



Departament de Ciències

Experimentals i de la Salut

Universitat Pompeu Fabra (UPF)

***Anàlisi de la diversitat del genoma
mitocondrial en poblacions humanes***

Memòria presentada per Stéphanie Plaza per optar al grau de doctora en Ciències Biològiques. Aquesta tesi ha estat realitzada sota la direcció del Dr. David Comas Martínez, a la Unitat de Biologia Evolutiva del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra, dins del programa de doctorat en *Ciències de la Salut i de la Vida (bienni 1999-2001)*.

DAVID COMAS MARTÍNEZ

STÉPHANIE PLAZA

Barcelona, gener del 2004

Dipòsit legal: B.32495-2004
ISBN: 84-688-7546-5

A mes parents, à Dédé, a Max.

*“Croire est une affaire de religion,
pour le reste on vérifie”*

Je ne me souviens plus de l'auteur de cette phrase entendue entre deux conférences, mais elle résume bien la situation.

Remerciements/Agraïments/Agradecimientos

Encara que no ho sembli, aquesta part no es el més fàcil. Són tantes persones que han participat directament o indirectament a la realització de la present tesi, que no se sap per on comença i que sempre ens deixem algú.

M'agradaria començar per agrair el Dr. Jaume Bertranpetit per haver-me donat la oportunitat de portar a terme aquest treball de recerca en el seu laboratori.

Un gran MERCI (en els dos idiomes) al meu director de Tesi, el Dr. David Comas, per la seva gran paciència i ajuda durant aquests quatre anys. T'agraeixo el teu interès, i suport constants al llarg de la present Tesi, els teus consells, la teva disponibilitat, les teves qualitats científiques, i grans qualitats humanes. Vaig haver de passar moltes vegades per "EL DESPATX", però en he après molt; i això es lo més important.

Je remercie spécialement le Dr Gérard Lefranc qui fut l'initiateur de cette aventure de quatre ans, sans qui rien de tout cela n'aurait été possible.

També voldria presenta un agraïment sincer als meus companys de laboratori, als que ja no hi són: la Eva, la meva ex veïna del barri de les eixamples, i que hauria pogut ser la meva mare africana, al Jordi Clarimon, tambien ex vecino, el craket del laboratorio, cantante, i maltratador de ordenador, por los buenos i malos momentos porque siempre se aprendre, als quatres magnífics de la "sala de pensar" per la seva gran ajuda: el Francesc, per la seva gran eficiència, l'Arcadi pels ànims, i l'atenció, i perquè em fa riure molt amb els seus comentaris masclistes, al Tomas per solucionar-me els problemes amb el Word, i altres programes, i un agradecimiento especial a

Oscar: gracias por tus mapas, tu ayuda con el ordenador, los programas, les charlas en los momentos de “depre”, i por ser tan buena persona; a la Monica, per facilitar-nos el dia a dia; a la Marta per el seu suport, i al Carles (que en aprendrà tant amb la Marta!!!), a l'Elena que em va ajuda molt quan vaig entrar al laboratori, a l'Anna G sobre todo al principio de su incorporación, al Josep, i también un gran merci a Aida con quien empecé, i que siempre me ofreció su ayuda i su apoyo moral en todos momentos, a los recientes incorporados, a Michi por su buen humor por las mañanas, a Lourdes por su apoyo, a Andres, siempre atento i dispuesto a ayudar, à Gemma, pour son intérêt et son grand enthousiasme.

Gracias a Toño, tambien conocido com el tío genial del laboratorio de Santiago de Compostela por su colaboración i su gran ayuda.

Al servei de seqüenciació a l'Anna P, al Roger, i a l'Anna petita per totes aquestes seqüències.

A ma famille, et en particulier à mes parents: les mots me manquent pour vous décrire ma reconnaissance, merci pour votre soutien, pour être à mes côtés à chaque instant, pour votre patience, et tous ces détails de la vie quotidienne.

A mon frère, à Dédé, parce que s'il est vrai qu'on ne choisit pas sa famille, il est plus que certain que je ne désire pour rien au monde un autre frère que toi. Merci pour ces moments passés ensemble, pour ton aide, pour ton soutien, bref, pour tout.

A mes amis, en particulier à Berny et Joaquim: c'est en grand part grâce à toi Berny que j'y suis arrivé !!!!! Merci pour ton aide précieuse à une époque où les choses n'étaient pas du tout faciles, merci à vous deux parce que j'ai toujours pu compter sur

vous, merci pour votre amitié; petit Simon a bien de la chance de vous avoir comme parents!!!

Als meus amics: a la Montse i la Carme per su amistad i su apoyo, al Miki perquè ser que sempre puc comptar amb tu, passi el que passi, a la Susi i a l'Agusti, per la vostre amistad, els vostres consells, merci a la Susi aguantar a la creuota de l'Agusti que aguantar la creueta de la Stéphanie; merci Agus per fer menys pesats aquests trajectes en aquestes hores tant intempestives al matí, per els croissanets calentets de la 7h00 que són com les magdalenes de Proust, un record que el temps mai podrà esborrar

I como lo mejor siempre se guarda para el final, a mi novio, Max: gracias por estar a mi lado, por tu apoyo día tras día, por tu paciencia, por tu comprensión , por dar-me fuerzas, por que contigo todo vale la pena, por todos estos momentos de felicidad, i todas estas cosas que solo pertenecen a nosotros dos.

Aquest treball l'he dut a terme mentre gaudia d'una beca de formació de personal investigador de la Generalitat de Catalunya (FI00696), de gener 2000 fins desembre 2003. A més el treball ha comptat amb les subvencions de la Dirección General de Investigación Científica y Técnica (PB98-1064 i BOS2001-0794), del Comissionat per a Universitats, de Recerca i Societat de la informació, i de la Generalitat de Catalunya com Grup de Recerca Consolidat (2001SGR00285).

~ INTRODUCCIÓ ~

ÍNDEX

INTRODUCCIÓ	1
I. Diversitat genètica humana	3
I.1. La mutació	4
I.2. La selecció	5
I.3. La migració	6
I.4. La deriva genètica	7
II. Els marcadors genètics emprats per l'anàlisi de les poblacions humanes	8
II.1. Els polimorfismes d'insercions <i>Alu</i>	9
II.2. Els microsatèl·lits o STRs (Short Tandem Repeat polymorphisms)	10
II.3. SNPs (Single Nucleotid Polymorphisms)	12
II.4. El mtDNA i el cromosoma Y	12
III. EL MITOCONDRI	14
III.1. Estructura	14
III.2. Semi-autonomia genètica	15
IV. El DNA mitocondrial (mtDNA)	17
IV.1. La regió codificant	17
IV.2. La regió no codificant	18
IV.3. Característiques del mtDNA	19
IV.3.a. Herència materna	19
IV.3.b. Taxa de mutació	21
IV.3.c. Homoplàsmia-Replicació segregativa-Heteroplàsmia	22
IV.3.d. Alt nombre de còpies per cèl·lula	23

V. El DNA mitocondrial: una eina genètica per aclarir la història de les poblacions humanes.	24
V.1. L'origen dels humans moderns	24
<i>V.1.a. La continuïtat multiregional</i>	25
<i>V.1.b. El model de l'Origen Recent Africà (ORA)</i>	26
V.2 L'avantpassat mitocondrial	28
V.3 Distribució dels llinatges mitocondrials i de les poblacions humanes.	34
<i>V.3a. Els llinatges mitocondrials a l'Àfrica</i>	34
<i>V.3b. Els llinatges mitocondrials a Europa</i>	35
<i>V.3b. Els llinatges mitocondrials a Àsia</i>	35
<i>V.3d. Els llinatges mitocondrials a les Amèriques</i>	36
V.4. Altres aplicacions del mtDNA	37
VI. Genealogia del gen	38
MATERIAL I MÈTODES	43
I. Polimorfismes del mtDNA	45
I.1. SNPs (Single Nucleotide polymorphisms)	45
<i>I.1.a. Regió de control</i>	45
<i>I.1.b Regió codificant</i>	45
I.2. Microsatèl·lit del mtDNA	46
I.3. Deleció de 9 parells de bases	46
II. Tècniques utilitzades	47
II.2. Purificació de les mostres	48
II.3 Seqüenciació automàtica	48
II.4. SNaPShot	49
II.5. Microsatèl·lit i deleció de 9 parell de bases	52

III. Poblacions estudiades	53
III.1. Les poblacions de l'oest del Mediterrani	54
<i>III.1.a. Els berbers de Tunísia</i>	57
III.2. La població d'Angola	59
III.3. Les poblacions d'Àsia Central	62
III.4. La població de l'illa de la Reunió	64
IV. Tractament estadístic	66
IV.1. Anàlisi molecular de la variància: AMOVA	66
IV.2. Anàlisi espacial de la variància molecular: SAMOVA	66
IV.3. Anàlisi de components principals	67
IV.4. Anàlisi de coordenades principals	67
IV.5. Anàlisi de correspondències.	68
IV.6. Anàlisi de <i>multidimensional scaling</i> o MDS	68
IV.7. Els <i>networks</i> filogenètics	68
IV.8. Anàlisi d'<i>admixture</i>	69
RESULTATS	71
CAPÍTOL I: <i>Genética e historia de las poblaciones del Norte de África y de la península Ibérica</i>	73
CAPÍTOL II: <i>Joining the Pillars of Hercules: mtDNA Sequences Show Multiregional Gene Flow in the Western Mediterranean</i>	83
CAPÍTOL III: <i>Mitochondrial DNA heterogeneity in Tunisian Berbers</i>	103
CAPÍTOL IV: <i>Insights into the western Bantu dispersal: mtDNA lineages analysis in Angola</i>	139

CAPÍTOL V: <i>Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages</i>	175
CAPÍTOL VI: <i>Admixture and sexual bias in the population settlement of La Réunion Island (Indic Ocean)</i>	187
DISCUSSIÓ	215
BIBLIOGRAFÍA	235

~ INTRODUCCIÓ ~

I. La diversitat genètica humana.

Salta a la vista que els humans som diferents els uns dels altres. Si agafem dos individus a l'atzar d'una mateixa ciutat, o de la mateixa família, trobarem diferències, que es poden atribuir, algunes, a fets biològics i, altres, a modificacions voluntàries (colors de cabells, tatuatges, piercings, lents de colors...). Tot i això, és fàcil veure quina és la part biològica d'aquestes diferències, és a dir, quines són determinades pel nostre genoma. Són els mecanismes de l'herència que creen les diferències entre els individus i que determinen el grau de semblança entre pares i fills.

Però, a més de les diferències que trobem comparant els individus, també hi ha diferències a nivell de les poblacions. Per exemple, es pot predir a quin continent pertanyen individus d'algunes poblacions només a partir de les seves característiques físiques. Aquests canvis en l'aspecte físic de les poblacions humanes són el resultat de l'adaptació a les condicions ambientals de l'entorn. Els humans som una espècie relativament jove i homogènia comparada amb altres espècies, amb una baixa diversitat nucleotídica: 0,1% (Chakravarti i col·laboradors. 1999; Jorde i col·laboradors. 2001; Przewoski i col·laboradors. 2000). Hom estima que el 85% de la diversitat genètica humana es trobat dins de les poblacions, i el 15% entre les poblacions humanes, repartides en un 10% entre els grans grups continentals, i en un 5% entre les poblacions d'un mateix continent (Lewontin 1972, Livshits i Nei 1990; Relethford i Harpending 1994; Batzer i col·laboradors. 1994; Jorde i col·laboradors. 1995; Barbujani i col·laboradors. 1997; Stoneking i col·laboradors. 1997). El patró de variació de les poblacions humanes depèn de la nostra història demogràfica, però també de les forces evolutives que han modulat el nostre genoma. Aquests motors evolutius de canvi són, bàsicament: la mutació, la selecció, la migració i la deriva genètica.

I.1. La mutació.

La mutació fa néixer la variació. La mutació és un canvi casual, font d'evolució. La majoria dels estudis de seqüenciació estimen que la diversitat nucleotídica mitjana en humans és de $\sim 1/1000$ (Libert i col·laboradors. 1998; Reich i col·laboradors. 2002). Això vol dir que si agafem dos individus a l'atzar trobarem poca diferència: una de cada 1000 nucleòtids. Les mutacions que afecten el nostre genoma poden tenir diversos efectes: les mutacions desavantajoses perjudiquen l'individu; les neutres no tenen cap efecte sobre cap funció; i les avantajoses milloren el funcionament de l'individu en les condicions particulars en les què viu.

Les mutacions desavantajoses són perjudicials per l'individu que les du ja que disminueixen les seves possibilitats de supervivència i/o redueixen la seva descendència. Poc després de la seva aparició, les mutacions desavantajoses són ràpidament eliminades per la selecció purificadora o selecció negativa.

Les mutacions avantajoses, al contrari, tenen un efecte positiu els individus que les duen, ja que augmenten la probabilitat de supervivència i/o la fertilitat, i per tant tenen més probabilitat de passar a la generació següent. Aquestes mutacions són relativament rares ja que quan apareixen, són ràpidament fixades en el nostre genoma sota l'efecte de la selecció positiva.

Les mutacions neutres no tenen cap efecte sobre la eficàcia biològica de l'individu, i són fixades en el genoma per deriva genètica. Les mutacions neutres són molt més freqüents que les mutacions avantajoses. La majoria de les diferències observades entre dues seqüències entre dos espècies o dintre de la mateixa espècie són el resultat de la fixació de mutacions neutres per deriva genètica. Això implica que l'evolució està condicionada per dos mecanismes principals: la selecció natural i la fixació de mutacions neutres per deriva genètica. Degut a la dificultat de demostrar l'empremta de la selecció natural, l'estudi de les mutacions neutres fixades en el genoma és l'eina principal en el estudis d'evolució i de genètica de poblacions.

I.2. La selecció.

La selecció natural modula la variació. És l'única força evolutiva que té conseqüències adaptatives, ja que és un fenomen que manté les mutacions favorables i elimina aquelles que poden ser perjudicials per l'individu. Tant la selecció com la mutació afecten el patró de variació en punts (loci) precisos del genoma. La selecció natural garanteix la supervivència dels individus més ben adaptats a les condicions ambientals, com el clima, l'alimentació, la resistència a les malalties, entre d'altres. Es fa referència a la selecció *negativa* quan disminueix la probabilitat de transmissió d'un gen portador d'una mutació a la generació següent, i a selecció *positiva* quan aquesta probabilitat augmenta. Els humans moderns, a partir del seu origen africà, es van establir en diferents regions del món i van haver d'adaptar-se a condicions ambientals diverses. La intervenció de la selecció natural va accentuar les diferències entre grups de poblacions, afavorint l'expansió dels al·lels de manera diferencial. Un exemple paradigmàtic de selecció natural en poblacions humanes és la resistència a la malària. Els individus portadors de determinades mutacions en el gen responsable d'alguna cadena de l'hemoglobina tenen una viabilitat variable en funció de l'ambient. Aquest és el cas de la variant S de l'hemoglobina en front de la variant A que és l'estàndard. En absència de malària, els individus homozigots normals i els heterozigots tenen la mateixa possibilitat de supervivència, i els individus homozigots per la mutació no arriben a l'edat sense tractament. En presència de malària, els individus heterozigots són seleccionats positivament i tenen més possibilitat de sobreviure que els individus homozigots normals que sucumbeixen a la malària. Els individus homozigots per la mutació segueixen tenint una eficàcia biològica reduïda. S'han descrit altres mutacions i altres gens que també han estat relacionats amb la resistència a la malària. L'efecte de la selecció natural està, en tot cas, restringit a alguns gens i algunes mutacions.

I.3. La migració

La migració modula la variació. Al contrari que la selecció, la migració afecta a tot el genoma. Els intercanvis genètics entre poblacions o grups d'individus contribueixen a reduir la divergència entre elles i a deixar una empremta de les seves històries respectives en el genoma. L'espècie humana no és genèticament homogènia, ja que com qualsevol altre espècie està subdividida en moltes poblacions que ocupen territoris geogràfics diferents. Existeixen dos patrons de migració: les migracions entre pobles no gaire distants l'un de l'altre, i les migracions que afecten un grup sencer, o a vegades pocs individus que deixen el seu lloc d'origen per establir-se en un lloc molt llunyà. El primer tipus de migració limita l'aïllament entre grups i disminueix la incidència de la deriva. Quan una nova mutació apareix en una comunitat d'individus, si no es perd per deriva genètica, pot arribar a una freqüència elevada en el lloc on s'ha generat. D'aquesta manera, la migració entre pobles veïns, i la migració associada amb el matrimoni d'individus d'origen diferents expandeix la nova mutació. L'altre tipus de migració implica el desplaçament d'un grup sencer d'individus cap a nous territoris, i que amb el temps es diferencien genèticament de la població original. Aquestes migracions són freqüents en temps de caresties, de desastres naturals, de guerra, de superpoblació. Sovint, les migracions en territoris inexplorats s'acompanyen d'una expansió ràpida de la població. La sortida d'Àfrica de l'home modern, que va afectar el patró de variació genètica de la nostra espècie, és l'exemple més significatiu d'aquest tipus d'esdeveniment. En temps antics quan la distància que separava la nova població de la població original era molt gran, els contactes entre les dues poblacions s'interrompien, i la deriva genètica i l'adaptació a nous ambients creaven una diferenciació entre els dos grups que podia ser extrema amb el temps. Aquest tipus de migracions augmenten la diferenciació entre els grups.

I.4. La deriva genètica

La deriva genètica constitueix amb la migració, la mutació, i la selecció natural una força evolutiva que influencia el patró del nostre genoma. En cas de la deriva genètica, es produeix una fluctuació aleatòria de les freqüències al·lèliques d'una generació a l'altra. Els efectes de la deriva genètica són particularment patents en poblacions amb un nombre reduït d'individus. Dins d'aquest fenomen de deriva tenen especial importància els processos de coll d'ampolla i els efectes fundadors. L'efecte fundador és un cas extrem de variació genètica aleatòria, on un grup petit d'individus que forma part d'una població més gran es separa per colonitzar noves terres. Aquest grup reduït d'individus pot ser portador de variants al·lèliques que no són necessàriament representatives de la població d'origen. Per tant, un al·lel que és comú en la població d'origen pot tenir una freqüència baixa o pot desaparèixer en la població fundadora. De la mateixa manera, un al·lel poc freqüent en la població original pot augmentar en freqüència en la població fundadora. Els processos de coll d'ampolla tindrien conseqüències similars: a partir d'una població nombrosa, es redueix dràsticament el número d'individus i posteriorment es recupera la mida poblacional. Aquest període de mida poblacional reduïda pot fer que aleatòriament les freqüències al·lèliques originals canviïn dràsticament en la població resultant.

II. Els marcadors genètics emprats per l'anàlisi de les poblacions humanes:

El patró de variació de la població humana depèn tant de la nostra història demogràfica com dels factors específics que afectin a la regió genòmica sota estudi. A partir de l'estudi d'aquests patrons de variació en les regions codificants i no codificants del genoma, es poden inferir els esdeveniments demogràfics i l'impacte de la selecció que han modulats la variació en el nostre genoma. Amb la recent publicació del genoma humà i la seva anotació s'està desenvolupant una nova àrea, la genòmica poblacional, per tal d'entendre com la història demogràfica de les poblacions i l'impacte de la selecció han modulats la variació en el nostre genoma. Entendre aquesta variació és important per entendre el que ens fa únics.

Existeix una gran varietat de marcadors genètics per tal de descriure la variació a nivell genòmic i determinar quina part de la variació genètica s'explica per la història demogràfica de les poblacions i quina està associada amb la variació fenotípica. Des dels estudis sobre el polimorfisme dels grups sanguinis i de les proteïnes (Lewontin 1972; Livshits i Nei 1990; Relethford i Harpending 1994), el desenvolupament del camp de la genètica de poblacions va integrar ràpidament diversos sistemes genètics com el mtDNA (Cann i col·laboradors. 1987; Vigilant i col·laboradors. 1991; Stoneking i Soodyall 1996; Ingman i col·laboradors. 2000; Bamshad i col·laboradors. 2001), el cromosoma Y (Hammer i col·laboradors. 1998; Seielstad i col·laboradors. 1999; Forster i col·laboradors. 2000; Bosch i col·laboradors. 2001), els microsatèl·lits autosòmics (Di Rienzo i col·laboradors. 1994; Deka i col·laboradors. 1995; Goldstein i col·laboradors. 1995; Jorde i col·laboradors. 1997; Perez-Lezaun i col·laboradors.

1997; Calafell i col·laboradors. 1998) i els SNPs (Stephens i col·laboradors. 2001; Gabriel i col·laboradors. 2002; Marth i col·laboradors. 2003) per entendre la variació genètica humana a nivell de les poblacions.

II.1. Els polimorfismes d'insercions *Alu*

Les insercions *Alu* són elements d'aproximadament 300 parells de bases, que es troben distribuïdes en el genoma de primats. Amb aproximadament un milió de còpies en el genoma humà (International Human Genome Consortium 2001; Batzer i Deininger 2002), els elements *Alu* són la classe més abundants de *SINEs* (Short Interspersed Nuclear Elements) del nostre genoma. Dintre de la família *Alu*, hom estima que uns 5000 elements *Alu* formen una subfamília recent que es va integrar en el genoma humà en el moment de la separació entre la nostra espècie i els grans simis fa entre 4 i 6 milions d'anys. Un subgrup d'elements *Alu* (~1200) es va integrar recentment en el genoma humà per retransposició, de tal manera que la seva inserció en lloc específic del genoma és polimòrfica entre cromosomes (Batzer i Deininger 2002). Els elements *Alu* es propaguen en els cromosomes també per retransposició a partir de una seqüència diana (Weiner i col·laboradors. 1986; Luan i col·laboradors. 1993; Jurka i col·laboradors. 1997; Esnault i col·laboradors. 2000; Kajikawa i Okada 2002). El polimorfisme d'insercions *Alu* és un marcador molt útil per reconstruir la història demogràfica de les poblacions humanes i per detectar processos de migracions i de diferenciació poblacional (Hammer i col·laboradors. 1994; Batzer i col·laboradors. 1994; Stoneking i col·laboradors. 1997; Jorde i col·laboradors. 2000; Bamshad i col·laboradors. 2001; Nasidze i col·laboradors. 2001) degut a la seves propietats: és selectivament neutre i el seu estat ancestral es conegut. De totes maneres, el baix grau de polimorfisme degut a que únicament hi ha dos al·lels (presència - absència de la

inserció) i la impossibilitat de reconstruir filogènies, limiten el seu ús en l'anàlisi de poblacions humanes.

II.2. Els microsatèl·lits o STRs (Short Tandem Repeat polymorphisms)

Els STRs són seqüències repetides en tàndem de 2 a 6 parells de bases (figura 1)

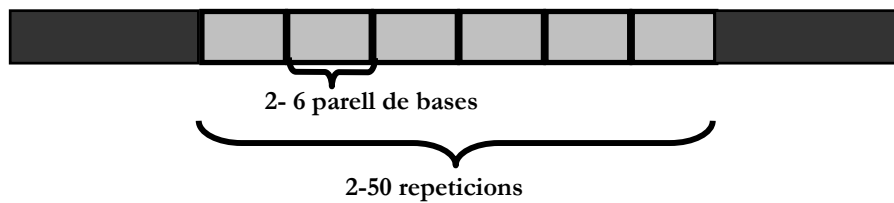


Figura 1: Estructura bàsica d'un microsatèl·lit.

Hom atribueix a un fenomen de *slippage* durant la replicació del DNA l'augment i la disminució del nombre de repeticions en una o poques unitats de l'estructura bàsica dels microsatèl·lits (Levinson i Gutman 1987; Weber 1990; Schlötterer i Tautz 1992). Aquest patró de mutació s'anomena model de mutació *stepwise* generalitzat (Di Rienzo i col·laboradors. 1994). La taxa de mutació dels microsatèl·lits és intrínseca de locus i depèn de factors com el nombre de repeticions (Weber 1990; Goldstein i Clark 1995; Brinkman i col·laboradors. 1998), qualitat de la seqüència repetida (Estoup 1995), sexe i edat dels portadors (Henke i Henke 1999), i de la longitud de la unitat repetida (Chakraborty i col·laboradors. 1997). D'una manera general, hom estima la seva taxa de mutació de l'ordre de 10^{-3} - 10^{-4} per locus, gàmeta i generació. Els STRs són selectivament neutres, altament variables i freqüents en tot el genoma (cada 30-50 Kb). Degut a les seves propietats, l'anàlisi de la variació dels microsatèl·lits és apropiat per

estudiar les relacions filogenètiques entre les poblacions humanes i inferir la història de les poblacions (Di Rienzo i col·laboradors. 1994; Dekker i col·laboradors. 1995; Goldstein i col·laboradors. 1995; Jorde i col·laboradors. 1997; Perez-Lezaun i col·laboradors. 1997; Calafell i col·laboradors. 1998). Els STRs també s'han revelat com una eina molt útil en el camp de la genètica forense tant pels tests de paternitat com per a la identificació individual (Hammond i col·laboradors. 1994; Urquhart i col·laboradors. 1994; Blouin i col·laboradors. 1996). Finalment, també s'han utilitzat àmpliament en estudis de lligament per situar la posició relativa de gens associats a malalties (Edwards i col·laboradors. 1991; Dib i col·laboradors. 1996).

II.3. SNPs (Single Nucleotid Polymorphisms)

Un *SNP* es defineix com la presència de dues possibles bases en una posició particular en el DNA. Hom distingeix un *SNP* d'una variant al·lèlica rara per la seva freqüència d'aparició: els SNPs estan estesos en tot el genoma humà amb una freqüència superior al 1%. Des del projecte genoma humà, l'interès pels *SNPs* va augmentant i està lligat a diverses àrees de recerca com l'anàlisi del genoma a gran escala, la bioinformàtica i el biocomputing; l'anàlisi genètic de malalties complexes, i la genètica de poblacions mundial humana. Els *SNPs* s'han revelat particularment informatius quan s'analitzen com haplotips, és a dir, com una combinació d'al·lels associats en una regió determinada d'un cromosoma, perquè es poden caracteritzar tant la diversitat haplotípica, com la recombinació i el desequilibri de lligament. Per aquestes raons s'estan desenvolupant grans projectes d'abast internacional com el projecte HapMap per tal de definir al llarg del genoma humà haplotips de *SNPs* que agrupin tota la informació útil del genoma.

II.4. El mtDNA i el cromosoma Y

El mtDNA i el cromosoma Y són dos marcadors complementaris ja que el primer caracteritza els llinatges femenins i el segon els llinatges masculins de les poblacions humanes. En aquest apartat, ens centrarem en les característiques del cromosoma Y, deixant de banda el mtDNA que serà l'objecte del capítol següent.

El cromosoma Y amb ~ 60 Mb és el cromosoma més petit del genoma humà. Es caracteritza per una regió d'heterocromatina a la part distal del braç llarg, de longitud variable entre individus, i d'una part d'eucromatina de 30 Mb que conté les regions de major interès genètic. El cromosoma Y conté el gen *SRY* (*Sex determining region*) que determina la masculinitat, i té un mode de transmissió uniparental, per via

paterna. La major part del seu genoma és no recombinant, a l'excepció de les regions pseudoautosòmiques, PAR1 i PAR2, situades a l'extrem de cada braç del cromosoma.

Les característiques del cromosoma Y (transmissió per via paterna i absència de recombinació) han permès identificar nombrosos polimorfismes (substitucions de bases, indels, microsatèl·lits, minisatèl·lits) per tal de caracteritzar els llinatges masculins en les poblacions i la seva distribució geogràfica. Els llinatges del cromosoma Y definits a partir de SNPs (Hammer i col·laboradors. 1998, 2001; Underhill col·laboradors. 2001; Bosch col·laboradors. 2001), presenten una genealogia molt ben definida i una distribució geogràfica restringida. La nomenclatura utilitzada va ser establerta pel Consorci del Cromosoma Y (The Y Chromosome Consortium 2002).

La regió no recombinant del cromosoma Y i el genoma mitocondrial es comporten com a loci únics. La mida efectiva del mtDNA i del cromosoma Y correspon a un quart de la dels cromosomes nuclears, i fa que siguin més sensibles als processos de deriva genètica i de selecció (si estan associats a malalties) que poden alternar els patrons de variació. La genealogia del mtDNA i del cromosoma Y són bàsicament similars però presenten diferències notables degudes a diferències socials que generen diferències en els patrons de migracions (Seielstad i col·laboradors. 1998). Per exemple, el predomini de la societat patriarcal explicaria les diferències que s'han pogut trobar en nombroses poblacions (Salem i col·laboradors. 1996; Oota i col·laboradors. 2001). La comparació dels patrons de variació entre aquests dos marcadors és adequada per estudiar les diferències en el patrons de migració entre les dones i els homes (Seielstad i col·laboradors. 1998). Les divergències entre les dades del mtDNA i del cromosoma Y també s'expliquen per diferències demogràfiques, com per exemple una mortalitat més gran en els homes que en les dones, i una freqüència més elevada de la poligàmia que de la poliàndria.

III. EL MITOCONDRI

El mitocondri és un petit òrgan cel·lular ($\sim 1 \mu\text{m}$ de diàmetre) present al citoplasma. Es troba en totes les cèl·lules dels organismes superiors que utilitzen l'oxigen com a font d'energia i són l'indret de nombrosos processos metabòlics crucials, com la fosforilació oxidativa. Per aquesta raó, es refereix al mitocondri com a la central elèctrica de la cèl·lula. La seva longitud i forma varien amb el tipus i l'activitat cel·lular. Hom estima que el nombre de mitocondris per cèl·lula varia entre 200 i 1700 depenent del tipus de teixit (Bogenhagen i col·laboradors. 1980; Robin i Wong 1988).

III.1. Estructura:

L'ultraestructura dels mitocondris es revela per microscopia electrònica ($0.5 \mu\text{m} - 10 \mu\text{m}$) (Figura 2a). Es compon de dues membranes: una externa llisa; i una d'interna, formada de crestes que segmenten el contingut mitocondrial, o matriu (Figura 2b). La composició química de les dues membranes és molt diferent. La membrana externa és molt semblant a la resta de les biomembranes, mentre que la membrana interna es compon de tres grups de proteïnes: transportadors transmembrana, els constituents de la cadena respiratòria, i un complex enzimàtic, l'ATP sintetasa mitocondrial. La matriu conté nombroses molècules iòniques i enzimàtiques, però també ribosomes, molècules de RNA i de DNA.

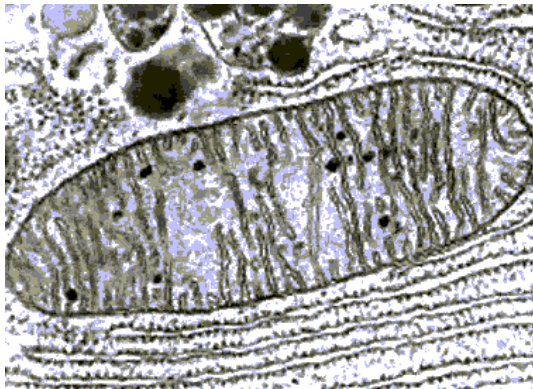


Figura 2a: micrografia electrònica d'un mitocondri (figura extreta de *Fancett A textbook of Histology, 1994*)

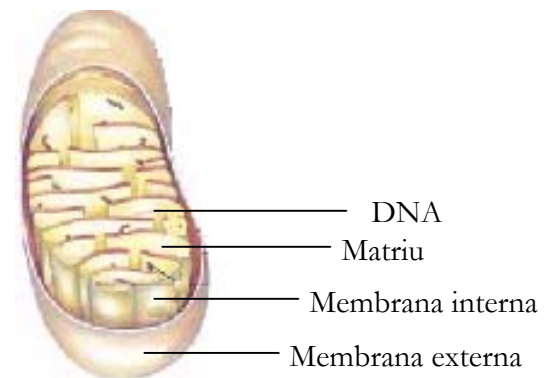


Figura 2b: esquema d'un mitocondri. (figura modificada a partir del llibre *Lehninger: Principles of Biochemistry 2000*)

III.2. Semi-autonomia genètica

Segons la teoria endosimbiòtica (Margulis, i col·laboradors. 1990), el mitocondri prové d'un bacteri que fa més de mil milions d'anys es va introduir a les cèl·lules, com a simbiota. L'avantpassat bacterià va perdre la capacitat de funcionar com un organisme independent, perdent la capacitat de créixer i duplicar-se tot sol, de tal manera que la majoria de les proteïnes funcionals del mitocondri estan codificades pels gens del nucli. La cèl·lula no pot prescindir dels mitocondris, i els mitocondris no poden prescindir de la cèl·lula. El mitocondri conserva, però, una certa independència respecte a la cèl·lula: és semi-autònom genèticament ja que està proveït d'un cromosoma de DNA. Un mitocondri conté de 2 a 10 molècules de DNA, i pot haver-hi fins a 1000 mitocondris per cèl·lula somàtica (Budowle i col·laboradors. 2003). S'estima, per exemple, que un oòcit madur conté més de 100.000 molècules de mtDNA. (Michaels i col·laboradors. 1982; Piko i Matsumoto, 1976). El mitocondri també es distingeix de la cèl·lula pel seu codi genètic diferent del que fa servir el nucli cel·lular. Per tant, el gens del mtDNA són indesxifrables pel sistema nucleocitosòlic

(Wallace 1982). UGA es llegeix com triptòfan en lloc de codó “stop”, AGA i AUU com un codó “stop”, en lloc d'arginina, AUA com a metionina en lloc de isoleucina, i AUA, AUU, i AUG es llegeixen com a codó d'iniciació (Anderson col·laboradors. 1981; Montoya col·laboradors. 1981).

El genoma mitocondrial humà és circular, de doble cadena, sense proteïnes associades. Està format per 16569 parells de bases i conté 37 gens, dels quals 13 són gens polipeptídics que codifiquen algunes de les sub-unitats essencials dels enzims de la fosforilació oxidativa (OXPHOS) que generen l'energia mitocondrial; 2 codifiquen RNA ribosomals (12S i 16S rRNA), i 22 codifiquen per els RNA de transferència (tRNA) necessaris per a la síntesis de proteïnes mitocondrials.

La majoria de les proteïnes OXPHOS i ribosomals, els DNA i RNA polimerases, i els factors de transcripció són tots codificats pels gens nuclears, sintetitzats en el citoplasma cel·lular, i, després, exportats dins del mitocondri (Shoffner & Wallace 1995, Wallace i col·laboradors. 1997a). La transcripció iniciada a partir de dos promotors (P_H i P_L) crea un RNA policistronic: la majoria de les seqüències de mRNAs estan flanquejades directament en la posició terminal 5' per una seqüència de tRNA (Montoya i col·laboradors. 1981). Els RNA missatgers (mRNA), ribosomal i de transferència són modificats després de la transcripció amb una cua de poliadenines pels mRNAs i rRNAs, i amb un codó CCA terminal en 3' de la seqüència dels tRNAs (Attardi i col·laboradors. 1982; Attardi i Montoya 1983, Clayton 1984, Wallace 1993, Taanman 1999). La transferència de seqüències de mtDNA cap al nucli és un procés continu (Wallace 1997, Hirano i col·laboradors. 1997), però no tots el gens mitocondrials que passen al nucli són funcionals. Hom estima que un centenar de seqüències mitocondrials no funcionals o pseudogenes es troben integrades al genoma nuclear humà (Tsuzuki i col·laboradors. 1983, Shay & Werbin 1992; Zischler i col·laboradors. 1998; Zischler 2000).

IV. El DNA mitocondrial (mtDNA)

Anderson i col·laboradors (1981) van descriure la seqüència completa i l'organització dels gens del genoma mitocondrial humà al 1981 (figura 3). La seqüència publicada es la seqüència referència estàndard: *the Cambridge Reference Sequence (CRS)*. EL DNA mitocondrial està format per dues cadenes, la cadena pesada H (*Heavy Strand*), rica en guanina, i la cadena lleugera L (*Light Strand*), rica en citosina. El genoma mitocondrial s'organitza en dues regions: la regió codificant i la regió no codificant, o regió de control (CR). La numeració de la seqüència estàndard de referència (CRS) es va establir segons Anderson i col·laboradors, a la qual es van afegir modificacions menors aportades ulteriorment per Andrews i col·laboradors (1999). La numeració comença arbitràriament al voltant de la meitat de la regió de control, en el origen de replicació de la cadena pesada (O_H), de tal manera que la regió de control s'expandeix des de la posició 16024 fins a la posició 16569, i segueix des de la posició 1 fins a la posició 576.

IV.1. La regió codificant:

La regió codificant representa 90% del genoma mitocondrial i conté 37 gens, 28 dels quals estan codificats per la cadena pesada (H), i 9 per la cadena lleugera (L). Totes les seqüències codificants són contínues al llarg del genoma mitocondrial i no hi ha introns (Anderson i col·laboradors. 1981, Wallace i col·laboradors. 1992, Zeviani i col·laboradors. 1998).

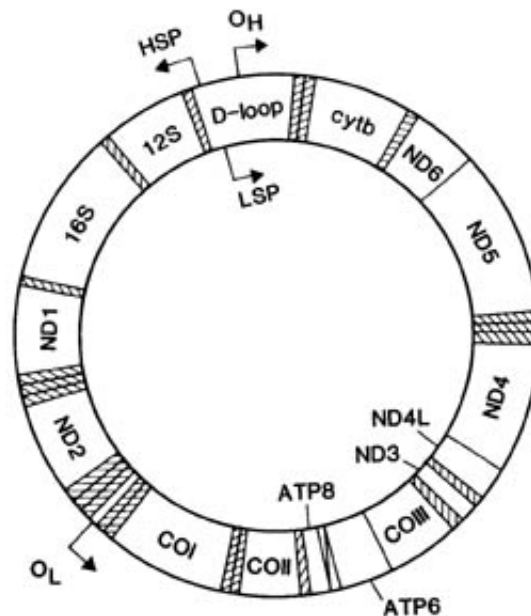


Figura 3: DNA mitocondrial humà. Regions codificants i no codificants, i organització dels gens (gens OXPHOS i rRNAs). Els tRNA es mostren amb una trama ratllada.

IV.2. La regió no codificant:

El DNA mitocondrial es caracteritza per una regió no codificant de 1121 parells de bases, la regió de control (CR). També s'anomena *D-loop* (desplaçament del bucle) per la estructura visible al microscopi electrònic que es forma durant la replicació del mtDNA. La regió de control conté l'origen de replicació (O_H) i les seqüències conservades relacionades amb l'inici de la replicació de la cadena H (Walberg i Clayton 1981). També conté les seqüències associades amb la finalització del desplaçament del bucle o D-loop (Chang i Clayton 1984), les seqüències de control de transcripció (Cantatore i Attardi 1980), i els promotors de transcripció (P_H i P_L) (Hixson i Clayton 1985). Dins de la regió de control, també trobem porcions de seqüències altament variables entre individus i selectivament neutres. Van ser definides per Vigilant i col·laboradors (1989) com el segment hipervariable I (HVS-I), i el

segment hipervariable II (HVS-II). El HVS-I s'expandeix des de la posició 16024 fins a la 16365, i el HVS-II des de la posició 73 fins a la 340.

IV.3. Característiques del mtDNA

Amb només 16569 parells de bases, el genoma mitocondrial representa el 0.00006% del genoma nuclear humà, però la seva contribució a la comprensió de la evolució humana té un pes considerable que compensa amplement la seva petita (o minúscula) contribució al nostre genoma. El mtDNA és una eina genètica potent pels estudis d'evolució humana i de genètica de poblacions humanes degut a les seves propietats.

IV.3.a. Herència materna

El mtDNA humà es transmet als descendents únicament per via materna (Giles i col·laboradors. 1980; Case i Wallace 1981) degut en gran part a una diferència numèrica. L'òcit madur conté aproximadament unes 100.000 molècules de mtDNA, mentre que l'espermatozoide conté només entre 100 i 1500 mtDNAs (Chen i col·laboradors. 1995, Manfredi i col·laboradors. 1997). D'altra banda, els mtDNAs de l'esperma arriben a introduir-se a l'ou en el transcurs de la fertilització (figura 4), però són eliminats en un estadi precoç de l'embriogènesi, entre el segon i el quart estadi cel·lular (Manfredi i col·laboradors. 1997). Tampoc es pot detectar el mtDNA patern en

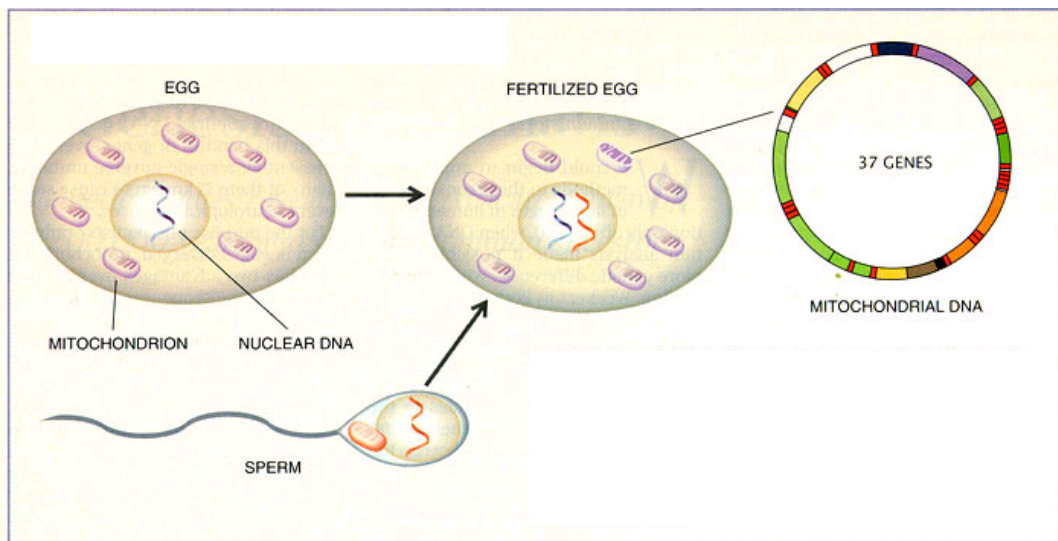


Figura 4: Herència mitocondrial

nounats nascuts després d'una fertilització *in vitro* per injecció intracitoplàsmica d'esperma (Danan i col·laboradors. 1999). L'eliminació de l'esperma en un estadi precoç de l'embriogènesi s'explica per la intervenció d'un mecanisme molecular específic. Es va observar que els mitocondris dels caps dels espermatozoides estan marcats amb ubiquïtines (Hopkin 1999), i per tant, són reconeguts i eliminats immediatament quan s'introdueixen al citoplasma de l'òvul. No obstant, la presència de mtDNA patern s'ha pogut trobar en alguns embrions humans anormals (poliploïdes) generats per fertilització *in vitro* i a partir de tècniques d'injecció intracitoplàsmica d'esperma (St John 2002).

El fet de presentar una herència estrictament materna vol dir que el mtDNA:

1. és un genoma haploide: totes les molècules de mtDNA en un individu són les mateixes (homoplasmia), excepte en els casos anòmals d'heteroplasmia

2. no recombina, de tal manera que les diferències entre dues seqüències mitocondrials representen només les mutacions que es van acumular des de la separació amb la seqüència ancestral.
3. de tots aquells individus que pertanyen a un mateix llinatge matern presenten la mateixa seqüència.

IV.3b. Taxa de mutació

La taxa d'aparició de noves mutacions en el mtDNA és de 10 a 17 vegades més elevada que en el genoma nuclear (Neckelmann i col·laboradors. 1987, Wallace col·laboradors. 1997a). Els segments hipervariables de la regió de control presenten una taxa de mutació encara més elevada que la de la regió codificant (Howell i col·laboradors. 1996). L'acumulació més ràpida de polimorfismes en el mtDNA es pot atribuir a algunes característiques úniques d'aquest sistema genètic:

- la mitocondria no disposa d'un sistema eficient de reparació de DNA (Bogehagen 1999), la qual cosa permet que s'acumulin més canvis introduïts per error.
- El mtDNA no està associat a proteïnes protectores com les histones, està associat a la membrana interna on es generen els radicals lliures, altament mutagènics (Richter i col·laboradors. 1988).
- Un metabolisme anormal de la mitocondria, com per exemple un increment del metabolisme oxidatiu, pot accelerar la taxa de mutació (Richter i col·laboradors. 1988).

IV.3.c. Homoplàsmia-Replicació segregativa-Heteroplàsmia

Totes les molècules del mtDNA d'una mateixa cèl·lula són idèntiques entre elles, és el que s'anomena *l'homoplàsmia*. En el transcurs de la divisió cel·lular, les mitocondries són distribuïdes aleatòriament en les cèl·lules filles; és el procés de la *replicació segregativa*. Arran de l'aparició d'una mutació, es crea una barreja intracel·lular de molècules normals i mutants, condició coneguda com *heteroplàsmia*. Malgrat el nombre elevat de molècules de mtDNA en els oòcits madurs i les poques divisions cel·lulars en la línia germinal materna, les variants de seqüències de mtDNA segreguen ràpidament entre generacions (Poulton i col·laboradors. 1998). Aquest fenomen s'atribueix a un coll d'ampolla genètic. La quantitat de molècules de DNA es redueix a un nombre molt petit durant l'oogènesis. Aquesta sub-població, que pot tenir una proporció de variants diferent de la de la població original, és transmesa després del coll d'ampolla i la població fundadora es replica fins a produir ~ 100 000 còpies de mtDNA en l'òvul madur. Aquest fenomen permetria la fixació d'una mutació en un llinatge únic en un nombre reduït de generacions o en una única generació.

Per tal de detectar variants de seqüències de mtDNA en un mateix individu es va dur a terme un estudi de comparació de seqüències d'un centenar de clons d'un mateix individu (Monnat i col·laboradors. 1986) que va concloure que l'heteroplàsmia era un fenomen poc freqüent i que els individus es podien considerar com homoplàsmics. El primer cas descrit d'heteroplàsmia de mutació puntal en la regió de control del mtDNA humà va ser en el cas d'identificació de les restes del Tsar Nicholas II (Gill i col·laboradors. 1994). Posteriorment, van seguir diversos treballs referint-se a l'heteroplàsmia de la regió de control (Comas i col·laboradors. 1995; Bendall i col·laboradors. 1996; Wilson i col·laboradors. 1997; Parsons i col·laboradors. 1997), però la mida mostral era massa petita per poder determinar amb precisió la freqüència d'heteroplàsmia en les posicions nucleotídiques afectades. També es va

descriure heteroplàsmia de longitud en la regió de control en dos tractes de policitosines: un al HVS-I (Bendall i Sykes 1995), i un altre al HVS-II (Marchington i col·laboradors. 1998).

La detecció d'heteroplàsmia de mutació puntual depèn bàsicament de la sensibilitat del mètode utilitzat. Es pot arribar a detectar una freqüència de 2 a 8% d'heteroplàsmia quan s'utilitza el mètode de seqüenciació, i fins a un 14% amb el mètode de gel d'electroforesi en gradient desnaturalitzant (DGGE). Una possible explicació per aquesta diferència de sensibilitat és que, en dades de seqüències, es fa difícil distingir l'heteroplàsmia del soroll de fons generat pels artefactes de la química de seqüenciació. Per aquesta raó, en molts estudis poblacionals humans basats en la seqüenciació de la regió de control s'assumeix l'homoplàsmia, i en les posicions on es pot haver produït l'heteroplàsmia, es considera el nucleòtid predominant per aquesta posició o bé es defineix aquesta posició com ambigua per causa desconeguda.

IV.3.d. Alt nombre de còpies per cèl·lula

Les cèl·lules somàtiques tenen dos còpies de qualsevol gen nuclear i de cent a mil còpies de mtDNA. en el citoplasma cel·lular (Robin i Wong 1988). Degut a la seva abundància, el mtDNA és més fàcil d'obtenir que el DNA nuclear. Aquesta característica converteix el mtDNA en marcador amb múltiples aplicacions: la genètica de poblacions, la medicina forense, i els anàlisis de DNA antic. En els casos on els marcadors nuclears no poden ser utilitzats perquè el material biològic és poc abundant (pèls sense bulb), o parcialment degradat (restes esquelètiques), el mtDNA és un marcador d'una gran utilitat ja que és l'únic marcador que es pot amplificar amb èxit.

V. El DNA mitocondrial: una eina genètica per aclarir la història de les poblacions humanes.

V.1. L'origen dels humans moderns

“Qui som? D'on venim?”, són preguntes fonamentals que s'ha fet la humanitat durant milers d'anys. La preocupació dels humans pels seus orígens es reflecteix, per exemple, en la mitologia judeo-cristiana amb Adam, Eva, i el jardí d'Eva. Les històries de la creació i de l'evolució de la nostra espècie semblen ser una característica universal de les cultures humanes.

La prova més directa del nostre passat està continguda en les restes fòssils que l'antropologia clàssica va explotar per tal d'aportar tota la informació possible sobre l'evolució dels humans. Però les dades extretes a partir de l'explotació dels fòssils no són del tot concloents per tal de resoldre la qüestió de l'origen i de l'evolució dels humans moderns. El registre fòssil és incomplet, hi ha molts forats i molta controvèrsia. Al principi dels anys 80, l'antropologia molecular va ampliar el seu camp de recerca al nivell del DNA. Les seqüències de DNA, mitocondrials i cromosomes del nucli cel·lular, guarden un record (polimorfisme) de la nostra experiència com a espècie. Desxifrar aquest record no és una tasca fàcil degut a nombrosos factors que poden influenciar el patró de seqüències del nostre DNA. Però, les variacions genètiques o polimorfismes que existeixen entre els humans ens fan a tots únics i l'anàlisi d'aquests polimorfismes ens permet resoldre la qüestió de l'origen de l'home modern i establir quines són les relacions genètiques entre nosaltres, o com som de propers o de llunyans genèticament.

Dues hipòtesis principals van ser proposades per explicar l'origen de l'home modern. Totes dues estan d'acord amb la sortida de Àfrica d'individus pertanyents al

gènere *Homo* entre 0.8 i 1.8 milions d'anys cap a Europa i Àsia però discrepen en el model de transició entre aquests primers *Homo* i *Homo sapiens*

V.1.a. La continuïtat multiregional

Aquest model suggereix que no hi ha un origen geogràfic únic dels humans moderns sinó que van evolucionar a partir dels primers homínids (*Homo erectus*) en diferents regions del món de manera simultània (figura 5). Els partidaris del model multiregional proposen una transició contínua entre les poblacions d'*Homo erectus* i els *Homo sapiens* de les diferents regions geogràfiques.

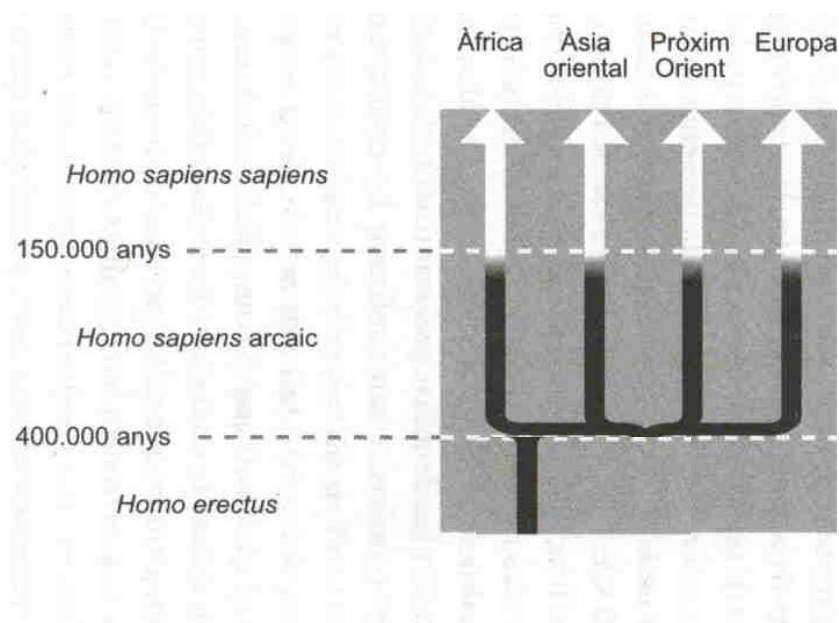


Figura 5: Hipòtesi en canelobre; continuïtat multiregional.
(Figura modificada a partir de Bertranpetit i Junyent 1998)

L'origen multiregional de l'home modern es va recolzar en les observacions de continuïtat regional d'alguns trets morfològics en restes fòssils. Perquè hagi estat possible una evolució paral·lela entre *H. erectus* i *H. sapiens* s'hauria d'haver produït un flux gènic considerablement important entre les poblacions. Per tant, els llinatges genètics presents en les poblacions de les diferents regions del món haurien de

tenir un temps de coalescència molt antic i el flux gènic hauria d'haver estat molt important i molt intens durant un gran període de temps i a través de grans regions geogràfiques. Una evolució multiregional implicaria una mida poblacional suficientment gran com per sostenir un flux gènic entre les poblacions repartides a través de les diferents regions del món.

V.1.b. El model de l'Origen Recent Africà (ORA)

El model de l'Origen Recent Africà, més conegut com el model de “*Out of Africa*” suggereix que els humans moderns van sortir de l'Àfrica fa entre cent i dos-cents mil anys, i que es van expandir i van colonitzar la resta del món sense (figura 6)

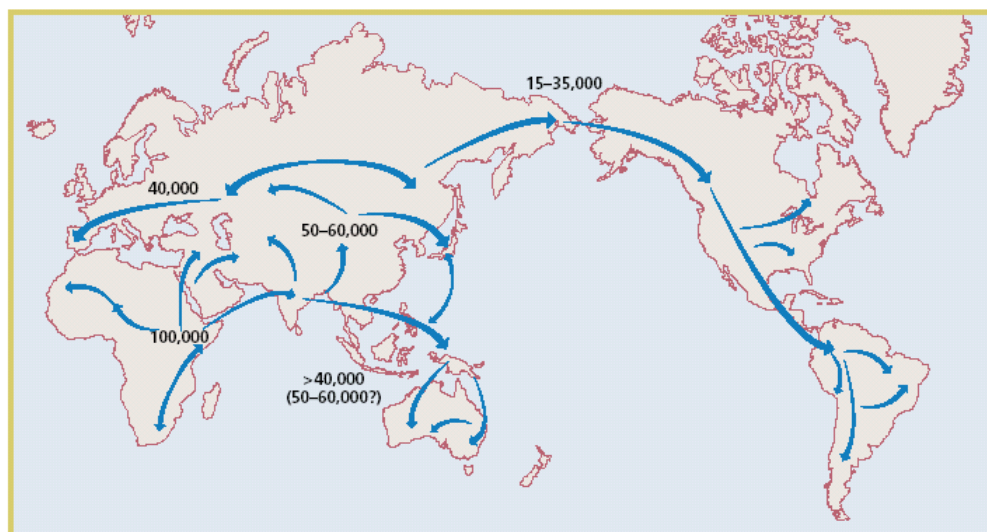


Figura 6: Hipòtesi del “*Out of Africa*”
(extreta de *nature genetics* volum 33; 2003)

que hi hagués una barreja genètica amb els humans arcaics presents fora de l'Àfrica, com els Neandertals (Stringer i Andrews 1988; Stringer 2002). Dades aportades per les restes fòssils i l'arqueologia donen suport a aquest model: restes fòssils entre 90 i 120

milers d'anys d'humans anatòmicament modern van ser trobades a l'Àfrica i a l'Orient Mitjà (Lahr i col·laboradors. 1996; Stringer i Andrews 1988; Hennessy i Stringer 2002), i les dades arqueològiques van datar l'aparició del comportament de l'home modern entre 70 i 90 milers d'anys (Henshilwoo i col·laboradors. 2002)

Una versió modificada del model “*Out of Africa*” es el “Weak Garden of Eden hypothesis (WGE)”. Aquest model proposa que després de la sortida d'Àfrica, es va produir un coll d'ampolla reduint la mida de la població, que es va expandir posteriorment fa 50 000 anys (Harpending i col·laboradors. 1993).

Aquests dos models, l'ORA i el WGE suggereixen que tots els llinatges en les poblacions humanes provenen d'un avantpassat comú africà. Si l'origen dels humans moderns es troba a l'Àfrica, i un subgrup d'aquesta població va sortir recentment de l'Àfrica, esperariem trobar una part de la diversitat genètica de les poblacions africanes actuals en les poblacions no africanes. Els nivells de diversitat en les poblacions no africanes depèn de l'amplitud del coll d'ampolla a la sortida d'Àfrica. El model WGE preveu alts nivells de subdivisions genètiques a través de diferents regions del món (Àfrica, Àsia, Europa), però no tan alts com aquells previstos pel model de continuïtat multiregional. En el model del “*Out of Africa*”, l'emergència de l'home modern no està necessàriament vinculat a un flux gènic a través de les regions del món. Models intermedis, com els models d'assimilació o d'hibridació, també han estat proposats. Suggereixen que el flux gènic entre les poblacions dels primers humans no va ser constant ni en el temps ni en l'espai. Els models d'hibridació assumeixen un flux gènic entre els humans moderns que van sortir de l'Àfrica i els primers homínids fora de l'Àfrica. D'aquesta manera l'evolució dels humans moderns és el resultat de la barreja entre els caràcters moderns de les poblacions africanes amb característiques de les poblacions arcaiques euroasiàtiques, com els Neandertals. Aquest model preveu una contribució variable dels gens de les poblacions arcaiques africanes i no africanes al *pool* genètic de les poblacions modernes.

La majoria de les dades genètiques donen suport al model del “*Out of Africa*” (Stringer i col·laboradors. 1988; Takahata i col·laboradors. 2001; Satta i Takahata 2002), però els models derivats del “*Out of Africa*” que preveuen un flux gènic entre les poblacions modernes que van sortir de l'Àfrica amb les poblacions d'humans arcaics que ja estaven fora de Àfrica són difícils de descartar basant-se en les dades disponibles (Wall 2000; Nordborg 1998, 2001; Templeton 2002).

V.2 L'avantpassat mitocondrial

Un tipus de dada genètica que va donar suport a la teoria de l'origen africà recent de l'home modern i va fer trontollar l'hipòtesi multiregional és el DNA mitocondrial.

El primer treball sobre la variació del genoma mitocondrial en les poblacions humanes va ser dut a terme per Wesley Brown (1980). Està basat en l'anàlisi de polimorfisme de longitud de fragments de restricció (RFPLs: Restriction Fragment Length Polimorphism), mitjançant 18 enzims de restricció en 21 individus. Brown va estimar que la diversitat del mtDNA present a les poblacions humanes actuals va començar a acumular-se fa 180.000 anys, ja que les variants detectades diferien de la seqüència ancestral en 0.18 per cent de les posicions i que la taxa de substitució nucleotídica era de 1 per cent per milió d'anys.

La filogènia de la variació global del mtDNA va ser establerta a partir d'estudis posteriors, també basats en RFLPs de baixa resolució (5 ó 6 enzims de restricció) i una mida mostral gran. L'arbre obtingut presentava una forma en estrella (“*starlike*”): d'un únic haplotip central comú a tots els individus d'on sorgeixen els altres haplotips, dels quals alguns són específics d'una població. L'haplogroup central o universal correspon a l'avantpassat comú més recent de tots els mtDNA del món. Això suggereix que totes les poblacions humanes comparteixen una història evolutiva

comuna des de fa molt de temps, i podria donar suport a l'idea que l'home va aparèixer en diferents parts del món a partir d'avantpassats arcaics (Excoffier & Langaney 1989; Templeton 1992).

El mtDNA va conèixer el gran èxit amb un treball publicat al 1987 per l'equip d'Allan Wilson, que va tenir un gran impacte en el món científic amb repercussions a nivell popular arrel del debat sobre l'origen de l'home modern i de l'Eva mitocondrial africana. Rebecca Cann, Mark Stoneking i Allan Wilson van analitzar 147 individus de cinc poblacions (africans, asiàtics, australians, caucasoides, nou guineans) mitjançant RFLPs d'alta resolució (12 enzims de restricció), per tal d'obtenir una filogènia del mtDNA més precisa. Els resultats obtinguts van ser representats en un arbre de màxima parsimònia (figura 7).

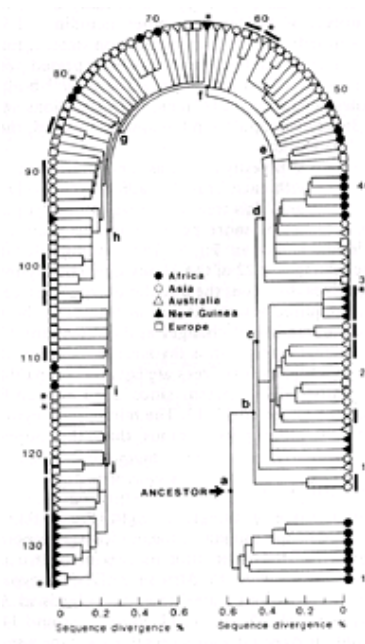


Figura 7: Arbre genealògic de 134 tipus de DNA humà; (Cann i col·laboradors 1987).

L'arbre del mtDNA proposat per Cann i els seus col·laboradors consisteix en dues branques: la més profunda presenta exclusivament mtDNA africans i l'altra mtDNA de les cinc poblacions. Els autors van concloure que tota la diversitat del mtDNA prové

d'una sola dona que va viure a l'Àfrica fa uns ~200.000 anys, amb un marge d'error tan elevat que la data estava compresa entre 140.000 i 290.000 anys. Aquest resultat va ser interpretat com l'evidència de l'origen recent africà (*Out of Africa*) de l'home modern. Cann i col·laboradors van designar el nostre avantpassat mitocondrial africà com l'"Eva mitocondrial". Va ser un nom molt sensacionalista per la premsa escrita i televisiva que va fer una àmplia difusió d'aquesta metàfora. Però el nom d'Eva va crear molta confusió i va ser mal interpretat, ja que evocava la falsa imatge de l'*Homo sapiens* provenint d'una mare única, la mare de tots nosaltres, i que no n'hi havia hagut cap altra abans.

La hipòtesi proposada per Cann i col·laboradors (1987) i defensada posteriorment per altres autors (Vigilant i col·laboradors. 1991; Chen i col·laboradors. 1995) es basa en el principi de coalescència (figura 8). La coalescència assumeix l'existència d'un origen comú a tots els organismes. A partir d'aquesta base, es dedueix que tota la variació d'un segment de DNA (mitocondrial o nuclear) en les generacions actuals prové d'un avantpassat únic. El cas del mtDNA, la reconstrucció dels llinatges en els temps és més fàcil ja que té una herència estrictament materna. Les dades de l'article publicat per Cann i els seus col·laboradors indiquen que:

- a) L'avantpassat mitocondrial no era l'únic individu viu, sinó que formava part d'una població on la resta dels llinatges mitocondrials es van perdre amb el temps.
- b) Aquest avantpassat no correspon a la primera dona apareguda de la nostra espècie, però sí al punt de partida de tots els llinatges mitocondrials actuals.

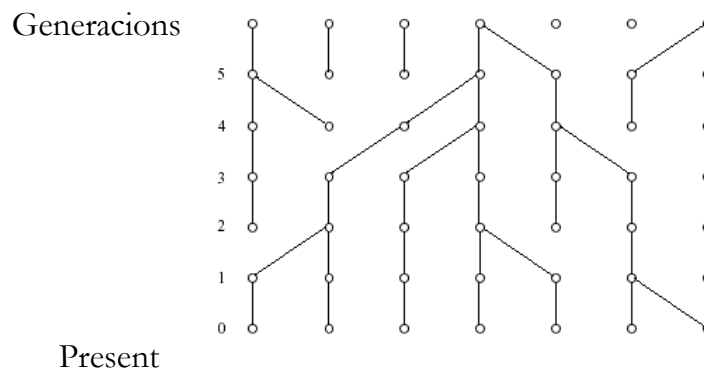


Figura 8: Il·lustració del principi de coalescència. Per tots els membres d'una població es pot trobar l'avantpassat matern. L'exemple mostra els llinatges mitocondrials de 7 dones d'una població estable. A cada generació, alguns llinatges proliferen i altres s'extingeixen. Per atzar, un dels llinatges materns substitueix tot els altres.

Aquest treball va ser l'objecte de nombroses crítiques: Primer, es va estimar que el nombre de llocs testats per l'anàlisi de RFLPs no era suficient per a aportar una informació genealògica robusta. Segon, el mètode utilitzat per designar l'arrel de l'arbre, agafant el punt mig de la branca més llarga, no era fiable. El tercer punt, és que van utilitzar individus afro-americans en representació de nadius africans. A més, l'arbre presentat en aquest estudi no era el més parsimoniós, i entre els més parsimoniosos alguns no exhibien una clara separació entre les poblacions africanes i la resta dels individus (Templeton 1992). Finalment, l'error estàndard de l'estima de la divergència entre els africans i no africans era tan gran que el temps de divergència podria ser de fins a 800.000 anys. L'estudi de Vigilant i col·laboradors (1991) basat en les seqüències de 189 individus de tot el món dels quals 121 eren nadius africans va confirmar l'origen africà del mtDNA (figura 9). A més, la utilització de seqüències de la regió de control de ximpanzé (Kocher i Wilson 1991; Foran i col·laboradors. 1988) per calibrar la taxa d'evolució del mtDNA va donar suport al treball de Cann i col·laboradors (1987).

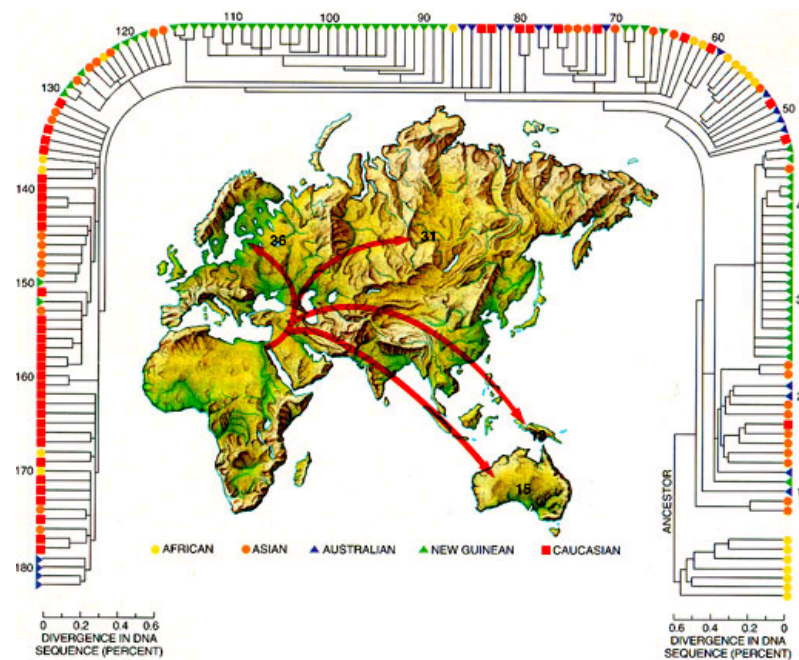


Figura 9: Arbre genealògic de 182 tipus de mtDNA humà que dóna suport a l'origen africà de l'avantpassat mitocondrial dels humans moderns (figura basada en el treball de Vigilant i col·laboradors 1991 i modificada a partir de l'article *The recent African genesis of Human; Scientific American 1992*)

Estudis posteriors (taula 1) es van basar en la seqüenciació de la regió de control, o d'alguns gens del mtDNA per tal de reduir l'error en l'estima del temps de divergència entre els africans i no africans.

Autors	Segment del mtDNA	Percentatge de divergència	d	TMRA (anys)
Cann et al., 1987	Molècula sencera	0.57	15.0	190.000
Kocher and Wilson 1991	ND4-5	0.33	9.6	172.000
Vigilant et al., 1991	Regió de Control	2.90	69.2	210.000
Ruvolo et al., 1993	COII	0.58	10.4	278.000
Chen et al., 1995	Molècula sencera	0.29	13.0	112.000
Horai et al., 1995	Molècula sencera	1.10	39.0	143.000
Horai et al., 1995	Regió de control	2.10	70.0	143.000
Watson et al., 1997	Regió de control	1.10	495.0	111.000
Chen et al., 2000	Molècula sencera	0.36	13.0	138.000
Ingman et al., 2000	Molècula sencera	0.58	17.0	171.500

Taula 1: Percentatge de divergència de seqüència entre dues seqüències humanes de mtDNA, diversitat de seqüència (d) estimada a partir dels mtDNA d'humà i ximpanzé, i temps de l'avantpassat comú més recent (TMRA) han estat obtinguts a partir de les dades dels diferents autors. Per l'homogeneïtat de les dades, s'assumeix un temps de divergència entre els llinatges d'humans i ximpanzés de 5 milions d'anys. (Taula extreta de Klein i Takahata 2001)

A més, l'anàlisi de seqüències del mtDNA de tres espècimens fòssils de Neandertal (Krings i col·laboradors. 1997, 1999, 2000; Ovchinnikov i col·laboradors. 2000) dona suport al model del "Out of Africa", ja que les dades indiquen una clara diferenciació entre el mtDNA neandertal i d'humà modern i el temps de divergència entre els dos tipus de llinatges s'estima entre ~ 300 i 850 milers d'anys.

V.3 Distribució dels llinatges mitocondrials i de les poblacions humanes.

El genoma mitocondrial ha servit no només per elucidar el origen de la nostra espècie, sinó que també s'ha utilitzat conjuntament amb altres marcadors genètics per seguir el rastre de les migracions de les poblacions humanes. El grup d'humans que va sortir d'Àfrica per ocupar nous territoris es va confrontar a condicions geogràfiques i climàtiques molt diferents. L'adaptació a un nou ambient i l'acció de la selecció natural van provocar canvis en el nostre genoma que van conduir a la diferenciació dels grups humans. Les mutacions es van acumular de manera seqüencial en els llinatges del mtDNA a mesura que els grups de poblacions van ocupar els diferents continents (figura.6), de tal manera que els llinatges del mtDNA tenen una distribució geogràfica que correspon a la de les poblacions humanes. La caracterització dels llinatges mitocondrials dels diferents continents permet entendre la variació genètica entre poblacions, i la diversitat genètica dintre de les poblacions. La genealogia del mtDNA va ser construïda a partir de la definició d'haplogrups, és a dir un grup de seqüències que comparteixen un mateix patró de mutació (figura10) i que són específics de continents (Wallace i col·laboradors. 1999).

V.3a. Els llinatges mitocondrials a l'Àfrica

A l'Àfrica, els tres haplogrups L0, L1, i L2 constitueixen el macrohaplogrup L, i són específics de l'Àfrica sub-sahariana (Chen i col·laboradors. 1995; Graven i col·laboradors. 1995). A partir del macrohaplogrup L es van formar el haplogrup africà L3, i els macrohaplogrup euroasiàtics M i N. Aquests dos haplogrups van aparèixer al nord est de l'Àfrica i es van expandir a Europa i Àsia quan els individus portadors d'aquests dos haplogrups van sortir d'Àfrica per colonitzar la resta

dels continents (Wallace i col·laboradors. 1999; Quintana-Murci i col·laboradors. 1999). L'anàlisi del mtDNA també ha revelat una gran heterogeneïtat dins d'Àfrica: per exemple, l'haplogrup M1 té el seu origen a l'Àfrica de l'Est (Quintana-Murci i col·laboradors. 1999), i l'haplogrup U6 és específic de l'Àfrica del Nord (Rando i col·laboradors. 1998, 1999). A més, les expansions Bantu i els flux gènics posteriors van contribuir a la formació de *pool* genètic diferents a les grans regions d'Àfrica (Bandelt i col·laboradors. 2001; Pereira i col·laboradors. 2001; Salas i col·laboradors. 2002).

V.3b. Els llinatges mitocondrials a Europa

Europa es caracteritza per la seva gran homogeneïtat genètica (Simoni i col·laboradors. 2000a; 2000b; Helgason i col·laboradors. 2000; Richards i col·laboradors. 2002), ja que els haplogrups H, I, J, N1b, T, U, V, W i X, tots derivats del macrohaplogrup N, constitueixen més del 98% dels llinatges mitocondrials europeus (Mishmar i col·laboradors, 2003). Els Samis i els Islandesos, a diferència del grup homogeni format per la resta d'europeus (Simoni i col·laboradors. 2000a), apareixen com poblacions aïllades. Aquestes poblacions presenten haplogrups que són comuns a la resta d'Europa (Finnila i col·laboradors. 2001; Wittig i col·laboradors. 2003), però la seva posició d'*ouliers* en el paisatge europeu és conseqüència de l'aïllament geogràfic i de la deriva genètica.

V.3c. Els llinatges mitocondrials a Àsia

Els llinatges mitocondrials coneguts com específics de l'Àsia que deriven del macrohaplogrup M són els haplogrups C, D, E, G, Z, i els que deriven del macrohaplogrup N són A, B, F, Y (Kivisild i col·laboradors., 2002). Degut a la

grandària del seu territori i a la complexitat del seu poblament, l'Àsia presenta una gran diversitat de freqüències haplotípiques. Podem destacar per exemple el cas de Sibèria (Derbeneva i col·laboradors. 2002) on els haplogrups A, C, D, G, Z i Y formen > 75% del *pool* genètic mitocondrial, i del Tibet (Torroni i col·laboradors. 1994) on aquests haplogrups representen només 14% dels llinatges mitocondrials. També podem destacar el cas de l'Índia que també presenta un gran diversitat en la composició dels seus llinatges mitocondrials. El seu *pool* genètic està format per haplogrups de l'oest euroasiàtic (HV, U, JT, R1; Macaulay col·laboradors, 1999), llinatges específics de l'est euroasiàtic (B, R9; Kivisild i col·laboradors, 2002), i per llinatges propis, com per exemple U2i i grups de llinatges dins el super-haplogrup M (Kivisild i col·laboradors, 1999; Bamshad i col·laboradors. 2001).

V.3d. Els llinatges mitocondrials a les Amèriques

En les poblacions de nadius americans, el cinc haplogrups A, B, C, D, i X engloben el 100% de la variació del mtDNA. Els haplogrups A, C, D representen 58% dels llinatges del Nord de Sibèria i la seva presència a les Amèriques seria la conseqüència de la travessia de l'estret de Bering des de Sibèria pels individus portadors d'aquests haplogrups. L'haplogrup B és present bàsicament a la costa asiàtica i en el Pacífic però pràcticament absent a Sibèria i poc freqüent a Amèrica del Nord. L'haplogrup X específic d'Amèrica es distribueix bàsicament a Amèrica del Nord, encara que la seva distribució no està ben establerta (Brown i col·laboradors, 1998).

V.4. Altres aplicacions del mtDNA

Com acabem de veure, el mtDNA ha permès entendre la diversitat genètica de les poblacions humanes generada per la colonització de les diferents regions geogràfiques. Però, el mtDNA també es fa servir en el camp de la genètica forense per els casos d'identificació de persones, com per exemple en el cas de la identificació de les restes del Tsar Nicolau II (Zhivotovsky, i col·laboradors. 1999) i de la identificació del presumpte cor de Lluís XVI, fill de Lluís XVI i de Marie-Antoinette (Jehaes i col·laboradors. 2001), entre molts d'altres.

VI. Genealogia del gen

Existeixen dues fonts d'error a l'hora de fer inferències sobre la genealogia a partir de seqüències. Una fa referència al nombre de posicions variables i l'altra al nombre de loci variables. En un genoma no recombinant, com més llarg és el segment seqüenciat, més precisa és l'estima de la distància evolutiva entre dues seqüències, i més robusta és la genealogia. Els resultats observats quan s'analitza un sol gen (o locus) poden ser atribuïts, o bé a la pròpia natura del locus (mutació, selecció, migració, deriva, mida poblacional, ...), o bé, reflecteixen els fenòmens demogràfics de la població estudiada. L'altra dificultat a tenir en compte és la taxa de mutació de les posicions variables estudiades. En el cas de el mtDNA, la taxa de mutació elevada de la regió de control (~10 vegades més que la regió codificant; Vigilant i col·laboradors. 1991) conjuntament amb la variació de la taxa de mutació per cada posició variable, significa que algunes posicions muten molt ràpid i que muten més d'una vegada en el transcurs de la genealogia. Aquesta situació provoca la possibilitat d'obtenir nombroses topologies d'arbres igualment plausibles. Per aquesta raó, moltes topologies d'arbres poden ser equivalents i no hi ha cap fonament per decidir entre elles. El model de "l'Out of Africa" va ser defensat principalment per l'arrel africana de l'arbre del mtDNA (Vigilant i col·laboradors. 1991). Però està clar que existeixen milers d'arbres possibles, molts dels quals no presenten una arrel africana (Hedges i col·laboradors. 1992; Templeton 1992). Per intentar resoldre aquest dilema es van desenvolupar noves eines per analitzar dades basades en el genoma mitocondrial. Una d'elles es el *network* filogenètic que intenta resumir tots els arbres possibles en un gràfic (Excoffier i Smouse 1994; Bandelt i col·laboradors. 1995).

Watson i col·laboradors. (1997) van aplicar el mètode de *network* al mtDNA en diverses poblacions africanes. Els resultats obtinguts concorden i donen suport a

l'estructura general de l'arbre del mtDNA presentat per Vigilant i col·laboradors (1991), i també va aportar informació complementària: tots els mtDNA eurasiàtics pertanyien a un sol grup, l'origen del qual es troba a l'Àfrica. A partir d'aquest argument es va proposar, seguint el model de "l'*Out of Africa*", que la població que va sortir de l'Àfrica per formar la població eurasiàtica va ser suficientment petita com per eliminar tots els tipus de mtDNA excepte un durant un període de temps. Quintana-Murci i col·laboradors (1999) suggereixen més aviat que els avantpassats de dos tipus principals de mtDNA van sobreviure a l'esdeveniment fundador del "*Out of Africa*".

A principis dels anys 90, l'aplicació de l'anàlisi de restricció d'alta resolució (amb 14 enzims de restricció) a la molècula de mtDNA va permetre cobrir del 15 al 20% de la seqüència del mtDNA. Aquesta tècnica està basada en la digestió de la molècula de mtDNA mitjançant enzims de restricció amb una separació a posteriori per electroforesi (RFLPs). Els estudis de RFLPs van revelar posicions polimòrfiques estables i antigues de la regió no codificant que defineixen grups monofilètics o haplogrups (Torroni i col·laboradors. 1996). Els haplogrups es defineixen com grup de seqüències que comparteixen un mateix patró de mutacions; mutacions que es van acumulant de manera seqüencial al llarg de la molècula i que són específiques de determinades regions geogràfiques. Així doncs, la caracterització dels haplogrups aporta la informació suficient per tal de poder definir les relacions interpoblacionals i inferir la història evolutiva de les poblacions. La tècnica de RFLPs requereix grans quantitats de DNA, una gran bateria d'enzims per obtenir una bona discriminació, un gran nombre de tampons donat que es necessita un tampó específic per cada enzim, i temps ja que és una tècnica no automatitzada. D'altra banda, la majoria de les anàlisis de la variabilitat del mtDNA s'han dut a terme mitjançant una altra estratègia: el mètode de seqüenciació aplicat al segment hipervariable I (HVS-I) de la regió de control. La tècnica de seqüenciació és adequada per detectar qualsevol tipus de mutació (mutació puntual i indels) en el mtDNA, i és una tasca ràpida amb la utilització de

seqüenciadors automàtics, i fàcil ja que la regió analitzada és una seqüència curta d'aproximadament uns 400 parells de bases.

La regió de control és una eina de doble tall degut a la alta taxa de mutació que exhibeix. Aquesta característica és força útil a l'hora de resoldre diferències entre seqüències relativament properes i de definir haplogrups, però, una alta taxa de mutació també implica la presència de mutacions recurrents i de substitucions paral·leles que poden ser la causa d'una manca de resolució a l'hora de construir i de resoldre filogènies. L'anàlisi conjunta de la informació aportada tant per la regió de control (variació de seqüències dels segments hipervariables I i II), com per la regió codificant (RFLPs d'alta resolució) va ser decisiu per definir amb precisió els haplogrups i establir una filogènia robusta del genoma mitocondrial (Figura 10; Macaulay i col·laboradors. 1999). Per una banda, la identificació d'haplogrups amb una genealogia robusta i una distribució geogràfica ben definida ha permès el desenvolupament d'una nova disciplina, la filogeografia. Aquesta està basada en l'estudi de la distribució geogràfica dels llinatges en un arbre genealògic per tal de resoldre qüestions sobre migracions, dispersions, colonitzacions, i per tant, inferir la història de les poblacions humanes. Per una altra banda, la comparació dels polimorfismes de seqüències de la regió de control amb els polimorfismes de longitud de fragments de restricció va complir dos objectius. El primer, va ser comprovar que hi havia una bona concordança entre els dos tipus de polimorfismes, i el segon, distingir entre les mutacions de la regió de control antigues i estables en la genealogia del mitocondrial, i per tant filogenèticament associades a un haplotip concret, de les mutacions recurrents (Bandelt i col·laboradors. 1995). Els estudis on es combinen els dos tipus de dades (seqüències de la regió de control i variants de la regió codificant) van tenir molt d'èxit i van proporcionar *networks* filogenètics acurats de les poblacions de l'oest d'Euràsia (Richards i col·laboradors. 1998; Macaulay i col·laboradors. 1999; Helgason i

col·laboradors. 2000), i de l'Àfrica sub-sahariana (Macaulay i col·laboradors. 1999; Quintana-Murci i col·laboradors. 1999).

Encara que el genoma mitocondrial va ser un dels primers genomes en ser seqüenciat en la seva totalitat, és a partir de l'any 2000, que es va començar a procedir a l'anàlisi de seqüències completes del mtDNA (Ingman i col·laboradors. 2000; Finnilä i col·laboradors. 2001; Maca-Meyer i col·laboradors. 2001; Herrnstadt i col·laboradors. 2002). Seqüenciar tot el genoma mitocondrial representa una gran quantitat de treball i esforç que els progressos tècnics han convertit en una tasca relativament fàcil de dur a terme. Els primers en iniciar de nou l'estudi de seqüències completes de mtDNA van ser Ingman i col·laboradors, que van analitzar 53 seqüències completes d'individus de diferents regions del món (Àfrica, Àsia, Europa). Aquest estudi va confirmar l'arrel africana de la filogènia del mtDNA i els *networks* filogenètics obtinguts a partir de les anàlisis de la regió de control/RFLPs conjunts (Macaulay i col·laboradors. 1999; Quintana-Murci i col·laboradors. 1999). L'estudi dut a terme per Finnilä i col·laboradors (2001) presenta una genealogia completa dels mtDNA europeus amb un *network* filogenètic representant la variació de la regió de control i un altre de la regió codificant. Així, Finnilä i col·laboradors van proporcionar nombrosos marcadors per distingir els diferents haplogrups europeus i els diferents llinatges dintre d'un mateix haplogrup. Aquest treball també va confirmar la informació dels resultats obtinguts pels mtDNA europeus a partir de les dades combinades de la regió de control i de variacions en la regió codificant, ja que totes les seqüències de Finnilä i col·laboradors van poder ser classificades en haplogrups prèviament definits a partir de les anàlisis de restricció d'alta resolució de la regió codificant. Els estudis duts a terme a partir de l'anàlisi de seqüències completes han proporcionat un gran nombre de posicions polimòrfiques que han permès resoldre sense ambigüitats la filogènia del mtDNA i obtenir una clara distribució geogràfica dels llinatges mitocondrials en el context de la filogeografia. Uns dels aspectes rellevants d'aquest tipus d'anàlisi és que van permetre

detectar els errors de topologia generats en les filogènies obtingudes a partir de les seqüències del HVS-I, i de les variants de RFLPs de la regió codificant, i l'altre és que van confirmar les relacions filogenètiques entre els haplogrups inferits prèviament amb les dades de les variants de la regió de control i/o de la regió codificant.

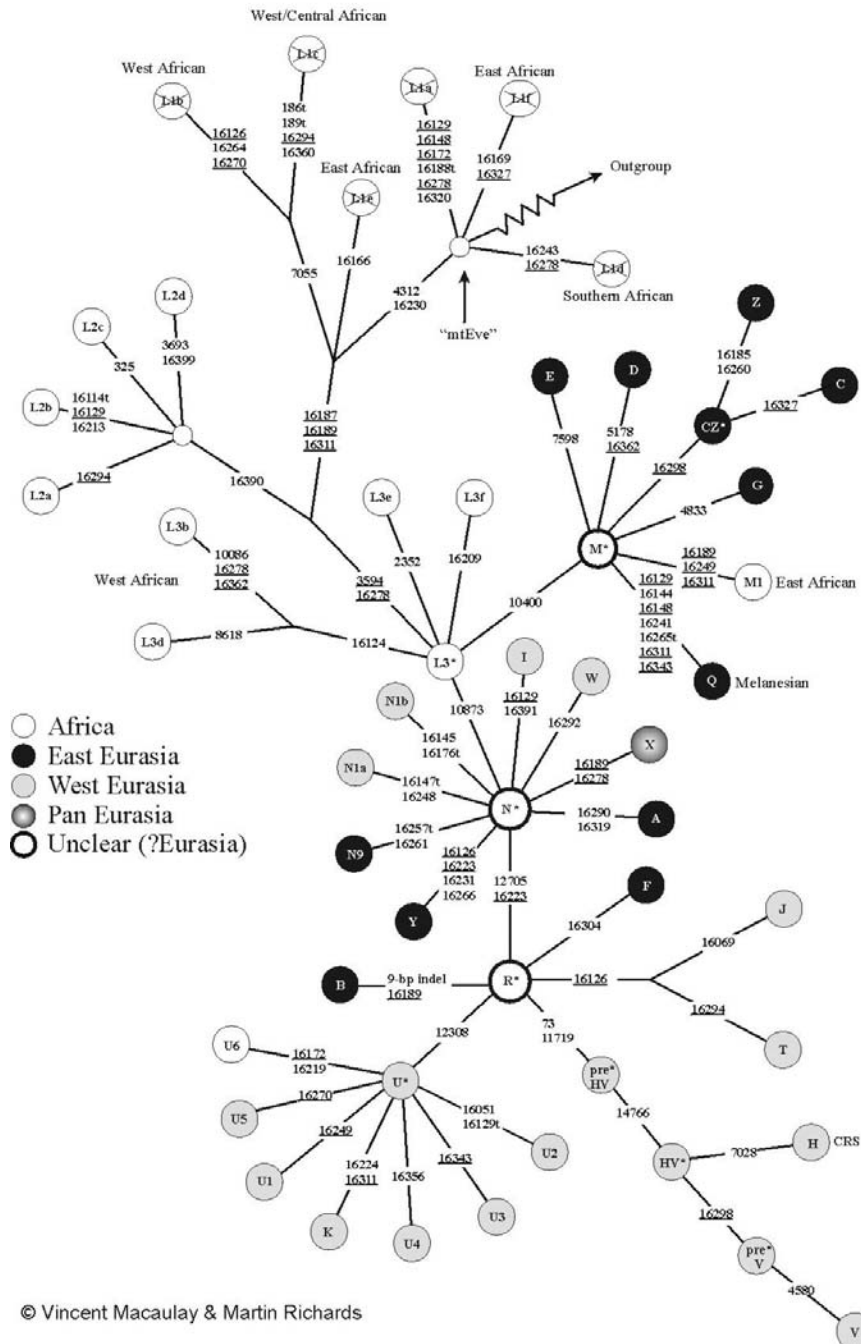


Figura 10: Filogènia del mtDNA construït a partir de dades RFLPs de la regió codificant i de dades de seqüències de la regió de control.

~ MATERIAL I MÈTODES ~

I. Polimorfismes del mtDNA

En la present tesi s'han analitzat diversos polimorfismes que caracteritzen el DNA mitocondrial:

I.1. SNPs (Single Nucleotide polymorphisms)

I.1.a. Regió de control:

Els SNPs de la regió de control s'han analitzat mitjançant seqüenciació directa, ja que en aquesta regió s'acumulen gran quantitat de SNPs (Greenberg i col·laboradors. 1983) degut a l'alta taxa de mutació de la regió de control respecte la regió codificant del mtDNA. El tipus de substitució més freqüent són les transicions, de tota manera, a part dels SNPs, també es poden detectar a la regió de control mitjançant seqüenciació directa polimorfismes del tipus indel (insersions/delecions) d'un nucleòtid.

I.1.b Regió codificant:

Aquests SNPs definits a partir de RFLPs i de seqüències de la regió de control per Macaulay i col·laboradors. (1999) i recentment pels estudis basats en anàlisi de seqüències completes del genoma mitocondrial (Ingman i col·laboradors. 2000; Finnilä i col·laboradors. 2001; Maca-Meyer i col·laboradors. 2001; Herrnstadt i col·laboradors. 2002), han permès definir una filogènia precisa i acurada dels llinatges mitocondrials. La determinació d'aquests SNPs és molt útil per poder definir a quins grans haplogroups pertanyen les mostres analitzades quan la regió de control no és prou informativa

I.2. Microsatèl·lit del mtDNA

Bodenteich i col·laboradors (1992) van descriure un microsatèl·lit entre les posicions 514 i 523 de la regió codificant. És un microsatèl·lit curt format per repeticions de dinucleòtids CA. La seqüència de referència del mtDNA presenta cinc repeticions CA (Anderson i col·laboradors. 1981).

I.3. Deleció de 9 parells de bases

En la seqüència de referència (Anderson i col·laboradors. 1981) s'ha descrit dos còpies de una seqüència de 9 parells de bases (CCCCCTCTA) situada en la regió intergènica de la sub-unitat de la citocrom oxidasa i del RNA de transferència de la lisina. Aquesta seqüència es va utilitzar com marcador específic de les poblacions d'origen asiàtic (Wrischnik i col·laboradors. 1987; Hertzberg i col·laboradors. 1989; Shields i col·laboradors. 1992) on s'havia descrit la deleció d'una de les dues còpies de la seqüència. S'havia postulat un origen únic asiàtic de la deleció, però es va trobar també en Pigmeus (Vigilant i col·laboradors. 1991; Watson i col·laboradors. 1996), així com en les poblacions sub-saharianes Kikuyu, Turkana, i Yoruba (Watson i col·laboradors. 1996). També es va descriure una triplicació de la seqüència de 9 parells de bases en algunes poblacions africanes Fulbe i Hausa (Watson i col·laboradors. 1996). Les deleccions i les triplicacions de les seqüències es van identificar en individus pertanyien a diferents haplogrups. Per tant, la informació aportada per aquest polimorfisme s'ha de considerar conjuntament amb la informació proporcionada amb les seqüències de regió de control i/o els SNPs de la regió codificant, ja que l'anàlisi d'aquest marcador en solitari no aporta gaire informació degut al seu origen recurrent en diferents llinatges mitocondrials.

II. Tècniques utilitzades

Per cadascun d'aquests polimorfismes es va utilitzar la tècnica més adequada en el moment de l'anàlisi. En aquest apartat, es comentaran els diferents mètodes utilitzats per el tractament de la mostra un cop extret el DNA.

II.1. Amplificació per seqüenciació directa

En el primer estudi realitzat, es va aplicar l'amplificació directa del segment hipervariable I amb els primers L15997 i H16401 (Vigilant i col·laboradors. 1989). Però per la resta dels treballs, es va amplificar tota la regió de control mitjançant els primers L15997 i H408 per tal de tenir l'opció de seqüenciar també la regió hipervariable II. Els primers utilitzats per l'amplificació i la seqüenciació estan descrits en la Taula 2. Les condicions d'amplificació descrites en la Taula 2 es van aplicar als dos tipus d'amplificació (regió hipervariable I i II). Es va comprovar cada amplificació mitjançant un gel d'agarosa al 2%.

Primers		Seqüència
HVS-I	L15997	5'-CAC CAT TAG CAC CCA AAG CT-3'
	H16401	5'-TGA TTT CAC GGA GGA TGG TG-3'
HVS-II	L29	5'-GGT CTA TCA CCC TAT TAA CCA C- 3'
	H408	5'-CTG TTA AAA GGTG CAT ACC GCC A-3'
Condicions d'amplificació		94°C 2' 30 cicles: 94°C 1'; 58°C 1'; 72°C 1' 72°C 5'

Taula 2: Conjunt de primers (nom i seqüència) utilitzats per l'amplificació i la reacció de seqüenciació de la regió de control, i descripció de les condicions de PCR per l'amplificació de la regió de control.

II.2.Purificació de les mostres

L'etapa de purificació abans de procedir a la seqüenciació és essencial per tal d'eliminar els dideoxinucleòtids no incorporats, els primers restants de l'amplificació, els productes de PCR parcialment amplificats, i tots aquells elements susceptibles d'interferir en la reacció de seqüenciació. Es van utilitzar dos mètodes per purificar les mostres depenent de la seva disponibilitat en el moment de la purificació:

- La purificació per precipitació química via l'ús del kit Gene Clean (BIO101)
- La purificació amb columnes GFX (Amersham Bioscience)

II.3 Seqüenciació automàtica

Les reaccions de seqüenciació es van fer sistemàticament a partir de la cadena lleugera mitjançant els primers L15997 i L29 (Taula 2) per la seqüenciació respectiva dels segments hipervariables I i II. Es van executar amb el kit de seqüenciació Big Dye Terminator (versió 3.0; Applied Biosystems) que utilitza ddNTPs marcats amb fluorocroms de diferents absorbància, i els productes de la reacció de seqüència van ser carregats en el seqüenciador A.B.I PRISM 3100 (Applied Biosystems).

El problema més freqüent és la heteroplàsmia de longitud. En la regió de control existeixen tractes de policitosines (poli Cs) que presenten una taxa de mutació elevada i són més propenses a presentar heteroplàsmia. Trobem tractes de poli Cs en el HVS-I entre les posicions 16184 i 16193, interromputs per una T en la posició 16189, i en el HVS-II, entre les posicions 303 i 315, interromputs per una T en la posició 310. Quan apareix una substitució en la posició 16189 i/o 310, s'obté una sèrie de 10 a 15

citòsines, i la seqüència que segueix a continuació no es pot caracteritzar per seqüenciació directa degut a un solapament de diversos patrons de lectures. En casos com aquest es va seqüenciar l'altra cadena per tal de caracteritzar el segment de seqüència que mancava.

II.4. SNaPShot

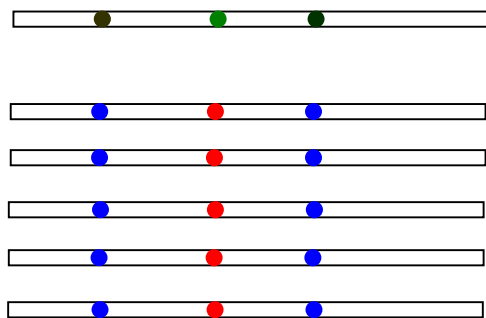
El SNaPshot és una de les nombroses estratègies disponibles per tipar SNPs, però particularment adequada en el cas del mtDNA on es poden multiplexar els SNPs a partir d'una única reacció de PCR, ja que en un fragment relativament curt de DNA s'acumulen gran quantitat de SNPs. És una tècnica automatitzada, ràpida i fàcil d'aplicar, basada en la extensió de primer en una sola base, que permet, a partir del producte de DNA amplificat, de caracteritzar de manera individual els SNPs d'interès (figura 11).

1) Preparació de la mostra

Doble cadena de DNA amb els SNPs d'interès



Amplificació

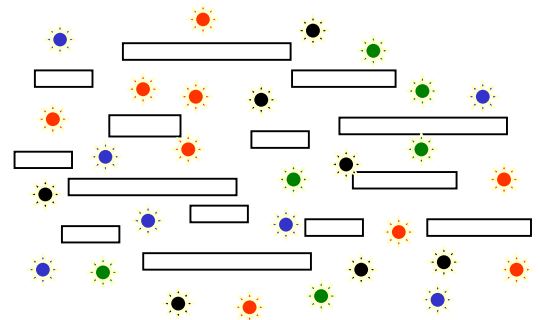


Eliminació dels primers i dNTPs no incorporats

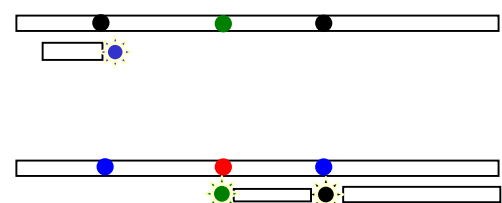
Tractament enzimàtic (digestió amb Exo/Sap)

2) Reacció d'extensió

● ddATP
● ddCTP
● ddGTP
● ddTTP



Extensió de primer



Tractament enzimàtic (digestió amb Exo/Sap)

3) Anàlisi de dades

Electroforesi en el ABI 377 ó 3100

Anàlisi de les dades amb el software GeneScan

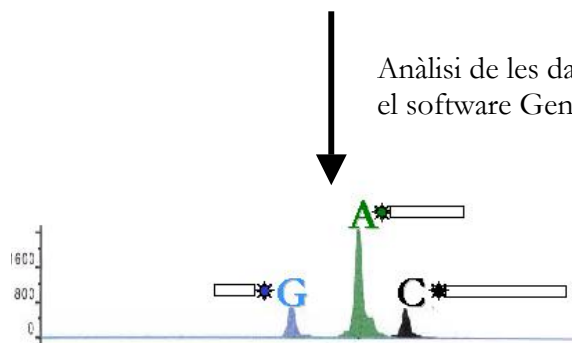


Figura 11: Descripció del protocol de SNaPshot amb tres SNPs qualsevol del mtDNA.

En el present treball, es van tipar quatre SNPs situats fora de la regió de control, que ens van permetre classificar sense cap ambigüitat les seqüències en haplogrups. Aquests SNPs s'han tipat només en els casos on la informació proporcionada per la regió de control no era suficient per caracteritzar el haplogrup de la seqüència. En una sola PCR es va amplificar la regió que contenia els quatre SNPs mitjançant els primers i les condicions de PCR descrits en la Taula 3, seguit de l'aplicació de la tècnica de SNaPshot per tipar els SNPs d'interès. El locus, la seqüència i longitud dels primers, el polimorfisme i l'haplogrup associat són descrits a la Taula 3.

Primers	Seqüència
L10373	5'-CCCTAAGTCTGGCCTATGAG-3'
H12744	5'-CGATGAACAGTTGGAATAGG-3'
Condicions de PCR	94°C: 5 35 cicles: 94°C 30"; 55°C 30"; 72°C 30" 72°C 5'

Taula 3: Primers (noms i seqüències) i condició de PCR utilitzats per l'amplificació prèvia a l'aplicació de la tècnica SNaPshot.

Locus	Seqüència (5'-3')	Longitud	Al 1els	Al 1el→Haplogrup
H10400X	TGTTTAAACTATATACCAATTC	22	G/A	A → M
L10873X	TTTT'TTTTCCACAGCCTAATTATTAGCATCATCCC	36	C/T	C → L i M
L12308X	CAGCTATCCATTGGTCTTAGGCCCAA	27	A/G	G → U
L12705X	AACATTAATCAGTTCTTCAAATATCTACTCAT	32	C/T	T → L, M i N

Taula 4: Locus i primers utilitzats en el protocol de SNaPshot.

Els productes amplificats van ser correguts en el seqüenciador automàtic 3100™ (Applied Biosystems). La combinació de primers de diferents grandàries permet la

detecció de més d'un SNP en cada carril del seqüenciador automàtic. L'assignació dels al·lels es va fer mitjançant el software GeneScan (Figura 11).

II.5. Microsatèl·lit i deleció de 9 parell de bases:

El microsatèl·lit i la seqüència amb les dues repeticions de 9 parell de bases (CCCCCTCTA) van ser amplificats conjuntament amb els primers i les condicions de PCR descrits a la Taula 5:

Polimorfisme	Primer	Seqüència
Deleció de 9 parell de bases	L8196	5'-ACAGTTTTCATGCCCATGGTC-3'
	H8297	5'ATGCTAAGTTAGCCTTACAG-3'
Microsatèl·lit	L483	5'-ACTCCCATACTACTAATCTC-3'
	H575	GAGGTAAGCTACATAAACTG-3'
Condicions de PCR	94°C: 5 35 cicles: 94°C 30"; 55°C 30"; 72°C 30" 72°C 5'	

Taula 5: Noms i seqüència dels primers i condicions de PCR per l'amplificació conjunta del marcador de 9 parell de bases i del microsatèl·lit del mtDNA.

Els primers L8196 i L483 van ser marcats respectivament amb els fluorocroms JOE i FAM que donen una fluorescència verda i blava i permeten detectar els dos marcadors en un mateix carril. Els estàndards ABI GS350 Rox o ABI GS500 TAMRA van ser utilitzats com marcadors interns de carril. Els productes amplificats es van fer córrer en el seqüenciador automàtic 377 i el anàlisi es va dur a terme amb el software GeneScan.

III. Poblacions estudiades

En la present tesi s'han estudiat 22 poblacions de diferents continents: vuit de l'Àfrica del nord oest (Algerians, Àrabs del Marroc, Berbers del Marroc i de Tunísia, Saharauis, Tunisians), dos de la Península Ibèrica (Andalusos i Catalans), un de l'Àfrica sub-sahariana (Angola), una de l'oceà Índic (Illa de la Reunió), i dotze de l'Àsia Central (Bukharan Arabs, Crimean Tatars, Iranians, Dungans, Karakalpaks, Kazaks, Khoremian Uzbeks, Kyrgyz, Tajiks, Turkmen, Uighurs i Uzbeks). L'entorn geogràfic i els esdeveniments històrics han condicionat els fenòmens demogràfics (expansió, migració, barreja, aïllament, deriva) que han patit aquestes poblacions i que es reflecteixen en la seva composició genètica. L'anàlisi de la diversitat genètica d'aquestes poblacions mitjançant el mtDNA ens permet resoldre la seva història demogràfica i contrastar hipòtesis proporcionades per altres marcadors i altres disciplines (lingüística, arqueologia).

El DNA de les poblacions estudiades va ser extret a partir de mostres sanguínies. Totes les mostres de DNA ja estaven disponibles al laboratori quan es va iniciar la present tesi. Per les poblacions del Nord d'Àfrica (menys els Berber de Tunísia) l'extracció de DNA es va fer en el laboratori mitjançant el mètode estàndard d'extracció amb fenol i cloroform. L'obtenció de la resta de les mostres va ser el resultat de col·laboracions amb diferents laboratoris. També es van extreure a partir de la literatura o bé a través de comunicacions personals seqüències de la regió de control del mtDNA de poblacions que es van revelar útils o essencials per dur a terme una anàlisi completa. A partir de les 22 poblacions citades anteriorment, es van dur a terme cinc estudis en quatre grans àrees geogràfiques de diferent interès.

III.1. Les poblacions de l'oest del Mediterrani

Aquesta anàlisi inclou onze poblacions de l'Àfrica del Nord oest, deu poblacions de la Península Ibèrica i cinc d'Itàlia (figura 12). Per la seva posició geogràfica a cada banda del Mediterrani i per el seu passat històric, les poblacions africanes i europees de l'oest del Mediterrani presenten característiques úniques, afinitats i diferències que han contribuït a modelar la seva composició genètica actual. Les poblacions del Nord i del Sud del Mediterrani tenen un àmbit geogràfic diferent que ha influenciat de manera diferent la seva estructura poblacional, però estan unides per la història ja que comparteixen esdeveniments històrics que van implicar moviments de poblacions humanes a través de les dues bandes del Mediterrani.



Figura 12: Localització geogràfica de les poblacions de l'oest del Mediterrani estudiades. Les rodones representen les mostres seqüenciades; els quadrats, les mostres seqüenciades i agafades de la literatura; i els triangles, les mostres agafades de la literatura.

La població autòctona del Nord d'Àfrica, els berbers, (Camps 1998) va experimentar onades successives d'invasions a partir del segon mil·lenni aC. En la antiguitat, hom els coneix sota la denominació de *númidas, mauros, getulos, libios* (Camps 1998). Els primers contactes nord-sud a través del Mediterrani es van establir primer amb l'arribada dels fenicis, fundadors de Cartago (814 aC) que van establir una ruta mercantil al llarg de la costa Nord africana pel tràfic de plata i d'estany amb la Península Ibèrica (Newman 1995). Al voltant del segle VI aC, Cartago es va convertir en el centre econòmic i polític més important de l'antiguitat. El segon gran moviment poblacional a través del Mediterrani va ser amb els romans a partir de 146 aC amb la caiguda de Cartago. L'imperi romà, que s'estenia des d'Egipte fins al Marroc, va portar el seu interès sobretot a la part oriental del Mediterrani (Egipte, Líbia, Tunísia). El tercer gran contacte nord-sud del mar Mediterrani va ser protagonitzat pels Vàndals, una tribu alemanya que va arribar al Nord d'Àfrica a través de l'estret de Gibraltar. El seu domini va durar fins a l'any 534 dC quan els bizantins va restaurar l'administració romana. Al final del segle VII, l'arribada dels Àrabs a l'Àfrica del Nord va provocar profunds canvis culturals. Es va assistir a una ràpida assimilació de la religió islàmica a tota la costa de l'Àfrica del Nord. Però l'impacte cultural no va ser acompanyat per un impacte demogràfic ja que les tropes àrabs eren un quants milers en front a milions de berbers autòctons (Camps, 1998). Els berbers convertits a l'Islam, s'incorporaven a les tropes àrabs i sota la direcció dels Àrabs van arribar a la Península Ibèrica al segle VIII, on van imposar la seva cultura. L'arabització i islamització de gran part dels berbers es va completar i confirmar al segle XI amb la invasió dels Beduïns. Els àrabs beduïns, que eren més nombrosos que els primers invasors, es van expandir demogràficament provocant la fugida dels berberòfons cap a les regions muntanyoses. En pocs segles, l'Àfrica del Nord poblada pels berbers en part romanitzats i cristianitzats es va transformar en un conjunt de països completament arabitzats i musulmans (Camps, 1998). La diversitat ètnica africana es va ampliar amb la presència de turcs otomans

durant els segles XVI-XIX (Newman, 1995). A partir del segle XV fins al segle XX, els invasors del Nord-oest d'Àfrica tenen el seu origen a Europa: portuguesos i espanyols al Marroc, francesos al Marroc, Algèria i Tunísia, i Italians a Líbia.

Històricament, la Península Ibèrica també va conèixer la influència de diferents cultures europees i africanes. Està vinculada a l'Àfrica del Nord des del segle VII dC amb la invasió islàmica que va aportar innovacions tecnològiques i que va tenir una predominant influència al sud de la Península. També s'ha proposat un origen comú entre els bascos i els Berbers en base als estudis del sistema HLA (Arnaiz-Villena i col·laboradors. 1995; 1997); que no van ser replicats amb la publicació noves dades i el reanàlisi de les anteriors (Comas i col·laboradors. 1998). Els estudis de marcadors clàssics van descriure els bascos com una població aïllada genèticament del conjunt genèticament homogeni format pels europeus (Calafell i Bertranpetit 1994).

L'àmbit geogràfic també va tenir un paper important en l'estructura de la població nord africana ja que va limitar els moviments de poblacions humanes. Malgrat la seva afiliació al continent africà, les poblacions del Nord d'Àfrica formen un conjunt totalment diferent de les poblacions sub-saharianes. El poblament del Nord oest d'Àfrica ha estat condicionat per diferents barreres geogràfiques: el gran desert del Sàhara al Sud, les muntanyes del Rif i de l'Atlas al Marroc, i el mar Mediterrani al Nord.

L'interès per saber com de permeables han estat aquestes barreres geogràfiques als moviments poblacionals i quin va ser l'impacte demogràfic de les diferents influències culturals, ha generat una sèrie d'estudis amb diversos marcadors genètics: marcadors clàssics, microsatèl·lits autosòmics, polimorfismes d'insercions Alu, polimorfismes del cromosoma Y, i llinatges del DNA mitocondrial. Molts dels estudis basats en l'anàlisi del mtDNA s'han centrat en la Península Ibèrica (Bertranpetit i col·laboradors. 1995; Corte-Real i col·laboradors. 1996; Salas i col·laboradors. 1998; Pereira i col·laboradors. 2000) o bé sobre l'Àfrica del Nord (Rando i col·laboradors. 1998; Brakez i col·laboradors. 2001), però fins a la present tesi no s'havia analitzat com

un conjunt les poblacions de l'oest del Mediterrani mitjançant les seqüències de mtDNA.

III.1.a. Els berbers de Tunísia

Un cas particular ha estat l'anàlisi dels berbers de Tunísia. A Tunísia, com a la resta del Nord d'Àfrica, la població autòctona eren els berbers que bàsicament es podien dividir entre els sedentaris i els nòmades. Els natius van patir onades successives d'invasions que es van iniciar amb els fenicis a l'antiguitat. Els regnes berbers que es van succeir al Nord d'Àfrica van anar canviant amb l'arribada i l'establiment de cada nou invasor. Fins a la invasió en massa dels Beduïns, els berbers no formaven una estructura social unida sinó que hi havia molts regnes berbers dispersos a tot el Nord d'Àfrica. Tot i això, els berbers conservaven un substrat cultural comú. Amb la conquesta àrab al segle VII, però sobretot amb l'arribada dels beduïns, la situació cultural i demogràfica dels berbers va canviar. La majoria dels berbers van assimilar la llengua àrab i a la religió islàmica. A partir d'aquesta reforma religiosa es va crear els dos regnes berbers islàmics més importants: el dels Almoràvids (1056-1147) que s'estenia en tota l'Àfrica del Nord i en l'Espanya musulmana (al-Andalus); i el dels Almohades (1121-1269) que va unificar tota l'Àfrica del Nord en un sol domini. La transformació etnosociològica dels Berbers va ser ràpida i efectiva. En dos segles es va imposar la llengua àrab i la religió musulmana, però algunes tribus berbers es va resistir a abandonar la seva llengua materna i la seva cultura. L'expansió demogràfica dels àrabs beduïns posterior a la colonització va forçar els berberòfons a fugir a les zones muntanyoses de la regió del país. L'arribada per l'est i la penetració a l'interior de Tunísia dels àrabs va provocar el desplaçament de les tribus berbers al sud del país que van trobar refugi en pobles de muntanya aïllats de les ciutats, per poder escapar de la dominació àrab, i de les posteriors invasions. Els turcs otomans, els

espanyols i els francesos van ocupar el territori tunisià entre els segles XVI i XX, i van contribuir al substrat genètic de la població tunisiana. Els berbers de Tunísia es distingeixen de la resta de la població únicament per la seva llengua, el Chelcha. Els berbers representen 1% de la població de Tunísia i estan repartits bàsicament en quatre pobles del Sud de Tunísia: Sened, Matmata, Chenini i Douiret; i en l'illa de Jerba. L'objectiu de l'anàlisi de les poblacions berbers del Sud de Tunísia era determinar el grau de heterogeneïtat entre aquests grups de berbers, comparar la seva composició genètica amb poblacions àrabs i berbers del Nord d'Àfrica, i determinar la contribució de les poblacions veïnes al substrat genètic dels berbers de Tunísia.

III.2. La població d'Angola

Amb aquesta població canviem de context per situar-nos a l'Àfrica subsahariana (figura 13).



Figura 13: Localització geogràfica d'Angola.

Angola entra a la història occidental a partir de 1400 amb la colonització europea. La història de la majoria dels països africans després del contacte europeu va ser força estudiada i documentada però no existeix cap document escrit de la història pre-colonial d'Angola. Els primers habitants d'Angola són els pobles Khoisànids. El terme Khoisan designa tant el poble com la seva llengua. El Khoisan es caracteritza per sons secs i es troba a la part sud i est de l'Àfrica: Botswana, Namíbia, Àfrica del Sud, el sud d'Angola, i el nord de Tanzània. Segons la classificació lingüística de Ruhlen (1987) basada en el treball de Greenberg (1963), el Khoisan és la més antiga de les famílies lingüístiques presents a l'Àfrica. Segon Rhulen (1987), les llengües africanes es classifiquen en quatre gran famílies: l'Afroasiàtica que es parla al Nord d'Àfrica i a l'Orient Mitjà, la Nilo-sahariana que és la segona més antiga, la Niger-Kordofanian que

s'estén a l'oest, centre i sud de l'Àfrica, i la Khoisan que té una distribució bàsicament sud-africana. Dins la família Niger-Kordofanian es classifiquen les llengües Bantu que van tenir un paper major en la transformació lingüística, cultural i demogràfica de l'Àfrica sub-sahariana. El cor de les llengües Bantu es troba a l'est del Níger i a l'oest del Camerun. Al final del Neolític (± 5000 aC) comença l'expansió Bantu, una de les migracions més importants que es va experimentar a l'Àfrica. Es va produir cap al sud del continent en dues direccions, l'est i l'oest (Newman 1995). Al voltants de l'any 3000 aC, els Bantu que havien agafat la direcció est van arribar a Uganda on van formar un nou nucli poblacional. A partir d'aquest segon nucli, els Bantu de l'est van tornar a iniciar el desplaçament cap al sud. A la mateixa època (3500-3000aC), la migració oest es va dividir en dos grups: un va vorejar la costa atlàntica, i l'altre va penetrar dins la selva equatorial on es va trobar amb la població autòctona, els Khoisan. El domini de l'agricultura i de les tecnologies basades en el ferro (cap a l'any 2000 aC) van conferir un avantatge decisiu als Bantu respecte les poblacions autòctones, de tal manera que van ocupar progressivament tota l'Àfrica central fins les sabanes del Sud (~ 1700 aC). Les dues expansions Bantu, est i oest, van coincidir probablement varies vegades en el temps, però els punts de trobada no estan ben caracteritzats. No es sap amb certesa si les dues expansions van convergir al sud de la selva tropical o bé si els contactes es van establir més tard en les zones de sabanes de més fàcil accés, o bé si es van produir ambdós esdeveniments. Les poblacions autòctones Khoisan eren caçadores-recol·lectores i van ser o bé assimilades per els Bantu o bé es van desplaçar fins al sud del continent africà. A partir de l'any 1400, els portuguesos van desenvolupar el tràfic d'esclaus a la costa atlàntica africana. Compraven esclaus als caps de les tribus africanes i posteriorment eren enviats a São Tomé i Brasil per treballar en les plantacions de canyes de sucre. Angola va proporcionar aproximadament quatre milions de persones (Thomas i col·laboradors. 1997) per al tràfic d'esclaus. D'aquests quatre milions, dos

milions d'esclaus van ser destinats a les Amèriques, i Brasil va rebre més del 50% dels esclaus angolesos.

Angola es una peça essencial en el paisatge genètic africà tant en temps pre-colonials, ja que representa la punta final de l'expansió Bantu de l'oest, com en la història moderna per haver proporcionat la major quantitat d'esclaus a Brasil i haver contribuït d'aquesta manera a la formació de la seva població. Sent el bressol de la humanitat, Àfrica va ser la diana, des del principi dels anys 90 fins a l'actualitat, dels estudis genètics mitjançant el cromosoma Y i del mtDNA. Moltes poblacions africanes van ser caracteritzades genèticament i es va poder definir marcadors Khoisan i Bantu tant per el cromosoma Y (Scozzari i col·laboradors. 1999; Underhill i col·laboradors. 2000; 2001; Thomas i col·laboradors. 2000) com pel mtDNA (Bandelt i col·laboradors. 1995; Bandelt and Foster 1999; Chen i col·laboradors. 1995; Salas i col·laboradors. 2002; Soodayll i col·laboradors. 1996; Watson i col·laboradors. 1997). L'estudi de Salas i col·laboradors aclareix el complex paisatge genètic africà descrivint quins són els llinatges mitocondrials que componen les grans regions africanes (Nord, Oest, Est, Centre, Sud-est i sud Àfrica) i com es relacionen genèticament mitjançant networks filogenètics. La única regió geogràfica mancant en aquest estudi és el Sud-oest de l'Àfrica representada per Angola. Amb l'estudi de la població angolesa en la present tesi es pot des d'ara completar el paisatge genètic africà revelat a partir de seqüències del mtDNA i aportar més informació sobre la contribució dels esclau angolesos a la composició genètica de la població brasilera.

III.3. Les poblacions d'Àsia Central.

L'Àsia Central és un ampli territori geogràficament definit per les grans estepes asiàtiques al Nord, la serralada de l'Hindu Kush al Sud, la Xina a l'est i el Caspi a l'oest. És una regió que ofereix una gran varietat de paisatges amb zones d'alta muntanya, extens deserts, desfavorables a la colonització i als desplaçaments de poblacions humanes, que contrasten amb rius i estepes favorables a la vida pastoral i a la cria de bestiar. El poblament de l'Àsia Central ha estat condicionat pel rigor de l'entorn geogràfic desfavorable a l'establiment i el desenvolupament de les poblacions humanes. Però la presència de vestigis del Paleolític mitjà (una resta Neandertal trobada a l'actual Uzbekistà, delimita l'extrem oriental de la distribució d'aquesta espècie) i superior, i del Mesolític, són prova de l'assentament antic i constant de les poblacions humanes en aquesta regió. El Neolític va arribar primer al Sud-oest de l'Àsia Central, relacionat amb les civilitzacions Mesopotàmiques i Iranianes, i un mil·lenni més tard al Nord-est associat amb els moviments poblacionals a les grans estepes asiàtiques. Però aquesta diferència cultural es va esborrar durant l'edat de Bronze. (Bowles, 1977). Els records històrics deixats pels grecs i els xinesos descriuen els primers habitants de l'Àsia Central, els Scythians (VII aC) i els Sarmatians, de llengua indo-europea, amb trets morfològics europeus. L'absència de fronteres amb l'Oest ha facilitat i consolidat els moviments de poblacions de l'Oest cap a l'Àsia Central i inversament. L'exemple més rellevant és l'exemple de la Ruta de la Seda que va connectar l'Est i l'Oest del continent durant diversos segles. Als pobles indo-europeus els van succeir els pobles túrquics, els xinesos, i els russos.



Els estudis d'antropologia clàssica mostren que les poblacions de l'Àsia Central presenten trets morfològics que són una barreja entre els de les poblacions de l'Est i de l'Oest (Bowles, 1977). Les dades genètiques clàssiques indiquen una posició intermèdia de l'Àsia Central entre l'Orient Mitjà i l'Àsia de l'Est (Cavalli-Sforza i col·laboradors. 1994). Les dades proporcionades pel mtDNA (Comas i col·laboradors. 1998), i els microsatèl·lits del cromosoma Y mostren que els habitants d'aquesta regió són el resultat d'una barreja genètica molt antiga que va produir una gran diversitat genètica. Els llinatges del cromosoma Y apunten a l'existència d'un gradient est-oest interromput per esdeveniments poblacionals específics, i a una diversitat genètica heterogènia (Zerjal i col·laboradors. 2002).

L'objectiu de l'anàlisi de 12 poblacions de l'Àsia Central a partir de seqüències de la regió de control i de SNPs de la regió codificant del mtDNA era aportar una millor resolució als llinatges mitocondrials definits prèviament (Comas i col·laboradors. 1998), per tal d'obtenir una descripció completa de la diversitat a l'Àsia Central i determinar si l'Àsia Central ha estat una zona de contacte entre les poblacions de l'Est i de l'Oest o bé si es una zona de barreja genètica, on les poblacions de l'est van reemplaçar parcialment les poblacions de l'Oest.

III.4. La població de l'illa de la Reunió

La població humana actual de l'illa de la Reunió, situada a l'oceà Índic a 200 Km a l'est de Madagascar i departament francès d'ultramar des de 1946, és el resultat d'una barreja de pobles deguda a les successives onades migratòries que ha rebut: europeus, malgaixos, africans, indis, xinesos formen part de la població barrejada actual de l'illa (figura 14). L'illa va estar deshabitada fins a mitjans del segle XVI quan navegants portuguesos i àrabs va fer estades temporals. Cap a mitjans del segle XVII els francesos prenen possessió de l'illa i s'hi instal·len els primers colons europeus, malgaixos i indis de la colònia portuguesa de Goa Posteriorment l'illa incrementarà el nombre d'habitants gràcies a l'esclavatge originari de Madagascar i també de l'Àfrica continental. Un cop abolit l'esclavatge, multitud de treballadors d'origen indi i xinès s'instal·len a l'illa per tal de treballar a les plantacions.

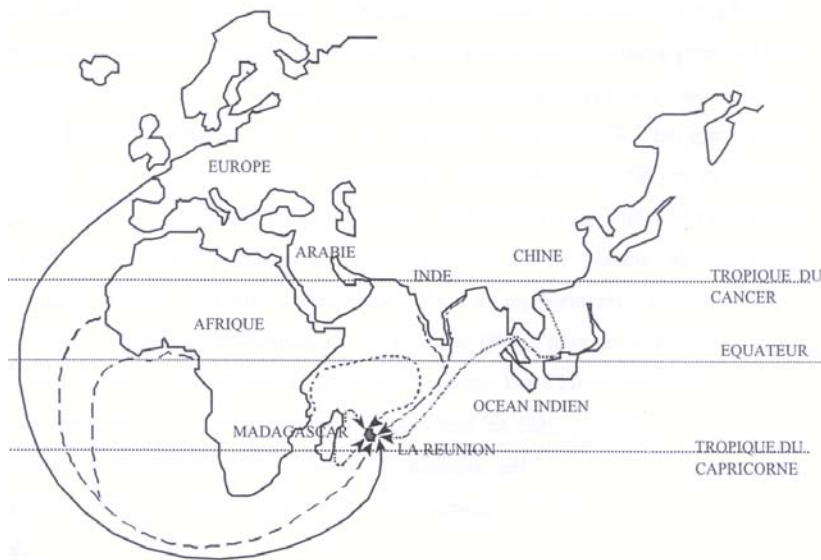


Figura 14: Origen de la població de l'illa de la Reunió (Figura extreta de Césari 1999).

L'objectiu de l'anàlisi genètica d'aquesta població és el d'intentar quantificar, mitjançant l'estudi del mtDNA i el cromosoma Y, l'aportació genètica de cadascun del pobles que van poblar l'illa i que es van barrejar per donar lloc a l'actual població reunionesa. Aquesta anàlisi té l'interès afegit de que part de la seva població prové de la veïna illa de Madagascar, els orígens de la qual són especialment interessants. Es postula que els pobladors inicials de Madagascar eren originaris d'Indonèsia i van arribar-hi mitjançant navegació a través de l'oceà Índic. Si aquesta hipòtesi fos certa, no seria estrany trobar llinatges mitocondrials del sud-est asiàtic a l'illa de la Reunió provinents de la veïna Madagascar.

IV. Tractament estadístic

Deixant de banda els paràmetres estadístics de diversitat genètica intra-poblacional, com la diversitat nucleotídica, la diversitat de seqüències, i la mitjana de *pairwise differences*, que s'han calculat en la present tesi amb el programa Arlequin 2000 (Schneider i col·laboradors. 1996), descriurem en aquest apartat els diferents mètodes estadístics utilitzats a partir de dades genètiques del mtDNA (matriu de distàncies genètiques i freqüències absolutes o relatives dels haplogrups) per establir les relacions filogenètiques entre les poblacions a partir de les dades .

IV.1. Anàlisi molecular de la variància: AMOVA

L'anàlisi molecular de la variància (AMOVA) permet estimar la diferenciació genètica dintre de poblacions, entre poblacions d'un mateix grup i entre grups, i testar les hipòtesis de la diferenciació (Excoffier i col·laboradors. 1992). La significació dels components de la variància s'obté a partir de tests de permutació no paramètrics. Aquest mètode s'aplica directament sobre diferents tipus de dades moleculars, com les freqüències al·lèliques, el contingut al·lèlic dels haplotips, i les seqüències de DNA., oferint d'aquesta manera un gran flexibilitat.

IV.2. Anàlisi espacial de la variància molecular: SAMOVA

El principi del SAMOVA consisteix en definir grups de poblacions geogràficament adjacents i genèticament homogènies, és a dir grups de poblacions que maximitzen la part de la variància genètica total explicada (Dupanloup i col·laboradors. 2002). El mètode es basa en un procediment de permutació d'agrupació. Indirectament, el SAMOVA permet identificar les barreres genètiques de la regió

estudiada. L'aplicació d'aquest mètode es fa a partir de dades haplotípiques i genotípiques i no fa cap assumptió sobre el equilibri Hardy-Weinberg dintre de les poblacions, ni sobre el desequilibri de lligament entre loci.

IV.3. Anàlisi de components principals

L'anàlisi de components principals és un mètode purament estadístic que pot aplicar-se a dades genètiques. Aquest mètode consisteix en simplificar la complexitat de les dades reduint el nombre de dimensions de l'espai en què es treballa, però amb una mínima pèrdua de la informació. En el cas d'aquest estudi, el mètode de components principals s'aplica sobre les freqüències relatives dels haplogrups del mtDNA trobats per cada població. Els components principals obtinguts no estan correlacionats i, per tant, es poden analitzar per separat. La quantitat de variabilitat explicada per cada component principal disminueix a mesura que s'avança en els diferents components: la primera component acumula el màxim de variació, el segon explica una quantitat de variació menor, etc... Els valors de components principals es poden representar en mapes sintètics on els principals patrons espacials estan resumits. Es dona una idea relativa de quins haplogrups provoquen les principals diferències entre les poblacions.

IV.4. Anàlisi de coordenades principals

L'anàlisi de coordenades principals es basa en el mateix principi que es va explicar en el anàlisi de components principals, però amb la diferència que s'aplica sobre la matriu de distàncies genètiques. La realització dels dos tipus d'anàlisi permet contrastar la informació aportada pels dos tipus de dades (matriu de distàncies genètiques i freqüències haplotípiques) a nivell poblacional.

IV.5. Anàlisi de correspondències.

L'anàlisi de correspondències és un mètode qualitatiu basat en el mateix principi que el anàlisi de components principals amb la diferència que s'aplica sobre les freqüències absolutes dels haplogrups i que permet identificar directament en el mapa sintètic l'haplogrup responsable de la distribució en el espai de les poblacions.

IV.6. Anàlisi de *multidimensional scaling* o MDS:

En el nostre cas, el *multidimensional scaling* (MDS) s'aplica sobre la matriu de distàncies genètiques però es pot utilitzar qualsevol tipus de distància o de matriu similar. El MDS no és un mètode exacte sinó una manera de reorganitzar les poblacions en l'espai per obtenir la configuració que millor s'aproximi a la matriu de distàncies genètiques. Per un procés de permutacions, el programa canvia la configuració de les poblacions en l'espai per definir el nombre adequat de dimensions i identificar la configuració que millor reproduceixi la matriu de distàncies genètiques. La mesura de estrès avalua com de bé la nova configuració reproduceix la matriu de distàncies genètiques.

IV.7. Els *networks* filogenètics

La construcció de *networks* filogenètics en el cas del mtDNA es fa a partir de les dades dels polimorfismes de restricció de longitud de fragments (RFLPs) o bé a partir de la variació de seqüència de la regió de control mitjançant el "*reduced-median algorithm*" (Bandelt i col·laboradors. 1995), desenvolupat en el programa network versió 3.0. El *network* representa la variabilitat genètica d'un haplogrup present en una o diverses poblacions. És un diagrama filogenètic representat per un gràfic format de

cercles (o nusos) connectats per línies. Les línies estan associades amb les posicions nucleotídiques on s'ha produït un canvi. La llargada de les línies és proporcional al nombre de canvis produïts. Cada cercle està associat amb un haplotip i per inferència els individus portadors d'aquest haplogrup. La mida del cercle es proporcional al nombre d'individus. A partir del *network* filogenètic, es pot estimar l'edat del haplogrup representat (Morral i col·laboradors. 1994; Saillard i col·laboradors. 2000). El mètode de datació es basa en considerar el nombre de mutacions acumulades des de la seqüència ancestral com una funció lineal de la taxa de mutació i del temps.

IV.8. Anàlisi d'*admixture*

En el cas de la formació d'una població híbrida a partir de dues o més poblacions parentals genèticament diferenciades per aïllament geogràfic o ecològic, es pot distingir l'aportació genètica de cadascuna de les poblacions parentals al *pool* genètic de la població híbrid. L'anàlisi d'*admixture* mitjançant el program *admix 2.0* (Dupanloup i Bertorelle 2001) permet estimar la proporció de variabilitat genètica aportada per cadascuna de les poblacions parentals tenint en compte les diferències de freqüències al·lèliques i la quantitat de diferenciació molecular entre els al·lèls. Aquest mètode s'aplica sobre les freqüències dels haplogrups del mtDNA.

~ RESULTATS ~

~ CAPÍTULO I ~

Genética e historia de las poblaciones del Norte de África y de la península Ibérica

E. Bosch, F. Calafell, S. Plaza, A. Perez-Lezaun, D. Comas,
J. Bertranpetit

Investigación y Ciencia n.º 317, págs: 62-69; 2003

Genética e historia de las poblaciones del norte de África y la península Ibérica

El análisis genético ha revelado que los amplios intercambios culturales producidos entre el Magreb y la península Ibérica no conllevaron grandes intercambios de poblaciones

E. Bosch, F. Calafell, S. Plaza, A. Pérez-Lezaun, D. Comas, J. Bertranpetit

Las poblaciones humanas se componen de individuos genéticamente distintos entre sí. Del estudio de la variabilidad de nuestra especie se ocupa la genética de poblaciones aplicada a escala mundial. A dicha disciplina le corresponde exponer la magnitud y distribución de la variabilidad genética humana.

Dos personas cualesquiera, tomadas al azar, se distinguen, en promedio, en un 0,1% de las bases nucleotídicas que conforman su ADN. Expresado de otro modo, discrepan en seis millones de pares de bases. (El ADN humano consta de unos 3000 millones de pares de bases en cada una de las dos dotaciones haploides, una procedente del padre y otra de la madre.)

Si del individuo pasamos a las poblaciones, las diferencias observa-

das explican, a lo sumo, un 15% de la disparidad genética total; un 10% se debe a las diferencias entre grandes grupos continentales y el 5% restante a las diferencias entre poblaciones de un mismo continente. Aun siendo pequeñas, estas últimas diferencias tienen que ver con la historia de cada población. Podemos apoyarnos en la diversidad genética entre poblaciones para reconstruir la historia demográfica.

El acervo genético de las poblaciones actuales es el resultado de la interacción entre diversas fuerzas evolutivas. Dependen éstas, a su vez, de la historia demográfica de las poblaciones, de las características intrínsecas de las regiones genómicas estudiadas y de la interacción entre genoma y factores ambientales.

Las características intrínsecas de cada región del genoma remiten a sus tasas y patrones de mutación y recombinación, así como a su modo de herencia. Se trata de parámetros que la ciencia conoce con razonable precisión.

La interacción entre la variabilidad de cada gen y el ambiente (tomado en sentido amplio, incluida, pues, la interacción con otros genes) puede promover la selección natural. Es decir, unas variantes pueden mostrarse más eficientes y verse

privilegiadas por la selección, en tanto que otras pueden ser desfavorables. Lo observamos en la hemoglobina. Algunas variantes de esta proteína confieren resistencia a la malaria; la selección prima su presencia en zonas palúdicas. En consonancia con ello, el estudio de la variabilidad de la hemoglobina nos informará sobre la distribución de la malaria con mayor rigor que la historia de las poblaciones.

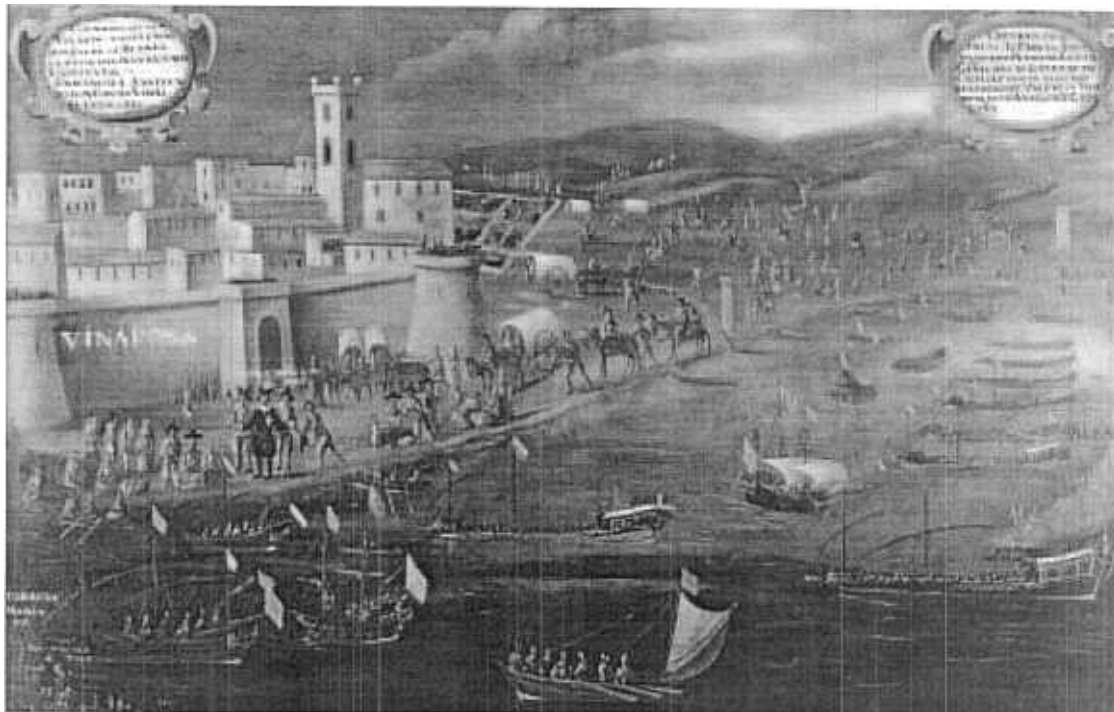
Conviene saber que sólo un 1,5% de la secuencia de ADN humano llega a expresarse, es decir, determina proteínas que se sintetizan y son objeto de selección natural. Por lo tanto, la mejor estrategia para conocer la historia de las poblaciones será la que se centre en la variabilidad presente en el 98,5% restante, cuyas probabilidades de verse afectada por la selección son mucho menores.

A partir de esta premisa podemos analizar dicha variación, explicar las diferencias genéticas neutras (polimorfismos) que encontramos entre los individuos de una población e interpretarlas en términos de historia de las poblaciones.

Disponemos de un amplio bagaje teórico, desarrollado desde los años cuarenta, gracias al cual, dada una historia demográfica, podemos predecir sus efectos sobre la diversidad genética. Podemos reconstruir

Los autores

E. BOSCH, F. CALAFELL, S. PLAZA, A. PEREZ-LEZAUN, D. COMAS y J. BERTRANPETIT han desarrollado la investigación que recoge su artículo en la Unidad de Biología Evolutiva de la Universidad Pompeu Fabra de Barcelona, cuya actividad se centra en el estudio de la diversidad genética en poblaciones humanas y en primates.



1. EMBARQUE DE LOS MORISCOS en el puerto de Vinaroz.

la historia demográfica a partir de la diversidad genética investigada en diversas regiones del genoma, que difieren en su velocidad de cambio y que permiten reconocer huellas genéticas a distintas profundidades de un tiempo pasado.

Deriva genética

¿Cómo influye la historia en la diversidad de las poblaciones? A través de dos mecanismos básicos: la deriva genética y el flujo génico. En la deriva se engloban todos los fenómenos de cambio genético aleatorio que se dan cuando una generación transmite sus genes a la siguiente. Así como hay apellidos que prosperan y otros que se pierden en razón del número de hijos varones procreados en cada generación, las variantes genéticas (o *alelos*) pueden también cambiar de frecuencia; en ambos casos se trata de fenómenos aleatorios. Tales oscilaciones serán tanto más intensas cuanto menor sea la población, por un simple efecto de muestreo. Las desviaciones respecto a la probabilidad teórica son mayores si reali-

zamos un número pequeño de ensayos, de la misma forma que al tirar una moneda al aire repetidas veces sólo se alcanza con seguridad la frecuencia esperada de 1/2 si se lanza muchísimas veces.

Se presenta un caso extremo de grandes cambios genéticos aleatorios, conocido por *efecto fundador*, cuando un grupo reducido de individuos establece una nueva población y se lleva consigo una muestra no necesariamente representativa de los genes de la población de origen. En la colonización sucesiva de las islas de la Polinesia, por ejemplo, se dio una secuencia clara de efectos fundadores; en el curso de la misma, un grupo limitado de individuos partía de una isla y se asentaba en la siguiente.

Las oscilaciones aleatorias de las frecuencias alélicas pueden llegar a la extinción de algunas de estas variantes. Puesto que dichas oscilaciones son más intensas en poblaciones pequeñas, se pierde variabilidad más fácilmente en éstas. A no ser que se dé una tasa de mutación extraordinaria, resulta muy poco probable que en las po-

blaciones pequeñas se regenere la variación perdida. Por lo tanto, al detectar una menor variabilidad genética en una población actual, podemos reconocer episodios de reducción de la población en el pasado (los llamados cuellos de botella), aunque la población actual se haya recuperado.

Además, habida cuenta de la naturaleza aleatoria de la deriva genética, las poblaciones pequeñas contiguas tenderán a diferir más entre sí que las mayores. Esta misma naturaleza aleatoria puede manifestarse de manera ligeramente diversa en regiones genómicas distintas. En consecuencia, los análisis basados en una sola región genómica pueden resultar poco fiables. Conviene siempre considerar la información procedente de un número razonable de regiones genómicas y extraer las tendencias medias.

Flujo génico

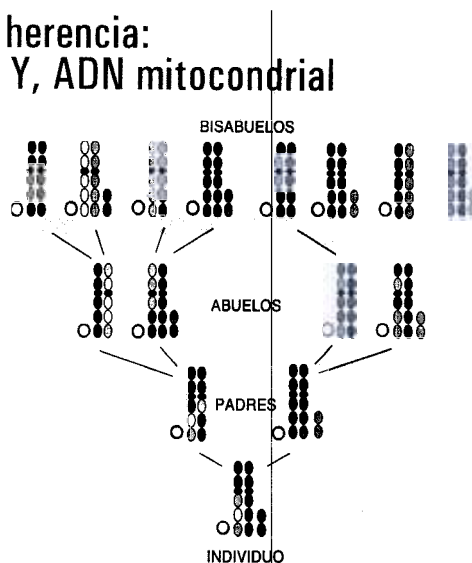
Las poblaciones que se hayan diferenciado, por deriva, en su composición genética y entren en contacto, pueden mezclarse y dar

Patrones de herencia: autosomas, cromosoma Y, ADN mitocondrial

Nuestro ADN se dispone en 23 pares de cromosomas. Cada miembro de un par es casi idéntico al otro en longitud y en la información que contiene; se trata de dos rasgos distintivos de cada par. Cada miembro de cada par de cromosomas nos viene de un progenitor; a cada uno de nuestros hijos le legaremos un solo miembro de cada pareja. Pero no es una transmisión fidedigna; en virtud del proceso de recombinación del material genético no heredamos el cromosoma original, sino una mezcla que contiene partes de cada miembro del par, tomadas al azar. Para ilustrarlo, la figura muestra un par de cromosomas que contiene fragmentos de distinta longitud de los cromosomas de los bisabuelos. Por eso resulta imposible predecir *a priori* de qué antepasado proviene un determinado fragmento de ADN autosómico.

Hay en el genoma dos regiones que presentan un patrón de herencia distinto. Nos referimos a los cromosomas sexuales y el ADN mitocondrial. A diferencia de los autosomas, los cromosomas sexuales (X e Y) son muy diferentes entre sí. El cromosoma Y determina la masculinidad a través de la acción de un único gen, SRY (*sex-determining region*); los cigotos con un cromosoma X y un cromosoma Y generan embriones masculinos, en tanto que los portadores de dos cromosomas X generan embriones femeninos. Por lo tanto, los varones heredan el cromosoma Y de su padre, que a su vez lo recibió del abuelo paterno, de la misma forma que se hereda el primer apellido. Lo vemos reflejado en la figura: de los cuatro bisabuelos varones, sólo el abuelo paterno del padre lega su cromosoma Y (*azul liso*) a su bisnieto.

Además de los cromosomas, que residen en el núcleo de las células, otros orgánulos contienen ADN. Se trata de las mitocondrias, que alojan decenas de



copias de una pequeña molécula circular de ADN. Este ADN mitocondrial (ADNmt) se hereda por vía materna: el ADNmt del embrión procede sólo del óvulo, porque el ADNmt del espermatozoide no llega a penetrar en el óvulo. Así, en la genealogía del ejemplo, el ADN mitocondrial del individuo (*magenta*) proviene de su madre, de su abuela materna, de la madre de ésta, y así sucesivamente.

Para entender esta figura sobre cromosomas y herencia, adviértase que las barras grandes representan autosomas (cromosomas no ligados al sexo), las pequeñas representan el cromosoma Y (cuya presencia denota un varón) y los círculos, el ADN mitocondrial. De abajo arriba se esquematiza un individuo, su madre y su padre, sus cuatro abuelos y sus ocho bisabuelos.

lugar a una población con características genéticas de las dos de partida. Este fenómeno de *flujo génico*, así se le llama, se debe a la migración. La propia migración en distancias cortas, habitual a través del matrimonio, puede promover, a largo plazo, el intercambio de genes a grandes distancias. Lo observamos, por ejemplo, en las poblaciones de Asia Central, que poseen características genéticas intermedias entre las de Europa y las de Asia Oriental; su peculiar constitución genética podría explicarse por su posición central y milenios de migraciones individuales de corto alcance.

La diferenciación entre poblaciones resultante de la deriva genética se acentúa con el paso del tiempo. Para medirla disponemos de un parámetro, la *distancia genética*, que indica el grado de diferenciación en-

tre pares de poblaciones para múltiples regiones del genoma. Si se trata de un conjunto de poblaciones, podemos representar su matriz de distancias genéticas mediante algoritmos; ofrecen éstos un paisaje genético que refleja las afinidades y diferencias dentro del conjunto poblacional. El paisaje compendia la historia de las poblaciones en términos de deriva genética y flujo génico.

Para trazar y cuantificar con razonable precisión los flujos génicos, disponemos de una nueva herramienta de análisis. Se trata de la filogeografía. Estudia ésta la genealogía del gen que ha dado origen a la variación existente dentro de una región del genoma y la distribución geográfica de dicha variabilidad.

Ante una diversidad genética dada, pensemos en una secuencia de ADN o en un conjunto de polimorfis-

mos, la herramienta mencionada se propone reconstruir el proceso evolutivo o filogenético que ha desembocado en la diferenciación observada a partir de un antepasado común. Al plasmar conjuntamente la diversificación del gen y la de las poblaciones, podemos anclar ciertas variantes genéticas (secuencias o haplotipos) en una rama del árbol evolutivo y en un origen geográfico. Aplicando ese método se ha cuantificado la aportación por vía paterna y materna de africanos, europeos y amerindios al acervo genético de la población brasileña contemporánea.

Para datar puntos concretos de la evolución humana se puede recurrir a los microsátélites, segmentos de ADN que contienen repeticiones de breves secuencias de dos a seis nucleótidos. Son marcadores de evo-

lución rápida. A partir de un determinado acontecimiento fundador, la cantidad de variación acumulada y medible es una función de la tasa de mutación (que podemos estimar) y del tiempo transcurrido, que es la incógnita que despejaremos. Por ejemplo, se observó que la mayoría de los judíos apellidados Cohen ("sacerdote") poseían cierto tipo de cromosoma Y. Concorre, además, que la condición de sacerdote, el apellido y el cromosoma Y se transmiten de padres a hijos exclusivamente por la línea masculina. Pues bien, de la variación acumulada en los microsátélites de este tipo de cromosoma Y se infiere un efecto fundador que operó hace unos 3000 años, coincidente con el establecimiento de una casta sacerdotal hebrea.

El Magreb y la península Ibérica

El análisis de la diversidad genética humana en poblaciones actuales ha arrojado luz sobre numerosas cuestiones históricas, en distintas escalas temporales y espaciales. Sabemos ya que la distribución y la antigüedad de la diversidad genética a escala mundial son compatibles con un origen reciente y africano de la humanidad actual. El punto de arranque, situado en África, se remontaría, a lo sumo, unos 150.000 años atrás. Por tanto, ni los habitantes del yacimiento de Atapuerca ni los neandertales serían antepasados nuestros.

A escala continental, se debate la proporción de genes de origen paleolítico (hace unos 30.000 años) y neolítico (hace 10.000) presentes en los europeos actuales. Los genes, por otro lado, apuntan a una fecha antigua (unos 30.000 años) y a un origen claramente nortasiático para la colonización de América.

A escala regional, podemos abordar también algunas cuestiones abiertas sobre la historia de las poblaciones. Se cuenta aquí con la colaboración de otras disciplinas; la arqueología, la paleoantropología o la lingüística suministran a menudo hipótesis que, en la medida que impliquen distintas historias demográficas, pueden verificarse mediante el estudio de la diversidad genética de las poblaciones actuales.

Desde la genética de poblaciones podemos abordar cuestiones que atañen a la península Ibérica (España y Portugal) y al noroeste de África (el Magreb: Marruecos, el Sahara Occidental, Mauritania, Argelia y Túnez). ¿Se puede hablar de un origen común para ambas poblaciones a sendas orillas del Mediterráneo? ¿Quedan en las poblaciones actuales rastros de un substrato paleolítico que represente el poblamiento inicial de los antepasados de las poblaciones actuales? ¿Es el mismo substrato para ambas regiones? ¿Cuál fue la aportación de la oleada de avance neolítica? ¿Qué fracción del acervo genético magrebí proviene de la invasión árabe? ¿Podemos identificar la contribución magrebí a las poblaciones peninsulares? ¿Es el Sahara una barrera impenetrable al intercambio de genes entre poblaciones?

Para resolver esa gavilla de cuestiones sobre el poblamiento y las relaciones genéticas entre la península Ibérica y el Magreb, hemos recurrido al análisis de marcadores clásicos, microsátélites autosómicos, inserciones Alu, secuencias de ADN mitocondrial, polimorfismos de un solo nucleótido (SNP) del cromosoma Y y microsátélites del cromosoma Y en muestras de poblaciones ibéricas, beréberes del norte, centro y sur de Marruecos y del centro de Argelia, árabes marroquíes, argelinos y tunecinos, y saharauis. No se estudiaron todas las poblaciones para todos los marcadores, aunque sí se investigó extensamente un núcleo fundamental.

En el caso de los marcadores clásicos, recopilamos la información publicada por otros equipos de trabajo. En otros casos, contrastamos nuestros resultados con los obtenidos por otros autores. De la investigación realizada se desprende una descripción, que creemos ajustada, de la historia de las poblaciones norteafricanas e ibéricas.

Orígenes remotos de norteafricanos y habitantes de la península Ibérica

Con pocas excepciones, todos los marcadores genéticos analizados muestran una separación clara entre las poblaciones magrebíes y

la de España y Portugal, incluidas, sin embargo, en el rango de la variación de las poblaciones *caucasoideas* (las de origen europeo, más las norteafricanas y medioorientales). Ahora bien, las distancias genéticas entre ibéricos y el resto de europeos son menores que entre ibéricos y magrebíes.

Este patrón mayoritario no se refleja en todos y cada uno de los genes analizados. Por ejemplo, en la región que lleva la información para la síntesis de los antígenos de los leucocitos humanos (HLA), los cuales definen la compatibilidad en trasplantes de órganos. Basándose en una sola región del genoma (y sometida a selección), algunos autores postularon un origen común de peninsulares y magrebíes. Pero la naturaleza aleatoria de la deriva genética y la acción de la selección pueden producir este tipo de desviaciones; para evitarlas, la interpretación debe apoyarse en la información conjunta del máximo número posible de genes y no en una sola región del genoma.

De acuerdo con nuestro análisis de marcadores clásicos, las distancias genéticas entre ibéricos y poblaciones del Oriente Medio son menores que la que existe entre ibéricos y magrebíes. Se da, además, una discontinuidad abrupta entre las orillas septentrional y meridional en el paisaje genético de la cuenca mediterránea, con una máxima pendiente en el estrecho de Gibraltar.

Estas y otras consideraciones nos llevaron a postular que en el Magreb pudo conservarse un substrato paleolítico, distinto del substrato paleolítico europeo. Supondría ello que, a diferencia de lo que parece haber sucedido en Europa, la transición al Neolítico norteafricano se produjo sin un recambio sustancial de genes. En esa hipótesis abunda la cultura capsense del Mesolítico norteafricano, que se prolonga hacia el Neolítico adoptando las nuevas formas de producción, aunque sin la drástica ruptura que se observa en gran parte de Europa. Planteamiento que viene avalado por investigaciones con secuencias Alu y microsátélites.

Hallamos una confirmación directa de la hipótesis anterior al analizar la filogeografía del cromosoma Y.

Los linajes del cromosoma Y, definidos a partir de SNPs, presentan una genealogía muy clara y una distribución geográfica que tiende a ser restringida. Las frecuencias del mismo en una región difieren de las observadas en la otra: en la península Ibérica predominan, lo mismo que en el resto de Europa Occidental, el linaje R1b* y sus inmediatos derivados (el conjunto R1b3, R1b6 y R1b8), en tanto que dos tercios de los cromosomas Y magrebíes pertenecen al linaje E3b2*.

Tras la investigación realizada sobre la variación en el cromosoma Y ha quedado patente que el grupo de linajes R1b se encuentra sólo en Oriente Medio y Europa. Ciertos linajes (R1b3, R1b6 y R1b8) se hallan circunscritos a la península Ibérica; habrían surgido aquí a partir del haplotipo fundador, sin apenas dispersarse allende sus fronteras. El grupo de linajes R1b, por su antigüedad y distribución geográfica, se habría originado en Oriente Medio y se habría difundido por

Europa con las colonizaciones iniciales del Paleolítico superior.

El linaje magrebí E3b2* se ha hallado en otras poblaciones, aunque con frecuencias mucho menores. Su antepasado más inmediato aparece entre los etíopes. Si atendemos a la acumulación de variabilidad en microsatélites asociada a E3b2*, se trataría de un linaje de más de 19.000 años de existencia. Podemos, pues, postular un escenario en que una expansión paleolítica desde África nororiental llevara al Magreb los antepasados del linaje E3b2*, que surgiría *in situ* después de la expansión. El linaje masculino E3b2* tiene un correlato en el linaje matrilineal mitocondrial U6, también de antepasados etíopes y difusión limitada al noroeste de África, aunque no alcanza las elevadas frecuencias de E3b2*. Por consiguiente, dos regiones genómicas independientes con filogeografías bien establecidas confirman la singularidad magrebí y sitúan sus raíces en el Paleolítico.

El Neolítico: un avance paralelo

Se admite que la agricultura y ganadería comenzaron en Oriente Medio hace unos 10.500 años. Esa fase prehistórica conllevó el crecimiento y la expansión consiguiente de la población en varias direcciones. Pero se debate si dicha expansión supuso un recambio genético en las poblaciones europeas o si se conserva en la actualidad un importante substrato genético paleolítico (anterior a la expansión del Neolítico), así como el grado en que ocurrió una cosa u otra.

Si atendemos al paisaje genético europeo, advertiremos un gradiente o clina desde el sudeste hacia el noroeste. Asimismo, aparece otra clina este-oeste en el norte de África, desde Egipto hasta Marruecos. No es fácil dar con una explicación de dichas clinas, pues se produjeron varios movimientos migratorios en esas mismas direcciones; por ejemplo, la primera colonización del Paleolítico y el avance del Neolítico en Europa, o el Neolítico y las in-

Tipos de polimorfismos genéticos

La variabilidad genética se presenta en múltiples formas. Con las técnicas disponibles nos es dado conocer distintas facetas de dicha variación. Podemos analizar directamente la variabilidad del ADN. Asimismo, los *marcadores genéticos clásicos* son los sistemas polimórficos detectados en los productos de expresión génica. A través de esta segunda vía, históricamente anterior a la primera, se nos revelaron los primeros sistemas polimórficos genéticos, por ejemplo, el de los grupos sanguíneos, descubiertos en 1900. Otros marcadores clásicos consisten en la variación de movilidad de proteínas plasmáticas, enzimas eritrocitarias y el sistema de antígenos leucocitarios humanos (antígenos HLA), que determinan la compatibilidad de los trasplantes.

La variabilidad o polimorfismo del ADN puede tomar varias formas. En la más sencilla, una determinada posición de nuestro genoma puede presentar dos variantes químicas (raras veces más de dos) de las cuatro posibles en las que se escribe el alfabeto del ADN (G, A, T y C). Se trata de un *SNP* (*Single Nucleotide Polymorphism*), de los que se calcula que hay unos tres millones en el genoma humano; se caracteriza por que en una posición concreta del genoma puede haber una u otra base (A o G, por ejemplo).

En vez de limitarnos a una sola posición del genoma, podemos definir un cierto segmento y determinar su secuencia en un conjunto de individuos.

Esta estrategia permite, además de reconocer las variantes existentes en ciertas posiciones, descubrir también toda la variación que hay en el ADN y reconstruir su genealogía, es decir, el proceso evolutivo que dio lugar a la variabilidad observada.

Encontramos otro tipo de variación genética en los *microsatélites* o *STR* (*Short Tandem Repeat Polymorphisms*). Estas secuencias de ADN consisten en la repetición en tándem de una unidad básica de entre 2 y 6 nucleótidos de longitud. Presentan una particularidad distintiva: el número de repeticiones de la unidad básica es muy variable entre individuos. Por ese motivo se han convertido en importantes marcadores en el campo de la genética forense (identificación de personas y de la paternidad).

Las *inserciones Alu* constituyen, por último, otro tipo interesante de polimorfismo. Estas secuencias, de unos 300 nucleótidos de longitud, utilizan la maquinaria celular de replicación del ADN para copiarse a sí mismas y reinsertarse en otras partes del genoma. Las nuevas inserciones, lo mismo que cualquier otro polimorfismo, pueden perderse o fijarse, pero algunas de ellas se hallan en una situación intermedia. Aunque esta tercera situación se da con cierta frecuencia, es un fenómeno variable de una población a otra. Además, caso único entre los polimorfismos, conocemos bien cuál es la dirección de la evolución: normalmente, de la ausencia (estado ancestral) a la presencia (estado derivado) del elemento Alu.

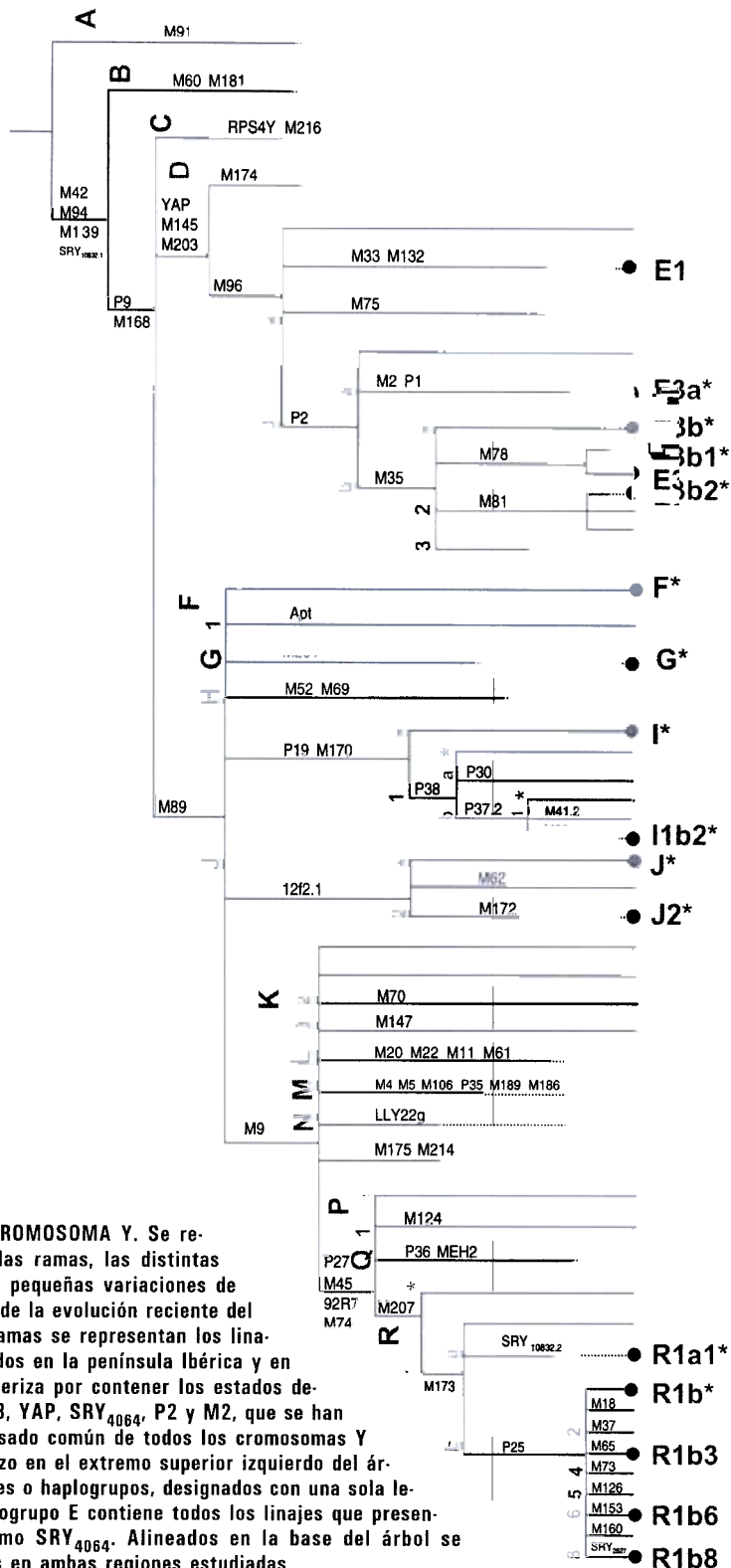
vasiones árabes en el norte de Africa.

Para nuestra fortuna, los métodos que permiten datar linajes acotan el intervalo de la difusión. Así, para el cromosoma Y en Europa los datos indican un impacto neolítico menor (con estimas en torno al 38,7%) frente a un substrato paleolítico mayor (en torno al 61,3%) en el conjunto europeo.

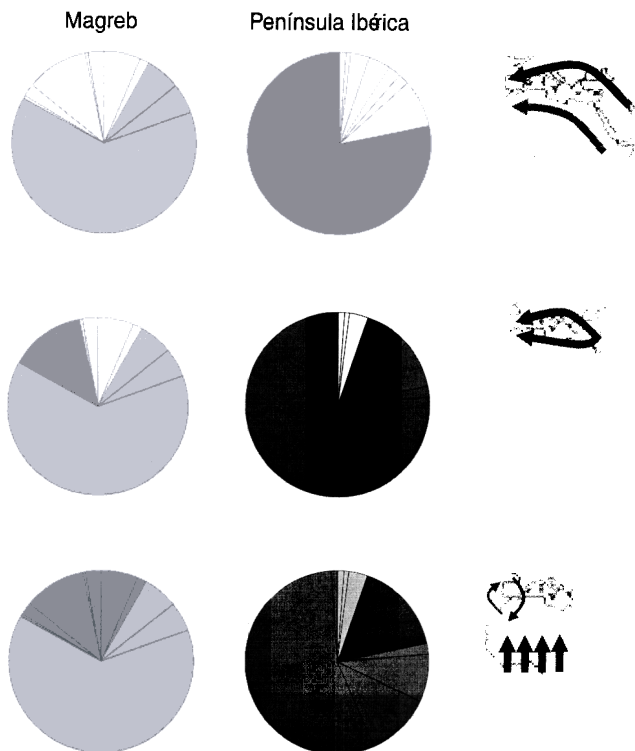
Hay razones para postular que los linajes F*, G*, J* y J2* del cromosoma Y se originaron en Oriente Medio. Desde allí se difundieron hacia el oeste por ambas riberas del Mediterráneo con la expansión del Neolítico. De acuerdo con nuestra investigación, la frecuencia de F* y G* es más elevada en la península Ibérica que en el NO de Africa; por el contrario, J* abunda más en el norte de Africa. Estas frecuencias dispares son compatibles con la hipótesis de la expansión del Neolítico, siguiendo pautas independientes, por ambas orillas del Mediterráneo; la península Ibérica y el Magreb representarían los extremos occidentales de ambas expansiones.

Colin Renfrew ha propuesto que en el Neolítico, además de los genes, se propagaron varias familias lingüísticas desde Oriente Medio: la familia indoeuropea hacia Europa, la afroasiática hacia Arabia y el norte de Africa, la elamodravídiciana hacia Irán y el subcontinente indio, y la altaica hacia Asia central. Habría, pues, un correlato génico de las expansiones lingüísticas que llevaron las lenguas indoeuropeas hacia Europa y las afroasiáticas hacia el norte de Africa. Esta hipótesis, muy controvertida, cuenta con escaso respaldo en su aplicación rigurosa.

2 GENEALOGIA DE LOS LINAJES DEL CROMOSOMA Y. Se representan con números, a lo largo de las ramas, las distintas mutaciones puntuales conocidas (SNP y pequeñas variaciones de longitud) que han aparecido en el curso de la evolución reciente del cromosoma Y. En los extremos de las ramas se representan los linajes o haplotipos del cromosoma Y hallados en la península Ibérica y en el Magreb. Así, el linaje E3a* se caracteriza por contener los estados derivados de los polimorfismos M42, M168, YAP, SRY₄₀₆₄, P2 y M2, que se han sucedido en este orden desde el antepasado común de todos los cromosomas Y existentes (indicado con un pequeño trazo en el extremo superior izquierdo del árbol). Además, se definen grupos de linajes o haplogrupos, designados con una sola letra, de la A a la R; por ejemplo, el haplogrupo E contiene todos los linajes que presentan el estado derivado del polimorfismo SRY₄₀₆₄. Alineados en la base del árbol se muestran los haplogrupos representados en ambas regiones estudiadas.



E. BOSCH, F. CALAFELL, S. PLAZA, A. PEREZ-LEZAIN, D. COMAS Y JAUME BERTRANDPÉTTI



3. HISTORIA DE LA POBLACION y linajes del cromosoma Y en el Magreb y la península Ibérica. (a) La primera colonización del Paleolítico se da independientemente en ambas regiones; introduce en el Magreb (*verde*) los linajes E3b*, E3b1* y E3b2*; en la península Ibérica (*rojo*), los linajes R1a1*, R1b* (que ulteriormente dio lugar a R1b8), R1b3 y R1b6. En los diagramas de sectores se muestra la frecuencia de dichos linajes en cada población. (b) La expansión del Neolítico, desde el Creciente Fértil y en paralelo por ambas riberas del Mediterráneo, aporta los linajes F*, J*, J2*, I* e I1b2* (*azul*). (c) Los fenómenos migratorios implican flujo génico desde la península hacia el Magreb (*en rojo*), en sentido contrario (*verde*), y desde más allá del Sahara hacia el Magreb (*malva*).

Arabes y beréberes

Los beréberes (o *imazighen*) constituyen los descendientes directos de una población ancestral que se extendía por gran parte del norte de Africa, desde Egipto hasta Senegal. Suman hoy 20 millones de personas, dispersas en pequeñas minorías de Egipto, Libia y Senegal. Sin embargo, muchos habitantes de Túnez, Argelia y Marruecos se definen a sí mismos como tales y hablan alguna de la veintena de lenguas beréberes, una rama de la familia afroasiática. El resto de la población habla y se considera árabe, y se supone descendiente de las invasiones, que desde el siglo VII y con especial intensidad en el XI,

llevaron el Islam desde la península Arábiga hasta el Magreb.

Así las cosas, podemos plantearnos si las invasiones árabes implicaron una aportación demográfica significativa o si, por el contrario, una elite numéricamente limitada pero culturalmente prestigiosa consiguió difundir una nueva lengua y religión, sin que ello conllevara una aportación de genes notable. Para resolver tal disyuntiva hemos de acudir al análisis genético de las poblaciones árabes y beréberes.

El análisis de gran cantidad de marcadores (inserciones Alu, microsátelites autosómicos y polimorfismos del cromosoma Y) nos revela una llamativa ausencia de

diferencias entre poblaciones árabes y beréberes. Sólo el ADN mitocondrial separa de los beréberes a los árabes argelinos y tunecinos (pero no marroquíes). Debemos concluir, pues, que la arabización del Magreb fue un fenómeno básicamente cultural, en que una reducida elite impuso su lengua y religión, sin que hubiera cambios sustanciales en la población local, incluso la actualmente arabófona.

(a)

(b)

(c)

Más allá del Sahara

La comparación de la diversidad genética con la hallada al sur del Sahara permite rastrear el flujo génico transahariano. En el estudio de las inserciones Alu se advierte con nitidez que las poblaciones más al sur de nuestra zona de trabajo (saharauis y beréberes del sur de Marruecos) muestran distancias genéticas más cortas con las poblaciones subsaharianas que las que se dan entre subsaharianos y poblaciones del norte del Magreb. Tal comprobación nos induce a pensar en un gradiente de flujo génico subsahariano; en el curso del mismo, las poblaciones del sur del Magreb habrían recibido una mayor aportación de genes subsaharianos, lo que, dada su posición geográfica y el conocido comercio de esclavos, parece verosímil.

Las regiones genómicas con una filogeografía bien establecida permiten cuantificar la aportación subsahariana. Así, en el norte de Africa aparecen en bajas frecuencias (un 8% en conjunto) los linajes E1* y E3a* del cromosoma Y, de origen subsahariano; no se han hallado en la península Ibérica.

En el caso del ADN mitocondrial, son de origen subsahariano los linajes L1, L2 y L3, que constituyen una media del 25% de los linajes magrebíes (con un rango entre 3% en rifeños y 40% en mauritanos). En la península Ibérica, presentan una frecuencia media del 3%, oscilando entre su ausencia en vascos y un 6% en portugueses del centro. En el caso de la Península, es difícil decidir si estos linajes proceden directamente de allende el Sahara, traídos con la trata de esclavos, o si, dada su frecuencia en el Magreb, llegaron a Iberia vía con-

tactos a través del estrecho de Gibraltar.

De la comparación entre los datos del ADN mitocondrial y los del cromosoma Y se desprende que los linajes subsaharianos heredados por vía materna se hallan a una frecuencia más elevada en magrebíes e ibéricos que los linajes paternos, lo que indicaría una diferencia entre sexos en la movilidad de los individuos desde el sur del desierto del Sahara. Esa observación genética debe contrastarse con datos sociales de movilidad y comercio de esclavos.

Tráfico en el estrecho de Gibraltar

¿Qué decir, por último, de las relaciones entre las poblaciones magrebíes y las peninsulares? Dejamos constancia al principio de la nítida separación entre ambas poblaciones, debido, probablemente, a un substrato paleolítico distinto. Ese hito temporal permite la detección del flujo génico a través del estrecho, así como su cuantificación a partir de linajes del cromosoma Y y del ADN mitocondrial.

Las personas y, si se reproducen, sus genes han cruzado el estrecho de Gibraltar con distinta intensidad a lo largo de la historia. En algunos períodos, dicho flujo aumentó. Además, se trata de una corriente bidireccional, pues también se dio un flujo génico de la Península al Magreb. De norte a sur, cruzaron el estrecho romanos, vándalos, judíos y moriscos. Los dos últimos grupos podrían haber difundido genes ibéricos. Del sur llegaron a Hispania los cartagineses; en el 711 arribaron los árabo-beréberes. En el siglo XII vinieron oleadas de almohades, almorávides y benimerines. Los datos genéticos no permiten precisar cuándo se produjo el tráfico. Sólo podemos descubrir el conjunto acumulado de los intercambios genéticos.

Como hemos comentado, el linaje E3b2* del cromosoma Y se originó en el Magreb, donde constituye unos dos tercios del total. En España y Portugal, su frecuencia se estima alrededor del 6%, con mínimos en el País Vasco y Cataluña y máximos en Extremadura y Andalucía occi-

dental. Dado que el flujo génico del Magreb hacia la Península acarrearía otros linajes, subestimaríamos la contribución genética magrebí de la Península si sólo consideráramos E3b2*; corrigiendo a tenor de la frecuencia de E3b2* respecto al total de linajes del cromosoma Y, la contribución norteafricana al acervo genético ibérico se puede estimar en un 8%. Al estudiar la variación de microsatélites dentro de este linaje, se observa una estrecha similitud entre los haplotipos ibéricos y los magrebíes. La variación observada en haplotipos peninsulares pudo haberse acumulado en un intervalo temporal que la hace compatible con las entradas del siglo VIII y, sobre todo, con las del siglo XII.

Por lo que respecta al ADN mitocondrial, encontramos un equivalente de E3b2* en U6, de origen magrebí, aunque menos frecuente. Se halla en un 10% de los magrebíes y en un 1,5% de los habitantes de nuestra península. La ausencia de linajes maternos específicamente norteafricanos a frecuencias moderadas o elevadas dificulta la estimación de la contribución femenina magrebí a la península; aunque existe, obviamente.

En un sentido inverso, los cromosomas Y del grupo R1b que hay en el Magreb pueden ser de origen europeo, si bien no podemos precisar que fuera específicamente ibérico. Su frecuencia, del 2,8% en norteafricanos, alcanza el 78,4% en ibéricos; ello supone una contribución europea del 3,6% al acervo genético magrebí. Por lo que respecta al ADN mitocondrial, el linaje V, de origen europeo, se en-

cuentra en una frecuencia del 6,8% en norteafricanos.

Conclusiones

La historia nos recuerda las intensas relaciones culturales y sociales que han existido, a lo largo de los siglos, entre el Magreb y la península Ibérica (España y Portugal). Elites o pueblo llano, mercaderes o guerreros, portaban una lengua, una religión, una cultura, en definitiva. Pero hasta ahora se nos mostraba esquivo el impacto demográfico ejercido por esos flujos, que los datos genéticos nos muestran existente pero moderado. Y todavía se debaten los movimientos relacionados con la expansión islámica.

Los datos genéticos, con sus limitaciones, han permitido trazar un primer marco comparativo entre ambas orillas del Mediterráneo, con la reconstrucción consiguiente de la historia e intercambios mutuos de sus poblaciones. La genética aporta el marco de la historia demográfica, en cuyo interior hemos de identificar las pruebas de intercambios suministradas por otras disciplinas.

Los cambios y sustituciones en el credo religioso, en la lengua o en los perfiles de las excavaciones arqueológicas nos hablan de interrelaciones y desplazamiento culturales. El alcance demográfico de los procesos demográficos asociados a esas transformaciones culturales hallan un correlato genético, cuya magnitud se va desentrañando merced al avance en el conocimiento del genoma. Es una de las múltiples sorpresas que la biología actual nos depara.

Bibliografía complementaria

- GENETIC STRUCTURE OF NORTHWESTERN AFRICA REVEALED BY STR ANALYSIS. E. Bosch, F. Calafell, A. Pérez-Lezaun, J. Clarimon, D. Comas, E. Mateu, R. Martínez, B. Morera, Z. Brakez, O. Akhayat, A. Sefrani, G. Hariti, A. Cambon-Thomsen y J. Bertranpetit en *European Journal of Human Genetics*, n.º 8, págs. 360-366; 2000.
- ALU INSERTION POLYMORPHISMS IN NW AFRICA AND THE IBERIAN PENINSULA: EVIDENCE FOR A STRONG GENETIC BOUNDARY THROUGH THE GIBRALTAR STRAITS. D. Comas, F. Calafell, N. Benchemsi, A. Helal, G. Lefranch, M. Stoneking, M. A. Batzer, J. Bertranpetit y A. Sajantila en *Human Genetics*, n.º 107, páginas 312-319; 2000.
- HIGH RESOLUTION ANALYSIS OF HUMAN Y-CHROMOSOME VARIATION SHOWS A SHARP DISCONTINUITY AND LIMITED GENE FLOW BETWEEN NORTHWESTERN AFRICA AND THE IBERIAN PENINSULA. E. Bosch, F. Calafell, D. Comas, P. J. Oefner, P. A. Underhill y J. Bertranpetit en *American Journal of Human Genetics*, n.º 68, págs. 1019-1029; 2001.

~ CAPÍTOL II ~

Joining the Pillars of Hercules: mtDNA Sequences Show Multiregional Gene Flow in the Western Mediterranean

S. Plaza, F. Calafell, A. Helal, N. Bouzerna, G. Lefranc, J. Bertranpetit,
D. Comas

Annals of Human Genetics (2003) 67, 312-328

Joining the Pillars of Hercules: mtDNA Sequences Show Multidirectional Gene Flow in the Western Mediterranean

S. Plaza¹, F. Calafell¹, A. Helal², N. Bouzerna³, G. Lefranc⁴, J. Bertranpetit¹ and D. Comas¹

¹Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona

²Laboratoire d'Immunogénétique, Faculté de Pharmacie, Monastir, Tunisia

³Laboratoire de Biochimie et Microbiologie Appliquée, Unité d'Immunobiologie, Université d'Annaba, Annaba, Algérie

⁴Institut de Génétique Humaine, UPR CNRS 1142, et Université Montpellier II, 34095 Montpellier Cedex 5, France

Summary

Phylogenetic analysis of mitochondrial DNA (mtDNA) performed in Western Mediterranean populations has shown that both shores share a common set of mtDNA haplogroups already found in Europe and the Middle East. Principal co-ordinates of genetic distances and principal components analyses based on the haplotype frequencies show that the main genetic difference is attributed to the higher frequency of sub-Saharan L haplogroups in NW Africa, showing some gene flow across the Sahara desert, with a major impact in the southern populations of NW Africa. The AMOVA demonstrates that SW European populations are highly homogeneous whereas NW African populations display a more heterogeneous genetic pattern, due to an east-west differentiation as a result of gene flow coming from the East. Despite the shared haplogroups found in both areas, the European V and the NW African U6 haplogroups reveal the traces of the Mediterranean Sea permeability to female migrations, and allowed for determination and quantification of the genetic contribution of both shores to the genetic landscape of the geographic area.

Comparison of mtDNA data with autosomal markers and Y-chromosome lineages, analysed in the same populations, shows a congruent pattern, although female-mediated gene flow seems to have been more intense than male-mediated gene flow.

Introduction

The western Mediterranean populations have experienced a long, intricated history that, too often, has been considered separately for the African and European shores, or from an exclusively European perspective. Both the African and the European shores have acted as termini of population expansions. The independent and parallel colonisation from the East of both areas by anatomically modern humans in Palae-

olithic times, and the expansion of farming during the Neolithic, have modelled the genetic landscape of both areas. Moreover other demographic events, such as the expansion of the Arabisation along the Maghrib, have also come from the East arriving in NW Africa.

Genetic diversity studies have provided a major insight into human evolution on a global scale, but they have also been useful in regional studies. Population processes such as expansions, migrations, dispersals and admixtures leave a footprint in the genetic composition of the groups that allow us to trace back population history. Several genetic markers have been analysed in the westernmost part of the Mediterranean in order to extricate such processes. The compilation of classical genetic markers (Bosch *et al.* 1997; Simoni *et al.* 1999)

*Correspondence: David Comas, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain. Tel: +34 93 542 28 02; Fax: +34 93 542 28 44. E-mail: david.comas@cexs.upf.es

has shown a clear genetic differentiation between the northern and southern coasts, attributed to independent parallel expansions along the two shores followed by little gene flow across the Mediterranean. Nevertheless, there is some contradictory data, based on HLA polymorphisms, on the degree of genetic relationship between both coasts in West Mediterranean populations (Arnaiz-Villena *et al.* 1995; Comas *et al.* 1998). Analyses of autosomal STRs (Bosch *et al.* 2000) and Alu insertion polymorphisms (Comas *et al.* 2000) confirmed the genetic difference between both groups of populations, also detecting some Sub-Saharan genetic flow into NW African populations. The high-resolution analysis of Y-chromosome biallelic and STR markers (Bosch *et al.* 2001) has revealed clear genetic differentiation due to a major independent Upper Palaeolithic contribution in both areas, followed by gene flow from the Near East during the Neolithic, and small bidirectional gene flow across the Mediterranean. Several mitochondrial DNA (mtDNA) analyses have focused in the structure of Iberian populations (Bertranpetit *et al.* 1995; C  rte-Real *et al.* 1996; Salas *et al.* 1998; Pereira *et al.* 2000), of NW African populations (Rando *et al.* 1998; Brakez *et al.* 2001), and their relation to the Canary Islands (Pinto *et al.* 1996). Nevertheless, no analysis has jointly considered the population relationships of Western Mediterranean populations using mtDNA sequences.

The analysis of mitochondrial DNA diversity has been one of the most successful tools applied to unravel regional population histories. Two different approaches have been followed in order to perform mtDNA analyses: the sequencing of the hypervariable segments of the non-coding part of the molecule, the control region, and the study of the coding region through high-resolution RFLPs. The joint analysis of both kinds of markers (control region sequences and RFLPs in the coding region) has proven to be a powerful tool in studying human diversity (Torrioni *et al.* 1996), and has led to the construction of robust phylogenies of mtDNA sequences (Macaulay *et al.* 1999), which allow one to elucidate human demographic scenarios.

In the present study, we have analysed the hypervariable segment I (HVSI) of the control region in several Western Mediterranean populations, and have added the information yielded by three SNPs in the mtDNA-

coding region in order to ascribe the mtDNA variation to specific branches of the gene genealogy. This analysis allows us to describe the genetic landscape of the geographic region, compare it to that obtained with other genomic regions (particularly those with a clear phylogeography, such as the Y-chromosome), and interpret it in terms of external gene flow and of exchanges between the northern and southern shores of the Mediterranean.

Material and Methods

Population Samples

A data set comprising sequences for the first hypervariable segment (HVSI) of the mtDNA control region (positions 16024 to 16383, according to the Cambridge Reference Sequence; CRS, Anderson *et al.* 1981; Andrews *et al.* 1999) in populations of the Western Mediterranean (defined as the Iberian and Italian Peninsulas and intervening islands, southern France, and NW Africa from Tunisia to Mauritania), was collected from the literature and from our own analyses. Sequences for a total of 1,719 individuals were collected (see populations, sample sizes and references in Table 1 and Figure 1); of those, we sequenced a total of 267 unrelated individuals: 172 from North-West Africa and 95 from the Iberian Peninsula. The NW African samples included 56 Saharawi, 18 Moroccans Arabs, 4 Berbers from North-Central Morocco, 47 Algerians and 47 Tunisians. The Iberian Peninsula samples comprised 49 Andalusians and 46 Catalans. Populations analysed were chosen in order to generate a complete picture of the region. These sequences are available at <http://www.upf.es/cexs/recerca/bioevo/index.htm>

MtDNA Amplification and Sequencing

Total DNA was extracted from fresh blood using standard phenol-chloroform methods after digestion with proteinase K. HVSI was amplified with primers and methods as described elsewhere (Mateu *et al.* 1997). The amplified product was purified with the Gene Clean kit (BIO 101) and sequencing reactions were performed using the Big Dye Terminator (version 3.0) Cycle Sequencing Kit, with AmpliTaq[®] DNA Polymerase (Applied Biosystems). The sequencing products were

Table 1 Diversity parameters for the HVRI in several populations

Populations	n	k	S	Sequence diversity	Mean pairwise differences	Nucleotide diversity	References
Northwest Africa	429						
Algerians	47	27	51	0.957±0.043	5.72±4.28	0.0158	1
Mauritanians	30	23	31	0.975±0.025	6.09±3.91	0.0169	2
Moroccan Arabs	50	44	68	0.993±0.007	7.04±2.96	0.0195	1,2
Moroccan Berbers	64	42	51	0.968±0.032	4.52±5.48	0.0125	1,2
Mozabites	85	29	35	0.942±0.058	4.73±5.27	0.0131	3
Saharawi	56	41	46	0.978±0.022	5.45±4.55	0.0151	1
Southern Berbers	50	34	38	0.941±0.059	4.60±5.40	0.0128	4
Tunisians	47	42	61	0.989±0.010	6.15±3.85	0.0171	1
Iberian Peninsula	887						
Andalusians	158	106	82	0.965±0.035	4.26±5.74	0.0118	1,5,6,7,8
Basques	173	71	64	0.942±0.058	3.03±6.97	0.0084	5,8,9,10
Catalans	78	48	45	0.938±0.062	3.66±6.34	0.0102	1,5,6,8
Central Spain	50	38	49	0.953±0.047	4.59±5.41	0.0128	6,8
Galicians	103	62	61	0.939±0.061	3.31±6.69	0.0092	6,8,11
Valencians	30	24	37	0.970±0.030	4.25±5.75	0.0118	8
Portuguese	54	38	40	0.934±0.066	3.60±6.40	0.0100	5
NPortuguese	100	67	71	0.953±0.047	4.78±5.22	0.0133	12
CPortuguese	82	62	66	0.977±0.023	4.87±4.13	0.0135	12
Sportuguese	59	41	55	0.943±0.057	4.54±5.46	0.0126	12
Italy	411						
Central Italy	83	63	61	0.974±0.012	4.78±5.41	0.0133	13
Sardinians	73	50	57	0.955±0.045	4.24±5.76	0.0118	14
Sicilians	169	97	92	0.936±0.064	4.03±5.97	0.0112	15,16
Southern Italy	37	31	47	0.969±0.031	4.86±5.14	0.0135	16
Tuscans	49	40	55	0.969±0.031	5.03±4.97	0.0140	17

n: number of individuals; k: number of different sequences; S: number of variable positions. NPortuguese: Northern Portuguese; CPortuguese: Central Portuguese; SPortuguese: Southern Portuguese. References: 1 Present study; 2 Rando *et al.* 1998; 3 Macaulay *et al.* 1999; 4 Brakez *et al.* 2000; 5 Côrte-Real *et al.* 1996; 6 Crespillo *et al.* 2000; 7 López-Soto *et al.* 2000; 8 A. Alonso (personal communication); 9 Bertranpetit *et al.* 1995; 10 Richards *et al.* 2000; 11 Salas *et al.* 1998; 12 Pereira *et al.* 2000; 13 Tagliabracci *et al.* 2001; 14 Di Rienzo & Wilson, 1991; O. Rickards (personal communication); 15 Cali *et al.* 2001; 16 O. Rickards *et al.* 2000 and personal communication; 17 Francalacci *et al.* 1996.

run in an ABI PRISM 3100 sequencer (Applied Biosystems).

Three positions in the mtDNA coding region (10400, 12308 and 12705, according to Anderson *et al.* 1981) were also determined in some of the sequenced individuals by using the SNaPshotTM ddNTP Primer Extension Kit (Applied Biosystems), as described elsewhere (Comas *et al.* in preparation), which implements a single-base primer extension protocol that uses labelled ddNTPs to interrogate SNPs.

Phylogenetic Analysis

Sequence alignment was performed using the ESEE program (Cabot, 1988). Each control region sequence

was assigned to a given haplogroup by comparison with the data sets where mtDNA had been typed for both RFLPs and HVSI sequences (Torroni *et al.* 1996; Watson *et al.* 1997; Rando *et al.* 1998; Macaulay *et al.* 1999) and the data sets of Richards *et al.* (2000), and their classification scheme was used with a single modification: a few sequences bearing a transition at position 16126 and not carrying transitions at 16069, 16294, 16296, or 16362, were classified as J/T, a denomination that should not necessarily imply that they are in a group ancestral to haplogroups J and T. When the information given by the nucleotide substitutions of the HVSI in those individuals sequenced in the present study was insufficient to assign a sequence to a given haplogroup, three positions outside the control region (10400, 12308

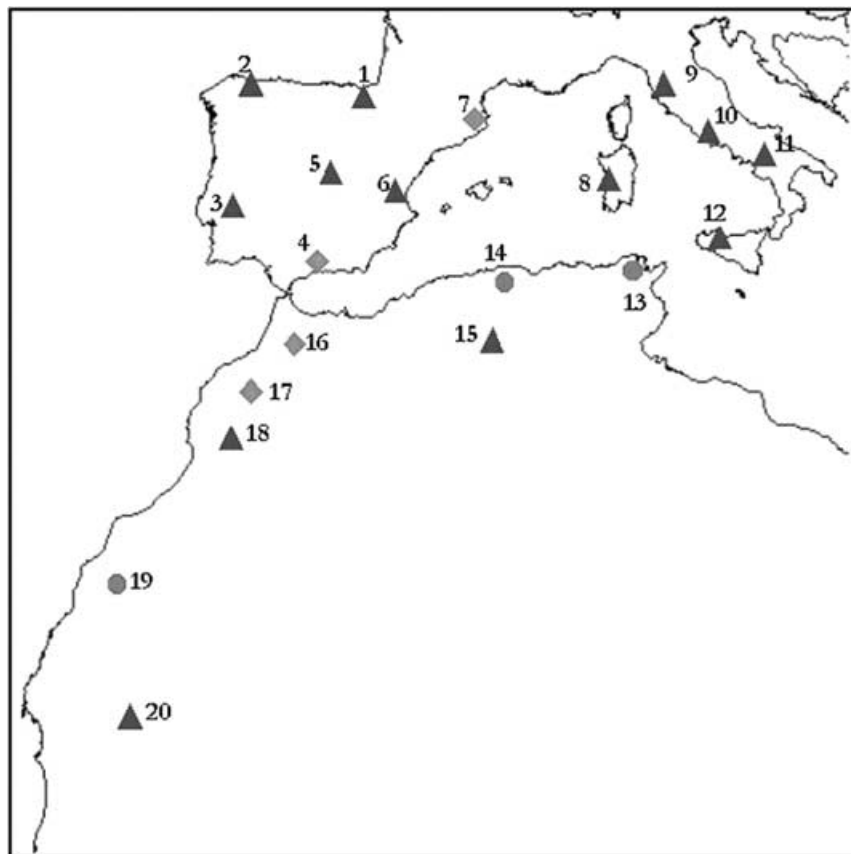


Figure 1 The Western Mediterranean. Geographical location of the samples analysed: 1: Galicians; 2: Basques; 3: Portuguese (include mixed Portuguese, Northern, Central and Southern Portuguese); 4: Andalusians; 5: Central Spain; 6: Valencians; 7: Catalans; 8: Sardinians; 9: Tuscans; 10: Central Italy; 11: Southern Italy; 12: Sicilians; 13: Tunisians; 14: Algerians; 15: Mozabites; 16: Moroccan Berbers; 17: Moroccan Arabs; 18: Southern Berbers; 19: Saharawi; 20: Mauritians. Dots represent samples sequenced in the present work, squares represent samples pooled from the literature as well as new sequences included, and triangles represent samples taken from the literature.

and 12705) were determined. These positions allowed us to assign the control region sequences to three different major haplogroups: 10400T defines the major M haplogroup, 12308G defines the major U haplogroup (including the K haplogroup), and 12705C defines the major R haplogroup, which includes a large set of haplogroups (H, V, J, T, U, B and F). Nevertheless, 2% of all sequences remained ambiguous or could not be typed for these three positions as they were taken from the literature, and they were classified as “other”.

The networks relating HVSI sequences within some of the haplogroups described were constructed by using a reduced-median algorithm (Bandelt *et al.* 1995) as implemented in the Network 3.0 program. The dating method employed (Morrall *et al.* 1994; Saillard *et al.*

2000) is based on the average number of mutations accumulated from an ancestral sequence as a linear function of time and mutation rate. This method was also performed with the Network 3.0 program.

Population Analysis

Population internal genetic diversity parameters (nucleotide diversity, sequence diversity and mean pairwise differences) were computed with the Arlequin 2000 program (Schneider *et al.* 1996).

Population genetic structure was tested through analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992), using the Arlequin 2000. Genetic distances between populations using the first mtDNA hypervariable region were calculated by intermatch-mismatch

pairwise differences according to the equation $D = d_{ij} - (d_{ii} + d_{jj})/2$ (Nei 1987), where d_{ij} is the mean pairwise differences between populations i and j , and d_{ii} and d_{jj} are the mean pairwise differences within populations i and j respectively. The distance standard errors were computed by resampling nucleotide positions with 1,000 bootstrap iterations (Efron, 1982). A principal co-ordinate plot (Gower, 1966) was also obtained from the distance matrix. Principal component analysis was performed from haplotype frequencies using the SPSS package.

Results

Phylogeographic Structure

Haplogroup frequencies estimated as described above are listed in Table 2. The phylogeographic structure of mtDNA in the Western Mediterranean can be summarised as five sets of haplogroups: 1) sub-Saharan haplogroup L (including L1, L2, L3); 2) haplogroups J, T, J/T; 3) haplogroups H, V, HV; 4) haplogroup U (including K); and 5) haplogroups W, I, X, and M.

L haplogroups are relatively infrequent in Italians (with a maximum of 8.1% in South Italians) and Iberians (with a maximum of 6.1% in Central Portuguese). On the contrary, L haplogroups are distributed in all North African populations at high frequencies (from 26% in South Berbers to 43.5% in Mauritians) with the exception of Mozabites (12.9%) and Moroccan Berbers (3.2%). In fact, the frequency of the L haplogroups in Moroccan Berbers is similar to that found in Iberians and Italians. The frequency of the L haplogroups might represent the sub-Saharan genetic flow into the populations analysed, which has shown to be substantial in NW Africa but very limited in European populations.

In the populations analysed, haplogroups J and T present their highest frequencies in the Italian samples, with values over 15%. Iberians showed a heterogeneous frequency distribution with values that range from 6.6% in Valencians to 18.7% in Southern Portuguese. NW Africans have similar J and T frequencies to Europeans, although it is worth noting that Saharawi and Mauritians, the southern NW African samples, differ from the other populations in that haplogroups J and T are almost absent.

Haplogroup U is found in all samples analysed at considerable frequencies. The most relevant aspect within this group of sequences is the presence of haplogroup U6, to which a North African origin has been attributed (Rando *et al.* 1998). Haplogroup U6 is largely distributed among Mozabites (28.2%) and Mauritians (20%). In other NW Africans, the frequency of U6 ranges from 4.2% in Tunisians to 8% in Moroccan Arabs, with the remarkable case of Algerians where haplogroup U6 is absent. In Italians, haplogroup U6 is practically absent, with only one sequence found among Sicilians. In the Iberian Peninsula U6 distribution is sparse. It is present in the south-western part of the Peninsula at low frequencies (<7%), and is absent in Basques, Catalans, Valencians, Central Portuguese, and Southern Portuguese. Few U6 sequences are found in other populations from different geographical regions: Sub-Saharan and NE Africa, the Middle East and the Canary Islands (with a frequency of 14%; Rando *et al.* 1998, 1999). The structure of the variation of U6 sequences is shown in Figure 2, from which the age of U6 can be estimated at $47,000 \pm 18,000$ years, similar to that first estimated by Rando *et al.* (1998). The network shows a clear structure in subhaplogroups within U6: U6a (characterised by 16278T; Rando *et al.* 1999), U6a1 (characterised by 16278T and 16189C; Richards *et al.* 2000), and U6b (characterised by 16311C; Rando *et al.* 1999). The present Iberian and NW African sequences are found within haplogroups U6a and U6a1, but haplogroup U6b contains no NW African sequences and is mainly composed of Canarian and Iberian sequences. Moreover, there is a group of sequences within U6b characterised by 16163T, which we named U6b1 (although unnamed, this was already discussed by Rando *et al.* 1999), which presents basically Canarian sequences. The age of this group of sequences (Canarians plus two Iberian sequences) is around $9,400 \pm 5,500$ years. The presence of U6a and U6a1 haplogroups in the Iberian Peninsula could be attributed to gene flow from NW Africa, and the most plausible origin for U6b1 lineages in Iberia is recent gene flow from the Canary Islands after the contact between Europeans and the Canary aborigines in the fifteenth century.

H and V represent the major group of sequences in Iberia and Italy. H is by far the most frequent haplogroup in western European populations, as it is in all of Europe

Table 2 Haplogroup frequencies (%) in Western Mediterranean Populations

Pop (n)	Alg (47)	Mau (30)	MA (50)	MB (64)	MB (64)	Moz (85)	Moz (85)	Sah (56)	SBer (50)	Tun (47)	NWA (429)	And (158)	Bas (173)	Cat (78)	CS (50)	Gal (103)	Val (30)	Port (54)	NPo (100)	CPo (82)	SPo (59)	IBE (887)	CIt (83)	Sard (73)	Sic (169)	SIt (37)	Tus (49)	ITA (411)
L1	6.4	23.4	10.0	1.6	-	3.6	6.0	-	6.4	-	-	-	-	-	2.0	-	-	-	1.0	1.3	1.7	0.6	-	1.4	-	2.7	-	0.8
L2	6.4	13.4	6.0	-	5.9	7.1	10.0	12.8	7.7	0.6	-	0.6	-	-	2.0	1.0	3.4	1.8	3.0	2.4	1.7	1.6	-	1.4	0.6	-	-	0.4
L3	14.9	6.7	16.0	1.6	7.0	23.3	10.0	14.9	11.8	1.3	-	1.3	-	1.3	-	1.0	-	1.8	1.0	2.4	1.7	1.0	1.2	-	-	5.4	2.0	1.7
D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.2	-	-	-	-	0.2
M1	12.8	-	2.0	-	4.7	-	-	-	4.2	3.0	1.9	-	-	-	-	-	-	-	-	1.3	-	0.3	-	1.4	1.8	-	-	0.6
M5	-	-	-	-	-	-	-	-	-	-	1.9	-	1.9	-	2.0	-	-	-	-	-	-	0.4	-	-	-	-	-	-
N	-	-	-	-	-	-	-	-	-	0.6	-	0.6	-	1.3	-	-	6.6	-	-	1.3	-	1.0	-	1.4	1.2	2.7	-	1.1
I	-	-	-	-	-	-	-	-	2.1	0.3	1.9	-	-	-	-	-	-	-	1.0	-	1.7	0.5	-	1.2	-	4.1	1.1	
W	-	-	-	-	-	-	-	-	-	1.3	-	1.3	-	5.1	2.0	1.9	-	-	2.0	1.3	-	1.4	1.2	1.4	1.8	5.4	2.0	2.4
X	2.1	-	4.0	-	-	-	-	-	2.1	1.0	3.2	1.7	2.6	2.0	1.0	-	-	-	-	3.6	1.7	1.6	3.6	1.4	2.9	5.4	6.1	3.9
R1	-	-	-	-	-	-	-	-	-	-	1.7	-	-	-	-	-	-	-	-	-	-	0.2	-	-	-	2.7	-	0.5
J/T	-	-	-	1.6	-	-	-	-	-	0.2	0.6	-	-	-	-	-	-	-	-	-	-	0.1	-	-	1.8	2.7	-	0.9
T	4.2	-	4.0	15.6	4.7	1.8	4.0	6.4	5.1	4.4	5.2	7.7	10.0	2.9	-	-	-	11.1	11.0	11.0	10.2	7.3	15.7	12.3	8.3	13.5	10.2	12.0
J	12.8	3.3	4.0	9.4	3.5	-	10.0	4.2	5.9	7.0	4.0	2.6	8.0	9.7	6.6	5.6	6.0	6.0	6.0	6.1	8.5	6.5	6.0	5.5	5.3	2.7	14.3	6.8
U*	2.1	3.3	12.0	6.2	12.9	8.9	8.0	6.4	7.5	10.1	13.3	9.0	20.0	12.6	16.6	9.3	16.0	14.6	11.9	13.1	9.6	12.3	4.7	5.4	12.2	8.8	-	-
U6	-	20.0	8.0	7.8	28.2	5.4	6.0	4.2	9.9	1.9	-	2.0	1.9	-	2.0	1.9	-	5.6	7.0	-	-	1.8	-	-	0.6	-	-	0.1
K	4.2	6.6	4.0	7.8	-	7.1	2.0	6.4	4.8	6.3	5.2	6.4	2.0	3.9	10.0	7.4	3.0	7.3	6.8	5.8	7.2	5.8	7.2	5.5	2.9	2.7	8.2	5.3
HV	-	-	-	-	-	-	7.1	2.0	6.4	1.9	2.5	-	-	1.3	2.0	1.0	3.4	3.7	-	-	-	1.4	1.2	2.7	-	-	2.0	1.2
H	34.0	20.0	26.0	42.2	24.7	17.9	32.0	23.4	27.5	46.2	57.8	56.4	46.0	59.2	53.3	48.1	41.0	37.8	44.1	49.2	47.0	50.7	50.3	45.9	38.8	46.5	-	-
V	-	3.3	4.0	6.2	8.2	17.9	10.0	-	6.2	5.7	10.4	5.1	-	2.9	-	3.7	8.0	7.3	6.8	5.0	4.8	2.7	5.9	2.7	5.9	2.7	-	3.2
Other	-	-	-	-	-	-	-	6.4	0.8	2.5	0.6	1.3	-	1.0	-	2.4	3.4	1.3	1.2	-	2.4	3.4	1.3	1.2	-	10.6	-	2.9

Alg: Algerians; Mau: Mauritians; MA: Moroccan Arabs; MB: Moroccan Berbers; Moz: Mozabites; Sah: Saharawis; SBer: South Berbers; Tun: Tunisians; And: Andalusians; Bas: Basques; Cat: Catalans; CS: Central Spain; Gal: Galicians; Val: Valencians; Port: Portuguese; NPo: North Portuguese; CPo: Central Portuguese; SPo: South Portuguese; CIt: Central Italy; Sard: Sardinians; Sic: Sicilians; SIt: South Italians; Tus: Tuscans. NWA: unweighted average frequencies in NW Africans; IBE: unweighted average frequencies in Iberians; ITA: unweighted average frequencies in Italians. (*): Excluding U6. N includes sequences carrying the HVRI substitutions diagnostic of either N1a or N1b.

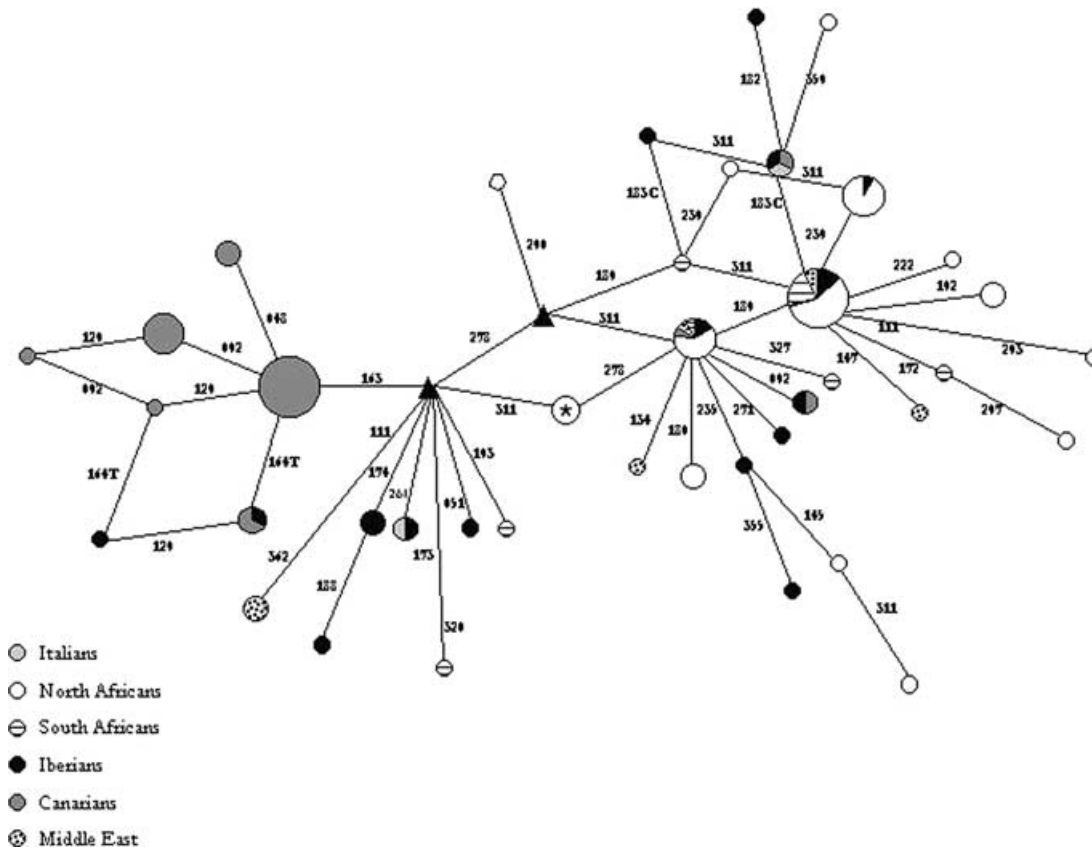


Figure 2 Phylogenetic network of haplogroup U6 HVRI sequences (present data and data from Richards *et al.* 2000). The size of the circles is proportional to the number of sequences. The node marked with an asterisk indicates the ancestral sequence (16172C, 16219G). Numbers along links refer to nucleotide positions in HVRI minus 16000; suffixes indicate a transversion. Subhaplogroups U6a and U6a1, to the right in the graph, are defined by positions 16278T and 16278T, 16189C respectively. Subhaplogroups U6b and U6b1, to the left in the graph, are defined by positions 16311C and 16163T, 16311C respectively.

(Simoni *et al.* 2000a; Richards *et al.* 2000). It has been suggested that haplogroup V originated and expanded from NE Iberia (Torroni *et al.* 1998; Torroni *et al.* 2001). In the European samples analysed, its frequency (which includes pre-V and V proper as defined by Torroni *et al.* 2001) ranges from 2.7% in Sardinia and Southern Italy to 10.4% in Basques, and is absent in Central Spaniards, Valencians, and Tuscans. Except in Algerians and Tunisians, haplogroup V has been found in all the samples analysed, with high frequencies among the Saharawi (17.9%) and Southern Berbers (10%). In order to elucidate the phylogenetic relationships between sequences, a network of V sequences was constructed (Figure 3). The network displayed a clear star-like pattern with all V sequences found in NW Africa close to the V sequence root type or with one or two added

substitutions, whereas Italian and Iberian V sequences show a wider distribution of substitutions. Out of the five different V haplotypes found in NW Africa, three were those that are most frequent in Europe, while only two were specific to NW Africa. A time depth for the haplogroup V of $13,700 \pm 3,000$ years was estimated when all sequences were included, similar to previous estimates (Torroni *et al.* 2001).

The last section of the mtDNA phylogeny considered includes the Eurasian haplogroups W, I, X, and haplogroup M. Haplogroups W, I, and X are basically found in continental Italy, and some traces are found in Iberians, Algerians, Tunisians and Moroccan Arabs. The M sequences found in the analysed populations can be sorted into two different phylogenetic groups: haplogroups M1 and M5. It has been suggested that

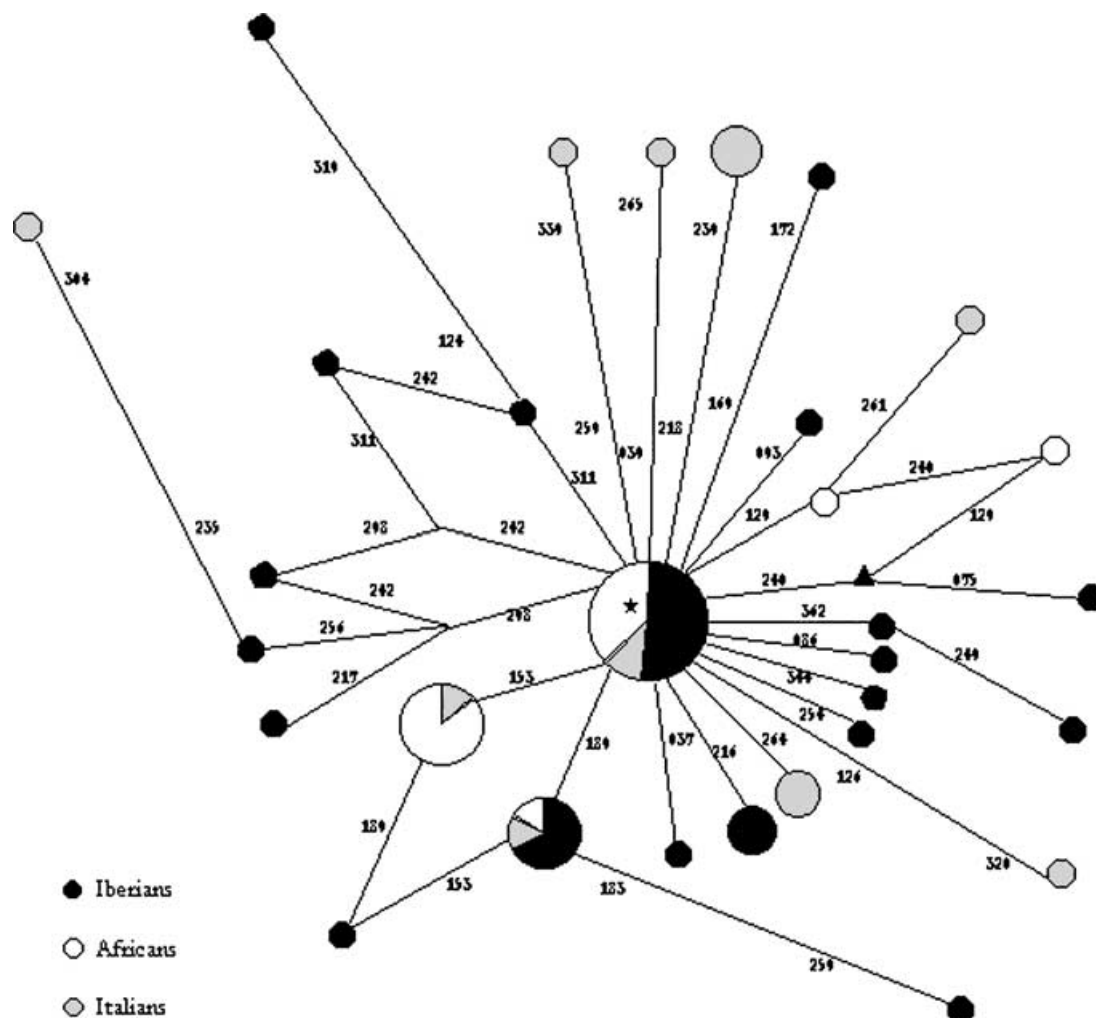


Figure 3 Phylogenetic network of haplogroup V HVRI sequences. The size of the circles is proportional to the number of sequences. The node marked with an asterisk indicates the ancestral sequence (16298C). Numbers along links refer to nucleotide positions in HVRI minus 16000.

haplogroup M1 originated in eastern Africa (Quintana-Murci *et al.* 1999), and it is almost absent in the European samples analysed. Nevertheless, it has been found at high frequencies in Algerians, and at a lower frequency in Tunisians, Mozabites and Moroccan Arabs, showing a slight east-west cline. On the contrary, haplogroup M5, defined by 16129A (Bamshad *et al.* 2001), which accounts for 97.3% of the M lineages in Gypsies (also known as Roma; Gresham *et al.* 2001), has only been found in Andalusians and Central Spaniards, which is not surprising given that Spain is one of the European countries where the Gypsy community is more numerous (~500,000 people; Liegeois, 1994).

Population Structure of Genetic Variation

Analyses of the molecular variance (AMOVA) were performed in order to detect any genetic structure within the present sample set (Table 3). Due to the difference observed in the contribution of L lineages in the populations studied, all the analyses described below were performed in duplicate: with the whole set of sequences, and without the L sequences. When all samples were treated as a single group, 97.4% of the variance was attributed to differences within populations and 2.6% ($p < 0.01$) represents differences among populations. This fraction was reduced to 2% when the

Table 3 Analyses of Molecular Variance (AMOVA) in West Mediterranean populations

Groups	Among groups		Among populations within groups		Within populations	
	with L	without L	with L	without L	with L	without L
All populations			2.56**	2.00**	97.44**	98.00**
NW Africans			3.53**	4.50**	96.47**	95.50**
SW Europeans			0.62**	0.62**	99.38**	99.38**
NW Africa vs SW Europe	2.62**	1.41**	1.48**	1.49**	95.90**	97.10**
Iberian Peninsula vs Italy	0.06 ns	0.11 ns	0.59**	0.57**	99.34**	99.32**
Eastern vs Western NW Africa ^a	0.94 ns	1.96*	1.29**	0.69 ns	97.76**	97.35**

** ($p < 0.01$); * ($p < 0.05$); ns : non-significant All the analyses were performed taking into account lineages belonging to L haplogroups (with L) and ignoring L lineages (without L).

^aTwo groups: Algerians and Tunisians versus the rest of NW African populations.

L lineages were removed. Considering separately the southern and northern populations, NW Africans are more heterogeneous: Φ_{ST} among NW Africans is 3.5% (4.5% without L sequences), as compared to 0.6% among SW Europeans.

When we grouped the samples according to their geographical area (SW Europeans versus NW Africans), 1.5% of the genetic variance was due to differences between samples of the same geographical area, and 2.6% was attributable to differences between geographical areas. The variance attributable to differences among geographical groups decreased to 1.4% when the L sequences were removed, whereas the variance attributable to differences within groups did not vary, showing that Sub-Saharan gene flow into NW Africa has in part been responsible for the differences between the two groups.

In order to establish a valid comparison between nuclear DNA, Y chromosomal, and mtDNA, we performed an AMOVA with Alu polymorphisms (Comas *et al.* 2000), the Y-chromosome lineages (Bosch *et al.* 2001) and mtDNA data among the same populations from NW Africa and the Iberia Peninsula as described in Bosch *et al.* (2001). We found that the proportion of the genetic variance that can be accounted for between the NW African and Iberian populations for mtDNA is 0.86% ($p = 0.053$), 1.89% ($p = 0.028$) for Alu insertion polymorphisms, and 35.2% ($p = 0.024$) for the Y chromosome. It is not surprising to find that the results show clear differences between male and female lineages due to the already described sexual differential migration

patterns for worldwide human populations (Seielstad *et al.* 1998). Autosomal markers, here represented by Alu insertion polymorphisms, show intermediate values between those found for the mtDNA and the Y chromosome, although Y-chromosome markers exhibit much greater differences between both geographical areas. The Y chromosome behaves as a single locus, and, as such, it is more prone to the vagaries of random drift than a set of independent loci such as the Alu polymorphisms. Moreover, sex-specific population structure (restricted gene flow with isolation) might have enhanced an initial increase of differentiation in male lineages compared to the other of genetic systems analysed.

When focusing in SW Europe, very small (0.1%), non-significant differences were found between Iberians and Italians, in agreement with the large mtDNA homogeneity described in Europe (Simoni *et al.* 2000a, 2000b; Helgason *et al.* 2000; see also Richards *et al.* 2002). Within NW Africa and grouping samples following a geographical east-west criterion (Tunisians and Algerians versus the other populations), no significant differences were found between groups, and 1.3% of the variation was attributed to differences among populations within groups. Nevertheless, when L lineages were removed, the variation attributed to differences among the two groups became significant (1.96%, $p < 0.05$), and the differences within groups did not differ from zero, which points to an east-west differentiation in NW Africa that may have been partially dampened by gene flow from sub-Saharan Africa to both subregions.

Genetic Landscape

The genetic relationship between NW Africans, Iberians and Italians was assessed through a principal co-ordinate analysis based on the distance matrix. The plot of the first two principal co-ordinates (Figure 4a) accounts for 65.8% of the genetic variance observed. The first co-ordinate (56.3%) separates NW African and European populations, except for Moroccan Berbers who are embedded within Europeans, placing Mozabites and Mauritians at one edge and Basques and Galicians at the opposite one. When L lineages are excluded from the analysis (Figure 4b), the plot clusters Europeans, Moroccan Berbers and Southern Berbers and even Tunisians in a group, whereas Algerians, Mozabites and Mauritians are more distant to this cluster.

The genetic relationships between the populations was also assessed through a principal component analysis based on the frequencies of the haplogroups displayed (Richards *et al.* 2002). We observed a similar general pattern to that displayed in the principal co-ordinates, but some differences were also found. The first two principal components (Figure 5a) account for 36.3% of the genetic variance observed and separate the NW African populations, characterised by high frequencies of L and U6 sequences (with absolute correlations of 0.619 for L1, 0.887 for L2, 0.781 for L3, and 0.663 for U6), from the rest of populations, which present high frequencies of H lineages (with an absolute correlation of 0.835). The second principal component encompassed 12.1% of the genetic variance observed and separated the Southern Italians, Tuscans and Sicilians from the rest of the SW Europeans by their low frequencies of K (absolute correlation of 0.458) and the presence of J/T lineages (absolute correlation of 0.735) in their genetic pool. When L sequences were removed from the analysis (Figure 5b), the first two principal components encompassed 33.4% of the genetic variance and separated most NW African populations from Italians, with the remaining populations lying between them.

Discussion

The phylogeographic analysis of mtDNA in the Western Mediterranean has shown the presence of a common set of haplogroups shared with the rest of Europe and

the Middle East (H, J, T, U, I, W, X), plus those of probable local origin (U6, V), and others introduced by gene flow from the south (L) and east (M). In this respect, our regional study, which has gathered published and new samples, not previously jointly analysed, confirms the basic frame described by Richards *et al.* (2000) for Europe and by Rando *et al.* (1998) for NW Africa. It should be noted, though, that inferring haplogroups from HVRI sequences and three coding-region SNPs could lead to slight imprecisions in the allocation of sequences to haplogroups. For instance, although we have assigned all CRS (Cambridge Reference Sequence) sequences to haplogroup H, 1.5% of all CRS sequences in West Eurasia belong to haplogroup HV* and 3.9% to U* (Richards *et al.* 2000). Typing of SNP 7028 could help in resolving this ambiguity, which nonetheless affects a relatively small number of sequences.

An additional caveat that should be taken into account throughout the discussion is that, although we define our area of study as the Western Mediterranean, for some areas, such as southern France, Corsica, northern Italy and the Kabyle in northern Algeria, no HVRI sequences are available. It is likely that such missing data would refine some of the conclusions we will reach below.

Now, we will discuss in detail the phylogeographic pattern for NW Africa, Iberia and Italy, and the trans-mediterranean gene flow.

Northwest African mtDNA Landscape

The main difference, found through the mtDNA analysis, between the populations of the two geographical areas studied is the presence of sub-Saharan L lineages in NW Africa compared to SW Europe, up to the point that, if L sequences were removed from the analyses, most NW African populations were genetically very close to SW Europeans. Since L sequences make up almost all mtDNA lineages in sub-Saharan Africa, and particularly in the areas just to the south of NW Africa, the frequency of L haplogroups in NW Africa can be read directly as a measure of gene flow. Thus, it can be estimated that $25.9 \pm 2.1\%$ of the NW African mtDNA pool has a sub-Saharan origin, under the assumption of negligible back flow from NW to sub-Saharan Africa. A similar estimation can be

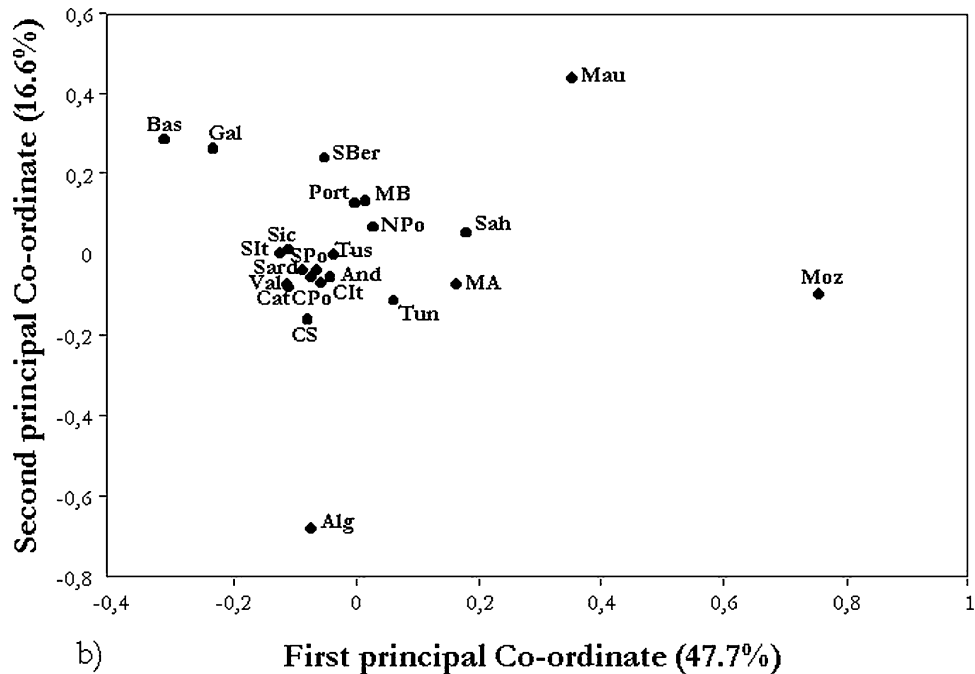
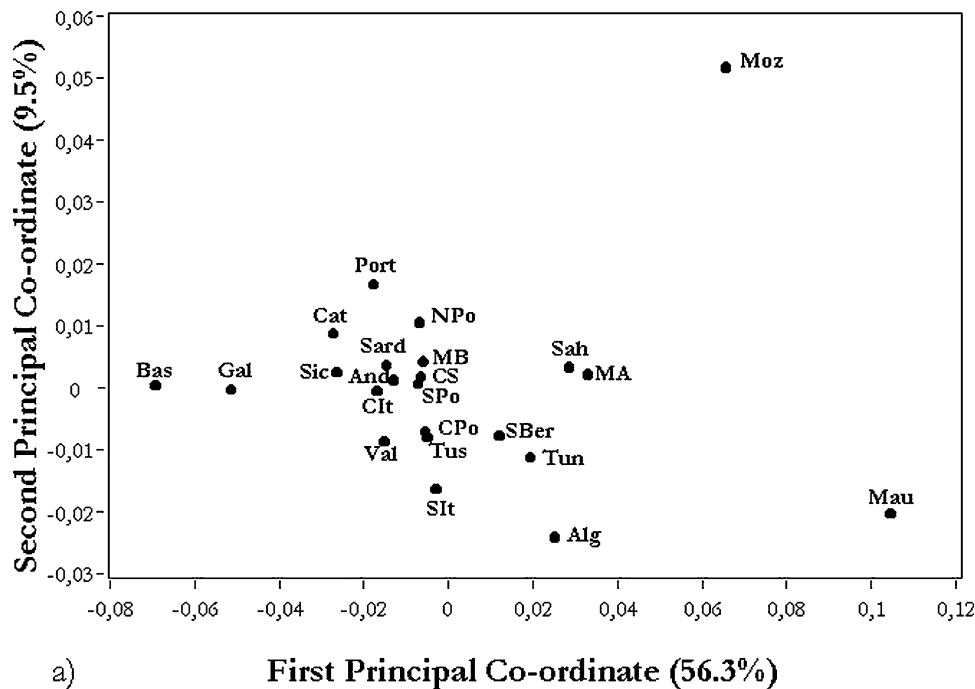
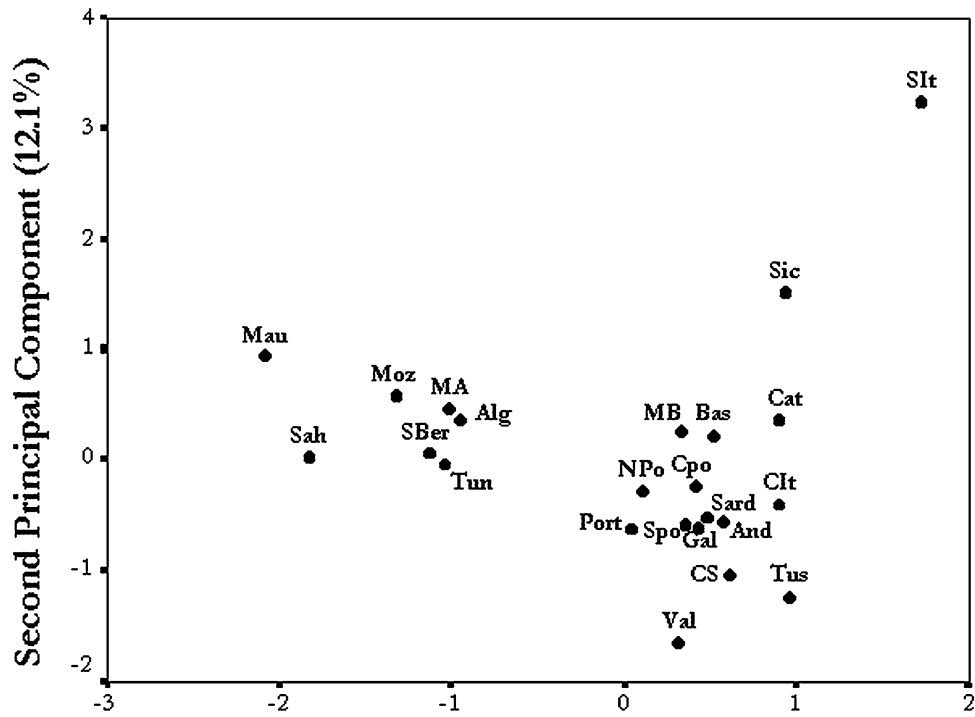
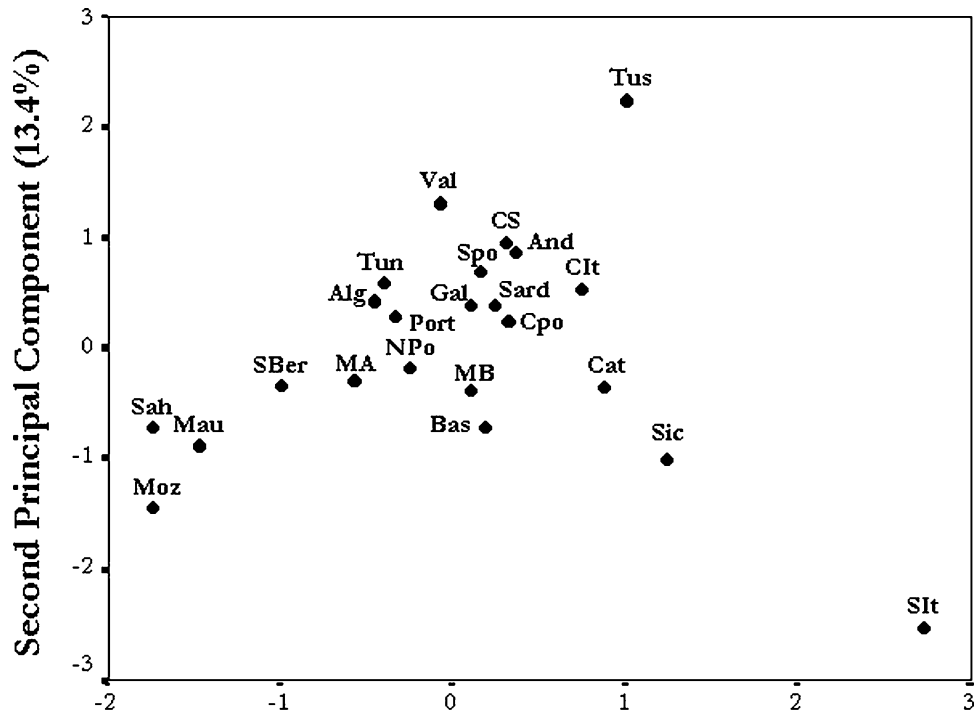


Figure 4 Plot of the two principal co-ordinate (PC) scores based on the genetic distance matrix of NW African, Iberian and Italian samples. a) PC analysis of populations considering all sequences, and b) PC analysis of populations without sequences belonging to haplogroups L. Abbreviations: Alg: Algerians; And: Andalusians; Bas: Basques; Cat: Catalans; CIt: Central Italians; Cpo: Central Portuguese; CS: Central Spaniards; Gal: Galicians; MA: Moroccan Arabs; Mau: Mauritians; MB: Moroccan Berbers; Moz: Mozabites; NPo: Northern Portuguese; Port: Portuguese; Sah: Saharawis; Sard: Sardinians; SBer: Southern Berbers; Sic: Sicilians; SIt: Southern Italians; Spo: Southern Portuguese; Tun: Tunisians; Tus: Tuscans; Val: Valencians.



a) First Principal Component (24.2%)



b) First Principal Component (20.0%)

Figure 5 Plot of the two first Principal component (PC) scores based on the haplogroup frequencies of NW African, Iberian and Italian samples. a) PC analysis of populations considering all sequences, and b) PC analysis of populations without sequences belonging to L haplogroups. Abbreviations as in Fig. 4.

performed for Y-chromosome lineages, since E1* and E3a* haplogroups (according to the nomenclature of the Y Chromosome Consortium, 2002) found in NW Africa at a frequency of $8.0\% \pm 2.0\%$ (Bosch *et al.* 2001), are of sub-Saharan origin. The female- and male-mediated estimates of sub-Saharan gene flow into NW Africa are clearly different, which could be a local consequence of a global trend to higher female than male migration (Salem *et al.* 1996; Seielstad *et al.* 1998; Pérez-Lezaun *et al.* 1999). Autosomal markers such as Alu insertion polymorphisms also show frequency patterns compatible with gene flow from sub-Saharan Africa into NW Africa (Comas *et al.* 2000), although the absence of a clear phylogeographic structure in that case prevents the estimation of gene flow without specifying a parental, non-admixed population for NW Africa.

Within NW Africa, L sequences are most frequent in Mauritians and Saharawi, whereas their frequency is lowest in northern populations. Alu insertion polymorphism analysis in NW Africa (Comas *et al.* 2000) has also shown that gene flow from sub-Saharan Africa in the southern part of this geographical area was more pronounced. A similar genetic gradient was also observed in NE Africa along the Nile valley from analysing Egyptian and Nubian mtDNA sequences (Kriings *et al.* 1999), where south-north migration (and vice versa) could be facilitated by the Nile.

Sequence frequency and diversity, and nucleotide diversity, point to NW Africa as the cradle of U6, with an estimated age of $47,000 \pm 18,000$ years. Such an ancient age contrasts with the limited spread of U6, which is found in N Africa, the Canaries and Iberia, and at very low frequencies in Italy, the Middle East, and the Sahel. This could be explained because, with the exception of the Moslem invasions of Iberia and Sicily, no large population expansion has been known to originate in NW Africa, and the gene tree structure for U6 does not seem compatible with a strong population expansion. U6 represents, thus, a local background in NW Africa. Its relatively low frequency ($\sim 10\%$ overall, although ranging from absence in Algeria to 28.2% in the Mozabites) is in stark contrast with the high frequency of Y-chromosome haplogroup E3b2* (64%; Bosch *et al.* 2001), which may also have originated (or expanded to such high frequency) locally in NW Africa. This discrepancy may be the result of ancient, random,

locus-specific drift, and/or of a male-biased bottleneck or migration. A locus-specific effect may be evidenced by the fact that AMOVA between Iberian and NW African populations is much higher for Y chromosome haplogroups than for multiple autosomal Alu insertion polymorphisms or mtDNA. Since men contribute their autosomes as well, the fact that population differentiation as demonstrated by autosomal loci is much closer to that for mtDNA than to that for the Y chromosome may be taken as evidence for ancient, random, locus-specific drift affecting the Y chromosome.

NW African populations are relatively heterogeneous in their mtDNA sequence pools. The eastern populations (Algeria and Tunisia) may have received more gene flow from the east, as evidenced by the frequencies of M1. This haplogroup originated in East Africa (Quintana-Murci *et al.* 1999) with a frequency $\sim 20\%$ in Ethiopians (Passarino *et al.* 1998), and declines north-westwards (Nubians $\sim 10\%$ and Egyptians $\sim 8\%$; Kriings *et al.* 1999), whereas its frequency in the Middle East is lower ($\sim 3\%$ in Jordanians from Amman, Richards *et al.* 2000; $\sim 2\%$ Israeli Palestinians, Richards *et al.* 2000; $\sim 2\%$ in Israeli Druze, Macaulay *et al.* 1999).

The major outlier within NW Africa are the Mozabites, a well-known Berber isolated group in Algeria, where drift may have altered haplogroup frequencies.

SW European mtDNA Landscape

The mtDNA homogeneity observed in Europe (Simoni *et al.* 2000a and 2000b; Helgason *et al.* 2000, see also Richards *et al.* 2002) is also seen in the present analysis of the West Mediterranean samples, and contrasts with the heterogeneity of NW African populations. All the European samples present the same set of haplotypes with similar frequencies, short genetic distances to each other, and no clear genetic structure, up to the point that populations from Iberia and Italy do not each form a neat group. It should be noted that this homogeneity is seen at the current level of phylogenetic resolution, and that a more fine-grained structure may emerge from the analysis of complete mtDNA sequences (Richards *et al.* 2002).

The most outstanding feature in the west Mediterranean genetic landscape is the outlier position of Sardinians and Basques shown by classical genetic markers

(Cavalli-Sforza *et al.* 1994; Calafell & Bertranpetit 1994; Cappello *et al.* 1996) and Y-chromosome polymorphisms (Caglià *et al.* 1997; Scozzari *et al.* 2001; Bosch *et al.* 2001), although not so pronounced in the Basques. Nevertheless, mtDNA data reveals no differences between these two populations and the rest of European populations. This has also been shown in Basques by analysis of 11 Alu insertion polymorphisms in west Mediterranean populations (Comas *et al.* 2000).

Genetic Exchange Through the Mediterranean

Each of the subregions analysed (NW Africa and SW Europe) shows sequences that originated on the opposite shore of the Mediterranean. This is particularly clear in the case of U6 and L in SW Europe. L sequences are found at frequencies $\sim 3\%$ in Iberia and $\sim 2.4\%$ in Italy. Given the relatively high frequencies of L sequences in NW Africa, it is not clear whether they were contributed by the historical populations movements from the south to the north of the Mediterranean (such as the Moslem invasions of the 7th–11th centuries), or whether its presence is associated with other processes not directly linked to NW Africa. Out of 23 different L sequences in Iberia, two were also found in NW Africa (as well as in sub-Saharan Africa), and 7 others were found in sub-Saharan Africa (in a dataset comprising 1,158 individuals from 20 populations; Graven *et al.* 1995, Pinto *et al.* 1996; Watson *et al.* 1996; Mateu *et al.* 1997; Rando *et al.* 1998; Krings *et al.* 1999; Pereira *et al.* 2001; Brehm *et al.* 2002) but not in NW Africa. Treating the set of L sequences in Iberia as if it were a population reveals genetic distances from some W African populations, such as the Senegalese and Yoruba, that are slightly smaller than those between L sequences in Iberia and NW Africa. Thus, it may be the case that gene flow from NW Africa is not entirely responsible for the presence of L sequences in Iberia.

This may be even clearer in Italy, where the frequency of U6 is much lower than in Iberia (one out of 411 individuals), and where none of the eight L sequences has been found in NW Africa. Three Italian L sequences have been described throughout Africa, and the remaining five are not found in $>1,000$ sub-Saharan individuals. Thus, the presence of L sequences cannot

be attributed to migration from NW Africa, and may instead represent gene flow from other sources, such as the Neolithic expansion or the Roman slave trade.

In contrast to mtDNA, no sub-Saharan Y chromosomal lineages were detected in Iberia (Bosch *et al.* 2001), or in Italy (Rosser *et al.* 2000), although sample sizes in these studies (97 and 99 chromosomes respectively) may not be sufficient to rule out their presence at low frequencies.

As hinted above, the presence of haplogroup U6 in Iberia may signal gene flow from NW Africa, and those of the subhaplogroup U6b1 recent gene flow from the Canary Islands. Haplogroup U6 is present at frequencies ranging from 0 to 7% in the various Iberian populations, with an average of 1.8%. Given that the frequency of U6 in NW Africa is $\sim 10\%$, the mtDNA contribution of NW Africa to Iberia can be estimated at 18%, with a 95% confidence interval of 8%–26% (estimated by sampling with replacement 10,000 times in populations having the same sample sizes and U6 frequencies as Iberia and NW Africa). This is larger than the contribution estimated with Y-chromosomal lineages (7%, 95% confidence interval 1%–14%, Bosch *et al.* 2001). However, it should be noted that the variance due to genetic drift is not included in the estimates, and this may have had a larger effect on U6, which has a much lower frequency in NW Africa than its Y-chromosome counterpart, E3b2*. In the same way, we can estimate the Canarian female contribution to the Iberian Peninsula: the subhaplogroup U6b1 is present at a frequency of 13% in the Canary Islands, and reached a frequency of 0.2% in the Iberian Peninsula. Thus, the mtDNA lineages of the Canary Islands contributed 1.5%, with a 95% confidence interval 0–4.7%, to the genetic pool of Iberia. The presence of lineages belonging to the U6b1 haplogroup in the Iberian Peninsula suggests recent gene flow from the Canary Islands, due to recent migration or to the enslavement and deportation of the native Canarians (also called Guanches) at the time of conquest by the kingdom of Castile (15th century).

With the present data, and in conjunction with other loci, we have glimpsed the palimpsest history of the Western Mediterranean; in that history, the geographical barriers imposed by the Sahara Desert and the Mediterranean Sea might not have been strong enough to prevent a certain degree of gene flow among already

differentiated populations, as they were not barriers to the flow of cultures, languages, and religions.

Acknowledgments

We express our appreciation to the blood donors who participated in the present study. We especially thank Antonio Alonso (Instituto Toxicológico, Madrid), Manuel Crespillo (Instituto Nacional de Toxicología, Barcelona), Manuel López-Soto (Instituto Nacional de Toxicología, Sevilla), Olga Rickards (Università Tor Vergata, Rome), and Martin Richards (Huddersfield University) for providing unpublished data. This work was supported by Dirección General de Investigación, Ministerio de Ciencia y Tecnología in Spain grant PB98-1064 and BOS2001-0794, and by Departament d'Universitats, Recerca i Societat de la Informació, Generalitat de Catalunya grant 2001SGR00285 and fellowship 2000FI00696 to SP.

References

- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R. & Young, I.G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.
- Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.
- Arnaiz-Villena, A., Benmamar, D., Alvarez, M., Diaz-Campos, N., Varela, P., Gomez-Casado, E. & Martinez-Laso, J. (1995) HLA allele and haplotype frequencies in Algerians. Relatedness to Spaniards and Basques. *Hum Immunol* **43**, 259–268.
- Bamshad, M., Kivisild, T., Watkins, W.S., Dixon, M.E., Ricker, C.E., Rao, B.B., Naidu, J.M., Prasad, B.V., Reddy, P.G., Rasanayagam, A., Papiha, S.S., Villems, R., Redd, A.J., Hammer, M.F., Nguyen, S.V., Carroll, M.L., Batzer, M.A. & Jorde, L.B. (2001) Genetic evidence on the origins of Indian caste populations. *Genome Res* **11**, 994–1004.
- Bandelt, H.J., Forster, P., Sykes, B.C. & Richards, M.B. (1995) Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753.
- Bertranpetit, J., Sala, J., Calafell, F., Underhill, P.A., Moral, P. & Comas, D. (1995) Human mitochondrial DNA variation and the origin of Basques. *Ann Hum Genet* **59**, 63–81.
- Bosch, E., Calafell, F., Perez-Lezaun, A., Comas, D., Mateu, E. & Bertranpetit, J. (1997) Population history of north Africa: evidence from classical genetic markers. *Hum Biol* **69**, 295–311.
- Bosch, E., Calafell, F., Perez-Lezaun, A., Clarimon, J., Comas, D., Mateu, E., Martinez-Arias, R., Morera, B., Brakez, Z., Akhayat, O., Sefiani, A., Hariti, G., Cambon-Thomsen, A. & Bertranpetit, J. (2000) Genetic structure of north-west Africa revealed by STR analysis. *Eur J Hum Genet* **8**, 360–366.
- Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A. & Bertranpetit, J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* **68**, 1019–1029.
- Brakez, Z., Bosch, E., Izaabel, H., Akhayat, O., Comas, D., Bertranpetit, J. & Calafell, F. (2001) Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Ann Hum Biol* **28**, 295–307.
- Brehm, A., Pereira, L., Bandelt, H.J., Prata, M.J. & Amorim, A. (2002) Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann Hum Genet* **66**, 49–60.
- Cabot *et al.* (1988) *ESEE: The Eyeball Sequence Editor, version 1.06*. Burnaby: University of British Columbia.
- Calafell, F. & Bertranpetit, J. (1994) Principal component analysis of gene frequencies and the origin of Basques. *Am J Phys Anthropol* **93**, 201–215.
- Cali, F., Le Roux, M.G., Anna, R., Flugy, A., De Leo, G., Chiavetta, V., Ayala, G.F. & Romano, V. (2001) MtDNA control region and RFLP data for Sicily and France. *Int J Legal Med* **114**, 229–231.
- Caglià, A., Novelletto, A., Dobosz, M., Malaspina, P., Ciminelli, B.M. & Pascali, V.L. (1997) Y-chromosome STR loci in Sardinia and continental Italy reveal islander-specific haplotypes. *Eur J Hum Genet* **5**, 288–292.
- Cappello, N., Rendine, S., Griffio, R., Marnelli, G.E., Succa, V., Vona, G. & Piazza, A. (1996) Genetic analysis of Sardinia: I. Data on 12 polymorphisms in 21 linguistic domains. *Am Hum Genet* **60**, 125–141.
- Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. (1994) *History and geography of human genes*. Princeton, NJ: Princeton University Press.
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A. & Bertranpetit, J. (1998) HLA evidence for the lack of genetic heterogeneity in Basques. *Ann Hum Genet* **62**, 123–132.
- Comas, D., Calafell, F., Benchemsi, N., Helal, A., Lefranc, G., Stoneking, M., Batzer, M.A., Bertranpetit, J. & Sajantila, A. (2000) Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* **107**, 312–319.
- Côrte-Real, H.B., Macaulay, V.A., Richards, M.B., Hariti, G., Issad, M.S., Cambon-Thomsen, A., Papiha, S., Bertranpetit, J. & Sykes, B.C. (1996) Genetic diversity in

- the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* **60**, 331–350.
- Crespillo, M., Luque, J.A., Paredes, M., Fernandez, R., Ramirez, E. & Valverde, J.L. (2000) Mitochondrial DNA sequences for 118 individuals from northeastern Spain. *Int J Legal Med* **114**, 130–132.
- Di Rienzo, A. & Wilson, A.C. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci USA* **88**, 1597–1601.
- Efron, B. (1982) *The Jackknife, the bootstrap and other resampling plans*. Philadelphia, Pa: Society for industrial and Applied Mathematics.
- Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.
- Francalacci, P., Bertranpetit, J., Calafell, F. & Underhill, P.A. (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am J Phys Anthropol* **100**, 443–460.
- Gower, J.C. (1966) Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika* **27**, 857–874.
- 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.
- Gresham, D., Mora, B., Underhill, P.A., Passarino, G., Lin, A.A., Wise, C., Angelicheva, D., Calafell, F., Oefner, P.J., Shen, P., Tournev, I., Pablo, R., Kucinskis, V., Perez-Lezaun, A., Marushiakova, E., Popov, V. & Kalaydjieva, L. (2001) Origins and divergence of the Roma (gypsies). *Am J Hum Genet* **69**, 1314–1331.
- Helgason, A., Sigurðardóttir, S., Gulcher, J.R., Ward, R. & Stefansson, K. (2000) MtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* **66**, 999–1016.
- Krings, M., Salem, A.E., Bauer, K., Geisert, H., Malek, A.K., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Utermann, G., Sajantila, A., Paabo, S. & Stoneking, M. (1999) MtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am J Hum Genet* **64**, 1166–1176.
- Liegeois, J.P. (1994) *Roma, Gypsies, Travellers*. Strasbourg, France: Council of Europe Press.
- López-Soto, M. & Sanz, P. (2000) Mitochondrial DNA polymorphisms in individuals living in Andalusia (south of Spain) and Extremadura (western Spain). In: *Cuadernos de Medicina Forense*, pp. 17–24.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B. & Torroni, A. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**, 232–249.
- Mateu, E., Comas, D., Calafell, F., Perez-Lezaun, A., Abade, A. & Bertranpetit, J. (1997) A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and Sao Tome, Gulf of Guinea. *Ann Hum Genet* **61**, 507–518.
- Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R., Macek, M. Jr. & Kalaydjieva, L., et al. (1994). The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat Genet* **7**, 169–175.
- Nei, M. (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.
- Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M. & Santachiara-Benerecetti, A.S. (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* **62**, 420–434.
- Pereira, L., Prata, M.J. & Amorim, A. (2000) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann Hum Genet* **64**, 491–506.
- 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. *Ann Hum Genet* **65**, 439–458.
- Perez-Lezaun, A., Calafell, F., Comas, D., Mateu, E., Bosch, E., Martinez-Arias, R., Clarimon, J., Fiori, G., Luiselli, D., Facchini, F., Pettener, D. & Bertranpetit, J. (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* **65**, 208–219.
- Pinto, F., Gonzalez, A.M., Hernandez, M., Larruga J.M. & Cabrera, V.M. (1996) Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann Hum Genet* **60**, 321–330.
- Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A.S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23**, 437–441.
- Rando, J.C., Pinto, F., Gonzalez, A.M., Hernandez, M., Larruga, J.M., Cabrera, V.M. & Bandelt, H.J. (1998) Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann Hum Genet* **62**, 531–550.
- Rando, J.C., Cabrera, V.M., Larruga, J.M., Hernandez, M., Gonzalez, A.M., Pinto, F. & Bandelt, H.J. (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet* **63**, 413–428.
- 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. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**, 1251–1276.
- 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.
- Rickards, O., Martínez labarga, C., Casalotti, R., Castellana, G., Tunzi sisto, A.M. & Mallegni, F. (2000) MtDNA variability in extinct and extant populations of Sicily and southern Italy. In: C. Renfrew & K. Boyle (Eds), *Archaeogenetics: DNA and the population prehistory of Europe*, pp. 175–183, McDonald Institute Monographs, McDonald Institute for Archaeological Research, Cambridge.
- Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D.G., Brede, G., Cooper, G., Cortereal, H.B., de Knijff, P., Decorte, R., Dubrova, Y.E., Evgrafov, O., Gilissen, A., Glisic, S., Golge, M., Hill, E.W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S.A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L.A., Malaspina, P., Maria, S., McElreavey, K., Meitinger, T.A., Mikelsaar, A.V., Mitchell, R.J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, P.C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M.J., Previdere, C., Roewer, L., Rootsi, S., Rubinsztein, D.C., Sallard, J., Santos, F.R., Stefanescu, G., Sykes, B.C., Tolun, A., Villems, R., Tyler-Smith, C. & Jobling, M.A. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* **67**, 1526–1543.
- Saillard, J., Magalhaes, P.J., Schwartz, M., Rosenberg, T. & Norby, S. (2000) Mitochondrial DNA variant 11719G is a marker for the mtDNA haplogroup cluster HV. *Hum Biol* **72**, 1065–1068.
- Salas, A., Comas, D., Lareu, M.V., Bertranpetit, J. & Carracedo, A. (1998) MtDNA analysis of the Galician population: a genetic edge of European variation. *Eur J Hum Genet* **6**, 365–375.
- Salem, A.H., Badr, F.M., Gaballah, M.F. & Paabo, S. (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *Am J Hum Genet* **59**, 741–743.
- Seielstad, M.T., Minch, E. & Cavalli-Sforza, L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet* **20**, 278–280.
- Schneider, S., Kueffer, J.M., Roessli, D. & Excoffier, L. (1996) *Arlequin (ver1.0): a software environment for the analysis of population genetics and Biometry Lab*. Geneva, Switzerland: University of Geneva.
- Scozzari, R., Cruciani, F., Pangrazio, A., Santolamazza, P., Vona, G., Moral, P., Latini, V., Varesi, L., Memmi, M.M., Romano, V., De Leo, G., Gennarelli, M., Jaruzelska, J., Villems, R., Parik, J., Macaulay, V. & Torroni, A. (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum Immunol* **62**, 871–884.
- Simoni, L., Gueresi, P., Pettener, D. & Barbujani, G. (1999) Patterns of gene flow inferred from genetic distances in the Mediterranean region. *Hum Biol* **71**, 399–415.
- Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. & Barbujani, G. (2000a) Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet* **66**, 262–278.
- Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. & Barbujani, G. (2000b) Reconstruction of prehistory on the basis of genetic data. *Am J Hum Genet* **66**, 1177–1179.
- Tagliabracci, A., Turchi, C., Buscemi, L. & Sassaroli, C. (2001) Polymorphism of the mitochondrial DNA control region in Italians. *Int J Legal Med* **14**, 224–228.
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.L. & Wallace, D.C. (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850.
- Torroni, A., Bandelt, H.J., Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M.L., 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**, 1137–1152.
- Torroni, A., Bandelt, H.J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martínez-Cabrera, V., Villems, R., Kivisild, T., Metspalu, E., Parik, J., Tolk, H.V., Tambets, K., Forster, P., Karger, B., Francalacci, P., Rudan, P., Janicijevic, B., Rickards, O., Savontaus, M.L., Huoponen, K., Laitinen, V., Koivumaki, S., Sykes, B., Hickey, E., Novelletto, A., Moral, P., Sellitto, D., Coppa, A., Al-Zaheri, N., Santachiara-Benerecetti, A.S., Semino, O. & Scozzari, R. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* **69**, 844–852.
- Watson, E., Bauer, K., Aman, R., Weiss, G., Haeseler, A. & Paabo, S. (1996) MtDNA sequence diversity in Africa. *Am J Hum Genet* **59**, 437–444.
- Watson, E., Forster, P., Richards, M. & Bandelt, H.J. (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**, 691–704.
- Y Chromosome Consortium, The (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* **12**, 339–348.

Received: 10 October 2002

Accepted: 19 March 2003

~ CAPÍTOL III ~

Mitochondrial DNA heterogeneity in Tunisian Berbers

K. Fadhlaoui-Zid, S. Plaza, F. Calafell, M. Ben Amor, D. Comas,

A. Bennamar El gaaied

D. Comas

Annals of Human Genetics (en premsa)

Mitochondrial DNA heterogeneity in Tunisian Berbers

K.Fadhlaoui-Zid^{1*}, S.Plaza^{2*}, F.Calafell², M. Ben Amor¹, D.Comas², A. Bennamar El gaaied¹

¹ Laboratoire de Génétique Moléculaire, Immunologie et Biotechnologie, Faculté des Sciences de Tunis, Université Tunis El Manar, 2092 Tunis, Tunisia

² Unitat de Biologia Evolutiva. Universitat Pompeu Fabra. Doctor Aiguader 80, 08003 Barcelona, Spain.

* Both authors contributed equally to the present work

Correspondence:

David Comas, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain.

Tel: +34 93 542 28 02; Fax: +34 93 542 28 44. E-mail: david.comas@upf.edu

Running head: mtDNA analysis in Tunisian Berbers

Summary

Berbers live in groups scattered across North Africa whose origins and genetic relationships with their neighbours are not well established. The first hypervariable segment of the mitochondrial DNA (mtDNA) control region was sequenced in a total of 155 individuals from three Tunisian Berber groups and compared to other North Africans. The mtDNA lineages found belong to a common set of mtDNA haplogroups already described in North Africa. Besides the autochthonous North African U6 haplogroup, a group of L3 lineages characterized by the transition at position 16041 seems to be restricted to North Africans, suggesting that an expansion of this group of lineages took place around 10500 years ago in North Africa and was spread to neighbouring populations. Principal components and the coordinate analysis show that some Berber groups, (the Tuareg, the Mozabite, and the Chenini-Douiret), are outliers within the North African genetic landscape. This outlier position is consistent with an isolation process followed by genetic drift in haplotype frequencies and with the high heterogeneity displayed by Berbers compared to Arab samples as shown in the AMOVA. Despite this Berber heterogeneity, no significant differences are shown between Berber and Arab samples, suggesting that the Arabization was mainly a cultural process rather than a demographic replacement.

INTRODUCTION

Berbers inhabit scattered places in North Africa, from the Moroccan western coast to the oasis Siwa in Egypt, and from Tunisia in the north to the oases in mid-Sahara. The origin of the Berber people is not clearly established. According to the archaeological record, North Africa was peopled since Upper Palaeolithic times (Newman 1995). The first well-defined Palaeolithic technology, the Aterian, dates back around 40,000 years ago, which is followed by the Iberomaurisian (~22,000 years ago; Ferembach, 1985; Close and Wendorf 1990). The archaeological record is followed by a Mesolithic culture, the Capsian (Brett & Fentress, 1996), which gave place to the Neolithic transition to agriculture that occurred around 9,500-7,000BC, spreading from the Near East to Egypt (Dupanloup 1993). Berbers may be the descendants of Mesolithic Capsian populations, and/or of the later Neolithic people who came from the Middle East via Egypt and who possibly introduced the Afro-Asiatic languages in North Africa (Renfrew 1991). Since then, the North African coast has known several invasions: Phoenicians, Romans, Vandals, Byzantines, Arabs, Ottomans, Spanish and French have occupied the territory, although their demographic impact is not well established.

In Tunisia, the first well-known post-Neolithic invasion was that of the Phoenicians coming from the East Mediterranean sea coast around 1,100BC. Nonetheless, their number was estimated at the end of their kingdom to be 100,000 for 500,000 Berbers living in Tunisia (Julien 1961). The long dominations such as Romans, Vandals and Byzantines had even a lesser demographic impact. The Arab conquest in Tunisia started in the 7th century and was followed by a massive Bedouin immigration during the 11th century. During this invasion, Berbers were restricted to Numidia (the

Center and South of Tunisia). The Arab expansion largely submerged the original Berber language and customs, except for the tribes that were forced back to the mountains and certain villages located in Southern Tunisia. Between the sixteenth and twentieth centuries Tunisia was ruled by Turks, Spaniards and French. Other immigrants came from the South (Sub-Saharan Africans slaves from Sudan). All these populations probably contributed to the present Tunisian gene pool.

Nowadays the only criterion based to distinguish the Berbers from the rest of the Tunisian population is the language called Chelha. Berbers who speak Chelha in Tunisia are 1% of the global population and are localized in four villages in the South of Tunisia (Sened, Matmata, Chenini and Douiret) and in the island of Jerba.

Mitochondrial DNA (mtDNA) is a powerful tool in reconstructing population history, because a fine-grained phylogeography has been defined for it. This is particularly relevant in Northern Africa, which is at the crossroads of Europe, the Middle East, and Sub-Saharan Africa, regions for which mtDNA phylogeography is known in detail (Macaulay et al., 1999; Richards et al., 2000; Salas et al., 2002).

Several genetic studies have been performed in North African populations, although very few in Tunisian Berbers. The compilation of classical genetic markers in North Africa (Bosch et al., 1997) showed a clear genetic differentiation between East and West, attributed to human expansions from the Middle East, such as Palaeolithic and/or Neolithic demographic expansions. Analyses of autosomal STRs (Bosch et al., 2000) and *Alu* insertion polymorphisms (Comas et al., 2000) showed some Sub-Saharan genetic flow into NW African populations. The high-resolution analysis of Y-chromosome biallelic and STR markers (Bosch et al., 2001) has

revealed a clear North African genetic differentiation compared to Europe due to a major independent Upper Palaeolithic contribution in both areas, followed by gene flow from the Near East during the Neolithic, and small bidirectional gene flow across the Mediterranean. Finally, mtDNA in North Africa has been analysed mainly in the Western coast (Rando et al., 1998; Brakez et al., 2001; Plaza et al., 2003) and Egypt (Klings et al., 1999), and it is known only for a non-Berber population in Tunisia (Plaza et al., 2003). Previous mtDNA analyses have suggested that modern Berbers are the descendants of the earlier groups living in North Africa in Palaeolithic times (Rando et al., 1998; Macaulay et al., 1999). The analysis of GM and KM haplotypes in Tunisian Berbers (Chaabani et al., 1984) suggested some heterogeneity within Berbers; thus the mtDNA analysis will contribute significantly to knowledge of the genetic pool of Tunisian Berbers.

In the present study we have analyzed mtDNA HVS-I sequences in three Berber isolates from Southern Tunisia, with the aim to evaluate the possible heterogeneity of these Berbers in relation to genetic drift, compare them to several Arabic and Berber-speaking populations from North Africa, and determine the gene contributions of surrounding populations in Tunisian Berbers to trace their population history.

MATERIALS AND METHODS

Samples

A total of 155 Berbers from Tunisia were analysed for the hypervariable segment I (HVS-I) of the mtDNA non-coding region. Blood samples were collected from four villages (Figure 1): 53 from Sened (also known as Sundia), 49 from Matmata (or Matmatia), and 53 from Chenini (Chenenaouia) and Douiret (Douiria). Chenini and Douiret are two neighbouring villages 20 km from each other; samples from these two villages were pooled and treated as a single population. Blood samples were collected according to geographic and linguistic criteria: donors were Berber Chelha speakers born in one of the four villages mentioned above. DNA extraction was performed using a standard phenol-chloroform method.

MtDNA amplification and sequencing

The HVS-I was amplified using the primers L15996 and H16401 as described in Vigilant et al.(1989). PCR products were purified with the QiAEX II KIT (Qiagen). The sequencing reaction was performed using the Big Dye Terminator (version 3.0) Cycle Sequencing Kit, with AmpliTaq® DNA Polymerase (Applied Biosystems). Sequences were run in an automatic Sequencer ABI377 (Applied Biosystems). Sequence analyses were performed from positions 16024 to 16391 according to the Cambridge Reference Sequence (CRS; Anderson et al., 1981; Andrews et al., 1999), and they are available in <http://www.upf.es/cexs/recerca/bioevo/index.htm>

In addition to the control region sequence, four single nucleotide polymorphisms (SNPs) in the mtDNA coding region (positions 10400, 10873, 12308, and 12705) were determined in those individuals that were difficult to

classify unambiguously into haplogroups based only in the information provided by the HVS-I. A single-base primer elongation method was used to genotype these four positions, and primers and PCR conditions are described elsewhere (Comas et al., submitted). Information yielded by each of the four SNPs allowed us to assign sequences to one of four major haplogroups: 10400T defines the major M haplogroup, 10873C defines the major L haplogroup, 12308G defines the major U haplogroup (including the K haplogroup), and 12705C defines the major R haplogroup, which includes a large set of haplogroups (H, V, J, T, U, B and F).

Phylogenetic analyses

DNA sequences were aligned using the CLUSTAL V program. Sequences were classified into haplogroups according to the nomenclature of Richards et al.(2000) and Salas et al. (2002). Genetic diversity measures were calculated (from position 16024 to 16383) with the Arlequin package 2.0 (Excoffier et al., 1992) and compared to those in a set of North African populations (Table 1). Analyses of molecular variance (AMOVA) were also performed with Arlequin 2.0. Genetic distances between populations using the HVS-I were calculated by intermatch-mismatch pairwise differences according to the equation $D = d_{ij} - (d_{ii} + d_{jj})/2$ (Nei 1987), where d_{ij} is the mean pairwise differences between populations i and j , and d_{ii} and d_{jj} are the mean pairwise differences within populations i and j respectively. The distance standard errors were computed by resampling nucleotide positions with 1,000 bootstrap iterations (Efron 1982). The distance matrix was also employed to build a multidimensional scaling using the STATISTICA 6.0 package. Principal component and correspondence analyses were obtained from haplotype frequencies using the SPSS package.

The network relating HVSI sequences within some of the haplogroups described was constructed by using a reduced-median algorithm (Bandelt et al., 1995) as implemented in the Network 3.0 program. The dating method employed (Saillard et al., 2000) is based on the average number of mutations accumulated from an ancestral sequence as a linear function of time and mutation rate. This method was also performed with the Network 3.0 program.

RESULTS

Haplogroup assignment was performed comparing HVS-I sequences with the data set of Richards et al (2000) and Salas et al (2002), and their classification was used. All Chenini-Douiret sequences were classified unambiguously using the information yielded by the HVS-I variable positions. One sequence from Sened and nine from Matmata, that could not unambiguously be assigned into haplogroups based only on the data from the control region, were typed for positions 10400, 10873, 12308, and 12705 in the coding region. From these, the sequence from Matmata with the following mutational pattern 16232-16293-16356 could not be typed for these four positions due to lack of DNA and was classified as “other”. Haplogroup frequencies are reported in Table2.

Sequence gene pool in Tunisian Berbers

The Tunisian Berber mtDNA gene pool is constituted by sequences belonging to several major haplogroups: the sub-Saharan African L group, the east African M1 haplogroup, and the West Eurasian major haplogroups N and R.

Sub-Saharan sequences, represented by haplogroups L1, L2, and L3, are found in the three Tunisian Berber populations at high frequencies: 26.6% in the Berbers from Sened, 24.3% in the Berbers from Matmata, 13.3% in the Berbers from Chenini-Douiret, showing a large sub-Saharan gene flow among these three Berber populations. Most of the sub-Saharan sequences found were not described in the data sets of Salas et al. (2002), although most of them differ from an already described sequence by one or two mutation steps. It is noteworthy the presence of eight sequences, out of the 17 L3 sequences found in Tunisian Berbers, that harbour

a transition at nucleotide position 16041. This transition has been shown in other Berber and Arab samples from North Africa and its presence is negligible in other populations (two South-Western Europeans, three West Africans and one East African), suggesting a North African origin for this group of sequences, although a sub-Saharan origin cannot be rejected. The network of sequences bearing the 16041 transition (Figure 2) shows a clear star-like phylogeny and a recent origin for this group dated around 10500 years (SE 3500 years).

The M1 haplogroup, to which an East-African origin is attributed (Quintana-Murci et al., 1999), is represented by a single sequence in six Sened Berbers (10.9%), in one individual from Matmata, and it is absent in Berbers from Chenini-Douiret. This sequence was found among Arabs from Tunisia but not in Algerians (Plaza et al., 2003). This sequence was found in other populations (Quintana-Murci et al., 1999; Richards et al., 2002) with a transversion at position 16183, which may be the result of the hypervariable length polymorphism in the poly-C tract of the control region (Bendall & Sykes, 1995). The presence of a single M1 sequence at high frequency in Tunisian Berbers could be explained by gene flow from East Africa followed by genetic drift.

U6 originated in N Africa ~40,000 years ago (Macaulay et al., 1999) and is found in Moroccan Berbers at 6-8%, reaching 28% in the isolated Mozabites from Algeria. In Tunisian Berbers it is found at 7.6% in Sened, 2.0% in Matmata and was absent in Chenini-Douiret; it has been found at 4.2% in Tunisian Arabs (Plaza et al., 2003). Thus, U6 frequencies in Tunisian Berbers are relatively low and may mark an eastward decline in the frequencies of this haplogroup. All U6 lineages found in Tunisian Berbers belong to the U6a (characterised by 16278T), and U6a1 (characterised by 16189C and by 16278T) subgroups.

Haplogroup U/K, excluding U6 sequences, was found in 9.5% of the Sened sequences, 10.2 % in the Matmata and 20.8% in Chenini-Douiret. Haplogroup U is represented by U1a, U3, U5, U5a1a, U7a, and K sequences. The high frequency of this haplogroup in the Chenini-Douiret sample is due to the presence of eight sequences belonging to haplogroup K. Three of them belong to the K root type, and the other five to the haplogroup K2, which is represented by two haplotypes. Haplogroup U7a reached a frequency of 4.1% in the Berbers from Matmata. This haplogroup is found in Middle Eastern populations such as, Iraqis, Palestinians, Armenians, Druze and Kurds; and some Southern and Eastern Europeans (Richards et al., 2000), but none of the already described sequences matched with those found in the Matmata Berbers. U7a sequences may have been integrated in Matmata Berbers after a wave of migration from the Middle East, as they differ by only one step mutation from the U7a Middle Eastern sequences, whereas three mutation steps separate them from the European U7a sequences. Haplogroup U3 is mainly found in populations from the Middle East (Richards et al., 2000), and was also reported in Chenini-Douiret Berbers at 5.7%, and in Matmata Berbers at 2.0%. But no match was found between Tunisian Berbers and Middle Eastern U3 sequences, nor between U3 sequences among Tunisian Berbers.

H is the most frequent haplogroup in most West Eurasian (Richards et al., 2000) and North African (Plaza et al., 2003) populations, as well as in Tunisian Berbers, where H sequences (and HV sequences, since sometimes they cannot be directly distinguished neither by HVS-I nor by the four coding positions typed in the present study, and the typing of position 7025 would be recommended) are found at a high frequency: 28.3% in the Berbers from Chenini-Douiret, 24.5% in the Berbers from Sened, and 26.5% in the Berbers from Matmata. Haplogroup V, which is largely distributed in Western Mediterranean populations (6% in NW Africa, 5% in Iberian

Peninsula, and 3,2% in Italy; Plaza et al., 2003), is displayed only in the Matmatia with a high frequency of 16.3%. Haplogroup V is absent in the two other Tunisian Berber populations, as it is also the case for Tunisian and Algerian Arabs. This value is comparable with those found in the Saharawi (17.9%) and higher than those found in the Basques (10.4%) and other Western European samples (Plaza et al., 2003). 8.2% of the V sequences of the Matmata Berbers bear only the variant 16298C, and the rest of the sequences bear one additional variant and did not match with previously described V sequences (Richards et al., 2000). It is also relevant to note the high frequency of the haplogroup HV1 (15.1%) in Berbers from Chenini-Douiret, represented by a unique sequence.

Haplogroup J is found at very low frequencies in the three samples, and except one Matmata sequence that can be classified into the J1 group, the rest of sequences remain in the J root type. All the T sequences found in the present sample set, except one individual from Sened and one from Chenini-Douiret who bear a T root type, belong to the T1 subgroup. The distribution of haplogroup T in the three Tunisian Berber samples is very different. The frequency of T sequences is very low in Matmata (4.1%) and Sened Berbers (3.8%), whereas haplogroup T is represented in the Chenini-Douiret sample by 17 individuals (32.1%) carrying six different haplotypes. Of those, the most frequent was found in nine individuals and contains a deletion at position 16193, which was also observed in a single one-step derivative. This sequence type, without the deletion, was shared with Moroccan Berbers, Mozabites and Egyptians, and is also frequent in Europe and the Middle East.

Other haplogroups that are mainly found in the Middle East, are also present in Tunisian Berbers: 8.2% of Matmata Berbers and one individual from Chenini-Douiret belong to N1b haplogroup, and 7.5% of Sened Berbers displayed I2 sequences.

Population structure

An analysis of the molecular variance (AMOVA) was performed with the sample data set described in Table 1. When all populations were considered as a single group, 4.14% of the variance was attributed to differences among populations, and 95.86% of the genetic variance was found within populations (Table 3). These populations were more diverse than sub-Saharan and West Mediterranean populations, where 98.8% (Salas et al., 2002) and 97.4% (Plaza et al., 2003) of the variance was found within the respective population sets. However, this could be due to the sampling of one or a few outlier populations that are not diluted in a high number of populations such those used in the sub-Saharan and West Mediterranean analyses.

Several population groups based on cultural and geographic criteria were performed in order to detect some population structure. Populations were first clustered according to linguistic and cultural criteria (Arab vs. Berber), and AMOVA was performed within each group. Berber populations were more diverse from each other than Arabs were: 6.24% of the genetic variance was due to differences among Berber samples whereas 1.54% was attributed to differences among North African Arabs. When this two groups were confronted, genetic variation between Arabs and Berbers was non-significantly different from 0, and 4.25% of the genetic variance was due to differences within Arabs and Berbers. When the analysis was also performed removing Mozabites and Tuaregs, known isolate Berbers where drift has acted strongly, the variation among groups remains non-significant, whereas the variation within Arabic and Berber groups decreases to 3.10% ($p < 0.01$). This result shows that, even when known outliers are removed, a certain degree of heterogeneity remains within groups, whereas no genetic differences are revealed between Arabs and Berbers.

We next classified Berber populations according to their geographical origin (Morocco, Algeria, and Tunisia), and no significant differences were found among groups. Nonetheless, the variance among populations of the same country was significantly high (4.98%, $p < 0.01$), reflecting heterogeneity among the Berber populations of the same geographical area. Since Mozabites are known outliers that could contribute disproportionately to interpopulation variance, they were removed from the analysis (Table 3). The apportionment of the variance due to differences within groups dropped to 4.79%, but were still highly significant ($p < 0.01$).

In order to detect a possible east-west differentiation in Northwest African samples suggested previously (Plaza et al., 2003), two geographical groups (Morocco versus Algeria-Tunisia) were formed, pooling Arabs and Berbers (Table 3). 4.36% ($p < 0.01$) of the genetic variance was attributed to differences within groups, whereas no significant differences were found between groups.

In summary, AMOVA showed that Tunisian Berbers were highly heterogeneous populations.

Tunisian Berbers within the North African genetic landscape

Measures of genetic diversity are reported in Table 1. Berbers from Chenini-Douiret present the lowest sequence diversity value of the geographical region considered, even lower than the diversity presented by the Mozabites, although its confidence interval overlaps widely with all others. However, the mean pairwise difference in Chenini-Douiret is not lower than in other populations. This result points to a micro geographic differentiation among the Tunisian Berbers due to genetic drift since some of the sequences (coming from a variety of haplogroups, and hence the relatively high average pairwise difference) have high frequencies in the sample. When

compared with the rest of the North African samples, the two other Tunisian Berber populations displayed haplotype diversity values that are in the range of the observed values.

In order to establish the genetic relationship between the Tunisian Berbers and the rest of North African populations, the genetic distance matrix between populations based on individual sequences was represented as a bidimensional plot by means of multidimensional scaling (MDS; Figure 3). The MDS plot isolates the Tuareg, the Mozabites and the Berbers from Chenini-Douiret from the rest of the populations, displaying each of them at one extreme of the plot. The Berbers from Matmata are close to the South and Centre Moroccan Berbers, whereas the Berbers from Sened are adjacent to Arab populations. A principal co-ordinate analysis, based on the same genetic distance matrix yielded very similar results (data not shown).

On the basis of the frequency of haplogroups, a correspondence analysis was performed (Figure 4). The Berbers from Chenini-Douiret are associated in the plot with haplogroups T, and HV, the Tuareg with L2; and Mozabites with U6 and V, each at one extreme of the plot, whereas the rest of populations are situated in the centre of the plot. A principal component analysis produced similar results (not shown). Thus, it seems that the heterogeneity detected by AMOVA is mostly contributed by the Chenini-Douiret sample.

DISCUSSION

The mtDNA haplogroup composition of Tunisian Berbers offered a similar picture to that in other Northern African populations (Rando et al., 1999; Plaza et al., 2003), with ~57% of sequences found in haplogroups of broad West Eurasian distribution, ~26% of sub-Saharan origin, ~14% of recent Middle Eastern origin, and ~3% locally originated in N Africa (that is, U6, although the L3 lineage with a transition at 16041 should probably be added to this category).

Few L sequences found in Tunisian Berbers were described in sub-Saharan African samples (Salas et al., 2002), but they differ from one or two positions from already described lineages, giving some evidence of sub-Saharan admixture. Nevertheless, the group of sequences belonging to L3 group with the substitution at position 16041 was also found in other North African samples, such as Moroccan Arabs, South Moroccan Berbers, and Algerians, but no match was found with sub-Saharan populations. This result points to a recent origin of this group of sequences, with an estimated origin around 10500 years, rather than to an ancient sub-Saharan gene flow. The presence of these L3 lineages and the U6 haplogroup suggests that human populations in North Africa have experienced several population expansion processes after the occupation of the region by anatomically modern humans, and the extant populations are not only the result of external migrations, from the Middle East, sub-Saharan Africa and Southern Europe into the area.

The high frequency of the East African haplogroup M1 in Sened is similar to that in Somalians (11%; Watson et al., 1997), Nubians (10%; Krings et al., 1999), and Ethiopians (10%; Thomas et al., 2002), nonetheless this haplogroup is represented by a single haplotype. The same pattern is shown by other Berber samples, with single

haplotypes reaching high frequencies, such as two sequences belonging to HV1 and T1 haplogroups in Chenini-Douiret, or two U6a1 sequences in the Mozabites. This result suggests that the introduction of these sequences was followed by drift probably due to isolation of Berber populations, which led to different haplogroup frequencies and yielded a high degree of heterogeneity among Berber groups.

The three Tunisian Berber populations analysed for the mtDNA control region are characterized by a high genetic heterogeneity, despite their geographic proximity and their common culture and history. Although they are distant by ~100 Km, they present similar haplogroups but at different frequencies. Tunisian Berbers are more heterogeneous than Moroccan Berbers, since an AMOVA within each group shows an apportionment of the variance in Moroccan Berbers of 2.47% ($p < 0.01$) compared to 6.67% ($p < 0.01$) in Tunisian Berbers. Nonetheless, when Chenini-Douiret Berbers are excluded from the analysis, similar values to those of Moroccan Berbers are found in Tunisian samples (3.97%, $p < 0.01$; data not shown). Therefore, the analyses performed point the Berbers from Chenini-Douiret as an outlier sample within Berbers, similar to what is observed in Mozabites. Nonetheless, the particularity of these Berber samples is not the result of the presence of unusual haplogroups and/or sequences in the region (Plaza et al., 2003). Therefore, the outlier position of these Berber groups may be the result of a relatively recent isolation process that have drifted sequence frequencies to unusual values compared to the rest of samples within the geographical area.

From the three Tunisian Berbers, the Berbers from Chenini-Douiret seem to be the most genetically isolated group with a high West Eurasian component in their mtDNA gene pool. This result is in accordance with studies based on the analysis of the

polymorphic Gm system (Chaabani & Cox, 1988; Helal et al., 1988) of the Tunisian populations, which has also a very defined phylogeography. Sub-saharan African Gm haplotypes were present at a low frequency (0.07) among the Berbers from Chenini-Douiret, whereas the West Eurasian haplotypes were much more frequent (0.62). The isolation of this Berber group may explain their variant haplogroup frequencies, low sequence diversity, and lower sub-Saharan African contribution.

The cultural differentiation present in North Africa between Berber and Arab samples seems not to reflect genetic differences between both groups as shown in the AMOVA analyses, and the MDS and PC analyses. If Arabs in Northern Africa were mostly descendants of Middle Eastern Arabs, the frequencies of haplogroups such as N, U1, U3, U7, and HV that are much more prevalent in the Middle East than elsewhere should be larger in N. African Arabs than in Berbers. However, the contrary is observed: these haplogroups add up to 5% in N. African Arabs but to 10% in Berbers. Drift in some of the more isolated Berber populations could explain this observation. The lack of differentiation between North African Arabs and Berbers has also been observed using other genetic markers such as classical markers (Bosch et al., 1997); autosomal STRs (Bosch et al., 2000), *Alu* insertion polymorphisms (Comas et al., 2000); and Y-chromosome lineages (Bosch et al., 2001). This pattern suggests that the Arabization of the area was mainly a cultural process rather than a demographic replacement of the Berber populations that inhabited the region where the Arabic expansion took place.

The present data has failed to confirm an east-west differentiation of North African populations as previously suggested using mtDNA sequences (Plaza et al., 2003) or other genetic markers (Bosch et al., 1997). The present mtDNA data show a more patchy genetic landscape of North Africa, with some Berber samples acting as

outliers in the general North African landscape. The lack of mtDNA data of large geographic regions like the Kabylie (Algeria) and Libya, and the large number of Berber isolated samples considered in the present analysis may decrease the power of finding the longitudinal differentiation shown by other studies.

Acknowledgements

We especially thank the blood donors for participating in the present study. We express our appreciation to David Goldstein for providing information from published sequences, Meherzia Ben Fadhel and Mònica Vallés for technical support. This work was supported by le Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de la Technologie de la Tunisie, by Dirección General de Investigación, Ministerio de Ciencia y Tecnología in Spain grant BOS2001-0794, and by Departament d'Universitats, Recerca i Societat de la Informació, Generalitat de Catalunya grant 2001SGR00285 and fellowship 2000FI00696 to S.P.

References

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, HJ., Forster, P., Sykes, BC., Richards, MB. (1995) Mitochondrial portraits of human populations using median networks. *Genetics*.**141**, 743-753.

Bendall, K.E. & Sykes, B.C. (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet.***57**,248-256.

Bosch, E., Calafell, F., Perez-Lezaun, A., Comas, D., Mateu, E., Bertranpetit, J. (1997) Population history of North Africa: evidence from classical genetic markers. *Hum. Biol.***69**, 295-311.

Bosch, E., Calafell, F., Perez-Lezaun, A., Clarimon, J., Comas, D., Mateu, E., Martinez-Arias, R., Morera, B., Brakez, Z., Akhayat, O., Sefiani, A., Hariti, G.,

Cambon-Thomsen, A., Bertranpetit, J. (2000) Genetic structure of north-west Africa revealed by STR analysis. *Eur. J. Hum. Genet.***8**, 360-366.

Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A., Bertranpetit, J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am. J. Hum. Genet.***68**, 1019-1029.

Brakez, Z., Bosch, E., Izaabel, H., Akhayat, O., Comas, D., Bertranpetit, J., Calafell, F. (2001) Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Ann. Hum. Biol.***28**, 295-307.

Brett M. & Fentress E. (1996) *The Berbers*. Blackwell, Oxford.

Chaabani H., Helal A.N., van Loghem E., Langaney A., Benammar Elgaaied A., Rivat Peran L., Lefranc G. (1984) Genetic study of Tunisian Berbers. I. Gm, Am and Km immunoglobulin allotypes and ABO blood groups. *J Immunogenet.* **11**, 107-113.

Chaabani H. & Cox D.W. (1988) Genetic characterization and origin of the Tunisian Berbers. *Hum Her.* **38**, 308-316.

Close A. E. & Wendorf F. (1990) *North Africa at 18000 BP. Vol 2: Low latitudes*. Unwim Hyman, London.

Comas, D., Calafell, F., Benchemsi, N., Helal, A., Lefranc, G., Stoneking, M., Batzer, M.A., Bertranpetit, J., Sajantila, A. (2000) Alu insertion polymorphisms in NW Africa

and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum. Genet.***107**, 312-319.

Côrte-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**, 331-350.

Dupanloup, I.(1993). *Effets de l'adoption de l'agriculture et de l'élevage sur les populations proche- orientales et africaines*. Thesis, University of Geneva.

Efron, B.(1982) *The Jackknife, the bootstrap and other resampling plans*. Philadelphia, Pa: Society for industrial and Applied Mathematics.

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.

Ferembach, D. (1985) On the origin of the Iberomaurusians (Upper Palaeolithic: North Africa). A new hypothesis. *J. Hum. Evol.* **14**: 393-397.

Helal AN, Rivat-Peran L, Van Loghem E, de Lange G, Langaney A, Lefranc G. (1988) Further contribution of common Gm*-Am* haplotypes and Km* alleles in the characterization of the Tunisian population. *Exp Clin Immunogenet.***5**, 1-14.

Julien, CA., (1961) Histoire de l'Afrique du Nord. Paris: Payot.

Krings, M., Salem, AE., Bauer, K., Geisert, H., Malek, AK., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Utermann, G., Sajantila, A., Paabo, S., Stoneking, M. (1999) MtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am. J. Hum. Genet.***64**, 1166-1176.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B., Torroni, A. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.***64**, 232-249.

Nei, M. (1987) *Molecular evolutionary genetics*. New York: Columbia University Press.

Newman, J.L. (1995) *The peopling of Africa: a geographic interpretation*. New Haven: Yale University Press.

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., Comas, D. (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. *Ann. Hum. Genet.* **67**, 312-328.

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**, 437-441.

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**, 531-550.

Rando JC, Cabrera VM, Larruga JM, Hernandez M, Gonzalez AM, Pinto F, Bandelt HJ. (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet.***63**, 413-428.

Renfrew, C. (1991) Before Babel: speculations on the origins of linguistic diversity. *Cambridge Archaeological Journal***1**, 3-23.

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. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.***67**, 1251-1276.

Richards M, Macaulay V, Torroni A, Bandelt HJ. (2002) In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet.* **71**, 1168-1174.

Saillard, J., Magalhaes, PJ., Schwartz, M., Rosenberg, T., Norby, S. (2000)

Mitochondrial DNA variant 11719G is a marker for the mtDNA haplogroup cluster HV.

*Hum. Biol.***72**, 1065-1068.

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, 1082-1111.

Thomas MG, Weale ME, Jones AL, Richards M, Smith A, Redhead N, Torroni A,

Scozzari R, Gratrix F, Tarekegn A, Wilson JF, Capelli C, Bradman N, Goldstein DB.

(2002) Founding mothers of Jewish communities: geographically separated Jewish

groups were independently founded by very few female ancestors. *Am J Hum Genet.* **70**,

1411-1420.

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC. (1989)

Mitochondrial DNA sequences in single hairs from a southern African population. *Proc*

Natl Acad Sci U S A. **86**, 9350-9354.

Watson, E. (1996) *Threads from the Past: a genetic study of African ethnic groups and*

human origins. PhD dissertation. Massey University, New Zealand.

Watson, E., Forster, P., Richards, M., Bandelt, HJ. (1997) Mitochondrial footprints of

human expansions in Africa. *Am. J. Hum. Genet.***61**, 691-704.

Table1. *Genetic diversity parameters in North African samples*

Populations	Code	Ref.	n	H	π
Matmata Berbers	TB M	1	49	0.964 \pm 0.021	5.05
Sened Berbers	TB S	1	53	0.975 \pm 0.011	7.53
Chenini-Douiret Berbers	TB CD	1	53	0.939 \pm 0.017	6.82
Tunisian Arabs	Tun	2	47	0.990 \pm 0.009	6.15
Mozabites	Moz	3	85	0.942 \pm 0.010	4.73
Algerian Arabs	Alg	2	47	0.956 \pm 0.014	5.72
Center Moroccan Berber	CB1	2, 4	64	0.968 \pm 0.013	4.51
Center Moroccan Berber	CB2	8	60	0.984 \pm 0.009	6.00
South Moroccan Berbers	SB	5	50	0.961 \pm 0.018	4.60
Moroccan Arabs	MA	2, 4	50	0.993 \pm 0.006	7.04
Mauritanians	Mau	4	30	0.975 \pm 0.017	6.09
Egyptians	Eg	6	68	0.992 \pm 0.005	7.06
Saharawi	Sah	2; 4	81	0.982 \pm 0.006	5.44
Tuareg	Tg	7	26	0.985 \pm 0.014	7.10

H: sequence diversity; π : average number of pairwise differences

References: 1: Present study; 2: Plaza et al. (2003); 3: Côté-Real et al. (1996), 4 : Rando et al. (1998), 5: Brakez et al. (2001), 6 : Krings et al. (1999), 7: Watson et al. (1996), 8 : Thomas et al. (2002)

Table 2. MtDNA haplogroup frequencies in Tunisia

Haplogroups	TB CD N= 53	TB S N= 53	TB M N= 49	Tunisian Arabs ^a N=47
L1b*	-	-	2.0	-
L1b1	-	5.7	-	-
L1c1	3.8	-	-	-
L2*	-	1.9	-	2.1
L2a	1.9	1.9	2.0	4.3
L2a1	-	3.8	-	4.3
L2a1a	1.9	3.8	2.0	-
L2b1	-	-	-	2.1
L3*	-	1.9	16.3	2.1
L3b	3.8	3.8	-	-
L3b1	-	1.9	-	-
L3d	-	-	2.0	2.1
L3e1	-	-	-	2.1
L3e2	-	1.9	-	4.3
L3f	1.9	-	-	4.3
M1	-	11.3	2.0	4.3
N1b	1.9	-	8.2	-
I2	-	7.5	-	-
W	-	-	-	2.1
X	-	-	-	2.1
pre-HV	-	5.7	-	-
HV	15.1	-	-	6.4
H*	13.2	24.5	26.5	23.4
V	-	-	16.3	-
J*	3.8	3.8	2.0	-
J1	-	-	2.0	2.1
J2	-	-	-	2.1
T*	1.9	1.9	-	-
T1	30.2	1.9	4.1	2.1
T2	-	-	-	2.1
T3	-	-	-	2.1
K*	5.7	-	4.1	4.3
K2	9.4	-	-	2.1
U*	-	1.9	-	-
U1a	-	3.8	-	4.3
U3	5.7	-	2.0	-
U5*	-	1.9	-	-
U5a1a	-	1.9	-	-
U5b	-	-	-	2.1
U6a*	-	1.9	2.0	4.3
U6a1	-	5.7	-	-
U7a	-	-	4.1	-
Other	-	-	2.0	6.4

^a Data from Plaza et al., 2003

Table 3. *Analyses of Molecular Variance (AMOVA) in North African samples*

Groups	Among groups	Among populations within groups	Within populations
All populations		4.14**	95.86**
Arabs		1.54**	98.46**
Berbers		6.24**	93.76**
Moroccan Berbers		2.47**	97.53**
Tunisian Berbers		6.67**	93.33**
Arabs vs Berbers	-0.21 ns	4.25**	95.95**
Arabs vs Berbers ^a	0.04 ns	3.10**	96.86**
Moroccan Berbers vs Algerian (Moz) vs Tunisian Berbers	0.60 ns	4.98**	94.42**
Moroccan Berbers vs Tunisian Berbers	-0.16 ns	4.79**	95.37**
Morocco vs Algeria-Tunisia	-0.34 ns	4.36**	95.98**

**($p < 0.01$); *($p < 0.05$); ns : not significant

^a Mozabites and Tuareg excluded from the Berber group

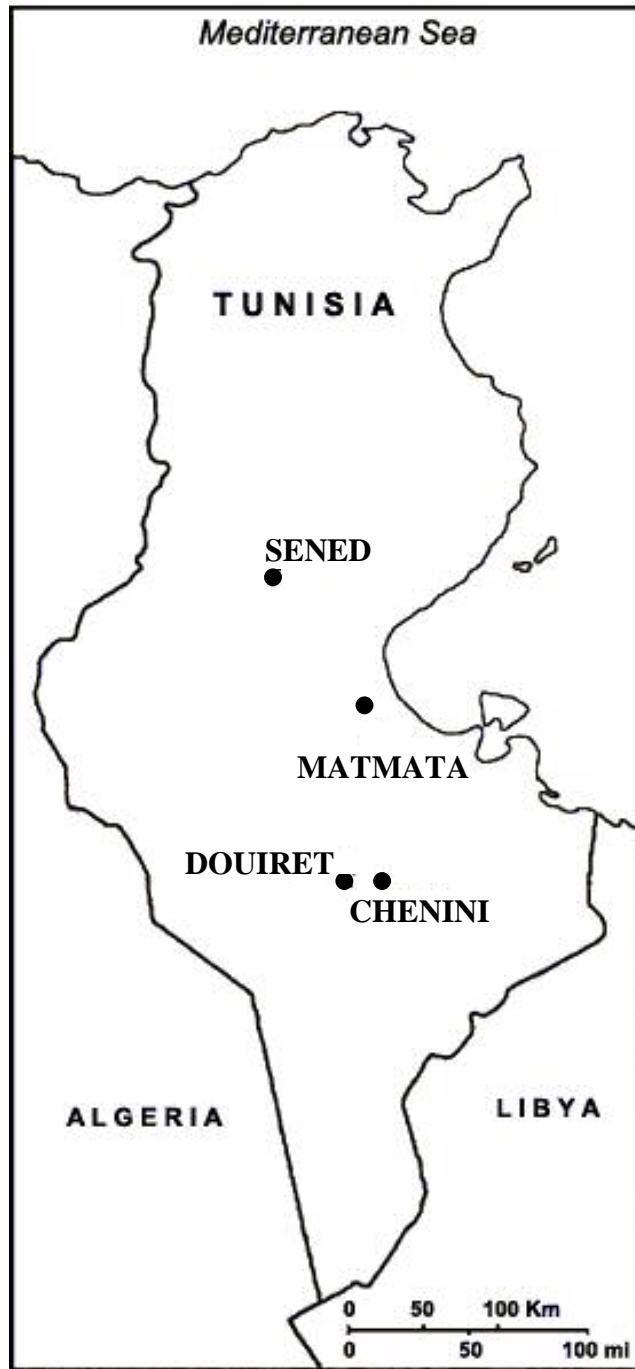
Legends to Figures

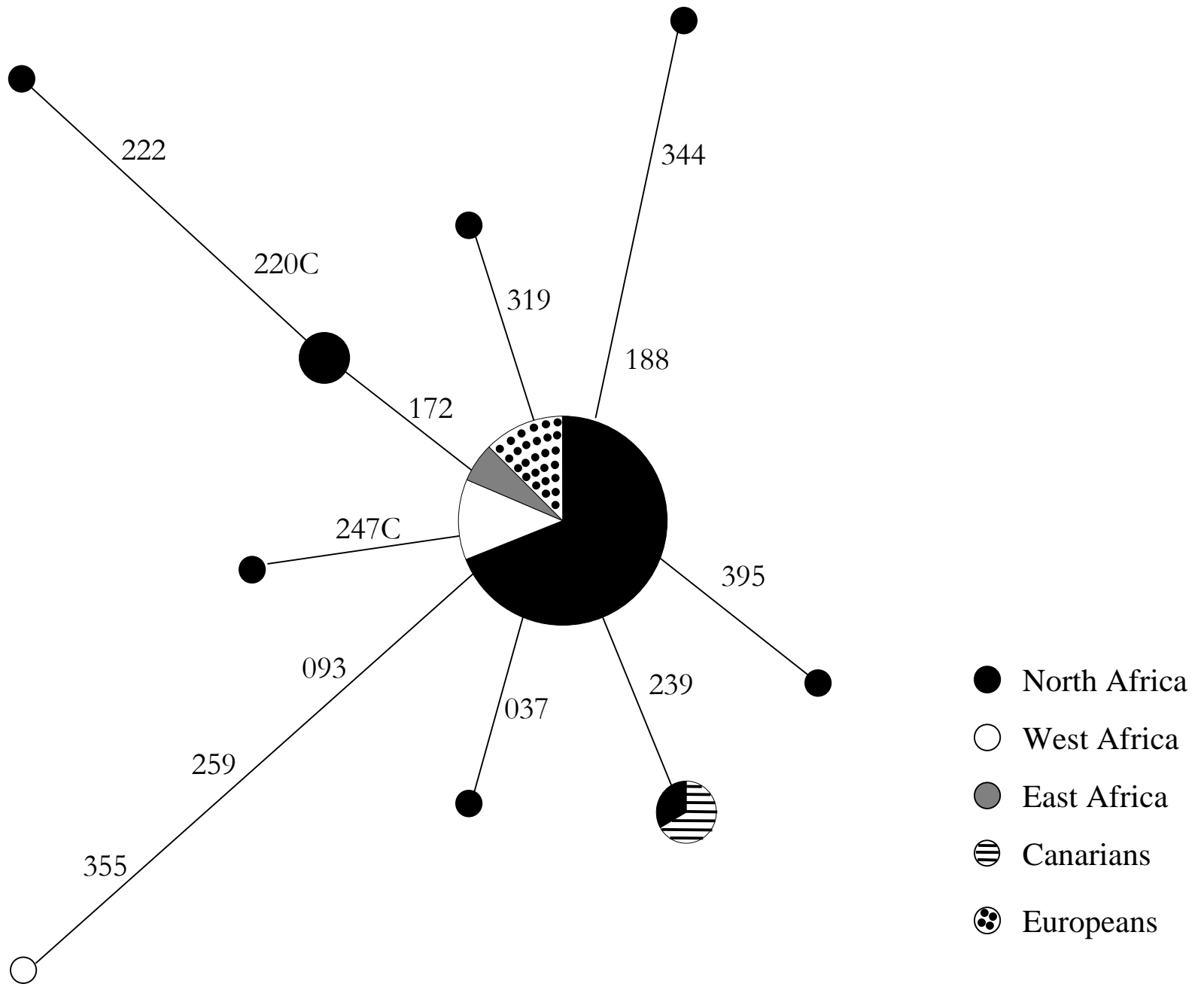
Figure 1. Geographical location of the localities sampled.

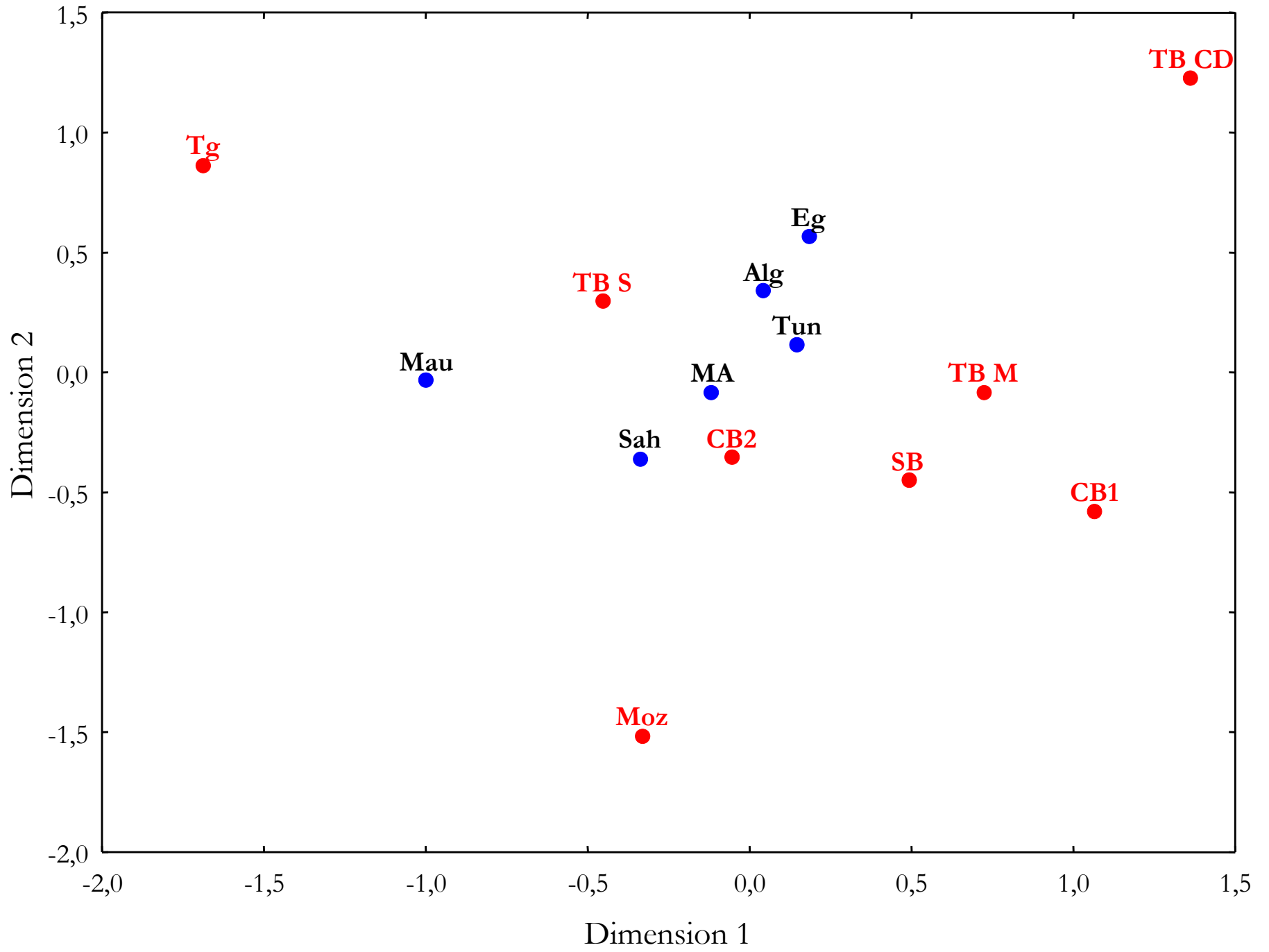
Figure 2. Phylogenetic network of a section of haplogroup L3 bearing a transition in position 16041. The size of the circles is proportional to the number of sequences. Mutated sites (minus 16000) are indicated along the lines. Transversions are indicated by the nucleotide after the number. The sequence marked with an asterisk differs from the Cambridge Reference Sequence by transitions at positions 16041 and 16223.

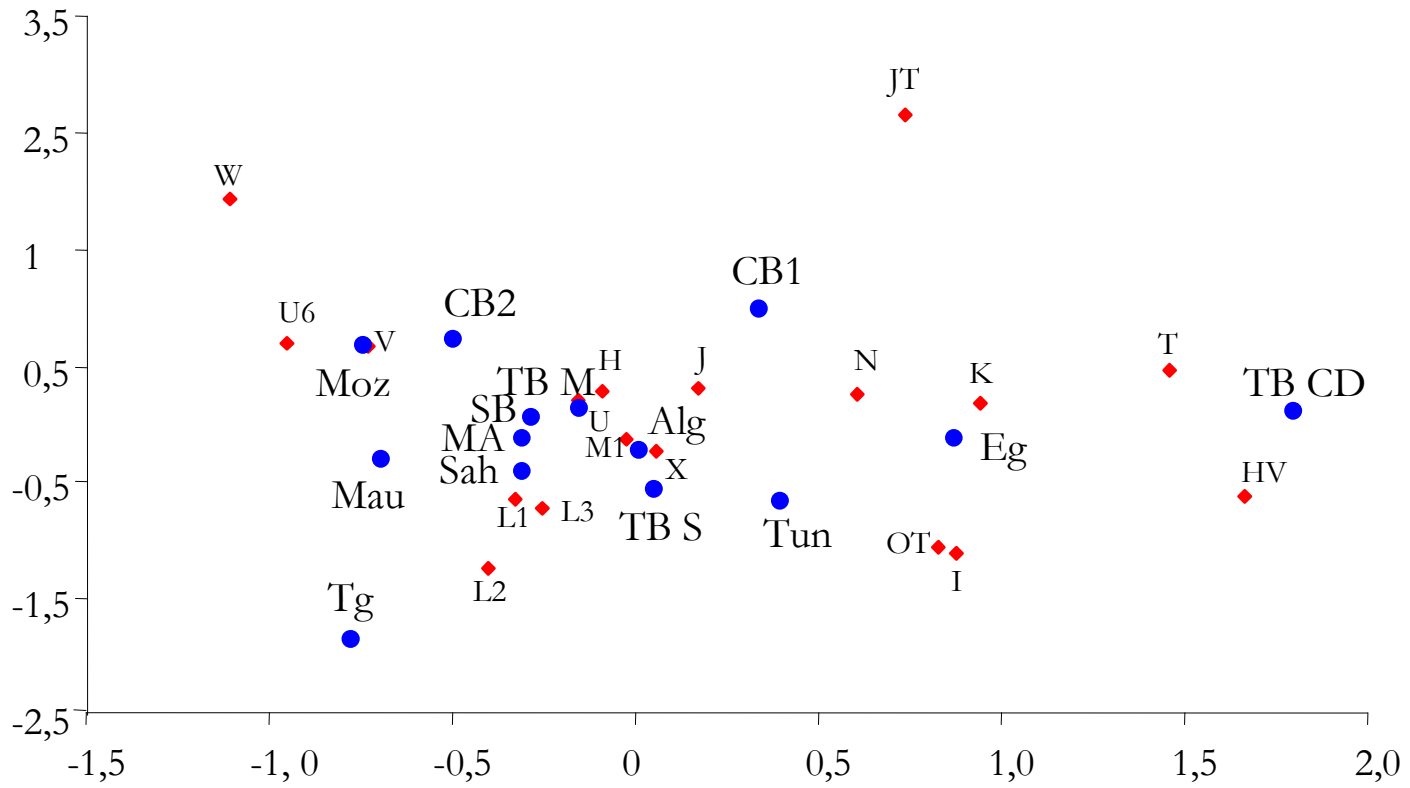
Figure 3. Multidimensional scaling (MDS) analysis based on the genetic distance matrix of North African samples. Abbreviations as in Table 1.

Figure 4. Plot of the analysis of correspondence based on the haplogroup frequencies of North African samples. Circles represent population samples and squares represent haplogroups. Abbreviations as in Table 1.









~ CAPÍTOL IV ~

Insights into the western Bantu dispersal: mtDNA lineages analysis in Angola

S. Plaza, A. Salas, F. Calafell, F. Cortereal, J. Bertranpetit, A. Carracedo,
D. Comas

(sotmès a consideració a Human Genetics)

Insights into the western Bantu dispersal: mtDNA lineages analysis in Angola

Stéphanie Plaza¹, Antonio Salas², Francesc Calafell¹, Francisco Corte-Real³, Jaume Bertranpetit¹, Ángel Carracedo², David Comas¹

¹ Unitat de Biologia Evolutiva. Universitat Pompeu Fabra. Doctor Aiguader 80, 08003 Barcelona, Spain.

² Unidad de Genética, Instituto de Medicina Legal, Universidad de Santiago de Compostela, Galicia, Spain.

³ Instituto de Medicina Legal, Servicio de Biología Forense, Coimbra, Portugal.

Correspondence:

David Comas, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain.

Tel: +34 93 542 28 02; Fax: +34 93 542 28 44. E-mail: david.comas@upf.edu

Key words: Mitochondrial DNA, haplogroup, Bantu expansion, Angola, Africa

ABSTRACT

African is the homeland of humankind and it is known to harbour the worldwide highest levels of human genetic diversity; however, many continental regions, especially in the sub-Saharan side, still remain largely uncharacterised (i.e. southwest and central Africa). Here, we examined the mitochondrial DNA (mtDNA) variation in a sample from Angola. The first and second mtDNA hypervariable segments as well as the 9-bp tandem repeat on the COII/tRNA^{lys} intergenic region have allowed us to allocate mtDNAs to common African haplogroups. Angola lies in the southern end of the putative western branch of the Bantu expansion, where it met the local Khoisan populations. Angolan mtDNA lineages show a basically Bantu substrate with no traces of Khoisan lineages. Roughly, more than a half of the southwestern mtDNA pool can be assigned to West Africa, ~25% to Central Africa and a significant 16% to East Africa, which points to the western gene pool having contributed most of the mtDNA lineages in Angola, but with extensive gene flow from Southeast Africa. This implies that the suggested eastern and western expansion Bantu expansion routes were not independent from each other, and were connected south of the rainforest and along the southern African savannah. In agreement with historical documentation, the analysis also showed that the Angola mtDNA genetic pool shows affinities to the African lineages from Brazil, the main American destination of the slavery from Angola, although not all lineages in Brazil can be accounted for by the Angolan mtDNA pool.

INTRODUCTION

Although the pre-colonial history of many part of Africa has been carefully researched, little is known on the southwestern region that forms contemporary Angola as it was before the arrival of the Europeans in the late 1400s. The area currently known as Angola has been inhabited since prehistoric times, and Khoisan people are thought to have been the first settlers of this territory. According to linguistics, one of the greatest expansion processes that has modelled the African landscape is the dispersal of Bantu languages. Current evidence suggests that the original Bantu homeland was located in the southeastern part of Nigeria (i.e. the Benue valley of southern Nigeria) and/or the northwestern part of Cameroon (i.e. the grassfields of western Cameroon) (Newman, 1995). The Bantu expansion probably coincided with the end of the Neolithic Age (about 5,000 BP) and was at least at some stage related to the diffusion of agriculture and iron metallurgy. The southward Bantu expansion split into two major paths: the western route expanding to the south along the Atlantic coast; and the eastern route, north of the rainforest, to the area of the Great Lakes, and subsequently to the south. Convergence between eastern and western Bantu routes might have occurred at different times and places although its extent is not well characterized. The settlement by Bantu-speaking farmers of the southwestern side of the subcontinent began around the second millennium B.C. , when yam-growers with Neolithic tools spread into the rainforest of Cameroon. Farmers speaking western Bantu languages gradually occupied all of Central Africa expanding over different sorts of terrain (coastal routes and through the rain forest). The dispersal was favoured by the adoption of an iron-based

technology (presumably accompanying the introduction of new crops). It is believed that the western Bantu expansion had important consequences on the demography of the native populations since it marks the first appearance of agriculture, which could have increased the carrying capacity by one order of magnitude (Ammerman and Cavalli-Sforza, 1984). Additionally, it seems that local languages were influenced, and ultimately replaced by Bantu languages (Vansina, 1995). Khoisan speakers may have been completely assimilated by the Bantu expansion or they may have moved towards the south and center of the continent, where they still inhabit part of southern Africa, including southern Angola, mostly in harsh environments like the Kalahari desert.

In colonial times, the Portuguese started slave trade in the African Atlantic coast buying slaves from African chiefs to work in sugar plantations in São Tomé and subsequently to America. Until the Portuguese abolished the slave trade, Angola became the source of as many as two million slaves for the Americas (Thomas, 1997). More than half of these went to Brazil, a third to the Caribbean, and from 10 to 15 percent to the Río de la Plata area on the southeastern coast of South America. As a result of the slave trade, the Angola territory may have lost around 4 million people (Thomas, 1997).

There is wide acceptance of the African origin of modern humans, a fact that increases the interest of the genetic knowledge of African populations, which show higher heterogeneity than any other geographical region and a complex population history that genetics is helping to unravel. A large compilation of classical markers across the African continent (Cavalli-Sforza et al., 1994) showed a clear differentiation between North African and sub-Saharan populations (first

Principal Component [PC]), the relationship between Ethiopian and Khoisan populations (second PC), and the similarity between Bantu populations (third and fourth PC). However, greater resolution has been achieved with the Y chromosome and the mtDNA, which are both uniparental markers with highly resolved phylogeographies. Since Vigilant et al., (1991), numerous African populations have been surveyed for mtDNA variation, whereas fewer studies have focused in the global African Y-chromosome variation (Scozzari et al., 1999; Underhill et al., 2000; 2001). The Y-chromosome landscape in sub-Saharan Africa has been characterized by haplogroups A, B, and part of haplogroup E (E3a) (nomenclature from the Y Chromosome Consortium, 2002), the last one related to the Bantu expansion and dated around 3,000-5,000 ago (Thomas et al., 2000).

The African mtDNA landscape is dominated by lineages belonging to L haplogroups (L0, L1, L2 and L3A; Chen et al., 1995; Watson et al., 1997; Salas et al., 2002); other African specific non-L haplogroups are M1 (with an East African origin; Quintana-Murci et al., 1999), and U6 (specific from North Africa; Rando et al., 1998; Plaza et al., 2003). Several L-mtDNA lineages are present in the Khoisan mtDNA pool (L0d and L0k; Bandelt and Foster; 1997) while others seem to have been dispersed along sub-Saharan Africa by Bantu farmers: L0a (Bandelt et al., 1995; Chen et al., 1995), L3b (Watson et al., 1997), L2, L3e, and L1e (Alves-Silva et al., 2000; Bandelt et al., 2001). The COII/tRNA^{Lys} intergenic 9-bp deletion related to part of the L0a haplogroup was also suggested as an important Bantu marker (Soodyall et al., 1996). Salas et al., (2002) confirmed these findings on the light of an analysis of the African mtDNA variation as a whole, but missing the Angola region. On the base of the composition of the Brazilian lineages available, Salas et al.,

(2002) speculated that the western Bantu expansion likely involved more assimilation of indigenous lineages in the forest zone (mainly in the form of L1c lineages) than the eastern stream. They also postulated the existence of four major founders of both West and East African origin involved in the eastern Bantu expansion (L0a1a, L0a2, L2a1a and L2a1b).

In the African survey of Salas et al., (2002), the overall mtDNA composition for the continent was described, except the uncharacterised Southwest. The analysis of Angola mtDNA could shed light to four main issues related to the African genetic diversity. First, characterize the mtDNA gene pool of southwestern Africa, from which little is known; second, establish the extent of the Bantu demographic expansion in its western part and the possible admixture with Khoisan lineages; third, determine the degree of differentiation between both Bantu expansion routes (West and East) by comparison with the Southeast region; and fourth, determine the contribution of Southwest Africa to the mtDNA pool of the main American destination of Angolan slaves, Brazil.

MATERIAL AND METHODS

Samples, mtDNA amplification and sequencing

A total of 44 unrelated individuals, whose maternal ancestors were known to be originally from Angola, were analysed for both hypervariable segments I (HVSI) and II (HVSII) of the mtDNA control region, and for the COII/tRNA^{lys} 9bp intergenic deletion (positions, 8281 – 8289; Anderson et al., 1981). Total DNA was extracted using a Chelex method (Lareu et al., 1994).

Both hypervariable segments were amplified in one reaction using primers L15996 and H408 (Vigilant et al., 1989), and the amplified product was purified with the Gene Clean kit (BIO 101). The primers used in the sequencing reaction were L15996 and H16401 for HVSI, and L29 and H408 for HVSII (Vigilant et al., 1989). Both hypervariable segments were sequenced with the Big Dye Terminator (version 3.0) Cycle Sequencing Kit, with AmpliTaq® DNA Polymerase (Applied Biosystems), and the sequence products were run in an ABI PRISM 377 sequencer (Applied Biosystems).

The 9-bp tandem repeat (CCCCCTCTA) of the COII/tRNA^{lys} intergenic region was amplified by PCR using the primers and methods described by Comas et al. (2004). The amplified product was run in an automatic sequencer ABI PRISM 377 and the fragment sizes were analysed with the GeneScan software analysis package.

Phylogenetic and population analyses

Sequences from positions 16024 – 16391 and 63 – 322 (according to Anderson et al., 1981) were used in the present analysis and are available in the following web site

(<http://www.upf.es/cexs>). HVSII sequences were not determined for two individuals, due to scarcity of DNA. The information provided by both the HVSI and HVSII was used to classify the sequences into haplogroups according to Salas et al., (2002; cf. Chen et al., 1995; Watson et al., 1997; Rando et al., 1998; Quintana-Murci et al., 1999; Alves-Silva et al., 2000; Bandelt et al., 2001; Pereira et al., 2001; Torroni et al., 2001). Following the suggestion of Richards and Macaulay, (2000) and Mishmar et al., (2003), L1 nomenclature has been changed according to the sequence scheme appeared in Salas et al. (2004).

Sequence diversity (h) was calculated as $h = [n/(n-1)] (1 - \sum_{i=1}^k p_i^2)$, where p is the frequency of each of the k different sequences in the sample, with the Arlequin 2.000 program (Schneider et al., 2000). In order to compare the present results with other populations, data for the first mtDNA hypervariable region (positions 16090-16365) from a number of Sub-Saharan population samples were taken from the literature (Table 1). Several groups and samples from Mozambique (Pereira et al., 2001; Salas et al., 2002) were pooled. Sequences from Cabo Verde islands (Brehm et al., 2002) were also considered as a single population. Other sub-Saharan and American samples appeared in Salas et al., (2002) were also used in parts of the analysis such as sequence sharing. For some analyses, population samples were grouped into major geographic areas: West, East, Central, Southeast and Southwest Africa (Table 1; Figure 1). L sequences from Brazil (Alves-Silva et al., 2000) were used as an additional population sample throughout the analyses.

Table1

Figure1

Population genetic structure was tested through analysis of molecular variance (AMOVA) (Excoffier et al., 1992), using the Arlequin 2000 program (Schneider et al., 2000). A spatial analysis of the molecular variance (SAMOVA) was

also performed using the SAMOVA 1.0 program (Dupanloup et al., 2002) by presetting different numbers of population groups. This approach defines groups of populations that are geographically homogeneous and maximizes the proportion of total genetic variance due to differences between groups.

An analysis of the correspondence was performed from haplogroup absolute frequencies using the SPSS package.

RESULTS

Angolan mtDNA genetic composition

Data for the HVSI and HVSII sequences, and the intergenic COII/tRNA^{lys} 9bp deletion, are shown in Table 2. The genetic diversity found in Angola for the HVSI (0.992 ± 0.007), the HVSII (0.982 ± 0.009) and both hypervariable regions together (0.997 ± 0.006) is similar to the diversity found in other sub-Saharan samples analysed previously (Salas et al., 2002; 2004).

Table2

All the sequences obtained in the present analysis can be assigned to the African specific L lineages (L0, L1, L2 and L3A) and have been classified into haplogroups according to Salas et al., (2002; submitted). Angolan L0, L1, L2 and L3 are found with a relative frequency of 13.6%, 22.7%, 29.5%, and 34.1%, respectively.

Within L0, which includes at least four haplogroups (L0a, L0d, L0k and L0f), only sequences belonging to haplogroup L0a, with the L0a1 and L0a2 subclades, have been found in Angola. The Khoisan subclades L0d and L0k, and the eastern African subclade L0f are not found in the present sample. The presence in Angola of the eastern African L0a1 subclade, which constitutes one tenth of the lineages found in East Africa, might be due to migration from eastern/southeastern Africa. On the other hand, a Central origin was proposed for the L0a2 subclade and has been associated with the COII/tRNA^{lys} 9bp deletion (Soodyall et al., 1996). Two of the three L0a2 sequences found in Angolans carried the intergenic COII/tRNA^{lys} 9bp deletion. The presence of L0a2 sequences in the Southwest (i.e. Angola) and

Southeast (i.e. Mozambique; Pereira et al., 2001; Salas et al., 2002) might be explained by migration from Central Africa.

Haplogroup L1b is known to be most frequent in West Africa (13%), and present as well in African Americans (10%) as a consequence of the African slave trade. Diffusion of this haplogroup seems to have been very limited in Southwest Africa since it is represented only by two individuals (4.5%) in Angola. This haplogroup is also scarce in Brazilian L-sequences (4.6%), where Angolan slaves were massively imported in the 16th-19th centuries.

The presence of the ancient haplogroup L1e is mainly limited to East Africa at a low frequency (4%). It is rare in other parts of Africa: only the subclade L1e2 has been found in two Mozambicans, one Mbuti, and one Egyptian. The single L1e sequence type found in Angola has no match with the rest of L1e African sequences.

The L1c haplogroup was postulated to have originated in Central Africa towards the Atlantic Coast (Salas et al., 2002), since it has been observed at relatively high frequencies in Central Africa as well as in African Americans (up to almost 23% in Brazilians, Alves-Silva et al., 2000), but it is rare elsewhere. As predicted, L1c lineages are frequent in Angola (16%), with three sublineages represented: L1c1 (4.5%), L1c2 (9.1%), and L1c3 (2.3%). Nonetheless, the network of the L1c sequences of Sub-Saharan, African American (Salas et al., 2002), and Angolan lineages (data not shown), locates Angolan L1c lineages at the tips of the branches. This seems to suggest that southwest Africa is not the homeland of L1c. Strikingly, no sequence matches were found between Angolan and African L1c sequence types.

Two of the four subclades of haplogroup L2 were present in Angola: L2a, the most common and widespread L2 subclade in Africa, which accounts for 25% of the Angolan lineages, and L2b (6.8%). In general, Angola presents a similar frequency pattern of haplogroup L2 compared to West Africa, except for the lack of L2c lineages in Angola.

L3A includes at least L3b, L3d, L3e, L3f, and L3g haplogroups (c.f. Salas et al., 2002). L3e is the second most frequent haplogroup in Angola with a frequency ~21%. Four subclades of L3e are found in Angola: L3e1 (6.8%), L3e1a (4.5%), L3e2b (2.3%), L3e3 (6.8%). L3e is not very abundant in East Africa (~ 3%), but it is more prevalent in West (~11%) and Central (~20%) Africa. Eastern Bantu expansion could have carried L3e at significant frequencies to Southeast (~15%), and South (~11%) Africa. As suggested for the L1c haplogroup, the Angolan slaves may have carried part of the diversity of Brazilian L3e haplogroup during the colonial period.

Sequence sharing

Angolan sequences were compared with the sequence dataset of Salas et al., (2002) and including Brehm et al., (2002). Most of the Angolan mtDNA were already found in other sub-Saharan populations. We examined how many identical HVSI sequences (positions 16090-16365) were shared between Southwest Africa and the rest of the African regions, as well as with American sequences. Variation at positions 16182-16185 and length polymorphism at the homopolymeric cytosine stretch were not considered. A total of twenty Angolan HVSI sequences were found in other African regions, being Southeast Africa (data pooled from Pereira et al.,

2001 and Salas et al., 2002) the region with the highest number of matches with Southwest Africa. Among these haplotypes, most were also found in other sub-Saharan populations, and only few are only present in Southwest, Southeast and America: one sequence (AN125, belonging to L2a) was shared only by Southwest and Southeast; three (AN94, AN9, and AN54/74; belonging to L0a1, L3e1, and L3e1a respectively) were found in these two regions and in America, and three haplotypes (AN37, AN5, and AN53; belonging to L0a2, L1c1, and L3g respectively) were shared only between Southwest and America. Only one sequence (AN40, belonging to L2a1a) was present only in West and Southwest regions.

Angola within the African mtDNA landscape

In order to place Angolans within the sub-Saharan African mtDNA framework, a correspondence analysis based on the absolute haplogroup frequencies (Table 3) was performed on the main African regions. The Khoisan (!Kung and Khwe) and the Pygmy (Biaka and Mbuti) samples are outliers due to their different haplogroup composition and sequence ambiguities in the Pygmy samples, and, therefore, were excluded in subsequent analyses. Only sequences belonging to the major sub-Saharan haplogroup L and M1 lineages were included in this analysis.

Table3

The analysis of the correspondence is shown in Figure 2. The first dimension (47.2%) separates the Southeast region from the rest of Africa, a pattern found also by Salas et al., (2002). This region appears isolated at one edge of the plot characterized by the L0a1a, L0a2, L0d, L1c2, L2a1a, L2a1b, L3e1, and L3e3 lineages. The second dimension (38.7%) shows a clear separation between West and

Figure2

East Africa. The eastern populations are mainly associated with lineages such as L0a1, L1e, L3a, L3g, and M1. Central Africa and the Southwest are placed between the other three regions of Africa. Southwest Africans are situated in the principal correspondence plot in between western and eastern African groups, but closer to the west, which suggests a contribution from both Bantu expansion routes to the Angolan genetic pool. A principal component analysis based on the relative haplogroup frequencies was also performed (data not shown) and displayed similar results.

Genetic variation and population structure

An analysis of molecular variance (AMOVA, Table 4) was performed on the sub-Saharan populations used for the correspondence analysis. When all populations were considered as a single group, 5.60% ($p < 0.001$) of the genetic variance was found between populations, showing significant genetic heterogeneity among these populations. In order to ascertain how this genetic structure was partitioned, different grouping criteria were applied. When linguistic affiliation was considered (Niger-Kordofanian/Afro-Asiatic/Nilo-Saharan/Portuguese-Creole), a non-significant 0.27% of the variance was attributed to differences among linguistic groups, which means that the genetic diversity is not structured according to linguistic classification. When the populations were roughly classified into western (western, central and southwestern samples in Table 1) and eastern Africa (eastern and southeastern samples in Table 1), 4.17% ($p < 0.05$) of the genetic variance was attributable to geographic groups, whereas 3.12% ($p < 0.001$) was due to differences among populations from the same geographical area. When the geographical area

Table4

was defined more precisely (western, eastern, central, southwestern, and southeastern Africa), the variance attributable to geographic groups increased to 4.84% ($p < 0.001$), and the differences between populations within the same geographical area decreased to 1.73% ($p < 0.001$). Finally, if the Angolans were grouped with the Central Africans, according to their position in the correspondence analysis, the differences among geographical areas increased to 4.93% ($p < 0.001$), and differences among populations within groups decreased to 1.68% ($p < 0.001$), suggesting a close genetic relationship between Angolans and Central Africans.

A SAMOVA was performed with the sub-Saharan African samples used in the AMOVA in order to define groups of populations that are geographically adjacent, and genetically homogeneous, and to maximize the proportion of genetic variance between them (Table 5). When two groups were sought, the maximum proportion of total genetic variance between groups (5.13%, $p < 0.001$) was found between the whole set of western populations (Table 1) plus São Tomeans, and the rest of populations. When the number of groups is set to four or five, Angola and Mozambique constitute one such group. The absence of any intervening sampled population makes Angola and Mozambique topological neighbours and allows the SAMOVA algorithm grouping them, which highlights their genetic relationship. This result stresses the affinity between the southeast and southwest regions, both at the end of the Bantu expansion routes.

Table5

DISCUSSION

Angola, and the southwestern part of Africa in general, was until now a missing piece in the African genetic puzzle. The analysis of mtDNA lineages in Angola has been used to address four issues: (1) the characterization of the southwestern Africa gene pool; (2) the detection of admixture between Bantu and Khoisan populations; (3) the location of Angola in the mtDNA African landscape; and (4) exploring the putative geographical origin of the African lineages detected in Brazil, the main Angolan slave trade destination.

Characterization of the southwestern Africa gene pool

As expected, all lineages found in Angola belong to the major haplogroup L (L0, L1, L2, and L3A). The Angolan mtDNA gene pool is similar to West/Central Africans with a minor eastern component as shown in the correspondence and the phylogeographic analyses. Moreover, AMOVA has shown that the amount of genetic variation between groups is higher when Angolans are jointly considered with Central Africans.

One of the most intriguing aspects of the Angolan gene pool is the L1c haplogroup. The frequency of L1c was high in Angola (15.9%), as it was expected since these lineages were also frequent in Brazilians and other Afro-American samples. They have also been found in high proportion (~22%) in Central Africa. L1c is much rarer elsewhere: from 0% in South Africa to ~5% in Southeast Africa. A putative Angolan (Alves-Silva et al., 2000), or Central African (Salas et al., 2002) origin for the haplogroup L1c has been postulated. Nonetheless, Angolan sequences

fall in the tips of the L1c network far from the root sequence, and are included in each of the three subclades of the L1c haplogroup. Moreover, none of the L1c Angolan sequence types matches with those described in African samples. Actually, matches for L1c sequences in African American have been mainly found in Central Africa. The Central African populations sampled so far are the islands of Bioko and São Tomé, which are, respectively, inhabited by an old western Bantu isolated group, and descendants of slaves (Mateu et al., 1997). In the mainland, populations studied are a small Fang sample (N=10) from Equatorial Guinea, and two Pygmy samples from the Central African Republic and the Congo Democratic Republic. Thus, most of mainland Central Africa remains to be sampled, including such large countries as Cameroon, Gabon and the Republic of the Congo, which add up to over one million sq. Km. and over 20 million people. A recent report (Destro-Bisol et al., 2004) shows that most of the mtDNA sequences in a western Pygmy population from Cameroon, the Mbenzele, belong to L1c (96.4%, being 9.1% L1c*, 29.1% L1c1a*, and 58.2% L1c1a1); the authors suggest a local origin for the L1c1a1 offshoot. In the same paper, the authors find 22.5% L1c in the non-Pygmy Ewondo (most of it L1c1a1 attributable to gene flow from the neighbouring Mbenzele), and 6% in the Bamileke, also from Cameroon. In summary, the heartland of L1c may lie in the still largely uncharacterised coastal facade of Central Africa (from Cameroon to Republic of the Congo), which may correspond to a secondary focus of the Bantu expansion, from where it may have been in part exported to the Americas. In addition, these results confirm the role of L1c in the southwestern Bantu expansion suggested by Salas et al., (2002) based on the Brazilian L-mtDNA composition.

Lack of Khoisan component in Southwest Africa

Khoisan peoples might have occupied a vast territory before the Bantu expansion, which gradually displaced or assimilated Khoisan speakers. As predicted by Alves-Silva et al., (2000), none of the Khoisan characteristic lineages (L0d or L0k) were found in Southwest mtDNA pool. In addition, note that L0d and L0k have not been found in the large African-American survey performed by Salas et al. (2004). In the extant Khoisan groups, !Kung and Khwe (Chen et al., 2000), L0d and L0k haplotypes constitute around the 36% and 24% respectively, of the lineages. The probability of not finding a particular sequence that is present in a population at a frequency f in a sample of size N is given by $\alpha=(1-f)^N$. Therefore, the maximum contribution of Khoisan lineages in Angolans compatible with the observation of the absence of L0d and L0k in a sample of 44 Angolans would be less than 10.8% (with a confidence of $p=0.05$), which is evidence for a dramatic (and almost complete) replacement of the Khoisan maternal lineages by the Bantu people. However, given the different carrying capacities associated with the hunter-gatherer (Khoisan) and farmer (Bantu) lifestyles, it is expected that, even if the Bantu absorbed all the local Khoisan people, the latter would not have contributed much to the admixture. A larger sample is needed to obtain sufficient power to discriminate between the two extreme hypotheses (no Khoisan admixture vs. complete assimilation). If the African lineages in the Brazilian are taken for proxies of the Angolan mtDNA pool (and actually, their haplogroup frequencies are not significantly different, Fisher's exact test, $p=0.164$), they could be pooled to increase the sample size to 113. In that case, the maximum possible contribution of Khoisans to the extant Angolan mtDNA pool would drop to 4.3%. Therefore, the

present gene pool of Angolans is basically the result of the Bantu expansion within the region with no contribution (or very small) by the Khoisan sequences.

Most of the lineages that are hypothesised to be dispersed by the Bantu are found in SW Africa, such as L0a1, L0a2, L3b, L3e, L2a1a, and L3e (Bandelt et al., 1995; Chen et al., 1995; Watson et al., 1997; Pereira et al., 2001; Salas et al., 2002). This fact added to the lack of Khoisan lineages in Angola, point to a basically Bantu substrate of the extant Angolan gene pool. Thus, the Bantu expansion was clearly more demic (in the sense of population replacement) in the southwest than in the southeast, where remnants of ancient settlers (related to extant Khoisan) are evident.

Angola in the African mtDNA landscape

The Southwest African lineages seem to have originated mostly from West/Central Africa. The analysis of the molecular variance show a clear grouping between Southwest and the Central Africa region. This result suggests a large contribution of the western stream of the Bantu expansion after dispersion and assimilation of indigenous lineages in the equatorial zone. The correspondence analysis displayed a clear separation between West and East Africa, but place the Southwest region in an intermediate position between West, East and Southeast Africa. This suggest that the western and eastern Bantu expansions were not independent, isolated events, but they rather met likely below the tropical forest zone and, then dispersed through the southern areas of Africa. This is also supported by the spatial analysis of the molecular variance (SAMOVA, Table 4), where Angola and Mozambique are jointly clustered in all the analyses with different

present numbers of groups, and they both constituted a separate cluster when samples are divided in four or more groups. Besides their geographic proximity, this reveals a genetic homogeneity between both regions and suggests that they shared a common set of haplogroups brought by the Bantu expansion. A local differentiation of the lineages followed by a gene flow between both regions may have also help to maintain a close relationship between both areas. These results are also in accordance with the African geographic landscape since the equatorial forest seems to have acted as a strong genetic barrier and limited the interaction between both western and eastern Bantu streams. And on the other hand, the southwest and southeast areas are separated by the savannas, easier to cross and more densely populated than the tropical rainforest and, therefore, more permeable to gene flow.

Following the same phylogeographic approach employed in Salas et al., (2002), West Africa would have contributed with ~60% of the southwestern mtDNA composition, Central Africa with a 23% and East Africa with a significant 16%.

African mtDNA contribution to the Brazilian population

As Angola was known to provide the major number of Africa slaves to Brazil (Thomas et al., 1997), inferences on the Angolan mtDNA composition were done on the basis of the Brazilian mtDNA pool (Alves-Silva et al., 2000). Our results confirm the prediction of Alves-Silva et al., (2000) on the Angolan haplogroup composition. Brazilian and Angolan samples share a low number of sequences, as they both displayed high haplotype diversity (0.992 ± 0.007 for Angolans and 0.994 ± 0.004 for Brazilians). When the Brazilian population was introduced as a

single group in the correspondence analysis based on haplogroup frequencies (data not shown), Brazil and Central Africa are clustered as a single group. Therefore, although haplogroup frequencies in Brazil are roughly similar to those in Angola, the African mtDNA pool of the Brazilian population did not come exclusively from Angola, but is likely the result of admixture of African slaves from different colonies distributed in West, Southwest, and Central Africa (Guinea Coast, Saõ Tomé).

Acknowledgements

We thank Mònica Vallés and Òscar Lao, Universitat Pompeu Fabra, for technical assistance and suggestions. The present study was supported by the Dirección General de Investigación, Ministerio de Ciencia y Tecnología, Spain (BOS2001-0794 and BFF2002-10206-E), the European Science Foundation (OMLL: Language, culture and genes in Bantu: a multidisciplinary approach of the Bantu-speaking populations of Africa), and Direcció General de Recerca, Generalitat de Catalunya (2001SGR00285). S.P. received a fellowship from the Direcció General de Recerca, Generalitat de Catalunya (2000FI00696). A.S. is supported by the Isidro Parga Pondal program (Xunta de Galicia).

Table 1. Samples used in the present study

Populations	Geographic Region	Code	Sample size	References
Angola	Southwest	An	44	Present study
Mozambique	Southeast	Mz	416	Pereira et al., 2001; Salas et al., 2002
Cabo Verde	West	CV	292	Bhrem et al., 2002
Mandenka	West	Mn	119	Graven et al., 1995
Fulbe	West	Fu	61	Watson et al., 1996
Hausa	West	Ha	20	Watson et al., 1996
Kanuri	West	Ka	14	Watson et al., 1996
Songhai	West	So	10	Watson et al., 1996
Tuareg	West	Tg	26	Watson et al., 1996
Yoruba	West	Yo	35	Watson et al., 1996; Vigilant et al., 1991
Senegalese	West	Sn	50	Rando et al., 1998
Serer	West	Sr	23	Rando et al., 1998
Wolof	West	Wo	48	Rando et al., 1998
Sudan	East	Su	76	Krings et al., 1999
Nubia	East	Nu	80	Krings et al., 1999
Kikuyu	East	Ki	25	Watson et al., 1996
Somali	East	Sm	27	Watson et al., 1996
Turkana	East	Tk	37	Watson et al., 1996
Ethiopian	East	Et	74	Thomas et al., 2000
Bubi	Central	Bu	45	Mateu et al., 1997
São Tomé	Central	ST	50	Mateu et al., 1997
Fang	Central	Fa	11	Pinto et al., 1996
Mbuti	Central	Mb	20	Vigilant et al., 1991
Biaka	Central	Bk	17	Vigilant et al., 1991
!Kung	South	Kg	67	Vigilant et al., 1991; Chen et al., 2000
Khwe	South	Kw	31	Chen et al., 2000
Brazil	America	Br	69	Alves-Silva et al., 2000

Table 3. L and M haplogroup frequencies found in African regions grouped according to Table 1

Haplogroup ^a	Southwest	West	Central	East	Southeast
M1	-	-	-	8.7	-
L0a*	-	0.3	-	1.6	0.5
L0a1	2.3	0.6	1.0	7.1	0.7
L0a1a	4.5	0.2	2.9	2.4	8.9
L0a2	6.8	-	4.8	3.2	14.7
L0d	-	-	-	0.4	5.1
L1b	4.5	2.0	2.9	0.4	0.5
L1b1	-	11.7	7.7	3.2	0.7
L1c*	-	0.5	1.0	-	0.7
L1c1	4.5	3.7	3.8	0.4	1.4
L1c2	9.1	-	1.9	-	1.9
L1c3	2.3	0.9	9.6	-	1.2
L1e	2.3	-	-	5.6	0.5
L1f	-	-	-	1.6	-
L2a	4.5	5.9	1.9	12.7	1.4
L2a1	13.6	14.5	14.4	8.7	5.1
L2a1a	6.8	0.9	1.0	-	10.1
L2a1b	-	0.6	-	-	16.4
L2b	4.5	4.9	1.0	0.8	1.4
L2c	-	15.1	4.8	-	0.7
L2d	-	1.1	6.7	0.4	0.7
L3a	-	3.2	1.0	22.2	1.0
L3b*	2.3	9.0	1.0	-	1.7
L3b1	-	2.3	-	0.8	-
L3b2	-	0.6	-	1.2	1.2
L3d	4.5	8.1	1.0	0.4	5.5
L3e1	11.4	0.5	9.6	1.2	10.4
L3e2	2.3	4.6	15.4	0.4	1.4
L3e3	6.8	0.6	1.0	0.4	3.4
L3e4	-	5.9	1.9	1.6	0.2
L3f*	2.3	0.8	2.9	4.8	1.7
L3f1	-	1.8	1.0	3.6	0.7
L3g	4.5	-	-	6.0	-

^a Haplogroups classified after Salas et al. (2002, 2004)

Table 4. Analysis of the molecular variance (AMOVA) in Sub-Saharan populations.

	Among groups	Among populations within groups	Within populations
All populations		5.60**	94.40**
Linguistic affiliation	0.27 ns	5.42**	94.31**
Geographical area			
West vs East	4.17*	3.12**	92.71**
Five geographical areas (W, SW, E, SE and Central)	4.84**	1.73**	93.43**
Four geographical areas (W, E, SE and SW+Central)	4.93**	1.68**	93.39**

* p<0.05

** p<0.001

Table 5. Spatial analysis of the molecular variance (SAMOVA) in Sub-Saharan populations. Abbreviations as in Table 1.

Number of groups	Groups	Amon g groups	Among populations within groups	Within populations
2	Western populations + ST The rest of populations	5.13**	2.52**	92.35**
3	Western populations + ST Sm + Et Ki + Su + Nu + Tk + An + Mz + Fg + Bu	5.22 **	2.23**	92.55**
4	Western populations + ST Eastern populations + Fg Bu An + Mz	5.40 **	1.56**	93.04**
5	Western populations + ST Bu Su + Sm + Et Ki + Nu + Tk + Fg An + Mz	5.63 **	1.30**	93.07**

** p<0.001

References

Ammerman AJ, Cavalli-Sforza LL (1984). The Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton, NJ.

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

Alves-Silva J, da Silva Santos M, Guimarões PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* 67:444-461

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753

Bandelt HJ, Forster P (1997) The myth of bumpy hunter-gatherer mismatch distributions. *Am J Hum Genet* 61:980-983

Bandelt HJ, Alves-Silva J, Guimarões 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 (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* 65:549-563

Brehm A, Pereira L, Bandelt HJ, Prata MJ, Amorim A (2002) Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann Hum Genet* 66:49-60

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) History and geography of human genes. Princeton University Press, Princeton, NJ.

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* 57:133-149

Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC (2000) MtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet* 66:1362-1383

Comas D, Plaza S, Wells RS, Yuldasheva N, Lao O, Calafell F, Bertranpetit J (2004) Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet* (in press).

Destro-Bisol G, Coia V, Boschi I, Verginelli F, Caglià A, Pascali V, Spedini G, Calafell F (2004) The analysis of variation of mtDNA hypervariable region-1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. *Am Nat* (in press).

- Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* 11:2571-2581
- 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
- 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
- Krings M, Salem AE, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Pääbo S, Stoneking M (1999) MtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am J Hum Genet* 64:1166-1176
- Lareu MV, Phillips CP, Carracedo A, Lincoln PJ, Syndercombe Court D, Thomson JA (1994) Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci Int* 66:41-52
- Mateu E, Comas D, Calafell F, Pérez-Lezaun A, Abade A, Bertranpetit J (1997) A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea. *Ann Hum Genet* 61:507-518
- 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 (2003) Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci* 100:171-176
- Newman JL (1995) *The peopling of Africa*. Yale University Press. New Haven, CT.
- 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. *Ann Hum Genet* 65:439-458
- Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM (1996) Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann Hum Genet* 60:321-330
- Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* 67:312-328
- 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* 23:437-441

Rando JC, Pinto F, González AM, Hernández 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* 62:531-550

Richards M, Macaulay V (2000) Genetic data and the colonization of Europe: genealogies and founders. In: Renfrew C, Boyle K (eds). *Archaeogenetics: DNA and the population prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge, pp 139-151

Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A (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, A (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum Genet* (in press).

Schneider S, Kueffer JM, Roessli D, Excoffier L (2000) Arlequin (ver 2.000): a software environment for the analysis of population genetics and Biometry Lab. University of Geneva. Geneva, Switzerland.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellito D, Arredi B, Destro-Bisol G, De Stefano G, Rickards O, Martinez-Labarga C, Modiano D, Biondi G, Moral P, Olckers A, Wallace DC, Novelletto A (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* 65:829-846

Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T (1996) mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* 58:595-608

Thomas H (1997). *The slave trade-the history of the Atlantic slave trade:1440-1870*. Macmillan, London.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa". *Am J Hum Genet* 66:674-686

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* 69:1348-1356

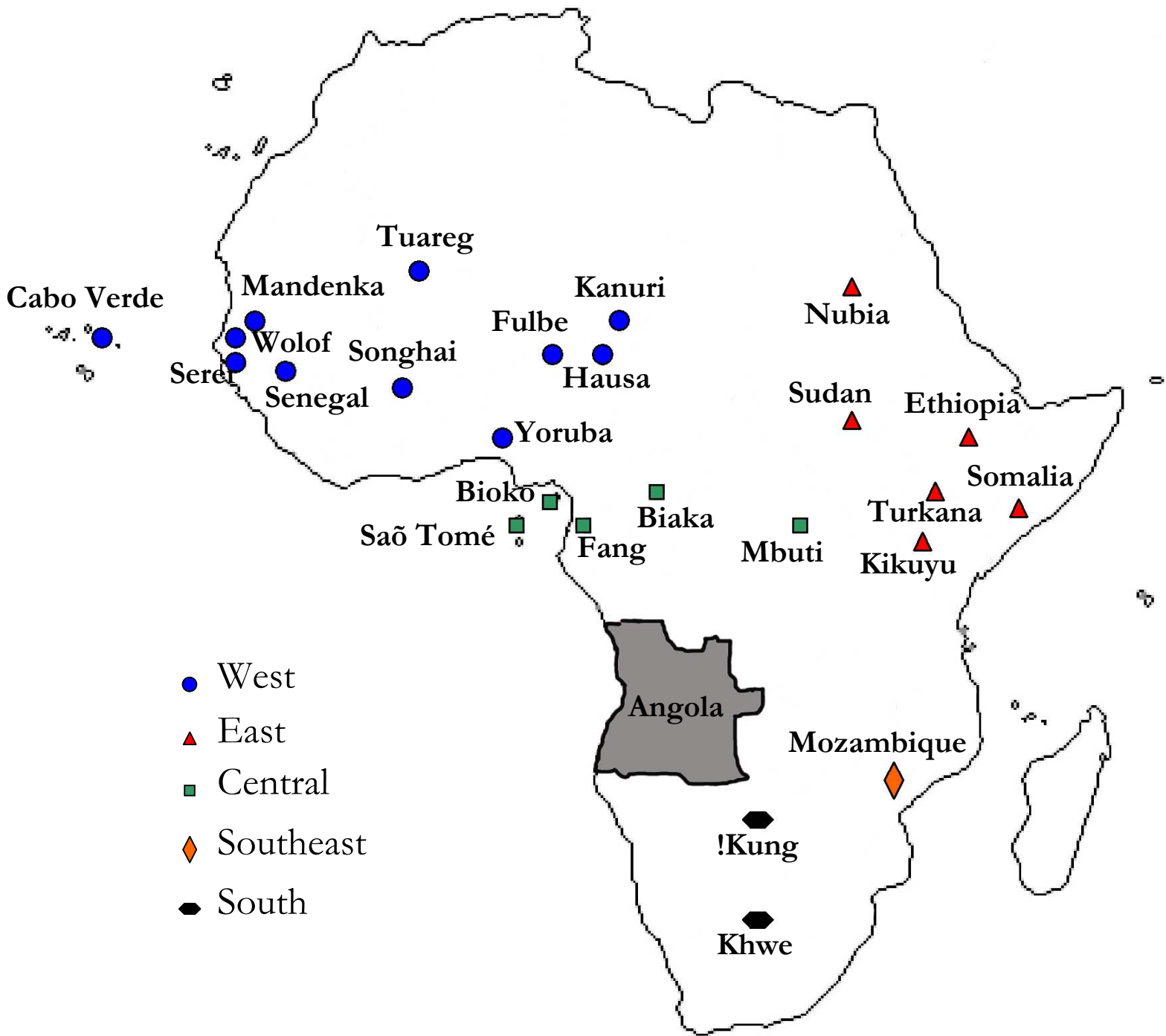
Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26:358-361

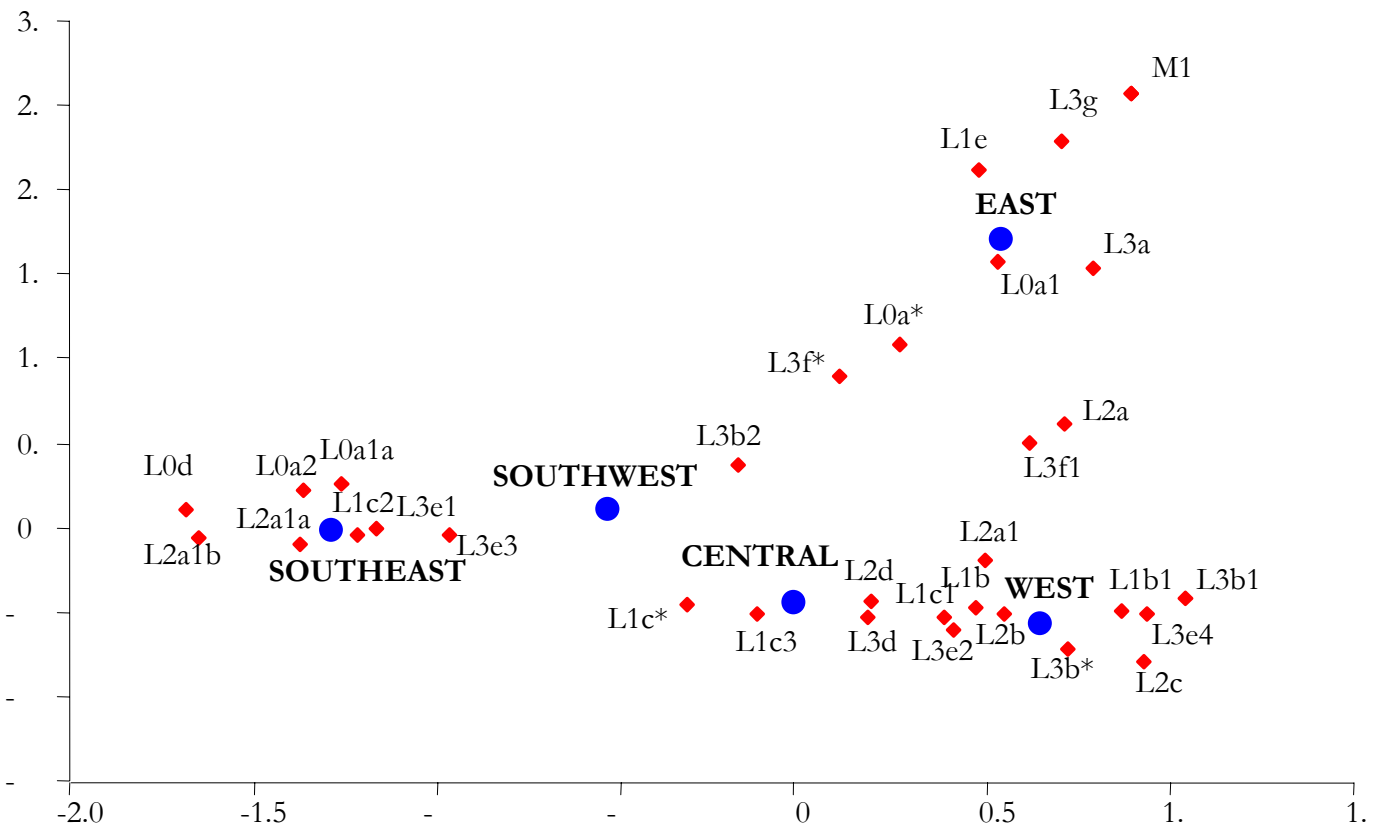
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65:43-62
- Vansina J (1995) New Linguistic evidence and the “Bantu expansion” *J Afr Hist* 36:173-195
- Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci* 86:9350-9354
- Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507
- 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
- Watson E, Forster P, Richards M, Bandelt HJ (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* 61:691-704
- Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 12:339-348

Legend to figures

Figure 1: Location of the samples used for reference and their regional ascription.

Figure 2: Two-dimension plot of the analysis of the correspondence based on the absolute L and M1 haplogroup frequencies of the West, East, Central, Southwest and Southeast African populations.





~ *CAPÍTOL V* ~

*Admixture, migrations, and dispersals in Central Asia:
evidence from maternal DNA lineages*

Comas D, Plaza S*, Wells R S, Yuldasheva N, Lao O, Calafell F, Bertranpetit J¹

*(en premsa a European Journal of Human
Genetics)*

ARTICLE

Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages

David Comas^{*,1,4}, Stéphanie Plaza^{1,4}, R. Spencer Wells², Nadira Yuldaseva^{2,3}, Oscar Lao¹, Francesc Calafell¹ and Jaume Bertranpetit¹

¹Unitat de Biologia Evolutiva, Departament de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona 08003, Spain; ²Wellcome Trust Center for Human Genetics, University of Oxford, Headington, UK; ³Institute of Immunology, Academy of Sciences, Tashkent, Uzbekistan

Mitochondrial DNA (mtDNA) lineages of 232 individuals from 12 Central Asian populations were sequenced for both control region hypervariable segments, and additional informative sites in the coding region were also determined. Most of the mtDNA lineages belong to branches of the haplogroups with an eastern Eurasian (A, B, C, D, F, G, Y, and M haplogroups) or a western Eurasian (HV, JT, UK, I, W, and N haplogroups) origin, with a small fraction of Indian M lineages. This suggests that the extant genetic variation found in Central Asia is the result of admixture of already differentiated populations from eastern and western Eurasia. Nonetheless, two groups of lineages, D4c and G2a, seem to have expanded from Central Asia and might have their Y-chromosome counterpart in lineages belonging to haplotype P(xR1a). The present results suggest that the mtDNA found out of Africa might be the result of a maturation phase, presumably in the Middle East or eastern Africa, that led to haplogroups M and N, and subsequently expanded into Eurasia, yielding a geographically structured group of external branches of these two haplogroups in western and eastern Eurasia, Central Asia being a contact zone between two differentiated groups of peoples.

European Journal of Human Genetics advance online publication, 11 February 2004; doi:10.1038/sj.ejhg.5201160

Keywords: mitochondrial DNA; genetic admixture; haplogroup; Central Asia

Introduction

Central Asia is a vast territory that has been crucial in human history due to its strategic location. Situated eastwards of the Caspian Sea, limited by the Hindu Kush and Altai mountain ranges to the east and by the great Asian Steppes to the north, this territory has been a complex assembly of peoples, cultures, and habitats.

The area has been occupied since Lower Paleolithic times, and there is evidence of Neanderthal skeletal

material in Teshik-Tash,¹ Uzbekistan. Nonetheless, the later expansion of Upper Paleolithic remains is far less clear.² Classical Greek and Chinese historic records cite the Scythians and Sarmatians, Indo-European-speaking people described as having European morphological traits, as the first inhabitants occupying the region. These historic citations raise the questions of the origin of the ancestors of the modern settlers across the region, and of the limits of western peoples in Asia. Several facts point to the presence of western peoples far east in Asia, such as an extinct Indo-European language (Tocharian) spoken during the latter half of the first millennium in Chinese Turkestan, the presence of mummified bodies with European facial traits in the Xinjiang region, the description of west Eurasian mitochondrial DNA lineages in Central Asia,³ and the suggested European affiliation of mitochondrial

*Correspondence: Dr D Comas, Universitat Pompeu Fabra Unitat de Biologia Evolutiva, Doctor Aiguader 80 Barcelona, Catalonia 08003, Spain. Tel: +34 93 5422844; Fax: +34 93 5422802; E-mail: david.comas@upf.edu

⁴These authors contributed equally to the present study.
Received 28 July 2003; revised 12 December 2003; accepted 17 December 2003

DNA sequences from ancient bones in an Eastern Chinese site.⁴ Besides Scythians and Sarmatians, other peoples left their influence in the area: Greeks, Chinese, Turkic tribes such as the Huns, and the Avars, Arabs, and others.

Physical anthropology has roughly defined Central Asian populations as presenting an admixture of eastern and western anthropometric traits.⁵ There are few genetic data about the human populations settled in the region. Classical genetic data⁶ have demonstrated an intermediate position of Central Asians between the Middle East and East Asia. As a general rule, the people inhabiting the area are the result of admixture between differentiated populations, which has produced a high genetic diversity.^{3,7-9} Nonetheless, recent data of Y-chromosome lineages in Central Asia¹⁰ have shown that genetic diversity is heterogeneous in the region, with some high-diversity populations contrasting with much reduced levels in others. This pattern has been interpreted as the occurrence of several bottlenecks or founder events in the area.

Mitochondrial DNA (mtDNA) lineages have been used to unravel past demographic scenarios due to their particular properties. Previous mtDNA analyses in Central Asia based on the sequence of the first hypervariable segment of the control region³ have shown that the mtDNA pool of three populations in Central Asia (the Kazakh, the Kirghiz, and the Uighur) is the result of admixture from east and west Eurasia. Although mtDNA control region sequences allowed the general distinction between the Eastern and Western sources, it did not allow full resolution into haplogroups and of the phylogeographic perspective. The knowledge provided by complete mtDNA sequences¹¹⁻¹⁵ and the refined definition of haplogroups both in West Eurasia¹⁶ and in East Asia^{15,17} provides a fine-grained phylogeography of the mtDNA lineage distribution, which might allow us to determine which mtDNA markers should

be determined to analyze the diversity of the present Central Asian samples.

The analysis of extant central Asians allows us to test several scenarios concerning the spread of western peoples in Asia and their interaction with eastern peoples. In this sense, we have analyzed 12 populations from all the major linguistic groups in the area, and have typed both hypervariable segments of the control region as well as some key SNPs in order to achieve a much finer phylogeographic resolution. This will allow a more complete description of the mtDNA diversity in Central Asia, and its interpretation in relation to human origins and dispersals into and out of Central Asia.

Material and methods

A total of 232 individuals from 12 different population groups were analysed: 20 Bukharan Arabs, 20 Crimean Tatars, 20 Iranians, 16 Dungans, 20 Karakalpaks, 20 Kazaks, 20 Khoremian Uzbeks, 20 Kyrgyz, 20 Tajiks, 20 Turkmen, 16 Uighurs, and 20 Uzbeks. Samples were collected in Uzbekistan and Kyrgyzstan, with informed consent; information about the origin of maternal ancestors was recorded in order to localize samples geographically, and their locations are shown in Figure 1.

DNA was extracted from blood samples using standard methods. Both mtDNA hypervariable regions (HVRI and HVRII) were amplified using primers L15996 and H408,¹⁸ and the amplification products were subsequently purified with the GenClean (BIO101) kit. The sequence reaction was performed for each strand, using primers L15996 and H16401 for the HVRI, and L29 and H408 for the HVRII,¹⁸ with the ABI PRISM dRhodamine Terminator Cycle Sequencing kit (Applied Biosystems) according to the supplier's recommendations. Sequences from positions

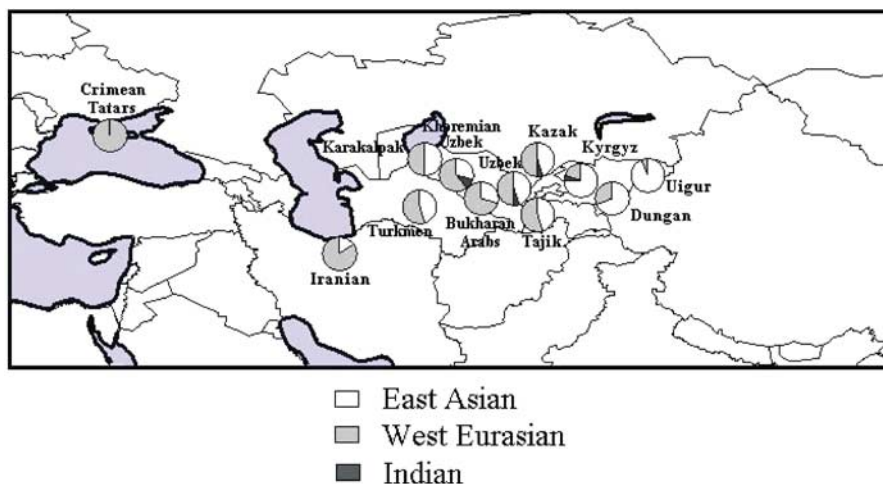


Figure 1 Geographic location of samples analyzed in the present study. Frequencies of East Asian, West Eurasian, and Indian lineages are shown in white, pale gray, and dark gray, respectively.

16024–16391 and 63–322, respectively,^{19,20} were obtained.

The 9-bp tandem repeat (CCCCCTCTA) of the COII/tRNA^{Lys} intergenic region was amplified using primers L8196 (5'-ACAGTTTCATGCCCATGGTC-3', labeled at 5' with JOE) and H8297 (5'-ATGCTAAGTTAGCCTTACAG-3'). The cycling conditions were as follows: 94°C for 2 min; followed by 30 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 1 min; and a final elongation step of 72°C for 5 min. The product was run in an ABI PRISM377 and GeneScan analysis software was used to measure the fragment sizes.

Three positions in the mtDNA coding region^{19,20} (10 400, 12 308, and 12 705) were also determined by using the SNaPshot™ ddNTP Primer Extension Kit (Applied Biosystems), which consists of a single-base primer extension which uses labeled ddNTPs to interrogate SNPs. The mtDNA region containing the three SNPs was amplified using primers L10373 (5'-CCCTAAGTCTGGCCTATGAG-3') and H12744 (5'-CGATGAACAGTTGGAATAGG-3'), with the following cycling conditions: 94°C for 5 min; 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and a final elongation step of 72°C for 5 min. The 2410-bp amplification products were purified using the QIAquick™ PCR Purification Kit (QIAGEN). The single-base primer extension was performed following supplier's recommendations using oligonucleotides H10400X (5'-TGTTAAACTATA-TACCAATTC-3'), L12308X (5'-CAGCTATCCATTGGTCT-TAGGCCCAA-3'), and L12705X (5'-AACATTAATCAGTT-CTTCAAATATCTACTCAT-3') in the same reaction. Unincorporated-labeled ddNTPs were removed by adding 1 U of CIP to the primer extension products for 1 h at 37°C, followed by an incubation of 15 min at 72°C to inactivate the enzyme. Products were run in an ABI PRISM377 and

GeneScan analysis software was used to measure fragment sizes.

Each mtDNA molecule was assigned to one haplogroup according to the following strategy. First, the combination of the three SNPs in the coding region was taken into account in order to classify the mtDNA molecules in one of the four major groups determined in the present work: R, U, M, or other (namely, L or N). Subsequently, the information yielded by the control region sequence was added in order to refine the classification into haplogroups^{15–17} (see Figure 2). Nonetheless, after this assignment strategy, some individuals were difficult to be classified as N or L3. For this reason, variation at position 10 873, distinguishing haplogroup N from L3, was also tested using the single-base primer extension approach with oligonucleotide L10873X (5'-TTTTTTTTTCCA-CAGCCTAATTATTAGCATCATCCC-3').

In order to compare the present results with other populations, HVRI data from several European, Middle Eastern, Indian, Central Asian, and East Asian populations were taken from the literature: Kazaks,³ Kyrgyz,³ Uighurs,³ Altaics,²¹ Mongolians,^{22,23} Daur,²³ Oroqen,²³ Turks,^{24–26} Han Chinese,^{17,27} Han Taiwanese,²⁸ Ainu,²⁸ Koreans,^{23,28,29} Japanese,^{28,30} Europeans,³¹ Middle Easterns,³¹ Caucasus populations,^{32,33} Thai,³⁴ Indians,³⁵ Russians,^{36,37} Ukrainians,³⁷ and Siberians.^{23,38,39}

The networks relating HVRI sequences within some of the haplogroups described were constructed by using a reduced-median algorithm⁴⁰ as implemented in the Network 3.0 program. The dating method employed⁴¹ is based on the average number of mutations accumulated from an ancestral sequence as a linear function of time and mutation rate. This method was also performed with the Network 3.0 program.

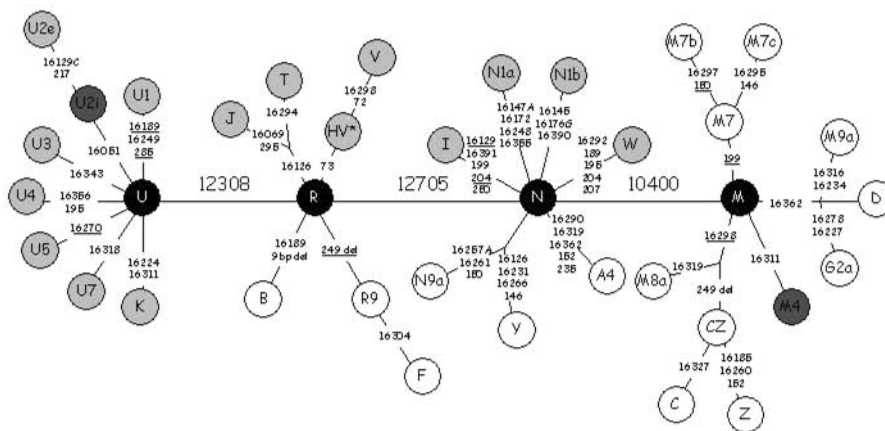


Figure 2 Phylogenetic reconstruction and geographic distribution of the haplogroups found in Central Asia. Numbers along the links indicate substitutions (transversions are indicated by the substituted nucleotide after the number), underlined numbers indicate recurrent events. East Asian, West Eurasian, and Indian lineages are shown in white, pale gray, and dark gray, respectively.

Program Admix 2.0⁴² was used to calculate the admixture proportions of the present samples based on the frequency of the haplogroups. As putative parental populations, we used four data sets that consisted of 258 Eastern Europeans³¹ (Bulgarians, Romanians, and Russians), 316 Middle Easterns³¹ (Bedouins, Syrians, and Turks), 190

Northern Indians, and Pakistanis³⁵ (regions of Uttar Pradesh, Rajasthan, Punjab, Kashmir, Haryana, and Pakistan), and 263 East Asians²⁷ (Han Chinese).

In order to detect the possible genetic structure among populations, an analysis of the molecular variance (AMOVA)⁴³ was performed using the Arlequin package.⁴⁴

Table 1 Haplogroup frequencies in the samples analyzed.

	CT	IR	TU	KR	KU	AR	UZ	TD	KZ	KG	DU	UI	Total
<i>R</i>													
HV*	4	3	6	5	5	7	3	3	4		1	1	42
V						1				1			2
J	2	2	1	1	1	2	1	1	1				12
T	6	1	1	1	1	1	1	1	1		3		17
B			1		2	1	1	1			1		7
F			1	1			1			3	1	4	11
R9				1			1						2
R*		2	1	1			3		1		1		9
<i>U</i>													
K	2	2			1		1	1					7
U1	1	1		2		2		1	1				8
U2e						1				1			2
U2i					2		1			1			4
U3	2												2
U4	1				2		1						4
U5		1			1			1	2				5
U7								1					1
<i>M</i>													
CZ				1									1
C		2	1			4	2		1	6	1		17
Z				1					1				2
D (D4c)		1	6 (5)	4	2		1	3 (2)	2 (1)	4	1	7(2)	31(10)
G2a				1			2	1	4				8
M4									1				1
M7b										1	2	1	4
M7c				1								1	2
M8							1			1		1	3
M9											1		1
<i>Other</i>													
I	1	1			1					1			4
W								2		1			3
N1a			1										1
N1b	1	3											4
N9a											2		2
A4					2			3			2	1	8
Y						1		1	1				3
N*		1	1										2
East	—	3	9	10	6	6	9	9	9	15	11	15	102
M	—	0.15	0.45	0.50	0.28	0.30	0.44	0.45	0.44	0.74	0.69	0.94	0.48 ^a
West	20	17	11	10	12	14	10	11	10	4	5	1	125
M	1	0.85	0.55	0.50	0.54	0.70	0.45	0.55	0.45	0.17	0.31	0.06	0.48 ^a
India	—	—	—	—	2	—	1	—	1	1	—	—	5
M	—	—	—	—	0.17	—	0.09	—	0.09	0.09	—	—	0.04 ^a
Total	20	20	20	20	20	20	20	20	20	20	16	16	232

In D haplogroup, in parenthesis, individuals belonging to D4c haplogroup. CT: Crimean Tatars, IR: Iranian, TU: Turkmen, KR: Karakalpak, KU: Khorezmi Uzbek, AR: Bukharan Arabs, UZ: Uzbek, TD: Tajik, KZ: Kazak, KG: Kyrgyz, DU: Dungan, and UI: Uighur, m: estimated contribution of each region (East Asia, West Eurasia, and India) to each Central Asian population, taking into account that the Indian mtDNA pool contains also East Asian and West Eurasian sequences, a Total contribution without taking into account Crimean Tatars.

Results

Phylogeographic structure

A total of 232 individuals have been analyzed for the HVRI and HVRII, for the presence of the 9-bp tandem repeat of the COII/tRNA^{Lys} intergenic region, and several SNPs in the mtDNA coding region. Individual data are available in the following web site (<http://www.upf.es/cexs/bioevo/index.html>).

Haplogroup frequencies by population are shown in Table 1. In all, 11 sequences were difficult to assign to a specific haplogroup and were named after the first major classification yielded by the coding SNPs (all belong either to R* or N*). The haplogroups found and the positions that define them are shown in Figure 2.

Within the present samples, no African lineages were found. No sub-Saharan L (L1, L2, and L3) lineages^{45,46} were present in Central Asian samples. Other haplogroups of African origin, such as U6 from North Africa,⁴⁷ or M1 from East Africa,⁴⁸ are not found in the present sample set.

Within major group R, mtDNA molecules analyzed belong either to West Eurasian haplogroups (H, V, J, and T) or to East Asian haplogroups (B, R9, and F). Within this group of lineages, the West Eurasian haplogroup HV* (including pre-HV, HV, and H) is the most numerous, and it is present in all the analyzed populations except the Kyrgyz. Two individuals belong to haplogroup V, which is likely to be of Western European origin.⁴⁹ Nevertheless, the range of haplogroup V extends far beyond Europe, into Northern Africa⁵⁰ and as far East as Central Asia.

MtDNA molecules belonging to major group U have their origin in West Eurasia and they have been found in most Central Asian populations. Nevertheless, Kivisild *et al*³⁵ distinguished two groups of lineages within haplogroup U2: West European U2e and Indian U2i. Within the present sample set, we have found both the U2 groups.

In continental Asia, lineages belonging to major group M have an Indian (M2, M3, M4, M5, and M6)⁵¹ or an East Asian (C, D, E, Z, M7, M8, M9, M10, and M11)^{15,17} origin. Only one M Indian lineage (belonging to the M4 haplogroup) has been found in the sample set, whereas the rest of M lineages have an East Asian origin. Haplogroup D is the most frequent haplogroup within this major group, followed by C lineages. Some mtDNA molecules belonging to E and G root lineages might have been classified as D since they are not distinguishable by control region sequence substitutions; this is not a major bias as all of them are of East Asian distribution.

Within D, a non-negligible fraction of sequences carry a transition at position 16245. This group may be a clear subclade within D, which, pending further coding-region characterization, we suggest to call D4c. D4c is highly frequent and diverse in Central Asia (25% in Turkmen, 10% in Tajik, 7% in Uighur, 2.7% in Kazak, and 0.9% in Kyrgyz) (present data and Comas *et al*³), and it is found at low

frequencies, in Turks (2.1%), Daur (8.9%, only two sequences), Mongolians (0.7%), southern Siberians (0.7%), Han Chinese (0.6%), and Koreans (0.5%). This group of sequences is absent in other East Asian, Indian, and Middle Eastern samples. The structure of the variation of these sequences is shown as a network in Figure 3, from which an age of 25 000 (SE 9600) years can be estimated.

All G lineages found in the present samples belong to the G2a group; thus, no G1 or G3 lineages were found. In fact, the presence of G2a lineages seems to be also restricted to Central Asia. This haplogroup characterized by the motif 16 223T, 16 227G, 16 278T, and 16 362C, has been found in Kazaks (9.3%), Kyrgyz (7.0%), Karakalpak (5.0%), Tajik (5.0%), and Uzbek (5.0%) (present data and Comas *et al*³). It has also been found in neighboring populations at lower frequencies, such as Mongolians (1.3%), Mansi from Siberia (6.1%, only one sequence), southern Siberians (2.4%), Ainu (3.9%), Japanese (0.7%), Daur (4.4%, two sequences), Han Taiwanese (3.0%), Korean (1.9%), Han Chinese (2.2%), and the Caucasus (0.6%). The structure of the variation of haplogroup G2a is shown in Figure 4, from which an age of 29 500 (SE 7000) years can be estimated.

Other haplogroups found in Central Asia are A, Y, and N9a, which have an East Asian origin, whereas haplogroups W, I, N1a, and N1b have been described in West Eurasian populations.

Admixture analysis

The presence in Central Asia of a high proportion of sequences originating elsewhere suggests that these populations have experienced intense gene flow. In order to quantify the apportionment of admixture in Central Asian samples, two different approaches were followed: a phylogeographic approach and an admixture approach based on haplotype frequencies. Crimean Tatars were excluded from the admixture analysis since their geographic position corresponds more to Europe rather than Central Asia, and their mtDNA pool is completely of West Eurasian origin.

Taking into account the phylogeography of the haplogroups described for West Eurasia¹⁶ and East Asia,^{15,17} these can be divided into three groups depending on their origins: West Eurasian, East Asian, and Indian (Table 1 and Figure 1). Whereas West Eurasian and East Asian populations contain almost exclusively locally originated mtDNA haplogroups, this is not the case for India. Then, admixture from India would also contribute West and East Eurasian sequences to Central Asia. Thus, estimated admixture proportions have been corrected with the frequencies of haplogroups of Indian (58.4%), West Eurasian (32.6%), and East Asian (8.9%) origins in a sample from India and Pakistan.³⁵ Standard deviations were estimated by sampling with replacement 100 000 times in samples having the same sizes and haplogroup frequencies as those in Central Asia and India, and computing each time the admixture proportion estimates. Considering all the

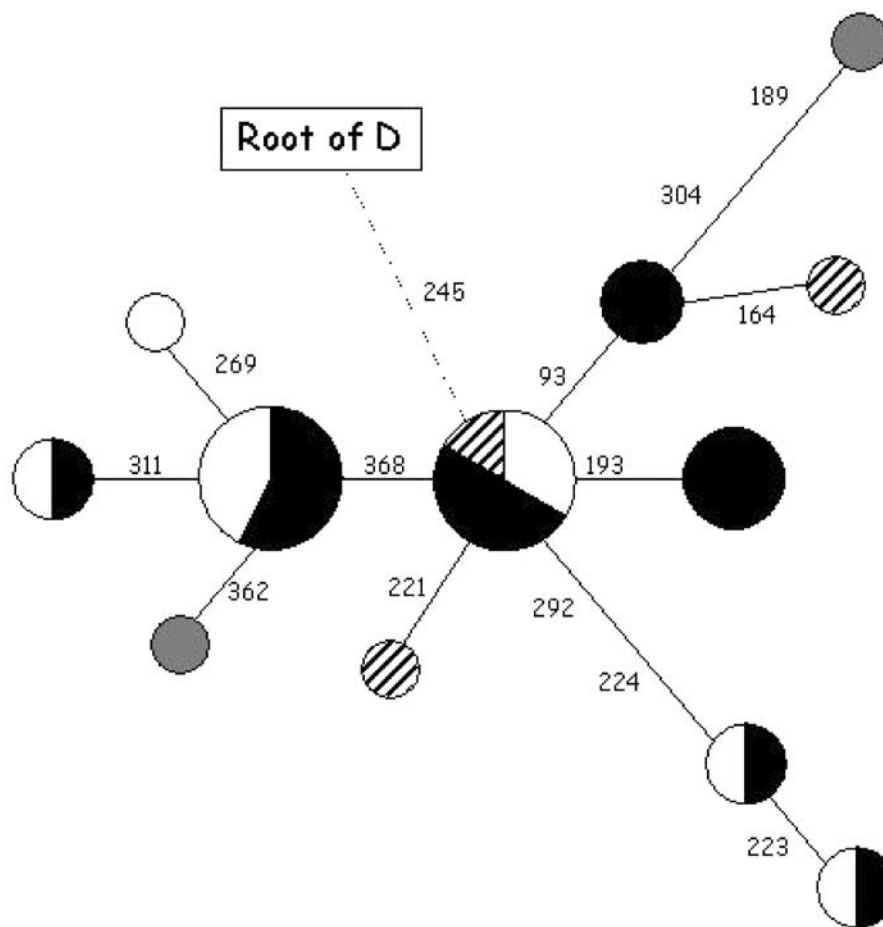


Figure 3 Phylogenetic network of a section of haplogroup D sequences (D4c). The size of the circles is proportional to the number of sequences. Central Asian samples are represented in black, East Asians in white, Turks in gray, and Siberians in striped gray. Mutated sites (minus 16 000) are indicated along the lines.

individuals as belonging to a single hybrid population, the estimated admixture proportions are 0.48 ± 0.04 West Eurasian, 0.48 ± 0.04 East Eurasian, and 0.04 ± 0.02 Indian. Given the sample sizes for individual populations, their admixture proportions (Table 1) carry large standard errors and are not discussed separately.

An admixture approach⁴² was performed using the method implemented in Admix 2.0 program, considering four putative parental populations. The apportionment for the whole sample set was 0.11 ± 0.24 European, 0.40 ± 0.25 Middle Eastern (which adds up to 51% for West Eurasia), 0.45 ± 0.05 East Asian, and 0.04 ± 0.04 Indian. Although this approach allowed us to use a larger number of parental populations, the standard deviation after 10 000 iterations is extremely high for some of the estimates.

Genetic structure of Central Asia

The genetic structure of Central Asian populations was investigated through AMOVA. When the 12 samples were considered as a single group, only 2.34% ($P < 0.0001$) of the

genetic variance was attributed to differences among populations. When samples were grouped according to language families (Afro-Asiatic, Altaic, Indo-European, and Sino-Tibetan), the fraction of the genetic variance found among groups was not significantly different from 0 ($P = 0.817$), whereas differences found among populations within language groups were statistically significant (2.9%, $P < 0.0005$), showing that the genetic variation found in the mtDNA was not structured according to language affiliation.

Discussion

The mtDNA genetic landscape of Central Asia contains four main differentiated lineage groups according to their phylogeographic origin: (i) a group of lineages originating in West Eurasian and comprising almost half of the mtDNA sequences in Central Asia; (ii) East Asia lineages, making almost the other half of lineages, (iii) two putatively locally expanded haplogroups, of East Asian origin, D4c and G2a,

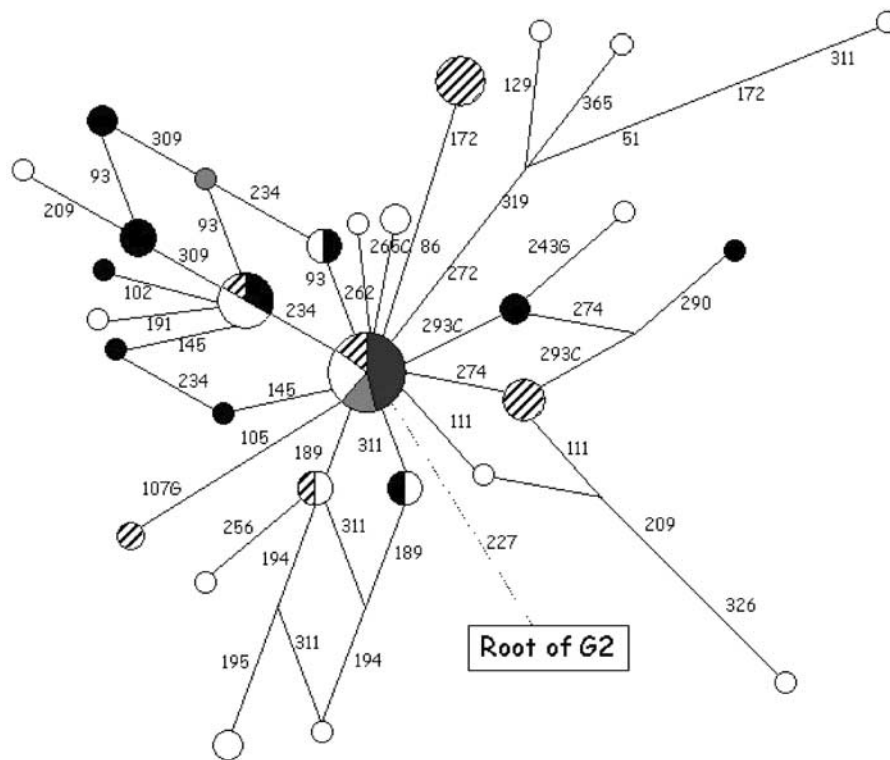


Figure 4 Phylogenetic network of haplogroup G2a. The size of the circles is proportional to the number of sequences. Central Asian samples and Mongolians are represented in back, East Asians in white, samples from the Caucasus in gray, and Siberians in stripped gray. Mutated sites (minus 16 000) are indicated along the lines.

accounting for a ~8% of the total sequences, and (iv) a tiny fraction of sequences of Indian origin.

We have detected some groups of sequences mainly restricted to this geographical area. This is the case of haplogroups G2a and D4c. The fact that these groups of lineages are localized in Central Asia at higher frequencies than in neighboring populations could be explained as a result of genetic drift during founder events that could have raised its frequency in this geographical area. Nevertheless, the high diversity found in Central Asia within both groups of sequences (Figures 3 and 4) supports an ancient origin of the founder mutations (around 30 000 and 25 000 years), an expansion of these lineages in Central Asia, and subsequent dispersal to neighboring populations. These ancient events represent ancient expansions originated in Central Asia and might have their Y-chromosome counterpart in lineages belonging to haplotype P(xR1a) that has a high frequency in Central Asia and is dated to $\approx 40\,000$ years.⁸ There is, thus, a fraction of the gene pool that can be considered Central Asian specific, which could reflect the remnants of the oldest peopling by modern humans.

Besides the specific cases of G2a and D4c lineages, no other lineages seem to have expanded in Central Asia, and the majority of lineages found have an Eastern or Western

origin, which are two mtDNA pools that do not overlap. This fact implies that both genetic pools were already differentiated when they met in Central Asia. Thus, the geographic distribution of mtDNA lineages in Europe and Asia is not compatible with a Central Asian origin of both mtDNA pools, in agreement with previous data.³

The presence of western sequences in Central Asia prompts the question of the eastern spread of western influence in Asia. The analyses performed of the ancient sites of Liangchun⁴ (2500 years old) and Yixi⁵² (2000 years old), eastern China, concluded that there was a drastic shift from a European-like population 2500 years ago, through an intermediate population 2000 years ago, to the present-day East Asian populations. Liangchun sequences are difficult to assign to haplogroups due to the short mtDNA sequence analyzed, and their ascription to the Western Eurasia gene pool has been challenged⁵³ up to the point that the latter authors do not interpret any Liangchun sequence as Western. On the other hand, most Yixi sequences belong to extant East Asian haplogroups such as D, C, or F, which suggests that the genetic composition of the 2000-year-old Yixi site presented no genetic traces of western influence. The genetic influence of western peoples across Asia is obvious in Central Asia, but there is no evidence of its presence in the easternmost regions

since no traces are found in extant or ancient East Asian populations. Even if Tocharian, an Indo-European language, was present in Eastern Asia, there is no evidence, from extant genetic variation in maternal lineages, of the Western Eurasia genetic contribution.

The presence of western and eastern sequences found in Central Asia leaves open questions about the mode and tempo of the generation of this admixture of lineages. Two scenarios could have produced this mtDNA pattern in Central Asia:

- (a) Western peoples inhabited Central Asia and were partially replaced by Eastern peoples, Central Asia being a hybrid zone.
- (b) Central Asia has been a 'contact zone' between two differentiated groups of peoples who originated in east and west Eurasia, respectively.

The revision of the ancient sequences from China⁵³ and the finding of specific Central Asian sequences clearly support the second. G2a and D4c haplogroups are 'twigs' (according to the terms devised by Kivisild *et al*¹⁷) belonging to the East Asian G and D 'limbs' of the M 'trunk'. The estimated ages of these haplogroups (around 30 000 and 25 000 years) point to the ancient presence of at least two different East Asian 'limbs' in Central Asia.

Kivisild *et al*¹⁷ showed considerable differences in the mtDNA lineages found in East Asia, A, C, D, G, Y, and Z being the haplogroups forming the pool of lineages in the northeast, whereas B and F were predominant in the southeast. Karafet *et al*,⁹ analyzing Y-chromosome markers, showed a closer genetic relationship between Central Asia and northeast Asia than with southeast Asia. Nevertheless, our mtDNA results show the presence of haplogroups represented in both northeast and southeast Asia, suggesting that the demographic scenario within Central Asia has been even more complex than previously stated.⁹

Contrary to the structure shown in Y-chromosome lineages in Central Asia, where 24% of the genetic variation could be attributed to differences between populations,¹⁰ mtDNA diversity is not structured, as shown by the AMOVA analysis. This discrepancy between the two uniparental genomic regions in Central Asia is in agreement with previous data in the region,⁷ and as a global trend in which higher female than male migration has been observed.⁵⁴

It is interesting to stress the lack of geographic structure of the basal branches of the non-African mtDNA (haplogroups M and N, called 'limbs'¹⁷), and a clear phylogeography in more external branches (haplogroups or sub-haplogroups; 'twigs'¹⁷) supports the existence of an ancestral population where the two main groups of lineages diverged. This could be related to the presence of a 'maturation phase', presumably in the Middle East or eastern Africa, of modern humans before the Upper Paleolithic expansion all across Eurasia, as proposed by

the fossil evidence⁵⁵ and other genetic data.⁵⁶ The lack of basal limbs in Central Asian samples and the presence of lineages belonging to external branches within the mtDNA phylogeny suggest that the mtDNA diversity found in Africa did not have its 'maturation phase' in Central Asia, and the diversity found in the region is mainly the result of admixture of already differentiated populations. The lack of mtDNA basal root types in Central Asia contrasts with the results of Y-chromosome analyses. Whereas the majority of extant Y lineages in Europe and Siberia appear to have expanded from the Middle East via Central Asia,⁸ the lack of deeply rooting mtDNA clades in Central Asia does not support the hypothesis that Central Asia is the maternal source population for the Upper Paleolithic colonization of Europe. This discrepancy might be the result of different sexual migration patterns in Central Asia, as noted above. Additional data from autosomal markers, such as SNP or SNPSTR haplotypes,⁵⁷ need to be gathered in order to clarify the genetic role of Central Asia in the settlement of modern humans in Europe and Siberia.

Acknowledgements

Some individual European and West Asian sequences (published as haplogroups) were kindly provided by Martin Richards, Huddersfield University, UK. We thank Miguel A Padilla and Mònica Vallés, Universitat Pompeu Fabra, for technical assistance. The present study was supported by the Dirección General de Investigación Científica y Técnica, Spain (BOS2001-0794), and Direcció General de Recerca, Generalitat de Catalunya (2001SGR00285). SP received a fellowship from the Direcció General de Recerca, Generalitat de Catalunya (2000FI00696).

References

- 1 Okladnikov AP: Nakhodka Neandertal'tsa v Uzbekistane. *Vestnik Drevnei Istorii* 1939; 1: 256–257.
- 2 Derevyanko AP, Zin'E L: Upper Palaeolithic cultures. in Danni AH, Masson VM (eds) *History of the Civilizations of Central Asia*. Paris: UNESCO; 1992, Vol I.
- 3 Comas D, Calafell F, Mateu E *et al*: Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am J Hum Genet* 1998; 63: 1824–1838.
- 4 Wang L, Oota H, Saitou N, Jin F, Matsushita T, Ueda S: Genetic structure of a 2500-year-old human population in China and its spatiotemporal changes. *Mol Biol Evol* 2000; 17: 1396–1400.
- 5 Bowles GT: *The People of Asia*. Birkenhead, Great Britain: Willmer Brothers Limited; 1977.
- 6 Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton: Princeton University Press; 1994.
- 7 Pérez-Lezaun A, Calafell F, Comas D *et al*: Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* 1999; 65: 208–219.
- 8 Wells RS, Yuldasheva N, Ruzibakiev R *et al*: The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA* 2001; 98: 10244–10249.
- 9 Karafet T, Xu L, Du R *et al*: Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* 2001; 69: 615–628.
- 10 Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C: A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet* 2002; 71: 466–482.

- 11 Ingman M, Kaessmann H, Paabo S, Gyllensten U: Mitochondrial genome variation and the origin of modern humans. *Nature* 2000; **408**: 708–713.
- 12 Finnila S, Lehtonen MS, Majamaa K: Phylogenetic network for European mtDNA. *Am J Hum Genet* 2001; **68**: 1475–1484.
- 13 Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM: Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* 2001; **2**: 13–20.
- 14 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.
- 15 Kong Q-P, Yao Y-G, Sun C, Bandelt H-J, Zhu C-L, Zhang Y-P: Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet* 2003; **73**: 671–676.
- 16 Macaulay V, Richards M, Hickey E *et al*: The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* 1999; **64**: 232–249.
- 17 Kivisild T, Tolk HV, Parik J *et al*: The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 2002; **19**: 1737–1751.
- 18 Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC: Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci USA* 1989; **86**: 9350–9354.
- 19 Anderson S, Bankier AT, Barrell BG *et al*: Sequence and organization of the human mitochondrial genome. *Nature* 1981; **290**: 457–465.
- 20 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.
- 21 Shields GF, Schmiechen AM, Frazier BL *et al*: MtDNA sequences suggest a recent evolutionary divergence for Beringian and Northern North American populations. *Am J Hum Genet* 1993; **53**: 549–562.
- 22 Kolman CJ, Sambuughin N, Bermingham E: Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics* 1996; **142**: 1321–1334.
- 23 Kong QP, Yao YG, Liu M *et al*: Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Hum Genet* 2003; **113**: 391–405.
- 24 Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L: From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet* 1996; **60**: 35–49.
- 25 Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bertranpetit J: Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations. *Mol Biol Evol* 1996; **13**: 1067–1077.
- 26 Richards M, C rte-Real H, Forster P *et al*: Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 1996; **59**: 185–203.
- 27 Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP: Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 2002; **70**: 635–651.
- 28 Horai S, Murayama K, Hayasaka K *et al*: MtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *Am J Hum Genet* 1996; **59**: 579–590.
- 29 Lee SD, Shin CH, Kim KB, Lee YS, Lee JB: Sequence variation of mitochondrial DNA control region in Koreans. *For Sci Int* 1997; **87**: 99–116.
- 30 Koyama H, Iwasa M, Maeno Y *et al*: Mitochondrial sequence haplotype in the Japanese population. *For Sci Int* 2002; **125**: 93–96.
- 31 Richards M, Macaulay V, Hickey E *et al*: Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 2000; **67**: 1251–1276.
- 32 Comas D, Calafell F, Bendukidze N, Fa nan s L, Bertranpetit J: Georgian and Kurd mtDNA sequence analysis shows a lack of correlation between languages and female genetic lineages. *Am J Phys Anthropol* 2000; **112**: 5–16.
- 33 Nasidze I, Stoneking M: Mitochondrial DNA variation and language replacements in the Caucasus. *Proc R Soc Lond B Biol Sci* 2001; **268**: 1197–1206.
- 34 Fucharoen G, Fucharoen S, Horai S: Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet* 2001; **46**: 115–125.
- 35 Kivisild T, Bamshad MJ, Kaldma K *et al*: Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol* 1999; **9**: 1331–1334.
- 36 Orekhov V, Poltoraus A, Zhivotovsky LA, Spitsyn V, Ivanov P, Yankovsky N: Mitochondrial DNA sequence diversity in Russians. *FEBS Lett* 1999; **445**: 197–201.
- 37 Malyarchuk BA, Derenko MV: Mitochondrial DNA variability in Russians and Ukrainians: implication to the origin of the Eastern Slavs. *Ann Hum Genet* 2001; **65**: 63–78.
- 38 Derenko MA, Gzybowski T, Malyarchuk A *et al*: Diversity of mitochondrial lineages in South Siberia. *Ann Hum Genet* 2003; **67**: 391–411.
- 39 Derbeneva OA, Starikovskaya EB, Wallace DC, Sukernik RI: Traces of early Eurasians in the Mansi of northwest Siberia revealed by mitochondrial DNA analysis. *Am J Hum Genet* 2002; **70**: 1009–1014.
- 40 Bandelt HJ, Forster P, Sykes BC, Richards MB: Mitochondrial portraits of human populations using median networks. *Genetics* 1995; **141**: 743–753.
- 41 Saillard J, Magalhaes PJ, Schwartz M, Rosenberg T, Norby S: Mitochondrial DNA variant 11719G is a marker for the mtDNA haplogroup cluster HV. *Hum Biol* 2000; **72**: 1065–1068.
- 42 Dupanloup I, Bertorelle G: Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol* 2001; **18**: 672–675.
- 43 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.
- 44 Schneider S, Kueffer JM, Roessli D, Excoffier L: *Arlequin (ver.1.0): A Software Environment for the Analysis of Population Genetics Data*. University of Geneva, Switzerland: Genetics and Biometry Lab; 1996.
- 45 Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC: Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet* 1995; **57**: 133–149.
- 46 Watson E, Forster P, Richards M, Bandelt HJ: Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* 1997; **61**: 691–704.
- 47 Rando JC, Pinto F, Gonzalez AM *et al*: Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann Hum Genet* 1998; **62**: 531–550.
- 48 Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS: Genetic evidence of an early exit of Homo sapiens from Africa through eastern Africa. *Nat Genet* 1999; **23**: 437–441.
- 49 Torroni A, Bandelt HJ, D’Urbano L *et al*: MtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 1998; **62**: 1137–1152.
- 50 Plaza S, Calafell F, Lefranc F, Helal A, Bertranpetit J, Comas D: Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. *Ann Hum Genet* 2003; **67**: 312–328.
- 51 Bamshad M, Kivisild T, Watkins WS *et al*: Genetic evidence on the origins of Indian caste populations. *Genome Res* 2001; **11**: 994–1004.
- 52 Oota H, Saitou N, Matsushita T, Ueda S: Molecular genetic analysis of remains of a 2000-year-old human population in China – and its relevance for the origin of the modern Japanese population. *Am J Hum Genet* 1999; **64**: 250–258.

- 53 Yao YG, Kong QP, Man XY, Bandelt HJ, Zhang YP: Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. *Mol Biol Evol* 2003; **20**: 214–219.
- 54 Seielstad MT, Minch E, Cavalli-Sforza LL: Genetic evidence for a higher female migration rate in humans. *Nat Genet* 1998; **20**: 278–280.
- 55 Stringer CB: The origin of early modern humans: a comparison of the European and non-European evidence. in Mellars P, Stringer CB (eds) *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans*. Princeton: Princeton University Press; 1989, pp 232–244.
- 56 Forster P, Torroni A, Renfrew C, Röhl A: Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol Biol Evol* 2001; **18**: 1864–1881.
- 57 Mountain JL, Knight A, Jobin M *et al*: SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res* 2002; **12**: 1766–1772.

~ CAPÍTOL VI ~

Admixture and sexual bias in the population settlement of La Réunion Island (Indic Ocean)

Stéphanie Plaza, Elena Bosch, Francesc Calafell, Gérard Lefranc, Maya Césari,

Jaume Bertranpetit, David Comas

(en preparació)

Admixture and sexual bias in the population settlement of La Réunion Island (Indic Ocean)

Stéphanie Plaza, Elena Bosch, Francesc Calafell, Gérard Lefranc, Maya Césari, Jaume Bertranpetit, David Comas

Introduction

La Réunion is one of the Mascarene Islands (Mauritius, Réunion and Rodriguez) located in the Indian Ocean, 690 Km east of Madagascar (Figure 1). The island is one of the overseas French departments with over 700.000 inhabitants (INSEE, Institut National de la Statistique et des Études Économiques, 1999). The island was uninhabited until settled by the French in 1642, although it was previously known to the Arabs and was first visited by the Portuguese in the early 16th century. There were in the island, in 1678, 46 men (44 French and 2 Portuguese) and 37 women (8 French, 15 Malgasy, and 14 Indo-Portuguese). The increase of population was the result of slave trade mainly from Madagascar, Mozambique, India, Senegal and Guinea. When slavery was abolished in 1848, a great influx of Indians from the Malabar Coast came to fill the vacancies created by newly freed slaves on plantations. Afterwards, there was a substantial immigration of Indians and Chinese from Canton. The present day inhabitants constitute a mixed population descendant of French settlers and East Africans, Indians, and Indochinese.

There are no population genetic studies performed in the Mascarene Islands and the only reference is the genetic analyses done in Madagascar. La Réunion is the Mascarene island closest to Madagascar, which archaeological and linguistic evidences suggest that its original settlers came from Indonesia around the fourth century AD and Africans arrived later (Vérin, 1986), although the opposite hypothesis has been also postulated (Ferrand, 1908). Nuclear genetic analyses on the beta-globine gene have shown that Malgasy individuals present a major component derived from central and east African populations, in particular, Bantu-speaking Negroids, although an Asian/Oceanic and Caucasoid components are present (Hewitt R, *AmJHumGenet* 58:1303-1308). The mtDNA analysis in Madagascar has shown that a Polynesian component is present mixed with African lineages (Soodyall *Nat Genet.* 1995 Aug;10(4):377-8).

Mitochondrial DNA (mtDNA) and Y chromosome have been used to elucidate past demographic scenarios due to their particular properties. The maternal and paternal contributions unravelled by mtDNA and Y-chromosome markers respectively, have been analysed in a large number of human population studies. The joint analysis of the non-coding region sequence of the molecule (i.e. the control region) and SNPs on the coding region has allowed establishing the phylogeography of mtDNA lineages. Moreover, the recent availability of complete mtDNA sequences in worldwide samples provides us the information needed to locate the origin of most mtDNA lineages. On the other hand, the phylogeny and new nomenclature provided by biallelic markers have allowed locating the geographic origin of Y-chromosome lineages even with a greater detail.

In order to describe and quantify the origins of the genetic components in the mixed population of the Réunion island, we have sequenced both hypervariable segments and

four SNPs located in the coding region of the mtDNA, and we have also typed several biallelic and STR markers in the Y-chromosome. This analysis will allow unravelling migrations, settlements and dispersal routes that formed the extant population of La Réunion.

Material and Methods

DNA from a total of 41 non-related individuals from La Réunion was extracted from blood samples using standard methods. All individuals were analysed for the mtDNA and the 17 males included in the sample were analysed for the Y-chromosome.

The 9-bp tandem repeat (CCCCCTCTA) of the COII/tRNA^{Lys} intergenic region and four positions in the mtDNA-coding region (10400, 10873, 12308 and 12705, according to Anderson et al., 1981) were also genotyped as described elsewhere (Comas et al., submitted). Both mtDNA hypervariable regions (HVRI and HVRII) were amplified using the primers, PCR conditions, purification methods and sequencing profile as described previously (Comas et al., in press). Sequences from positions 16024 – 16391 and 63 – 322 respectively (according to Anderson et al., 1981) were used in the present analysis and are available in the following web site (<http://www.upf.es/cexs>). In order to compare the present results with other populations, data for the first mtDNA hypervariable region (positions 16024-16383) from several population samples were taken from the literature.

Sequence diversity was calculated as $[n/(n-1)] (1 - \sum_{i=1}^k p_i^2)$, where p is the frequency of each of the k different sequences in the sample, with the Arlequin 2.000 program (Schneider et al., 2000).

Y-chromosome biallelic polymorphisms were typed in a hierarchic manner using three different multiplex reactions. All male individuals were analysed for markers M89, M172, M69, M201, M170, M9, 12f2 and M145 (multiplex I). Subsequently, those individuals assigned to clade K (M9 derived) according to the Y Chromosome Consortium (2002) were further characterized for markers M173, M45, SRY831, M207, M17 and PN25 (multiplex II). Finally, those individuals belonging to clade DE (YAP

derived branch) were further characterized for markers M96, P2, M123, M75, M78, M81, M33 and M35 (multiplex III). Amplification in multiplex was carried out in a two step PCR using only locus-specific amplification primers with a common 5'-end universal sequence at very low concentration and adding a high concentration of universal zip code primers after 15 cycles. Conditions and amplification primer sequences were slightly modified from Paracchini et al. (2002) except for markers M69, M201, P2, M207, PN25, and M75 for which new amplification primers were designed (Bosch et al., in preparation) and for polymorphisms 12f2 and SRY10831 whose primer sequences were modified from Blanco et al. (2000) and Whitfield et al. (1995) respectively, by adding universal code sequences at the 5' end. The afore mentioned biallelic markers were typed by a single-nucleotide primer extension using the SNaPshot Multiplex Kit (Applied Biosystems) following manufacturer's instructions. Oligonucleotide sequences will be available elsewhere (Bosch et al., in preparation). In order to compare the present results with other populations, Y-chromosome data from African (Cruciani et al., 2002), South Asian (Underhill et al., 2000; Kivisild et al., 2003), and European (Semino et al., 2000) populations were used.

Amplification of 11 YSTRs was performed within two multiplex reactions (MS1: DYS19, DYS388, DYS390, DYS391, DYS392 and DYS393; and EBF: DYS385, DYS389, DYS460, DYS461, DYS462 and amelogenin) as described in Bosch *et al.* (2002). PCR products were mixed with 400HD ROX standard and run on a ABI3100. Allele analysis and designation was carried out using haplotyped reference controls and the GeneScan Analysis Software v.3.7. Allele nomenclature in table 3 follows nomenclature used in the Y-STR Haplotype Reference Database (<http://ystr.charite.de>).

Nine Y-STR loci haplotypes were searched in the YSTR database (Roewer et al 2001, Kayser et al. 2002, Lessig et al 2003) using the Worldwide Search option. As of 17th December 2003, it contained 19,443 haplotypes in a set of 172 populations, 93 within Europe, 35 from Asia including North and South Caucasus, Iran, Pakistan, Iran and Northern India, plus four African populations from Cameroon, West Africa, Egypt, Mozambique. If no match was found, all haplotypes differing by one repeat at one locus were investigated.

Results

MtDNA lineage classification

The joint analysis of four SNPs within the coding region (10400-10873-12308-12705), the 9bp deletion of the COII/tRNA^{Lys} intergenic region and the sequence of both hypervariable segments of the mtDNA control region (HVRI and HVRII) allowed us to define a total of 18 different mtDNA lineages in the present sample set (Table 1). The 9bp deletion was typed as it is a useful marker to trace migrations out of southeast Asia and the Pacific (Redd et al., 1995), although the deletion might have arose several times as it has also been found in African populations (Soodyall et al., 1996).

Nonetheless, the 9bp deletion was not found in the present sample. The sequence diversity found (0.896 ± 0.030) is reduced compared to other African populations due to the fact that lineages LRN05 and LRN09 are found in ten and eight individuals respectively. One lineage (LNR07) is found in four individuals, two lineages (LRN06 and LRN13) are found in three individuals, while the rest are unique.

As a result of the refined phylogeography knowledge available for the mtDNA molecules, the present lineages were classified in four broad geographical regions: Africa, India, East Asia, Europe (Table 1). Indian lineages constitute the major contribution to the maternal genetic pool in La Réunion (~44%), followed by East Asian (~27%), European (~19%) and African lineages (~10%).

Within haplogroup M, the east and central Asian sub-haplogroups D, E, G, C, Z, M7, M8, M9, and M10 constitute the bulk of M lineages (Kivisild et al., 2002; Yao et al., 2002). The phylogeny of haplogroup M in Indian populations differs profoundly from that observed in east and central Asian populations (Bamshad et al., 2001), where

M2, M3, M4, M5, and M6 are found. Moreover, sub-haplogroup M1 has been found in East Africa (Quintana-Murci et al., 1999). Since the classification of sequence LRN09, belonging to haplogroup M, was difficult to refine taking into account the control region sequence, positions 7598 and 10384 of the mtDNA coding region were determined by direct sequencing in several individuals, and confirmed their adscription to haplogroup E1 (Kivisild et al., 2002). Therefore, in the present analysis, Indian lineages belonging to haplogroups M2 and M6, and East Asian lineages belonging to M7c, D5a and E1 haplogroups are found.

Four lineages belonging to the major African L haplogroup (Watson et al., 1997; Chen et al., 2000; Ingman et al., 2000; Salas et al., 2002) have been found in La Réunion. These L sequences (L1b1, L2a1, L3b, and L3e1) have a wide African distribution.

The N super-haplogroup is constituted by different group of lineages with a heterogeneous distribution: N1, N2, W, and I have a west Eurasian distribution (Richards et al., 2000); whereas N9, Y, and A have an East Asian origin (Yao et al., 2002). Moreover, sub-haplogroup X, has been found distributed at low frequencies across Eurasia and even in the Americas (Brown et al., 1998). In the present sample set only one sequence belonging to haplogroup I has been found.

The R haplogroup has also a Eurasian distribution: R9, B, and F sub-haplogroups are found in East Asia, whereas H, V, J, and T are the most frequent sub-haplogroups in west Eurasia. Only one F lineage is present in La Réunion.

Finally, haplogroup U is not present in eastern Asia, but is frequent in European populations and among Indians, although Indian U lineages differ substantially from those observed in Europe (Kivisild et al., 1999). Sub-haplogroups U2i and U7 constitute more than 90% of Indian U lineages, whereas U5 is the most frequent one in Europe.

Besides these European and Indian sub-haplogroups, U6 has been found in North Africa (Rando et al., 1998; Plaza et al., 2003). In the present analysis, European (U2e) and Indian sequences (U2i) have been found.

MtDNA sequence sharing

The African contribution is scant with only four lineages. In order to find shared sequences with other African populations, the compiled data from Salas et al. (2002) and Brehm et al. (2002) was used. LRN01, belonging to haplogroup L1b, was also found in one Fang and one Brazilian. LRN02 was found in one Bubi and one individual from Cabo Verde only differ in position 16092 and was classified within haplogroup L2a. LNR03 was found in four Fulbe and two Senegalese, and classified as L3b.

Indian lineages found in the Réunion were compared to sequences reported in Mountain et al. (1995), Kivisild et al. (1999), and Bamshad et al. (2001). LRN06, found in three individuals from La Réunion and classified as M6, is found in three Lobana individuals from the Indian region of Punjab. The two lineages classified as U2i are found in one Lambadi individual from Andra Pradesh (LRN07); and one Bogsa from Uttar Pradesh, and two individuals from Andra Pradesh (LRN08). It is noteworthy that the most frequent lineage in La Réunion, classified as M2, (LRN05) is not found in the data set used for comparison.

East Asian lineages were compared to a data set of published sequences (Horai et al., 1996; Lum et al., 2000; Fucharoen et al., 2001; Yao et al., 2002). Lineages LRN09, LRN10 and LRN12 were not found in the data set used for comparison. LRN10 and LRN12 lineages belong to haplogroups D5a and F respectively, which are widely spread in continental Asia. No matches were found for sequence LRN09, which

belong to haplogroup E1 and has been far only in Southeast Asia (Ballinger et al., 1992; Kivisild et al., 2002). Lineage LRN11, haplogroup M7c, deserves a special mention. Although it is not found in mainland Asia, except for a southeastern Asian individual (Thai, Fucharoen et al., 2001), it is found in the southeast Asian islands (Borneo, Java, Philippines) and Central-Eastern Micronesia, but it is not found neither in Western Micronesia nor in Polynesia (Lum et al., 2000).

The European lineages, represented by sequences belonging to sub-haplogroups H, J, T1, I, and U2e, are found at high frequencies in Europe and are scattered across Europe with no defined pattern. Nevertheless, the frequency pattern observed in La Réunion differs from that observed in Europe (Richards et al., 2000).

Y-chromosome lineages

The Y-chromosome biallelic markers analysed in the 17 males of the sample classify them into six different branches or clades of the Y phylogeny: 11 chromosomes belong to R1b (PN25 derived), two to KxP (M9 derived and M45 ancestral), and one single chromosome to each of the rest of clades: E3b3 (M123 derived), R1a1 (M17 derived), I (M170 derived), and G (M201 derived).

On the basis of the global geographical distribution of the Y chromosome biallelic haplotypes (or haplogroups) found in the Reunion Island, the origins of these chromosomes were investigated (Table 2). African Y chromosomes are quite distinct and specific African haplogroups are not found in the Reunion sample. Although the haplogroups found widely overlap in their geographical distribution between Europe and India, it is also true that their frequencies are not the same in the two regions: Y chromosomes belonging to clades R1b, G and I are more frequent in Europe than in

India, while the contrary is true for R1a1 and KxP. In order to ascertain with greater confidence the origin of these Y chromosomes, a search was conducted of their STR haplotypes against a world-wide Y STR database (<http://ystr.charite.de>). Matching patterns are shown on Table 3. In summary, 14 out of 17 chromosomes are well represented in Europeans while apparently absent in the south Asian samples of the Y-STR database. It is tempting to ascribe a European origin to these chromosomes. As for the remaining lineages, the two KxP chromosomes lacked any matches but one-step mutation derivatives were found in Han Chinese. Finally, the only E3b3 chromosome also did not present any matches but one-step mutation derivatives were present in Europe and Asia. Therefore, we may conclude that the Y chromosomes analysed from the Reunion Island come mostly from Europe with a sizeable Asian component. Given the small sample size, we have not attempted to quantify these admixture components or to generalize these conclusions.

Discussion

The analysis of maternal (mtDNA) and paternal (Y chromosome) lineages in the island of La Réunion has demonstrated that the extant population of the island is a melting pot of peoples that have settled in the island after its first colonization by the Europeans in the 17th century. The extant mtDNA gene pool in La Réunion is roughly composed by ~ 44% Indian, ~27% East Asian ~19% European, and ~10% African lineages, whereas the paternal contribution is mainly European although it is difficult to quantify due to the small sample size analysed.

Focusing in the mtDNA, the Indian contribution is the most frequent, although the diversity within this group of lineages is very low due to the large number of individuals who bear the same lineage. The entrance of these lineages in the island might be the result input of slaves from the Portuguese colonies such as Goa, and the influx of south Indians from the Malabar coast who came to fill the vacancies created by newly freed slaves on sugar cane plantations.

The African sequences found in the present sample are also found in some West African populations and not found in the closest continental African coast, Mozambique (Pereira et al., 2002; Salas et al., 2002), suggesting that the influx of African sequences might have been introduced directly by slavery from western Africa or through Madagascar.

It is worth to note the case of lineages LRN09 and LRN11, whose origin can be traced to South East Asian islands or Micronesia. In Madagascar, a substantial Pacific substrate is found and these lineages might have arrived to La Réunion from Madagascar, also via slavery and perhaps with other lineages of African origin. The

other two East Asian lineages might have arrived to the island with the Chinese migration.

The present analysis shows a clear bias of maternal and paternal lineages in La Réunion. This differential sexual pattern has also been observed in other human populations, such as the population of Brazil (Carvalho-Silva *Am J Hum Genet*), suggesting that the female and male contributions to extant populations have been different not only in number, being the migration rate higher in females than males as a general pattern (Seielstad et al., 1998), but also in origin.

References

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. Sequence and organization of the human mitochondrial genome. *Nature*. 1981 Apr 9; 290(5806): 457-65.

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB. Genetic evidence on the origins of Indian caste populations. *Genome Res*. 2001 Jun; 11(6): 994-1004.

Blanco P, Shlumukova M, Sargent CA, Jobling MA, Affara N, Hurles ME. Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J Med Genet*. 2000 Oct; 37(10): 752-8.

Bosch E, AC Lee, F Calafell, E Arroyo, P Henneman, P de Knijff and MA Jobling. High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic Science International* 125: 42-51 (2002)

Brown MD, Hosseini SH, Torroni A, Bandelt HJ, Allen JC, Schurr TG, Scozzari R, Cruciani F, Wallace DC. mtDNA haplogroup X: An ancient link between Europe/Western Asia and North America? *Am J Hum Genet*. 1998 Dec; 63(6): 1852-61.

Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet.* 2001 Jan; 68(1): 281-6.

Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet.* 2000 Apr; 66(4): 1362-83.

Comas D, Plaza S, Wells RS, Yuldasheva N, Lao O, Calafell F, Bertranpetit J. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet* (in press)

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet.* 2002 May; 70(5): 1197-214.

Ferrand G (1908) L'Origine africaine des Malgaches. *Journal Asiatique* 10: 353-500.

Fucharoen G, Fucharoen S, Horai S. Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet.* 2001; 46(3): 115-25.

Hewitt R, Krause A, Goldman A, Campbell G, Jenkins T. Beta-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from Bantu-speaking Negroids. *Am J Hum Genet.* 1996 Jun; 58(6): 1303-8.

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park KS, Omoto K, Pan IH. mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan.

Am J Hum Genet. 1996 Sep; 59(3): 579-90.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000 Dec 7; 408(6813): 708-13.

Kayser M, Brauer S, Willuweit S, Schdlich H, Batzer MA, Zawacki J, Prinz M, Roewer L, Stoneking M (2002) Online Y-chromosomal short tandem repeat (STR) haplotype reference database for U.S. populations. J For Sci 47 (3): 513-519

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. Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages.

Curr Biol. 1999 Nov 18; 9(22): 1331-4.

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. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet. 2003 Feb; 72(2): 313-32.

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol. 2002 Oct; 19(10): 1737-51.

Lessig R, Willuweit S, Krawczak M, Wu F-C, Pu C-E, Kim W, Henke L, Henke J, Miranda J, Hidding M, Benecke M, Schmitt C, Magno M, Calacal G, Delfin FC, De Ungria MCA, Elias S, Augustin C, Tun Z, Honda K, Kayser M, Gusmao L, Amorim A, Alves C, Hou Y, Keyser C, Ludes B, Klintschar M, Immel UD, Reichenpfader B, Zaharova B, Roewer L (2003) Asian Online Y-STR Haplotype Reference Database. *Legal Medicine* 5: 160-163.

Lum JK, Cann RL. mtDNA lineage analyses: origins and migrations of Micronesians and Polynesians. *Am J Phys Anthropol.* 2000 Oct; 113(2): 151-68.

Mountain JL, Hebert JM, Bhattacharyya S, Underhill PA, Ottolenghi C, Gadgil M, Cavalli-Sforza LL. Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet.* 1995 Apr; 56(4): 979-92.

Paracchini S, Arredi B, Chalk R, Tyler-Smith C. Hierarchical high-throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Res.* 2002 Mar 15; 30(6): e27.

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 Sep; 65(Pt 5): 439-58.

Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D. Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet.* 2003 Jul; 67(Pt 4): 312-28.

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 Dec; 23(4): 437-41

Rando JC, Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM, Bandelt HJ. Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann Hum Genet.* 1998 Nov; 62 (Pt 6): 531-50.

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro AS, Stoneking M. Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol.* 1995 Jul; 12(4): 604-15.

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. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet.* 2000 Nov; 67(5): 1251-76.

Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, Anslinger K, Augustin C, Betz A, Bosch E, Cagli A, Carracedo A, Corach D, Dekairelle A-F, Dobosz T, Dupuy BM, F|redi S, Gehrig C, Gusmao L, Henke J, Henke L, Hidding M, Hohoff C, Hoste B, Jobling MA, Krgel H-J, De Knijff P, Lessig R, Liebeherr E, Lorente M,

MartPascali VL, Penacino G, Ploski R, Rolf B, Sala A, Schmidt U, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Kayser M (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *For Sci International* 118 (2-3): 106 - 113.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A. The making of the African mtDNA landscape. *Am J Hum Genet.* 2002 Nov; 71(5): 1082-111.

Schneider S, Roessli D, Excoffier L. Arlequin ver 2.000: a software for population genetics data analysis. Genetics and Biometry Lab, University of Geneva, Switzerland, 2000.

Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans. *Nat Genet.* 1998 Nov; 20(3): 278-80

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science.* 2000 Nov 10; 290(5494): 1155-9.

Soodyall H, Jenkins T, Hewitt R, Krause A, Stoneking M (1996). The peopling of Madagascar. In *Molecular biology and human diversity*. AJ Boyce and CGN Mascie-Taylor editors. Cambridge University Press. Cambridge, UK.

Soodyall H, Jenkins T, Stoneking M. 'Polynesian' mtDNA in the Malagasy. *Nat Genet.* 1995 Aug; 10(4): 377-8.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ. Y chromosome sequence variation and the history of human populations. *Nat Genet.* 2000 Nov; 26(3): 358-61.

Vérin P (1986) *The History of Civilization in North Madagascar.* Rotterdam and Boston: Balkema

Watson E, Forster P, Richards M, Bandelt HJ. Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet.* 1997 Sep; 61(3): 691-704.

Whitfield LS, Sulston JE, Goodfellow PN. Sequence variation of the human Y chromosome. *Nature.* 1995 Nov 23; 378(6555): 379-80.

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet.* 2002 Mar; 70(3): 635-51.

Table 1. MtDNA lineages in La Réunion

Lineage	N	10400	10873	12308	12705	HVRI sequence ^a	HVRII sequence ^b	Haplogroup
African								
LRN01	1	C	A	T	C	111 126 187 189 223 239 270 278 293 311	73 146 151 152 182 185T 189 247	L1b1
LRN02	1	C	A	T	C	092 223 278 294 309 390	73 146 152 195	L2a1
LRN03	1	C	A	T	C	093 223 278 362	73	L3b
LRN04	1	C	A	T	C	093 223 311 327	73 150 185 189	L3e1
Indian								
LRN05	10	T	A	T	C	086 148 223 259 278 319	73 150 200	M2
LRN06	3	T	A	T	C	223 231 311 356 362	73	M6
LRN07	4	C	G	C	T	051 206C 230 311	73	U2i
LRN08	1	C	G	C	T	051 93A 154 206C 230 311	73 309.1 309.2	U2i
East Asian								

LRN09	8	T	A	T	C	221 223 291 362 390	73 309.1	M- E1
LRN10	1	T	A	T	C	164 182C 183C 189 223 266 362	73 150 207 309.1 309.2	M-D5a
LRN11	1	T	A	T	C	223 295 362	73 146 199 309.1	M7c
LRN12	1	C	A	C	T	266 278 304 309 356	73 152	F
European								
LRN13	3	C	G	C	T	051 129C 179 182C 183C 189 362	73 114 152 217 263 310	U2e
LRN14	1	C	G	C	T	051 93 129C 179 182C 183C 189 362	73 114 152 217 263 310	U2e
LRN15	1	C	A	C	T	126 163 186 189 294	64 73 152 195 309.1	T1
LRN16	1	C	A	C	T	069 126	73 295	J
LRN17	1	C	A	C	T	183 311	CRS	H
LRN18	1	C	A	T	T	93 129 223 311 355 391	73 199 204 250 309.1	I

Position numbers according to the CRS (Anderson et al., 1981). Unless indicated, substitutions are transitions.

^a Position numbers are given without the prefix “16”

^b Insertion of one or two cytosines are indicated by “.1” and “.2” respectively. All sequences have 263G and 315.1

Table 2. Y-chromosome lineage frequencies (%) found in La Réunion and comparison with different population samples.

	La Réunion ^a	Africa ^b	Pakistan-India ^c	India ^d	Europe ^e	French ^e
Y-chromosomes	17 (N)	473	88	325	1007	22
R1b	64.7 (11)	-	6.8	6.5	2.0 – 88.9	52.2
KxP	11.7 (2)	0.6	14.8	0.6	0 – 41.7	-
R1a1	5.9 (1)	13.1	31.8	27.1	3.7 – 60.0	-
I	5.9 (1)	-	-	-	2.2 – 44.8	17.4
G	5.9 (1)	-	8.0	0.3	0 – 30.1	-
E3b3	5.9 (1)	-	3.4	-	2.2 – 25.8	8.7

^a Present results

^b Haplogroup frequencies in Africa were obtained pooling all African populations analysed in Cruciani et al 2002 except Arabs, Berbers and Ethiopian Jews. R1b and KxP within Africa were actually confined to Cameroon population samples.

^c Underhill et al., 2000

^d Kivisild et al., 2003

^e Semino et al., 2000

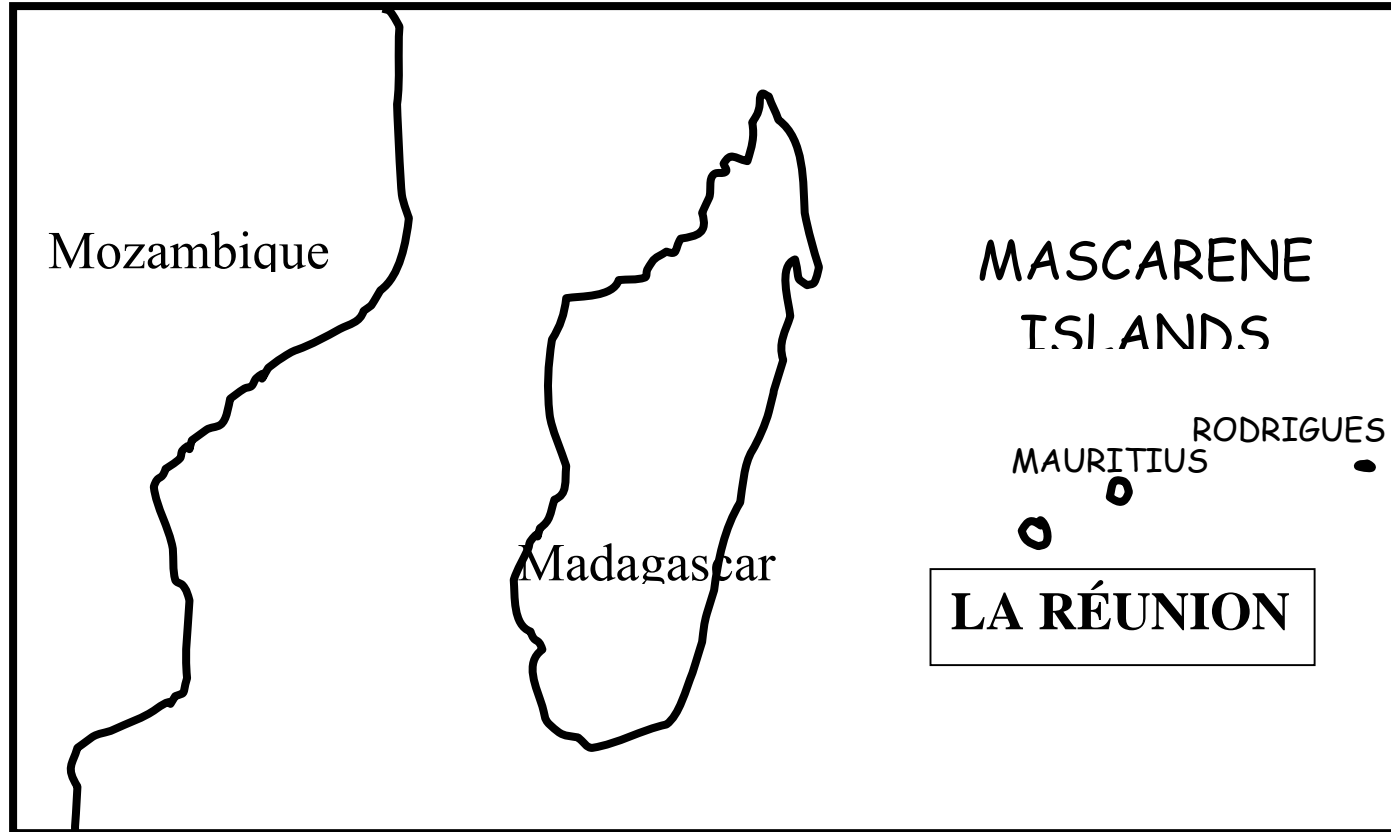
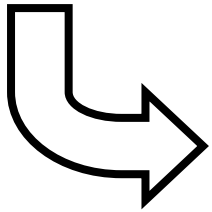
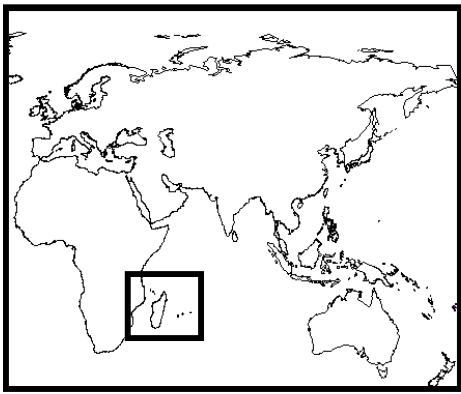
Table 3. Y-STR haplotypes found in La Réunion and matches within Europe, Asia and Africa in the Y STR database (<http://ystr.charite.de>T).

HG	9 Y-STR haplotype ^a	La Reunión	Europe	Asia	Africa
		N=17	N=13,253	N=2,912	N=328
R1a1	16-13-31-25-10-11-13-12/14	1	1 Ukraine	-	-
R1b	14-13-29-24-10-13-13-11/14	1	171	1 Iran	-
	14-13-29-24-11-12-13-11/14	1	12	-	-
	14-13-28-24-10-13-13-11/14	1	20	-	-
	14-13-29-24-11- ? - ? -11/14	1	491	8 Caucasus, 2 Turkey	
	? -12-28-24-11-14-13-11/14	1	6	-	-
	14-13-29-24-10-13-14-11/14	1	12	-	-
	14-14-30-23-11-13-13-11/14	2	35	-	-
	14-13-29-24-10-13-14-11/14	1	12	-	-
	15-13-30-24-10-13-13-11/14	1	8	-	-
	14-13-29-23-11-11-13-12/14	1	1 Northern Germany	1 North Caucasus	
E3b3	13-12-30-23-10-11-13-15/17	1	1 N Spain (1 step)	1 Turkey (1 step)	-
KxP	15-13-32-23-11-13-12-12/19	1	-	1 Han (1step)	-

	16-12-29-25- 9-13-12-13/20	1	-	1 Han	-
				(1 step)	
I	16-12-28-22-10-11-13-14/14	1	3	-	-
G	14-12-29-22-10-11-14-14/15	1	9	-	-

^a DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS385I/II

?: not determined



~ DISCUSSIÓ ~

I. Les poblacions de l'oest del Mediterrani

L'anàlisi del genoma mitocondrial a les poblacions de l'oest del Mediterrani ha permès caracteritzar la composició dels llinatges materns de les poblacions al Nord oest d'Àfrica i al Sud oest del Mediterrani, detectar i quantificar el flux genètic a cada banda del Mediterrani, i comparar els resultats obtinguts amb els patrons genètics trobats amb altres marcadors genètics com els marcadors clàssics, els microsatèl·lits autosòmics, els polimorfismes d'inserció Alu, i els polimorfisme del cromosoma Y.

El substrat genètic de les poblacions del Nord oest d'Àfrica i del Sud oest del Mediterrani tenen en comú tota una sèrie d'haplogrups que es troben majoritàriament a la resta d'Europa i de l'Orient Mitjà (H, J, T, U, I, W, X), fet que posa de manifest que aquestes poblacions tenen un origen comú recent, diferent del d'altres regions geogràfiques properes com per exemple l'Àfrica sub-Sahariana. De tota manera, s'observen diferències en alguns dels haplogrups a ambdues ribes del Mediterrani, a més de migració sub-Sahariana al nord d'Àfrica tal com s'exposarà seguidament.

I.1. Heterogeneïtat genètica de les poblacions del nord-oest d'Àfrica.

La principal diferència entre els dos grups de poblacions és la presència de l'haplogrup sub-saharià L al nord d'Àfrica. Hem quantificat en un 26% la contribució sub-sahariana al *pool* genètic de les poblacions del nord d'Àfrica, mentre que es troben en un ~3.0% a la Península Ibèrica i en un ~2.4% a Itàlia.. Els llinatges materns sub-saharians dintre del nord-oest d'Àfrica tenen una distribució en gradient:

les poblacions situades en una posició més meridional del nord-oest d'Àfrica, els mauritans, saharauis i berbers del sud del Marroc presenten 43.5%, 34% i 26% de seqüències L, respectivament, mentre que aquestes freqüències es troben en un ~3.0% i un ~13% en els berbers del nord del Marroc i en els mozabites. El mateix patró s'ha pogut observar amb el polimorfisme de les insercions Alu (Comas i col·laboradors. 2000), on un flux gènic sub-saharià ha tingut més impacte en les poblacions més meridionals del nord-oest d'Àfrica. L'estima de la presència dels llinatges sub-saharians, E1* i E3a* del cromosoma Y, mostra que ~8.0% dels llinatges paterns al nord-oest d'Àfrica tenen un origen sub-saharià (Bosch i col·laboradors. 2001). La contribució més elevada (26%) del llinatges femenins sub-saharians a la població del nord-oest africà és conseqüència d'una diferència entre sexes en el patró de migració des del sud del desert del Sàhara. Aquesta diferència segueix una tendència global on, per raons segurament socio-culturals, la dona ha migrat més que l'home (Seielstad i col·laboradors. 1998; Pérez-Lezaun i col·laboradors. 1999, entre d'altres).

També s'ha pogut detectar un flux gènic provinent de l'Àfrica de l'est mitjançant la presència de l'haplogrup M1. L'origen de l'haplogrup M1 es troba a l'Àfrica de l'Est (Quintana-Murci i col·laboradors., 1999) i s'han detectat freqüències més elevades d'aquest haplogrup a la part més oriental de l'Àfrica del nord-oest: a Algèria amb una freqüència de 12,8% i a Tunísia amb un 4.0%. Per la resta de les poblacions estudiades, només s'ha trobat en àrabs del Marroc amb un 2.0% i en mozabites amb un 4.7%. Però les freqüències dels haplogrups en els mozabites, coneguts com a grup berber aïllat, s'han de considerar amb precaució ja que poden haver estat molt afectades per la deriva genètica. Algèria presenta una freqüència de seqüències M1 semblant a les que es troben en els Etiòps (10%) i en els Nubis (8%) (Klings i col·laboradors. 1999).

El llinatge característic del Nord d'Àfrica és l'haplogrup U6 (Rando i col·laboradors. 1998). Tot i així, s'ha detectat en baixes freqüències en les poblacions estudiades: de 0% a Algèria fins a 28.2% en els mozabites. L'haplogrup U6

també es troba a la Península Ibèrica (~2.0%), a Itàlia (0.1%), a les illes Canàries, (13%), a l'Orient Mitjà, (<1.0%), i a l'Àfrica sub-sahariana (<2.0%). La limitada difusió de l'haplogrup U6 ofereix un contrast amb la seva antiga edat estimada al voltant dels $47\,000 \pm 18\,000$ anys. El fet que l'haplogrup U6 no hagi estat mai associat amb grans expansions poblacionals podria explicar aquesta distribució. L'única expansió de poblacions coneguda des del Nord d'Àfrica fins a la Península Ibèrica i Sicília va ser la invasió àrab al segle VII i va tenir un impacte demogràfic molt limitat (Camps, 1998). L'equivalent de l'haplogrup U6, el llinatge E3b2* amb un origen al Magrib representa 2/3 dels llinatges masculins al nord-oest d'Àfrica (Bosch i col·laboradors. 2001), fet que posa de manifest altre cop la diferència en els comportaments migracionals entre sexes, fent que hi hagi gran dispersió en els llinatges mitocondrials però no en els del cromosoma Y.

Tot aquest conjunt de diferències en els llinatges mitocondrials en les poblacions del nord-oest africà provoquen certa heterogeneïtat en les poblacions analitzades que contrasta amb l'alta homogeneïtat observada en les poblacions europees.

I.2. Homogeneïtat genètica de la poblacions del sud-oest europeu.

Front a l'heterogeneïtat genètica de les poblacions del nord-oest d'Àfrica s'ha de destacar l'homogeneïtat genètica de les poblacions del sud-oest d'Europa. Les poblacions de la Península Ibèrica i de la Península italiana presenten una homogeneïtat genètica similar a la que es pot observar per la resta de les poblacions europees tant pel mtDNA (Simoni i col·laboradors. 2000a; 2000b; Helgason i col·laboradors. 2000; Richards i col·laboradors, 2002) com per marcadors genètics clàssics (Cavalli-Sforza i col·laboradors, 1994) i el cromosoma Y (Semino i

col laboradors. 2000; Quintana-Murci i col laboradors. 2003). Aquesta homogeneïtat ha estat explicada per l'origen recent de les poblacions europees que malgrat tenir un origen al Paleolític Superior, han estat profundament influïdes per l'onada neolítica des de l'Orient Mitjà, encara que el grau d'influència d'aquesta onada està en discussió.

Els resultats obtinguts amb el mtDNA, conjuntament amb els dels polimorfismes d'insercions Alu (Comas i col laboradors. 2000), difereixen dels observats amb els marcadors genètics clàssics (Cavalli-Sforza i col laboradors. 1994; Calafell i Bertranpetit 1994; Cappello i col laboradors. 1996), i amb els polimorfismes del cromosoma Y (Caglià i col laboradors. 1997; Scozzari i col laboradors. 2001; Bosch i col laboradors. 2001) on els bascos i els sards apareixen com dos poblacions genèticament diferents de la resta de les poblacions europees. Aquest fet podria explicar-se per la major migració femenina, la qual cosa explicaria l'homogeneïtat observada pel mtDNA, però no pot explicar aquesta homogeneïtat mostrada pels marcadors autosòmics tipus Alu.

Els haplogrups majoritaris trobats al sud-oest de Europa són els haplogrups: H (50% a la Península Ibèrica ; 47% a Itàlia) que és també el més freqüent a Europa; U (13% a la Península Ibèrica; 9% a Itàlia); i T (7% a la Península Ibèrica; 12% a Itàlia). L'haplogrup V, pel qual s'ha proposat un origen a la Península Ibèrica (Torroni i col laboradors. 1998; 2001), presenta un rang de freqüències des del 2.7% en els sards fins 10.4% en els bascos. Però també s'ha detectat al nord oest d'Àfrica amb freqüències elevades en els saharuis (18%), i els berbers del sud del Marroc (10%). Mitjançant un network filogenètic vam poder determinar que el origen del haplogrup V és clarament europeu, encara que és difícil designar un origen geogràfic d'aquest haplogrup dins l'Europa meridional. El fet de trobar altes freqüències de l'haplogrup V en algunes poblacions del nord-oest africà pot explicar-se per deriva gènica.

I.3. Detecció de flux gènic a través de les dues ribes del Mediterrani.

Entre els llinatges detectats al nord oest d'Àfrica i al sud oest d'Europa alguns són específics de una banda del Mediterrani. Per tant la seva presència a l'altre riba del Mediterrani permet la detecció de flux gènic cap a una banda o l'altre del Mediterrani. S'han pogut detectar flux genètics des de la riba sud del Mediterrani cap a la riba nord i viceversa amb els haplogrup U6 i V. Les seqüències U6 presents a la Península Ibèrica i a Sicília pertanyen a llinatges d'origen diferent. S'han trobat seqüències específiques dels sub-haplogrups U6a i U6a1; i dels sub.-haplogrups U6b i U6b1. Els haplogrups U6a i U6a1 són específics del Nord d'Àfrica, però el haplogrup U6b1 s'ha trobat únicament a les illes Canàries i a la Península Ibèrica. Per tant, les freqüències dels haplogrup U6a, U6a1, i U6b1 permet detectar i quantificar flux genètics provenint de l'Àfrica del Nord i de les illes Canàries. S'estima les poblacions del nord oest d'Àfrica van contribuir en 18% a la composició dels llinatges mitocondrials de les poblacions de Península. En comparació, els llinatges del cromosoma Y específics del Nord oest (E3b2*) d'Àfrica componen 7% del llinatges masculí de la Península Ibèrica. La presència del llinatge U6b1 podria explicar-se per una migració recent dels Guanches (nadius de les illes Canàries) durant el regne de Castella al segle XVI. Aquesta aportació genètica de la illes Canàries al pool genètic mitocondrial s'eleva en un 1.5%. De la mateixa manera, el flux gènic de la banda nord cap a la banda sur del Mediterrani s'ha detectat i quantificat mitjançant el haplogrup V. El haplogrup europeu V es troba en una freqüència del 6.2% en les poblacions del Nord oest d'Àfrica, i el llinatge europeu R1b del cromosoma Y en una freqüència del 2.8%.

Els resultats obtinguts amb l'anàlisi del mtDNA ens permeten concloure que el mar Mediterrani i el desert del Sàhara no semblen haver actuar com fortes barreres al flux gènic entre les poblacions de l'oest del Mediterrani.

I.4. Els berbers de Tunísia

Els descendents de la població autòctona del Nord d'Àfrica, els berbers, estant presents avui en dia a Egipte, Líbia, Senegal, Tunísia, Algèria, i Marroc. Les poblacions àrabs i berbers es distingeixen només pel criteri lingüístic. Existeixen fins a 30 llengües berber i 20 milions de berberòfons. Els berbers de Tunísia parlen el Chelha i representen el 1% de la població global tunisiana. Es troben en quatre pobles del Sud de Tunísia, Sened, Matmata, Chenini, i Douiret; i a l'illa de Jerba. L'anàlisi de les poblacions berbers del Sud de Tunísia ens ha permès caracteritzar la seva composició genètica, i comparar-la amb la de la resta de les poblacions del Nord d'Àfrica.

I.4.a. Els llinatges mitocondrials dels Berbers de Tunísia

Els grups berbers analitzats estan separats per pocs quilòmetres de distància geogràfica, però es caracteritzen per una alta heterogeneïtat genètica. Presenten freqüències molt diferents dels mateixos haplogrups.

La majoria de les seqüències (~57%) dels berbers de Tunísia tenen un origen a l'oest d'Euràsia (la majoria d'haplogrups dins el super-haplogroup R), ~24% dels llinatges provenen de l'Àfrica sub-sahariana (haplogrups L1, L2, L3), ~16% de l'Orient Mitjà (haplogrup N, U3, U7, HV), i ~3% són autòctons de l'Àfrica del Nord (haplogrup U6). Entre les seqüències L3 trobades, un 50% duen una transició a la posició 16041. L'origen recent d'aquest grup de seqüències (13200 ± 3900 anys), la seva presència en altres poblacions nord africanes i la seva escassa presència a l'Àfrica sub-sahariana semblen indicar un origen nord africà d'aquest grup de seqüències, i no una aportació antiga per flux gènic sub-saharià.

El *pool* genètic dels berbers de Tunísia es caracteritza per la presència d'una freqüència elevada d'haplotips únics. És el cas, per exemple, de l'haplogrup M1 que es troba en una freqüència del ~11% en el berbers de Sened però que està representat per una única seqüència. També s'observa el mateix patró en els berbers de

Chenini-Douiret amb les seqüències de l'haplogrup HV1 i T. La presència de freqüències elevades d'haplotips únics en les poblacions berbers de Tunísia indica l'efecte de la deriva genètica que altera les freqüències dels haplogrups. Aquest patró genètic es correspon amb la història demogràfica d'aquestes poblacions que es van refugiar en la regions muntanyoses del sud de Tunísia per escapar a la dominació àrab. Aquesta deriva genètica consegüent a l'aïllament de les poblacions berbers de Tunísia ha provocat la heterogeneïtat genètica observada en els berbers de Tunísia. Entre els diferents grups berbers de Tunísia, els berbers de Chenini-Douiret apareixen com els més aïllats genèticament. És el grup que presenta el component eurasiàtic més elevat en el seu *pool* genètic, i la més baixa diversitat genètica. El cas dels berbers de Chenini-Douiret és similar al dels Mozabites on la deriva genètica consegüent a l'aïllament poblacional ha provocat un augment en freqüència de algunes variants de seqüències.

I.4.b. Els berbers de Tunísia en el context genètic del Nord d'Àfrica.

La composició haplotípica dels llinatges mitocondrials dels berbers de Tunísia és molt semblant a la que s'ha pogut observar en les poblacions del Nord d'Àfrica amb un fort component eurasiàtic, una aportació genètica sub-sahariana amb l'haplogrup L, i la presència de l'haplogrup U6 específic de l'Àfrica del Nord. La comparació entre berbers de Tunísia, d'Algèria i del Marroc mitjançant un AMOVA (anàlisi molecular de la variància) mostra que no hi ha cap diferència significativa entre els diferents grups berbers però que existeix una heterogeneïtat elevada (4.86%, $p < 0.01$) dintre dels grups berbers. Aquesta heterogeneïtat s'explica en gran part per la presència de grups berbers com els Mozabites, i berbers de Chenini-Douiret on la deriva genètica ha tingut un fort impacte genètic. La comparació entre poblacions àrabs i berbers, mitjançant els anàlisis de AMOVA, MDS (multidimensional scaling), i PC (components principals), revela una major heterogeneïtat dels berbers però cap

diferència significativa entre els grups àrabs i berbers. Aquest resultat confirma l'absència de diferència genètica revelada anteriorment amb els marcadors clàssics (Bosch i col·laboradors, 1997), els microsatèl·lits autosòmics (Bosch i col·laboradors, 1999), els polimorfismes d'insercions Alu (Comas i col·laboradors, 2000), i els llinatges del cromosoma Y (Bosch i col·laboradors, 2001). Aquesta manca de diferència entre poblacions àrabs i berbers confirma la hipòtesi de que l'arabització del Magrib va ser un fet bàsicament cultural amb poc impacte demogràfic.

II. La població d'Angola.

La població d'Angola, i en general el sud-oest del continent africà, fins a la present tesi no s'havia caracteritzat genèticament. Aquesta regió forma un peça important en el paisatge genètic africà. Angola va contribuir a la història de les poblacions humanes per dos grans esdeveniments. El primer és l'expansió Bantu que va tenir un impacte cultural (i segurament demogràfic) important sobre la població original d'Angola, i de tota l'Àfrica sub-sahariana, els Khoisan. El segon esdeveniment correspon a la formació de les poblacions d'Amèrica a l'època de l'Europa colonial durant la qual Angola va proporcionar dos milions d'esclaus al Brasil. L'anàlisi del mtDNA de la població anglesa ha permès determinar la composició genètica del sud-oest d'Àfrica, quantificar l'impacte de les migracions Bantu sobre la població original d'Angola, situar Angola en el paisatge genètic mitocondrial africà, i deduir la contribució dels llinatges del mtDNA anglesos a la població del Brasil.

II.1. Caracterització dels llinatges mitocondrials de la població anglesa.

Totes les seqüències de mtDNA trobades a Angola es classifiquen dins els haplogrups L0, L1, L2, i L3A, específics de l'Àfrica sub-sahariana. La composició genètica de la població d'Angola, basada en les freqüències d'aquests haplogrups, és genèticament semblant a les poblacions de l'Àfrica del Centre i de l'oest. Les anàlisis filogenètiques dutes a terme donen suport a una proximitat genètica més important amb les poblacions de l'Àfrica Central.

Una de les expectatives era trobar una gran proporció de llinatges L1c a Angola, ja que és un component important del *pool* genètic de la població del Brasil i s'havia proposat un origen angolès pels llinatges L1c trobats al Brasil (Alves-Silva i col·laboradors. 2000). Salas i col·laboradors (2002) van suggerir un origen a l'Àfrica

central de l'haplogrup L1c on la seva freqüència s'eleva fins un 22%. Com era d'esperar, Angola presenta un freqüència elevada (15.9%) de llinatges L1c, però cap de les seqüències L1c angolesees estan compartides amb les seqüències L1c trobades a la resta de Àfrica. La integració de les seqüències L1c d'Angola en un *network* filogenètic construït a partir de totes les seqüències L1c africanes (Salas i col·laboradors. 2002) ens ha permès descartar un possible origen angolès de l'haplogrup L1c. Un estudi recent (Destro-Bisol i col·laboradors. en premsa) sobre els llinatges del mtDNA a les poblacions del Camerun, ha revelat que més del 95% de les seqüències de mtDNA trobades a la població pigmea de Camerun (Mbenzele) i 22% d'aquelles en la població no pigmea (Ewondo) pertanyen a l'haplogrup L1c. Per tant, el origen de l'haplogrup L1c, ja suggerit per Salas i col·laboradors (2002) i recolzat amb els presents resultats, podria situar-se a l'Àfrica central en les regions de Camerun, Gabon i la República del Congo que encara no s'han caracteritzat genèticament.

II.2. L'expansió Bantu i el seu impacte genètic a les poblacions Khoisan

L'altre aspecte rellevant d'aquest estudi era d'intentar discriminar entre dos hipòtesis sobre l'expansió Bantu: una assimilació completa de les població Khoisan o bé un desplaçament de les poblacions Khoisan sense intercanvi genètic. En la mostra estudiada no s'ha trobat cap seqüència, L0d i L0k, característiques dels llinatges mitocondrials Khoisan. La contribució màxima dels Khoisan a la població Angola s'estima en un 10.8% (amb un interval de confiança del 0.05) i correspon a un reemplaçament dels llinatges mitocondrial Khoisan per els llinatges Bantu. Però la mida de la mostra (44 individus) no és suficientment gran com per poder discriminar entre una assimilació total del Khoisan o entre una absència d'intercanvis genètics entre les dues poblacions. L'absència de llinatges mitocondrials Khoisan i la presència de la

majoria dels llinatges Bantu (Bandelt i col·laboradors. 1995; Chen i col·laboradors. 1995; Watson i col·laboradors. 1997; Pereira i col·laboradors. 2001; Salas i col·laboradors. 2002) indiquen que la població anglesa està formada per un substrat genètic bàsicament Bantu.

II.3. Angola en el paisatge mitocondrial africà

Els llinatges del mtDNA de les diferents regions d'Àfrica descrits en el estudi de Salas i col·laboradors (2002) ens han permès situar Angola en un context genètic africà més ampli. L'anàlisi molecular de la variància (AMOVA) mostra que una agrupació d'Angola amb la regió del Centre d'Àfrica resulta en un augment de la variació genètica entre les diferents regions d'Àfrica (4.93, $p < 0.001$). Aquest resultat suggereix una major contribució de la migració Bantu oest al *pool* genètic d'Angola. Per una altra banda, l'anàlisi de correspondències, basat en les freqüències absolutes dels haplogrups, situa Angola en una posició intermèdia entre les regions oest, est i sud-est africanes. Aquests resultats indiquen que les dues expansions Bantu no s'han produït de forma independent sinó que possibles contactes i intercanvis entre les dues poblacions s'han produït entre les dues expansions Bantu. A més, la clara separació entre l'oest i l'est d'Àfrica en el anàlisi de correspondències suggereix que l'intercanvi s'hauria produït un cop les poblacions del sud-oest i sud-est es van diferenciar, segurament per la barrera de la selva tropical. L'anàlisi espacial de variància molecular (SAMOVA) revela una relativa homogeneïtat genètica entre les dues regions, Angola i Moçambic, més properes geogràficament. Deixant de banda la situació geogràfica, Angola ha rebut una major influència de l'expansió Bantu oest, mentre que l'expansió Bantu est ha tingut un impacte major a Moçambic. Però la proximitat geogràfica entre les dues ha permès l'intercanvi dels diferents llinatges mitocondrials i ha reduït les diferències entre els dos *pools* genètics. Aquestes dades genètiques concorden amb el

paisatge geogràfic africà: la selva tropical sembla haver actuat com una forta barrera geogràfica resultant en una clara separació entre les regions est i oest africanes. En canvi, les sabanes que separen el sud-oest i el sud-est d'Àfrica ha estat més fàcil de creuar i han estat més permeables als moviments poblacionals.

II.4. La contribució genètica d'Angola a la població del Brasil.

L'aportació de gran nombre d'esclaus angolesos a la població del Brasil suggereix un paper important d'Angola en la composició genètica dels brasilers. La comparació de seqüències de mtDNA revela que les dues poblacions comparteixen molt poques seqüències ja que totes dues presenten una gran diversitat genètica: 0.992 ± 0.007 pels angolesos i 0.994 ± 0.004 pels brasilers. En canvi les freqüències dels haplogrups trobats a Brasil i Angola són bastant similars. Però, Brasil forma un sol grup amb les poblacions de l'Àfrica central quan s'efectua un anàlisi de correspondències. Aquest resultat suggereix que la població del Brasil és el resultat d'una barreja genètica entre els esclaus provinents de les diferents colònies de l'Àfrica sub-sahariana situades a Saõ Tomé i al llarg de la costa de Guinea.

III. Les poblacions de l'Àsia Central.

L'Àsia Central ocupa una posició estratègica, a la frontera entre les poblacions europees de l'oest i la poblacions asiàtiques. Ha sigut l'encreuament de nombroses rutes comercials, i de cultures i de poblacions. El poblament d'aquest ampli territori és el resultat d'una barreja de poblacions, i aquesta complexitat es reflecteix en la composició genètica de la seva població.

L'anàlisi del genoma mitocondrial en 12 poblacions d'Àsia central revela que aproximadament la meitat del seu substrat genètic està format, a grans trets, per llinatges originaris a l'oest eurasiàtic i l'altre meitat està constituïda per llinatges de l'Àsia de l'est. Els llinatges asiàtics presents a l'Àsia Central són comuns tant al nord-est com al sud-est asiàtic. Aquest resultat i la diferència genètica entre el nord-est i el sud-est d'Àsia descrita per Kivisild i col·laboradors (2002) confirmen la complexitat del poblament de l'Àsia Central. Els haplogrups de l'oest trobats a l'Àsia Central estan dispersos per tota Europa i l'Orient Mitjà. La presència de l'haplogrup V a l'Àsia Central indicaria un possible contacte entre les poblacions de l'Europa del nord-oest i les de l'Àsia Central, com també ho indica els llinatges del cromosoma Y (Zerjal i col·laboradors. 1997).

Deixant de banda aquest gran gruix de llinatges de l'est i l'oest d'Euràsia a les poblacions de l'Àsia Central, en el present treball s'han definit alguns llinatges que poden haver-se originat a l'Àsia Central (6% per dos haplogrups, D4c i G2a) i una petita proporció (<5%) de seqüències d'origen a l'Índia. Els dos grups de seqüències D4c i G2a estan presents gairebé exclusivament a l'Àsia Central, encara que també es troben a baixes freqüències en algunes poblacions veïnes. La representació en un *network* filogenètic dels llinatges D4c i G2a mostra la gran diversitat d'aquestes seqüències en les poblacions de l'Àsia Central i la estima de la seva edat (~25.000 anys per l'haplogrup D4c; i ~30.000 anys per l'haplogrup G2a) indiquen que va haver-hi una expansió antiga al Paleolític superior d'aquests llinatges a l'Àsia Central, amb una

dispersió posterior cap a les poblacions veïnes. S'ha pogut observar el mateix patró amb els llinatges R1a del cromosoma Y (Wells i col·laboradors. 2001): s'estima un edat de 40.000 anys amb una localització restringida a l'Àsia Central.

La presència a l'Àsia Central de seqüències de l'oest eurasiàtic qüestiona la influència dels pobles de l'oest eurasiàtic a Àsia: quins són els límits d'aquesta influència de l'oest eurasiàtic? L'anàlisi de DNA antic en restes fòssils procedents de jaciments de l'est de la Xina (Wang i col·laboradors. 2000; Oota i col·laboradors. 1999) d'antiguitats entre 2500 i 2000 anys, va suggerir en un començament que la població inicial d'aquesta regió era similar a la població de l'oest eurasiàtic i que posteriorment va ser substituïda per poblacions de l'est. Però la reanàlisi de les dades (Yao i col·laboradors. 2003) ha revelat l'absència de seqüències d'origen de l'oest eurasiàtic en aquests estudis de DNA antic. Per tant, l'extensió de les poblacions de l'oest eurasiàtic, clarament presents a l'Àsia Central, no hauria assolit la regió més oriental del continent.

Els presents resultats ens permeten posicionar-nos sobre les hipòtesis prèvies que s'havien suggerit respecte al procés generador de la co-existència de seqüències de l'oest i de l'est a Àsia Central. Va ser l'Àsia Central una zona ocupada inicialment per poblacions de l'oest que van ser gradualment reemplaçades per poblacions de l'est, o bé aquesta regió ha estat una regió de contacte de dues poblacions genèticament diferenciades? L'anàlisi filogeogràfica de les seqüències de mtDNA revela que els haplogrup G2a i D4c, específics de l'Àsia Central, formen part dels haplogrups de l'Àsia de l'est G i D respectivament (Kivisild i col·laboradors. 2002). A més, l'antiga presència de G2a i D4c (~30.000 i ~25.000, respectivament) dóna suport a la hipòtesi d'una zona de contacte entre les poblacions europees i asiàtiques des de la època del Paleolític. Per una altra banda, vam intentar determinar quina dinàmica poblacional va generar la presència dels llinatges d'origen europeu i asiàtic a l'Àsia Central. La co-existència dels dos tipus de llinatges es pot explicar per migracions d'un conjunt de poblacions europees i asiàtiques en poc temps (moviment de grans grups poblacionals), o bé per un flux gènic constant de l'oest i de l'est durant moltes generacions. L'alt grau

de diversitat genètica dins de cada grup de seqüències dóna suport a la segona possibilitat.

Els llinatges del cromosoma Y (Zerjal i col·laboradors. 2002) descriuen un gran diversitat genètica a l'Àsia central amb zones de molt baixa i molt alta diversitat genètica. Aquestes dades també indiquen que la estructura patriarcal de les poblacions ha accentuat l'efecte produït per el fenomen de coll d'ampolla patit per algunes poblacions. En canvi, l'absència de estructura genètica del mtDNA en les poblacions d'Àsia Central mostra un patró de migració diferent entre homes i dones de l'Àsia Central, fet ja suggerit en un estudi anterior (Pérez-Lezaun i col·laboradors. 1999)

IV. La composició genètica de l'Illa de la Reunió

La composició genètica de l'illa de la Reunió és un bon exemple de barreja de poblacions diferenciades. Es té coneixement històric de la barreja d'africans, malgaixos, indis, xinesos i europeus a l'illa, encara que la proporció de cada contribució és difícil de determinar. El coneixement de la filogeografia dels llinatges mitocondrials, i en aquest cas particular també dels del cromosoma Y, ha permès fer la quantificació d'aquestes poblacions parentals. Segons les dades del mtDNA, la composició de llinatges materns és d'un 44% d'origen indi, un 27% est asiàtic, un 19% europeu i finalment un 10% africà. Les seqüències africanes que es troben a la Reunió són molt diverses i els seus orígens es remunten a localitzacions molt diferents dins del continent africà, la qual cosa reforça la idea de que la presència de llinatges africans és resultat del tràfic d'esclaus de diferents punts del continent africà i no de migració directa de la costa est africana. La presència de llinatges provinents del Pacífic a la Reunió suggereix que aquests van arribar a l'illa a través de migració des de l'illa veïna de Madagascar, la població inicial de la qual és d'origen indonesi.

Un aspecte interessant del poblament de l'illa de la Reunió és l'asimetria sexual observada gràcies a la comparació de llinatges masculins i femenins. Malgrat el baix nombre de llinatges masculins analitzats en el present treball, la gran majoria són d'origen europeu, fet que contrasta amb la composició del mtDNA. Aquest fet posa de manifest que les diferències sexuals que observem en les poblacions actuals no només són quantitatives (hi ha hagut una major mobilitat femenina) sinó que també apunten a una diferència qualitativa en les poblacions barrejades (diferents orígens en les contribucions masculines i femenines).

L'anàlisi poblacional generalitzat del mtDNA iniciat als anys 90, ha permès respondre qüestions relacionades amb l'origen de la nostra espècie, la colonització de

diferents continents i regions geogràfiques, i l'origen de poblacions més específiques. L'acumulació d'informació al llarg d'aquesta dècada de seqüències de control i de RFLPs en diferents poblacions ha possibilitat l'establiment d'una filogeografia força acurada dels llinatges mitocondrials humans. S'ha posat de manifest la necessitat de no limitar l'anàlisi del mtDNA únicament a l'estudi de la regió de control o d'alguns SNPs de la regió codificant, sinó que cal combinar ambdues estratègies i augmentar el nombre de marcadors analitzats. Actualment, els estudis poblacionals del mtDNA s'han de centrar en l'anàlisi de la regió de control juntament amb una àmplia bateria de SNPs a la regió codificant, i finalment, en un futur no gaire llunyà, tal com ja han començat a fer alguns treballs, l'anàlisi poblacional del mtDNA es farà mitjançant l'estudi de seqüències completes.

~ BIBLIOGRAFIA ~

- Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt HJ, 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
- Arnaiz-Villena A, Benmamar D, Alvarez M, Diaz-Campos N, Varela P, Gomez-Casado E, Martinez-Laso J (1995) HLA allele and haplotype frequencies in Algerians. Relatedness to Spaniards and Basques. *Hum Immunol* 43:259-268
- Arnaiz-Villena A, Martinez-Laso J, Alonso-Garcia J (1999) Iberia: Population genetics, anthropology, and linguistics. *Hum Biol* 71:725-743
- Attardi G, Chomyn A, Montoya J, Ojala D (1982) Identification and mapping of human mitochondrial genes. *Cytogenet Cell Genet* 32:85-98
- Attardi G, Montoya J (1983) Analysis of human mitochondrial RNA. *Methods Enzymol* 97:435-469
- Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. *Genome Res* 11:994-1004
- Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-753
- Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37-48
- 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 (2001) Phylogeography of the human mitochondrial haplogroup L3c: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* 65:549-563
- Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al. (1994) African origin

- of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci U S A* 91:12288-12292
- Batzler MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3:370-379
- Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet* 57:248-256
- Bendall KE, Macaulay VA, Baker JR, Sykes BC (1996) Heteroplasmic point mutations in the human mtDNA control region. *Am J Hum Genet* 59:1276-1287
- Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. *Ann Hum Genet* 59 (Pt 1):63-81
- Bertranpetit J, Junyent C (1998) *Viatge als orígens* (ed) Bromera
- Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Mol Ecol* 5:393-401
- Bodenteich A, Mitchell LG, Polymeropoulos MH, Merrill CR (1992) Dinucleotide repeat in the human mitochondrial D-loop. *Hum Mol Genet* 1:140
- Bogehagen DF, Sakonju S, Brown DD (1980) A control region in the center of the 5S RNA gene directs specific initiation of transcription: II. The 3' border of the region. *Cell* 19:27-35
- Bogehagen DF (1999) Repair of mtDNA in vertebrates. *Am J Hum Genet* 64:1276-1281
- Bosch E, Calafell F, Perez-Lezaun A, Comas D, Mateu E, Bertranpetit J (1997) Population history of north Africa: evidence from classical genetic markers. *Hum Biol* 69:295-311
- Bosch E, Calafell F, Santos FR, Perez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, Bertranpetit J (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* 65:1623-1638
- Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* 68:1019-1029

- Bowles GT (1977) *The people of Asia*. Willmwer Brothers Limited, Birkenhead, Great Britain.
- Brakez Z, Bosch E, Izaabel H, Akhayat O, Comas D, Bertranpetit J, Calafell F (2001) Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Ann Hum Biol* 28:295-307
- Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62:1408-1415
- Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci U S A* 77:3605-3609
- Brown MD, Hosseini SH, Torroni A, Bandelt HJ, Allen JC, Schurr TG, Scozzari R, Cruciani F, Wallace DC. (1998) mtDNA haplogroup X: An ancient link between Europe/Western Asia and North America? *Am J Hum Genet.* 63(6): 1852-1861.
- Budowle B, Allard MW, Wilson MR, Chakraborty R (2003) Forensics and mitochondrial DNA: applications, debates, and foundations. *Annu Rev Genomics Hum Genet* 4:119-141
- Caglia A, Novelletto A, Dobosz M, Malaspina P, Ciminelli BM, Pascali VL (1997) Y-chromosome STR loci in Sardinia and continental Italy reveal islander-specific haplotypes. *Eur J Hum Genet* 5:288-292
- Calafell F, Bertranpetit J (1994) Principal component analysis of gene frequencies and the origin of Basques. *Am J Phys Anthropol* 93:201-215
- Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet* 6:38-49
- Camps G (1998) *Los Bereberes: de la orilla del Mediterráneo al límite meridional del Sáhara*. Enciclopedia del Mediterráneo Icaria edition Barcelona
- Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325:31-36
- Cantatore P, Attardi G (1980) Mapping of nascent light and heavy strand transcripts on the physical map of HeLa cell mitochondrial DNA. *Nucleic Acids Res* 8:2605-2625
- Cappello N, Rendine S, Griffo R, Mameli GE, Succa V, Vona G, Piazza A (1996) Genetic analysis of Sardinia: I. data on 12 polymorphisms in 21 linguistic domains. *Ann Hum Genet* 60 (Pt 2):125-141

- Case JT, Wallace DC (1981) Maternal inheritance of mitochondrial DNA polymorphisms in cultured human fibroblasts. *Somatic Cell Genet* 7:103-108
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) History and geography of human genes. Princeton University Press, Princeton, NJ
- Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 33 Suppl:266-275
- Césari M (1999) Les marqueurs génétiques associés au diabète insulino-dépendant dans la population réunionnaise - Typage des gènes TAP et quantification des ARNm HLA DQ These
- Chakravarti A (1999) Population genetics--making sense out of sequence. *Nat Genet* 21:56-60
- Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc Natl Acad Sci U S A* 94:1041-1046
- Chang DD, Clayton DA (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. *Cell* 36:635-643
- 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* 57:133-149
- Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC (2000) mtDNA variation in the South African Kung and Khwe and their genetic relationships to other African populations. *Am J Hum Genet.* 66(4): 1362-1383.
- Chen X, Prosser R, Simonetti S, Sadlock J, Jagiello G, Schon EA (1995) Rearranged mitochondrial genomes are present in human oocytes. *Am J Hum Genet* 57:239-247
- Clayton DA (1984) Transcription of the mammalian mitochondrial genome. *Annu Rev Biochem* 53:573-594
- Comas D, Paabo S, Bertranpetit J (1995) Heteroplasmy in the control region of human mitochondrial DNA. *Genome Res* 5:89-90
- Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bertranpetit J (1998) HLA evidence for the lack of genetic heterogeneity in Basques. *Ann Hum Genet* 62 (Pt 2):123-132

- Comas D, Calafell F, Benchemsi N, Helal A, Lefranc G, Stoneking M, Batzer MA, Bertranpetit J, Sajantila A (2000) Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. *Hum Genet* 107:312-319
- Comas D, Plaza S, Wells RS, Yuldasheva N, Lao O, Calafell F, Bertranpetit J. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet* (in press)
- Côrte-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 (Pt 4):331-350
- Danan C, Sternberg D, Van Steirteghem A, Cazeneuve C, Duquesnoy P, Besmond C, Goossens M, Lissens W, Amselem S (1999) Evaluation of parental mitochondrial inheritance in neonates born after intracytoplasmic sperm injection. *Am J Hum Genet* 65:463-473
- Deka R, Shriver MD, Yu LM, Ferrell RE, Chakraborty R (1995) Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. *Electrophoresis* 16:1659-1664
- 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* 70:1009-1014
- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380:152-154
- Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci U S A* 91:3166-3170
- Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol* 18:672-675
- Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* 11:2571-2581
- Ellis N, Hammer M, Hurles ME, Jobling MA, Karafet T, King TE, de Knijff P, Pandya A, Redd A, Santos FR, Tyler-Smith C, Underhill P, Wood E, Thomas M, Cavalli-Sforza L, Ellis N, Jenkins T, Kidd J, Kidd K, Forster P, Zegura S,

- Kaplan M (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research* 12:339-348
- Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 24:363-367
- Estoup A, Garnery L, Solignac M, Cornuet JM (1995) Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* 140:679-695
- Excoffier L, Langaney A (1989) Origin and differentiation of human mitochondrial DNA. *Am J Hum Genet* 44:73-85
- 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
- Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics* 136:343-359
- Fawcett DW (1994) *A textbook of Histology* (ed) Chapman & Hall.
- Fucharoen G, Fucharoen S, Horai S (2001) Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet* 46(3): 115-125.
- Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. *Am J Hum Genet* 68:1475-1484
- Foran DR, Hixson JE, Brown WM (1988) Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. *Nucleic Acids Res* 16:5841-5861
- Forster P, Rohl A, Lunnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B (2000) A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet* 67:182-196
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225-2229
- Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 77:6715-6719

- Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K (1994) Identification of the remains of the Romanov family by DNA analysis. *Nat Genet* 6:130-135
- Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A* 92:6723-6727
- Goldstein DB, Clark AG (1995) Microsatellite variation in North American populations of *Drosophila melanogaster*. *Nucleic Acids Res* 23:3882-3886
- 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
- Greenberg JH (1963) *The Language of Africa*. Bloomington: Indiana University Press
- Greenberg BD, Newbold JE, Sugino A (1983) Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene* 21:33-49
- Hammer MF (1994) A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol* 11:749-761
- Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* 15:427-441
- Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18:1189-1203
- Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet* 55:175-189
- Harpending H, Sherry ST, Rogers AR, Stoneking M (1993) Genetic structure of ancient human populations. *Curr Anthro* 34:483-496
- Hedges SB, Kumar S, Tamura K, Stoneking M (1992) Human origins and analysis of mitochondrial DNA sequences. *Science* 255:737-739
- Helgason A, Sigurethardottir S, Gulcher JR, Ward R, Stefansson K (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* 66:999-1016

- Henke J, Henke L (1999) Mutation rate in human microsatellites. *Am J Hum Genet* 64:1473
- Hennessy RJ, Stringer CB (2002) Geometric morphometric study of the regional variation of modern human craniofacial form. *Am J Phys Anthropol* 117:37-48
- Henshilwood CS, d'Errico F, Yates R, Jacobs Z, Tribolo C, Duller GA, Mercier N, Sealy JC, Valladas H, Watts I, Wintle AG (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science* 295:1278-1280
- 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* 70:1152-1171
- Hertzberg M, Mickleson KNP, Serjeantson SW, Prior JF, Trent RJ (1989) An Asian-Specific 9-Bp Deletion of Mitochondrial-DNA Is Frequently Found in Polynesians. *American Journal of Human Genetics* 44:504-510
- Hirano M, Shtilbans A, Mayeux R, Davidson MM, DiMauro S, Knowles JA, Schon EA (1997) Apparent mtDNA heteroplasmy in Alzheimer's disease patients and in normals due to PCR amplification of nucleus-embedded mtDNA pseudogenes. *Proc Natl Acad Sci U S A* 94:14894-14899
- Hixson JE, Clayton DA (1985) Initiation of transcription from each of the two human mitochondrial promoters requires unique nucleotides at the transcriptional start sites. *Proc Natl Acad Sci U S A* 82:2660-2664
- Hopkin K (1999) Molecular biology - Death to sperm mitochondria. *Scientific American* 280:21-21
- Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A* 92:532-536
- Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713
- Jehaes E, Pfeiffer H, Toprak K, Decorte R, Brinkmann B, Cassiman JJ (2001) Mitochondrial DNA analysis of the putative heart of Louis XVII, son of Louis XVI and Marie-Antoinette. *Eur J Hum Genet* 9:185-190
- Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. *Am J Hum Genet* 56:11-14

- Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC (1997) Microsatellite diversity and the demographic history of modern humans. *Proc Natl Acad Sci U S A* 94:3100-3103
- Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 66:979-988
- Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics* 10:2199-2207
- Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* 94:1872-1877
- Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* 111:433-444
- 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 (1999) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages *Curr Biol.* 9(22): 1331-1334.
- 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* 19:1737-1751
- Klein J, Takahata N (2002) Where do we come from? The Molecular Evidence for Human Descent Springer Edition
- Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In Osawa S, Honjo T (eds). *Evolution of Life: Fossils, Molecules, and Cultures* Tokyo: Springer-Verlag pp391-413
- Krings M, Salem AE, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Paabo S, Stoneking M (1999) mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am J Hum Genet* 64:1166-1176
- Krings M, Geisert H, Schmitz RW, Krainitzki H, Paabo S (1999) DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. *Proc Natl Acad Sci U S A* 96:5581-5585

- Krings M, Capelli C, Tschentscher F, Geisert H, Meyer S, von Haeseler A, Grossschmidt K, Possnert G, Paunovic M, Paabo S (2000) A view of Neandertal genetic diversity. *Nat Genet* 26:144-146
- Lehninger M (2000) Principles of Biochemistry-3rd edition (ed) Nelson DL, Cox MM.
- Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203-221
- Lewontin RC (1972) Uses of Knowledge. *Science* 177:386-&
- Libert F, Cochaux P, Beckman G, Samson M, Aksenova M, Cao A, Czeizel A, et al. (1998) The *deltacr5* mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Hum Mol Genet* 7:399-406
- Livshits G, Nei M (1990) Relationships between Intrapopulational and Interpopulational Genetic Diversity in Man. *Annals of Human Biology* 17:501-513
- Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72:595-605
- Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* 2:13
- Macaulay V, Richards M, Sykes B (1999) Mitochondrial DNA recombination-no need to panic. *Proc R Soc Lond B Biol Sci* 266:2037-2039; discussion 2041-2032
- 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* 61:953-960
- Marchington DR, Macaulay V, Hartshorne GM, Barlow D, Poulton J (1998) Evidence from human oocytes for a genetic bottleneck in an mtDNA disease. *Am J Hum Genet* 63:769-775
- Margulis L, Olendzenski L, Afzelius BA (1990) Endospore-forming filamentous bacteria symbiotic in termites: ultrastructure and growth in culture of *Arthromitus*. *Symbiosis* 8:95-116
- Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S,

- Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc Natl Acad Sci U S A* 100:376-381
- Michaels GS, Hauswirth WW, Laipis PJ (1982) Mitochondrial DNA copy number in bovine oocytes and somatic cells. *Dev Biol* 94:246-251
- 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 (2003) Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A* 100:171-176
- Monnat RJ, Jr., Reay DT (1986) Nucleotide sequence identity of mitochondrial DNA from different human tissues. *Gene* 43:205-211
- Montoya J, Ojala D, Attardi G (1981) Distinctive features of the 5'-terminal sequences of the human mitochondrial mRNAs. *Nature* 290:465-470
- Morrall N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek M, Jr., Kalaydjieva L, et al. (1994) The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat Genet* 7:169-175
- Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M (2001) Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* 9:267-272
- Neckelmann N, Li K, Wade RP, Shuster R, Wallace DC (1987) cDNA sequence of a human skeletal muscle ADP/ATP translocator: lack of a leader peptide, divergence from a fibroblast translocator cDNA, and coevolution with mitochondrial DNA genes. *Proc Natl Acad Sci U S A* 84:7580-7584
- Newman J, *The peopling of Africa: Ageographic interpretation*. University Press, New Haven, CT
- Nordborg M (1998) On the probability of Neanderthal ancestry. *Am J Hum Genet* 63:1237-1240
- Nordborg M (2001) In *Handbook of statistical Genetics*, ed. Balding DJ, Bishop MJ, Canning C, UK: Wiley and sons pp. 179-212
- Oota H, Saitou N, Matsushita T, Ueda S (1999) Molecular genetic analysis of remains of a 2,000-year-old human population in China-and its relevance for the origin of the modern Japanese population. *Am J Hum Genet* 64:250-258

- Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet* 29:20-21
- Ovchinnikov IV, Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, Goodwin W (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* 404:490-493
- Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, Holland MM (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet* 15:363-368
- 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
- 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. *Ann Hum Genet* 65:439-458
- Perez-Lezaun A, Calafell F, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Allele frequencies for 20 microsatellites in a worldwide population survey. *Hum Hered* 47:189-196
- Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* 65:208-219
- Piko L, Matsumoto L (1976) Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. *Dev Biol* 49:1-10
- Poulton J, Macaulay V, Marchington DR (1998) Mitochondrial genetics '98 is the bottleneck cracked? *Am J Hum Genet* 62:752-757
- Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. *Trends Genet* 16:296-302
- Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of *Homo sapiens* from Africa through eastern Africa. *Nat Genet* 23:437-441
- Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES (2003) Genetic structure of Mediterranean populations revealed by Y-chromosome haplotype analysis. *Am J Phys Anthropol* 121:157-171

- 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 (Pt 6):531-550
- Rando JC, Cabrera VM, Larruga JM, Hernandez M, Gonzalez AM, Pinto F, Bandelt HJ (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet* 63 (Pt 5):413-428
- Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32:135-142
- Relethford JH, Harpending HC (1994) Craniometric variation, genetic theory, and modern human origins. *Am J Phys Anthropol* 95:249-270
- Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* 62 (Pt 3):241-260
- Richards M, Macaulay V, Torroni A, Bandelt HJ (2002) In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet* 71:1168-1174
- Richter C, Park JW, Ames BN (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc Natl Acad Sci U S A* 85:6465-6467
- Robin ED, Wong R (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *J Cell Physiol* 136:507-513
- Ruhlen M (1987) *A Guide to the World's Languages: Volume 1. Classification* London. Edward Arnold 76- 124
- Ruvolo M, Zehr S, von Dornum M, Pan D, Chang B, Lin J (1993) Mitochondrial COII sequences and modern human origins. *Mol Biol Evol* 10:1115-1135
- Saillard J, Magalhaes PJ, Schwartz M, Rosenberg T, Norby S (2000) Mitochondrial DNA variant 11719G is a marker for the mtDNA haplogroup cluster HV. *Hum Biol* 72:1065-1068
- 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
- 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* 71:1082-1111

- Salem AH, Badr FM, Gaballah MF, Paabo S (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *Am J Hum Genet* 59:741-743
- St John JC (2002) The transmission of mitochondrial DNA following assisted reproductive techniques. *Theriogenology* 57:109-123
- Satta Y, Takahata N (2002) Out of Africa with regional interbreeding? Modern human origins. *Bioessays* 24:871-875
- Schlotterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20:211-215
- Schneider S, Roessli D, Excoffier L. (2000) Arlequin ver 2.000: a software for population genetics data analysis. Genetics and Biometry Lab, University of Geneva, Switzerland.
- Scozzari R, Cruciani F, Pangrazio A, Santolamazza P, Vona G, Moral P, Latini V, Varesi L, Memmi MM, Romano V, De Leo G, Gennarelli M, Jaruzelska J, Villems R, Parik J, Macaulay V, Torroni A (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum Immunol* 62:871-884
- Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, Arredi B, Destro-Bisol G, De Stefano G, Rickards O, Martinez-Labarga C, Modiano D, Biondi G, Moral P, Olckers A, Wallace DC, Novelletto A (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* 65:829-846
- Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet* 20:278-280
- Seielstad M, Bekele E, Ibrahim M, Toure A, Traore M (1999) A view of modern human origins from Y chromosome microsatellite variation. *Genome Res* 9:558-567
- Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* 290:1155-1159
- Shay JW, Werbin H (1992) New evidence for the insertion of mitochondrial DNA into the human genome: significance for cancer and aging. *Mutat Res* 275:227-235

- Shields GF, Hecker K, Voevoda MI, Reed JK (1992) Absence of the Asian-Specific Region-V Mitochondrial Marker in Native Beringians. *American Journal of Human Genetics* 50:758-765
- Shoffner JM., Wallace DC (1995) Oxidative phosphorylation diseases. In: Scriver CR; Beaudet A.L; Sly WS; Valle D. (eds.) : *The Metabolic and Molecular Bases of Inherited Disease*. Vol. 1. New York: McGraw-Hill (7th ed.). Pp. 1535-1609.
- Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000a) Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet* 66:262-278
- Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000b) Reconstruction of prehistory on the basis of genetic data. *Am J Hum Genet* 66:1177-1179
- Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T (1996) mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* 58:595-608
- Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293:489-493
- Stoneking M, Soodyall H (1996) Human evolution and the mitochondrial genome. *Curr Opin Genet Dev* 6:731-736
- Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res* 7:1061-1071
- Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science* 239:1263-1268
- Stringer C (2002) Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci* 357:563-579
- Taanman JW (1999) The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta* 1410:103-123
- Takahata N, Lee SH, Satta Y (2001) Testing multiregionality of modern human origins. *Mol Biol Evol* 18:172-183
- Templeton AR (1992) Human origins and analysis of mitochondrial DNA sequences. *Science* 255:737

- The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 12:339-348
- Thomas H (1997) *The slave trade-the history of the Atlantic slave trade: 1440-1870*, Macmillan, London
- Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N, Goldstein DB (2000) Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa". *Am J Hum Genet* 66:674-686
- Torrioni 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* 93:189-199
- Torrioni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics* 144:1835-1850
- Torrioni 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* 62:1137-1152
- Torrioni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, et al. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 69:844-852
- Tsuzuki T, Nomiya H, Setoyama C, Maeda S, Shimada K (1983) Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. *Gene* 25:223-229
- Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26:358-361
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65:43-62

- Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med* 107:13-20
- Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial-DNA Sequences in Single Hairs from a Southern African Population. *P Natl Acad Sci USA* 86:9350-9354
- Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507
- Walberg MW, Clayton DA (1981) Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. *Nucleic Acids Res* 9:5411-5421
- Wall JD (2000) Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154:1271-1279
- Wallace DC (1982) Structure and evolution of organelle genomes. *Microbiol Rev* 46:208-240
- Wallace DC, Lott MT, Shoffner JM, Brown MD (1992) Diseases resulting from mitochondrial DNA point mutations. *J Inherit Metab Dis* 15:472-479
- Wallace DC (1993) Mitochondrial diseases: genotype versus phenotype. *Trends Genet* 9:128-133
- Wallace DC (1997) Mitochondrial DNA in aging and disease. *Sci Am* 277:40-47
- Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD (1997a) Ancient mtDNA sequences in the human nuclear genome: a potential source of errors in identifying pathogenic mutations. *Proc Natl Acad Sci U S A* 94:14900-14905
- Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. *Gene* 238:211-230
- Wang L, Oota H, Saitou N, Jin F, Matsushita T, Ueda S (2000) Genetic structure of a 2,500-year-old human population in China and its spatiotemporal changes. *Mol Biol Evol* 17:1396-1400
- Watson E, Bauer K, Aman R, Weiss G, vonHaeseler A, Paabo S (1996) mtDNA sequence diversity in Africa. *American Journal of Human Genetics* 59:437-444

- Watson E, Forster P, Richards M, Bandelt HJ (1997) Mitochondrial footprints of human expansions in Africa. *American Journal of Human Genetics* 61:691-704
- Weiner DB, Watson SR, Babcock G, Keller SJ (1986) Expression of human T antigens in interspecies hybridomas. *Cell Immunol* 100:197-209
- Weber JL (1990) Human DNA polymorphisms and methods of analysis. *Curr Opin Biotechnol* 1:166-171
- Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, et al. (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A* 98:10244-10249
- Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. *Nature* 378(6555): 379-380
- Wilson MR, Polansky D, Replogle J, DiZinno JA, Budowle B (1997) A family exhibiting heteroplasmy in the human mitochondrial DNA control region reveals both somatic mosaicism and pronounced segregation of mitotypes. *Hum Genet* 100:167-171
- Wittig H, Aschenbrenner A, Poetsch M, Krause D, Lignitz E (2003) Variability of mitochondrial DNA in a population sample from Iceland. *Leg Med (Tokyo)* 5 Suppl 1:S173-176
- Wrischnik LA, Higuchi RG, Stoneking M, Erlich HA, Arnheim N, Wilson AC (1987) Length mutations in human mitochondrial DNA: direct sequencing of enzymatically amplified DNA. *Nucleic Acids Res* 15:529-542
- Yao YG, Kong QP, Man XY, Bandelt HJ, Zhang YP (2003) Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. *Mol Biol Evol* 20:214-219
- Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *American Journal of Human Genetics* 71:466-482
- Zeviani M, Tiranti V, Piantadosi C (1998) Mitochondrial disorders. *Medicine (Baltimore)* 77:59-72
- Zhivotovsky LA (1999) Recognition of the remains of Tsar Nicholas II and his family: a case of premature identification? *Ann Hum Biol* 26:569-577

Zischler H, Geisert H, Castresana J (1998) A hominoid-specific nuclear insertion of the mitochondrial D-loop: implications for reconstructing ancestral mitochondrial sequences. *Mol Biol Evol* 15:463-469

Zischler H (2000) Nuclear integrations of mitochondrial DNA in primates: inference of associated mutational events. *Electrophoresis* 21:531-536