 plex and dedicated 23 plex/29 plex amplifications to assess this effect. All SNP analyses used the following modifications to the published protocol [10]: PCR volume was reduced to 13.5 μl, PCR annealing time was 50 s, extension time 40 s. Primer extension was at a reaction volume of 6 μl comprising: 2 μl SNaPshot™ Ready Reaction mix, 1.5 μl extension primer mix and 2 μl purified PCR product. Electrophoresis was as for STRs (AB3130, POP-7™, 50 cm capillary) injecting: 2 μl extension product, 9.5 μl of HI-DI™ formamide and 0.3 μl of AB LIZ™ 120 internal size standard. The 34 plex SNP set used the same reaction conditions and modifications as the 52 plex. Detailed SNP typing protocols are available as a download from: http://www.snpforid.org.

2.5. Mini-STR typing

Three short-amplicon mini-STR sets were used; Mini-NC01 and Mini-SGM, both developed by the National Institute of Standards and Technology (NIST) [12,13 respectively] and the commercial AmplSTR® MiniFiler™ kit (AB). All typing followed the recommended guidelines of NIST (http://www.cstl.nist.gov/biotech/strbase/miniSTR/updated_NC01_protocol.pdf) and AB. Electrophoresis conditions were the same as standard STR analysis.

3. Results

3.1. Bone extraction systems and reproducibility of genotype profiles

Summary genotyping results obtained with each DNA extraction system for SNPs and with ancient DNA extraction for STRs are outlined in Tables 1 and 2, respectively. The most complete electropherograms obtained from each of the short-amplicon genotyping approaches are shown in Fig. 2, which in all cases resulted from the ancient DNA extraction system for STR sets: MiniFiler™, Mini-NC01 and Mini-SGM, and for SNP sets: 23 plex, 29 plex (i.e. the reduced-scale PCRs of the 52 plex) and 34 plex. In particular the SNP results, summarized as total loci typed in Table 1, highlighted the improvement in quality and completeness of profiles obtained from the ancient DNA protocol and suggested enhanced sensitivity with this modified extraction method.

Because short-amplicon systems Mini-SGM and notably MiniFiler™ gave indications that more complete profiles might be obtainable, each of these marker sets was analyzed with

Table 2
Genotyping success using short-amplicon STRs

STR multiplex	Locus	Ladder allele range	Size range	Femur	Daughter
MiniFiler™	CSF1PO	6–15	88–124	10	10
	D16S539	5/8–15	91–119	–	9,13
	Amelogenin	*X–Y*	104/109	*X(Y)*	*XX*
	D13S317	8–15	106–134	11	8,11
	D2S1338	15–28	123–175	16,21	21
	D18S51	7–27	128–208	16	12,16
	FGA	17–33.2/53.2	151–213/293	OL	23,28
	D7S820	6–15	153–189	–	8,12
	D21S11	24–38	190–250	(24,29)	27,29
Mini-NC01	D22S1045	8/10–19	74–109	16	16
	D10S1248	9–19	83–123	14,16	14,16
	D14S1434	9–16	85–106	14	14
Mini-SGM	TH01	3–14	58–102	6,9.3	7,9.3
	D16S539	5–17	79–126	(OL,11)	9,13
	D2S1338	15–30	94–154	–	21
	D18S51	5–29	101–197	–	12,16
	Amelogenin	*X–Y*	124/130	*XY*	*XX*
	FGA	17–33.2/53.2	141–207/287	–	23,28

Individual genotypes are shown for ancient DNA extraction only, listed from shortest to longest first allele. Results in brackets represent peaks in reference positions but with signal strength between 30–50 RFU, these genotypes were not used in paternity calculations. OL: off-ladder peaks outside of reference position.

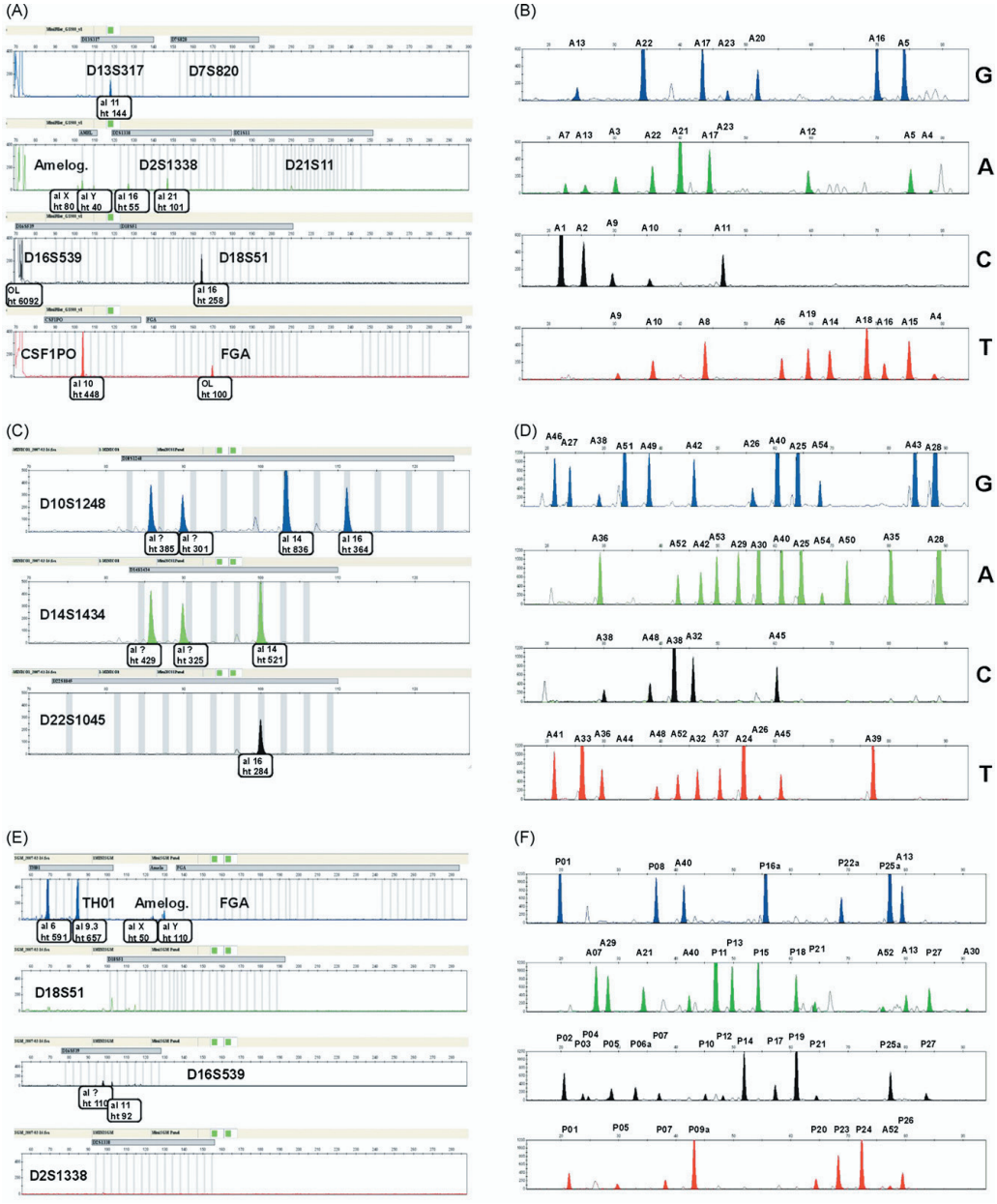


Fig. 2. Genotype results obtained from the femur using the ancient DNA extraction protocol. (A) AmplSTR® MiniFiler™ STR set, (B) SNP*for*ID Auto1 23 plex SNP set using a dedicated 23 plex PCR, (C) NIST Mini-NC01 STR set, (D) SNP*for*ID Auto2 29 plex SNP set (dedicated 29 plex PCR), (E) NIST Mini-SGM STR set, (F) SNP*for*ID 34 plex ancestry indicative marker SNP set. Solid-colour peaks denote identified genotypes except the off-ladder (OL) peaks at FGA (plot A) and D16S539 (E). Solid peaks of STR sets in plots A, C, E are above a pre-determined minimum signal intensity of 50 RFU, solid peaks of SNP sets in plots B, D, F are those within ±0.5 bp of pre-determined size positions. All 52 plex identification set SNPs in (A) and (D) labeled as in the SNP*for*ID browser (http://www.bioinformatics.cesga.es/snpforid/). 34 plex ancestry set loci labeled as follows: P01-rs2304925, P02-rs5997008, P03-rs1321333, P04-rs2814778, A07-rs917118, A29-rs1024116, P05-rs7897550, P06a-rs10843344, A21-rs722098, P07-rs239031, P08-rs12913832, A40-rs2040411, P09a-rs1978806, P10-rs773658, P11-rs10141763, P12-rs182549, P13-rs1573020, P14-rs896788, P15-rs2065160, P16a-rs2572307, P17-rs2303798, P18-rs2065982, P19-rs3785181, P20-rs881929, P21-rs1498444, P22a-rs1426654, P23-rs2026721, P24-rs4540055, A52-rs1335873, P25a-rs16891982, P26-rs730570, A13-rs1886510, P27-rs5030240, A30-rs727811.

duplicate amplifications for the ancient DNA extraction, three analyses per marker set in total. Although Fig. 2 shows the best results, product peaks were consistent amongst analyses in all cases but weaker on other occasions. All SNP sets were analyzed with duplicated amplifications.

3.2. Performance of conventional STRs, short-amplicon STRs, mtDNA and SNPs

Each of the conventional STR sets failed to amplify detectable alleles, including the shortest amplified products of amelogenin, across multiple analyses, comprising duplicate amplifications of each dilution and from both extraction systems. In this identification case the normal recourse of extracting from different bones within the same skeleton or from different segments of the long bones was precluded by the condition of the submitted material and use of most of the best material for two parallel extractions.

Short-amplicon STRs gave detectable alleles in all three marker sets used, although NIST Mini-SGM only gave reliable genotypes for the shortest amplicons of TH01. MiniFiler™ showed partial profiles for both extracts and the signal strength generally correlated with amplicon length with the exception of D16S539: the STR that performed poorly in all cases despite a relatively short amplicon size. The typing results of the NIST sets suggest that while Mini-SGM performs poorly, the Mini-NC01 STRs amplify well with degraded DNA, although extra non-specific peaks are evident in Fig. 2C at two loci. Similarly off-ladder (OL) peaks above 50 RFU are evident in D16S539 in Mini-SGM and FGA in MiniFiler™. Although the D16S539 OL peaks are not particularly strong compared to the TH01 peaks in the same set, the FGA peak appears to be more akin to a well-defined product peak just outside a standard allele bin. Analysis of the FGA peak shown in Fig. 2A and a second replicate indicated a size estimate ~0.2 bp slower than the allele bin left "edge" and ~0.65 bp slower than the bin mid-point, while the co-electrophoresed CSF1PO peak ran to the midpoint of its bin. This indicates that the FGA peak was more likely to be artifactual than a slow running 22 allele or a rare intermediate 21.3 allele running fast, both of which would exclude the daughter. The OL peaks in D16S539 and FGA were observed in both ancient DNA extraction replicates but not in the phenol–chloroform extraction.

HVI mtDNA sequencing success is not included in the table but sequencing gave good quality sequence profiles in all analyses with no detectable reduction in peak quality given the challenging condition of the DNA. This is in line with widely published casework results where mtDNA has previously been the most successful means to identify badly burnt human remains.

It is evident that in the analyses of this case SNP genotyping gives the most successful performance. Firstly, the subdivision of the 52 plex into a dedicated 23 plex and 29 plex gave complete profiles compared to using a single combined PCR prior to primer extension but improvements in both success and signal strength were marginal. Secondly, previous observations of the relative performance of Auto1 and Auto2 in a range of degraded DNA analyses indicated that, regardless of the PCR multiplex used, Auto2 is consistently more robust and sensitive as well as providing better quality peak patterns in nearly all cases [8,10]. This is consistent with the results obtained in this study and is illustrated by adjacent electropherograms B and D in Fig. 2. Lastly, the ancestry-indicative 34 plex gave near identical success rates to the subdivided Auto1 and Auto2 PCR indicating that a slightly larger-scale PCR and primer extension reaction does not unduly affect performance with highly degraded DNA. Although some non-specific peaks were seen in each of the primer extension electropherograms, in only a few instances did these reach comparable heights to the SNP alleles and all were well separated from the expected sizes of the alleles. Although the most complete SNP profiles are shown in Fig. 2, ~5% of loci failed in the duplicate amplification runs (slightly less with 34 plex) and in keeping with SNP analysis we have performed in other challenging cases, no consistently good or bad performers were evident amongst the 52 and 34 loci.

3.3. Informativeness of marker sets

Comparison of the genotype profiles from the femur with those of the claimed relative gave a paternity index (PI) of 4625 (probability = 99.978%) for STRs, a PI of 42,645 (probability = 99.998%) for the identification SNP set and a combined PI of 197,350,337 (probability = 99.999999%). The genotyping data from the femur and their assessments of paternity listed above were therefore reported as positive identification of the skeleton as the remains of the man missing from the area 10 years back. The PI and probability values obtained indicate that when a full set of SNP genotypes is obtained the collective power is better than a partial STR profile despite the characteristic that SNPs, as binary polymorphisms, have much lower informativeness per locus. In fact the increased power of SNP sets to differentiate individuals compared to MiniFiler™ STRs is evident from a comparison of complete profiles for each marker set. The eight STRs of MiniFiler™ give an average random match probability of 6.5×10^{-11} for African Americans and 8.2×10^{-11} for Europeans (source: AB product information) compared to 1.0×10^{-18} and 3.7×10^{-21} as equivalent values for the 52 SNPs (source: SNP*for*ID allele frequency browser: http://www.bioinformatics.cesga.es/snpforid/search.php).

The 34 plex ancestry informative set was not used in the paternity analysis since these SNPs were chosen to give little or no variability amongst individuals from the same population-group. However, information about likely population of origin may be useful in the analysis of multiple deceased individuals particularly when applied to the identification of mass disaster victims. For this reason it is important to use the opportunity to assess the 34 plex SNP set with challenging DNA cases and to test the informativeness for ancestry if partial data is obtained [12]. Using a Bayesian classification system imbedded within an interpretative web-portal (http://www.mathgene.usc.es/snipper/) to derive a probability of European, African or East Asian ancestry, expressed as −log likelihoods, (i.e. a smaller value is more suggestive of ancestry), indicated the 34 plex

Table 3
Classification probabilities for SNP profile from femur using the SNP*for*ID 34-SNP ancestry indicative marker set

Training set for calculation	−log likelihood of SNP profile	Likelihoods (as exponentials)	Likelihood ratio	Probability expressed as a verbal predicate
European	41.1095	1.401E−18	European not African:1.644E+11	164 Billion times more likely European than African
African	66.9353	8.518E−30		
European	41.1095	1.401E−18	European not Asian:4.456E+10	44 Billion times more likely European than Asian
Asian	65.6296	3.144E−29		
Asian	65.6296	3.144E−29	Asian not African:3.69	
African	66.9353	8.518E−30		

Ancestry assignment is based on a three population-group comparison and 120 sample training set from each group for the classification algorithm. Likelihood of SNP profile denotes the probability of the individual ancestry matching that of the training set (the lowest −log likelihood value equates to the highest probability).

profile from the femur was 164 billion times more likely to be European in origin than African and 44 billion times more likely European than East Asian (using a three-way differentiation and the three training sets of the classification portal). The statistical values obtained from the Bayesian analysis are outlined in Table 3.

To gauge the reliability of the 52 plex SNP results in the absence of reference genotypes from the deceased, the observed number of heterozygotes was compared to an expected number estimated from a population sample of NW Spain listed in the SNP*for*ID frequency browser. The Auto1 + 2 profiles showed 21 of 52 heterozygous SNPs compared to an expected 24 heterozygotes based on a 46% heterozygosity estimate for NW Spain. This was interpreted as indicating an absence of allele dropout in the SNP profiles obtained from the femur. Although there were insufficient loci in the small-amplicon STR sets to allow the same check to be realistically made with the STR profiles these evidently showed lower heterozygosity than allele frequencies would suggest. An alternative approach to assessing allele dropout is to estimate the expected exclusion rate from the markers used and in the case of the SNP set this is 99.98% in Europeans [9]. Therefore, the lack of exclusions in our analyses suggests an absence of allele dropout.

4. Discussion

Although this is a single challenging DNA case a noticeable difference in typing performance between short and standard amplicon length markers is evident in the DNA analyses described. It is likely from the circumstances of the case that the DNA extracted was severely compromised from the combined processes of decomposition and heat-induced degradation. Therefore, our observations of this case, showing unusually severe degradation, together with other challenging analyses [8] can provide a better way to gauge the success of new marker sets and extraction procedures than trying to reproduce degradatory effects in the laboratory. Considerable effort was made in the original SNP sets primer design process [11,12] to create multiplexes producing amplicons below a 120 bp size limit (average and range for the 52 plex: 88 bp, 59–115 bp and for the 34 plex: 86 bp, 61–117 bp) and the comparable near-complete profiles and peak quality shown in both sets suggests that this is a key factor in reducing locus dropout when typing highly degraded DNA. The reduced-length STR marker sets all provided genotypes from degraded DNA when none could be obtained with standard length amplicons. Although it is inappropriate to draw conclusions from the performance of individual STRs in one case, generally those amplifying above a size ranging from 150 to 180 bp appear to fail more readily in these analyses. A recent study led to the recommendation that short-amplicon STR products should aim to be smaller than 150 bp to ensure the best sensitivity [2]. In our analysis the amelogenin peaks of MiniFiler™ showed a degree of size related imbalance in contrast to the other heterozygous locus of D2S1338 suggesting that stochastic effects between alleles cannot be ruled out as complicating factors in the interpretation of peaks from larger STRs or those showing broader allele size ranges (e.g. FGA). Furthermore FGA in MiniFiler™ and D16S539 in Mini-SGM showed unassigned peaks with good signal strength.

One problem that can limit the proper comparison of marker set performance in real casework is the inability to control for allele dropout when assessing identification cases based solely on bodily remains. Without a reference profile the paternity index for a surviving relative provides a statistical likelihood that both individuals are related as claimed and in this case the failure to detect any exclusion in the 52 SNPs, coupled with a heterozygosity that is a reasonable match to the population as a whole gives persuasive evidence that the SNP profiles obtained from the femur represent the true genotypes of the deceased. Although the possibility of SNP allele dropout cannot be completely discounted, we have observed across a range of challenging DNA analyses [8] that the primer extension reactions of the SNaPshot assays used do not reveal noticeable differences in PCR efficiency, only some imbalance in peak heights related to variation in dye signal strength or extension efficiency between alleles of the same SNP. Furthermore although the background baseline signal can be higher with DNA extracts from degraded sources, these extra peaks always occupy positions well separated from the size bins used to identify the SNP alleles. Therefore, SNP typing tends to show more consistent differences in peak height between loci in the same electropherogram largely independent of the DNA quality and it is difficult to isolate particular SNPs as weak performers with degraded DNA.

The low individual informativeness of SNPs has been a factor hindering their widespread adoption amongst the forensic

community as first-choice markers for degraded DNA analysis. However, this characteristic may become largely irrelevant if full SNP profiles can be reliably obtained from highly degraded material using large-scale multiplexes and simple, easily adopted genotyping systems. This case suggests that short-amplicon approaches offer improved success when typing highly degraded DNA, while SNPs, as part of a range of such marker sets available to the forensic practitioner, could be valuable additions to the more established STRs.

Acknowledgments

Two grants from the Xunta de Galicia PGIDIT06PXIB208079PR and PGIDTIT06PXIB228195PR given to AS and MVL, respectively, a grant from the Fundación de Investigación Médica Mutua Madrileña awarded to AS and a grant of the Ministerio de Educación y Ciencia (BIO2006-06178) given to MVL supported this project.

References

[1] D.T. Chun, J. Drabek, K.L. Opel, J.M. Butler, B.R. McCord, A study on the effects of degradation and template concentration on the amplification efficiency of the STR Miniplex primer sets, J. Forensic Sci. 49 (2004) 733–740.

[2] L.A. Dixon, A.E. Dobbins, H.K. Pulker, J.M. Butler, P.M. Vallone, M.D. Coble, W. Parson, B. Berger, P. Grubwieser, H.S. Mogensen, N. Morling, K. Nielsen, J.J. Sanchez, E. Petkovski, Á. Carracedo, P. Sanchez-Diz, E. Ramos-Luis, M. Brion, J.A. Irwin, R.S. Just, O. Loreille, T.J. Parsons, D. Syndercombe-Court, H. Schmitter, B. Stradmann-Bellinghausen, K. Bender, P. Gill, Analysis of artificially degraded DNA using STRs and SNPs – results of a collaborative European (EDNAP) exercise, Forensic Sci. Int. 164 (2006) 33–44.

[3] P. Wiegand, M. Kleiber, Less is more – length reduction of STR amplicons using redesigned primers, Int. J. Legal Med. 114 (2001) 285–287.

[4] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA, J. Forensic Sci. 48 (2003) 1054–1064.

[5] P. Gill, An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, Int. J. Legal Med. 114 (2001) 204–211.

[6] B. Sobrino, M.Á. Brion, Á. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic Sci. Int. 154 (2005) 181–194.

[7] C. Lalueza-Fox, J. Bertranpetit, J.A. Alcover, N. Shailer, E. Hagelberg, Mitochondrial DNA from *Myotragus balearicus*, an extinct bovid from the Balearic Islands, J. Exp. Zool. 288 (2000) 56–62.

[8] M. Fondevila, C. Phillips, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L.M. Fernández, Á. Carracedo, M.V. Lareu, Challenging DNA: assessment of a range of genotyping approaches for highly degraded forensic samples, Forensic Sci. Int. Genet. Supplement Series 81 (2008) 1–3.

[9] V. Álvarez-Iglesias, J.C. Jaime, Á. Carracedo, A. Salas, Coding region mitochondrial DNA SNPs: targeting East Asian and Native American Haplogroups, Forensic Sci. Int. Genet. 1 (2007) 44–55.

[10] J.J. Sanchez, C. Phillips, C. Børsting, K. Balog, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, Á. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.

[11] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevil, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNP*for*ID Consortium, inferring ancestral origin using a single multiplex assay of autosomal ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.

[12] M.D. Coble, J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA, J. Forensic Sci. 50 (2005) 43–53.

[13] C.R. Hill, M.C. Kline, J.J. Mulero, R.E. Lagace, C.W. Chang, L.K. Hennessy, J.M. Butler, Concordance study between the AmpFlSTR MiniFiler PCR amplification kit and conventional STR typing kits, J. Forensic Sci. 52 (2007) 870–873.

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics Supplement Series 1 (2008) 26–28
www.elsevier.com/locate/FSIGSS

Research article

Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples

M. Fondevila, C. Phillips *, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L.M. Fernández, Á. Carracedo, M.V. Lareu

Institute of legal Medicine, University of Santiago de Compostela, Spain

Received 5 September 2007; accepted 11 October 2007

Abstract

It is common in forensic casework to encounter highly degraded DNA samples from a variety of sources. In this category bone and teeth samples are often the principal source of evidential material for criminal investigations or identification of long-deceased individuals. In these circumstances standard STRs are prone to fail due to their long amplicon sizes (since DNA becomes progressively more fragmented as it degrades). To successfully resolve such cases alternative markers can be used and until recently the only other tool available was mitochondrial DNA, which despite being more resistant to degradation, is much less informative. A rapidly developing approach to analyzing degraded DNA is the typing of loci from short-amplicon PCR products based on markers such as mini-STRs and autosomal SNPs. We have performed an analysis of several cases with naturally degraded DNA using established STRs plus mini-STRs and autosomal SNPs in order to make an objective comparison of the performance of each method using challenging DNA. The main aim was to establish the benefits and drawbacks of each marker set to help the practitioner choose the DNA analysis method most suited to the circumstances of each case.

Keywords: STRs; Degraded DNA; Short-amplicon STR analysis; SNP typing

1. Introduction

It is common to encounter highly degraded DNA samples in a criminal investigation or the identification of long-deceased individuals. Most forensic laboratories have experienced situations where the DNA is so degraded that normal PCR amplification gives inconclusive results. The quality of genotyping depends largely on the degradation processes the sample has been exposed to and these affect typing success in different ways: agents can affect the DNA structure itself through the action of nucleases or oxidative damage, while indirectly, the inhibitory effects of co-extracted agents such as humic acid or degradation by-products can hinder the PCR reaction. The extent of the degradation process depends on two factors: time and environmental conditions [1]. Degradative processes accumulate with time while environmental conditions (temperature, humidity, pH, soil chemistry) modify the rate and aggressiveness of degradation. Both factors interact in a complex way so there is no direct correlation between time since death/deposition of material and the extent of degradation, making it very difficult to set specific rules for the treatment of any one sample. Fortunately several short-amplicon genotyping approaches have been recently developed specifically for the analysis of degraded DNA. This study compares the performance of established and novel genotyping methods in a range of challenging cases with naturally degraded DNA.

2. Materials and methods

We analyzed 15 separate skeletal samples from submitted case-work, in a range of conditions, each assessed from the performance of standard STR typing. All cases originated from NW region of Spain: characterized by high annual rainfall, mild temperatures, and organic-rich, acidic soil. Surviving relatives were analyzed to obtain reference profiles when available. Prior to DNA extraction, bones or teeth were thoroughly cleaned with a scalpel/sandpaper. Except for the most degraded samples, extraction was based on the phenol–chloroform method with *Centricon*® column purification. For the most degraded

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).

1875-1768/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigss.2007.10.057

material we used a second method based on the ancient DNA protocol of Lalueza-Fox et al. [2] modified to enhance DNA yield. All extracts were quantified with the real-time PCR *Quantifiler*® human DNA kit (Applied Biosystems: AB). An internal PCR control (IPC) in each *Quantifiler*® reaction identifies material containing PCR inhibitors. Standard STR typing comprised *AmpFlSTR Identifiler*® (AB) and *Power-Plex16* (Promega) with extract dilutions: neat, 1 in 2, 1/4, 1/8 and 1/16 in tandem PCRs. Two mini-STR sets were: *MiniNC01*, developed at NIST [3], and *AmpFlSTR MiniFiler*® (AB). SNP typing consisted of two assays: the SNP*for*ID 52plex human identification SNP set [4] and the SNP*for*ID 34plex population informative AIM-SNP set [5]. Both sets use an initial PCR followed by a multiplexed primer extension (52plex comprises tandem 23 and 29plex extension reactions). SNP assays used undiluted DNA samples throughout.

3. Results

Table 1 outlines quantification data and % genotyping success from each sample. The most significant findings came from a comparison of genotyping success with data from the *Quantifiler*® IPC. IPC values higher than 28 indicate the presence of inhibitors which themselves can affect the quantification, so values obtained in these circumstances are less indicative of the actual DNA levels in an extract. Furthermore, values suggesting high DNA concentrations do not often equate to sufficient quantities of intact high molecular weight target to ensure successful PCR of standard STRs. We countered inhibitory effects by diluting extracts so although less DNA was applied, inhibition was reduced. Most samples gave complete profiles with each multiplex once inhibition–reducing methods were applied. Table 1 shows that simple dilution of the extract is, in nearly all cases, enough to enhance the genotyping success rate suggesting that inhibitors play a critical role in reducing PCR efficiency in severely degraded samples. However long amplicon systems (leftmost in Table 1) appear to be more affected by PCR inhibitors, either due to a lower initial undegraded target concentration, a less efficient DNA polymerization, or both. In contrast SNP typing shows relative immunity from PCR inhibition effects since these systems are likely to benefit from a higher initial concentration of intact target plus more efficient (shorter) polymerization steps.

If amplicon length is an important factor with or without inhibition control then the success rate should be expected to rise as amplicon size diminishes. This was observed with the standard amplicon STRs, since *Powerplex16* showed the highest overall failure rate. This trend continued as the degree of degradation rose, and longer *Identifiler*® loci failed to amplify (amplicons up to 380 bp) followed by *MiniFiler*® (amplicon up to 300 bp). The performance of mini-STR loci between 100 and 300 bp does not differ markedly, however STR products below 100 bp, notably those of *MiniNC01* (all amplicons $<$120 bp) appear to be resistant to the most aggressive degradation in the samples of this study. It can also be suggested that a small-scale triplex PCR using amplicons much shorter than average is likely to be more efficient when intact target DNA is at low levels in the extract. In comparison SNP multiplexes clearly do not need to be small-scale in scope to achieve efficient amplification of highly degraded DNA. An additional advantage of the SNP assays developed by SNP*for*ID is that the genotyping reaction is separated from the initial PCR so amplicons of similar length, often much shorter than 100 bp, can be combined with ease.

The above findings suggest that for most degraded material standard STR typing methods will suffice if inhibition is properly assessed and controlled prior to PCR. However a further level of degradation, characterized by extremely aggressive environmental conditions over long periods of time can present the most challenging analyses. This applies to two femurs we analyzed: *78p03* was from a 35-year internment in a tomb with warm, damp, acid conditions, both epiphyses were missing and the bone had a loose, sawdust form with extensive

Table 1
Genotyping success of 15 cases involving challenging DNA, with extracts in order of increasing success rate using Powerplex16 (this assay comprises longest overall amplicon sizes)

Sample	IPC value	Quantity DNA (ng/μl)	Powerplex16		Identifiler®		MiniFiler®		52plex auto1 (23plex PCR)		34plex AIM-SNP set		52plex auto2 (29plex PCR)		MiniNC01	
			Inhibition	Corrected	Inhibition	Corrected	Inhibition	Corrected	Inhibition	Corrected	Inhibition	Corrected	Inhibition	Corrected	Inhibition	Corrected
70-06	ND	0.04	0	0	0	0	30	50	85	100	90	100	93.7	100	100	100
78P03	ND	0.53	0	0	9.3	9.3	38.8	38.8	95.6	100	100	100	96.5	100	100	100
71P04	28.03	0.67	62.5	62.5	78.12	87.5	100	100	82.6	100	100	100	100	100	100	100
126P04	28.43	1.03	78.1	78.12	62.5	100	100	100	84.8	100	95.6	95.6	100	100	100	100
122P04	ND	5.88	87.5	87.5	0	100	100	100	97.8	100	100	100	93.1	100	100	100
23P04	27.73	0.03	87.5	93.75	31.25	100	100	100	95.6	100	100	100	100	100	100	100
72P03	31.69	0.6	93.75	93.75	81.25	84.37	44.4	44.4	95.6	100	100	100	93.1	100	100	100
77P04	28.56	0.9	87.5	96.87	93.75	100	100	100	67.4	100	100	100	100	100	100	100
50p04	27.76	0.12	87.5	100	43.75	100	88.9	88.9	100	100	97	97	100	100	100	100
12P05	28	0.16	100	100	87.5	100	100	100	82.6	100	100	100	96.6	100	100	100
45P04	28.12	0.72	96.9	100	100	100	88.9	88.9	100	100	100	100	100	100	100	100
24P05	28.3	0.98	100	100	87.5	100	100	100	86.9	100	100	100	100	100	100	100
11P05	35.55	2.93	93.75	100	56.25	100	100	100	71.7	100	100	100	100	100	100	100
105-05	35.15	8.63	100	100	100	100	100	100	100	100	100	100	100	100	100	100
5P04	35.49	19.51	93.75	100	25	100	100	100	73.8	100	100	100	100	100	100	100

Values for each multiplex denote % success as proportion of full genotypes observed in undiluted extract (inhibition) and at optimum dilution (corrected). Multiplexes arranged in ascending order left to right of total genotyping success. ND = not determined, IPC = internal PCR control.

mould growth; *70–06* was from skeletal remains half buried for 10 years in forest soil (damp, acid and organic-rich conditions) that were discovered after a severe forest fire which badly charred the remaining tissue/bone surfaces. High molecular weight DNA was absent from samples and all standard STRs failed, together with most mini-STR loci. In contrast, successful amplification of the shortest amplicon mini-STRs and all SNP multiplexes indicated that short, fragmented DNA was present in enough quantity for efficient PCR of these systems. Run in parallel for these samples; the standard protocol DNA extracts amplified poorly compared to those from the ancient DNA protocol (respectively: 22.2% vs. 44% success with *MiniFiler*® and 85% vs. 100% with SNP typing). The ancient DNA extraction method had a major impact on DNA recovery and typing success for *MiniNC01* and autosomal SNPs. Improved DNA yield with the ancient DNA protocol also helped to control inhibition in all systems by allowing greater levels of dilution.

4. Conclusions

An optimum analysis method for degraded DNA can be chosen between standard STRs and short-amplicon STRs combined with SNP typing depending on several key factors: the likely state of degradation, initial visual exploration of the sample upon receipt, volume of extract obtained and quantification results, particularly IPC values detected. Generally DNA fragments up to 300 bp should be present (even in low concentration), and these can be successfully amplified with any approach if inhibition is properly managed. Our analysis of two cases with extremely degraded material indicates that SNPs and small-scale mini-STR assays amplifying DNA from extraction procedures optimized for both yield and inhibition, offer the best approach.

Conflict of interest

None.

References

[1] J. Burger, S. Hummel, B. Herrmann, W. Henke, Electrophoresis 20 (1999) 1722–1728.
[2] C. Lalueza-Fox, J. Bertranpetit, J.A. Alcover, N. Shailer, E. Hagelberg, J. Exp. Zool. 288 (2000) 56–62.
[3] D. Michael, M. Coble, J.M. Butler, J. Forensic Sci. 50 (2005) 43–53.
[4] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, Electrophoresis 27 (2006) 1713–1724.
[5] C. Phillips, A. Salas, J.J. Sanchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaz, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNP*for*ID Consortium, Forensic Sci. Int. Genetics 1 (2007) 273–280.

Forensic Science International: Genetics 3 (2009) 261–264

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

Testing the performance of mtSNP minisequencing in forensic samples

A. Mosquera-Miguel [a,1], V. Álvarez-Iglesias [a,1], M. Cerezo [a], M.V. Lareu [a], Á. Carracedo [a,b], A. Salas [a,*]

[a] Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia 15782, Spain
[b] Fundación Pública Galega de Medicina Xenómica (FPGMX), CIBERER, Hospital Clínico Universitario, Universidade de Santiago de Compostela, 15706 Galicia, Spain

ARTICLE INFO

Article history:
Received 26 December 2008
Received in revised form 20 March 2009
Accepted 14 April 2009

Keywords:
mtDNA
Coding region
HVS-I and HVS-II
SNP
Haplotype
SNaPshot
Phylogeny
Forensics
Haplogroup H

ABSTRACT

There is a growing interest among forensic geneticists in developing efficient protocols for genotyping coding region mitochondrial DNA (mtDNA) SNPs (mtSNPs). Minisequencing is becoming a popular method for SNP genotyping, but it is still used by few forensic laboratories. In part, this is due to the lack of studies testing its efficiency and reproducibility when applied to real and complex forensic samples. Here we tested a minisequencing design that consists of 71 mtSNPs (in three multiplexes) that are diagnostic of known branches of the R0 phylogeny, in real forensic samples, including degraded bones and teeth, hair shafts, and serial dilutions. The fact that amplicons are short coupled with the natural efficiency of the minisequencing technique allow these assays to perform well with all the samples tested either degraded and/or those containing low DNA amount. We did not observe phylogenetic inconsistencies in the 71 mtSNP haplotypes generated, indicating that the technique is robust against potential artefacts that could arise from unintended contamination and/or spurious amplification of nuclear mtDNA pseudogenes (NUMTs).

1. Introduction

Developing new protocols to genotype coding region mtSNPs is a topic of growing interest among forensic geneticists. The common practice of sequencing the first (and sometimes the second) hypervariable segment (HVS-I) is most of the times insufficient due to its limited power of discrimination in forensic casework samples [1]. Unfortunately, the analysis of mtDNA is often the only choice when dealing with highly degraded samples (and/or those containing low amounts of DNA), and hair shafts, basically because the presence of high copy number of mtDNA molecules per cell in comparison to the diploid copy of the nuclear DNA.

Sequencing strategies based on the amplification of small amplicons are needed when dealing with highly degraded samples, but these sequencing strategies are very time-consuming, costly, and require intense lab work. A variety of screening and genotyping strategies has been proposed in the past to analyze the variation at the coding region of the mtDNA molecule (see for instance [2,3]). During these last few years, minisequencing procedures have become popular among geneticists because this technique shows several advantages when compared to other SNP genotyping methods: (a) a solid design based on phylogenetic criteria can substantially increase the discrimination power of the mtDNA test or can be powerful for pure screening purposes [4], (b) it allows to design multiplexes involving large amounts of SNPs located at distant parts of the mtDNA molecule, (c) it is rapid and cost-effective, and (d) multiplex genotyping prevents artefactual recombinations [5–8] which can easily occur when genotyping one SNP at a time.

Several multiplex minisequencing assays were proposed to genotype coding region mtSNPs. Quintáns et al. [9] reported a multiplex protocol to target SNPs defining main West European branches. A similar multiplex assay was independently developed by Vallone et al. [10]. Brandstätter et al. [11] designed three assays aimed to genotype 45 coding region SNPs representing major and minor sub-lineages within haplogroup H, which makes-up more than 40% of the mtDNAs in European populations. SNPs representing the main trunks and twigs of the East Asian and Native American phylogenies were compiled in other multiplex assays [4,12]. Recently, Endicott et al. [13] showed the suitability of the minisequencing technique to genotype historic samples from Andaman Islanders. Finally, minisequencing assays were also developed to successfully genotype autosomal SNPs [14] of forensic interest.

The performance of a minisequencing multiplex assay is highly dependent on amplicon sizes (as it is also with other techniques and markers, e.g. STRs); it is well known that small amplicons

* Corresponding author. Tel.: +34 981 582327; fax: +34 981 580336.
E-mail address: antonio.salas@usc.es (A. Salas).
[1] Both authors contributed equally to this work.

1872-4973/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2009.04.006

Table 1
List of SNPs genotyped in the present study.

Marker	Allele	Multiplex	Marker	Allele	Multiplex	Marker	Allele	Multiplex
709	G-A	1	951	G-A	2	1,438	A-G	3
750	A-G	1	961	T-G	2	2,259	C-T	3
2581	A-G	1	3915	G-A	2	5250	T-C	3
2706	A-G	1	3936	C-T	2	5263	C-T	3
3010	G-A	1	3992	C-T	2	8869	A-G	3
3796	A-G	1	4310	A-G	2	8994	G-A	3
3847	T-C	1	4336	T-C	2	9336	A-G	3
4550	T-C	1	4727	A-G	2	10166	T-C	3
4580	G-A	1	4745	A-G	2	10211	C-T	3
6253	T-C	1	4769	A-G	2	11140	C-T	3
6296	C-T	1	4793	A-G	2	11719	G-A	3
6365	T-C	1	7028	C-T	2	12308	A-G	3
6776	T-C	1	7645	T-C	2	12438	T-C	3
7337	G-A	1	8269	G-A	2	12705	C-T	3
10810	T-C	1	8271	A-T	2	13101	A-C	3
12858	C-T	1	8473	T-C	2	13105	A-G	3
12957	T-C	1	8592	G-A	2	14869	G-A	3
13708	G-A	1	8598	T-C	2	14872	C-T	3
13759	G-A	1	8602	T-C	2	15452	C-A	3
14365	C-T	1	9066	A-G	2	15773	G-A	3
14470	T-A	1	9088	T-C	2	15833	C-T	3
14766	C-T	1	9150	A-G	2	15904	C-T	3
14770	C-T	1	10044	A-G	2			
15218	A-G	1	10394	C-T	2			
			13404	T-C	2			

generally yield better results in low quality forensic samples. This requirement is only reached by some previous assays, but covering different parts of the phylogeny [4] or with a more broad phylogenetic coverage (e.g. West European branches [9]). Recently, we developed a minisequencing design that considers 71 mtSNPs defining different branches of macro-haplogroup R0, which embraces the common West European haplogroup H [15]. The main goal of the present study is to evaluate its performance when applied to real forensic samples, given the fact that these assays were deliberately designed to yield small amplicon sizes.

2. Materials and methods

2.1. Forensic samples

We have analyzed three different hair shafts of 1 cm long collected from three unrelated donors. DNA extraction was carried out using the Bio Robot EZ1 robot (Qiagen; Hilden, Germany) using the manufacture protocols. In addition, two bone and two teeth samples were extracted using a phenol–chlorophorm protocol from the Instituto de Medicina Legal of Santiago de Compostela (Spain). These samples were highly degraded, as indicated by the several unsuccessful attempts to amplify autosomal STRs using the popular commercial kits in the field (e.g. PowerPlex from Promega and Identifiler from Applied Biosystem).

The samples were quantified using the Applied Biosystems' Quantifiler human DNA quantification kit (Applied Biosystems, Foster City, US), and also the IPC (internal PCR control) PCR inhibitor detecting feature. Quantification values for the forensic samples were as follows (see Table S1 for sample labels): 0.981 ng/μl for Femur_1; 1.03 ng/μl for Tooth_1; 0.05 ng/μl for Tooth_2. The other samples yielded undetermined values.

2.2. Serial DNA dilution samples

Four different blood samples were selected for serial dilution experiments. The samples were extracted and quantified as indicated above.

The reactions were run on an Applied Biosystems 7300 real time PCR device following the manufacturer's specifications.

Finally, all the samples were diluted as follows: 1000, 500, 200, 100, 50, 25, and 10 pg/μl, and genotyped for the whole set of SNPs considered in the present report.

2.3. SNP selection and SNaPshot design

All the details concerning SNP selection and SNaPshot™ (Applied Biosystem) design are given in Álvarez-Iglesias et al. [15]. The design consists of three different multiplex reactions aimed to genotype 71 SNPs that are diagnostic of different haplogroup R0 branches (Table 1).

Positive and negative controls were used during the whole extraction, amplification and genotyping processes. Negative controls are mock extractions run with the same reagents as the sample extractions, but with not sample added.

2.4. Automatic sequencing

All the forensic samples and the serial dilution sample of 500 pg/μl were sequenced in forward and reverse directions for the HVS-I segment. An indirect indication of the high level of DNA degradation of the forensic samples was the fact that the standard protocols based on PCR of large amplicons (~400 bp) such as those described in Álvarez-Iglesias et al. [4] did not perform well. Most of the samples were therefore amplified and sequenced using small amplicons (~250 bp) in at least two independent and overlapping reactions. The PCR was performed in 22 μl of reaction mixture containing 50 μM of each dNTP, 2.5 U Taq DNA Polymerase, recombinant (Invitrogen, Carlsbad, CA, USA), 1× PCR Buffer, 1.5 mM of Magnesium Chloride, 0.16 μM of BSA, 0.2 μM of each primer forward and reverse (see also Álvarez-Iglesias for primers that yield smaller amplicons [4]) up to 22 and 3 μl sample template. PCR amplification was carried out in a thermocycler GenAmp PCR System 9700 (Applied Biosystems) with the following conditions: one cycle of 95 °C for 1 min; then 36 cycles of 95 °C for 30 s, 55 °C for 1 min and 72 °C for 30 s, and ending 15 °C for 10 min. PCR products and negative controls were checked in poliacrilamide gel and visualized with silver staining. Then PCR products were purified to remove excess of primers and un-incorporated dNTPs with a treatment with ExoSAP-IT (GE

Healthcare, Uppsala, Sweden): 2.15 μl of PCR product was incubated with 0.85 μl for 15 min at 37 °C followed by 15 min at 80 °C for enzyme inactivation. Sequencing reaction was performed in 11.5 μl of reaction mixture, containing 2.5 μl of sequencing buffer (5×), 0.5 μl of BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems), 1 μl of the corresponding primer (final concentration was 1 μM), 3 μl of the purified PCR product and water up to 11.5 μl. Sequencing reaction was carried out in a thermocycler GenAmp PCR System 9700 (Applied Biosystems) with one cycle of 96 °C for 3 min and then 25 cycles of 96 °C for 30 s, 50 °C for 15 s and 60 °C for 4 min. Sequencing products were doubled purified following a treatment with SAP (GE Healthcare). The final volume (11.5 μl) was treated with 1 μl of SAP for 60 min at 37 °C followed by 15 min at 80 °C for enzyme inactivation, followed by purification using Montage™ SEQ96 Sequencing Reaction Cleanup Kit (Millipore) according to manufacturer protocols. MtDNA automatic sequencing was carried out in a capillary electrophoresis ABI3130xl™ (Applied Biosystems).

2.5. Control for genotyping errors

We have used the mtDNA worldwide phylogeny as a reference for checking phylogenetic inconsistencies and avoiding as much as possible artefactual results and documentation errors [5,8,15–19].

3. Results and discussion

A total of 71 coding region SNPs covering the internal variability of macro-haplogroup R0 were genotyped in a set of real casework forensic samples. The average amplicon size of these multiplexes is short (135 bp, SD =35), ranging from 66 to 195 bp, increasing their efficiency and applicability to forensic samples. Thus, the success rate of the multiplex assays was ~100% for a well-preserved DNA control sample; that is, a sample having acceptable DNA amount (>1–2 ng) and DNA of good quality. The minisequencing assays performed equally well for the forensic samples considered in the present study (see Table S1). Fig. 1 shows two minisequencing profiles (Fig. 1A=haplogroup R; Fig. 1B=haplogroup H1) that correspond to samples #Femur_1 and #Tooth_1 (Table S1), respectively. Reagent blank extraction controls associated with the casework extracts were also investigated. We did not observe false positives due to contamination. Figure S1 shows the electropherogram of a reagent blank sample.

Serial dilution experiments indicate that the average genotyping success (measured by the relative number of SNPs genotyped unambiguously) was very high in those samples containing 1000–50 pg (average genotyping success = 99.6%). Genotyping success was significantly lower for dilutions containing 25 pg (~72% on average). Dilutions of 10 pg experienced a clear lost of electrophoretic signal and in fact only 45% (on average) of the SNPs could

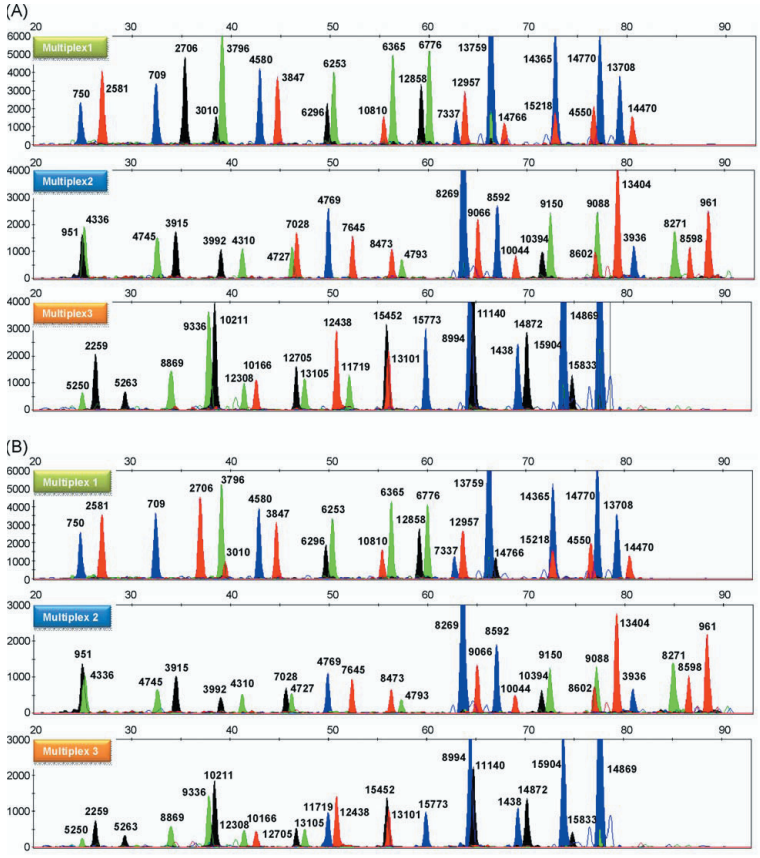


Fig. 1. Haplogroup R* (A) and H1 (B) SNP profiles for samples Femur_1 and Tooth_1, respectively (see Table S1).

be genotyped unambiguously. We are aware that the relative amount of mtDNA and nuclear DNA varies between tissues [20] and therefore the extrapolation of the present figures to other tissues has to be taken with care.

We observed small correlation (Pearson's correlation coefficient; $r^2 = 0.47$) between the three multiplex designs and their corresponding genotyping success. This low correlation was apparently unrelated to the average amplicon size of the multiplexes or their respective number of SNPs.

Finally, SNP typing allowed the allocation of most of the samples to their specific sub-haplogroups; that is, to a much higher phylogenetic resolution that the one provided by the control region alone (see Table S1).

To summarize, we have successfully minisequenced 71 SNPs spanning the mtDNA molecule in highly degraded, low amount DNA, and serial dilution samples. However, below 25 pg/μl of template, the efficiency of the minisequencing assays employed here is moderately reduced. These mtSNPs increase considerably the discrimination power of the mtDNA test in samples of European ancestry. Apart from being more cost-effective than many other alternative techniques (e.g. RFLP genotyping), minisequencing seems to be reliable when genotyping suboptimum samples; in fact, we did not observed suspicious SNP patterns that could indicate, e.g. artificial sample mix-up [20,21] or involuntary amplification of NUMTs [4,22].

Acknowledgement

We would like to thak Meli Rodriguez, Raquel Calvo, and Manuel Fondevila for their technical assistance. This work was supported by grants from the Xunta de Galicia (Grupos Emerxentes; 2008/037), Ministerio de Ciencia e Innovación (SAF2008-02971), and Fundación de Investigación Médica Mutua Madrileña (2008/CL444) given to AS.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2009.04.006.

References

[1] A. Salas, H.-J. Bandelt, V. Macaulay, M.B. Richards, Phylogeographic investigations: the role of trees in forensic genetics, Forensic Sci. Int. 168 (2007) 1–13.
[2] F. Barros, M.V. Lareu, A. Salas, A. Carracedo, Rapid and enhanced detection of mitochondrial DNA variation using single-strand conformation analysis of superposed restriction enzyme fragments from polymerase chain reaction-amplified products, Electrophoresis 18 (1997) 52–54.
[3] A. Salas, E.M. Rasmussen, M.V. Lareu, N. Morling, A. Carracedo, Fluorescent SSCP of overlapping fragments (FSSCP-OF): a highly sensitive method for the screening of mitochondrial DNA variation, Forensic Sci. Int. 124 (2001) 97–103.
[4] V. Álvarez-Iglesias, J.C. Jaime, Á. Carracedo, A. Salas, Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups, Forensic Sci. Int.: Genet. 1 (2007) 44–55.
[5] H.-J. Bandelt, Q.-P. Kong, W. Parson, A. Salas, More evidence for non-maternal inheritance of mitochondrial DNA? J. Med. Genet. 42 (2005) 957–960.
[6] H.-J. Bandelt, A. Salas, C.M. Bravi, Problems in FBI mtDNA database, Science 305 (2004) 1402–1404.
[7] H.-J. Bandelt, A. Salas, S. Lutz-Bonengel, Artificial recombination in forensic mtDNA population databases, Int. J. Legal Med. 118 (2004) 267–273.
[8] A. Salas, Y.-G. Yao, V. Macaulay, A. Vega, Á. Carracedo, H.-J. Bandelt, A critical reassessment of the role of mitochondria in tumorigenesis, PLoS Med. 2 (2005) e296.
[9] B. Quintáns, V. Álvarez-Iglesias, A. Salas, C. Phillips, M.V. Lareu, Á. Carracedo, Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, Forensic Sci. Int. 140 (2004) 251–257.
[10] P.M. Vallone, R.S. Just, M.D. Coble, J.M. Butler, T.J. Parsons, A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome, Int. J. Legal Med. 118 (2004) 147–157.
[11] A. Brandstätter, A. Salas, H. Niederstätter, C. Gassner, Á. Carracedo, W. Parson, Dissection of mitochondrial superhaplogroup H using coding region SNPs, Electrophoresis 27 (2006) 2541–2550.
[12] K. Umetsu, M. Tanaka, I. Yuasa, N. Adachi, A. Miyoshi, S. Kashimura, K.S. Park, Y.H. Wei, G. Watanabe, M. Osawa, Multiplex amplified product-length polymorphism analysis of 36 mitochondrial single-nucleotide polymorphisms for haplogrouping of East Asian populations, Electrophoresis 26 (2005) 91–98.
[13] P. Endicott, M. Metspalu, C. Stringer, V. Macaulay, A. Cooper, J.J. Sánchez, Multiplexed SNP typing of ancient DNA clarifies the origin of Andaman mtDNA haplogroups amongst south Asian tribal populations, PLoS ONE 1 (2006) e81.
[14] J.J. Sánchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 13–24.
[15] V. Álvarez-Iglesias, A. Mosquera-Miguel, M. Cerezo, B. Quintáns, M.T. Zarrabeitia, I. Cuscó, M.V. Lareu, O. García, L. Pérez-Jurado, Á. Carracedo, et al., New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0, PLoS ONE 4 (2009) e5112.
[16] Y.-G. Yao, V. Macaulay, T. Kivisild, Y.-P. Zhang, H.-J. Bandelt, To trust or not to trust an idiosyncratic mitochondrial data set, Am. J. Hum. Genet. 72 (2003) 1341–1346 (author reply 1346–1349).
[17] A. Salas, Á. Carracedo, V. Macaulay, M. Richards, H.-J. Bandelt, A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics, Biochem. Biophys. Res. Commun. 335 (2005) 891–899.
[18] H.-J. Bandelt, P. Lahermo, M. Richards, V. Macaulay, Detecting errors in mtDNA data by phylogenetic analysis, Int. J. Legal Med. 115 (2001) 64–69.
[19] H.-J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, Am. J. Hum. Genet. 71 (2002) 1150–1160.
[20] M. Montesino, A. Salas, M. Crespillo, C. Albarran, A. Alonso, V. Alvarez-Iglesias, J.A. Cano, M. Carvalho, D. Corach, C. Cruz, et al., Analysis of body fluid mixtures by mtDNA sequencing: an inter-laboratory study of the GEP-ISFG working group, Forensic Sci. Int. 168 (2007) 42–56.
[21] L. Prieto, A. Alonso, C. Alves, M. Crespillo, M. Montesino, A. Picornell, A. Brehm, J.L. Ramírez, M.R. Whittle, M.J. Anjos, et al., 2006 GEP-ISFG Q1 collaborative exercise on mtDNA. Reflections about interpretation, artefacts, and DNA mixtures, Forensic Sci. Int. Genet. 2 (2008) 126–133.
[22] Y.-G. Yao, Q.-P. Kong, A. Salas, H.-J. Bandelt, Pseudo-mitochondrial genome haunts disease studies, J. Med. Genet. 45 (2008) 769–772.

IV.3 CLINICAL GENETICS

IV.3.1. Article 13: High mitochondrial DNA stability in B-Cell Chronic Lymphocytic Leukemia *PLoS ONE*

OPEN ACCESS Freely available online

PLoS one

High Mitochondrial DNA Stability in B-Cell Chronic Lymphocytic Leukemia

María Cerezo[1], Hans-Jürgen Bandelt[2], Idoia Martín-Guerrero[3], Maite Ardanaz[4], Ana Vega[5], Ángel Carracedo[1], África García-Orad[3], Antonio Salas[1]*

1 Unidade de Xenética, Instituto de Medicina Legal, and Departamento de Anatomía Patolóxica y Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain, **2** Department of Mathematics, University of Hamburg, Hamburg, Germany, **3** Laboratorio Interdepartamental de Medicina Molecular, Departamento de Genética Antropología Física y Fisiología Animal, Facultad de Medicina, Universidad del País Vasco- Euskal Herriko Unibertsitatea, Leioa, Spain, **4** Servicio de Hematología, Hospital Txagorritxu, Vitoria, Spain, **5** Fundación Pública Galega de Medicina Xenómica (FPGMX), Hospital Clínico Universitario, Universidad de Santiago de Compostela, Galicia, Spain

Abstract

Background: Chronic Lymphocytic Leukemia (CLL) leads to progressive accumulation of lymphocytes in the blood, bone marrow, and lymphatic tissues. Previous findings have suggested that the mtDNA could play an important role in CLL.

Methodology/Principal Findings: The mitochondrial DNA (mtDNA) control-region was analyzed in lymphocyte cell DNA extracts and compared with their granulocyte counterpart extract of 146 patients suffering from B-Cell CLL; B-CLL (all recruited from the Basque country). Major efforts were undertaken to rule out methodological artefacts that would render a high false positive rate for mtDNA instabilities and thus lead to erroneous interpretation of sequence instabilities. Only twenty instabilities were finally confirmed, most of them affecting the homopolymeric stretch located in the second hypervariable segment (HVS-II) around position 310, which is well known to constitute an extreme mutational hotspot of length polymorphism, as these mutations are frequently observed in the general human population. A critical revision of the findings in previous studies indicates a lack of proper methodological standards, which eventually led to an overinterpretation of the role of the mtDNA in CLL tumorigenesis.

Conclusions/Significance: Our results suggest that mtDNA instability is not the primary causal factor in B-CLL. A secondary role of mtDNA mutations cannot be fully ruled out under the hypothesis that the progressive accumulation of mtDNA instabilities could finally contribute to the tumoral process. Recommendations are given that would help to minimize erroneous interpretation of sequencing results in mtDNA studies in tumorigenesis.

Citation: Cerezo M, Bandelt H-J, Martín-Guerrero I, Ardanaz M, Vega A, et al. (2009) High Mitochondrial DNA Stability in B-Cell Chronic Lymphocytic Leukemia. PLoS ONE 4(11): e7902. doi:10.1371/journal.pone.0007902

Editor: Iris Schrijver, Stanford University, United States of America

Received July 31, 2009; **Accepted** October 20, 2009; **Published** November 18, 2009

Funding: Two grants from Fundación de Investigación Médica Mutua Madrileña (2006/CL370 and 2008/CL444), a grant "Grupos Emerxentes" from Xunta de Galicia (2008/XA122), and a grant from the Ministerio de Ciencia e Innovación (SAF2008-02971) awarded to A.S. supported this project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: antonio.salas@usc.es

Introduction

Chronic Lymphocytic Leukemia (CLL) becomes manifest in progressive accumulation of lymphocytes in the blood, bone marrow, and lymphatic tissues [1]. B-CLL is the most frequent form of leukemia in Western countries and represents 30% of all leukemic cases [2]. Studies on ethnic distribution of CLL in the world have shown considerable variation [3], Thus for instance, according to the American Cancer Society (http://www.cancer.org), about 15,490 new cases of CLL will be diagnosed in the United States during 2009, and about 4,390 people will die of CLL in this country during this year. For patients with progressing CLL, treatment with conventional doses of chemotherapy is not curative; selected patients treated with allogeneic stem cell transplantation achieved prolonged disease-free survival [4,5]. The median survival for all patients ranges from 8 to 12 years in older trials [6]. CLL occurs primarily in middle-aged and elderly adults, with increasing frequency in successive decades of life. The clinical course of this disease progresses from an indolent lymphocytosis without other evident disease to one of generalized lymphatic enlargement with concomitant pancytopenia. Complications of pancytopenia, including haemorrhage and infection, represent a major cause of death in these patients [7]. Prognostic factors that may help predict clinical outcome include cytogenetic subgroup, immunoglobulin mutational status, ZAP-70, and CD38 (see for instance [1]). Staging is useful in CLL to predict prognosis and also to stratify patients to achieve comparisons for interpreting specific treatment results. Anemia and thrombocytopenia are the major adverse prognostic variables. Although CLL has no standard staging system, the most common ones are the Rai staging system and the Binet classification [1,8]. New prognostic markers have been proposed to the clinician and investigator [9].

The potential role of the mtDNA genome in CLL was first approached by Carew et al. [10] analyzing a sample of 20 patients. According to these authors, chemotherapy with DNA-damaging agents could cause mtDNA mutations in primary leukemia cells,

which often exist in heteroplasmic condition. These findings were later put into question by Meierhofer et al. [11], who showed that platelet transfusion can mimic somatic mtDNA mutations. He et al. [12] sequenced the entire mtDNA from both normal tissue (buccal epithelial cells) and cells extracted from bone marrow in 24 patients with adult-onset leukemia. They reported mtDNA mutations in nine tumoral mtDNAs (one mutation per patient) and, in particular, inferred pathogenic implications of the mutation A15296G in leukemia. Grist et al. [13] analyzed mtDNA mutations in 22 patients with acute myeloid leukemia (AML) and 26 patients with acute lymphoblastic leukemia (ALL). The authors found (multiple) mtDNA control-region mutations in 36% of the AML patients and in 58% of the ALL patients, but the sequence information was not provided for closer examination. It was pointed out by the authors that most of the mutations tended to appear at hotspots: "*several hotspots were at sites of poly C tracts, but there was no single-sequence motif which seemed to be associated with mutations*" [13]. In a comprehensive review, Gattermann [14] analyzed the implications of mtDNA mutations in leukemogenesis. Recently, Yao et al. [15] studied mtDNA sequence variation in more than 3,500 single normal cells and individual blasts from 18 patients with leukemia and 10 healthy donors. Further they found that the somatic mutation process in leukemia is complex and generally leads to diverse levels of genetic alterations. These authors also observed that the somatic mutation events in single hematopoietic cells are prone to occur at well-characterized mutation hot spots, thus corroborating the results previously obtained in tumors of the central nervous system [16].

He et al. [12] conducted the first mtDNA study that aimed at comparing mtDNA extracted from leukemic cells with mtDNA extracted from buccal mucosa cells from the same patient, observing that 40% of their patients bear somatic point mutations in their mtDNAs. Other studies[17,18] compared mtDNA from clonal bone marrow disorders with mtDNA obtained from healthy individuals; as advanced by Gattermann [14], these studies however may be futile since these comparisons just reflect the biological differences that exist between mtDNA lineages within or between human population groups (see also [19]). It has been claimed in numerous studies that the mtDNA molecule is in general prone to instability in tumorigenesis, and even hypothesized that mtDNA genome instability could play some active role during the development of the tumor in several types of cancer. Most of these studies, however, are most problematic in view of the patterns of recorded mtDNA alterations between putatively matched tissues, which are more akin to the result of sample mixing [19,20,21]. This prompted us to take a closer look at the previous mtDNA sequencing results in patients with CLL, AML, and ALL (especially those from Carew et al. [10]) in order to exhibit patterns that would clearly point to artefacts or other shortcomings.

To shed more light on the debatable issue of mtDNA alterations in leukemia we analyzed mtDNA instability in a large number of B-CLL patients from the same geographic area (Basque country) under strict laboratory conditions. Since previous studies always found alterations in the control-region, we sequenced the entire mtDNA control-region in lymphocytes obtained from blood samples of the B-CLL patients. Since the myeloid line and, in particular, the granulocytes should not be directly affected in B-CLL patients, these cells were used as the corresponding control samples for instability in the lymphoid line of each patient. In contrast to previous attempts, our design involved the comparison of two different cell lines that have very recent common cell ancestors (the colony deriving from the pluri-potent stem cells) in order to maximally rule out the effect of tissue-mediated instabilities. Such instabilities exist even when comparing different fragments of hairs from the same (healthy) individual [22,23].

Materials and Methods

Ethical Statement

The study was conducted according to the Spanish Law for Biomedical Research (Law 14/2007- 3 of July) and complied with the Declaration of Helsinki. The study and the use of archive samples for this project was approved by the Ethics Committee of the University of Santiago de Compostela where the study was carried out. Written informed consent was obtained for all patients. All the samples were collected anonymously.

Sample Collection and DNA Extraction

A total of 146 blood samples from patients diagnosed of B-CLL were collected (2–4 ml) by venous puncture using EDTA as anticoagulant. The patients belong to the sanitary area of the Hospital de Tagorritxu and Basurto in the Basque country. Written informed consent was obtained for all individuals.

Granulocyte and lymphocyte cells were separated using a gradient Ficoll-Paque Plus (Amhersham Biosciences) using manufacture protocols (see also Carew et al. [10] for a similar approach). According to manufacture instructions, we did not remove excess Ficoll-Paque PLUS in order to minimize contamination of the lymphocytes fraction with granulocytes. Platelets were removed as indicated in those instructions. These protocols minimize the presence of non-lymphocyte cells in the lymphocyte fraction and lead to the purification of the lymphoid extraction for the presence of lymphocytes.

We also recovered the granulocyte cell fraction. According to manufacture instructions, the efficiency of the separation is as follows: lymphocytes ~95% of cells present in fraction are mononucleocytes, and ~60% recovery of lymphocytes from the original blood sample. There are other cells in the extract: ~3% granulocytes, ~3.5% erythrocytes, and <0.5% of total platelets in the original blood sample. By using washing protocols, we intended to further improve the cell separation and purification of the lymphocyte and granulocyte cell fractions. Note also that standard sequencing protocols cannot detect the presence of a minor non-lymphoid cell component in the lymphocyte extract if this component is below ~10% of the mixture. In any case, the presence of such cell mixture would not mask the mutational difference (if any) between both cell lines; this difference would be generally detected as a heteroplasmic-like pattern.

After the two fractions were obtained, we extracted the DNA using a standard phenol-chloroform protocol [24]; quantification was performed using GeneQuant Pro (Pharmacia), which showed variable concentrations varying between 10 to 100 ng/µl.

PCR Amplification and Automatic Sequencing of the mtDNA Control-Region

The PCR was performed in 22 µl of reaction mixture containing 50 µM of each dNTP (200 µM of GeneAmp® 10 mM dNTP Mix with dTTP, Applied Biosystems [AB], Foster City, CA, USA), 2,5U *Taq* DNA Polymerase, recombinant (Invitrogen, Carlsbad, CA, USA), 1X PCR Buffer, 1,5 mM of Magnesium Chloride, 016 µM of BSA, 0,2 µM of each primer forward and reverse (15997L and 017H for HVS-I, and 16555L and 599H for HVS-II; Table 1), water-up to 22 µl and 3 µl sample template. PCR amplification was carried out in a thermocycler GenAmp PCR System 9700 (AB) with the following conditions: one cycle of 95°C for 1 minute; then 35 cycles of 95°C for 10 seconds, 55°C for 30 seconds and 72°C for 30 seconds, and ending

Table 1. Set the primers used for PCR amplification and sequencing.

mtDNA segment	PCR primer (5′ to 3′)		Sequencing primers (5′ to 3′)	
HVS-I (16024-16569)	15997L	CACCATTAGCACCCAAAGCT	15997L	CACCATTAGCACCCAAAGCT
	–	–	16254L	CACATCAACTGCAACTCCAAA
	–	–	16236H	CTTTGGAGTTGCAGTTGATG
	–	–	16401H	TGATTTCACGGAGGATGGTG
	–	–	16450H	CAAGTGTTATGGGCCCGGAGC
	–	–	16380L	TCAGATAGGGGTCCCTTGAC
	017H	CCCGTGAGTGGTTAATAGGGT	017H	CCCGTGAGTGGTTAATAGGGT
HVS-II (1-578)	16555L	CCCACACGTTCCCCTTAAAT	16555L	CCCACACGTTCCCCTTAAAT
	–	–	172L	ATTATTTATCGCACCTACGT
	–	–	285H	GGGGTTTGGTGGAAATTTTTTG
	332L	CCCGCTTCTGGCCACAGCAC	332L	CCCGCTTCTGGCCACAGCAC
	–	–	370L	CCCTAACACCAGCCTAACCA
	–	–	408H	CTGTTAAAAGTGCATACCGCCA
	599H	TTGAGGAGGTAAGCTACATA	599H	TTGAGGAGGTAAGCTACATA
	901H	ACTTGGGTTAATCGTGTGACC	901H	ACTTGGGTTAATCGTGTGACC

doi:10.1371/journal.pone.0007902.t001

15°C for 10 minutes. For those samples showing insertions at 573, we additionally used the amplification pair of primer 332L [15] and primer 901H [25] in order to obtain forward and reverse reading (Table 1).

After this reaction, PCR products and negative controls were checked in polyacrylamide gel and visualized with silver staining. Then PCR products were purified to remove excess of primers and un-incorporated dNTPs in MultiScreen®PCRμ_{96} plates (Millipore, Bedford, MA, USA) according to the manufacturer protocol.

Sequencing reaction was performed in 11.5 µl of reaction mixture, containing 2.5 µl of sequencing buffer (5X), 0.5 µl of BigDye Terminator v3.1 Cycle Sequencing Kit (AB), 1 µl of the corresponding primer (final concentration was 1 µM), 3 µl of the purified PCR product and water up to 11.5 µl. Sequencing reaction was carried out in a thermocycler GenAmp PCR System 9700 (AB) with one cycle of 96°C for 3 minutes and then 25 cycles of 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes or was carried out in a 9800 Fast Thermal Cycler (AB) with one cycle of 96°C for 1 minute then 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 1 minute. To obtain 'clean' electropherograms, the sequencing product was doubly purified, first using Montage^{TM} SEQ$_{96}$ Sequencing Reaction Cleanup Kit (Millipore) according to manufacturer protocols, followed by purification with Sephadex^{TM} G-10 (Amersham Biosciences, Uppsala, Sweden), the latter also according to manufacturer protocol. MtDNA automatic sequencing was carried out in a capillary electrophoresis ABI3730 (AB). Each pair of samples showing differences between granulocytes and lymphocytes were sequenced in both forward and reverse directions. In order to obtain clear pattern of instability or simply to allow the reading of length variability, additional *ad hoc* primers were used (the full list of primers is given in Table 1).

All instabilities found at this stage were first rechecked in the same laboratory through all the steps involving PCR amplification and sequencing in forward and reverse direction for both the granulocytes and the lymphocytes counterparts by using a slightly different protocol which allows to minimize potential technical artefacts. The PCR was performed in 10 µl of reaction mix, containing 4 µl of *Taq* PCR Master Mix (Qiagen, Hilden, Germany), 0.5 µl 1 µM of each primer, 1 µl sample template and 4 µl of water. This PCR was carried out in a thermocycler GenAmp PCR System 9700 (AB) with one cycle of 95°C for 15 minutes and then 35 cycles of 94°C for 30 seconds, 58°C for 90 seconds and 72°C for 90 seconds with a full extension cycle of 72°C for 10 minutes. The PCR product was purified with ExoSAP-IT (Amersham Biosciences): 2.15 µl of PCR product was incubated with 0.85 µl ExoSAP-IT for 20 min at 37°C followed by 15 min at 80°C for enzyme inactivation. The next steps (e.g. sequencing, purification) were carried out using the protocol described above.

Assessing Sequence Quality

The SeqScape v.2.5 (AB) was set up to automatically detect the presence of heteroplasmic-like patterns involving at least 15% of the signal for the minor variant; in addition, all the electropherograms were inspected visually. In order to avoid erroneous interpretation of seeming DNA instabilities [19] we have followed a phylogenetic framework [19,26,27,28] that allowed us to detect some errors committed during the analytical and documentation process. Indeed, the analysis of DNA (usually through automatic sequencing) has always been prone to errors of different nature [26,27,29,30,31,32,33,34,35].

All sequence instabilities detected in our set of samples were finally confirmed by sequencing the forward and reverse strains and replicated in a different laboratory by a different analyst. The laboratory where the replication was carried out was not informed about the sequence profiles expected for the DNA samples in order to rule out any bias in reading and interpreting sequencing electropherograms. Control samples (those without any apparent instability) were also submitted to the second lab as well as samples showing seeming instabilities in the forward but not replicated in the reverse sequencing. All the probable instabilities observed in the first laboratory were confirmed in the second laboratory. The sequencing results obtained for all the samples are presented in Data S1.

Table 2. Clinico-pathological characteristics of patients with B-CLL.

Age (years)		
	range	33 to 92
	mean	69
Gender		
	Male	~57%
	Female	~43%
Origin		
	Araba (Basque Country, Spain)	~90%
	Gupuzkoa (Basque Country; Spain)	~1%
	Burgos (Castilla; North-central Spain)	~1%
Transplant		
	none	100%
Inmuno-phenotype		
	19+ and 5+	~95%
	FMC7-	~95%
	79b-	~92%
	23+	~91%
	38-	~85%
	10-	~76%
	22-	~66%
	k+	~44%
	k-	~33%
	a-	~29%
	a+	~26%
	22+	~25%
	38+	~8%
	79b+	~4%
CD38		
	>30%	~6%
	<30%	~94%
Serological markers		
	LDH high	~14%
	LDH normal	~86%
	B2MG high	~45%
	B2MG normal	~55%
Tissue morphology		
	typical	~82%
	atypical	~18%
Survival		
	alive	~92%
	exitus	~8%
RAI classification		
	stage 0	~54%
	stage I	~31%
	stage II	~6%
	stage III	~1%
	stage IV	~8%
Binet staging		
	stage A	~88%
	stage B	~2%
	stage C	~10%
Adenopaties		
	yes	~32%
	no	~68%
Marrow infiltration		
	non-diffuse	~87%
	diffuse	~13%
Electromagnetic radiation		
	yes	~7%
	no	~83%
Treatment		
	none	~68%
	Leukeran	~18%
	Ciclosfosfamida	~9%
	Fluradabina	~8%
	Chop	~8%
	Clorambucil	~6%
	anti CD20	~6%
	Prednisona	~4%

doi:10.1371/journal.pone.0007902.t002

Identification of the Same Donor for Each Pair of Lymphocyte and Granulocyte Samples

For all those samples showing instability-like patterns we genotyped a set of microsatellites in the lymphocyte fraction and their counterpart granulocytes. This (along with sharing the same mtDNA profile) allowed us to corroborate a common biological source for these pairs of samples and therefore practically rule out potential sample mix-up. We have followed the protocols of the Instituto de Medicina Legal of the Universidad de Santiago de Compostela. The following STR autosomal markers were analyzed using PowerPlex® 16 System (Promega; Madison, USA): D21S11, D3S1358, PENTA-E, D16S539, CSF1PO, FGA, PENTA-D, TPOX, TH01, vWA, D8S1179, D18S51, D5S818, D7S820, D13S317, and amelogenin. Samples were additionally genotyping using PowerPlex® which contains the same STRs as PowerPlex® (with the exception of PENTA-D and PENTA-E) but has in addition the STR markers D19S433 and D2S1338.

Phylogenetic Analysis and Database Comparisons

Polymorphisms are referred with respect to the revised Cambridge Reference Sequence (rCRS [36]). Haplogroup classification was carried out *alter alia* according to ref. [37,38,39,40,41,42,43,44], and using the most up to date mtDNA tree from http://www.phylotree.org/. A worldwide mtDNA database of control-region sequences published in the literature and/or Genbank was used for searching mtDNA profiles; this database also contains Basque and other Iberian profiles (e.g. [45,46,47,48]).

Statistical Analysis

Pearson's Chi-squared test with Yates' continuity correction was computed in order to test for possible association between clinical-pathological variants (Table 2) and the amount of instabilities

found and haplogroup status. In order to adjust *P*-values for multiple tests, we applied Bonferroni correction and the procedure [49] to control the False Discovery Rate at the level of $\alpha = 0.05$: (i) for m tests, the *P*-values are ranked in ascending order $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$, (ii) denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$; let k be the largest i for which

$$P_{(i)} \leq \frac{\alpha}{m} i$$

and (iii) all null hypotheses $H_{(1)} \ldots H_{(k)}$ are rejected.

Statistical evaluation of the common biological source for pairs of tumor/non-tumor samples showing mtDNA instabilities has been carried out according to standard forensic conventions for sample identification. Thus, the software Familias 1.81 [50] was used to compute the likelihood ratio (LR) that considers the probability of the evidence given two alternative hypotheses: the samples came from the same donor *versus* the samples came from different donors). LR was in all the cases higher than 10^{13} (Data S2).

Results

Reassessment of the Data from Carew et al. (2003)

Carew et al. [10] analyzed four mtDNA fragments from peripheral blood samples from 20 patients with B-CLL (10 untreated and 10 treated with chemotherapy). According to the authors, their results "*revealed that primary CLL cells from patients with prior chemotherapy has a significantly higher frequency of heteroplasmic mutations than did those from untreated patients*" [10]. Note however, that no attempt was made to determine the effect of chemotherapy in one and the same patient, and no attempt was made to distinguish mutations in the blood fraction (lymphocytes, in particular) from germline mutations. Therefore, from the outset, the design of the experiments is not really appropriate for understanding the role of mtDNA mutations in leukemia since there is no way to know whether the differences observed between treated *versus* untreated patients were due to normal variation between individuals rather than a higher rate of instability in treated patients.

The mtDNA regions that could be analyzed for mutations are the amplified fragments minus primer locations, i.e. 35–464 (covering the second hypervariable segment, HVS-II), 3324–3806, 7665–8296, 8560–9039, 11424–11905, and 15281–15752. The obtained sequence fragments were apparently compared to some version of the Cambridge Reference Sequence (CRS) but obviously not to the sequence NC_18007.4 as asserted by Carew et al. (2003), which e.g. bears the changes A73G, C150T, T195C, A263G, 309+C, 315+C, and T408A relative to the rCRS. The rCRS is a member of haplogroup H2a2, the root of which is distinguished from rCRS by the four changes A263G, 315+C, A8860G, and A15326G [51]. These changes should have been observed in almost all mtDNA sequences under study. However, the latter two mutations were nowhere recorded in those tables [10] and, in addition, the former two are lacking in patients UT7 and UT8.

The authors seem to have then interpreted their sequencing results under the wrong premise that the amount of differences to the rCRS could be indicative of cancer [19]. However, the worldwide phylogeny clearly shows that the rCRS is just one particular mtDNA lineage typically differing from other lineages by dozens of changes. The obtained sequencing results can be compared, one by one, to the entire database of complete mtDNA sequences. Leaving heteroplasmic changes aside for the moment, all 20 sequences except one testify to lineages of West Eurasian (and European, in particular) ancestry.

The mtDNA lineages of patients T1, T4, T5, T9, T10, UT1, UT2, and UT5 may all belong to haplogroup H (or at least to the larger haplogroup HV). In particular, T1, T5, and T9 bear the mutation C456T characteristic of haplogroup H5. Moreover, the three changes T195C, A257G (which is a quite rare event), and 309+C found in patient UT1 are shared with a particular lineage from haplogroup H1 (GenBank accession number EF177411 [52]). Patients T2 and T3 bear mtDNA lineages belonging to a specific branch of haplogroup K1a1a that have C114T; evidently, A73G and A263G were overlooked in Patient T3. Patients UT7 and UT8 have haplogroup K1a4a1 lineages, for which A263G and 315+C were not recorded; the nucleotide information for UT7 in column "Change" of the table is mis-documented. Patients T6, T7, and T10 have haplogroup U lineages, which cannot be further specified except for excluding subhaplogroup status K, U1b, U2e, U3, U4′9, U5b, and U6, so that U5a status would be most probable. In fact, mutation T15565C has been detected sporadically in some haplogroups, including haplogroup U5a (http://freepages.genealogy.rootsweb.ancestry.com/~ncscotts/mtDNA/GenBank%20Mutation%20Lists/hg%20U/mtDNA_hg_U_Mutation_Distribution.htm). The mtDNA lineage of patient UT9 is a J1c1 lineage, whereas the UT3 lineage is a specific J1c lineage not belonging to J1c1 but instead sharing the entire motif from the region 35–464 plus the infrequent mutation G8865A with the J1c lineage of GenBank accession number EU573192 (submitted by the company 'Family Tree DNA'); the lack of G11719A likely represents an oversight. The mtDNA of patient UT6 is of non-European ancestry; this profile belongs to the East Asian haplogroup F3 as inferred from the rare combination 249del and A3434G.

Finally, the reported mtDNA variation of the patients T8 and UT4 is of mosaic nature. Namely, all mutations except for 309+C in region 35–464 match sequence no. 19 from ref. [41], which in particular encompass the rare mutation pair T279C, C285T that (in combination) is absolutely specific to haplogroup U1b. The coding-region variation, however, perfectly matches haplogroup K1a1b1 members. The most parsimonious explanation for this puzzling pattern is sample mix-up resulting in an inadvertent exchange of the amplicon for the control-region fragment. Similarly, patient T8 is assigned a mixed variation pattern, where the control-region fragment unambiguously indicates a haplogroup I lineage and the coding-region fragments point to haplogroup U (A11467G) instead of haplogroup I (with expected mutation G8251A).

The heteroplasmic variation is so immense and bizarre that it cannot reflect natural variation: most seeming heteroplasmies cluster in quite narrow stretches of the amplified fragments, especially in 3356–3410. Such patterns are indicative of artefacts termed phantom mutations. The amount of DNA and the number of PCR cycles were in fact too high, which together with a low temperature for the primers would invite all kinds of artefacts [53]. Similarly inadvertent amplification and sequencing conditions (following a forensic protocol that was widely used at the time) led to an excess of seeming heteroplasmies in hair roots [54], which was later admitted and corrected by the authors [55].

In summary, nearly all the heteroplasmic variation in Tables 1 and 2 of Carew et al. [10] is very likely induced by suboptimal amplification and sequencing protocols and thus artefactual. In contrast, the reported homoplasmic variation does not exhibit any excess or unexpected mutations: by and large, it matches the natural variation in the general population. A few mutations were actually not recorded, either by systematic employment of a wrong reference sequence or through oversight. Moreover, two samples were contaminated or confused at the amplification steps as testified by two clear-cut cases of artificial recombination.

Apparent Instabilities and Data Quality Assessment

Heteroplasmic-like patterns frequently show up in the electropherograms as a result of sequence background noise. Some of these artefacts occur within the same sequencing plate and are not reproducible when replicating the sequencing in different plates. For instance, Data S3 shows an example affecting position 220 in HVS-II in one plate that was not replicated in a second sequencing round carried out in a different plate. Some of these unspecific artefacts appear as arrays of positions (e.g. 16105-16305-16310). Many of these spurious changes did not constitute typical phantom mutations that would occur at well-known sites [29,56]. We have also observed that several positions appear to be instable (with heteroplasmic-like patterns) in forward but not in reverse sequence electropherograms or *vice versa*. With single-strand reading there is a high risk of taking these artefacts as face value of real mtDNA instability.

We also detected a case of sample mix-up due to an erroneous labelling of samples: LL79 was initially labelled as LL80 and *vice versa*. Since these two samples belong to two different haplogroups (with different sets of diagnostic motifs in the control-region), the error could be easily detected: the mtDNA profile of 79LG was C16111T-A16220C-T16362C-T16519C-A263G-309+C-315+C (belonging to haplogroup HV), while the profile of its seeming counterpart 79LL (80LL in reality) was C16069T-T16126C-G16145A-T16172C-T16231C-C16261T-A73G-C150T-T152C-T195C-A215G-A263G-C295T-310+T-315+C-T319C-T489C-G513A (haplogroup J2a1a1).

This kind of 'artefactual instability' due to sample mixing could go unnoticed in case the two samples would share very similar variation (e.g. when they belong to the same narrow sub-haplogroup). Therefore, the best way to determine whether both samples belong to the same biological source or not is to genotype a set of autosomal markers as was carried out in the present study.

Confirmed Mutation Instabilities in B-CLL Patients

We considered confirmed mtDNA instabilities those sequence patterns that could be replicated in the first round of sequencing analysis in both forward and reverse strands, and additionally in an independent laboratory by a different analyst. Moreover, granulocyte and the counterpart lymphocyte mtDNA profiles matched phylogenetically in all the cases. Identification analysis carried out by means of STR autosomal profiling (as commonly exercised in the forensic field;), further corroborated the common biological source of each pair of samples (lymphocytes/granulocytes) taken from the patients analyzed in the present study.

We have detected instability-like patterns in a total of 20 patients (Table 3). Data S1 shows the full list of haplotypes obtained for the 146 patients analyzed in the present study. Figure 1 shows the electropherogam of a single example of instability whereas Data S4 shows the electropherograms for the full set of instabilities observed. Most of the instabilities (76%) appear associated with the homopolymeric C-stretch located in HVS-II around position 310. These instabilities are well known in forensic science. Further known hotspots affected by instabilities were positions 152 and 16093, the dinucleotide variation between 514 and 523, as well as the C-homopolymeric tract around position 570.

Most haplogroups characteristic of West Eurasia were present in our sample of B-CLL. The instabilities observed did not cluster in any particular haplogroup.

Testing for statistical association between haplogroup and clinical-pathological variants and the incidence of mtDNA instabilities

The best *P*-value value for the potential association between the amount of instabilities and the haplogroup status was found to be for haplogroup H, (*P*-value = 0.0024; Chi-square). Adjustment for multiple testing either using Bonferroni or FDR test (adjusted $\alpha = 0.0022$ for both tests) indicates a lack of statistical association between the presence of instabilities with the HG status and the clinical-pathological variants (Table 2).

Discussion

Our experimental design explored the accumulation of mutations in two different cell lines that are biologically very closely related: both kinds of cell lines have their origin in single pluri-potent stem cells, so that the differences that could arise in the mtDNA of these two cells would be due mainly to the tumor condition of the lymphocyte line and not predominantly to somatic differences that arose during the differentiation process of these cell lines. We have detected a total of twenty cases carrying mtDNA instabilities, most of them affecting well-known hotspots in the mtDNA genome. This lends additional support to the hypothesis formulated in Vega et al. [16]; briefly, mtDNA instability occurs predominantly at natural hotspots and is at least in a first stage neutral to DNA function. We did not find any mtDNA instability that would follow a pathway in the mtDNA phylogeny and mimic a haplogroup motif (*contra* Linnartz et al. [57]).

Thus, an important amount of our analytical effort was devoted to the sequence quality and assessment of mtDNA instability. We only consider the existence of instability if it is confirmed in the forward, the reverse strands, in different DNA amplicons, and replicated in different laboratories. These aspects are particularly important in tumor studies since the use of relaxed criteria to assess instability can alter significantly results and conclusions. Despite the precautions for avoiding contamination, we detected an instance of sample mix-up that could have led to an erroneous interpretation of multiple instabilities (see above). Re-analysis of these samples uncovered the source of the error and fully ruled out the apparent evidence of instability.

In conclusion, instabilities observed in B-CLL patients seem to be neutral to DNA function and likely do not contribute to the tumor development.

We here advance some recommendations that would help to minimize erroneous interpretation of sequencing results in mtDNA studies in tumorigenesis:

- No instance of seeming heteroplasmy should be interpreted as real instability unless it is fully confirmed with forward and reverse sequencing.
- It is highly recommended to replicate positive results of instabilities in a second laboratory.
- Samples showing instability patterns (e.g. tumor and its counterpart 'healthy' control sample) should be genotyped for a set of autosomal markers (e.g. microsatellites) in order to preclude erroneous assignment of samples to an individual.
- Reverse and forward electropherograms for the instabilities found should be presented in manuscripts.
- Full mtDNA sequence results should be presented in the text [16] for the whole set of individuals analyzed rather than mere summary statistics that would only report the number of instabilities observed at individual positions [19].

Supporting Information

Data S1 MtDNA sequencing results for the entire control-region of the mtDNA genome in 146 pairs of samples of granulocytes (HG or LG) and lymphocytes (HL or LL) obtained from B-CLL patients.

Table 3. Summary of the mutational differences found between granulocytes (G) and their counterpart lymphocytes (L).

NUCLEOTIDE POSITION	REGION	CHANGE (with respect to rCRS)	Frequency	Cell type
16093	MT-HV1	T→C	1 (homoplasmic)	G
		T→C>T	1 (heteroplasmic)	L
		T→T>C	1 (heteroplasmic)	G
		T→T≫C	1 (heteroplasmic)	L
16224	MT-HV1	T→C	1 (homoplasmic)	G
		T→C>T	1 (heteroplasmic)	L
16235	MT-HV1	A→G	1 (homoplasmic)	L
		A→G>A	1 (heteroplasmic)	G
16270	MT-HV1	C→T≫C	1 (heteroplasmic)	G
		C→T>C	1 (heteroplasmic)	L
16302	MT-HV1	A→A = G	1 (heteroplasmic)	G
		A→G>A	1 (heteroplasmic)	L
16362	MT-HV1	T→T>C	1 (heteroplasmic)	G
		T→C>T	1 (heteroplasmic)	L
73	MT-HV2	A→G	1 (homoplasmic)	L
		A→G≫A	1 (heteroplasmic)	G
152	MT-HV2, MT-OHR	T→T≫C	1 (heteroplamic)	L
		T→T>C	2 (heteroplasmic)	G
		T→T = C	1 (heteroplasmic)	L
228	MT-HV2, MT-OHR, MT-CSB1	G→A	1 (homoplasmic)	L
		G→A≫G	1 (heteroplasmic)	G
303-309	MT-HV2, MT-OHR, MT-CSB2	7C→8C	1 (homoplasmic)	L
		7C→8C>7C	2 (heteroplasmic)	G
		7C→8C = 9C	3 (heteroplasmic)	G and L
		7C→8C = 9C>7C	1 (heteroplasmic)	G
		7C→8C>9C	4 (heteroplasmic)	G and L
		7C→8C<9C	2 (heteroplasmic)	L
		7C→9C>10C	1 (heteroplasmic)	G
		7C→9C<10C	1 (heteroplasmic)	L
514-523		$(CA)_5$→$(CA)_5$>$(CA)_4$	1 (heteroplasmic)	G
568-573		6C→7C	1 (homoplasmic)	L
		6C→8C>6C	1 (heteroplasmic)	G
		6C→9C	1 (heteroplasmic)	G
		6C→10C	1 (homoplasmic)	L
		6C→10C>9C	1 (heteroplasmic)	L

doi:10.1371/journal.pone.0007902.t003

Found at: doi:10.1371/journal.pone.0007902.s001 (0.05 MB XLS)

Data S2 LR values for the identification of the common biological source for those pairs of tumor/non-tumor samples showing mtDNA instabilities.
Found at: doi:10.1371/journal.pone.0007902.s002 (0.02 MB XLS)

Data S3 Sequence electropherograms of the mtDNA heteroplasmic-like pattern at position 220 in HVS-II showing up as seemingly heteroplasmic in one plate but not replicated in a second round of sequencing analysis carried out in a different plate.
Found at: doi:10.1371/journal.pone.0007902.s003 (0.07 MB DOC)

Data S4 Full set of sequence electropherograms showing the mtDNA instabilities detected in the present study. For each pairs of samples (indicated right below each tetrad of electropherograms together with the description of the instability observed), we indicate the forward (top pair electropherogram) and the reverse (bottom pair of electropherograms) sequences.
Found at: doi:10.1371/journal.pone.0007902.s004 (2.22 MB DOC)

Author Contributions

Conceived and designed the experiments: IMG AV AGO AC AS. Performed the experiments: MC. Analyzed the data: MC HJB AS. Contributed reagents/materials/analysis tools: IMG MA AV AGO AC AS. Wrote the paper: HJB AS.

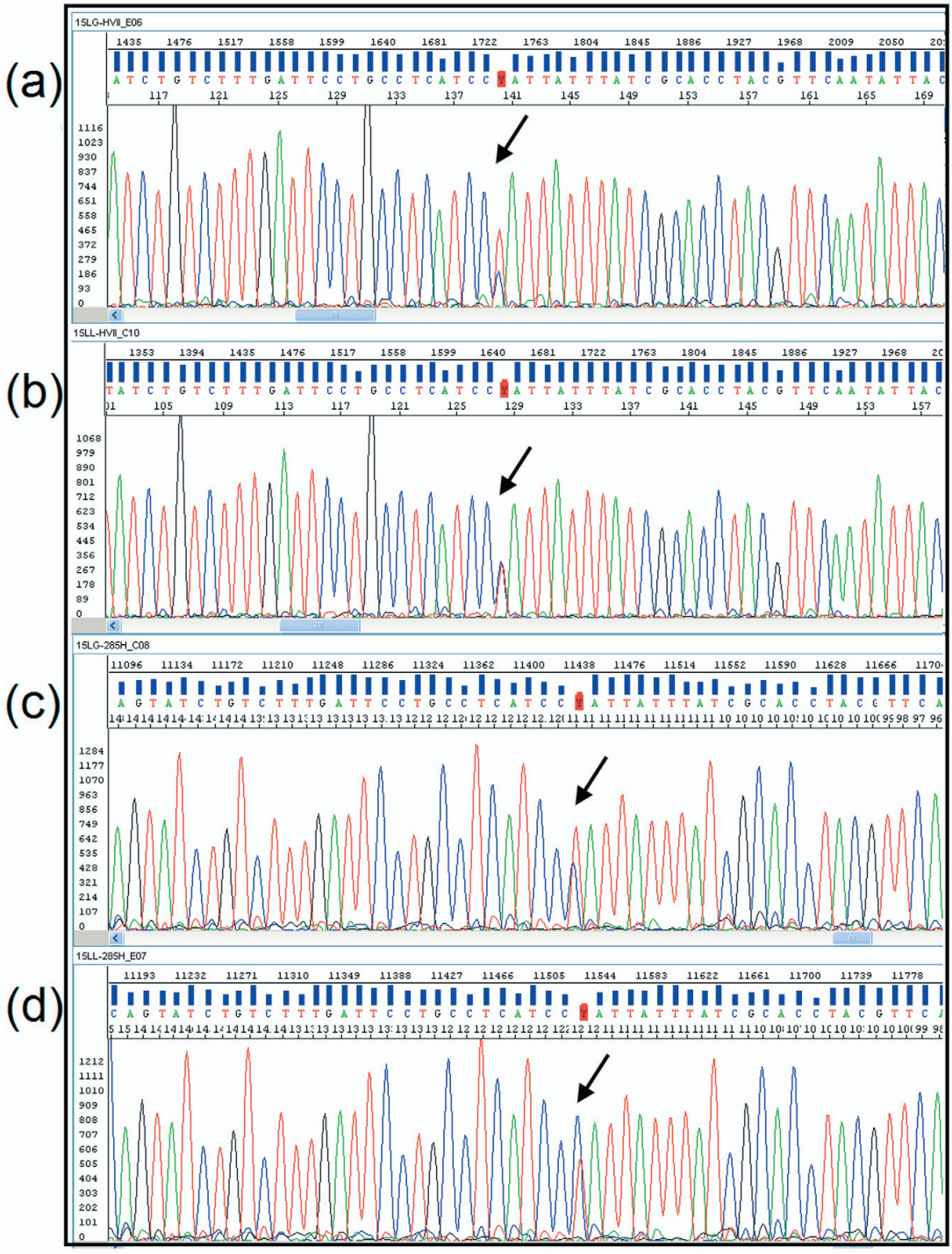


Figure 1. Example of mtDNA instability observed at position 152 in HVS-II; forward in granulocytes (a) and lymphocytes (b) and the reverse patterns in granulocytes (c) and lymphocytes (d).
doi:10.1371/journal.pone.0007902.g001

References

1. Chiorazzi N, Rai KR, Ferrarini M (2005) Chronic lymphocytic leukemia. N Engl J Med 352: 804–815.
2. Kipps TJ (1995) Chronic lymphocytic leukemia and related diseases; WJ W, editor. New York: McGraw-Hill. pp 1017–1003.
3. Shvidel L, Shtarlid M, Klepfish A, Sigler E, Berrebi A (1998) Epidemiology and ethnic aspects of B cell chronic lymphocytic leukemia in Israel. Leukemia 12: 1612–1617.
4. Rawstron AC, Green MJ, Kuzmicki A, Kennedy B, Fenton JA, et al. (2002) Monoclonal B lymphocytes with the characteristics of "indolent" chronic lymphocytic leukemia are present in 3.5% of adults with normal blood counts. Blood 100: 635–639.
5. van Besien K, Keralavarma B, Devine S, Stock W (2001) Allogeneic and autologous transplantation for chronic lymphocytic leukemia. Leukemia 15: 1317–1325.
6. Rozman C, Montserrat E (1995) Chronic lymphocytic leukemia. N Engl J Med 333: 1052–1057.
7. Anaissie EJ, Kontoyiannis DP, O'Brien S, Kantarjian H, Robertson L, et al. (1998) Infections in patients with chronic lymphocytic leukemia treated with fludarabine. Ann Intern Med 129: 559–566.
8. Cheson BD, Bennett JM, Grever M, Kay N, Keating MJ, et al. (1996) National Cancer Institute-sponsored Working Group guidelines for chronic lymphocytic leukemia: revised guidelines for diagnosis and treatment. Blood 87: 4990–4997.
9. Esteve J, Villamor N, Colomer D, Cervantes F, Campo E, et al. (2001) Stem cell transplantation for chronic lymphocytic leukemia: different outcome after autologous and allogeneic transplantation and correlation with minimal residual disease status. Leukemia 15: 445–451.
10. Carew JS, Zhou Y, Albitar M, Carew JD, Keating MJ, et al. (2003) Mitochondrial DNA mutations in primary leukemia cells after chemotherapy: clinical significance and therapeutic implications. Leukemia 17: 1437–1447.
11. Meierhofer D, Ebner S, Mayr JA, Jones ND, Kofler B, et al. (2006) Platelet transfusion can mimic somatic mtDNA mutations. Leukemia 20: 362–363.
12. He L, Luo L, Proctor SJ, Middleton PG, Blakely EL, et al. (2003) Somatic mitochondrial DNA mutations in adult-onset leukaemia. Leukemia 17: 2487–2491.
13. Grist SA, Lu XJ, Morley AA (2004) Mitochondrial mutations in acute leukaemia. Leukemia 18: 1313–1316.
14. Gattermann N (2004) Mitochondrial DNA mutations in the hematopoietic system. Leukemia 18: 18–22.
15. Yao Y-G, Ogasawara Y, Kajigaya S, Molldrem JJ, Falcao RP, et al. (2007) Mitochondrial DNA sequence variation in single cells from leukemia patients. Blood 109: 756–762.
16. Vega A, Salas A, Gamborino E, Sobrido MJ, Macaulay V, et al. (2004) mtDNA mutations in tumors of the central nervous system reflect the neutral evolution of mtDNA in populations. Oncogene 23: 1314–1320.
17. Shin MG, Kajigaya S, Levin BC, Young NS (2003) Mitochondrial DNA mutations in patients with myelodysplastic syndromes. Blood 101: 3118–3125.
18. Ivanova R, Lepage V, Loste MN, Schachter F, Wijnen E, et al. (1998) Mitochondrial DNA sequence variation in human leukemic cells. Int J Cancer 76: 495–498.
19. Salas A, Yao Y-G, Macaulay V, Vega A, Carracedo Á, et al. (2005) A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2: e296.
20. Salas A, Yao Y-G, Bandelt H-J (2006) Reply to Bora Baysal: Mitochondria: more than mitochondrial DNA in cancer. PLoS Med, 05 January.
21. Bandelt HJ, Salas A (2009) Contamination and sample mix-up can best explain some patterns of mtDNA instabilities in buccal cells and oral squamous cell carcinoma. BMC Cancer 9: 113.
22. Salas A, Lareu MV, Carracedo A (2001) Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. Int J Legal Med 114: 186–190.
23. Tully G, Barritt SM, Bender K, Brignon E, Capelli C, et al. (2004) Results of a collaborative study of the EDNAP group regarding mitochondrial DNA heteroplasmy and segregation in hair shafts. Forensic Sci Int 140: 1–11.
24. Valverde E, Cabrero C, Cao R, Rodríguez-Calvo MS, Díez A, et al. (1993) Population genetics of three VNTR polymorphisms in two different Spanish populations. Int J Legal Med 151: 251–256.
25. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373–387.
26. Bandelt H-J, Achilli A, Kong Q-P, Salas A, Lutz-Bonengel S, et al. (2005) Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem Biophys Res Commun 333: 122–130.
27. Bandelt H-J, Kong Q-P, Parson W, Salas A (2005) More evidence for non-maternal inheritance of mitochondrial DNA? J Med Genet 42: 957–960.
28. Yao Y-G, Salas A, Bravi CM, Bandelt H-J (2006) A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum Genet 119: 505–515.
29. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150–1160.
30. Bandelt H-J, Salas A, Bravi CM (2004) Problems in FBI mtDNA database. Science 305: 1402–1404.
31. Bandelt H-J, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267–273.
32. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891–899.
33. Salas A, Prieto L, Montesino M, Albarrán C, Arroyo E, et al. (2005) Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02-03 proficiency testing trial. Forensic Sci Int 148: 191–198.
34. Yao Y-G, Kong Q-P, Salas A, Bandelt H-J (2008) Pseudo-mitochondrial genome haunts disease studies. J Med Genet 45: 769–772.
35. Salas A, Bandelt H-J, Macaulay V, Richards MB (2007) Phylogeographic investigations: The role of trees in forensic genetics. Forensic Sci Int 168: 1–13.
36. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.
37. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276.
38. Bandelt H-J, Kivisild T, Parik J, Villems R, Bravi CM, et al. (2006) Lab-specific mutation processes; H.-J. Bandelt MR, V. Macaulay, editor. Berlin-Heidelberg: Springer-Verlag. pp 119–150.
39. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373–387.
40. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, et al. (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251–257.
41. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, et al. (2005) Saami and Berbers–an unexpected mitochondrial DNA link. Am J Hum Genet 76: 883–886.
42. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.
43. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences from the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152–1171.
44. Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, et al. (2008) Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS ONE 3: e2062.
45. González AM, Brehm A, Pérez JA, Maca-Meyer N, Flores C, et al. (2003) Mitochondrial DNA affinities at the Atlantic fringe of Europe. Am J Phys Anthropol 120: 391–404.
46. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, et al. (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE 4: e5112.
47. Bertranpetit J, Cavalli–Sforza L (1991) A genetic reconstruction of the history of the population of the Iberian Peninsula. Annals of Human Genetics 55: 51–67.
48. Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A (1998) mtDNA analysis of the Galician population: a genetic edge of European variation. Eur J Hum Genet 6: 365–375.
49. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57: 289–300.
50. Egeland T, Mostad PF, Mevåg B, Stenersen M (2000) Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int 110: 47–59.
51. Roostalu U, Kutuev I, Loogväli E-L, Metspalu E, Tambets K, et al. (2007) Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol 24: 436–448.
52. Pereira L, Gonçalves J, Franco-Duarte R, Silva J, Rocha T, et al. (2007) No evidence for an mtDNA role in sperm motility: data from complete sequencing of asthenozoospermic males. Mol Biol Evol 24: 868–874.
53. Brandstätter A, Parson W (2003) Mitochondrial DNA heteroplasmy or artefacts–a matter of the amplification strategy? Int J Legal Med 117: 180–184.
54. Grzybowski T (2000) Extremely high levels of human mitochondrial DNA heteroplasmy in single hair roots. Electrophoresis 21: 548–553.
55. Grzybowski T, Malyarchuk BA, Czarny J, Miscicka-Sliwka D, Kotzbach R (2003) High levels of mitochondrial DNA heteroplasmy in single hair roots: reanalysis and revision. Electrophoresis 24: 1159–1165.
56. Brandstätter A, Sänger T, Lutz-Bonengel S, Parson W, Béraud-Colomb E, et al. (2005) Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis 26: 3414–3429.
57. Linnartz B, Anglmayer R, Zanssen S (2004) Comprehensive scanning of somatic mitochondrial DNA alterations in acute leukemia developing from myelodysplastic syndromes. Cancer Res 64: 1966–1971.

V DISCUSSION

For the present dissertation mtDNA variability was analysed in more than 3000 new samples, and genotyping data including more than 40000 profiles were collected from the literature. Most of the studies carried out in this project concern population genetics of African populations but also European and Native-American ones. In addition, several studies were also carried out in the context of forensic and medical genetics.

Nowadays sequencing complete genome is the best way for the description of new lineages and their correct allocation within the mtDNA phylogeny. A growing number of articles describing new lineages have been published in the last years based on complete sequencing projects. As part of the laboratory effort carried out in the present PhD project; a protocol to sequence entire genomes was developed. This has allowed us to reduced costs and narrow the time-frame for entire genome sequencing. On the other hand, a consensus nomenclature for haplogroup designation is mandatory in order to prevent problems of different nature. The scientific community is now considering Phylotree (www.phylotree.org) as the reference worldwide mtDNA tree, both, phylogenetically and from the point of view of haplogroup nomenclature. Here we have also contributed to improve some branches of the phylogeny and followed Phylotree as the nomenclature standard.

V.1 Human mtDNA variability in Europe

Patterns of mtDNA variability in West Eurasia (mainly in Europe) have been analyzed in detail in the literature. Several haplogroups have been described in Europe, including H, HV, I, J, K, U, T, V, W and X; haplogroup H accounts for ~40% of all mtDNAs in this continent. Moreover, most of the R0 sub-lineages (where H, V, and HV are nested) have not good diagnostic at the control region. The genotyping of selected coding region SNPs allowed us to discriminate between R0 samples bearing the same control region. Thus, a set of 71 coding region SNPs were selected in order to discriminate between 71 different R0 profiles and three multiplexes were designed accordingly. In our studied samples, 40 different R0 sub-lineages were present; from a total of 237 samples belonging to R0 sub-lineages, 163 carried a unique haplotype when HVI plus coding region SNPs were analyzed together in contrast to the 124 different haplotypes that could be resolved with the HVI alone. The SNPs provide also a solid background for haplogroup assignation in contrast to the weak signal provided by the control region alone.

Several branches of the phylogeny have been improved as a consequence of the work carried out in the present study. Thus, for instance, the phylogeny of haplogroup H2 has now a new West-Eurasian sub-lineage, named in our study as H2a5 and that seems to be autochthonous from the Basque Country. Eight samples that carried a transition with respect to rCRS at position 4592 were whole genome sequenced. All of them shared the sequence motif 1842, 4592 13708 and 16291. Five samples bearded exactly the same profile sharing the basal motif while the other three samples carried three new variants. Nowadays, following the latest Phylotree version (Build 11), this lineage has been renamed as H2a5a. The differences between the phylogeny described in (Alvarez-Iglesias et al. 2009) and the phylogeny described according to the data present in Phylotree Build 11 are represented in Figure 42.

Complete genome sequencing has also allowed to discover several new branches of the African phylogeny, which are present in Europe and America (studies in preparation).

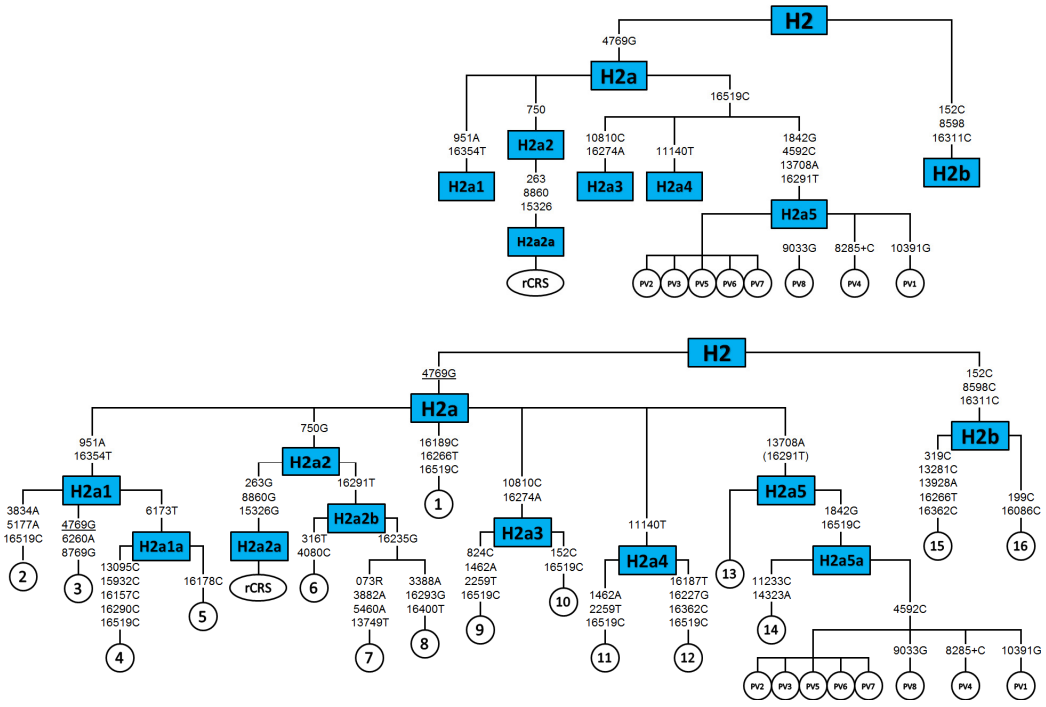


Figure 42 Differences in H2a phylogeny. At the top, it is the phylogeny described in (Alvarez-Iglesias et al. 2009). At the bottom, it is the phylogeny according Phylotree Build 11 (7 February 2011).

V.2 Human mtDNA variability in America

America was the last continent settled by humans. The modern Native American populations belong to only five main mtDNA haplogroups (although minor lineages are being discovered in the last few years). However, since colonial times (e.g. with the arrival of Europeans and slaves from Africa) to present the mtDNA landscape of America has changed dramatically, such that today, with the exception of only few regions, most of the American variability is non-Native.

Colombia provides a paradigmatic case of a melting-pot of inter-continental admixture. Thus, nowadays, the Colombian population is the result of several processes of admixture occurred between Europeans, Africans, and Native Americans to different degrees, depending on the Colombian region and social strata considered. Although the census data does not include ethnicity, more than 50 different indigenous ethnic groups have been described in the country to date. We have investigated the mtDNA ancestry of admixed groups whose individuals were collected according to their ‘‘self-reported ethnicity’’ in order to demonstrate that the ancestry descriptors are not supported by genetic variation (at least in this populations).

A total of 183 Colombian individuals with self-reported ethnicity including *“Mestizos”, “Mulatos”* and *“Afro-Colombians”* and Native Americans were genotyped for the HVI and a subset of them were additionally genotyped for the HVII. As a whole, there are 99 Native American mtDNA haplotypes and 56 mtDNA of recent Sub-Saharan ancestry. Our results showed that *“Mestizos”* have a larger Native American component (97%) whereas *“Mulatos”* and *“Afro-Colombians”* have a main sub-Saharan African ancestry (81% and 73% respectively). These results differ significantly from other previous studies that targeted the same population groups. We concluded that self-reported ethnicity is a misleading indicator of the genetic background of the individual (at least for the mtDNA). The Native American component of our Colombian population is more closely related to Central and northern South America. The sub-Saharan component seems to be more closely related to West-Central, Southeast and South West Africa. Both results are consistent with previous results and historical data.

The amount of mtDNA information currently available on Central America is very limited compared to other American regions. The study of the mtDNA variation in a population sample from El Salvador contributed to improve our knowledge on the

colonization of this region and its role in the peopling of the New World. The control region was analyzed in 90 individuals from the general population in El Salvador. Contrarily to what we observed in the Colombian population, less than 5% of the mtDNA lineages belonged to African and European mtDNA haplogroups, the rest could be attributed to lineages of Native American origin (most to haplogroup A2). We carried out an admixture analysis in order to determine which one of the source populations (North and South America) contributed more to the extant population in El Salvador. Our results indicated that North America accounts for the 76-92% (depending on the method) of the lineages present in El Salvador, while South America contribute the remaining 8-24%.

About 91% of the Salvador mtDNAs belonged to haplogroup A2, whereas haplogroups B2 and C1 accounted for 2% each. A founder analysis was carried out assuming a single migration into El Salvador, and North America considered as the unique source population. Seven founders were observed in our Salvadorian population, all of them present in North American populations. Other potentially founders were discarded because they were found also in Central America but not in the North (as result of recent migration between of neighbor populations) or were detected in North America but related with samples assigned to self-reported “Hispanic” with not information about location or other ethnic information. The coalescence age of haplogroup A2 in El Salvador was estimated in about 12.600±4900 y.b.p. The estimated age of A2 in Salvadorians is relatively close to the value calculated for A2 for all populations (14.600 ybp), suggesting that the peopling of the region (and Central America in general) could have occurred soon since the initial colonization of the Americas.

Colombia and El Salvador represent therefore two totally different American population models, although they are relatively close geographically. Both have suffered from different pre and post-colonial processes that have altered in different directions their patterns of genetic variation.

V.3 Human mtDNA variability in Africa

An important effort of the present PhD project has been devoted to the analysis of African mtDNA variability. One of these projects focuses in one of the most enigmatic and interesting populations of the Sahara, the Tuareg. The Tuareg are nomads that nowadays live along the Saharan desert; their ancestral origin is enigmatic and it cannot be attributed to only a single African location (as it generally occurs with most of the human

population groups). The aim of our study was to dig into the mtDNA genetic origin of these nomads and a deep analysis of their genetic relationships with other neighbours.

We carried out the genotyping of the control region in 90 samples collected from three different groups from Burkina Faso, Niger and Mali, all of them self-reported as Tuaregs. Individuals from Burkina and Niger live within the bend of Niger, whereas individuals from the Republic of Niger lived far away. We also carried out analyses of the 71 SNPs in 35 samples belonged to European lineages in order to improve haplogroup classification in this part of the phylogeny. Moreover, we also carried out the genotyping of complete genome of three individuals belonging to a new M1a2 sub-lineage.

A total of 48% of the mtDNA haplotypes observed in this study belonged to Sub-Saharan haplogroups; whereas 39% were members of West Eurasian lineages and 13% fit within the East African haplogroup M1. The three Tuareg populations vary in their haplogroup frequency make-up. Sample from the Republic of Niger showed the highest level of sub-Saharan haplogroups (81%), whereas in the other two populations L-haplogroups accounted for about 27% of the component (26% in case of the Tuareg from Burkina and 28% in case of those from Mali). For the West Eurasian component, both samples located near the bend of Niger showed similar frequencies (around 52%) while in the population from the Republic of Niger this component was lower (16%). Similar results were obtained for the East African haplogroup M1, with the sample from the Republic of Niger showing the lowest frequency (3%) whereas the other two populations showing frequencies about 18%. These differences in the haplogroup distribution are statistically significant when comparing the Tuareg from the Republic of Niger to the other two Tuareg populations. The West Eurasian component of the three Tuareg populations mostly belongs to H1, H3 and V haplogroups; with a most likely geographical origin in Iberia. These haplogroups were most likely originated in the Franco-Cantabrian refugee, and spread to Europe right after the LMG with the retreat northward of ice sheets. Migrations southwards of the Iberian Peninsula and later towards the North of Africa could have driven these lineages to the Tuareg more than 10000 years ago. Other interesting finding of the study was the absence of haplogroups J and T in the Tuareg. These haplogroups spread into Europe during the Neolithic expansion and they have been detected in other North Africa populations. The Sub-Saharan component of the Tuaregs is basically haplogroup L2 (58.1%), followed by haplogroups L3 (23.3%) and L1 (14%). When we tried to determine the most likely African origin of the Sub-Saharan component, West

and West-Central African lineages appears to be the most predominant in the extant three Tuareg populations. The East African haplogroup M1 was also observed in the Tuareg. It was deeply investigated by way of complete genome sequencing; all of them resulted to belong to a new sub-lineage baptized here as M1a2a. The three complete genomes were identical, therefore suggesting recent genetic drift. Although sample sizes were relatively small for the establishment of a consistent hypothesis on the demography of the Tuareg, it is interesting to note that each one of the three populations showed a different sub-branch of M1. Thus, in the population from Burkina Faso there were seven individuals belonging to M1a2 lineages, while there were four individuals belonging to M1b1 in Tuaregs from Mali, and just one sample belonging to M1b2 in the Republic of Niger. Also interesting is the lack of complete absence of U6 in the Tuareg, in opposition to haplogroup M1 (both lineages have been hypothesized as being the result of back migration into Africa from western Asia).

Other studies of the present project have been focused in the African variation. Most of previous efforts in the literature have been devoted to the coding region. We have developed a method that allows a quick assignation of a sample to a specific sub-branch of the African mtDNA phylogeny, based on MALDI-TOF. The technique allows to genotype 230 SNPs that were meticulously selected in order to represent all major and minor branches of the African mtDNA phylogeny. An initial methodological study was carried out in order to show the principles of the technique as well as its accuracy for mtDNA SNP genotyping. In this study, 542 samples belonging to 12 ethnic populations from the Lake Chad Basin (*Borgor Fulani, Buduma, Chad Arabs, Fali, Hide, Kanembu, Kanuri, Kotoko, Mafa, Masa, Shuwa Arabs* and *Tcheboua Fulani*) were genotyped; most of these samples appeared in other studies, but 94 of them were not previously analyzed for the control region. All the samples were genotyped for the full set of 230 mtSNPs using MALDI-TOF MS. For most of the populations, haplotype diversity was higher for HVI than for mtSNPs profiles whereas nucleotide diversity was near twice as large for mtSNPs as for HVI. This result most likely mirrors the fact that haplotype diversity is enriched in HVI for the presence of rare variants, whereas the agglomeration of identical sequences into different haplogroups enriches the nucleotide diversity. For the AMOVA, most of the genetic variation (~96%) was found to occur within populations. The Principal Component Analysis (PCA) shows that the three first components accounts for ~40% of the variation and reveals important divergence between the different ethnic groups of this region. The first principal component (PC1), which accounts for 15% of the variation, locates Mafa

and Kanuri in one side of the plot, and the Fulani populations in the opposite side. PC2 (13%) shows again Mafa in one pole and Kanembu in the other extreme, whereas PC3 (12%) shows Mafa in one side of the plot and Kotoko in the opposite one. There are not unique features in Mafa that makes them different from the other populations, but the separation could be due to an accumulative effect of differences in haplogroup frequencies.

There are other two articles that were elaborated as part of the present project but are still under preparation. Both are two ambitious projects focusing on the analysis of African variation at a pan-African scale and the implications on the Trans-Atlantic slave trade in America and in Europe. A first study consists of a meta-analysis that considers different sources of data (i) high throughput data generated in this project for more than 2000 African and African-American samples, (ii) control region data collected from the literature (>15.000 profiles from Africa and African-Americans), and (iii) complete genome sequences collected from the literature and public resources such as GenBank.

The first study aims to reveal new aspects concerning the origin of humans, as the cradle of all human beings is still a question of intense debate, where archaeological findings pointing to North Africa instead of East or South Africa. Demographic aspects of the African continent are also debated in this article. The data reveals that Western Africa contributed more than any other region to the Atlantic slave trade, and we discuss the role of several sub-regions within Western Africa.

A second study focuses on the analysis of sub-Saharan variation in Europe given that near 1-2% of Sub-Saharan mtDNA lineages present into European population belong to African L-haplogroups. The genotyping of complete mitochondrial genomes of several individuals from Europe revealed new branches of the African phylogeny that were still not found in the African continent. Some of them showed European specific divergence pointing to a local origin outside Africa long time ago. As a whole, the results indicated that there have been almost continuous contacts between Africa and Europe since at least 15000 years ago, although most of the sub-Saharan lineages entered into Europe during the last two millenniums during the Romanization period and the Islamic conquest of the Iberian Peninsula, and more recently, during the slave trade period.

V.4 Forensic genetics

In forensic casework, the mtDNA test only allows the identification of lineages but not the identification of single persons (as it is usually achieve using nuclear DNA markers); however, the mtDNA test can be still useful for excluding suspects or including specific persons under certain pre-formulated hypothesis. On the other hand, quality controls are important in all aspects of mtDNA research, but can be particularly relevant in forensic casework. The present PhD project has carried out some effort under this perspective. Thus, all the samples were analyzed in order to resolved any ambiguity using different methodological approaches (e.g. SNP minisequencing, standard sequencing, MALDITOF MS), or using a posteriory approaches based on the present knowledge of the mtDNA global phylogeny. We have also contributed to the analysis of the discrimination power of the mtDNA test. For instance, effort has been done in order to test the ability of minisequencing to resolved one of the most ambiguous and frequent branches of the European phylogenetic branches, namely R0, and its performance for the analysis of complex and degraded samples or carrying limited amount of DNA.

In a forensic context, SNaPshot minisequencing performs very well with complex samples and can be used for both, screening large amount of samples or as a discriminative tool. Moreover, all the population studies carried out in the present PhD project contributed to improve the databases available for forensic use. However, given that most of the datasets available are control region biased, we consider that this segment has to be generated in order to allow estimation of haplotype frequencies. We foresee future situations where entire mtDNA genomes would be carried out in forensic casework, however, novel methodological approaches would be needed in order to treat complex forensic samples.

V.5 Medical studies

We have carried out the analysis of mtDNA instability in leukaemia patients comparing tumor mtDNA profiles with the profiles obtained from normal counterpart samples extracted from the same individual. These analyses were carried out for the first time under the standards usually required for the analysis of forensic DNA samples.

Contrary to what other authors have reported in the literature, we have observed that the mtDNA is very mutationally stable in leukaemia, at least as stable as observed in

nature (healthy individuals). Therefore, it is difficult to reconcile these results with an active role of the mtDNA in the tumorigenesis process. We conclude that most of the studies in the literature dealing with mtDNA instability in tumors are critically affected by sequencing errors and therefore 'false positives' of instabilities, confirming the theory of previous studies based on theoretical grounds.

The results of the different studies carried out in the present project all indicate that standards for mtDNA analysis are necessary in order to avoid artifacts of different nature (sampling mix-up, contamination, documentation errors, etc). This is equally important in all fields of research but the consequences can be critical in forensic practice, erroneously driven the evidence in favour of the defence or the accusation.

VI SUMMARY AND CONCLUSIONS

VI.1 SUMMARY

The results of the present project indicate that the analysis of the mtDNA variation can be useful in medical, forensic, and population genetic studies. The particular features of the mtDNA, including high copy number, lack of recombination, and high average mutation rate; also determine its usefulness and limitations in genetic studies. For instance, the reconstruction of the phylogeny is straightforward because the lineages are passed through the matriline with the only changes generated by mutation. However, this is a single marker and only tells the history of female population, which not necessarily match the demography of the whole population.

We have applied these principles to the analysis of several human populations, to the forensic field, and to some medical study. All of them have many aspects in common, indicating also the important interplay that should be always mandatory in all mtDNA studies. For instance, one cannot carry out a forensic or medical genetic study ignoring population variation patterns or the important heterogeneity that exists regarding site specific mutation rates.

We have contributed to improve our knowledge of the variation in several African, European, and American populations. In this project we have also focussed our attention in several aspects of forensic interest, concerning the analysis of degraded and low DNA amount samples. And finally, we have tried to establish a necessary bridge between the different fields of research, indicating that proper quality standards can help to avoid false positives of instabilities in cancer studies, erroneous conclusions in forensic casework, or errors in datasets that could have consequences in population studies or indirectly in forensic or medical genetic ones.

VI.2 CONCLUSIONS

VI.2.1 Population genetics

1. Some European sub-lineages are poorly defined in the control region. The genotyping of 71 diagnostic coding region SNPs with SNaPshot minisequencing allows reaching a higher resolution within the most prevalent European macro-haplogroup, namely, R0.

2. The genotyping of the complete genome of eight individuals from the Basque Country carrying the characteristic transition at position 4592 allows describing a new recent autochthonous Basque clade called H2a5 (nowadays renamed in Phylotree Build 11 as H2a5a), with an estimated coalescence of 1,285±161 years.

3. Analysis of an admixed population from Colombia lead us to conclude that ''self-reported ethnicity'' is not supported by genetic variation. Colombian populations are the result of a complex and heterogeneous admixture of African (from the Slave trade), Native American and European (from colonization process) mtDNA lineages. 'Mulatos' and 'Afro-Colombians' have a dominant African mtDNA component whereas 'Mestizos' carry predominantly Native American haplotypes in Colombian population.

4. In El Salvador most of the population belongs to indigenous Native American mtDNA component. A2 is the main haplogroup in El Salvador with a coalescence age close to A2 in America suggesting that the settlement of El Salvador was close to initial expansion into the Americas

5. mtDNA variability present in *Tuaregs* is different from other northern African populations: 48% lineages are of sub-Saharan ancestry, 39% of West Eurasia ancestry and 13% from East Africa (haplogroup M1). There are higher levels of West Eurasian contribution than in other North African populations but there are no *Tuareg* mtDNA lineages connected with the Neolithic expansion from the Near East. The presence of European lineages allowed establishing a link between the Tuareg peoples and the Franco-Cantabrian refugee expansions occurring after the LGM.

6. MALDITOF MS is an appropriate methodology that allows to genotype large number of SNPs in large number of samples, reducing costs and time in regards to other more standard techniques. We have demonstrated that the technique is reliable by way of analyzing two test samples from the Chad and Mozambique.

7. Different ethnic groups from the Chad were genotyping using MALDITOF-MS, revealing new features of the demography of these populations, indicating for instance a clear difference in the mtDNA background of nomads and agriculturalist populations from the Chad.

8. A meta-analysis of African and 'African-American' samples has allowed to reveal new aspects of the African demography, the cradle of the human beings, and the impact of the Trans-Atlantic slave trade. In agreement with recent archaeological findings, there is growing evidence that East Africa or the South are not necessarily the best places to locate the origin of humans, and perhaps the western or the North-West of Africa could now be considered as good candidates for future studies.

9. Complete genome sequencing efforts have allowed for the first time to established new ancestral links between Europe and Africa. We have contributed new clades to the African phylogeny that have most likely evolved within the European continent.

VI.2.2 Forensic genetics

1. The development of new genotyping techniques has allowed improving the discrimination power of the mtDNA test in forensic casework. The mini-sequencing approach *via* SNaPshot allows the characterization of complex forensic samples containing low amount and/or degraded DNA where other standard techniques (such as Sanger standard sequencing) fail.

2. The quality control exercises organized by the GEP-ISFG have helped to provide new insights into interpretation, and error monitorization of the mtDNA test. The mtDNA has also demonstrated to be useful to unravel the presence of multiple contributors to a mixture sample.

VI.2.3 Clinical genetics

1. We have not found evidences supporting molecular mtDNA instability as responsible for B-LLC. However, a secondary role cannot be rejected considering the hypothesis that the mtDNA instability accumulation could contribute to the tumoral process.

2. Quality standards are important in medical studies in order to avoid false positives claims of mtDNA instability and therefore erroneous conclusions about the role of the mtDNA in tumorigenesis.

VII REFERENCES

(2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12: 339-48

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, Chen YT, Chu J, Cutiongco-de la Paz EM, De Ungria MC, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho SF, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim HL, Kim K, Kim S, Kim WY, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee JY, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu JY, Xiao H, Xu S, Yang JO, Shugart YY, Yoo HS, Yuan W, Zhao G, Zilfalil BA (2009) Mapping human genetic diversity in Asia. Science 326: 1541-5

Abu-Amero KK, Gonzalez AM, Larruga JM, Bosley TM, Cabrera VM (2007) Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC Evol Biol 7: 32

Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, Woodward SR, Salas A, Torroni A, Bandelt HJ (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3: e1764

Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, Magri C, Scozzari R, Babudri N, Santachiara-Benerecetti AS, Bandelt HJ, Semino O, Torroni A (2005) Saami and Berbers--an unexpected mitochondrial DNA link. Am J Hum Genet 76: 883-6

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogvali EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910-8

Akouchekian M, Houshmand M, Hemati S, Ansaripour M, Shafa M (2009) High rate of mutation in mitochondrial DNA displacement loop region in human colorectal cancer. Dis Colon Rectum 52: 526-30

Alonso A, Albarran C, Martin P, Garcia P, Capilla J, Garcia O, de la Rua C, Izaguirre N, Pereira F, Pereira L, Amorim A, Sancho M (2006) Usefulness of microchip electrophoresis for the analysis of mitochondrial DNA in forensic and ancient DNA studies. Electrophoresis 27: 5101-9

Alonso A, Salas A, Albarran C, Arroyo E, Castro A, Crespillo M, di Lonardo AM, Lareu MV, Cubria CL, Soto ML, Lorente JA, Semper MM, Palacio A, Paredes M, Pereira L, Lezaun AP, Brito JP, Sala A, Vide MC, Whittle M, Yunis JJ, Gomez J (2002) Results of the 1999-2000 collaborative exercise and proficiency testing program on mitochondrial DNA of the GEP-ISFG: an inter-laboratory study of the observed variability in the heteroplasmy level of hair from the same donor. Forensic Sci Int 125: 1-7

Alvarez-Iglesias V, Barros F, Carracedo A, Salas A (2008) Minisequencing mitochondrial DNA pathogenic mutations. BMC Med Genet 9: 26

Alvarez-Iglesias V, Jaime JC, Carracedo A, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44-55

Alvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintans B, Zarrabeitia MT, Cusco I, Lareu MV, Garcia O, Perez-Jurado L, Carracedo A, Salas A (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS One 4: e5112

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457-65

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147

Arnold RS, Sun CQ, Richards JC, Grigoriev G, Coleman IM, Nelson PS, Hsieh CL, Lee JK, Xu Z, Rogatko A, Osunkoya AO, Zayzafoon M, Chung L, Petros JA (2009) Mitochondrial DNA mutation stimulates prostate cancer growth in bone stromal environment. Prostate 69: 1-11

Asari M, Tan Y, Watanabe S, Shimizu K, Shiono H (2007) Effect of length variations at nucleotide positions 303-315 in human mitochondrial DNA on transcription termination. Biochem Biophys Res Commun 361: 641-4

Attardi G, Schatz G (1988) Biogenesis of mitochondria. Annu Rev Cell Biol 4: 289-333

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann Rev Ecol Syst 18: 489-522

Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286: 2524-5

Bai RK, Wong LJ (2005) Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease. J Mol Diagn 7: 613-22

Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previdere C, Wilson IJ, Tyler-Smith C, Jobling MA (2010) A predominantly neolithic origin for European paternal lineages. PLoS Biol 8: e1000285

Balter M (2009) Archaeology. Ancient DNA says Europe's first farmers came from afar. Science 325: 1189

Balter M (2010) Archaeology. Of two minds about Toba's impact. Science 327: 1187-8

Ballinger SW, Schurr TG, Torroni A, Gan YY, Hodge JA, Hassan K, Chen KH, Wallace DC (1992) Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. Genetics 130: 139-52

Bandelt HJ, Alves-Silva J, Guimaraes PE, Santos MS, Brehm A, Pereira L, Coppa A, Larruga JM, Rengo C, Scozzari R, Torroni A, Prata MJ, Amorim A, Prado VF, Pena SD (2001a) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Annals Of Human Genetics 65: 549-63

Bandelt HJ, Forster P (1997) The myth of bumpy hunter-gatherer mismatch distributions. Am J Hum Genet 61: 980-3

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141: 743-53

Bandelt HJ, Kivisild T (2006) Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. Ann Hum Genet 70: 314-26

Bandelt HJ, Kong QP, Parson W, Salas A (2005) More evidence for non-maternal inheritance of mitochondrial DNA? J Med Genet 42: 957-60

Bandelt HJ, Lahermo P, Richards M, Macaulay V (2001b) Detecting errors in mtDNA data by phylogenetic analysis. Int J Legal Med 115: 64-9

Bandelt HJ, Macaulay V, Richards M (2006) Human Mitochondrial DNA and the Evolution of *Homo sapiens*. Springer, Heidelberg

Bandelt HJ, Parson W (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. Int J Legal Med 122: 11-21

Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150-60

Bandelt HJ, Salas A, Bravi C (2004a) Problems in FBI mtDNA database. Science 305: 1402-4

Bandelt HJ, Salas A, Lutz-Bonengel S (2004b) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267-73

Bandelt HJ, Salas A, Taylor RW, Yao YG (2009) Exaggerated status of "novel" and "pathogenic" mtDNA sequence variants due to inadequate database searches. Hum Mutat 30: 191-6

Bar W, Brinkmann B, Budowle B, Carracedo A, Gill P, Holland M, Lincoln PJ, Mayr W, Morling N, Olaisen B, Schneider PM, Tully G, Wilson M (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Legal Med 113: 193-6

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13: 2277-90

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F (2007) Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Molecular Phylogenetics And Evolution 43: 635-44

Batini C, Lopes J, Behar DM, Calafell F, Jorde LB, van der Veen L, Quintana-Murci L, Spedini G, Destro-Bisol G, Comas D (2010) Insights into the demographic history of African Pygmies from complete mitochondrial genomes. Mol Biol Evol

Behar DM, Metspalu E, Kivisild T, Rosset S, Tzur S, Hadid Y, Yudkovsky G, Rosengarten D, Pereira L, Amorim A, Kutuev I, Gurwitz D, Bonne-Tamir B, Villems R, Skorecki K (2008a) Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS ONE 3: e2062

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S (2008b) The dawn of human matrilineal diversity. Am J Hum Genet 82: 1130-40

Berger C, Parson W (2009) Mini-midi-mito: adapting the amplification and sequencing strategy of mtDNA to the degradation state of crime scene samples. Forensic Sci Int Genet 3: 149-53

Bergstrom CT, Pritchard J (1998) Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. Genetics 149: 2135-46

Bianchi NO (2010) Mitochondrial genome instability in cancer. Cytogenet Genome Res 128: 66-76

Bianchi NO, Bianchi MS, Richard SM (2001) Mitochondrial genome instability in human cancers. Mutat Res 488: 9-23

Bogenhagen D, Clayton DA (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. J Biol Chem 249: 7991-5

Bogenhagen D, Clayton DA (1977) Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle. Cell 11: 719-27

Bogenhagen DF, Applegate EF, Yoza BK (1984) Identification of a promoter for transcription of the heavy strand of human mtDNA: in vitro transcription and deletion mutagenesis. Cell 36: 1105-13

Bogenhagen DF, Clayton DA (2003a) Concluding remarks: The mitochondrial DNA replication bubble has not burst. Trends Biochem Sci 28: 404-5

Bogenhagen DF, Clayton DA (2003b) The mitochondrial DNA replication bubble has not burst. Trends Biochem Sci 28: 357-60

Brandstatter A, Niederstatter H, Parson W (2004) Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. Int J Legal Med 118: 47-54

Brandstatter A, Salas A, Niederstatter H, Gassner C, Carracedo A, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27: 2541-50

Brandstatter A, Sanger T, Lutz-Bonengel S, Parson W, Beraud-Colomb E, Wen B, Kong QP, Bravi CM, Bandelt HJ (2005) Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis 26: 3414-29

Brandstatter A, Zimmermann B, Wagner J, Gobel T, Rock AW, Salas A, Carracedo A, Parson W (2008) Timing and deciphering mitochondrial DNA macro-haplogroup R0 variability in Central Europe and Middle East. BMC Evol Biol 8: 191

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Paabo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325: 318-21

Bromham L, Eyre-Walker A, Smith NH, Smith JM (2003a) Mitochondrial Steve: paternal inheritance of mitochondria in humans. TRENDS in Ecology and Evolution 18: 3

Bromham L, Eyre-Walker A, Smith NH, Smith JM (2003b) Mitochondrial Steve: paternal inheritance of mitochondrial in humans. Trends in Ecology and Evolution 18: 3

Brown DT, Herbert M, Lamb VK, Chinnery PF, Taylor RW, Lightowlers RN, Craven L, Cree L, Gardner JL, Turnbull DM (2006) Transmission of mitochondrial DNA disorders: possibilities for the future. Lancet 368: 87-9

Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc Natl Acad Sci U S A 77: 3605-9

Brown WM, George M, Jr., Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci U S A 76: 1967-71

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18: 225-39

Brown WM, Shine J, Goodman HM (1978) Human mitochondrial DNA: analysis of 7S DNA from the origin of replication. Proc Natl Acad Sci U S A 75: 735-9

Brucato N, Cassar O, Tonasso L, Tortevoye P, Migot-Nabias F, Plancoulaine S, Guitard E, Larrouy G, Gessain A, Dugoujon JM (2010) The imprint of the Slave Trade in an African American population: mitochondrial DNA, Y chromosome and HTLV-1 analysis in the Noir Marron of French Guiana. BMC Evol Biol 10: 314

Budowle B, Fisher CL, Polanskey D, Den Hartog BK, Kepler RB, Elling JW (2008) Stabilizing mtDNA sequence nomenclature with an operationally efficient approach. Forensic Science International: Genetics Supplement Series 1: 3

Burgart LJ, Zheng J, Shu Q, Strickler JG, Shibata D (1995) Somatic mitochondrial mutation in gastric cancer. Am J Pathol 147: 1105-11

Campbell MC, Tishkoff SA (2010) The evolution of human genetic and phenotypic variation in Africa. Curr Biol 20: R166-73

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31-6

Capelli C, Tschentscher F, Pascali VL (2003) "Ancient" protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis. Forensic Sci Int 131: 59-64

Carracedo A, Bar W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M (2000) DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 110: 79-85

Castri L, Tofanelli S, Garagnani P, Bini C, Fosella X, Pelotti S, Paoli G, Pettener D, Luiselli D (2009) mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. Am J Phys Anthropol 140: 302-11

Clayton DA (1984) Transcription of the mammalian mitochondrial genome. Annu Rev Biochem 53: 573-94

Corte-Real HB, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60: 331-50

Craven L, Tuppen HA, Greggains GD, Harbottle SJ, Murphy JL, Cree LM, Murdoch AP, Chinnery PF, Taylor RW, Lightowlers RN, Herbert M, Turnbull DM (2010) Pronuclear transfer in human embryos to prevent transmission of mitochondrial DNA disease. Nature 465: 82-5

Cree LM, Samuels DC, de Sousa Lopes SC, Rajasimha HK, Wonnapinij P, Mann JR, Dahl HH, Chinnery PF (2008) A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nat Genet 40: 249-54

Crespillo M, Paredes MR, Prieto L, Montesino M, Salas A, Albarran C, Alvarez-Iglesias V, Amorin A, Berniell-Lee G, Brehm A, Carril JC, Corach D, Cuevas N, Di Lonardo AM, Doutremepuich C, Espinheira RM, Espinoza M, Gomez F, Gonzalez A, Hernandez A, Hidalgo M, Jimenez M, Leite FP, Lopez AM, Lopez-Soto M, Lorente JA, Pagano S, Palacio AM, Pestano JJ, Pinheiro MF, Raimondi E, Ramon MM, Tovar F, Vidal-Rioja L, Vide MC, Whittle MR, Yunis JJ, Garcia-Hirschfel J (2006) Results of the 2003-2004 GEP-ISFG collaborative study on mitochondrial DNA: focus on the mtDNA profile of a mixed semen-saliva stain. Forensic Sci Int 160: 157-67

Crews S, Ojala D, Posakony J, Nishiguchi J, Attardi G (1979) Nucleotide sequence of a region of human mitochondrial DNA containing the precisely identified origin of replication. Nature 277: 192-8

Croteau DL, Stierum RH, Bohr VA (1999) Mitochondrial DNA repair pathways. Mutat Res 434: 137-48

Chandrasekar A, Kumar S, Sreenath J, Sarkar BN, Urade BP, Mallick S, Bandopadhyay SS, Barua P, Barik SS, Basu D, Kiran U, Gangopadhyay P, Sahani R, Prasad BV, Gangopadhyay S, Lakshmi GR, Ravuri RR, Padmaja K, Venugopal PN, Sharma MB, Rao VR (2009) Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. PLoS ONE 4: e7447

Chang DD, Clayton DA (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. Cell 36: 635-43

Chang DD, Clayton DA (1985) Priming of human mitochondrial DNA replication occurs at the light-strand promoter. Proc Natl Acad Sci U S A 82: 351-5

Chang DD, Clayton DA (1987a) A mammalian mitochondrial RNA processing activity contains nucleus-encoded RNA. Science 235: 1178-84

Chang DD, Clayton DA (1987b) A novel endoribonuclease cleaves at a priming site of mouse mitochondrial DNA replication. EMBO J 6: 409-17

Chaubey G, Metspalu M, Kivisild T, Villems R (2007) Peopling of South Asia: investigating the caste-tribe continuum in India. Bioessays 29: 91-100

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57: 133-49

Cheng B, Tang W, He L, Dong Y, Lu J, Lei Y, Yu H, Zhang J, Xiao C (2008) Genetic imprint of the Mongol: signal from phylogeographic analysis of mitochondrial DNA. J Hum Genet 53: 905-13

Chinault AC, Shaw CA, Brundage EK, Tang LY, Wong LJ (2009) Application of dual-genome oligonucleotide array-based comparative genomic hybridization to the molecular diagnosis of mitochondrial DNA deletion and depletion syndromes. Genet Med 11: 518-26

Chinnery PF, Taylor GA, Howell N, Brown DT, Parsons TJ, Turnbull DM (2001) Point mutations of the mtDNA control region in normal and neurodegenerative human brains. Am J Hum Genet 68: 529-32

Dawid IB, Blackler AW (1972) Maternal and cytoplasmic inheritance of mitochondrial DNA in Xenopus. Dev Biol 29: 152-61

Den Hartog BK, Elling JW, Budowle B (2009) The impact of jumping alignments on mtDNA population analysis and database searching. Forensic Science International: Genetics Supplement Series 2: 315-316

Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, Wallace DC (1981) Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci U S A 78: 5768-72

Derbeneva OA, Starikovskaya EB, Wallace DC, Sukernik RI (2002) Traces of early Eurasians in the Mansi of northwest Siberia revealed by mitochondrial DNA analysis. Am J Hum Genet 70: 1009-14

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee HK, Vanecek T, Villems R, Zakharov I (2007) Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet 81: 1025-41

Derenko MV, Grzybowski T, Malyarchuk BA, Dambueva IK, Denisova GA, Czarny J, Dorzhu CM, Kakpakov VT, Miscicka-Sliwka D, Wozniak M, Zakharov IA (2003) Diversity of mitochondrial DNA lineages in South Siberia. Ann Hum Genet 67: 391-411

Doda JN, Wright CT, Clayton DA (1981) Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc Natl Acad Sci U S A 78: 6116-20

Egeland T, Salas A (2008) Estimating haplotype frequency and coverage of databases. PLoS ONE 3: e3988

Elson JL, Andrews RM, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (2001) Analysis of European mtDNAs for recombination. Am J Hum Genet 68: 145-153

Endicott P, Ho SY, Metspalu M, Stringer C (2009) Evaluating the mitochondrial timescale of human evolution. Trends Ecol Evol 24: 515-21

Eyre-Walker A, Smith NH, Smith JM (1999) How clonal are human mitochondria? Proc Biol Sci 266: 477-83

Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA, Jr., Zago MA, Ribeiro-dos-Santos AK, Santos SE, Petzl-Erler ML, Bonatto SL (2008) Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82: 583-92

Fernandez-Silva P, Enriquez JA, Montoya J (2003) Replication and transcription of mammalian mitochondrial DNA. Exp Physiol 88: 41-56

Fish J, Raule N, Attardi G (2004) Discovery of a major D-loop replication origin reveals two modes of human mtDNA synthesis. Science 306: 2098-101

Fisher RP, Topper JN, Clayton DA (1987) Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. Cell 50: 247-58

Fondevila M, Phillips C, Naveran N, Fernandez L, Cerezo M, Salas A, Carracedo A, Lareu MV (2008) Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur. Forensic Sci Int Genet 2: 212-8

Forster L, Forster P, Gurney SM, Spencer M, Huang C, Rohl A, Brinkmann B (2010) Evaluating length heteroplasmy in the human mitochondrial DNA control region. Int J Legal Med 124: 133-42

Forster P (2004) Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. Philos Trans R Soc Lond B Biol Sci 359: 255-64; discussion 264

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

Forster P, Torroni A, Renfrew C, Rohl A (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. Mol Biol Evol 18: 1864-81

Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, Deka R, Yanagihara R, Merriwether DA (2005) Expanding Southwest Pacific mitochondrial haplogroups P and Q. Mol Biol Evol 22: 1506-17

Garcia O, Fregel R, Larruga JM, Alvarez V, Yurrebaso I, Cabrera VM, Gonzalez AM (2010) Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge. Heredity

Gaspari M, Larsson NG, Gustafsson CM (2004) The transcription machinery in mammalian mitochondria. Biochim Biophys Acta 1659: 148-52

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A 77: 6715-9

Gilkerson RW, Schon EA, Hernandez E, Davidson MM (2008) Mitochondrial nucleoids maintain genetic autonomy but allow for functional complementation. J Cell Biol 181: 1117-28

Gillum AM, Clayton DA (1978) Displacement-loop replication initiation sequence in animal mitochondrial DNA exists as a family of discrete lengths. Proc Natl Acad Sci U S A 75: 677-81

Gillum AM, Clayton DA (1979) Mechanism of mitochondrial DNA replication in mouse L-cells: RNA priming during the initiation of heavy-strand synthesis. J Mol Biol 135: 353-68

Goios A, Prieto L, Amorim A, Pereira L (2008) Specificity of mtDNA-directed PCR-influence of NUclear MTDNA insertion (NUMT) contamination in routine samples and techniques. Int J Legal Med 122: 341-5

Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24: 757-68

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

Graf WD, Marin-Garcia J, Gao HG, Pizzo S, Naviaux RK, Markusic D, Barshop BA, Courchesne E, Haas RH (2000) Autism associated with the mitochondrial DNA G8363A transfer RNA(Lys) mutation. J Child Neurol 15: 357-61

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S (2010) A draft sequence of the Neandertal genome. Science 328: 710-22

Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prufer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajkovic D, Kucan Z, Gusic I, Wikstrom M, Laakkonen L, Kelso J, Slatkin M, Paabo S (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134: 416-26

Gyllensten U, Wharton D, Josefsson A, Wilson AC (1991) Paternal inheritance of mitochondrial DNA in mice. Nature 352: 255-7

Gyllensten U, Wharton D, Wilson AC (1985) Maternal inheritance of mitochondrial DNA during backcrossing of two species of mice. J Hered 76: 321-4

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK (1999) Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proceedings Biological Sciences / The Royal Society 266: 485-92

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK (2000) Evidence for mitochondrial DNA recombination in a human population of island Melanesia: Correction. Proc Biol Sci 267: 1

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95: 1961-7

Hayashi JI, Yonekawa H, Gotoh O, Watanabe J, Tagashira Y (1978) Strictly maternal inheritance of rat mitochondrial DNA. Biochem Biophys Res Commun 83: 1032-8

Hazkani-Covo E, Graur D (2007) A comparative analysis of numt evolution in human and chimpanzee. Mol Biol Evol 24: 13-8

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLoS Genet 6: e1000834

He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz Jr LA, Kinzler KW, Vogelstein B, Papadopoulos N (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature

Henn BM, Gignoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol Biol Evol 26: 217-30

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152-71

Herrnstadt C, Preston G, Howell N (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am J Hum Genet 72: 1585-6

Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet 69: 1113-26

Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, Rissler L, Victoriano PF, Yoder AD (2010) Phylogeography's past, present, and future: 10 years after Avise, 2000. Mol Phylogenet Evol 54: 291-301

Hixson JE, Wong TW, Clayton DA (1986) Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. J Biol Chem 261: 2384-90

Holt IJ (2009) Mitochondrial DNA replication and repair: all a flap. Trends Biochem Sci 34: 358-65

Holt IJ, Harding AE, Morgan-Hughes JA (1988) Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. Nature 331: 717-9

Holt IJ, He J, Mao CC, Boyd-Kirkup JD, Martinsson P, Sembongi H, Reyes A, Spelbrink JN (2007) Mammalian mitochondrial nucleoids: organizing an independently minded genome. Mitochondrion 7: 311-21

Holt IJ, Jacobs HT (2003) Response: The mitochondrial DNA replication bubble has not burst. Trends Biochem Sci 28: 355-6

Holt IJ, Lorimer HE, Jacobs HT (2000) Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. Cell 100: 515-24

Horai S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, Tajima K (1993) Peopling of the Americas, founded by four major lineages of mitochondrial DNA. Mol Biol Evol 10: 23-47

Howell N, Elson JL, Turnbull DM, Herrnstadt C (2004) African Haplogroup L mtDNA sequences show violations of clock-like evolution. Mol Biol Evol 21: 1843-54

Howell N, McCullogh DA, Kubacka I, Halvorson S, Mackey D (1992) The Sequence of Human mtDNA: The Question of Errors versus Polymorphisms. American Journal of Human Genetics 50: 6

Hutchison CA, 3rd, Newbold JE, Potter SS, Edgell MH (1974) Maternal inheritance of mammalian mitochondrial DNA. Nature 251: 536-8

Iborra FJ, Kimura H, Cook PR (2004) The functional organization of mitochondrial genomes in human cells. BMC Biol 2: 9

Ingman M, Gyllensten U (2003) Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res 13: 1600-6

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708-13

Innan H, Nordborg M (2002) Recombination or mutational hot spots in human mtDNA? Mol Biol Evol 19: 1122-7

Jin HJ, Tyler-Smith C, Kim W (2009) The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. PLoS ONE 4: e4210

Jorde LB, Bamshad M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931

Kaneda H, Hayashi J, Takahama S, Taya C, Lindahl KF, Yonekawa H (1995) Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. Proc Natl Acad Sci U S A 92: 4542-6

Kang D, Hamasaki N (2002) Maintenance of mitochondrial DNA integrity: repair and degradation. Curr Genet 41: 311-22

Kayser M (2010) The human genetic history of Oceania: near and remote views of dispersal. Curr Biol 20: R194-201

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhovel W, Stoneking M (2008) The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol 25: 1362-74

Kitchen A, Miyamoto MM, Mulligan CJ (2008) A three-stage colonization model for the peopling of the Americas. PLoS ONE 3: e1596

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R (1999a) Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9: 1331-4

Kivisild T, Kaldma K, Metspalu M, Parik J, Papiha SS, Villems R (1999b) The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World. Kluwer Academic/Plenum Publishers

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752-70

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72: 313-32

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373-87

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19: 1737-51

Kivisild T, Villems R (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol 13: 464-73

Ko LW, Sheu KF, Thaler HT, Markesbery WR, Blass JP (2001) Selective loss of KGDHC-enriched neurons in Alzheimer temporal cortex: does mitochondrial variation contribute to selective vulnerability? J Mol Neurosci 17: 361-9

Kondo R, Satta Y, Matsuura ET, Ishiwa H, Takahata N, Chigusa SI (1990) Incomplete maternal transmission of mitochondrial DNA in Drosophila. Genetics 126: 657-63

Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP (2003) Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73: 671-6

Kraytsberg Y, Schwartz M, Brown TA, Ebralidse K, Kunz WS, Clayton DA, Vissing J, Khrapko K (2004) Recombination of human mitochondrial DNA. Science 304: 981

Kumar S, Hedrick P, Dowling T, Stoneking M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931

Kumar S, Padmanabham PB, Ravuri RR, Uttaravalli K, Koneru P, Mukherjee PA, Das B, Kotal M, Xaviour D, Saheb SY, Rao VR (2008) The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. BMC Evol Biol 8: 230

Lambeck K, Esat TM, Potter EK (2002) Links between climate and sea levels for the past three million years. Nature 419: 199-206

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M (2008) The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. BMC Genomics 9: 267

LeDoux SP, Driggers WJ, Hollensworth BS, Wilson GL (1999) Repair of alkylation and oxidative damage in mitochondrial DNA. Mutat Res 434: 149-59

Legros F, Malka F, Frachon P, Lombes A, Rojo M (2004) Organization and dynamics of human mitochondrial DNA. J Cell Sci 117: 2653-62

Liu VW, Yang HJ, Wang Y, Tsang PC, Cheung AN, Chiu PM, Ng TY, Wong LC, Nagley P, Ngan HY (2003) High frequency of mitochondrial genome instability in human endometrial carcinomas. Br J Cancer 89: 697-701

Lombard J (1998) Autism: a mitochondrial disorder? Med Hypotheses 50: 497-500

Loogvali EL, Kivisild T, Margus T, Villems R (2009) Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS ONE 4: e8260

Loogvali EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, Metspalu E, Tambets K, Reidla M, Tolk HV, Parik J, Pennarun E, Laos S, Lunkina A, Golubenko M, Barac L, Pericic M, Balanovsky OP, Gusar V, Khusnutdinova EK, Stepanov V, Puzyrev V, Rudan P, Balanovska EV, Grechanina E, Richard C, Moisan JP, Chaventre A, Anagnou NP, Pappa KI, Michalodimitrakis EN, Claustres M, Golge M, Mikerezi I, Usanga E, Villems R (2004) Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21: 2012-21

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol 39: 174-90

Lum JK, Cann RL (2000) mtDNA lineage analyses: origins and migrations of Micronesians and Polynesians. Am J Phys Anthropol 113: 151-68

Lutz-Bonengel S, Sanger T, Pollak S, Szibor R (2004) Different methods to determine length heteroplasmy within the mitochondrial control region. Int J Legal Med 118: 274-81

Maca-Meyer N, Gonzalez AM, Pestano J, Flores C, Larruga JM, Cabrera VM (2003) Mitochondrial DNA transit between West Asia and North Africa inferred from U6 phylogeography. BMC Genet 4: 15

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034-6

Macaulay V, Richards M, Sykes B (1999) Mitochondrial DNA recombination-no need to panic. Proc Biol Sci 266: 2037-9; discussion 2041-2

Majumder PP (2010) The human genetic history of South Asia. Curr Biol 20: 184-7

Malka F, Lombes A, Rojo M (2006) Organization, dynamics and transmission of mitochondrial DNA: focus on vertebrate nucleoids. Biochim Biophys Acta 1763: 463-72

Malyarchuk BA, Rogozin IB, Berikov VB, Derenko MV (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum Genet 111: 46-53

Manfredi G, Thyagarajan D, Papadopoulou LC, Pallotti F, Schon EA (1997) The fate of human sperm-derived mtDNA in somatic cells. Am J Hum Genet 61: 953-60

Mao CC, Holt IJ (2009) Clinical and molecular aspects of diseases of mitochondrial DNA instability. Chang Gung Med J 32: 354-69

Mason PA, Matheson EC, Hall AG, Lightowlers RN (2003) Mismatch repair activity in mammalian mitochondria. Nucleic Acids Res 31: 1052-8

Maximo V, Lima J, Soares P, Botelho T, Gomes L, Sobrinho-Simoes M (2005) Mitochondrial D-Loop instability in thyroid tumours is not a marker of malignancy. Mitochondrion 5: 333-40

Maynard S, de Souza-Pinto NC, Scheibye-Knudsen M, Bohr VA (2010) Mitochondrial base excision repair assays. Methods

Melton T, Clifford S, Martinson J, Batzer M, Stoneking M (1998) Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes. Am J Hum Genet 63: 1807-23

Mellars P (2006) Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proceedings Of The National Academy Of Sciences Of The United States Of America 103: 9381-6

Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martinez-Fuentes A, Comas D (2008) Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol 8: 213

Merriweather DA, Kaestle FA (1999) Mitochondrial recombination? (continued). Science 285: 837

Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS (2005) Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc Natl Acad Sci U S A 102: 13034-9

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R (2004) Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 5: 26

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat 23: 125-33

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003a) Natural selection shaped regional mtDNA variation in humans. Proceedings Of The National Academy Of Sciences Of The United States Of America 100: 171-6

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003b) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171-6

Montoya J, Christianson T, Levens D, Rabinowitz M, Attardi G (1982) Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. Proc Natl Acad Sci U S A 79: 7195-9

Morten KJ, Ashley N, Wijburg F, Hadzic N, Parr J, Jayawant S, Adams S, Bindoff L, Bakker HD, Mieli-Vergani G, Zeviani M, Poulton J (2007) Liver mtDNA content increases during development: a comparison of methods and the importance of age- and tissue-specific controls for the diagnosis of mtDNA depletion. Mitochondrion 7: 386-95

Mourier T, Hansen AJ, Willerslev E, Arctander P (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol 18: 1833-7

Mulligan CJ, Hunley K, Cole S, Long JC (2004) Population genetics, history, and health patterns in native americans. Annu Rev Genomics Hum Genet 5: 295-315

Namslauer I, Brzezinski P (2009) A mitochondrial DNA mutation linked to colon cancer results in proton leaks in cytochrome c oxidase. Proc Natl Acad Sci U S A 106: 3402-7

Nei M, Tajima F (1981) DNA polymorphism detectable by restriction endonucleases. Genetics 97: 145-63

Nishigaki Y, Ueno H, Coku J, Koga Y, Fujii T, Sahashi K, Nakano K, Yoneda M, Nonaka M, Tang L, Liou CW, Paquis-Flucklinger V, Harigaya Y, Ibi T, Goto YI, Hosoya H, Dimauro S, Hirano M, Tanaka M (2010) Extensive screening system using suspension array technology to detect mitochondrial DNA point mutations. Mitochondrion

Normile D (2009) Genetics. SNP study supports southern migration route to Asia. Science 326: 1470

O'Rourke DH, Raff JA (2010) The human genetic history of the Americas: the final frontier. Curr Biol 20: R202-7

Ohno K, Tanaka M, Suzuki H, Ohbayashi T, Ikebe S, Ino H, Kumar S, Takahashi A, Ozawa T (1991) Identification of a possible control element, Mt5, in the major noncoding region of mitochondrial DNA by intraspecific nucleotide conservation. Biochem Int 24: 263-72

Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt HJ, Torroni A (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314: 1767-70

Pakendorf B, Novgorodov IN, Osakovskij VL, Stoneking M (2007) Mating patterns amongst Siberian reindeer herders: inferences from mtDNA and Y-chromosomal analyses. Am J Phys Anthropol 133: 1013-27

Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6: 165-83

Pala M, Achilli A, Olivieri A, Kashani BH, Perego UA, Sanna D, Metspalu E, Tambets K, Tamm E, Accetturo M, Carossa V, Lancioni H, Panara F, Zimmermann B, Huber G, Al-Zahery N, Brisighelli F, Woodward SR, Francalacci P, Parson W, Salas A, Behar DM, Villems R, Semino O, Bandelt HJ, Torroni A (2009) Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians. Am J Hum Genet 84: 814-21

Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75: 966-78

Palanichamy MG, Zhang CL, Mitra B, Malyarchuk B, Derenko M, Chaudhuri TK, Zhang YP (2010) Mitochondrial haplogroup N1a phylogeography, with implication to the origin of European farmers. BMC Evol Biol 10: 304

Palmieri L, Persico AM (2010) Mitochondrial dysfunction in autism spectrum disorders: Cause or effect? Biochim Biophys Acta 1797: 1130-1137

Parson W (2007) The art of reading sequence electropherograms. Ann Hum Genet 71: 276-8; author reply 279-280

Parson W, Bandelt HJ (2007) Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1: 13-9

Parson W, Dur A (2007) EMPOP--a forensic mtDNA database. Forensic Sci Int Genet 1: 88-92

Parsons TJ, Irwin JA (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931

Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong QP, Myres NM, Salas A, Semino O, Bandelt HJ, Woodward SR, Torroni A (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1-8

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH, Carossa V, Ekins JE, Gomez-Carballa A, Huber G, Zimmermann B, Corach D, Babudri N, Panara F, Myres NM, Parson W, Semino O, Salas A, Woodward SR, Achilli A, Torroni A (2010) The initial peopling of the Americas: A growing number of founding mitochondrial genomes from Beringia. Genome Res

Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Annals Of Human Genetics 65: 439-58

Pereira L, Richards M, Goios A, Alonso A, Albarran C, Garcia O, Behar DM, Golge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res 15: 19-24

Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, Martins H, Soares P, Behar DM, Richards MB, Macaulay V (2010) Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol 10: 390

Pham XH, Farge G, Shi Y, Gaspari M, Gustafsson CM, Falkenberg M (2006) Conserved sequence box II directs transcription termination and primer formation in mitochondria. J Biol Chem 281: 24647-52

Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet 1: 273-80

Pierson MJ, Martinez-Arias R, Holland BR, Gemmell NJ, Hurles ME, Penny D (2006) Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. Mol Biol Evol 23: 1966-75

Piko L, Taylor KD (1987) Amounts of mitochondrial DNA and abundance of some mitochondrial gene transcripts in early mouse embryos. Dev Biol 123: 364-74

Podini D, Vallone PM (2009) SNP genotyping using multiplex single base primer extension assays. Methods Mol Biol 578: 379-91

Poloni ES, Naciri Y, Bucho R, Niba R, Kervaire B, Excoffier L, Langaney A, Sanchez-Mazas A (2009) Genetic evidence for complexity in ethnic differentiation and history in East Africa. Ann Hum Genet 73: 582-600

Prieto L, Montesino M, Salas A, Alonso A, Albarran C, Alvarez S, Crespillo M, Di Lonardo AM, Doutremepuich C, Fernandez-Fernandez I, de la Vega AG, Gusmao L, Lopez CM, Lopez-Soto M, Lorente JA, Malaghini M, Martinez CA, Modesti NM, Palacio AM, Paredes M, Pena SD, Perez-Lezaun A, Pestano JJ, Puente J, Sala A, Vide M, Whittle MR, Yunis JJ, Gomez J (2003) The 2000-2001 GEP-ISFG Collaborative Exercise on mtDNA: assessing the cause of unsuccessful mtDNA PCR amplification of hair shaft samples. Forensic Sci Int 134: 46-53

Pyle A, Foltynie T, Tiangyou W, Lambert C, Keers SM, Allcock LM, Davison J, Lewis SJ, Perry RH, Barker R, Burn DJ, Chinnery PF (2005) Mitochondrial DNA haplogroup cluster UKJT reduces the risk of PD. Ann Neurol 57: 564-7

Qian YP, Chu ZT, Dai Q, Wei CD, Chu JY, Tajima A, Horai S (2001) Mitochondrial DNA polymorphisms in Yunnan nationalities in China. J Hum Genet 46: 211-20

Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, van Helden PD, Hoal EG, Behar DM (2010) Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. Am J Hum Genet

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596-601

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet 23: 437-41

Quintans B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251-7

Rajkumar R, Banerjee J, Gunturi HB, Trivedi R, Kashyap VK (2005) Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol Biol 5: 26

Rando JC, Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM, Bandelt HJ (1998) Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann Hum Genet 62: 531-50

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro AS, Stoneking M (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12: 604-15

Reilly JG, Thomas CA, Jr. (1980) Length polymorphisms, restriction site variation, and maternal inheritance of mitochondrial DNA of Drosophila melanogaster. Plasmid 3: 109-15

Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. PLoS Biol 2: E273

Richard SM, Bailliet G, Paez GL, Bianchi MS, Peltomaki P, Bianchi NO (2000) Nuclear and mitochondrial genome instability in human breast cancer. Cancer Res 60: 4231-7

Richards M, Macaulay V (2001) The mitochondrial gene tree comes of age. Am J Hum Genet 68: 1315-20

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251-76

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62: 241-60

Roberti M, Musicco C, Polosa PL, Milella F, Gadaleta MN, Cantatore P (1998) Multiple protein-binding sites in the TAS-region of human and rat mitochondrial DNA. Biochem Biophys Res Commun 243: 36-40

Roostalu U, Kutuev I, Loogvali EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R (2007) Origin and expansion of haplogroup H, the dominant human

mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol 24: 436-48

Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14: 255-74

Salas A, Acosta A, Alvarez-Iglesias V, Cerezo M, Phillips C, Lareu MV, Carracedo A (2008a) The mtDNA ancestry of admixed Colombian populations. Am J Hum Biol 20: 584-91

Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168: 1-13

Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005a) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891-9

Salas A, Carracedo A, Richards M, Macaulay V (2005b) Charting the ancestry of African Americans. American Journal Of Human Genetics 77: 676-80

Salas A, Jaime JC, Alvarez-Iglesias V, Carracedo A (2008b) Gender bias in the multiethnic genetic composition of central Argentina. J Hum Genet 53: 662-74

Salas A, Prieto L, Montesino M, Albarran C, Arroyo E, Paredes-Herrera MR, Di Lonardo AM, Doutremepuich C, Fernandez-Fernandez I, de la Vega AG, Alves C, Lopez CM, Lopez-Soto M, Lorente JA, Picornell A, Espinheira RM, Hernandez A, Palacio AM, Espinoza M, Yunis JJ, Perez-Lezaun A, Pestano JJ, Carril JC, Corach D, Vide MC, Alvarez-Iglesias V, Pinheiro MF, Whittle MR, Brehm A, Gomez J (2005c) Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02-03 proficiency testing trial. Forensic Sci Int 148: 191-8

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082-111

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454-65

Salas A, Richards M, Lareu MV, Sobrino B, Silva S, Matamoros M, Macaulay V, Carracedo A (2005d) Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. American Journal Of Physical Anthropology 128: 855-60

Satoh M, Kuroiwa T (1991) Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res 196: 137-40

Scozzari R, Torroni A, Semino O, Sirugo G, Brega A, Santachiara-Benerecetti AS (1988) Genetic studies on the Senegal population. I. Mitochondrial DNA polymorphisms. Am J Hum Genet 43: 534-44

Schlebusch CM, Naidoo T, Soodyall H (2009) SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis 30: 3657-64

Schurr TG, Sherry ST (2004) Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am J Hum Biol 16: 420-39

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P, Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463: 943-7

Schwartz M, Vissing J (2002) Paternal inheritance of mitochondrial DNA. N Engl J Med 347: 576-80

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol 20: 2-15

Singer-Sam J, Tanguay RL, Riggs AD (1989) Use of chelex to improve the PCR signal from a small number of cells. Amplifications: A Forum for PCR users 3: 1

Slate J, Gemmell NJ (2004) Eve 'n' Steve: recombination of human mitochondrial DNA. TRENDS in Ecology and Evolution 19: 3

Soares P, Achilli A, Semino O, Davies W, Macaulay V, Bandelt HJ, Torroni A, Richards MB (2010) The archaeogenetics of Europe. Curr Biol 20: R174-83

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740-59

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo JH, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards MB (2011) Ancient voyaging and Polynesian origins. Am J Hum Genet 88: 239-47

Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB (2008) Climate change and postglacial human dispersals in southeast Asia. Mol Biol Evol 25: 1209-18

Stevanovitch A, Gilles A, Bouzaid E, Kefi R, Paris F, Gayraud RP, Spadoni JL, El-Chenawi F, Beraud-Colomb E (2004) Mitochondrial DNA sequence diversity in a sedentary population from Egypt. Ann Hum Genet 68: 23-39

Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet 67: 1029-32

Stoneking M, Delfin F (2010) The human genetic history of East Asia: weaving a complex tapestry. Curr Biol 20: R188-93

Stoneking M, Jorde LB, Bhatia K, Wilson AC (1990) Geographic variation in human mitochondrial DNA from Papua New Guinea. Genetics 124: 717-33

Stoneking M, Nasidze I (2006) The patient is not dead yet: premature autopsy of a mtDNA data set. Ann Hum Genet 70: 327-31

Stoneking M, Soodyall H (1996) Human evolution and the mitochondrial genome. Current Opinion In Genetics & Development 6: 731-6

Sun C, Kong QP, Palanichamy MG, Agrawal S, Bandelt HJ, Yao YG, Khan F, Zhu CL, Chaudhuri TK, Zhang YP (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol Biol Evol 23: 683-90

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G (1999) Ubiquitin tag for sperm mitochondria. Nature 402: 371-2

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G (2000) Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. Biol Reprod 63: 582-90

Suzuki H, Hosokawa Y, Nishikimi M, Ozawa T (1991) Existence of common homologous elements in the transcriptional regulatory regions of human nuclear genes and mitochondrial gene for the oxidative phosphorylation system. J Biol Chem 266: 2333-8

Sweet D, Lorente M, Valenzuela A, Lorente JA, Alvarez JC (1996) Increasing DNA extraction yield from saliva stains with a modified Chelex method. Forensic Sci Int 83: 10

Tanaka M, Cabrera VM, Gonzalez AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14: 1832-50

Taylor RW, Turnbull DM (2005) Mitochondrial DNA mutations in human disease. Nat Rev Genet 6: 389-402

Thieme M, Lottaz C, Niederstatter H, Parson W, Spang R, Oefner PJ (2009) ReseqChip: automated integration of multiple local context probe data from the MitoChip array in mitochondrial DNA sequence assembly. BMC Bioinformatics 10: 440

Thompson WE, Ramalho-Santos J, Sutovsky P (2003) Ubiquitination of prohibitin in mammalian sperm mitochondria: possible roles in the regulation of mitochondrial inheritance and sperm quality control. Biol Reprod 69: 254-60

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Molecular Biology And Evolution 24: 2180-95

Topf AL, Gilbert MT, Dumbacher JP, Hoelzel AR (2006) Tracing the phylogeography of human populations in Britain based on 4th-11th century mtDNA genotypes. Mol Biol Evol 23: 152-61

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22: 339-45

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137-52

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumaki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001a) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69: 844-52

Torroni A, Miller JA, Moore LG, Zamudio S, Zhuang J, Droma T, Wallace DC (1994) Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. Am J Phys Anthropol 93: 189-99

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001b) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? American Journal Of Human Genetics 69: 1348-56

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53: 563-90

Torroni A, Schurr TG, Yang CC, Szathmary EJ, Williams RC, Schanfield MS, Troup GA, Knowler WC, Lawrence DN, Weiss KM, et al. (1992) Native American mitochondrial DNA analysis

indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130: 153-62

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80: 71-7

Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ, Lee ZY, Lin M (2005) Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol 3: e247

Tsuzuki T, Nomiyama H, Setoyama C, Maeda S, Shimada K (1983) Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. Gene 25: 223-9

Tully G, Bar W, Brinkmann B, Carracedo A, Gill P, Morling N, Parson W, Schneider P (2001) Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. Forensic Sci Int 124: 83-91

Turner C, Killoran C, Thomas NS, Rosenberg M, Chuzhanova NA, Johnston J, Kemel Y, Cooper DN, Biesecker LG (2003) Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Hum Genet 112: 303-9

Vallone PM, Jakupciak JP, Coble MD (2007) Forensic application of the Affymetrix human mitochondrial resequencing array. Forensic Sci Int Genet 1: 196-8

van der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC, Welsh-Bohmer KA, Saunders AM, Roses AD, Small GW, Schmechel DE, Murali Doraiswamy P, Gilbert JR, Haines JL, Vance JM, Pericak-Vance MA (2004) Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neurosci Lett 365: 28-32

van Eijsden RG, Gerards M, Eijssen LM, Hendrickx AT, Jongbloed RJ, Wokke JH, Hintzen RQ, Rubio-Gozalbo ME, De Coo IF, Briem E, Tiranti V, Smeets HJ (2006) Chip-based mtDNA mutation screening enables fast and reliable genetic diagnosis of OXPHOS patients. Genet Med 8: 620-7

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: E386-94

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503-7

Wai T, Teoli D, Shoubridge EA (2008) The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. Nat Genet 40: 1484-8

Walberg MW, Clayton DA (1983) In vitro transcription of human mitochondrial DNA. Identification of specific light strand transcripts from the displacement loop region. J Biol Chem 258: 1268-75

Wallace D, Ye J, Neckelmann S, Singh G, Webster K, Greenberg B (1987) Sequence analysis of cDNAs for the human and bovine ATP synthase b subunit: mitochondrial DNA genes sustain seventeen times more mutations. Curr Genet 12: 10

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238: 211-30

Wallace DC, Singh G, Lott MT, Hodge JA, Schurr TG, Lezza AM, Elsas LJ, 2nd, Nikoskelainen EK (1988) Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science 242: 1427-30

Wallis GP (2000) Mitochondrial recombination or coevolution of sites? Trends Ecol Evol 15: 1

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A (2007) Genetic variation and population structure in native Americans. PLoS Genet 3: e185

Wang SB, Weng WC, Lee NC, Hwu WL, Fan PC, Lee WT (2008) Mutation of mitochondrial DNA G13513A presenting with Leigh syndrome, Wolff-Parkinson-White syndrome and cardiomyopathy. Pediatr Neonatol 49: 145-9

Ward RH, Frazier BL, Dew-Jager K, Paabo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci U S A 88: 8720-4

Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, Lu D, Chakraborty R, Jin L (2004) Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet 74: 856-65

White DJ, Gemmell NJ (2009) Can indirect tests detect a known recombination event in human mtDNA? Mol Biol Evol 26: 1435-9

White DJ, Wolff JN, Pierson M, Gemmell NJ (2008) Revealing the hidden complexities of mtDNA inheritance. Mol Ecol 17: 4925-42

Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002) Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Int 129: 35-42

Woischnik M, Moraes CT (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res 12: 885-93

Wolff JN, Gemmell NJ (2008) Lost in the zygote: the dilution of paternal mtDNA upon fertilization. Heredity 101: 429-34

Xing J, Chen M, Wood CG, Lin J, Spitz MR, Ma J, Amos CI, Shields PG, Benowitz NL, Gu J, de Andrade M, Swan GE, Wu X (2008) Mitochondrial DNA content: its genetic heritability and association with renal cell carcinoma. J Natl Cancer Inst 100: 1104-12

Xue F, Wang Y, Xu S, Zhang F, Wen B, Wu X, Lu M, Deka R, Qian J, Jin L (2008) A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. Eur J Hum Genet 16: 705-17

Yang MY, Bowmaker M, Reyes A, Vergani L, Angeli P, Gringeri E, Jacobs HT, Holt IJ (2002) Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. Cell 111: 495-505

Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. Forensic Sci Int 141: 1-6

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002a) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70: 635-51

Yao YG, Kong QP, Salas A, Bandelt HJ (2008) Pseudomitochondrial genome haunts disease studies. J Med Genet 45: 769-72

Yao YG, Nie L, Harpending H, Fu YX, Yuan ZG, Zhang YP (2002b) Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am J Phys Anthropol 118: 63-76

Yao YG, Ogasawara Y, Kajigaya S, Molldrem JJ, Falcao RP, Pintao MC, McCoy JP, Jr., Rizzatti EG, Young NS (2007) Mitochondrial DNA sequence variation in single cells from leukemia patients. Blood 109: 756-62

Yao YG, Salas A, Logan I, Bandelt HJ (2009) mtDNA data mining in GenBank needs surveying. Am J Hum Genet 85: 929-33; author reply 933

Yoza BK, Bogenhagen DF (1984) Identification and in vitro capping of a primary transcript of human mitochondrial DNA. J Biol Chem 259: 3909-15

Yu M, Zhou Y, Shi Y, Ning L, Yang Y, Wei X, Zhang N, Hao X, Niu R (2007) Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients. IUBMB Life 59: 450-7

Zouros E, Freeman KR, Ball AO, Pogson GH (1992) Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel Mytilus. Nature 359: 412-4

VIII APPENDIX

VIII.1 RESUMEN TESIS CASTELLANO

Las mitocondrias son orgánulos celulares cuya principal función es la producción de energía para la célula. Dependiendo de las necesidades energéticas, las mitocondrias pueden estar presentes en cantidades variables. Así entre las células que presentan mayor número de mitocondrias figuran las de tejido cardíaco y las de tejido cerebral. Una característica particular de las mitocondrias es que poseen su propio genoma, el cual también está presente en número variable dentro de cada mitocondria.

En 1981 se secuenció por primera vez el genoma mitocondrial humano (Anderson et al. 1981), y dicha secuencia fue tomada como referencia denominándose *Reference Cambridge Sequence* (CRS). De acuerdo con este trabajo se estableció que la numeración fuese en base a la cadena L (*light strand*, denominada así debido a que es rica en bases pirimidínicas). En 1999 fue resecuenciada (Andrews et al. 1999) detectándose algunos errores, esta nueva secuencia se denomina *revised Cambridge Reference Sequence* (rCRS) y es la utilizada actualmente como referencia.

El genoma mitocondrial (ADNmt) es circular, cerrado, de doble cadena, está compuesto por aproximadamente 16569 b.p.(pair of bases) con 37 genes, 13 de ellos codifican para proteínas estructurales de la cadena respiratoria, 2 codifican ARN ribosómicos y los 22 restantes codifican ARN transferentes. Además contiene una región no codificante, conocida como región control que presenta una alta variabilidad.

Además de la localización fuera del núcleo y su polihaploidía (hay múltiples copias por cada célula), el ADNmt presenta herencia materna, no recombina, es altamente codificante y carece de protección por histonas. Esta última característica hace que presente una tasa de mutación mayor que el ADN nuclear pudiendo llegar a ser 100-200 veces superior en el caso de la denominada región control (que es la no codificante).

En ocasiones existe más de una variante mitocondrial y este estado se conoce como heteroplasmía. Esta mezcla puede heredarse o bien puede surgir debido a una mutación somática. En el caso de estar presente, el grado de heteroplasmía no tiene porqué ser constante, pudiendo ser diferente dentro de un individuo o dentro de un tejido en el transcurso de la división celular. Esto se debe a que en la célula continuamente se produce la replicación relajada de las mitocondrias, que es un proceso aleatorio por el cual las mitocondrias se fusionan y después sufren una bipartición.

Debido a todas las peculiaridades que presenta el ADNmt, previamente mencionadas, el análisis de su variabilidad se puede aplicar a diferentes campos de estudio:

- En el campo de la genética de poblaciones: permite analizar el origen y la dispersión de las poblaciones humanas modernas
- En el campo de la genética forense: permite investigar muestras forenses degradadas o sin ADN nuclear (como huesos y cabellos sin bulbo).
- En el campo de la genética clínica: permite la identificación de posibles mutaciones que pueden causar una enfermedad o participar de ella; además de detectar una posible estratificación en los estudios caso-control que podría suponer el establecimiento de una relación erróneas entre una variante genética y una enfermedad.

Para detectar la variabilidad en los diferentes tipos de muestras se pueden llevar a cabo diversas técnicas. Tradicionalmente, se empezó a analizar la variabilidad entre individuos por medio de RFLPs (*Restriction Fragment Length Polymorphisms*) que son los diferentes patrones de corte que producen los enzimas de restricción dependiendo de si está presente o no la variante que hace que el enzima reconozca la zona de corte.

Posteriormente, con el desarrollo de la PCR (*Polymerase Chain Reaction*) y los métodos de secuenciación, el estudio de la variabilidad encontrada por RFLPs se combinó con el de la detección de variantes de secuencia por medio de la secuenciación. En un principio se comenzó secuenciando pequeños fragmentos de la región control, posteriormente fueron ampliadas las zonas genotipadas de la región control y más recientemente se ha extendido el análisis de la variabilidad del ADNmt a la secuenciación del genoma completo.

En el presente proyecto de tesis, además del desarrollo de un protocolo para la obtención del genoma completo, se han desarrollado diferentes métodos de minisecuenciación, que consisten en la incorporación de una sola base en una posición conocida que presenta variabilidad entre individuos, como son SNaPShot (Applied Biosystems, Foster City, CA, USA) que detecta la variante por fluorescencia o por primera vez por medio de MALDI-TOF MS (Sequenom, San Diego, CA, USA) que detecta la variante por espectrometría de masas.

VIII.1.1 OBJETIVOS

Aunque desde la primera secuenciación en 1981, se ha avanzado mucho en el conocimiento de la variabilidad del ADNmt, en parte gracias a las mejoras de las técnicas para su obtención, aún quedan diferentes cuestiones por resolver.

Como objetivo general, con el presente proyecto de tesis se ha pretendido mejorar el conocimiento de la variabilidad del ADNmt por medio del desarrollo y mejora de diferentes técnicas de genotipado. La aplicación de este conocimiento a los diferentes campos permitió establecer objetivos más concretos dependiendo de cada tipo de estudio.

En el caso de la genética de poblaciones:

La población de Europa es una de las más estudiadas, por ello se conoce que alrededor del 40% de la población pertenece al mismo grupo mitocondrial (haplogrupo). Por ello es necesario el desarrollo de técnicas de genotipado de la región codificante que permitan discriminar muestras que en un principio pueden ser clasificadas juntas en base a su región control.

Existen dentro de la filogenia del ADNmt diferentes ramas no bien definidas debido a la escasa información disponible. Por ello es necesario el desarrollo de un método de genotipado de todo el genoma mitocondrial de diferentes muestras correspondientes a dichas ramas.

La población Nativo Americana es la que ha experimentado mayor deriva y menor tiempo de divergencia desde la salida de África. Sin embargo, eventos poblacionales recientes como la colonización de América o el tráfico de esclavos han cambiado drásticamente esta condición. El estudio de diferentes grupos poblacionales pertenecientes a múltiples regiones y diferentes países, es necesario para identificar la influencia que dichos eventos ha tenido en las poblaciones actuales.

La población Africana es la que presenta mayor diversidad entre toda la población mundial. Por ello se considera necesario un estudio a mayor profundidad que los realizados anteriormente, desarrollando un método que permita un genotipado a gran escala. Al realizar dicho genotipado en diferentes regiones de los continentes africano y americano además de mejorar el conocimiento de la distribución de la variabilidad mitocondrial en dichas zonas, se puede llegar a discriminar mejor las zonas que fueron el origen del tráfico de esclavos.

La frecuencia de haplogrupos mitocondriales sub-saharianos en Europa no supera el 1-2% en los diferentes estudios realizados hasta la fecha. Sin embargo, no está presente en la literatura un estudio a fondo de dicha presencia. Por ello se planteó la necesidad de realizar un análisis de genomas completos de todos aquellos individuos disponibles, pertenecientes a estos linajes además de una completa revisión de la literatura de diferentes linajes para la actualización de la filogenia.

En el caso de la genética forense:

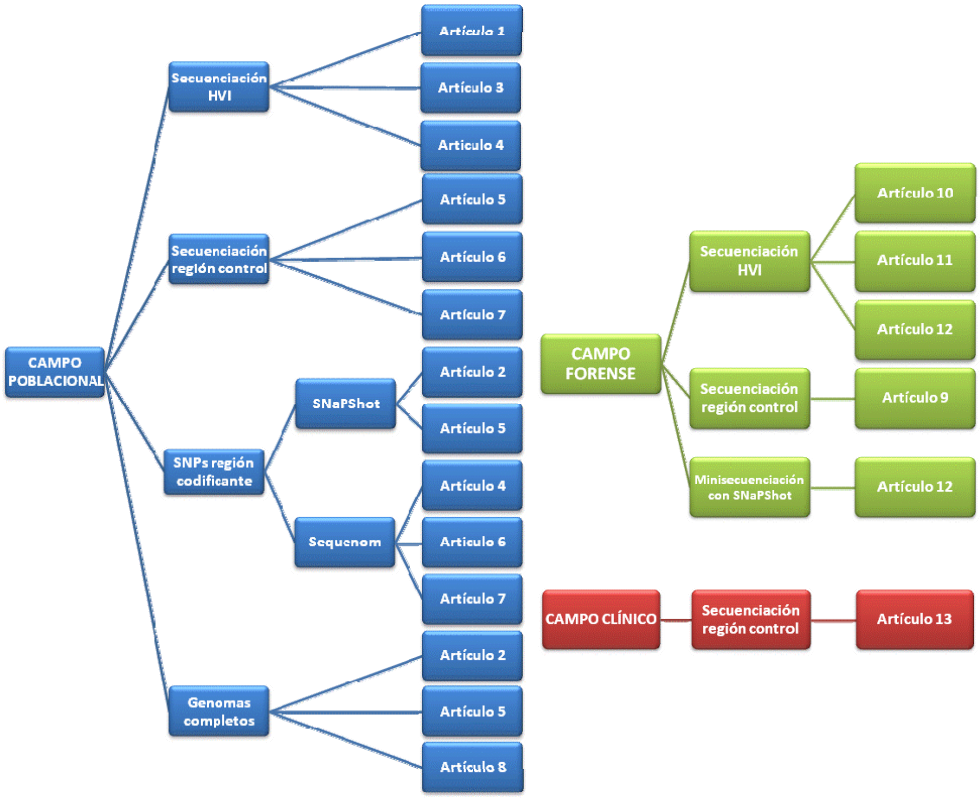
En ocasiones la obtención de un genotipado que permita la inclusión o exclusión de una muestra en un caso es difícil debido a que en muchas ocasiones se trata de muestras con poco ADN o éste se encuentra degradado. Por ello es prioritario el desarrollo de técnicas que optimicen la obtención de resultados y la validación de éstas para el uso forense al cumplir unos controles de calidad.

En el caso de la genética clínica:

La mejora en la obtención de genotipos ha resultado en un crecimiento muchas veces exponencial del número de artículos que tratan de relacionar una variante mitocondrial con una enfermedad o su predisposición a ella. Muchas veces la aparición de estos resultados es debida a contaminaciones, errores de genotipado, errores de transcripción de las posiciones que cambian... etc. Por ello es necesario el establecimiento de unos criterios de calidad además de la consideración de las características del ADNmt y su contexto filogenético que permitan descartar esos resultados erróneos.

VIII.1.2 MATERIAL Y MÉTODOS

Con el fin de alcanzar los objetivos antes mencionados se llevaron a cabo diferentes técnicas de genotipado que iban desde la secuenciación de fragmentos de la región control hasta la secuenciación de todo el genoma mitocondrial, con la detección por fluorescencia en electroforesis capilar. Por otro lado también se llevaron a cabo técnicas de genotipado cuyos resultados informan de variaciones puntuales por medio de minisecuenciaciones y una posterior detección por fluorescencia (en el caso de SNaPShot) o la detección que detecta la variante por espectrometría de masas MALDI-TOF MS. Dependiendo del estudio y la calidad de la muestra, se llevaron a cabo alguna de las técnicas antes mencionadas, lo cual viene esquematizado en la siguiente figura



VIII.1.3 RESULTADOS

VIII.1.3.1 GENÉTICA DE POBLACIONES

VIII.1.3.1.1 Artículo 1: La ancestralidad de las poblaciones colombianas mezcladas American Journal of Human Biology.

Se llevó a cabo la secuenciación de la región HVI del ADN mitocondrial en un total de 185 individuos de Colombia. Adicionalmente, se llevó a cabo la secuenciación de la región HVII en parte de las muestras. Dichas muestras fueron recogidas junto con información de etnicidades autoasignadas por los propios individuos, que incluían “mestizos”, “mulatos” y “Afro-colombianos”. Además para realizar inferencias filogeográficas y comparaciones poblacionales fueron utilizadas bases de datos que englobaban más de 4.300 linajes de Nativo-Americanos, 6.800 Africanos y 15.600 Europeos

Se observó que mulatos y Afro-colombianos tienen un componente mitocondrial africano dominante, mientras que los mestizos portan mayoritariamente haplotipos Nativo Americanos. Todas las poblaciones analizadas presentan altos índices de diversidad nucleotídica sin signos de bruscos episodios de deriva genética. América Central y Sudamérica figuran como el origen más probable de los linajes Nativo Americanos presentes en esta población, mientras que el centro-oeste, el sudeste y el suroeste Africano serían el origen más probable de los linajes africanos presentes entre dichos individuos.

Estos resultados difieren de estudios previos con grupos poblacionales similares en cuanto a las frecuencias haplotípicas. Esto puede deberse a varios aspectos: (a) en Colombia la ancestralidad autoasignada no es adecuada para indicar la etnicidad de los individuos. (b) nuestros resultados no apoyan el uso de las descripciones anacrónicas de razas (mestizos, mulatos …) principalmente porque no se corresponden con ningún grupo genéticamente homogéneo y (c) los estudios que se apoyan en estos términos para describir el grupo poblacional al que pertenece un individuo, tienden a tratar dichos grupos como genéticamente homogéneos, con lo que se eleva las posibilidades de error de tipo I (falsos positivos) en los estudios médicos en este país así como interpretaciones erróneas en el caso de investigaciones forenses.

VIII.1.3.1.2 Artículo 2: Nueva población y características filogenéticas de la variación interna del macro-haplogrupo R0 PLOS One Abril

El macro-haplogrupo R0 engloba el linaje de ADN mitocondrial (ADNmt) más común entre las poblaciones en Eurasia occidental, el haplogrupo H (~40%), y otros clados menos frecuentes. Los linajes de R0 están muy mal caracterizados en la región control, lo cual hace necesario el análisis de polimorfismos de la región codificante para aumentar la resolución al inferir eventos demográficos en Europa occidental.

Se llevó a cabo la secuenciación de la región HVI del ADN mitocondrial de 518 individuos procedentes de diferentes regiones del norte de la Península Ibérica. En aquellas muestras que pertenecían a R0 (57%) fueron además genotipados 71 SNPs codificantes con el fin de caracterizar la filogenia de R0, tanto las ramas principales como las subramas.

Se encontró que el norte de la Península Ibérica muestra niveles moderados de estratificación poblacional, por ejemplo, el haplogrupo V presenta los mayores niveles en Cantabria pero menores en Galicia y Cataluña. Cuando se compararon con otras poblaciones europeas y de Oriente Medio, los haplogrupos H1 H3 y H5a muestran picos de frecuencia en la región Franco-Cantábrica disminuyendo de Oeste a Este y Sur de Europa.

Además, se genotipó el genoma mitocondrial completo de un nuevo clado de H autóctono del País Vasco, denominado H2a5. Su edad de coalescencia se estimó en 1285±161 años. Su reciente edad junto con episodios locales de aislamiento, podrían explicar el hecho de que H2a5 haya permanecido confinado a este área hasta la actualidad.

VIII.1.3.1.3 Artículo 3: "Ecos" mitocondriales del primer asentamiento y continuidad genética en El Salvador PLoS ONE

Se llevó a cabo la secuenciación de la región control del ADN mitocondrial de 90 individuos de El Salvador. Además se llevó a cabo la recopilación de más de 3.985 perfiles de la región control de dominio público y de la literatura para poder llevar a cabo comparaciones entre poblaciones.

Los resultados revelaron un componente mitocondrial Nativo Americano predominante, el mayoritario con hasta el 90% en este país es A2 a diferencia de otras poblaciones del Norte Centro o Sur de América. El haplogrupo A2 muestra una filogenia en estrella y es muy diverso en una zona del ADNmt (45% entre 16.030-16.365), algo que aún no se ha observado en otras poblaciones americanas. Se usaron dos diferentes aproximaciones bayesianas para estimar las proporciones de mezcla en el Salvador, mostrándose que la mayor parte de los linajes observados proceden de América del Norte. Con un análisis fundador preliminar se estableció que el asentamiento en el Salvador sucedió hace 13.400±5.200 años. La edad de fundación de A2 en el Salvador es similar a la de A2 en todo América, lo cual sugiere que la colonización de esta región ocurrió unos pocos miles de años después de la entrada inicial en las Américas.

Los resultados son compatibles con la hipótesis de que la variabilidad actual de A2 en El Salvador representa a gran parte del componente indígena actual de la región. Además se observó una pequeña proporción de ADN mitocondrial procedente de Eurasiay África (5%) lo cual indica que el tráfico de esclavos del Atlántico tuvo un impacto demográfico muy bajo, en contraste con la prevalencia de linajes mitocondriales sub-Saharianos en poblaciones vecinas de la costa caribeña.

VIII.1.3.1.4 Artículo 4: Aplicaciones de la tecnología MALDI-TOF MS a estudios poblacionales a gran escala. Electrophoresis

El análisis de la variabilidad del ADN mitocondrial es comúnmente llevado a cabo en diferentes campos de la investigación biomédica. Se propuso llevar a cabo el análisis de la variación de SNPs de la región codificante del ADN mitocondrial a un alto nivel de resolución filogenética mediante espectrometría de masas (MS) por MALDI-TOF. Para comprobar la aplicabilidad de la técnica, fue elegida la filogenia africana aunque cualquier otra parte de la filogenia mundial (u otro grupo de SNPs) podría haber sido igualmente adecuada para el genotipado por MALDI-TOF MS.

Así la selección de los SNPs intentó cubrir todas las ramas principales y menores que definen el árbol mitocondrial africano, incluyendo el macrohaplogrupo L y los haplogrupos M1 y U6. Se seleccionaron 230 SNPs, que fueron genotipados en 30 individuos de Mozambique y en 60 de la cuenca del Chad. Se utilizaron diferentes controles internos (como la correspondencia filogenética y la secuenciación automática) para evaluar la reproducibilidad de la técnica, la cual resultó del 100% al usar muestras que previamente se habían sometido a la secuenciación del genoma mitocondrial completo.

Las ventajas de la técnica son además tratadas en comparación con otros métodos como la minisecuenciación, dada la relevancia de su naturaleza de alto rendimiento, lo cual es particularmente adecuado para estudios clínicos de caso-control, bases de datos forenses o estudios poblacionales.

VIII.1.3.1.5 Artículo 5: Conectando los componentes genéticos Sub-Sahariano y Europeo: herencia materna y paterna de los nómadas Tuaregs del Sahel Africano. European Journal of Human Genetics

Actualmente los *Tuareg* viven entre el Sáhara y el Sahel. Escritores de la antigüedad sugirieron que sus ancestros fueron los Garamantes del desierto de Libia. Sin embargo, vvidencias biológicas, basadas en marcadores genéticos clásicos, han indicado parentesco con los Bejas del Este de Sudán.

Nuestro estudio de la región control y de diferentes SNPs de la región codificante del ADN mitocondrial junto con SNPs del cromosoma Y de tres diferentes grupos de *Tuareg* procedentes de Mali, Burkina Faso y de la República de Niger ha revelado que dentro de su composición genética hay componentes del Oeste de Eurasia junto con el Norte de Africa.

Los datos muestran que ciertos linajes podrían no haber sido introducidos en estas poblaciones antes de hace 9.000 años mientras que las expansiones locales establecen una fecha mínima de alrededor de 3.000 años. Algunos de los haplogrupos encontrados en la población *Tuareg* estuvieron involucrados en la expansión humana post-glacial desde los refugios de la Península Ibérica hacia el resto de Europa y el norte de África. Es interesante destacar el hecho de que en la población de estudio no hay linajes mitocondriales de Oriente Próximo conectados con la expansión Neolítica.

Por otro lado los datos de los SNPs del cromosoma Y muestran que los linajes paternos pueden probablemente ser rastreados desde las expansiones Neolíticas procedentes de Oriente Próximo a través del Norte de África, un periodo que por otro lado es concordante con la expansiones mitocondriales antes mencionadas

El periodo para la migración de los *Tuareg* a través del cinturón del Sahel coincide con los cambios climáticos del Holoceno temprano a lo largo del Sáhara (desde unas condiciones óptimas de vegetación hace 10.000 años a la actual aridez que comenzó hace 6.000 años) y las migraciones de otras poblaciones africanas nómadas en el área

VIII.1.3.1.6 Artículo 6: Nuevas perspectivas de la estructura poblacional en la Cuenca del Lago Chad reveladas por medio del genotipado de alto rendimiento de SNPs codificantes del ADN mitocondrial

Se llevó a cabo el análisis de 230 SNPs de la región codificante del ADN mitocondrial en 542 muestras procedentes de 12 poblaciones de la cuenca del lago Chad por medio de la tecnología de espectrometría de masas MALDI-TOF.

Este conjunto de SNPs permite obtener una mejor resolución filogenética que los estudios previos realizados sobre esta región geográfica, permitiendo conocer mejor su historia poblacional. Se ha observado una alta heterogenidad de haplogrupos en la zona reflejando las diferentes historias demográficas de estos grupos étnicos.

Las poblaciones nómadas mostraron valores de diversidad menores que aquellos grupos sedentarios, lo cual puede sugerir que hubo una deriva génica en sus poblaciones ancestrales. Sin embargo, algunas poblaciones nómadas retuvieron mayor diversidad haplotípica en su segmento hipervariable I (HVS-I), pero no en los SNPs de la región codificante (mtSNPs), lo cual puede indicar una etnogénesis más ancestral.

Mientras que las poblaciones del norte (Fulani nómadas) muestran una mayor influencia mediterránea, lo cual viene representado principalmente por los sublinajes M1, R0, U6, y U5; las poblaciones del Sur muestran un patrón más consistente con las influencias Sub-Saharianas. Aunque el estilo de vida puede haber influido en los patrones de diversidad y en la composición de haplogrupos, el análisis de varianza molecular (AMOVA) no detecta esas diferencias. El presente estudio indica que el análisis de SNPs a una alta resolución puede ser una rápida y extensa aproximación para el cribado de la variación en estudios poblacionales donde técnicas que requieren mucho trabajo como la secuenciación de todo el genoma mitocondrial continúan siendo impracticables.

VIII.1.3.1.7 Artículo 7 (en preparación): El meta-análisis de la variación del ADNmt Africano proporciona nuevas pistas sobre la demografía continental en el pasado y los patrones del tráfico de esclavos Trans-Atlántico

Se llevó a cabo el análisis de 230 SNPs de la región codificante del ADN mitocondrial en 2426 muestras de diferentes regiones, localizadas tanto en África como en América, por medio de la tecnología de espectrometría de masas MALDI-TOF. Se genotiparon por secuenciación tres SNPs correspondientes a nuevas ramas de la filogenia que fueron apareciendo después del primer diseño del ensayo.

Además se llevó a cabo un meta-análisis de los datos existentes en la literatura correspondientes a la región control con el objetivo de inferir parámetros demográficos y de diversidad a lo largo de ambos continentes.

A la luz de los datos obtenidos, parece ser que la zona centro y oeste de África pudo jugar un papel más importante que el que se le ha asignado hasta ahora en cuanto a lo que se ha venido conociendo como cuna de la humanidad. De la misma forma que esta zona sigue siendo la que pudo contribuir en mayor proporción a la variabilidad de linajes de ADNmt sub-saharianos existente en el continente americano debido al proceso del tráfico de esclavos. El presente estudio muestra además la necesidad ampliar el número de genomas completos para poder asignar un origen geográfico concreto en África a diferentes muestras Afro-Americanas.

VIII.1.3.1.8 Artículo 8 (en preparación): Reconstruyendo con el ADNmt la relación de la dispersión Africana en Europa.

La mayor parte de los linajes de ADNmt presentes en África pertenecen al macro-haplogrupo L. En Europa estos linajes representan menos del 2% y previamente se ha sugerido que con mayor probabilidad llegaron a Europa durante el periodo del tráfico de esclavos. Esta hipótesis se basó en que los datos de la región control no mostraban señales de evolución dentro del continente europeo. Un análisis a una mayor resolución podría sin embargo revelar diferentes patrones en los linajes africanos presentes en Europa y podría permitir establecer la datación de diferentes linajes específicos.

Se llevó a cabo el análisis de genomas completos de 69 individuos europeos y 2 africanos pertenecientes a diferentes linajes del macro-haplogrupo L. Dicho análisis refleja la existencia de diferentes linajes que pueden haber evolucionado dentro de Europa. El sub-haplogrupo L1b es el más frecuente de los encontrados dentro de Europa y tiene su mayor frecuencia mundial en el Centro y en el Oeste de África, por ello se llevó a cabo un análisis completo de su filogenia con los datos del presente trabajo y los que están presentes en la literatura.

Los linajes africanos encontrados en Europa parecen indicar que el África Sub-sahariana y Europa han mantenido diferentes contactos esporádicos desde hace al menos 15000 años, pero que la mayor parte de los linajes africanos probablemente son debidos a movimientos mucho más recientes, como pueden ser el proceso de Romanización de Europa, la conquista de la Península Ibérica por parte de los árabes o el tráfico de esclavos en época colonial.

VIII.1.3.2 GENÉTICA FORENSE

VIII.1.3.2.1 Artículo 9: Ejercicio de colaboración en ADN mitocondrial GEP-ISFG 2006: reflexiones sobre interpretación, artefactos y mezclas de ADN Forensic Science International: Genetics

En este artículo, se presentan los resultados de la séptima edición de los ejercicios de colaboración de ADN mitocondrial del grupo español y portugués (GEP) de la Sociedad Internacional de Ciencias Forenses (ISFG).

Las muestras enviadas a los diferentes laboratorios consistían en manchas de sangre para un caso de maternidad y muestras forenses simuladas, incluyendo un caso de mezcla. La tasa de éxito para las manchas de sangre fue moderado (77%) aunque un tercio de los errores fue remitido por 4 laboratorios inexpertos. Un éxito similar fue el que se alcanzó en el caso del análisis de las muestras mezcladas (78,8% para la mezcla de pelo y saliva y un 69,2% en el caso de la mezcla de saliva y saliva).

La mayor parte de los errores fueron debidos a problemas en la lectura y a errores en la interpretación de los electroferogramas, demostrando una vez más que la ausencia de una sólida aproximación experimental es la principal causa de error en las pruebas de ADN mitocondrial.

VIII.1.3.2.2 Article 10: Case Report: Identificación de restos óseos usando el análisis de marcadores de fragmentos cortos de un fémur calcinado y descompuesto Forensic Science International: Genetics

Se aplicaron dos protocolos de extracción para obtener ADN a partir de un fémur quemado, que fue recuperado después de un incendio forestal. Se utilizó una batería de marcadores forenses que habían sido recientemente desarrollados que incluían mini-STRs y SNPs para genotipar la muestra y confirmar la identificación al compararlos con una hija del fallecido.

La identificación de los restos sugirió que el individuo había muerto hacía 10 años y por ello probablemente el ADN estaría degradado debido a los efectos de descomposición y exposición a temperaturas elevadas.

Se usaron una serie de nuevos marcadores desarrollados específicamente para analizar ADN altamente degradado, que incluían fragmentos de amplificación de tamaño reducido tanto para una batería de STRs como para multiplexes de SNPs autosómicos. Además se utilizó el ADN mitocondrial que es un marcador que presenta mejores resultados en estas condiciones de degradación. Esto proporcionó la oportunidad de asignar la efectividad de cada marcador a la hora de obtener resultados en el caso de muestras altamente degradadas.

Los resultados además exhibieron que la modificación de un protocolo de extracción de ADN antiguo ofrecía mejoras en cuanto al éxito de obtención de resultados en el caso de material de huesos degradados.

VIII.1.3.2.3 Artículo 11: ADN desafiante: Evaluación de diferentes aproximaciones de genotipado para muestras altamente degradadas. Forensic Science International: Genetics Supplement Series

Es frecuente encontrar en el caso de investigaciones forenses muestras con ADN altamente degradado a partir de una amplia variedad de procedencias. En esta categoría las muestras de huesos y dientes son frecuentemente el principal recurso para obtener ADN para investigaciones criminales o identificaciones de individuos fallecidos hace tiempo. En estas circunstancias los STRs estándar son propensos a fallar debido a que sus fragmentos de amplificación son grandes (desde que comienza la degradación el ADN es fragmentado progresivamente). Para resolver con éxito dichos casos, pueden utilizarse diferentes marcadores, aunque hasta hace poco el ADN mitocondrial era la única herramienta disponible, a pesar de que aun siendo más resistente es menos informativo.

Una aproximación metodológica rápidamente desarrollada para analizar ADN degradado es el genotipado a partir de fragmentos de amplificación cortos, basados en marcadores como los mini-STRs y los SNPs autosómicos. Se llevó a cabo un análisis de diferentes casos con ADN degradado de manera natural usando STRs ya establecidos junto con mini-STRs, SNPs autosómicos y ADN mitocondrial. El objetivo principal fue establecer las ventajas y desventajas de cada marcador a la hora de ayudar a los profesionales a elegir el método de análisis de ADN más adecuado dependiendo de las circunstancias de cada caso.

VIII.1.3.2.4 Artículo 12: Examinando el rendimiento de la minisecuenciacion de SNPs mitocondriales en muestras forenses. Forensic Science International Genetics

Se ha dado entre los genetistas forenses un creciente interés en el desarrollo de protocolos eficientes para el genotipado de SNPs de la región codificante del ADN mitocondrial (mtSNPs). La minisecuenciación se ha convertido en un método popular para el genotipado de SNPS, pero aún es utilizada por pocos laboratorios forenses. Esto es debido en parte la ausencia de estudios que examinen su eficiencia y reproducibilidad cuando se aplican a muestras forenses reales y complejas.

Se examinó un diseño de minisecuenciación que consistía en 71 mtSNPs (en tres multiplexes) que son diagnósticos de ramas conocidas de la filogenia de R0 en muestras forenses reales que incluían huesos degradados, dientes y pelos así como sus diluciones seriadas. El hecho de que los fragmentos de amplificación fuesen cortos junto con la eficiencia natural de la minisecuenciación permitió que se obtuviesen resultados en todas las muestras examinadas que presentaban poca cantidad de ADN y/o ADN degradado.

No se observaron inconsistencias filogenéticas en los haplotipos generados con los 71 mtSNPs, indicando la robustez de la técnica frente a potenciales artefactos que podrían provenir de contaminaciones y/o amplificaciones espúreas de pseudogenes mitocondriales que se han insertado en el genoma nuclear (NUMTs).

VIII.1.3.3 GENÉTICA CLÍNICA

VIII.1.3.3.1 Artículo 13: Alta estabilidad del ADN mitochondrial de células B en laleucemia linfocítica crónica. PLoS One

La leucemia linfocítica crónica (LLC) conduce a la acumulación progresiva de linfocitos en sangre, médula espinal y tejidos linfáticos. Estudios previos han sugerido que el ADN mitocondrial podría jugar un papel importante en la LLC.

Se llevó a cabo la secuenciación de la región control del ADN mitocondrial en 146 pacientes diagnosticados de LLC (todos procedentes del Pais Vasco) de los extractos celulares de linfocitos y se compararon con sus homólogos de granulocitos. Los mayores esfuerzos se centraron en excluir artefactos metodológicos que podrían hacer incrementar la tasa de falsos positivos para inestabilidades mitocondriales y así conducir al establecimiento de interpretaciones erróneas de las inestabilidades mitocondriales.

Únicamente se confirmaron 20 inestabilidades, la mayor parte de ellas en la zona del tracto homopolimérico de la región hipervariable II (HVII) alrededor de la posición 310. Esta posición es un punto caliente de polimorfismo de longitud y las demás son frecuentemente observadas en la población general.

Al realizar una revisión crítica de los resultados encontrados en estudios previos, se pudo encontrar una falta general de los estándares metodológicos necesarios, los cuales pueden conducir a interpretaciones erróneas del papel del ADNmt en el origen tumoral de la LLC.

VIII.2 PRIMER TABLES

**The rest of sequence primers as well as additional information concerning to other genotyping protocols and the tables with genotyping results are online for each one of the publications

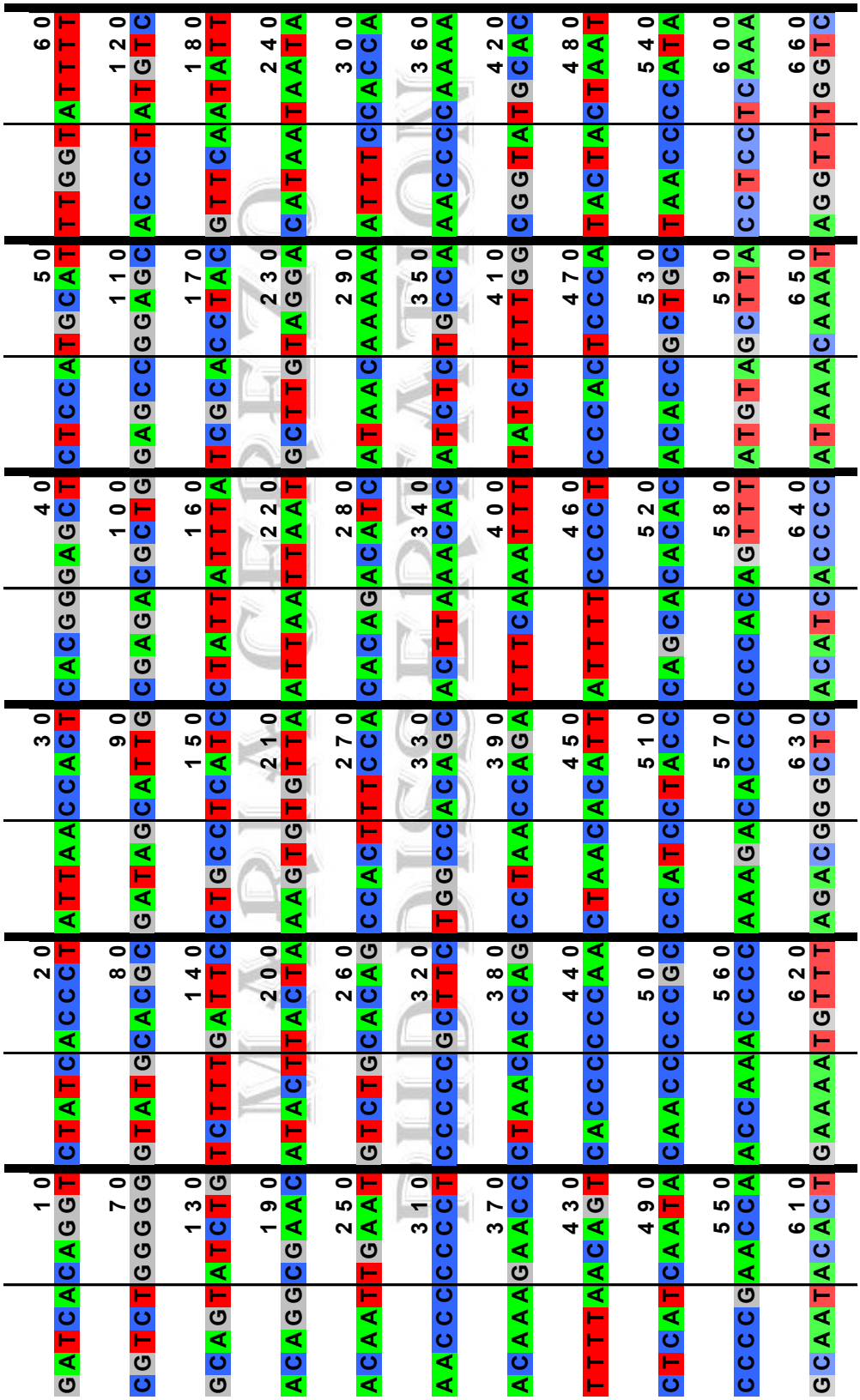
Appendix Table 1: List of primers used to carry out the sequencing of the control region

	PRIMER	SEQUENCE
HVI (16024-16569)	15997L	CACCATTAGCACCCAAAGCT
	16121L	TACTGCCAGCCACCATGAAT
	16157H	ACTACAGGTGGTCAAGTATTTATGGT
	16159L	TACTTGACCACCTGTAGTAC
	16183H	TTTTGATGTGGATTGGGTTTT
	16213L	CATGCTTACAAGCAAGTACAG
	16236H	CTTTGGAGTTGCAGTTGATG
	16254L	CACATCAACTGCAACTCCAAA
	16281H	TTGGTATCCTAGTGGGTGAGG
	16365H	CACGGAGGATGGTGGTCAAG
	16380L	TCAGATAGGGGTCCCTTGAC
	16313H	CTATGTACGGTAAATGGCTTTATG
	16401H	TGATTTCACGGAGGATGGTG
	017H	CCCGTGAGTGGTTAATAGGGT
	16517H	CATCTGGTTCCTACTTCAGG
HVII (1-576)	16555L	CCCACACGTTCCCCTTAAAT
	029L	GGTCTATCACCCTATTAACCAC
	047L	CTCACGGGAGCTCTCCATGC
	048L	CTCACGGGAGCTCTCCATGC
	172L	ATTATTTATCGCACCTACGT
	285H	GGGGTTTGGTGGAAATTTTTTG
	332L	CCCGCTTCTGGCCACAGCAC
	370L	CCCTAACACCAGCCTAACCA
	408 H	CTGTTAAAAGTGCATACCGCCA
	586H	TGTATTGCTTTGAGGAGGTAAGC
	599H	TTGAGGAGGTAAGCTACATA
	611H	CAGTGTATTGCTTTGAGGAGG
	649H	TTTGTTTATGGGGTGATGTGA
	16450H	CAAGTGTTATGGGCCCGGAGC
	15971L	TTAACTCCACCATTAGCACC
	16400H	GTCAAGGGACCCCTATCTGA
	613H	TCAGTGTATTGCTTTGAGGAGGT

Appendix Table 2: List of primers used to carry out the sequencing of the complete genome. The set of primers were previously published in (Kivisild et al. 2006; Torroni et al. 2001)

PCR TORRONI	SEQ TORRONI	PCR AND SEQ KIVISILD	
1	1a 14948F		39R 15185R
14897F	1b 15564F	40F 15071F	40R 15519R
155R	1c 131R	41F 15463F	41R 16022R
2	2a 16522F		
16488F	2b 584F	1F 435F	1R 901R
1677R	2c 1060F	2F 888F	2R 1380R
3	3a 1445F	3F 1366F	3R 1845R
1404F	3b 2047F	4F 1838F	4R 2317R
	3c 2509F	5F 2262F	5R 2713R
3235R		6F 2686F	6R 3125R
4	4a 3085F	7F 3119F	3R 3477R
2900F	4b 3598F	8F 3460F	8R 3867R
	4c 4010F	9F 3873F	9R 4268R
4683R		10F 4244F	10R 4658R
5	5a 4410F	11F 4631F	11R 5064R
4381F	5b 5014F	12F 5055F	12R 5439R
	5c 5544F	13F 5419F	13R 5841R
6151R		14F 5825F	
6	6a 6041F		14R 6270R
5871F	6b 6600F	15F 6258F	15R 6657R
		16F 6495F	16R 6928R
7617R		17F 6829F	17R 7276R
7	7a 7336F	18F 7248F	18R 7678R
7239F	7b 7937F	19F 7619F	19R 8082R
	7c 8459F	20F 7956F	20R 8371R
9218R		21F 8355F	21R 8746R
8	8a 8975F	22F 8729F	22R 9208R
8910F	8b 9589F	23F 9123F	23R 9605R
	8c 10147F	24F 9329F	24R 9834R
		25F 9770F	25R 10259R
10649R		26F 10193F	26R 10640R
9	9a 10498F	27F 10622F	27R 11028R
10457F	9b 11081F	28F 11014F	28R 11520R
	9c 11644F	29F 11475F	29R 11965R
12225R		30F 11970F	
10	10a 12114F		30R 12487R
12014F	10b 12600F	31F 12414F	31R 12818R
	10c 13134F	32F 12806F	32R 13250R
13829R		33F 13139F	33R 13603R
11	11a 13568F	34F 13530F	34R 13715R
13477F	11b 14103F	35F 13648F	35R 13936R
	11c 14603F	36F 13855F	36R 14010R
		37F 13959F	37R 14352R
		38F 14311F	38R 14789R
15349R		39F 14723F	

HVI
15970 15980 15990 16000 16010
16030 16040 16050 16060 16070
16090 16100 16110 16120 16130
16150 16160 16170 16180 16190
16210 16220 16230 16240 16250
16270 16280 16290 16300 16310
16330 16340 16350 16360 16370
16390 16400 16410 16420 16430
16450 16460 16470 16480 16490
16510 16520 16530 16540 16550
16569 10 20 30 40
16020 16080 16140 16200 16260 16320 16380 16440 16500 16560


HVII
10 20 30 40 50 60
70 80 90 100 110 120
130 140 150 160 170 180
190 200 210 220 230 240
250 260 270 280 290 300
310 320 330 340 350 360
370 380 390 400 410 420
430 440 450 460 470 480
490 500 510 520 530 540
550 560 570 580 590 600
610 620 630 640 650 660