Diversitat genòmica a les poblacions del nord d'Àfrica

Elena Bosch Fusté

2000

Dipòsit legal: B.24348-2003
ISBN: 84-688-2285-X

Departament de Ciències Experimentals
i de la Salut (en constitució)

UNIVERSITAT POMPEU FABRA

Diversitat genòmica a les poblacions del nord d'Àfrica

Memòria presentada per Elena Bosch Fusté
per optar al grau de Doctor per la Universitat Pompeu Fabra

Jaume Bertranpetit i Busquets
Director de tesi

Elena Bosch Fusté

Barcelona, febrer del 2000

Als meus pares, al meu germà i a en Francesc

Agraïments

Des de finals d'agost de l'any 1995 són moltes les persones que amb paciència i bon humor m'han animat, recolzat i ajudat a realitzar aquest treball. Vull expressar-los ara el meu reconeixement.

En primer lloc, m'agradaria agrair a en Jaume Bertranpetit, director d'aquesta tesi, l'oportunitat que em donà per introduir-me a aquest món de la ciència. Puc dir ben alt i amb seguretat que m'honora haver-lo tingut com a mestre. He gaudit enormement amb el seu entusiasme i gran tafaneria científica.

També un agraïment ben carinyós a en Francesc Calafell per la seva extraordinària paciència, tant dins com fora el laboratori, i per la seva inestimable ajuda en l'estadística, redacció i discussió de tot el que ha passat entre les nostres mans.

Voldria agrair de manera molt especial l'ajuda desinteressada, actitud diplomàtica en tota mena de reunions o comitès d'ètica d'hospitals i el gran esperit aventurer de la Cristina Junyent; el bon rotllo i les explicacions per a realitzar les meves primeres tècniques de laboratori per part d'en David Comas, també company de viatge juntament amb la Cristina i en Jaume; les pacients classes de tot el que necessitava saber i més sobre els microsatèl·lits i el seu tipatge, així com el bon humor i exageracions catastròfiques de l'Anna Pérez; els petits consells i secrets sobre alguns protocols a la Blanca Gutiérrez i a la Neus Valveny; l'ajuda de la Rosa Martínez i l'Eva Mateu en l'expedició de València i sortides de matinada cap a Vic; el gran esforç realitzat per la mateixa Rosa, l'Araceli Rosa i en Jordi Clarimón per mimar al màxim els cultius cel·lulars, així com l'ajuda d'en Jordi en l'extracció de DNA de les darreres mostres que obtinguerem del Marroc.

El meu sincer agraïment també a tota la resta de companys de laboratori, per la vostra companyia, acudits, i aventures en els moments de feina, així com en els àpats i sobretaules compartides. A la Unitat d'Antropologia (UB) són la Bàrbara Arias, la Mireia Esparza, l'Esther Esteban, l'Antonio González, en Carles Lalueza, en Toni López, i en Marc Via; i ara, a Biologia Evolutiva (UPF), l'Aida Andrés, en Xavier Domingo, en Josep Marmi, en Bernal Morera, l'Stephanie Plaza, i la Mònica Vallès.

M'agradaria reconèixer també el suport rebut en dues estades a l'estranger. A Anglaterra, al Departament de Bioquímica de la Universitat d'Oxford on vaig trobar la càlida companyia d'en Fabrício R. Santos, de l'Arpita Pandya, de la Giovanna Arpidia, d'en Williams, d'en John, de la Milli, d'en Carlos, de l'Angela, d'en Leslie i del professor Chris Tyler-Smith. Als Estats Units, al Departament de Genètica de la Universitat d'Stanford on vaig conèixer en Peter A. Underhill, l'Anna Hurlbut, en Peidong Shen, l'Alice Lin, en Peter J. Oefner, la Phillys, la Silvia, en Titus, l'Adrianne i la Wie. Gràcies també a la Kay, la Cher, la Paula, la Marcie, i la Marga per la seva companyia a Montain View.

També vull agrair el suport dels professors de la Unitat d'Antropologia de la Universitat de Barcelona, Lourdes Fañanás, Clara García Moro, Miquel Hernández, Pedro Moral, Pasqual Moreno, Txomin Toja i Daniel Turbón. I un record especialment carinyós al doctor Josep Pons, company d'aventures científiques del meu avi, temps enrere.

Tenen el meu reconeixement també tots aquells donants de sang que acceptaren participar en aquest estudi. Sens dubte, gràcies a ells aquest treball ha estat possible. Igualment, voldria agrair la col·laboració inestimable d'investigadors de diverses universitats i de metges, infermeres i ajudants de diferents hospitals i dispensaris durant la recol·lecció de mostres. Són Omar Akhayat, Zahra Brakez, Hassan Izaabel, Omar Ouakrim (Agadir); Oriol Vall (Barcelona); Noufissa Benchemsi (Casablanca); Xavier Balanzó, Alba Bosch, Jordi Colomer, Josep Lluís Fernández Roure, Carme i Glòria (Mataró); Abdelaziz Sefiani (Rabat); Khadietu, Adda Brahim, Omar Mansur i Baha Mustafa (Sàhara Occidental); Elisabeth Pintado (Sevilla); Anne Cambon-Thomsen, Jean-Michel Dugoujon, Ghania Hariti (Toulouse); Enric Bufill, Assumpta i Montse (Vic); Enric Padrés, Felip Pi, doctora Prats i Pere Simonet (Viladecans), entre altres.

Gràcies també a l'Amaia, l'Anna, la Carme i en Ramon de la Unitat de Seqüenciació del Servei Científico-Tècnic de la Universitat de Barcelona per la seva assistència, ajuda i consells, setmanana rera setmana.

Un record per a les amigues Anna Campalans i Sílvia Busquets, antigues companyes de carrera, per tots els dinars dels dimecres on sovint rèiem de les nostres petites potineries al laboratori.

Finalment, també vull agrair als meus pares i al meu germà el seu suport continu.

Aquest treball l'he dut a terme mentre gaudia d'una beca de formació de personal investigador del Comissionat per a Universitats i Recerca, de gener de 1996 a desembre de 1999 (FI/96- 1.153). A més, el treball ha comptat amb les subvencions de la Dirección General de Investigación Científica y Técnica als projectes "Variación del genoma humano: análisis de diversas regiones genómicas de especial interés" (PB95-0267-CO2-01); i del Comissionat per a Universitats i Recerca com a Grup de Recerca Consolidat (1996SGR00141, 1998SGR00009).

ÍNDEX

MATERIALS I MÈTODES

RESULTATS

PRESENTACIÓ

El treball que presento a continuació pretén estudiar la variabilitat genètica de les poblacions del nord d'Àfrica a partir de l'anàlisi de diverses regions genòmiques per tal d'entendre les poblacions analitzades per una banda, i compendre la dinàmica del genoma per l'altra. Comença amb una *Introducció* on es caracteritzen les poblacions del nord d'Àfrica, així com les diferents eines genètiques emprades per al seu estudi. Segueixen els *Objectius* on presento les principals fites que es pretenien assolir en aquesta tesi. A continuació, a *Materials i Mètodes* s'introdueixen les poblacions estudiades, els marcadors tipificats, les tècniques emprades per a la seva anàlisi i els principals mètodes estadístics utilitzats al llarg del treball.

Els *Resultats* obtinguts s'exposen en cinc capítols que corresponen a un article publicat a la revista *Human Biology* on es realitza una anàlisi preliminar de la variació genètica al nord d'Àfrica a partir de marcadors *clàssics* (*Capítol I*); a dos articles acceptats, respectivament, a les revistes *European Journal of Human Genetics* i *International Journal of Legal Medicine* en els quals es treballa amb la informació genètica aportada pels microsatèl·lits autosòmics (*Capítol II*) i pels microsatèl·lits específics del cromosoma Y (*Capítol III*); a un article acceptat a la revista *American Journal of Human Genetics* on plantegem com s'estructura la variació dels microsatèl·lits en el cromosoma Y (*Capítol IV*); i a un article en preparació basat en la informació aportada per marcadors bial·lèlics específics del cromosoma Y tipats mitjançant DHPLC (*Capítol V*).

Finalment, la *Discussió* és un assaig d'interpretació conjunta de tots els resultats obtinguts.

Noteu que la numeració de taules i figures comença de nou a cada apartat.

INTRODUCCIÓ

L'estudi de la variació genètica entre els humans està experimentant un fort impuls en els darrers anys. Això ha estat possible gràcies al gran desenvolupament i aplicació de diverses tècniques del camp de la biologia molecular. Actualment, sabem que entre dos individus qualssevol hi ha diferències genètiques i que aquestes diferències es troben desigualment repartides en el nostre genoma i entre els humans de diferents poblacions. Hom està observant, a més, que gran part d'aquesta variació genètica és deguda a processos històrics de la formació de les poblacions. Sens dubte, cada dia són més nombrosos els exemples on hom mostra que la interpretació de la variació genètica és fonamental per a arribar a reconstruir parts del nostre passat encara no del tot resoltes des d'altres disciplines com l'arqueologia o la lingüística. En el present treball pretenc aplicar aquest tipus d'inferència històrica i poblacional a l'estudi de les poblacions del nord d'Àfrica. Per altra banda, el tipatge de diferents tipus de marcadors genètics que evolucionen a velocitats diferents al cromosoma Y ha permès d'analitzar l'estructura de la variació genètica en aquest cromosoma.

Introduirem, a continuació, les poblacions del nord-oest d'Àfrica, llur història i varietat lingüística, així com les diferents eines genètiques que hem emprat per intentar reconstruir llur història poblacional.


PORTUGAL
ESPAÑA
Sierra Morena
Sevilla
Córdoba
Granada
Málaga
Almería
Murcia
Alicante
Cartagena
Huelva
Jerez de la Frontera
Cádiz
Tánger
Tetuán
Orán
Mostaganem
Sidi-Bel-Abbès
Tlemcen
Oujda
Fès
Meknès
Kenitra
Salé
RABAT
Dar el Beida (Casablanca)
El Jadida
Safi
Marrakech
MARRUECOS
Essaouira
Agadir
Madeira (Port.)
Funchal
Islas Canarias
Santa Cruz de Tenerife
Las Palmas
Tenerife
Gran Canaria
El Aaiún
Hamada du Draa
Sahara Occidental
MAURITANIA
El Djouf
Aklé Aouana
Hodh
SENEGAL
BAMAKO
OUAGADOUGOU


ITALIA
Palermo
Messina
Reggio di Calabria
Sicilia
Catania
Siracusa
Ragusa
Mar Jónico
GRECIA
Mar Mediterráneo
MALTA
AL JAZĀ'IR (ARGEL)
Blida
Ech Cheliff
Sétif
Annaba
Skikda
Constantina
Batna
Biskra
Tebessa
TUNEZ
Bizerta
Sousse
Al Qayrawān
TUNICIA
Sfax
Qafsah
Qābis (Gabès)
Golfo de Gabès
Djerba
Chott el Djerid
Laghouat
Ghardaïa
Ouargla
Touggourt
El Oued
TARĀBULUS (TRÍPOLI)
Misrātah
Khalīj Sirt (Golfo de Sirte)
Benghazi (Bengasi)
Al Jabal al Akhdar
Jabal Nafūsah
Tripolitania
Gran Erg Oriental
Al Hamādah al Hamrā'
Hamada de Tinrhert
Atlas Sahariano
Meseta de Tademaït
In Salah
Ghāt
Sabhā
Fezzān
Dunas de Murzuq
Tassili N-Ajjer
Ahaggar
Tamanrasset
Hoggar
Adrar des Iforas
Aïr
Arlit
Agadez
Erg de Ténéré
Gran Erg de Bilma
Erg Brusset
Tibesti
Emi Koussi
Desierto del Tibesti
Desierto de Libia
NIGER
Zinder
Maradi
NIAMEY
Tahoua
Sokoto
Katsina
Kano
Zaria
Maiduguri
Lago Chad
N'DJAMENA
Kanem
Borkou
Bodélé
CHAD
LIBIA
ARGELIA

1. Nord d'Àfrica

1.1 Marc geogràfic

La zona estudiada engloba tota la regió natural del nord-oest d'Àfrica, que s'estén pel Sàhara Occidental (antic Sàhara Espanyol, des de 1976 annexat al Marroc), Marroc i Algèria, entre la Mediterrània i el Sàhara. Juntament amb Tunísia, aquest territori rep el nom de Magreb, denominació que prové dels geògrafs àrabs, els quals sovint consideraven aquesta regió de forma metafòrica com l'illa de ponent (*Djezira el Magreb*). És una regió de característiques mediterrànies, formada per un important conjunt de terres altes que bàsicament comprenen el Rif al nord del Marroc, el Gran Atlas (Haut Atlas), l'Atlas Mitjà (Moyen Atlas) i el Petit Atlas (Anti Atlas) que travessen de sud-oest a nord-est el Marroc, més l'Atlas Saharià, l'Aurès i el Tell Atlas al nord d'Algèria i les muntanyes d'Ahaggar al sud; però que a la vegada engloba, com a primers indicis de la presència del gran desert del Sàhara, àmplies extensions desèrtiques, els anomenats *ergs* i *regs*, ocasionalment puntejades per algun oasi. Són notables també les valls dels rius Dra, Souss, Oum er-Rbia, Sebou, Mouloudauya i Chelif, entre altres.

Si bé l'àmbit geogràfic sota estudi és la part occidental del Magreb, sovint trobarem que aquesta regió està clarament condicionada i lligada a la geografia, poblament i història d'una regió més àmplia, la del nord d'Àfrica, entenent com a tal, l'àrea compresa entre l'Atlàntic i el Mar Roig, i entre la Mediterrània i el Sahel. L'aïllament geogràfic, imposat per la Mediterrània i especialment pel desert del Sàhara, sembla haver condicionat fortament el poblament del nord d'Àfrica, limitant els principals moviments humans a una direcció est-oest. Tanmateix, l'extensió demogràfica d'aquests moviments al nord-oest resta, en principi, desconeguda.

Al llarg del pleistocè i l'holocè, hom distingeix en el Sàhara un mínim de cinc períodes àrids i plujosos. Diferents evidències mostren que, en èpoques humides, va estar més extensament ocupat que en l'actualitat. L'últim període humit engloba part del paleolític superior i el neolític i es considera que finalitza al tercer mil·leni abans de

Crist. A partir d'aquest moment, comença el període àrid present (Said i Faure, 1990). Tot i la presència de poblacions nòmades i l'existència de les rutes comercials trans-sahari anes, en aquest punt el desert clarament limita el contacte de la regió sota estudi amb la resta del continent africà. Al contrari, al llarg de la història són nombrosos els exemples en què la Mediterrània ha permès el contacte del nord d'Àfrica amb d'altres pobles, resultant més un vincle que no pas una barrera entre poblacions.

Per altra banda, la muntanyosa orografia, juntament amb el fet de poder trobar sovint la possibilitat de petits assentaments envoltats per grans extensions àrides, han afavorit una certa fragmentació de la població nativa dins la regió sota estudi, que s' ha vist reforçada per una estructura social tradicional de tipus tribal, tot i la relativa uniformitat que algunes cultures intentaren imposar.

1.2 El poblament prehistòric

Les primeres evidències d'ocupació humana a la zona (Aïn Hanech, Ternifine i Sidi Abderrahman) daten fins a uns 700.000 anys i corresponen a restes classificades com a *Homo erectus*, el qual probablement travessà la barrera sahariana per la vall del Nil (McEvedy, 1980). S'han trobat diferents exemples de la cultura acheuliana (paleolític inferior) distribuïts al llarg d'una gran regió d'Àfrica que s'estén des del Magreb fins al cap de Bona Esperança (Newman, 1995). Per altra banda, es troba ben documentat, també, el desenvolupament d'indústries posteriors, com la mosteriana, que evolucionà de forma local al nord d'Àfrica donant lloc a la cultura ateriana (paleolític mitjà), amb restes al Magreb i gran part del Sàhara (Newman, 1995). Les poques restes humanes que hom hi troba associades són neandertaloides i presenten una certa diferenciació est-oest. A les indústries íbero-marusianes considerades del paleolític superior (Camps, 1974) que s'estengueren per la costa del nord d'Àfrica, segueix la capsiana (de Capsa o Gafsa), d'origen incert. Tot i que és probable que aquesta cultura vingués d'Europa, els seus orígens es podrien trobar més cap a l'est (Desanges, 1990). Amb el nom de tradició capsiana del neolític va perviure durant l'expansió de la cultura neolítica. El neolític entrà al nord d'Àfrica des de l'est, on sens dubte contribuí al sorgiment del Regne d'Egipte, i s'estengué, lentament, al llarg de la costa mediterrània cap al Magreb (McEvedy, 1980). Hi ha evidències que la cultura neolítica del Magreb no fou introduïda per invasió, sinó per l'acceptació de les noves

tecnologies per part dels pobles capsians, la cultura dels quals es troba fins el primer mil·leni abans de Crist (aC). La quantitat i abast geogràfic del flux gènic, si n'hi ha hagut, associat amb l'expansió del neolític és altament controvertit.

1.3 Història de les poblacions del nord d'Àfrica

Exceptuant Egipte, no hi ha una història escrita pels propis pobles del nord d'Àfrica fins les primeres invasions musulmanes del segle VII. Tot el que en sabem prové dels relats dels pobles que conquereixen o colonitzen la regió, que són nombrosos. Els habitants més antics que han viscut a la mateixa regió des de l'inici de la història són els que actualment coneixem com a berbers (Desanges, 1990).

El nord d'Àfrica entra a la història de la Mediterrània amb l'arribada dels fenicis, els quals establiren un imperi centrat a Cartago (814 aC), que monopolitzà el comerç de tota la Mediterrània occidental. Tot i que la influència fenícia fou considerable, les relacions amb la població nativa -els components ètnics de la qual ja estaven ben establerts a començaments del primer mil·leni aC- es limitaren bàsicament a l'esclavitud, pagament d'impostos i subministrament de tropes. És en el període de decadència de l'imperi cartaginès (segles III-II aC) quan apareixen els primers regnes natius, *mauri* (Marroc), *numidae* (Algèria), *gaetuli* (Tunísia) i el regne dels libis (a Líbia), que foren posteriorment romanitzats.

La civilització romana (del 146 aC al 250 de la nostra era) s'estengué pel nord d'Àfrica desenvolupant-se fonamentalment a la part més oriental (Tunísia i nord d'Algèria). La magnitud de la romanització és controvertida, amb escàs impacte a les regions de muntanya (Newman, 1995). L'arribada dels vàndals (Gaiseric, l'any 429) que creuaren l'estret de Gibraltar procedents de la Península Ibèrica, i posteriorment la dels bizantins (Belisari, l'any 533) van tenir un efecte purament administratiu i polític més que no pas demogràfic.

El trencament entre el món antic i medieval arriba al Magreb amb les primeres expedicions musulmanes (l'any 643), inicialment confinades a Egipte. Al segle VII, la majoria d'habitants del Magreb eren berbers, amb un substrat cultural (llengua i civilització) comú, però, amb una estructura social tribal notablement fragmentada. Després de la fundació d'al-Qayruan (Túnisia) al 670, tota nord Àfrica fou ràpidament islamitzada (McEvedy, 1980). Sovint, el poble berber convertit formava part de l'exèrcit

àrab. És en aquest període, concretament al 711, quan Tarik passa a la Península Ibèrica (al-Andalus en àrab), una incursió en part facilitada per la crisi interna del poder visigot a la Península (Hitti, 1990).

Tanmateix, la primera onada d'invasors àrabs no fou realment un reemplaçament poblacional; només n'arribaren uns quants milers, que romangueren majoritàriament en el que ara és Tunísia i est d'Algèria, i a les grans ciutats (McEvedy, 1980; Newman 1995). La resta de la regió va seguir habitada per berbers però organitzada en diversos regnes sota diferents graus de control musulmà com el de Sigilmassa (772-977), el d'Idrissid (788-926, Fes i nord del Marroc), el Rustamid de Tahart (787-911, Magreb central), i el dels *aghlabids* (800-909, a Tunísia).

El xiisme penetrà al Magreb al segle IX i fou ràpidament adoptat pels berbers, els quals prestaren suport al nou poder fatimita en la seva expansió nord-africana (Kasule, 1998). Després d'un període de guerres religioses, els fatimites succeïren els *aghlabids* (909) i conqueriren Egipte (964).

A partir del 1051, tingueren lloc les invasions dels àrabs beduïns *Banu Hilal* (a Túnisia, est d'Algèria i Cirenaica), *Banu Sulaym* (Tripolitània) i *Ma'gil*. A meitat del segle XIII, el poder beduí ocupava tota la part oriental de la regió, mentre la meitat occidental es mantenia sota altres dinasties berbers. Els àrabs beduïns, que foren molt més nombrosos que els primers invasors, sí que realment afectaren la situació demogràfica, arraconaren la població berber i arabitzaren la regió. S'inicià també el moviment cap al sud dels pobles nòmades saharians (Newman 1995).

Entre els segles XI i XV sorgeixen dues importants dinasties berbers islàmiques: la dels *almoràvids* (1056-1147) i la dels *almohades* (1121-1269), que unificaren sota el seu control tot el Magreb des de Sous a Trípoli. És en aquest període quan el domini musulmà conjunt, sobre el nord d'Àfrica i la Península Ibèrica, assoleix la seva màxima expansió. Després de la caiguda dels *almohades*, aparegueren les darreres dinasties medievals del nord d'Àfrica: els *hàfsids* (Tunísia), els *Abd al-Wadid* de Tilimsem (oest d'Algèria) i els *merínids* (Marroc), entre altres (McEvedy, 1980).

Els portuguesos, al llarg de la costa atlàntica, i els espanyols a la costa mediterrània, fundaren els primers forts a partir del segle XV, coincidint amb la caiguda dels *merínids*, que foren succeïts pels *wattàsids*. Comença a aquesta època l'establiment de les dues dinasties xarifianes que han existit al Marroc: la saadi (1525-1659) i la alawita (des de 1659). La presència de turcs otomans, durant els segles XVI-

XIX, comporta un nou element de diversitat ètnica al nord d'Àfrica, especialment altre cop als centres urbans (Newman 1995).

Finalment, entre els darrers invasors del nord d'Àfrica a partir del segle XV s'inclouen els europeus: portuguesos i espanyols al Marroc, francesos al Marroc, Algèria i Tunísia, i italians a Líbia. No fou fins el 1830 que la presència colonial europea començà a crear les condicions que donaren origen als moviments nacionalistes que, finalment, portaren a la independència dels països actuals (Líbia 1950, Tunísia i Marroc 1956, Mauritània 1960, Algèria 1962) excepte pel cas del Sàhara Occidental, antic Sàhara Espanyol que fou annexat al Marroc al 1976 i resta a la espera de la celebració del referèndum d'autodeterminació.

1.4 Pobles berber, àrab i sahrauí

El poble berber representa una de les civilitzacions més antigues de la Mediterrània. Són els antics pobladors de tot el Magreb, amb grups discontínuament distribuïts des de l'oasi de Siwa a Egipte fins a l'Atlàntic, incloent l'antiga població guanxe de les Canàries, i de fet, es consideren el poble natiu del nord d'Àfrica. Viuen en grups compactes, generalment a la muntanya, al Marroc (Rif, Atlas), a Algèria (Gran i Petita Kabília, Aurès i Mzab), i en algunes zones del desert de Egipte, Líbia, Tunísia i Mali.

En temps clàssics els berbers també foren anomenats libis o númides. Herodot (446 aC) en fa una de les primeres referències en parlar dels pobladors de Líbia, la zona que s'estén des d'Egipte fins a les columnes d'Hèrcules. La paraula berber, però, és d'origen greco-llatí i prové de la paraula *barbarus*. El poble berber, en canvi, sol utilitzar el mot *amazight*, en plural *imazighen*, que significa "home lliure". La forma femenina, *tamazight*, s'empra per a fer referència a la seva llengua.

No és fins el segle VII que el poble àrab, provinent de l'Orient Mitjà, realitza les primeres incursions al nord d'Àfrica i comença a imposar la seva llengua i religió sobre la població berber, que acabà convertint-se a l'islam. Tanmateix, es considera que la veritable onada àrab amb impacte demogràfic es donà a meitat del segle XI.

La diferenciació entre àrabs i berbers no sempre és evident, sinó que es basa essencialment en la llengua: es considera berber aquell que en conserva la llengua. Molts marroquins àrabs, per exemple, originàriament eren berbers (o formaven part

d'una població mixta), que amb el transcurs del temps han perdut la seva llengua i s'han arabitzat. Actualment, les ciutats més importants són àrabo-parlants, cosa que no vol dir que no hi hagi ciutats absolutament berbers (com Fes i Agadir, respectivament, al nord i sud del Marroc o Ghardaïa a Algèria). Per altra banda, el món rural es divideix en zones de parla àrab, a la vora de les ciutats, i zones de parla berber, normalment més lluny de la influència urbana, tot i que sense un aïllament total (Camps, 1994).

El poble sahrauí, població natural del Sàhara Occidental, és d'origen mixt, bàsicament àrabo-berber, amb una important contribució de les tribus berbers dels Sanhaja. La seva llengua és el *hassani*, un dialecte de l'àrab diferent a l'àrab parlat a la resta de la regió. El sentit d'identitat pròpia és molt patent en la població sahraui, i té la seva expressió en la lluita per l'independència del seu territori.

1.5 Les llengües del nord d'Àfrica

Sovint, l'evolució de la llengua es dóna de manera paral·lela a l'evolució genètica. Quan una comunitat creix i s'expandeix a noves regions, els grups que la conformen es van separant de la seva regió d'origen i s'estableixen en llocs nous, dels quals sorgeixen altres grups, que de nou en expandir-se, continuen el seu camí cap a llocs més distants. En algun punt serà impossible mantenir el contacte amb els llocs i la població d'origen. L'aïllament de molts grups que s'han format d'aquesta manera ha determinat dos fenòmens inevitables: la formació de diferències genètiques i la formació de diferències lingüístiques. Tot i seguir els seus camins propis i la seva pròpia dinàmica, la història de les separacions, que són la causa de la diferenciació genètica i lingüística, és comuna en ambdós casos (Cavalli-Sforza, 1994). Aquest fenòmen ens porta a parlar de lingüística dins un estudi de variabilitat genètica.

La gran majoria de les llengües que actualment es parlen al nord d'Àfrica pertanyen a la família afroasiàtica. Segons Renfrew (1991) les llengües afroasiàtiques (abans denominades camito-semítiques) s'haurien expandit a l'Àfrica des de l'Orient Mitjà amb l'anomenada difusió dèmica del neolític, juntament amb altres tres famílies lingüístiques: l'altaica cap a l'Àsia Central, la dravidiana cap a l'Índia i la indoeuropea cap a Europa. Tanmateix, Barbujani i col·laboradors (1994) descriuen que l'evidència

genètica més dèbil per a una expansió demogràfica de les famílies lingüístiques des de l'Orient Mitjà correspon, justament, a l'afroasiàtica.

Dins la família afroasiàtica es troben representades, al nord d'Àfrica, dues grans branques: la semítica i la berber. La branca semítica al nord d'Àfrica inclou diversos dialectes de l'àrab. Arribà a la zona amb l'expansió de l'islam a partir de finals del segle VII, s'estengué com a vehicle de la fe musulmana i aviat arraconà les llengües autòctones de la branca berber, que fins llavors es parlaven per tot el nord d'Àfrica. Amb tot, encara existeixen fins a unes 30 llengües de la branca berber amb un total aproximat de 20 milions de parlants, que es divideixen en quatre grans grups (vegeu figura 1.5).

i) les tres llengües tuaregs (el *tamahaq*, el *tamazhig* i el *tamasheq*, que sembla ser l'únic parlar que ha conservat fins avui l'escriptura anomenada *tifinagh*, segurament derivada de l'antiga escriptura líbica) al Sàhara central i Níger.
ii) el grup oriental, que inclou les diferents llengües berbers de Líbia i Egipte (*nefusi*, *zwara*, *siwa*, *awjila*, *sokna* i *ghadames*).
iii) el grup occidental, que comprèn una llengua berber aïllada, el *zenaga*, parlat al Senegal i sud de Mauritània,
iv) la branca central, que engloba la majoria de les llengües berbers d'Algèria (amb el *shawiya*, *gourara*, *kabila* o cabilenc, el *mzab*, i l'*ouargla* entre altres), Marroc (amb el *tarifit* o *rifeny* propi del nord a la regió del Rif, el *tamazigh* parlat al nord de l'Atlas marroquí, i el *tahelhit* o *shilha* a la vall del Sous i a la banda occidental del Gran i Petit Atlas) i Tunísia (*jerba*, *seud*, *tamazret, taoujjout*, *tmagourt* i *zawa*).

Dos grups extingits pertanyen també a la branca berber: el guanxe parlat pels primers habitants de les Canàries i l'antic libi. Les llengües berbers es consideren parles àgrafes; bàsicament, es transmeten i conserven per mitjà de l'ús i de la tradició oral. Recentment, s'intenta recuperar el seu alfabet mil·lenari, el *tifinagh*, en la versió actualitzada anomenada *neo-tifinagh*, donat que ha estat enriquit amb alguns signes complementaris per adaptar-lo als usos actuals. Tanmateix, el seu ús per la majoria de la població resta encara molt llunyà.

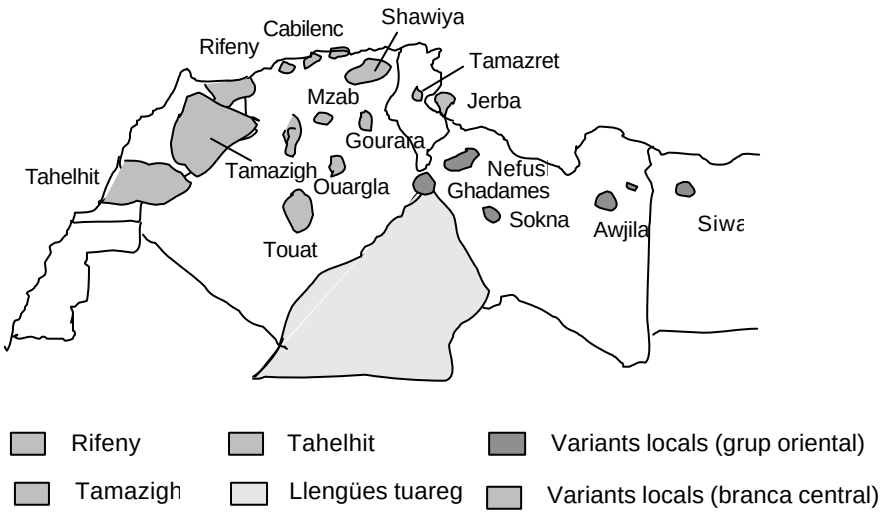


Figura 1.5. Distribució geogràfica de les principals llengües berbers.

A la perifèria sud de l'àrea considerada trobem, a més, dues altres grans famílies lingüístiques: la nilo-sahariana, amb dues subbranques, la *sahariana* i la *songhai*, a Líbia, Chad, Níger i Mali, i la del nígero-kordofanès, que s'estén des del Sahel fins a Sudàfrica. Ambdues famílies s'associen a pobles de pell negra i poden ser considerades intrusions sud-saharianes.

2. Marcadors *clàssics*

2.1 Descripció i principals característiques

Informalment, s'anomenen marcadors genètics *clàssics* aquells sistemes polimòrfics detectats directament en els productes d'expressió gènica, en oposició a aquells en què s'analitza directament la variació en el DNA. Tanmateix, per a la majoria d'aquests sistemes hom ja coneix la variació de la seqüència de DNA subjacent a la variabilitat en el producte d'expressió. Convencionalment, hom divideix aquests polimorfismes en quatre grans blocs:

i) grups sanguinis: són proteïnes de la membrana eritrocitària, sovint amb cadenes glucosídiques, que són capaces d'induir la formació d'anticossos en determinades circumstàncies (en transfusions, en animals de laboratori o en dones multípares). El polimorfisme en els grups sanguinis es troba en el nombre de sucres que conformen la cadena glucosídica (com a conseqüència d'una activitat diferencial dels diversos al·lels que codifiquen per l'enzim responsable d'addicionar aquestes unitats glucosídiques) o bé en la composició de la cadena peptídica. L'anàlisi d'aquests polimorfismes es realitza mitjançant anticossos comercials que aglutinaran els eritròcits segons posseeixin l'antigen corresponent.

ii) proteïnes plasmàtiques: la seva anàlisi es fa mitjançant electroforesi. Sovint, les diferències aminoacídiques es tradueixen en diferents mobilitats electroforètiques que hom pot resoldre bé amb tècniques d'isoelectroenfocament. Posteriorment a l'electroforesi, hom pot revelar la proteïna estudiada mitjançant una tinció inespecífica de proteïnes o a partir de les propietats enzimàtiques o immunològiques específiques de la proteïna.

iii) enzims eritrocitaris: són analitzats també per electroforesi. En el revelat hom empra la reacció enzimàtica que catalitzen per a produir productes acolorits o que es poden tenyir específicament.

iv) sistema d'antígens leucocitaris humans o antígens HLA: juguen un paper importantíssim en el sistema immunitari. La seva detecció es basava, en general, en la disponibilitat d'anticossos específics. Tanmateix, actualment hom els analitza sobre la seqüència de DNA, ja sigui directament o per tècniques indirectes (ARMS o *Amplification Refractory Mutation System* i ASO o *Allele Specific oligonucleotide*, entre d'altres).

2.2 Aplicacions a la genètica

Els anomenats marcadors *clàssics* foren els primers sistemes polimòrfics genètics de què hom disposà per a la descripció i caracterització de la variabilitat genètica de les poblacions humanes. A partir del descobriment dels grups sanguinis a principis de segle, s'iniciaren els estudis de les freqüències dels diferents al·lels d'aquests marcadors en diversos grups humans dels cinc continents. Tanmateix, la seva explotació i interpretació en termes de genètica de poblacions foren posteriors. De fet, no és fins els anys 30 que hom comença a disposar d'una descripció matemàtica per als quatre principals agents de l'evolució: la mutació, la selecció, la deriva genètica i la migració. Primer, fou necessari reconèixer que gran part d'aquests marcadors *clàssics* eren neutres, fet que implica que l'efecte de la selecció natural sobre ells és nul o molt baix i que les freqüències poblacionals d'aquests marcadors deuen haver variat bàsicament per deriva genètica i/o migració. Cal assenyalar que la mutació és un esdeveniment molt rar en aquest tipus de polimorfisme.

La caracterització genètica d'una població es pot fer fàcilment a partir de les freqüències dels al·lels d'un nombre elevat de marcadors *clàssics*. Hom troba la culminació de l'aplicació dels polimorfismes *clàssics* a l'estudi de la variabilitat genètica humana en la magna obra de L. Luca Cavalli-Sforza, Paolo Menozzi i Alberto Piazza (1994). En aquesta obra es compilen les dades existents en una vastíssima bibliografia sobre polimorfismes clàssics a tot el món, s'analitzen i s'interpreten en termes d'història de poblacions. En analitzar diverses poblacions dins una àrea geogràfica, hom pot produir representacions geogràfiques de llurs freqüències al·lèliques. La informació continguda en aquests conjunts de mapes es pot sintetitzar mitjançant anàlisi de components principals (vegeu Materials i Mètodes, apartat 5.2). A partir d'aquí és fàcil observar que una fracció de la variació genètica presenta estructura

geogràfica. Una de les conclusions més importants dins la genètica de poblacions humanes en els darrers anys ha estat la interpretació d'aquesta variació. Aquesta és conseqüència de la història de les pròpies poblacions, especialment, de les grans expansions demogràfiques sovint lligades a canvis tecnològics, com el descobriment de l'agricultura o la domesticació del cavall (Cavalli-Sforza et al. 1993). En definitiva, és el resultat d'una història demogràfica que podrem identificar.

3. Els microsatèl·lits

3.1 Descripció i principals característiques

Els microsatèl·lits, també anomenats STRs o *Short Tandem Repeats*, són seqüències de DNA consistents en la repetició en tàndem d'una unitat bàsica, d'entre dos i sis nucleòtids de longitud. Es troben de forma ubiqua distribuïdes al llarg del genoma humà i poden ser extraordinàriament polimòrfiques entre individus o inclús entre els al·lels d'un mateix individu. A més d'en el genoma humà, s'han trobat també en tots els genomes eucariotes i, encara que de forma molt menys abundant, també en procariotes.

Tot i que s'ha descrit que en determinades bactèries patògenes els microsatèl·lits poden promoure l'aparició de propietats noves que capaciten a aquests organismes per a sobreviure davant canvis de l'entorn potencialment letals, la funció d'aquestes seqüències repetitives en el genoma humà resta encara desconeguda (Moxon i Wills, 1999). De fet, es considera que els microsatèl·lits formen part de l'anomenat DNA *escombraria*, és a dir, sense funció coneguda. Com a conseqüència directa d'aquesta manca de funció coneguda, hom pressuposarà que la variació existent en aquestes seqüències repetitives és selectivament neutra.

El polimorfisme en els microsatèl·lits consisteix principalment en la variació del nombre d'unitats repetides. Aquesta variabilitat es genera mitjançant un patró de mutació peculiar: en cada mutació, el microsatèl·lit guanya o perd una o diverses unitats bàsiques de la seva estructura, sent els guanys o pèrdues d'una sola repetició els més freqüents (Weber i Wong, 1993; Amos i Rubinsztein, 1996). Podrem analitzar les diferències en nombre de repeticions directament a partir de les llargades en parells de bases dels fragments obtinguts en utilitzar una parella de *primers* o encebadors que amplifiquin de manera específica cada microsatèl·lit. Convé que els *primers* emprats siguin propers a la seqüència repetitiva per evitar detectar canvis de llargada o de patró de migració electroforètica que no corresponguessin a variacions

en la seqüència del microsatèl·lit sinó a variacions en les regions flanquejants d'aquest.

Hom pot classificar els microsatèl·lits en base al nombre de nucleòtids de la unitat bàsica o segons la complexitat de la seqüencia repetitiva. Segons el primer criteri, quan el motiu de la repetició consta, respectivament, de dos, tres, quatre, cinc o sis nucleòtids s'anomenaran els respectius microsatèl·lits dnucleòtids, trinucleòtids, tetranucleòtids, pentanucleòtids i hexanucleòtids. Tanmateix, l'anàlisi de la seqüència d'aquestes repeticions en tàndem indica que l'estructura de molts d'aquests *repeats* pot no ser tan simple com hom creia prèviament, en el sentit que, en determinats casos, els al·lels poden diferir en la seva composició de bases així com en la seva longitud. D'aquí sorgí el segon criteri de classificació en el qual es distingeix entre microsatèl·lits perfectes o purs (quan la variació en els al·lels consisteix exclusivament en el nombre de repeticions d'una única unitat bàsica), imperfectes o interromputs (quan es presenten interrupcions en la unitat repetitiva) i complexos o compostos (quan apareixen diferents tipus d'unitats bàsiques).

Tipus	Exemples
Segons el nombre nucleòtids del motiu repetit	
Dinucleòtid	CTTTGACCTA (**CA**)$_n$ GAGGTACAT
Trinucleòtid	CTTTCACCGA (**GAC**)$_n$ TAGCCATT
Tetranucleòtid	CATTG (**TTTC**)$_n$ TACCGGATAACGC
Pentanucleòtid	TCGATATTGCA (**TACCA**)$_n$ GGTCA
Segons la complexitat de la seqüència repetida	
Perfectes o purs	TTGACACCTA (**GATA**)$_n$GGTTCATA
Imperfectes o interromputs	GACC (**GACA**)$_n$ TTTT(**GACA**)$_n$TATA
Complexos o compostos	TGGA (**TCTG**)$_n$ (**TCTA**)$_n$ TCCAACTA

Taula 3.1. Exemples de diferents tipus de microsatèl·lits. Nota: n indica un nombre qualsevol de repeticions.

Cal assenyalar que quan algunes d'aquestes seqüències (totes elles trinucleòtids) muten produint expansions fora del rang habitual d'al·lels (és a dir, en passar de poques desenes fins a centenars o milers de repeticions) són causa directa d'importants malalties genètiques de tipus neurològic i/o neuromuscular (Ashey i Warren, 1995; Richards i Sutherland, 1994). Entre les més conegudes destaquen la

malaltia de Huntington, la distrofia miotònica i l'atàxia de Friedreich. Es tracta de neuropaties rares originades per la disrupció que causa l'expansió d'un trinucleòtid en el gen que el conté o en un de pròxim; en molts casos el trinucleòtid codifica per un aminoàcid i l'expansió dóna lloc a una proteïna amb un tracte molt llarg de repetició de l'aminoàcid. De fet, hom utilitza els propis microsatèl·lits responsables per al diagnòstic d'aquestes malalties neurològiques i per a la detecció de persones amb risc de presentar-les. També s'ha comprovat que els microsatèl·lits canvien de longitud en les fases primerenques de certs càncers, cosa que els converteix en valuosos marcadors per al diagnòstic precoç d'aquests (Rampino et al. 1997; Perucho, 1998).

3.2 Mutació en microsatèl·lits

L'augment o disminució del nombre de repeticions en una o poques unitats de l'estructura bàsica en els microsatèl·lits molt probablement es deu a fenòmens d'*slippage* o lliscament durant la replicació del DNA (Levinson i Gutman, 1987; Weber, 1990; Schlötterer i Tautz, 1992). Aquest patró de mutació ha estat formalitzat matemàticament en el que s'anomena model de mutació *stepwise* generalitzat (Di Rienzo et al. 1994). Tanmateix, no es descarta que els grans salts mutacionals observats en algunes distribucions d'al·lels de microsatèl·lits i especialment en les dramàtiques expansions de triplets en algunes neuropaties puguin ser generats per altres processos mutacionals com una recombinació desigual (Freimer i Slatkin, 1996).

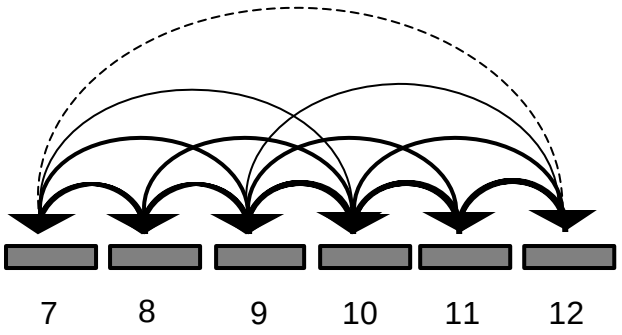


Figura 3.2. Representació del model de mutació *stepwise* generalitzat. Les mutacions poden implicar guany o pèrdua de més d'una unitat de repetició, encara que els més probables són els salts que impliquen una sola repetició.

Durant molt de temps s'ha discutit quin és el valor de la taxa de mutació (μ) que presenten aquests marcadors en humans (Weber i Wong, 1993; Edwards et al. 1992; Cooper et al. 1996; Heyer et al. 1997; Kayser et al. 1997; Bianchi et al. 1998). La pregunta resta encara oberta i tan sols sabem que dependrà de molts factors. Parlar d'una única taxa, per tant, és potser una falàcia. Segurament la taxa de mutació en microsatèl·lits és intrínseca de locus i depèn de factors com el nombre de repeticions (Weber 1990; Goldstein i Clark, 1995; Brinkman et al. 1998), perfecció o imperfecció dels repeats (Estoup, 1995), sexe i edat dels progenitors (Henke i Henke, 1999), i longitud de la unitat repetitiva (Chakraborty et al. 1997) tot i que en alguns d'aquests casos tampoc està clar cap a quina direcció. Inclús hom ha suggerit que la seqüència de les regions flanquejants podrien influir també en la taxa de mutació d'aquestes seqüències. En general, les estimacions filogenètiques i geneològiques donen taxes de l'ordre de 10^{-3}-10^{-4} per locus, gàmeta i generació.

L'homoplàsia (o coexistència d'al·lels que són idèntics en estat sense ser idèntics per descendència) serà una conseqüència directa del model i l'elevada taxa de mutació en aquestes seqüències repetitives. Dos al·lels són idèntics per descendència si descendeixen sense mutació d'un mateix al·lel ancestral. Òbviament, dos al·lels de microsatèl·lits poden presentar la mateixa grandària (ser idèntics en estat), i no ser idèntics per descendència. Tot i derivar d'un mateix al·lel ancestral poden provenir de camins o històries diferents.

3.3 Aplicacions a la genètica

La seva naturalesa altament polimòrfica, el fet que, si més no la gran majoria, són polimorfismes selectivament neutres, juntament amb les característiques de trobar-se àmpliament estesos per tot el genoma humà, mostrar unes bases genètiques de variabilitat conegudes i modelables i poder ser fàcilment tipificats mitjançant tècniques com la PCR, han convertit els microsatèl·lits en importants marcadors genètics per a la localització de gens mitjançant anàlisi de lligament, la medicina forense i la genètica de poblacions.

La relativa velocitat amb què hom pot obtenir resultats per a un gran nombre de microsatèl·lits i el fet de poder disposar d'aproximadament un STR cada 30-50 Kb han facilitat en gran part la seva utilització com a fites per situar la posició relativa de gens

de malalties en estudis de lligament, on, sens dubte, són els marcadors genètics per excel·lència (Edwards et al. 1991; Dib et al. 1996).

Dins el camp de la genètica forense els microsatèl·lits s'empren tant en tests de paternitat com per a la identificació individual (Hammond et al. 1994; Urquhart et al. 1994; Blouin et al. 1996). Cal assenyalar que, en casos d'identificació de criminals o de reconeixement de víctimes en accidents, el material a analitzar sovint es troba en poca quantitat i degradat. Són casos on és extremadament útil l'aplicació de marcadors que puguin ser genotipats conjuntament en una mateixa reacció de PCR, tal com possibiliten els conjunts de microsatèl·lits de diversos *kits* comercials. Mentre existeixen milers de microsatèl·lits mapats al llarg del genoma humà, només una vintena de loci han estat estandarditzats per a la seva aplicació en genètica forense. Els microsatèl·lits tetranucleòtids són els més utilitzats en genètica forense ja que presenten una separació de grandàries més gran entre al·lels, cosa que en facilita la identificació (Kimpton et al. 1993), i són més abundants que els pentanucleòtids i els hexanucleòtids. Entre les característiques més desitjables per als sistemes STRs utilitzats en genètica forense s'inclouen una elevada heterozigositat, una unitat de repetició regular, al·lels distingibles i capacitat d'amplificació robusta mitjançant PCR (Gill et al. 1995).

Per altra banda, són diversos els exemples en la literatura on s'analitza la variació del microsatèl·lits en poblacions humanes per a intentar inferir la història evolutiva dels humans anatòmicament moderns, així com per estudiar les relacions filogenètiques entre poblacions i/o espècies properes (Bowcock et al. 1994, Di Rienzo et al. 1994, Deka et al. 1995, Jorde et al. 1995, Pérez-Lezaun et al. 1997 i Calafell et al. 1998).

Els microsatèl·lits s'empren també en consell genètic, per estudiar la segregació de cromosomes portadors de mutacions rares per malalties mendelianes recessives. Finalment, com hem comentat anteriorment, aquestes seqüències repetitives també són emprades en la detecció precoç d'alguns càncers i en el diagnòstic de determinades neuropatologies rares.

4. El cromosoma Y

4.1 Caracterització

El cromosoma Y, un dels més petits del genoma humà, és una molècula de DNA lineal, d'aproximadament 60 Mb, amb una seqüència encara no del tot coneguda. Citològicament, comprèn una regió d'heterocromatina a la part distal del braç llarg, que és de longitud variable entre els individus, i una part de grandària constant, d'eucromatina, d'unes 30 Mb, que conté les regions de major interès genètic.

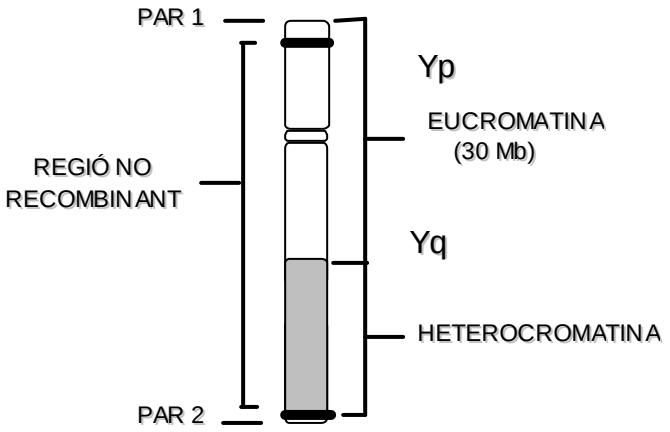


Figura 4.1. Esquema del cromosoma Y humà.

El cromosoma Y conté els gens que determinen la masculinitat, s'hereta uniparentalment, per via paterna, i en la major part de la seva longitud no presenta recombinació. L'excepció es troba a les regions pseudoautosòmiques dels extrems, anomenades PAR1 i PAR2, les quals recombinen amb el cromosoma X. Per tant, la major part del cromosoma (l'anomenada regió no recombinant) es comportarà com un únic bloc o grup de lligament en cada transmissió pare - fill baró. Ambdues característiques, herència paterna i absència de recombinació, el converteixen en una molècula d'especial interès filogenètic, donat que, en retenir el registre

d'esdeveniments mutacionals ocurreguts al llarg del temps, permet resseguir l'evolució dels llinatges masculins en les poblacions.

El seu escàs contingut en gens ha fet que inclús s'arribés a considerar un cromosoma desert. Aproximadament, hi ha descrits uns 26 gens. Tanmateix, destaca la coherència en l'organització d'aquests en dos grans grups segons la seva funció. Els del primer grup, anomenats gens de manteniment (*housekeeping genes*), s'expressen en gran quatitat d'òrgans i tenen homòlegs en el cromosoma X inactivat que escapen de la inactivació. El segon grup, format per famílies de gens del cromosoma Y que s'expressen de manera específica en teixit testicular, i que englobaria els gens determinants del sexe masculí, podria explicar la infertilitat entre els homes amb delecions en el cromosoma Y (Lahn i Page, 1997).

La mida efectiva dels cromosomes Y (nombre de cromosomes Y que passen a la descendència) en una població qualsevol és quatre vegades més petita que la de qualsevol autosoma. La conseqüència directa d'aquesta reducció és que processos com la deriva genètica tindran un impacte molt més fort sobre les regions lligades al cromosoma Y que no pas sobre cap altra part del nostre genoma, exceptuant el DNA mitocondrial (Pérez-Lezaun et al. 1997). Aquest efecte, a vegades, permetrà de discernir colls d'ampolla no aparents amb d'altres marcadors genòmics i poder estudiar la diferenciació genètica entre poblacions properes geogràficament o que hagin divergit en temps recents (de Knijff et al. 1997).

Cal destacar, també, que al llarg de la història, la dinàmica de la població masculina sovint pot haver estat diferent a la de la població femenina. Activitats específiques en quant al sexe, tals com la guerra, la caça o la mateixa poliginia, en determinats casos, poden haver reduït el nombre efectiu de cromosomes Y a les poblacions i, per tant, incrementar encara més la seva predisposició a la deriva genètica (Seielstad et al. 1998).

Per altra banda, donada l'absència de recombinació en la major part de la longitud del cromosoma, qualsevol mutació selectivament avantatjosa en els caràcters lligats al cromosoma Y podria haver conduït a una selecció simultània, per efecte *hitch-hiking*, de tots els altres al·lels presents en l'haplotip seleccionat. Tot i que aquesta situació podria arribar a comportar una dràstica reducció en els nivells de diversitat genètica del cromosoma Y, només podrem detectar les seves conseqüències en determinats casos, quan l'escombratge selectiu (*selective sweep*) hagués estat suficientment important, i sempre que s'hagués donat en temps relativament recents ja

que, en cas contrari, hom espera que la pròpia mutació hauria tingut temps de regenerar els nivells de polimorfisme en els sistemes més polimòrfics, com els microsatèl·lits.

4.2 Tipus de polimorfismes

Històricament, s'ha donat un retard important en la detecció de variació en el cromosoma Y. La dificultat tecnològica que suposa treballar amb grans quantitats de DNA repetitiu, que dificulta enormement les tecnologies moleculars, i el fet de presentar pocs gens que poguessin cridar l'atenció als genètics clínics són factors que endarreriren, sens dubte, la detecció de polimorfismes en aquest cromosoma. Tanmateix, la recerca sistemàtica de polimorfismes, mitjançant la tecnologia convencional i, sobretot, amb l'aplicació del DHPLC o *Denaturing High Performance Liquid Chromatograph*y, està generant gran quantitat de nous marcadors en els darrers temps.

Els grans tipus de polimorfismes que es troben en el cromosoma Y inclouen: substitucions de base, insercions, duplicacions o delecions, rearranjaments complexos, microsatèl·lits, minisatèl·lits i DNA satèl·lit (vegeu exemples a taula 4.2). Mentre la gran majoria són fàcilment tipificables mitjançant tècniques com la PCR (*Polymerase Chain Reaction*) seguida de l'anàlisi d'RFLPs (*Restriction Fragment Length Polymorphisms*) o de l'aplicació del DHPLC, hom també troba alguns marcadors de major dificultat de detecció mitjançant hibridació de sondes o amb electroforesi de camp pulsant.

En un determinat punt del genoma, esdeveniments mutacionals com les substitucions de base, les insercions, i les duplicacions o delecions són extremadament rars i, probablement només s'han donat una vegada en l'evolució. Per la seva pròpia naturalesa, aquestes mutacions generen polimorfismes bial·lèlics, és a dir, amb dos estats al·lèlics. En absència de recurrència, el tipatge en primats no humans d'aquests polimorfismes permet de determinar quin al·lel és ancestral i quin és derivat.

Per altra banda, en contraposició als marcadors bial·lèlics amb una taxa d'evolució lenta, hom troba que el cromosoma Y també presenta polimorfismes amb elevades taxes d'evolució. Entre els més utilitzats en genètica de poblacions,

destacarien els microsatèl·lits específics del cromosoma Y i el minisatèl·lit MSY1. Com veurem, gràcies a aquesta gran varietat de taxes de mutació en els seus polimorfismes, el cromosoma Y ha estat emprat per estudiar fenòmens evolutius a diferents escales temporals.

Tipus	Exemples	Mètode de detecció més usual
Substitucions de base	SRY-2627, M9	PCR, RFLPs, DHPLC
Insercions	YAP	PCR
Duplicacions o delecions	12f2	Hibridació en filtre, PCR
Rearranjaments complexos	49 a/f	Hibridació en filtre
Microsatèl·lits	DYS19, YCA1	PCR
Minisatèl·lits	MSY1	MVR-PCR
DNA satèl·lit	Yα1	PFGE, hibridació en filtre

Taula 4.2. Exemples de tipus de polimorfismes en el cromosoma Y. Notes: MVR-PCR, *minisatellite variant repea*t PCR; PFGE, *pulsed-field gel electrophoresis*.

4.3 Estructura de la diversitat genètica: haplogrups i haplotips

Gràcies a l'absència de recombinació i donat que només hi ha una sola dotació per individu, la combinació dels estats al·lèlics de diferents polimorfismes en el cromosoma Y permet de reconstruir directament el seu haplotip (genotip haploïde definit per diversos polimorfismes lligats entre ells).

Sovint, hom empra el mot *haplogrup* per a designar el conjunt de cromosomes que presenta un mateix haplotip de marcadors bial·lèlics. Donada la naturalesa única (o extremadament rara) de la mutació en els polimorfismes que els caracteritzen, hom espera que els haplogrups siguin grups de cromosomes relacionats per descendència, és a dir, grups de cromosomes que deriven d'un avantpassat comú.

Per contra, els haplotips definits per marcadors amb una major taxa de mutació, com els microsatèl·lits, poden ser idèntics per descendència o per estat. La natura recurrent de la mutació en els microsatèl·lits permet que un mateix haplotip es generi més d'una vegada. Com hem esmentat en parlar de microsatèl·lits en general, aquest fenòmen rep el nom d'homoplàsia. Tanmateix, com més gran sigui el nombre

de microsatèl·lits considerats, menys probable serà que dos haplotips idèntics per estat no siguin idèntics per descendència. L'elevada taxa de mutació en microsatèl·lits comporta també que els haplotips de microsatèl·lits puguin ser genèticament molt diversos, inclús en cromosomes filogenèticament propers. Aquesta propietat els fa idonis per a ser emprats en identificació individual o en l'estudi de poblacions molt properes. Per contra, l'estabilitat dels haplogrups permetrà d'utilitzar-los per investigar fenòmens poblacionals del passat remot o basats exclusivament en deriva i migració.

4.4 Aplicacions a la genètica

Són remarcables les possibilitats que aporta el cromosoma Y dins la genètica forense. La disponibilitat de diferents marcadors altament polimòrfics i tipificables per PCR, sobretot els de tipus microsatèl·lit, ha permès en gran part aquesta aplicació en el camp forense, on sovint les quantitats i qualitat del DNA de les mostres a identificar no són les òptimes. Tanmateix, cal recordar les limitacions dels resultats obtinguts a partir del cromosoma Y com a prova suficient d'inclusió en els casos de genètica forense. Donada la possibilitat que cada haplotip concret de polimorfismes del cromosoma Y sigui compartit entre tots els descendents barons d'una mateixa família, sovint hom troba que per arribar a obtenir un perfil genètic suficientment informatiu, cal combinar els resultats dels polimorfismes del cromosoma Y amb els dels polimorfismes autosòmics. Per altra banda, l'exclusivitat masculina del cromosoma Y resulta especialment útil en casos de violació, on permet l'estudi directe en barreges de teixit femení i de l'agressor masculí.

La variació en el cromosoma Y és d'especial interès també en estudis on es busca si certs *backgrounds* genètics en el cromosoma Y confereixen susceptibilitat a la infertilitat masculina. Aquests seria el cas, per exemple, si certs polimorfismes bial·lèlics predisposessin a l'aparició de grans delecions en el cromosoma. Com que la regió no recombinant del cromosoma Y forma un únic haplotip sense recombinació, hom espera que sigui fàcil identificar-hi *backgrounds* de susceptibilitat. També s'ha aplicat a estudis d'associació en càncers testiculars. Si algun gen del cromosoma Y està implicat en la etiologia d'aquests càncers, la manca de recombinació en el cromosoma Y pot donar associacions significatives amb polimorfismes al llarg del cromosoma.

Com ja hem comentat anteriorment, l'absència de recombinació i l'herència exclusivament paterna, converteixen a la regió no recombinant del cromosoma Y en una important eina filogenètica per a traçar i comparar els llinatges paterns de les poblacions humanes, de manera similar a com el DNA mitocondrial s'empra per a l'estudi dels llinatges femenins. Tanmateix, la presència en el cromosoma Y de diferents tipus de polimorfismes amb diferents mecanismes i taxes de mutació permeten d'augmentar encara més el seu potencial en els estudis d'evolució humana, ja que aquest podrà ser informatiu a diferents escales temporals i/o geogràfiques. Tots aquests factors han provocat que, subseqüentment a la descripció de polimorfismes en el cromosoma Y, els estudis sobre evolució humana basats en aquest cromosoma, tant a nivell d'espècie (Dorit et al. 1995; Whitfield et al. 1995) com a nivell poblacional, s'hagin incrementat notablement en els darrers anys (Hammer 1995; Jobling and Tyler-Smith 1995; Cooper et al. 1996; Deka et al. 1996; de Knijff et al. 1997; Hammer et al. 1998).

Alguns dels nombrosos casos en els que el cromosoma Y ha estat decisiu dins aquests àmbits de la genètica inclouen des de l'origen i dispersió de l'home anatòmicament modern (Hammer et al. 1997, 1998) fins a l'estudi de certs llinatges masculins particulars, com els dels jueus amb el cognom Cohen (Thomas et al. 1998) o el de Thomas Jefferson, tercer president dels EEUU, a qui s'atribueix la paternitat del fill d'una de les seves cambreres (Foster et al. 1998). Destaca també, especialment, el paper protagonista que ha tingut per a resoldre qüestions sobre el poblament de determinades regions geogràfiques i l'origen de la contribució masculina en la composició d'algunes poblacions actuals. Entre els exemples més estudiats, en aquest cas, hi trobem el poblament d'Amèrica (Pena 1995; Underhill et al. 1996, Karafet et al. 1997; Santos et al. 1995, 1996, 1999), l'origen de la població japonesa (Hammer and Horai 1995), la contribució masculina asiàtica a les poblacions nordeuropees (Zerjal et al. 1997), la colonització de Polinèsia i posterior contribució europea (Hurles et al. 1998), i la colonització dels hàbitats muntanyosos a l'Àsia central (Pérez-Lezaun et al. 1999).

Dins aquesta línia, en la present tesi s'intentarà, mitjançant l'ànalisi de diferents marcadors en el cromosoma Y, la investigació de la composició actual i origen dels llinatges masculins de les poblacions del nord-oest d'Àfrica, així com l'estudi de la possible influència masculina que deixaren les migracions que tingueren lloc, en el

passat, entre aquesta regió i la Península Ibèrica; i l'estructura de la variabilitat genètica dels microsatèl·lits en diferents rerefons definits per polimorfismes estables.

OBJECTIUS

Els objectius de la present tesi doctoral han estat:

- Recopilar i interpretar totes les dades disponibles sobre polimorfismes *clàssics* en el nord d'Àfrica per tal d'establir un marc descriptiu i interpretatiu preliminar de la genètica de les poblacions d'aquesta regió.

- Analitzar tota la variabilitat genètica coneguda en el cromosoma Y incloent-hi marcadors de diferents taxa d'evolució com microsatèl·lits i polimorfismes que poden representar esdeveniments evolutius únics, mitjançant tècniques moleculars emergents com el DHPLC.

- Genotipar microsatèl·lits autosòmics i combinar-ne la informació amb l'obtinguda a partir del cromosoma Y per tal de verificar diferents hipòtesis sobre la història de les poblacions del nord d'Àfrica.

- Verificar diferents hipòtesis sobre el primer poblament del nord-oest d'Àfrica; en concret, intentar establir el pes relatiu del substrat paleolític i de l'expansió neolítica.

- Reconèixer l'abast de la contribució àrab al nord-oest d'Àfrica i verificar si es correspon a l'abast de la influència lingüística, cultural i religiosa.

- Mesurar la contribució dels pobles sud-saharians a la composició genètica actual de les poblacions nordafricanes.

- Intentar establir l'abast de la contribució genètica nordafricana a les poblacions de la Península Ibèrica, ja sigui vinculada a l'expansió de l'islam o a altres moviments poblacionals.

- Establir una base de dades de referència per a l'ús de microsatèl·lits del cromosoma Y aplicats a la genètica forense en poblacions del nord-oest d'Àfrica.

- Estudiar l'estructura de la variabilitat de microsatèl·lits específics del cromosoma Y en el rerafons genètic definit per marcadors d'evolució molt més lenta. En particular, entendre l'estratificació genètica dels marcadors del cromosoma Y, especialment l'estructuració per llinatges, en relació a la que diferencien les poblacions.

MATERIALS I MÈTODES

1. Poblacions estudiades

L'obtenció de mostres procedents del nord d'Àfrica no fou senzilla. Implicà una intensa tasca d'establir acords de col·laboració amb diferents universitats, així com amb diversos hospitals i ambulatoris, de Barcelona i rodalia, Andalusia, França i Marroc. Els fruits d'aquestes col·laboracions s'obtingueren de forma gradual al llarg dels quatre anys d'aquest treball.

Podem distingir quatre fonts diferents de mostres:

i) a partir de col·laboracions amb d'altres laboratoris que disposaven de mostres de DNA per a aquestes poblacions. És el cas de part de la mostra d'àrabs marroquins (obtinguda pel professor Abdelaziz Sefiani de la Universitat de Rabat), de la mostra de mozabites (obtinguda pels professors Jean-Michel Dugoujon, Ghania Hariti i Anne Cambon-Thomsen de la Facultat de Medicina de Toulouse) i de la mostra de *tahelhits* (obtinguda pels professors Omar Akhayat, Hassan Izaabel i Zahra Brakez de la Universitat Ibnou Zohr d'Agadir).

ii) a partir d'immigrants nordafricans residents a Catalunya i Andalusia, gràcies a la col·laboració amb el doctor Enric Bufill de l'Hospital General de Vic; amb els doctors Felip Pi, Pere Simonet, Enric Padrés i doctora Prats de Hospital i de l'ambulatori Nou de Viladecans; amb el doctors Xavier Balanzó, Jordi Colomer i Alba Bosch del Consorci de Mataró i amb el doctor Josep Lluís Fernández Roure del dispensari Cirera-Molins de Mataró; i de la professora Elisabeth Pintado de la Facultat de Medicina de Sevilla.

iii) a partir d'una campanya de recollida de mostres sanguínies a València on hi passaven l'estiu un grup de refugiats sahrauís, la qual fou efectuada per membres del nostre laboratori gràcies a l'ajuda d'Oriol Vall, Khadietu, Omar Mansur, Baha Mustafa i Adda Brahim, entre altres.

iv) a partir de dues expedicions al Marroc realitzades per membres del nostre laboratori l'abril i el juny de 1997. En aquest cas, gràcies a la col·laboració amb la doctora Noufissa Benchemsi del Centre Nacional de Transfusió Sanguínia del Marroc, hom passà a disposar d'un ampli espectre geogràfic de mostres sanguínies que inclouen diversos grups poblacionals àrabs i berbers.

Cal assenyalar que l'obtenció de mostres a partir de ii), iii) i iv) va permetre desenvolupar tècniques d'immortalització cel·lular per tal de poder disposar d'una font indefinida del DNA d'aquestes poblacions. En tots els casos, hom ha recollit el consentiment informat dels individus participants i informació diversa referent a la procedència geogràfica, llengua materna i sexe de cada donant, així com dels seu pares i quatre avis.

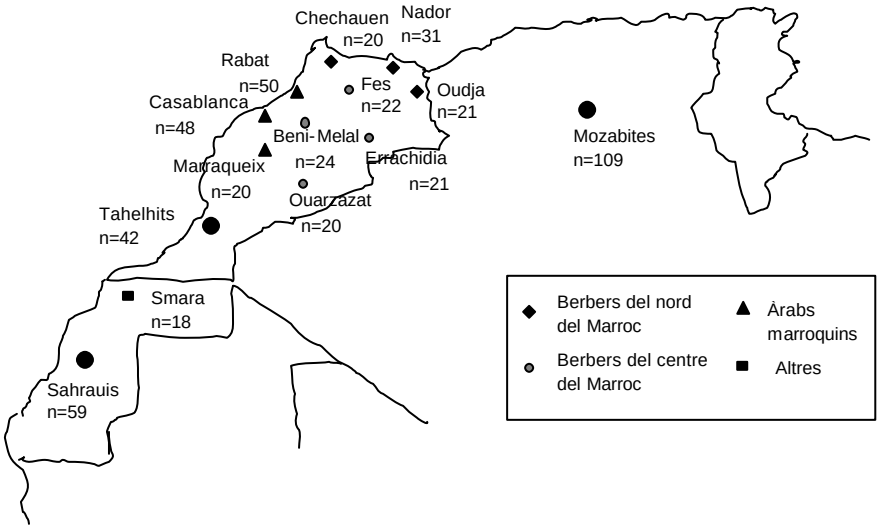


Figura 1. Localització geogràfica i grandària mostral del conjunt de mostres recollides. Les mostres indicades amb un cercle gran i negre constitueixen poblacions analitzades com a unitats independents. En canvi, les mostres indicades amb altres símbols han estat reunides per formar les agrupacions poblacionals esmentades en la llegenda de la figura.

El nombre de cromosomes analitzats per cada agrupació poblacional nordafricana considerada en els treballs corresponents als capítols II-V s'indiquen a la taula 1. Cada agrupació poblacional s'ha definit en funció de la procedència geogràfica de les mostres i segons criteris culturals o lingüístics.

El conjunt de mostres de la Península Ibèrica analitzades en l'últim capítol d'aquest treball es trobava disponible en forma de DNA en el nostre laboratori. Comprèn 37 andalusos, 44 bascos i 16 catalans.

Marcadors genètics estudiats	Àrabs Marroc	Berbers NC Marroc	Berbers S Marroc	Mozabites	Sahrauís
Capítol II: 21 STRs autosòmics	94-160	50-126	84-96	88	104-118
Capítol III: 8 STRs específics del cromosoma Y	44	--	42	68	29
Capítol IV: 8 STRs i 11 marcadors bial·lèlics cromosoma Y (set 1)	44	13	44	--	29
Capítol V: 41 marcadors bial·lèlics cromosoma Y (set 2 o DHPLC)	44	63	40	--	29

Taula 1. Nombre de cromosomes analitzats en cada agrupació poblacional nordafricana considerada en els diferents estudis.

En total, hom ha diferenciat fins a vuit poblacions:

i) Àrabs marroquins. Inclou individus que s'han autodefinit com a àrabs i que procedeixen de les grans ciutats de la costa marroquina, com Rabat i Casablanca, així com dels seus voltants.

ii) Berbers del nord i centre del Marroc. Inclou individus que s'han autodefinit com a berbers i que procedeixen de pobles i ciutats del nord i del centre del Marroc com Nador, Oudja, Chechauen, Taza, Fez, Beni-Melal, Errachidia i Ouarzazat; o bé individus que s'han autoidentificat com a rifenys o com a *tamazights*, és a dir, com a parlants de les llengues berbers predominants, respectivament, al nord i centre del Marroc.

iii) Berbers del sud del Marroc o *tahelhits*. Inclou individus que s'han autodefinit com a berbers i que procedeixen de la regió del riu Souss o bé individus berbers que s'han autoidentificat directament com a *tahelhits*, és a dir, parlants de la llengua berber predominant al sud del Marroc.

iv) Mozabites. Inclou individus berbers procedents de la ciutat de Ghardaia a Algèria, els quals s'autoidentificaren com a membres d'aquest grup berber culturalment ben definit.

v) Sahrauís. Inclou individus sahrauis procedents del Sàhara Occidental però actualment localitzats en diversos camps de refugiats al sud d'Algèria.

vi) Bascos. Inclou individus autòctons de la província de Guipúscoa, al País Basc.

vii) Catalans. Inclou individus de les comarques de l'Alt Empordà, el Baix Empordà, el Gironès, el Pla de l'Estany, la Selva i la Garrotxa.

viii) Andalusos. Inclou individus d'arreu d'Andalusia.

2. Obtenció, purificació i quantificació de DNA

El DNA de les poblacions s'ha obtingut a partir de mostres sanguínies extretes amb *vacutainers* de 5 ml que contenien citrat sòdic. En tots els casos, es realitzà una extracció de DNA mitjançant un mètode estàndard d'extracció amb fenol i cloroform (Sambrook et al. 1989) a partir de la lisi de limfòcits i posterior digestió cèl·lular amb proteinasa K en presència d'EDTA i un detergent, i en el qual s'inclouen una sèrie d'extraccions amb fenol i cloroform, a fi i efecte d'eliminar les proteïnes de la mostra. Seguidament, es precitava el DNA extret amb etanol.

La determinació de la quantitat de DNA obtinguda per a cada mostra es feu mesurant l'absorbància en un espectrofotòmetre a 260 nm (una unitat d'absorbància a 260 nm equival a 50 ng/µl de DNA de doble cadena). Es calculà també la relació d'absorbàncies a 260 nm i 280 nm ($D.O._{260}/D.O._{280}$) per a determinar la relació d'àcids nucleics i proteïnes de cada mostra (1.8 és el valor òptim que hom espera, mentre que un valor inferior a 1.8 indica una contaminació per excés de proteïnes o restes de fenol). Posteriorment, les mostres es prepararen a una solució de treball de 100 ng/µl.

3. Anàlisi de microsatèl·lits

Amb l'excepció de dos trinucleòtids del cromosoma Y, tots els microsatèl·lits analitzats en la present tesi són tetranucleòtids. Aquesta elecció respon bàsicament a raons pràctiques, donada llur abundància i ubiqüitat en el genoma i que, com s'ha comentat a la introducció, l'assignació dels al·lels en les repeticions de quatre nucleòtids mostra menys problemes que la d'altres microsatèl·lits amb longituds de repetició menor.

Cal assenyalar també, que certs problemes com la presència de bandes paràsites (Weber 1990) i l'addició o subtracció d'algunes repeticions per la Taq polimerasa (Smeets et al. 1989) durant l'amplificació d'aquestes seqüències en el procés de PCR, són fenòmens més freqüents en dinucleòtids que en tetranucleòtids, cosa que els confereix un avantatge pràctic indiscutible.

Per altra banda, l'elecció de quins *loci* concrets estudiar vingué condicionada en gran part a la gran disponibilitat d'informació que hom troba en la literatura per a aquests marcadors en d'altres poblacions que poguéssim utilitzar de referència en el nostre treball.

3.1 Caracterització dels microsatèl·lits autosòmics estudiats

El *locus*, motiu repetit, localització cromosòmica, localització gènica, nombre d'al·lels i rang de repeticions descrites així com la seqüència dels *primers* i/o *kit* comercial emprats per al seu tipatge, juntament amb la longitud dels productes PCR obtinguts en parells de bases per al conjunt de 21 microsatèl·lits autosòmics estudiats es presenten a les taules 3.1.1 i 3.1.2.

Locus	Motiu repetit	Localització cromosòmica	Nombre d'al·lels	Rang de repeticions	*Kit* comercial emprat	Longitud producte PCR (pb)
D3S1358	$(TCTA)_n$	3p	13	9-20	Profiler Cofiler	114-142
FGA	$(CTTT)_n$	4q28	21	16-29	Profiler	219-267
D8S1179	$(TCTA/G)_n$	8	10	8-17	Profiler	128-168
D21S11	$(TCTG/TA)_n$	21	25	24.2-38	Profiler	189-243
D13S317	$(GATA)_n$	13q22-31	11	5-15	Profiler	206-234
D16S539	$(AGAT)_n$	16q24-qter	11	5-15	Cofiler	234-274
D18S51	$(AGAA)_n$	18q21.3	18	9-23	Profiler	273-341
CSF1PO	$(AGAT)_n$	5q33.3-34	11	6-15	Cofiler Green I	281-317

Taula 3.1.1. Característiques principals dels microsatèl·lits autosòmics estudiats (I). Nota: per a la majoria d'aquests loci la localització gènica és desconeguda. Excepcions: FGA, Alpha fibrinogen (intró 3, nucleòtid 2912); CSF1PO, proto-oncogen c-fms pel receptor CSF-1.

Locus	Motiu repetit	Localització cromosòmica	Localització gènica	Nombre d'al·lels	Rang de repeticions	Seqüència dels *primers* (5'-3')	Longitud producte PCR (pb)[1]	*Kit* comercial emprat	Longitud producte PCR (pb)[2]
D5S818	$(AGAT)_n$	5q 21–31	Desconeguda	11	7-16	GGGTGATTTTCCTCTTTGGT TGATTCCAATCATAGCCACA	133-169	Profiler	135-171
D7S820	$(GATA)_n$	7q	Desconeguda	11	6-15	TGTCATAGTTTAGAACGAACTAACG CTGAGGTATCAAAAACTCAGAGG	198-234	Profiler Cofiler	258-294
D11S2010	$(GATA)_n$	11	Desconeguda	7	9-15	TTTTCAGGCTTTATCTCATTCA GGGACATATGAGGGCTCTCT	107-131	--	--
D13S767	$(GATA)_n$	13	Desconeguda	8	8-15	AGTGTTTCTAATGTAGGTTGATGC TTTCTGTGCCATGAGCAGTA	155-179	--	--
D14S306	$(GATA)_n$	14q12–q13	Desconeguda	9	9-17	TGACAAAGAAACTAAAATGTCCC AAAGCTACATCCAAATTAGGTAGG	186-218	--	--
D18S848	$(GATA)_n$	18	Desconeguda	9	5-13	TTGGTACATATGATACATTGGATG GAATTTTGCGAACAAACTGG	80-112	--	--
D2S1328	$(GATA)_n$	2	Desconeguda	12	5-12	GTGGCTTTGGAGGAACACTA TGGCACATGTACACCAGAAC	135-171	--	--
D4S243	$(AGAT)_n$	4	Desconeguda	11	8-19	TCAGTCTCTCTTTCTCCTTGCA TAGGAGCCTGTGGTCCTGTT	165-205	--	--
F13A1	$(AAAG)_n$	6p24.2–p23	Factor coagulació XIII (248pb, intró A)	12	3.2-17	GAGGTTGCACTCCAGCCTTT ATGCCATGCAGATTAGAAA	180-244	--	--
FES / FPS	$(ATTT)_n$	15q25-qter	fes/fps proto-oncogen (4713pb, intró V)	7	8-14	GGAAGATGGAGTGGCTGTTA CTCCAGCCTGGCGAAAGAAT	143-167	--	--
TH/THO1	$(CATT)_n$	11p15.5	Tyrosina hidroxilasa (1170pb, intró 1)	9	5-11	GTGGGCTGAAAAGCTCCCGATTAT ATTCAAAGGGTATCTGGGCTCTGG	183-207	Cofiler Green I	169-189
TPO	$(AATG)_n$	2p25–p24	Peroxidasa tiroidea (intró 10)	8	6-13	CACTAGCACCCAGAACCGTC CCTTGTCAGCGTTTATTTGCC	106-134	Cofiler Green I	218-242
VWF/VWA	$(TCTA)_n$	12p13.3–p13.2	Factor Von Willebrand (intró 40)	12	11-22	CCCTAGTGGATGATAAGAATAATC GGACAGATGATAAATACATAGGATGGATGG	126-162	Profiler	157-197
D9S926	$(GATA)_n$	9	Desconeguda	8	7-16	TCCTCAGCCTACAATTCCTG GACTGAAGCACAGCTAAGCC	198-228	--	--

Taula 3.1.2. Característiques principals dels microsatèl·lits autosòmics estudiats (II). Notes: Alguns dels loci foren amplificats o amb els primers indicats o amb els continguts en els *kits* comercials. (1) Longitud producte PCR obtingut mitjançant els *primers* indicats; (2) Longitud producte PCR obtingut a partir del corresponent *kit* comercial.

3.2 Caracterització dels microsatèl·lits específics del cromosoma Y estudiats

El *locus*, motiu repetit, nombre d'al·lels i rang de repeticions descrites, seqüència dels *primers* emprats per al seu tipatge, així com la longitud en parells de bases dels productes de PCR obtinguts per als vuit microsatèl·lits específics del cromosoma Y estudiats es presenten a la taula 3.2.

Locus	Motiu repetit	Nombre d'al·lels	Rang de repeticions	Seqüència dels *primers* (5'-3')	Longitud producte PCR (pb)
DYS19	$(CTAT/C)_n$	10	10-19	CTACTGAGTTTCTGTTTATAGT ATGGCATGTAGTGAGGACA	174-210
DYS388	$(ATT)_n$	7	11-17	GTGAGTTAGCCGTTTAGCGA CAGATCGCAACCACTGCG	125-143
DYS389I	$(TCTG/TA)_n$	7	7-13	CCAACTCTCATCTGTATTATCTATG TCTTATCTCCACCCACCAGA	239-263
DYS389II	$(TCTG/TA)_n$	9	23-31	CCAACTCTCATCTGTATTATCTATG TCTTATCTCCACCCACCAGA	353-385
DYS390	$(CTG/AT)_n$	10	18-27	TATATTTTACACATTTTTGGGCC TGACAGTAAAATGAACACATTGC	191-227
DYS391	$(CTAT)_n$	6	8-13	CTATTCATTCAATCATACACCCA GGATTCTTTGTGGTTGGGTCTG	275-295
DYS392	$(ATT)_n$	8	7-16	TCATTAATCTAGCTTTTAAAAACAA AGACCCAGTTGATGCAATGT	236-263
DYS393	$(GATA)_n$	6	9-15	CCAACTCTCATCTGTATTATCTATG TCTTATCTCCACCCACCAGA	108-132

Taula 3.2. Característiques principals dels microsatèl·lits específics del cromosoma Y estudiats.

3.3 Tipatge

El conjunt de microsatèl·lits autosòmics inclosos en els *kits* comercials (PE Applied Biosystems) AmpF*l*STR Profiler Plus, AmpF*l*STR Green I, i AmpF*l*STR Cofiler foren analitzats mitjançant una amplificació conjunta i determinació en multiplex seguint les recomanacions dels propis fabricants. Alguns microsatèl·lits autosòmics inclosos en els *kits* comercials, els microsatèl·lits autosòmics restants no inclosos i tot el conjunt de microsatèl·lits específics del cromosoma Y foren amplificats en reaccions

individuals utilitzant *primers* marcats amb fluorocroms que donen fluorescència blava, verda o groga.

Les condicions de la reacció d'amplificació i de les condicions de PCR en el termociclador (Perkin Elmer 9600) s'indiquen a la taula 3.3. Els productes amplificats es van fer córrer en un sequenciador automàtic ABI 377™ (PE Applied Biosystems). La combinació de les diferents grandàries dels fragments amplificats i la utilització de diferents fluorescències van permetre fer córrer més d'un marcador en cada carril. Hom emprà els estàndards ABI GS500 Rox o ABI GS500 Tamra (Perkin Elmer) com a marcadors interns de carril. Les imatges i resultats de cada gel foren recollides pel software ABI Collection™ (PE Applied Biosystems). Posteriorment, empràrem els paquets de software GeneScan 672™ i Genotyper 1.1™ i 2.1x3™ per a analitzar la grandària dels al·lels en parells de bases i realitzar l'assignació d'al·lels segons el nombre de repeticions a partir de la comparació amb escales seqüenciades específiques de *locus*.

Marcadors utilitzats	Condicions reacció PCR	Condicions termociclador (PE 9600)
AmpF/STR Profiler Plus, AmpF/STR Green I, AmpF/STR Cofiler	25 µl totals amb 0.5 ng DNA, 10.5 µl AmpF/STR PCR *Reaction Mix*, 0.5 µl Ampli*Taq* Gold DNA *polymerase*, 5.5 µl AmpF/STR Profiler Plus/ Green I/ Cofiler Primer Set	95 C 11 min; 28 cicles a 94 C 1 min, 59 C 1 min, i 72 C 1 min; 60 C 45 min
Altres microsatèl·lits autosòmics	10 µl totals amb 100 ng DNA, 50 mM KCl, 10 mM tris-HCl (pH 8.3), 1.5 mM $MgCl_2$ (2.5 mM per F13A1), 250 µM dNTPs, 0.2 µM de cada *primer* i 1 U *Taq* DNA polimerasa	94 C 1 min; 14 cicles a 94 C 20 s, 63 C 1 min (-0.5 C/cicle), i 72 C 1 min; 20 cicles a 94 C 20 s, 56 C a 45 s, i 72 C 1 min; 72 C 5 min.
DYS390, DYS392	10 µl totals amb 100 ng DNA, 50 mM KCl, 10 mM tris-HCl (pH 8.3), 1.5 mM $MgCl_2$, 250 µM dNTPs, 0.2 µM de cada *primer* i 1 U *Taq* DNA polimerasa	94 C 2 min; 8 cicles a 94 C 20 s, 58 C 30 s (-0.5 C/cicle), i 72 C 30 s; 27 cicles a 94 C 20 s, 54 C a 30 s, i 72 C 30 min; 72 C 10 min.
DYS19, DYS388, DYS389I, DYS389II, DYS391, DYS393	10 µl totals amb 100 ng DNA, 50 mM KCl, 10 mM tris-HCl (pH 8.3), 1.5 mM $MgCl_2$ (2.5 mM per DYS19), 250 µM dNTPs, 0.2 µM de cada *primer* i 1 U *Taq* DNA polimerasa	94 C 1 min; 14 cicles a 94 C 20 s, 63 C 1 min (-0.5 C/cicle), i 72 C 1 min; 20 cicles a 94 C 20 s, 56 C a 45 s, i 72 C 1 min; 72 C 5 min.

Taula 3.3. Sinopsi de les condicions d'amplificació emprades.

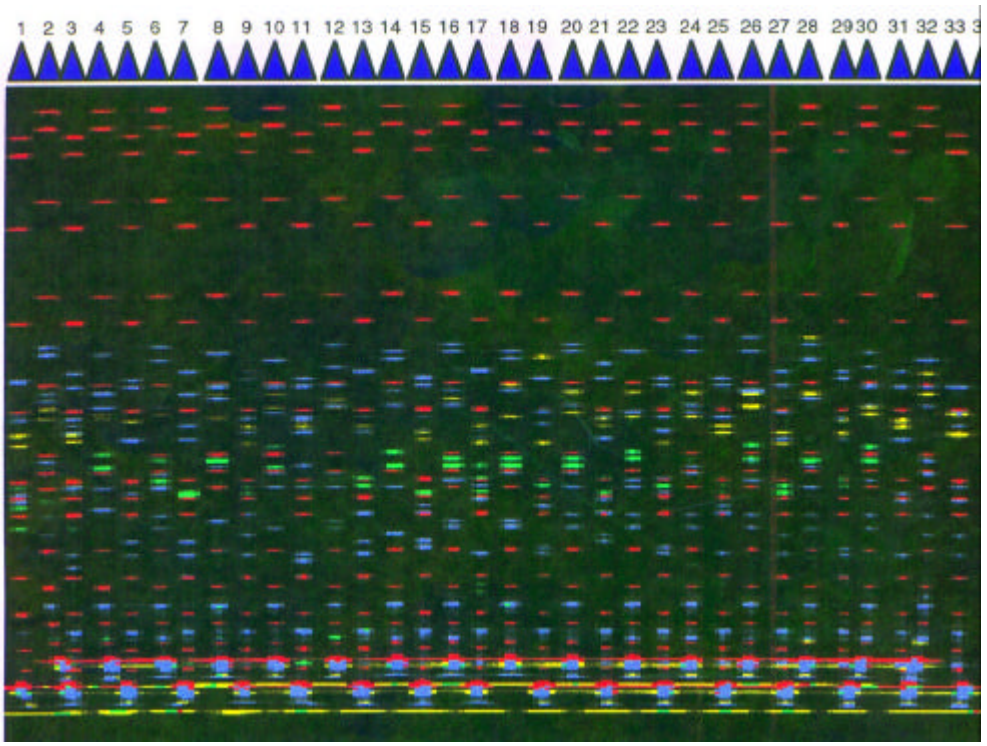


Figura 3.3.1 Imatge d'una electroforesi realitzada mitjançant un seqüenciador automàtic.

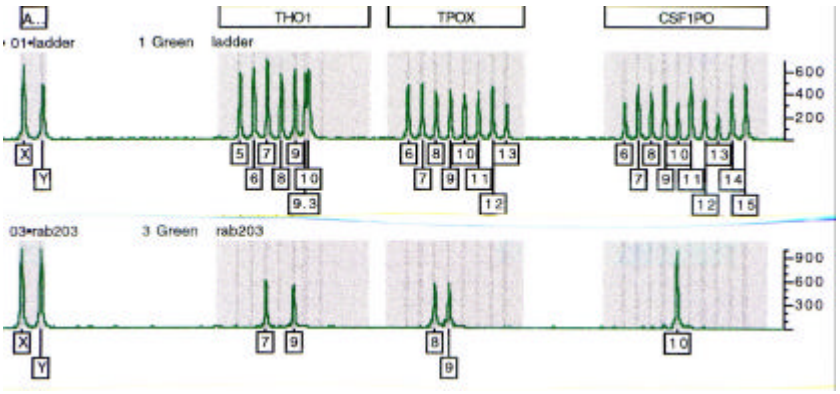


Figura 3.3.2 Electroferograma dels loci amelogenina (marcador de sexe), THO1, TPOX i CSF1PO per a l'individu rab63 (carril 3) amb les respectives escales al·lèliques (carril 1).

4. Anàlisi de polimorfismes bial·lèlics

Mentre la gran majoria de polimorfismes bial·lèlics tipificats en el present treball poden ser fàcilment analitzats per tècniques com la PCR seguida de l'anàlisi d'RFLPs o de l'aplicació de la metodologia del DHPLC, hom també ha utilitzat alguns marcadors que requereixen ser detectats mitjançant hibridació per sondes específiques.

4.1 Polimorfismes analitzats per RFLP o *Restriction Fragment Length Polymorphism*

Els polimorfismes bial·lèlics analitzats mitjançant RFLP es presenten a la taula 4.1.1. Després de la seva corresponent amplificació mitjançant PCR, foren detectats a partir del canvi que generen en la diana d'un enzim de restricció (vegeu condicions a la taula 4.1.2). L'excepció es troba en l'element YAP, on la presència de la inserció Alu pot analitzar-se directament en agarosa després de la seva amplificació, donada la diferència de 305 pb entre l'al·lel amb la inserció Alu i l'al·lel sense inserció. Posteriorment a la seva separació mitjançant una electroforesi en gel d'agarosa o acrilamida, la detecció dels fragments amplificats i si s'esqueia digerits es feu mitjançant tinció amb plata, bromur d'etidi o *Sybr Green* segons el cas (vegeu exemple a la figura 4.1).

Polimorfisme	Tipus	Seqüència dels *primers* (5'-3')	RFLP
92R7	C → T Substitució de base	TGCATGAACACAAAAGACGTA GCATTGTTAAATATGACCAGC	Pèrdua diana *Hind*III
SRY-2627	C → T Substitució de base	CGCGGCTTTGAATTTCAAGCTCTG CCAGGGCCCCGAGGGACTCTT	Pèrdua diana *Ban*I
sY81	A → G Substitució de base	AGGCACTGGTCAGAATGAAG AATGGAAAATACAGCTCCCC	Pèrdua diana *Nla*III
SRY-1532	G → A Substitució de base	TCCTTAGCAACCATTAATCTGG AAATAGCAAAAAATGACACAAGGC	Guany diana *Dra*III ,
SRY-8299	G →A Substitució de base	ACAGCACATTAGCTGGTATGAC TCTCTTTATGGCAAGACTTACG	Pèrdua diana *BsrB*I
YAP	+ → -Alu Inserció Alu	CAGGGGAAGATAAAGAAATA ACTGCTAAAAGGGGATGGAT	--

Taula 4.1.1. Polimorfismes bial·lèlics del cromosoma Y analitzats mitjançant RFLP

Polimorfisme	Condicions reacció PCR i condicions del termociclador	Enzim emprat	Longitud dels fragments	Detecció
92R7	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 2 min 94 C ;30 cicles a 60 C 20 s, 72 C 30 s i 94 C 20 s; més 72 C 5 min	*Hind*III	55 (0) ~25+30 (1)	PAGE 8% Tinció plata
SRY-2627	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 3 min 94C; 30 cicles a 63 C 30 s, 72 C 60 s i 94 C 30 s; més 72 C 5 min	*Ban*I	264+86+41(0) 350+41(1)	Agarosa 1.5% BrEt
sY81	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 3 min 94C; 30 cicles a 60 C 20 s, 72 C 30 s i 94 C 20 s; més 72 C 5 min	*Nla*III	~102+65+42 (0) ~144+65 (1)	Nusieve/Agarose Seakem (3:1) 4% *SybrGreen*
SRY-1532	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 3 min 94C; 30 cicles a 60 C 20 s, 72 C 30 s i 94 C 20 s; més 72 C 5 min	*Dra*III	167 (0) ~55+112 (1)	Nusieve/Agarose Seakem (3:1) 4% *SybrGreen*
SRY-8299	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 3 min 94C; 30 cicles a 60 C 30 s, 72 C 60 s i 94 C 30 s; més 72 C 5 min	*BsrB*I	~147+362 (0) 509 (1)	Agarosa 1.5% BrEt
YAP	50 ng DNA, 200 µM dNTPs, 1.5 mM $MgCl_2$, 1 µM *primers* 0.5 U/tub *Taq* polimerasa 3 min 94C; 30 cicles a 51 C 60 s, 72 C 60 s i 94 C 60 s; més 72 C 5 min	--	150 (0) 455 (1)	Agarosa 1.5% BrEt

Taula 4.1.2. Sinopsi del tipatge dels polimorfismes bial·lèlics del cromosoma Y analitzats mitjançant RFLP

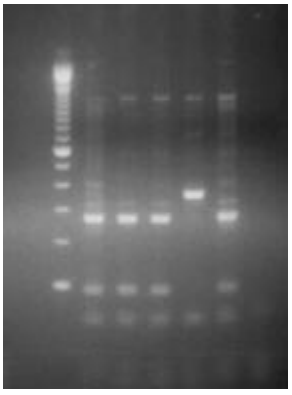


Figura 4.1. Exemple de tipatge de la mutació SRY-2627. Carril esquerre, marcador de longitud de seqüència. Els tres carrils següents i el darrer contenen DNA d'individus amb l'estat ancestral d'aquesta mutació, mentre que el segon per la dreta és un exemple d'al·lel derivat.

4.2 Polimorfismes analitzats mitjançant hibridació amb sondes específiques

Els polimorfismes 12f2, 50f2P i 50f2I foren analitzats mitjançant hibridació amb sondes específiques seguint els mètodes descrits per Casanova et al. (1985) i Jobling (1994). A partir de la digestió completa del DNA de cada individu a analitzar i posterior precipitació del DNA amb etanol, es realitzà una electroforesi en agarosa que fou seguida d'una transferència dels fragments digerits i separats a un filtre per *Southern blot*. Aquests filtres foren prehibridats amb esperma de salmó i posteriorment hibridats amb les corresponents sondes específiques dels polimorfismes a estudiar, les quals havien estat prèviament marcades amb α^{32}P mitjançant *random priming*. Després de diferents rentats per eliminar l'excés de sonda dels filtres, les bandes específiques de cada polimorfisme foren visualitzades per autoradiografia mitjançant l'exposició dels filtres a films sensibles.

Polimorfisme	Tipus	Detecció
12f2	Duplicació/deleció 10 Kb $\rightarrow$ 8Kb	Taq I (EcoRI) Hibridació
50f2P	Substitució de base 8.5 Kb $\rightarrow$ 3.1 Kb	Taq I Hibridació
50f2I	Substitució de base 8.5 Kb $\rightarrow$ 4 Kb	Taq I Hibridació

Taula 4.2. Característiques dels polimorfismes bial·lèlics del cromosoma Y detectats mitjançant hibridació amb sondes específiques.

4.3 DHPLC o *Denaturing High Performance Liquid Chromatography*

Recentment, Peter A. Underhill i Peter J. Oefner, un genetista i un bioquímic del Departament de Genètica de la Universitat d'Stanford (USA), han desenvolupat una nova estratègia molt eficient i automatitzada per estudiar la variació bial·lèlica en general, amb una aplicació inicial en el cromosoma Y. Aquesta nova estratègia, basada en l'aplicació de DHPLC per a la detecció d'heterodúplexs creats artificialment,

serveix tant per a descobrir nous polimorfismes com per a tipificar mostres poblacionals. Aquesta darrera és l'aplicació que n'he fet en aquest treball.

De manera esquemàtica aquest nou mètode es pot resumir en els següents quatre passos:

i) amplificació per PCR del segment de DNA que es vol estudiar ja sigui perquè s'hi vol intentar identificar nous polimorfismes o perquè conté una mutació que ja ha estat caracteritzada prèviament. En el cas de voler tipar dues mostres A i B (figura 4.3.1) caldrà amplificar per a cada una d'elles la regió que conté la mutació així com un control positiu (individu amb l'al·lel derivat) i una mostra de referència (individu amb l'al·lel ancestral).

ii) barreja equilibrada de cada mostra amplificada amb el producte PCR de referència de seqüència coneguda. És molt important per a una correcta anàlisi en el DHPLC que el nombre de molècules de DNA amplificades de cada mostra sigui comparable amb el de la mostra de referència. És per això que, com a pas previ a aquesta barreja, es verifica en agarosa que la intensitat dels fragments amplificats sigui comparable.

iii) desnaturalització (pujant la temperatura fins a uns 95 C) i reanellament (baixant gradualment la temperatura fins a 65 C) de cada barreja individual de manera que sigui possible la formació de molècules de DNA de doble cadena d'origen mixt, és a dir, molècules de doble cadena on una de les cadenes correspongui a la mostra a tipar i l'altra a la mostra de referència.

iv) per a cada mostra, anàlisi mitjançant el DHPLC de la possible presència, entre totes les molècules de DNA de doble cadena creades artificialment, d'aquelles molècules on les dues cadenes no siguin complementàries (és a dir, dels heterodúplexs).

El DHPLC explota la retenció diferencial de les molècules de DNA, segons siguin heterodúplex o homodúplex, en una columna cromatogràfica. Permet detectar d'aquesta manera els desaparellaments creats per un únic nucleòtid així com petites insercions o delecions dins un fragment de DNA amplificat de l'ordre de centenars de parells de bases de longitud quan es compara amb un producte PCR de referència.

Sobretot, està demostrant ser d'especial interès en aquelles regions del genoma en què el nombre de nucleòtids variables és baix (fet força comú en el cas del cromosoma Y), ja que en aquests casos els patrons d´heterodúplex observats poden ser clarament caracteritzats i diferenciats per a cada polimorfisme.

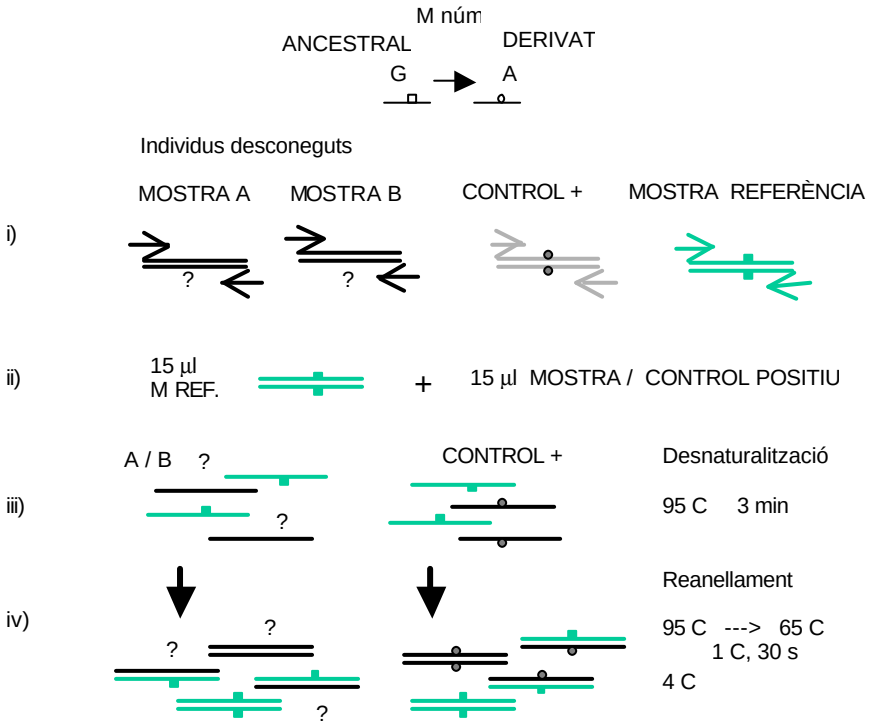


Figura 4.3.1. Tipatge per DHPLC d'una mutació qualsevol en el cromosoma Y.

En tots els tipatges, primer s'analitza el producte de referència o homodúplex que donarà un únic pic, el control positiu (en general s'esperen dos o més pics) i només si són correctes (els esperats donada la mutació a estudiar) es passa a analitzar les mostres. La temperatura en la que s'analitza cada mutació és crítica i variarà en cada cas en funció de la composició de la seqüència amplificada. Cal destacar que, només en aquells casos excepcionals on es mostra un heterodúplex diferent al característic determinat prèviament per a cada marcador caldrà seqüenciar per a identificar exactament la naturalesa d'un possible nou polimorfisme.

Mitjançant aquesta nova estratègia, l'equip investigador d´Stanford ha arribat a identificar i caracteritzar fins el moment un total de 166 polimorfismes bial·lèlics específics de la regió no recombinant del cromosoma Y (Underhill et al. 1999). Per

altra banda, la determinació dels estats ancestrals d'aquests marcadors mitjançant el tipatge de primats no humans i la caracterització de fins a 1.062 cromosomes d'origen ètnic divers, els han permès d'establir un nou arbre filogenètic del cromosoma Y molt detallat, compost per 115 llinatges masculins diferents que es reparteixen en deu grans grups (Underhill et al. 1999). Sens dubte, la seva descripció i interpretació detallada en cada regió demostraran ser de vital importància dins l'estudi de la diversitat genètica del cromosoma Y en les poblacions humanes. Els resultats obtinguts en el present treball mostren que és una eina amb un poder de resolució enorme per als estudis de diversitat i inferència evolutiva en humans.

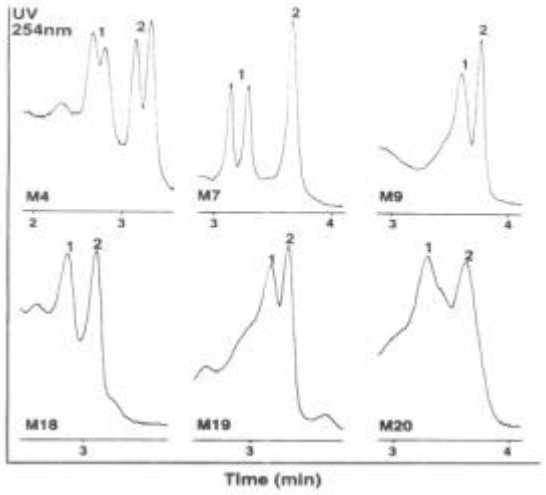


Figura 4.3.2. Exemples de cromatogrames on es mostra com es distingeixen els homodúplex (pics 2) dels heterodúplex (pics 1) quan es barregen els productes de PCR d'un individu baró de referència amb individus que presenten els estats derivats dels polimorfismes M4, M7, M9, M18, M19 i M20, respectivament (Underhill et al. 1997).

Hom preveu que tot aquest nou conjunt de marcadors bial·lèlics específic del cromosoma Y estarà disponible en una base de dades pública en la qual s'indicarà, per a cada mutació, les condicions d'amplificació, seqüència dels *primers* utilitzats, grandària del fragment a amplificar, posició on es troba la mutació respecte el final 5' del *primer* F, estat ancestral i al·lel derivat, més la temperatura utilitzada per a la detecció del polimorfisme mitjançant DHPLC.

En el present treball, el tipatge d'aquest nou conjunt de marcadors bial·lèlics es realitzà de manera seqüencial i ordenada, de més antics a més moderns segons la

seva posició en les branques de l'arbre filogenètic definit amb el total de 166 polimorfismes descrits fins el moment (figura 4.4.1). D'aquesta manera, sempre ha estat possible inferir l'estat al·lèlic d'aquells marcadors no tipats i, per tant, obtenir per a cada individu la informació completa de la combinació dels estats al·lèlics de tots els marcadors en forma d'haplotip.

Mutació	Polimorfisme	Mutació	Polimorfisme	Mutació	Polimorfisme
YAP positius					
M1	- -> + Alu	M107	A → G	M66	A → C
M40	G → A	M165	Anc. → der.	M155	G → A
M35	G → C	M33	A → C	M149	G → A
M34	G → T	M44	G → C	M58	C → T
M78	C →T	M75	G -> A	M154	T → C
M148	A → G	M2	A → G	M120	T → C
M81	C → T	M116	A → C	M124	C → T
YAP negatius					
M1	- → + Alu	M65	A → T	M52	A → C
M89	C → T	M153	T → A	M172	T → G
M9	C → G	M167	C → T	M67	A → T
M45	G → A	M37	C → T	M12	G → T
M173	A → C	M17	4 G's → 3 G's	M170	A → C
M150	A → C	M73	+ → - 2 pb	M72	A → der.
M126	+ → -4 pb	M70	A → C	M26	G → A
M18	+ → - 2 pb	M62	T → C	M21	A → der.

Taula 4.3.1. Conjunt de marcadors bial·lèlics específics del cromosoma Y analitzats mitjançant DHPLC en el present treball. Anc.: ancestral; der.: derivat, en els casos de polimorfismes per als quals hom no disposa d'aquesta informació.

4.4 Filogènia de la variació en el cromosoma Y

L'anàlisi en el DNA de la diversitat genètica ha suposat moltes millores i avantatges en relació a la genètica de poblacions *clàssica*. Entre aquests aventatges, sovint s'esmenta que hi ha un gran nombre de marcadors en el DNA, se'n coneixen les bases moleculars, es poden tipificar mitjançant mètodes estàndards de biologia molecular, i ens podem endinsar a esbrinar-ne les freqüències i patrons de mutació. Però, probablement, el que representa una gran millora és la possibilitat de conèixer la

filogènia de tota la variació existent. Això, de moment, és factible per a les regions no recombinants del nostre genoma: mtDNA i cromosoma Y.

Per al mtDNA, hi ha una quantitat extraordinària de treballs que intenten inferir el procés evolutiu que ha donat lloc a la variació observada. Però, probablement, la visió que en tenim s'haurà de revisar quan es conegui millor la variació a la regió més conservada de la molècula.

Per al cromosoma Y, hi ha hagut diversos intents i tot fa pensar que la descripció que tenim ara de 166 marcadors bial·lèlics és robusta i que presenta una filogènia amb alta correlació amb la que hipotèticament donaria la seqüència de tot el cromosoma Y.

A la figura 4.4.1 es presenta l'arbre filogenètic del conjunt de marcadors bial·lèlics presentats als apartats 4.1 i 4.2. Les relacions entre haplogrups foren determinades mitjançant un criteri de màxima parsimònia, que és especialment fàcil d'aplicar en una regió no recombinant i a polimorfismes amb una baixa taxa de mutació. En efecte, només SRY-1532 ha experimentat homoplàsia. Els estats al·lèlics ancestrals i derivats s'indiquen al llarg de les branques mitjançant sagetes i foren determinats mitjançant el tipatge en primats no humans. Dins els cercles, s'indica mitjançant codis numèrics els haplogrups determinats a partir dels estats al·lèlics dels polimorfismes inclosos.

A la figura 4.4.2 es mostra l'arbre filogenètic del cromosoma Y adaptat a partir de Underhill et al. (1999). Aquest arbre inclou fins a 166 polimorfismes descoberts mitjançant DHPLC. La construcció d'aquest arbre es féu amb els mateixos criteris que l'anterior: màxima parsimònia i determinació dels estats ancestrals a partir de primats no humans. Cada polimorfisme s'indica amb M més un número. Els haplotips definits per la combinació dels polimorfismes s'indiquen al final de les branques amb H més un número. Les xifres romanes designen els grups d'haplotips considerats per Underhill et al. (1999).

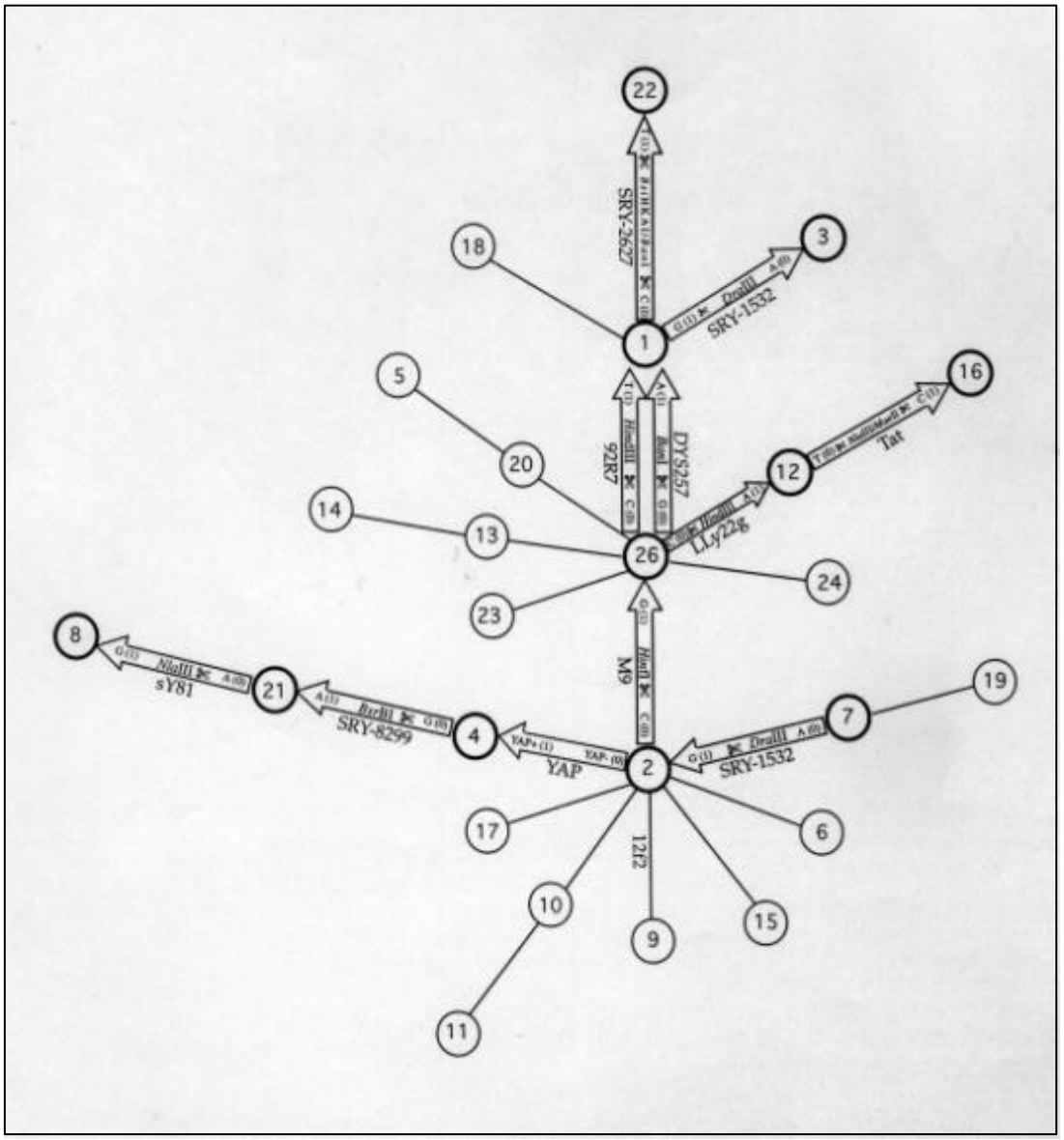


Figura 4.4.1 Arbre filogenètic del conjunt de marcadors bial·lèlics presentats als apartats 4.1 i 4.2.

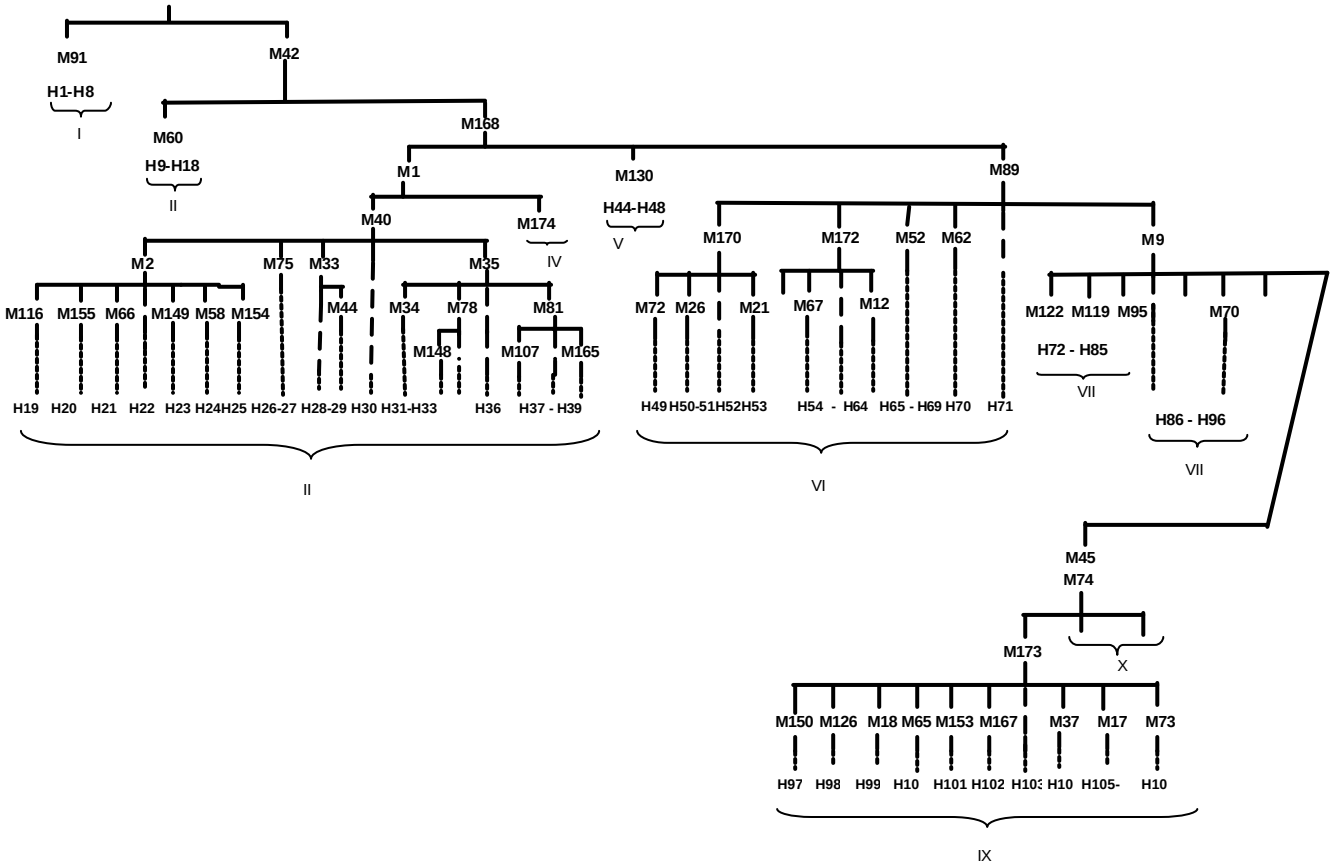


Figura 4.4.2. Arbre filogenètic del cromosoma Y adaptat a partir de Underhill et al. 1999.

5. Tractament estadístic

5.1 Elaboració d'una base de dades de freqüències al·lèliques recopilada de la literatura

Com a primer pas per a la descripció de la variabilitat genètica al nord d'Àfrica es recopilà una base de dades en Dbase III plus (Ashton-Tate) que contenia tota la informació disponible sobre polimorfismes genètics *clàssics* en mostres de poblacions d'aquesta regió. Diverses recopilacions (Mourant et al. 1976; Steinberg and Cook 1981; Tills et al. 1983; Roychoudhury and Nei 1988 i Cavalli-Sforza et al. 1994) foren utilitzades com a punt de partida i actualitzades a partir de dades publicades a les principals revistes del camp (*American Journal of Human Genetics*, *American Journal of Physical Anthropology*, *Annals of Human Biology*, *Annals of Human Genetics*, *Gene Geography*, *Human Heredity*, *Human Genetics*, *Humangenetik*, *Journal of Human Evolution*, *Tissue Antigens*, *Human Biology* i *Vox Sanguinis*) fins desembre de 1995.

Cada registre o entrada correspon a un polimorfisme estudiat en una mostra poblacional i, a més de les freqüències al·lèliques trobades, conté diversos camps amb informació relativa a la referència bibliogràfica de l'estudi, localització geogràfica de la mostra, així com de l'etnicitat, llengua, grandària mostral, i informació del polimorfisme estudiat. En total, la base de dades elaborada conté 1213 registres. El seu abast geogràfic engloba els territoris del Marroc, Algèria, Tunísia, Líbia, Egipte, Sàhara Occidental i Mauritània. Els polimorfismes inclosos són detectats, en la seva majoria, en els productes d'expressió gènica. Hom els anomena ja com a *clàssics* (per oposició als del DNA) i bàsicament inclouen quatre grans grups: enzims eritrocitaris, proteïnes plasmàtiques, antígens HLA i grups sanguinis.

5.2 Components principals: concepte estadístic i aplicació en genètica de poblacions

L'ànalisi de components principals, així com altres mètodes derivats (anàlisi de coordenades principals, anàlisi multidimensional, *biplot*, etc), són mètodes estadístics utilitzats per a simplificar dades multivariants amb pèrdua mínima d'informació, a

través de la reducció de la dimensió. En disminuir la complexitat inherent a les dades, permeten d'identificar els patrons de variació presents en aquestes, els quals, en el nostre cas, podran ser posteriorment interpretats en termes de genètica de poblacions. Si considerem un nombre gran de variables, n, mesurades en diferents punts (individus o poblacions), cadascun d'aquests punts es podrà representar en un espai de n-dimensions, prenent com a coordenades els valors de les variables. Intuïtivament, l'anàlisi de components principals consistirà en trobar quina és la recta en aquest espai n-dimensional tal que la suma de les distàncies dels punts originals a ella sigui mínima. Projectant els punts sobre aquesta recta s'obtenen els valors de la primera component principal. Algebraicament, aquesta transformació no és més que una combinació lineal de les variables originals:

$$PC = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n$$

on x són les variables i a els coeficients que cal derivar. A partir d'aquests coeficients hom pot trobar la correlació entre les variables originals i la component principal. Aquesta propietat serà molt útil alhora d'identificar quines són les variables que contribueixen a crear les diferències més grans dins la mostra. Evidentment, no tota la variació original és continguda en una component principal, però és la màxima que hom pot resumir en una dimensió. És possible, a més, calcular quina fracció de la variabilitat de la mostra és continguda en una component principal.

El següent pas per explorar la variabilitat restant consisteix en trobar la segona component principal, altra vegada complint la propietat que la suma de les distàncies dels punts originals a aquesta recta és mínima. Per contrucció, la segona component principal té correlació zero amb la primera i explica una menor fracció de la variabilitat. El mateix procés es pot repetir obtenint les següents components principals. Hom para l'extracció de components principals quan assoleix l'equilibri entre la variabilitat explicada per les components principals i el disposar d'un nombre suficient de components principals que puguem manejar i visualitzar fàcilment.

Les diferents components principals o factors obtinguts mitjançant aquest procediment tenen dues propietats principals. La primera és que el primer factor acumula el màxim percentatge de variació possible i que, avançant en les diverses components, la quantitat de variació que s'explica cada vegada és menor. I la segona és que les diferents components principals són incorrelacionades. Això significa que els processos evolutius que generaren la variabilitat mostrada per cada component

principal són diferents, i que, per tant, hom pot analitzar i interpretar cada component principal separadament.

En aplicar l'anàlisi de components principals a la genètica de poblacions humanes, les variables més emprades són les freqüències genètiques mesurades en diverses poblacions. Tanmateix, hom també troba exemples amb seqüències de nucleòtids en el DNA mitocondrial i amb dos *loci* del cromosoma Y (Cavalli-Sforza and Minch, 1997).

Si tenim dades per a cada polimorfisme a cada població, podrem aplicar directament aquests resultats a l'ànalisi de components principals i obtenir els diferents valors de les components principals per cada població. Aquests resultats poden després representar-se en una gràfica bi o tri-dimensional, amb cada població situada amb les seves coordenades principals. Malauradament, hom sovint troba i especialment amb els anomenats polimorfismes *clàssics*, que la cobertura geogràfica de les dades genètiques varia molt de polimorfisme a polimorfisme i de població a població. L'ànalisi de components principals no admet valors *missing*. Tanmateix, hom pot solucionar aquest problema recorrent a la interpolació de les freqüències al·lèliques als nodes d'una xarxa imaginària estesa sobre l'àrea geogràfica sota consideració. Els valors interpolats són llavors introduïts a l'anàlisi de components principals per a cada punt de la xarxa.

Posteriorment, hom pot utilitzar aquests mateixos nodes per a generar un paisatge genètic: una representació de les coordenades geogràfiques més el valor de component principal com a tercera dimensió. Cal recordar en aquest punt que el signe d'un valor de PC és intercanviable, i que, per tant, una ombra clara o fosca (si utilitzem una gradació d'intensitats d'ombres per a representar la tercera dimensió) o un pic o una vall (en una superfície tridimensional) no tenen cap significat intrínsec; és tot el paisatge genètic el que cal interpretar.

Resumint, l'anàlisi de components principals permet d'obtenir tres tipus bàsics de resultats:

i) els valors de components principals a cada node permeten d'obtenir mapes sintètics. Aquests resumiran els principals patrons espacials de la variabilitat genètica, els quals, podran ser interpretats en termes de la història de la població (migració i deriva).

ii) la correlació entre les freqüències genètiques i la component principal. Aquest resultat ens permet d'identificar quins al·lels creen les principals diferències en la nostra mostra de dades.

iii) el percentatge de variabilitat explicat per cada component principal. Ens dóna una idea de la importància relativa del procés que va causar cada component principal.

Cal tenir present que l'anàlisi de components principals és un mètode purament estadístic que, contràriament al que fan les distàncies genètiques, no incorpora cap model evolutiu particular a la diferència en freqüències al·lèliques entre poblacions.

5.3 Distàncies genètiques

Les distàncies genètiques s'utilitzen per mesurar de forma global la diferència genètica entre dues poblacions. Hom ha proposat moltes equacions i models diferents per a calcular aquesta diferència genètica. Tanmateix, tot i considerar diferents assumpcions teòriques i, en alguns casos, haver estat inicialment ideades per a ser aplicades a determinats tipus de marcadors genètics, trobem sovint que les diverses distàncies genètiques mostren altes correlacions entre elles (Cavalli-Sforza et al. 1994).

Al llarg del present estudi he emprat majoritàriament una d'aquestes mesures genètiques, la distància F_{st}, que es defineix com la variància estandaritzada de les freqüències al·lèliques (Wright 1951).

$$F_{st} = \frac{\mathrm{var}(p_i)}{\overline{p}_i(1-\overline{p}_i)}$$

on var(p_i) és la variància de les freqüències genètiques per a un conjunt de poblacions i $\overline{p}_i$ és la freqüència genètica mitjana en aquestes poblacions. Cal assenyalar que F_{st} representa directament la fracció de la variació genètica total que es troba entre poblacions en oposició a la fracció que es troba dins les poblacions.

Reynolds i col·laboradors (1983) proposaren el paràmetre *coancestry coefficient* (també anomenat distància F_{st} per Cavalli-Sforza i col·laboradors, 1994) com a estimador de la distància F_{st} per un conjunt de loci i de poblacions que permetia

estimar el temps de separació entre poblacions que haguéssin divergit per deriva genètica. En aquest cas,

$$D = -\log_e(1 - F_{st}) = t/2N$$

on *N* és la gradària efectiva de la població.

El *coancestry coefficient* (*D*) serà idèntic a la distància F_{st} per una parella de poblacions excepte pel que fa a la correcció per les diferents grandàries de les mostres poblacionals.

5.4 Representació gràfica en arbre de les distàncies genètiques

Per a visualitzar en forma d'arbre una matriu de distàncies genètiques, en la que hi ha el valor obtingut per a cada parell de comparacions possible, hom pot utilitzar diferents algoritmes estadístics.

En el present estudi he emprat l'algoritme *neighbor-joining* proposat per Saitou i Nei el 1987, el qual genera arbres sense arrel i tendeix a minimitzar la longitud total de les branques de l'arbre. En permetre que els extrems de les branques se situïn a longituds diferents no pressuposa una taxa constant d'evolució. Aquesta darrera característica permet, a més, obtenir millors correlacions entre les distàncies originals i les distàncies al llarg de l'arbre. El *neighbor-joining* és un mètode àmpliament emprat per la construcció d'arbres, el qual combina la velocitat de càlcul amb la singularitat del resultat: la majoria d'implementacions donen un únic arbre. Aquestes dues característiques (és a dir, obtenir un únic arbre, ràpid) l'han fet semblar especialment atractiu. El *neighbor-joining* és un mètode d'agrupació més que no pas un mètode d'optimització i, per tant, presenta la limitació que no optimitza un criteri d'ajust entre l'arbre i les dades. Tanmateix, és un bon mètode heurístic per estimar l'arbre de mínima evolució. Una estratègia per a trobar l'arbre de mínima evolució és primer calcular l'arbre *neighbor-joining,* i llavors veure si cap reordenació local d'aquest produeix un arbre més curt. Tanmateix, a la pràctica l'arbre *neighbor-joining* és sovint el mateix o molt similar a l'arbre de mínima evolució (Page i Holmes, 1998).

5.5 Validació de distàncies genètiques i arbres: *bootstrap*

Per a obtenir una estimació de l'error estadístic associat a les distàncies genètiques i per trobar una mesura de la robustesa estadística dels arbres genètics obtinguts, s'ha recorregut a la metodologia del *bootstrap*. Aquest mètode, inicialment descrit per Efron (1982) per a la estimació de la variància d'estadístics la distribució dels quals és desconeguda, es basa en el remostratge. A grans trets, l'estratègia del bootstrap consisteix en construir remostres dels al·lels que intervenen en el càlcul de les distàncies genètiques entre poblacions, de tal manera, que en cada remostra (o matriu *bootstrap*) hi ha al·lels originals absents i d'altres representats més d'un vegada però sempre conservant el nombre total d'al·lels originals. A partir de cada matriu *bootstrap* (se'n generen de l'ordre de 10.000) obtinguda d'aquesta manera es passa a generar llavors cada nova matriu de distàncies entre poblacions i el seu corresponent arbre.

La desviació estàndar de cada distància de les matrius *bootstrap* és un estimador directe de l'error de la distància. Per altra banda, cada vegada que un determinat *cluster* apareix entre els diferents arbres *bootstrap* generats es compta i es dóna com a percentatge en l'arbre original. Aquests percentatges seran els indicadors directes de la robustesa estadística de cada *cluster*.

5.6 Anàlisi de coordenades principals

L'anàlisi de coordenades principals forma part d'un conjunt de mètodes estadístics multivariants destinats a la reducció de la dimensió. Permet representar la matriu de distàncies en l'espai, tot adjudicant a cada població un petit nombre de coordenades (normalment restringides per tal d'obtenir una representació bidimensional), de tal manera que la distància entre aquests punts en el pla sigui el més semblant possible a la matriu de distàncies genètiques original. L'anàlisi de coordenades principals permet visualitzar les relacions contingudes en una matriu de distàncies genètiques sense pressuposar un model de fissions successives (com fan els arbres genètics), que és poc versemblant en poblacions humanes genèticament properes.

5.7 Detecció de fronteres genètiques

La detecció de fronteres genètiques està demostrant ser una eina especialment útil per a identificar possibles patrons geogràfics d'aïllament, sigui per raons culturals o per barreres geogràfiques, entre poblacions. A partir de la localització geogràfica de mostres poblacionals que han estat caracteritzades genèticament, hom pot unir parelles de poblacions contigües mitjançant una xarxa de Delaunay (Brassel i Reif, 1979) i associar cada un dels segments d'aquesta xarxa amb la distància genètica existent entre les dues poblacions que uneix. Per tal d'identificar les zones de canvi genètic més brusc, és a dir, detectar fronteres genètiques, cal començar a traçar una línia perpendicular al segment de la xarxa corresponent a la major distància genètica. Aquesta línia serà l'origen de la primera frontera genètica. Posteriorment s'estendrà la frontera iniciada tallant els segments adjacents que presentin les distàncies genètiques més grans fins arribar a sortir del límit de la xarxa. El mateix procés es pot repetir fins a obtenir el nombre de fronteres genètiques que hom cregui convenient per a la interpretació de les dades.

5.8 *Analysis of Molecular Variance* o AMOVA

L'anàlisi de la variància molecular permet d'analitzar l'estructura genètica a diferents nivells jeràrquics (dintre d'individus, dintre de poblacions, entre poblacions d'un mateix grup i entre grups) i testar la significació dels components de la variància obtinguts per aquests diferents nivells possibles d'estructura genètica mitjançant procediments permutacionals no paramètrics (Excoffier et al. 1992). És un mètode d'anàlisi molt flexible en el sentit que es pot aplicar a diferents tipus de dades genètiques, directament sobre les freqüències al·lèliques, o a partir del contingut al·lèlic dels haplotips, o en seqüències de DNA. Aquest mètode s'ha emprat per a testar la significació genètica de grups de poblacions on s'hi ha definit una estructura jeràrquica a partir de criteris geogràfics, lingüístics o culturals.

5.9 Paràmetres emprats per a la descripció de la variació en haplotips de microsatèl·lits

Per a caracteritzar la variació dels microsatèl·lits dins cada població (o *haplogroup* segons el cas) hom ha emprat diferents paràmetres de diversitat com són el nombre d'al·lels diferents, el nombre d'haplotips diferents, la diversitat haplotípica, la diversitat genètica mitjana per locus, el nombre mitjà d'al·lels diferents i el nombre mitjà de diferències en el nombre de repeticions.

El nombre d'al·lels diferents s'obté en comptar el nombre d'al·lels diferents presents en els haplotips de la població mentre que el nombre d'haplotips diferents s'obtindrà comptant directament el nombre d'haplotips diferents en a la població.

La diversitat haplotípica (també denominada com a índex de diversitat gènica) és equivalent a l'heterozigositat esperada per dades diploides. Es defineix com la probabilitat que dos haplotips escollits al atzar en a la mostra siguin diferents. S'estima a partir de la següent equació:

$$\hat{H} = \frac{n}{n-1}(1-\sum_{i=1}^{k} p_i^2)$$

on n és el nombre de còpies gèniques en a la mostra, k el nombre d'haplotips, i p_i la freqüència en a la mostra de l'i-èssim haplotip.

El nombre mitjà d'al·lels diferents s'obté comptant el nombre d'al·lels diferents entre totes les possibles parelles d'haplotips i divindint pel nombre total d'aquests en la mostra. El nombre mitjà de diferències en el nombre de repeticions s'estimarà de la mateixa manera però en aquest cas comptant la diferència en el nombre de repeticions entre totes les possibles parelles d'haplotips. Cal assenyalar, però, que aquest últim paràmetre no només ens indica quants al·lels són diferents sinó quant diferents són. Per tant, sota el model mutacional per microsatèl·lits *stepwise*, hom espera que mostri una millor correlació amb el temps de separació entre poblacions.

RESULTATS

CAPÍTOL I

Population History of North Africa: Evidence from Classical Genetic Markers

Elena Bosch, Francesc Calafell, Anna Pérez-Lezaun, David Comas, Eva Mateu i Jaume Bertranpetit

Human Biology (1997) 69: 295-311

Population History of North Africa: Evidence from Classical Genetic Markers

E. BOSCH,[1] F. CALAFELL,[1,3] A. PÉREZ-LEZAUN,[1] D. COMAS,[1] E. MATEU,[1,2] AND J. BERTRANPETIT[1,2]

Abstract After an intensive bibliographic search, we compiled all the available data on allele frequencies for classical genetic polymorphisms referring to North African populations and synthesized the data in an attempt to reconstruct the populations' demographic history using two complementary methods: (1) principal components analysis and (2) genetic distances represented by neighbor-joining trees. In both analyses the main feature of the genetic landscape in northern Africa is an east-west pattern of variation pointing to the differentiation between the Berber and Arab population groups of the northwest and the populations of Libya and Egypt. Moreover, Libya and Egypt show the smallest genetic distances with the European populations, including the Iberian Peninsula. The most plausible interpretation of these results is that, although demic diffusion during the Neolithic could explain the genetic similarity between northeast Africa and Europe by a parallel process of gene flow from the Near East, a Mesolithic (or older) differentiation of the populations in the northwestern regions with later limited gene flow is needed to understand the genetic picture. The most isolated groups (Mauritanians, Tuaregs, and south Algerian Berbers) were the most differentiated and, although no clear structure can be discerned among the different Arab- and Berber-speaking groups, Arab speakers as a whole are closer to Egyptians and Libyans. By contrast, the genetic contribution of sub-Saharan Africa appears to be small.

Genetic evidence can contribute to reconstructing the history of a population. The information contained in allele frequencies and in DNA data may allow us to identify the ancestors of the present inhabitants of an area and to understand the demographic processes that led to the current patterns and levels

[1] Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Catalonia, Spain.
[2] Institut de Salut Pública de Catalunya, Barcelona, Catalonia, Spain.
[3] Current address: Department of Genetics, Yale University School of Medicine, New Haven CT.

Human Biology, June 1997, v. 69, no. 3, pp. 295–311.

KEY WORDS: NORTH AFRICA, ELECTROPHORETIC POLYMORPHISMS, GENE FREQUENCIES, PRINCIPAL COMPONENTS ANALYSIS, GENETIC DISTANCES

of genetic diversity. In this endeavor linguistic and archeological tools provide independent evidence against which the models suggested by genetic analysis can be tested.

Northern Africa is a particularly interesting region. Although it belongs to continental Africa, its light-skinned populations were classified as Caucasoid by nineteenth-century anthropologists. The peopling of northern Africa appears to be conditioned by the barriers imposed to the north by the Mediterranean Sea and to the south by the Sahara Desert, which constrains human movement to an east-west direction. The harsh landscape, in which mountainous areas are surrounded by arid extensions, favors a dispersed, fragmented pattern of human settlement.

The first evidence of hominid occupation (Ain Hanech, Ternifine, and Sidi'Abd ar-Rahman) dates to more than 200,000 years ago and consists of remains classified as *Homo erectus* (Newman 1995). According to some controversial views (Stringer and Gamble 1993), the transition from archaic *Homo sapiens* to fully anatomically modern humans can be traced in northern Africa through fossils from Jebel Irhoud I (100,000–200,000 years ago), Dares-Soltan V (70,000 years ago; similar to the most ancient anatomically modern humans from the Levant), and Nazlet Khater (33,000 years ago; linked to European anatomically modern humans). These findings attest to the antiquity and continuity of human occupation in northern Africa.

Little is known about human population movements during the North African Upper Paleolithic. Neolithic populations diffused into the region from the east, where they contributed to the rise of the Egyptian kingdom (McEvedy 1980). In the west new production techniques appear to be associated with elements of a previous culture: the Capsian (7000–5000 B.C.) (Desanges 1990). The amount and geographic range of gene flow, if any, associated with the appearance of Neolithic populations is highly controversial.

Phoenicians (814 B.C.) and Romans (from 146 B.C.) occupied part of coastal northern Africa with limited population contributions, which were even less significant for Vandals (A.D. 429) and Byzantines (A.D. 533) (Newman 1995). The first Arab invasion, initially confined to Egypt, started in A.D. 643 and may have involved only a few thousand individuals (McEvedy 1980). The Arabs began to impose their religion and language over the Berber population, a process that culminated with the second and more numerous Arab wave in which the Bedouin reached the Maghreb (northwest Africa) in the eleventh century. The Islamic expansion went on to engulf the Iberian Peninsula (A.D. 711) and to occupy Sicily for about two centuries (A.D. 827–1016) (Hitti 1990). The later arrivals to northern Africa in colonial times include Europeans (Portuguese and Spanish in Morocco; French in Morocco, Algeria, and Tunisia; Italians in Libya) and Ottoman Turks, mainly in Egypt.

Most of the languages spoken today in northern Africa belong to the Afro-Asiatic (formerly named Hamito-Semitic) family (Ruhlen 1991). The origins of Afro-Asiatic are controversial: Starostin (1990) lexicostatistically

dates the Afro-Asiatic divergence at 15,000 years ago, whereas Militarev's lexical reconstruction (personal communication, 1992) places proto-Afro-Asiatic in the Natufian culture of the Levant, ca. 10,000 years ago. According to Renfrew (1991), the Afro-Asiatic languages expanded into Africa from the Levant with the demic diffusion of Neolithic populations. Three other language families also may have expanded in a similar process: Altaic toward Central Asia, Dravidian toward India, and Indo-European toward Europe. Barbujani et al. (1994) found that the weakest genetic evidence for all four expansions was for Afro-Asiatic.

In northern Africa two major branches of the Afro-Asiatic family are found: Arabic (composed of several dialects) and Berber. Berber embraces 30 extant languages with 11 million speakers divided into 4 main groups: (1) the central branch, which includes most Berber languages in Morocco, Algeria, and Tunisia; (2) the Tuareg languages; (3) the eastern group, which includes four little used languages in Libya and Egypt; and (4) the western group, represented by one language, Zenaga, spoken in southern Mauritania and Senegal [see Ruhlen (1991)].

Two other linguistic families are spoken on the edge of North Africa: Nilo-Saharan (with languages in southern Libya, Chad, Niger, and Mali) and the vast Nigero-Kordofanian, which stretches from the Sahel to South Africa. Both families are associated with dark-skinned people and can be considered sub-Saharan intrusions.

As outlined, several questions about the demographic history of northern Africa remain open, such as the levels of gene flow associated with population movements (e.g., the Neolithic and Arabic expansions), the relative effect of genetic drift, the extent of sub-Saharan admixture, and the contribution of northern African peoples to the gene pool of the northern Mediterranean shores. We address some of these problems through a study of the available genetic information (compiled from the literature) for northern Africans, synthesizing the data by means of genetic distances and principal components analyses and comparing them with European and sub-Saharan data.

Materials and Methods

Database. After an intensive bibliographic search, we compiled a database containing all the available information on allele frequencies in northern Africa. The geographic scope of our database includes Morocco, Algeria, Tunisia, Libya, Egypt, the Western Sahara, and Mauritania (Figure 1). Several published compilations were used as starting points for our database (Mourant et al. 1976; Steinberg and Cook 1981; Tills et al. 1983; Roychoudhury and Nei 1988; Cavalli-Sforza et al. 1994), which was updated with data published through December 1995. Each record corresponds to a polymorphism studied

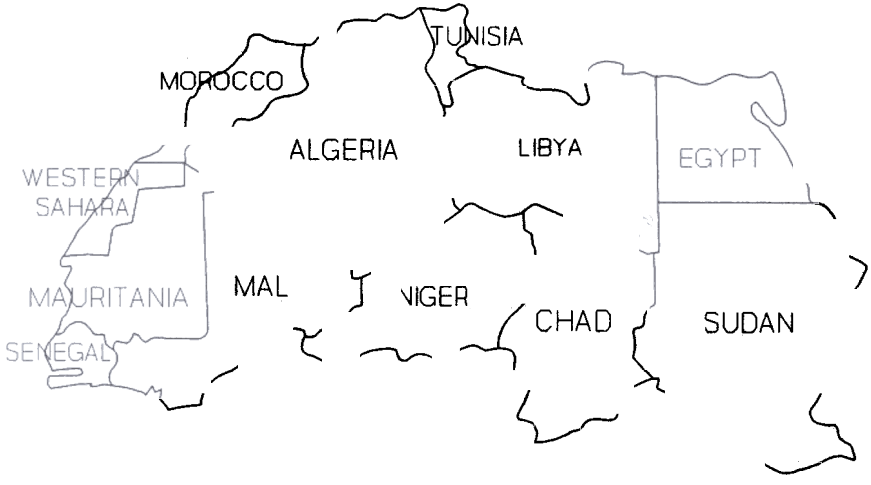


Figure 1 Political boundaries in North Africa.

in one population sample. Data on the polymorphism and its allele frequencies and on sample size, ethnicity, language, and geographic location were recorded. Genetic polymorphisms include blood groups, red cell enzymes, serum proteins, and HLA antigens (i.e., the so-called classical polymorphisms). The database contains 1213 records with data on 62 loci, although the number of systems with an adequate geographic coverage is much smaller. The database is available through anonymous ftp at porthos.bio.ub.es, subdirectory /pub/teu/nafrica.

Principal Components Analysis. The geographic coverage of genetic data varies greatly from polymorphism to polymorphism. To overcome this problem, we interpolated allele frequencies into a grid delimited by parallels 20°N and 39°N and by meridians 18°W and 36°E and spaced 1° apart. The estimated frequency for each node was obtained as an average weighted by the inverse squared distances between the node and the location of each sample. Interpolation was performed using the Surfer (v. 415) package (Golden Software, Golden, Colorado). Samples with fewer than 50 individuals and alleles with mean frequencies less than 0.01 were not taken into account because of their high sampling errors.

Seventeen polymorphisms composed of 33 alleles (Table 1) were included in a principal components analysis. Two basic selection criteria were followed: For each system samples had to be available for at least four of the analyzed countries, and differential selection could not be acting on these genes. These criteria excluded *G6PD* deficiency.

Table 1. Genetic Systems Used to Generate the Synthetic Maps

System	*Number of Alleles*[a]	*Samples*
Blood groups		
ABO	3	259
A1A2 (*ABO*)	2	30
RH		114
CDE		67
MN		94
MNSs (haplotypes)	4	44
KEL		23
FY		14
P1	3	25
JK		29
Proteins		
HP	1	47
PI	3	4
GC	1	14
Enzymes		
ACP1	3	28
PGD	1	16
GLO1		7
PGM1		16
Total	33	831

a. The number of alleles for each locus is that used in the principal components analysis, which is lower than the total number known or usually considered. Only alleles with a frequency higher than 1% have been taken into account.

Principal components analysis was performed on the correlation matrix. For the top three principal components, scores in every node were charted as contour plots by dividing the score range into six equal intervals and displaying them as different shades. In this analysis high and low principal component scores have no intrinsic meaning, and thus light and dark shades are interchangeable.

Genetic Distances and Trees. As opposed to the *continuous* nature of principal components analysis on an interpolated grid, genetic distances were computed in *discrete* population units. In a compromise between the number of populations and the number of genetic systems that have been typed in them, we devised two different levels of analysis. In the first approach we maximized the number of populations studied by including several Berber, Arab, and Tuareg groups for which information on 9 loci (*ABO*, *ACP1*, *HP*, *KEL*, *MN*, *P*, *PGM1*, *RH*, and *CDE*) with 29 alleles was available. The data matrix contained 17% missing values. In the second approach, to maximize the genetic information for each population unit, we pooled all Berber groups

in one unit and pooled the Arab samples from Morocco, Algeria, and Tunisia in a different group ("northwest Arabs"). This allowed us to consider 16 loci (*ABO*, *ACP1*, *AK1*, *ESD*, *FY*, *GC*, *HP*, *KEL*, *MN*, *MNS*, *P*, *PGD*, *PGM1*, *RH*, *CDE*, and *TF*) with 49 alleles. We also included populations from Europe (especially the Mediterranean), the Middle East, and sub-Saharan Africa to explore these populations' genetic relationships with northern Africa. The populations were Sicily, Sardinia, Greece, and Saudi Arabia (Cavalli-Sforza et al. 1994); the Basque Country, Andalusia, and the Balearic Islands (Calafell 1995); and West Africa (Cavalli-Sforza et al. 1994). Only 6% of the allele frequencies were missing.

An F_{ST}-based distance, the coancestry coefficient [Reynolds et al. 1983; see also Cavalli-Sforza et al. (1994, pp. 26–27)], was computed using the maximum number of loci available for every pair of populations. From the distance matrix genetic trees were computed by means of the neighbor-joining algorithm (Saitou and Nei 1987). A few small negatives branches were obtained; these were set to zero. Bootstrap analysis was used to estimate both distance errors and tree robustness. Genes were resampled with replacement, producing 1000 bootstrap data sets; we computed the corresponding distance matrix and neighbor-joining tree for each data set. The standard deviation of each distance across the bootstrap matrices is an estimator of the distance error (Efron 1982). In the neighbor-joining trees every occurrence of a particular cluster was recorded and given as a percentage of the 1000 bootstrap trees (Felsenstein 1985). Percentages above 50% were regarded as indications of statistical robustness for a cluster. The PHYLIP 3.5c package (Felsenstein 1989) was used throughout this analysis.

A Delaunay network (Brassel and Reif 1979) was used to define pairs of continuous samples. In this way the 10 estimated localities for the population groups used in Figure 5 were connected by 18 edges. The geographic baricenter of each population was estimated as the mean of the geographic coordinates of the samples included in each population, weighted by sample size. Each edge was associated with a genetic distance value. To identify the zones of sharpest genetic change, or genetic boundaries, we initially traced a perpendicular line across the edge showing the highest genetic distance, which was the origin of the first boundary. The boundary was then extended across the adjacent edges showing the highest genetic distances until it reached the limits of the network. Because of the small number of samples, we chose to define only the two most significant boundaries, repeating the procedure twice.

Results

Principal Components Analysis. Principal components analysis was performed on 33 allele frequencies (see Table 1). The first principal component

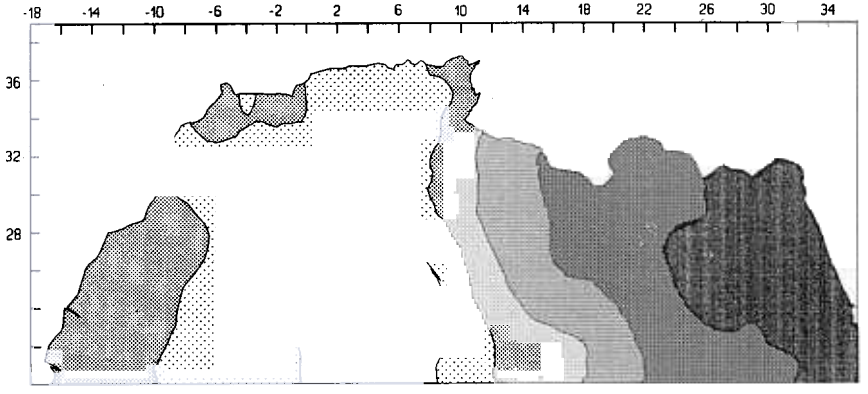


Figure 2. First principal component of gene frequencies in North Africa. This factor explains 36.5% of the variation.

(Figure 2) shows an east-west pattern of genetic differentiation with extremes in Egypt and southern Algeria. The alleles that showed the highest absolute correlation with the first principal component are listed in Table 2. The first principal component explains 36.5% of the total variation. The second principal component (Figure 3) explains 15.1% of the total variation and separates southern Libya from all other regions. Finally, the third principal component (Figure 4), which accounts for 12.8% of the genetic variation, can be interpreted as an irregular north-south gradient. The three top principal components explain 64.4% of the genetic variation.

Genetic Distances and Trees. Genetic distances were computed according to the method of Reynolds et al. (1983) for 10 northern African populations (Table 3). Their standard errors were estimated from 1000 bootstrap iterations and are shown above the diagonal in Table 3. The corresponding neighbor-joining tree (Figure 5) shows most populations, both Arab and Berber speakers, in a central cluster with no internal structure, as reflected by the short interpopulation branches and low bootstrap values. However, several populations seem to depart from this homogeneity: The south Algerian Berbers, the Tuaregs, and the Mauritanians are linked to the tree through long branches, whereas the Libyans and Egyptians cluster together, joining the rest of the populations with the most statistically robust branch (found in 78% of the bootstrap replications).

The strongest genetic boundaries in North Africa (Figure 6), as identified by overlaying the genetic distances on a Delaunay network, were found to encircle Libya and Egypt and the Tuareg, in accordance with both principal components analysis and genetic trees.

Table 2. Alleles Most Involved in the Three Main Principal Components[a]

Correlation	*Positive*	*Negative*
First principal component		
$\|r\| > 0.9$	*ABO*A1*	*HP**
	*FY*A*	
	K	
$0.9 > \|r\| > 0.8$	*cdE*	*JK*A*
	*P1*S*	
$0.8 > \|r\| > 0.7$		*GLO*1*
		*GC*1*
		*PGD*A*
		*ABO*O*
Second principal component		
$0.9 > \|r\| > 0.8$	*cdE*	
	Cde	
$0.8 \quad \|r\| \quad 0.7$	*CDE*	
$0.7 \quad \|r\| \quad 0.6$	*ACP1*A*	*ACP1*B*
	MS	
Third principal component		
$0.8 > \|r\| > 0.7$		*CDe*
$0.7 > \|r\| > 0.6$	*cDE*	
$0.6 > \|r\| > 0.5$	*P1*1*	*cde*
	NS	

Alleles are listed in descending order of the correlation coefficient (positive or negative) with each of the three main principal component axes.

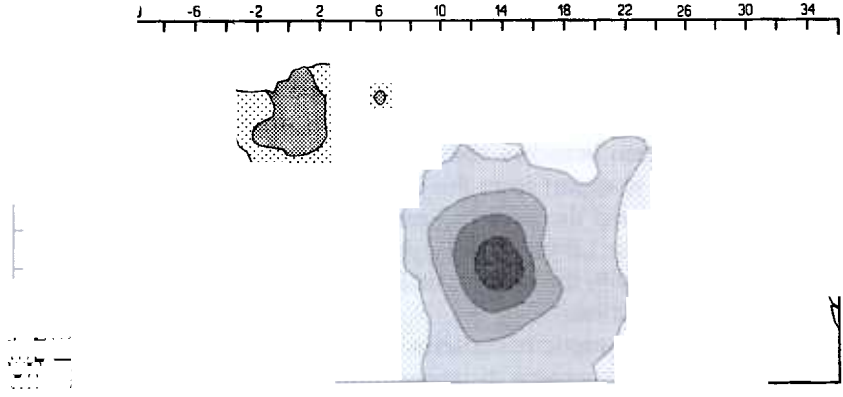


Figure 3. Second principal component of gene frequencies in North Africa. This factor explains 15.1% (51.6% accumulated) of the variation.

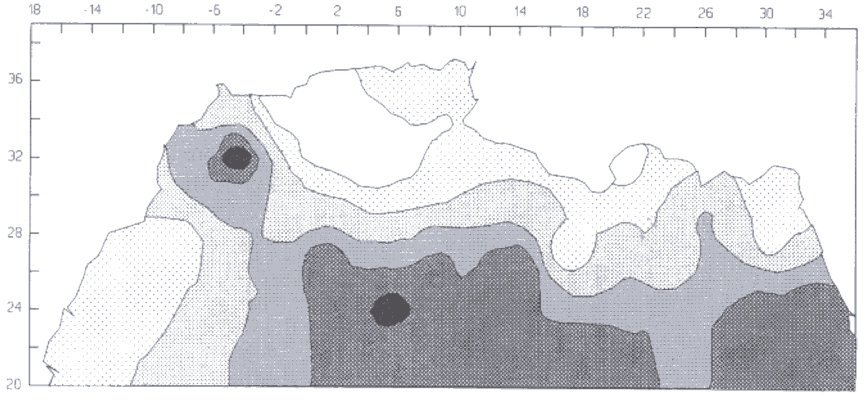


Figure 4. Third principal component of gene frequencies in North Africa. This factor explains 12.8% (64.4% accumulated) of the variation.

A larger number of genes and several European populations were added to the analysis to have a broader context of the genetic variation within North Africa. The genetic distances and their errors are shown in Table 4. The neighbor-joining tree (Figure 7) shows that the Libyans and Egyptians are close to both European Mediterraneans and Saudi Arabians, whereas the northwest Arabs, Berbers, and Tuaregs are separated from both the cluster of Europeans and northeast Africans, and the West Africans.

Discussion

Some caution has to be taken when interpreting principal components maps and genetic trees. Principal components are based on interpolated allele frequencies, which can create spurious clines (Sokal 1995); on the other hand, given the complex population history of humans and the pattern of migration and genetic admixture in this relatively small area, genetic trees should be regarded only as a reflection of the genetic distance matrix and should not be interpreted as literal schemes of branching and divergence between populations. However, it is reassuring to find strong coincidences in the results provided by both methods.

The first principal component, both neighbor-joining trees, and the genetic boundaries detected through the Delaunay network show a differentiation between Libya and Egypt versus the Berber and Arab populations of Morocco, Algeria, Tunisia, the Western Sahara, and Mauritania (to which we refer collectively as northwest Africa or the Maghreb). Several factors could have produced such a pattern.

Table 3. Genetic Distance Matrix and Standard Errors for the Populations Used in Figure 5[a]

Population	*Algerian Arab*	*Egyptian*	*Libyan*	*Mauritanian*	*Moroccan Arab*	*North Moroccan Berber*	*Atlas Berber*	*South Algerian Berber*	*Tuareg*	*Tunisian Arab*
Algerian Arab						33		65		
Egyptian	313					55		202		
Libyan	267	143				60		282		
Mauritanian	202	278	267			66		134		
Moroccan Arab	102	203	230	235		51		55		
North Moroccan Berber	117	130	165	159	84			130		
Atlas Berber	112	264	166	194	80	126		70		
South Algerian Berber	175	455	514	319	151	237	259			
Tuareg	302	416	472	317	213	282	259	533		
Tunisian Arab	87	124	92	123	69	47	59	262	222	

a. Below diagonal, Reynolds distances; above diagonal, standard errors estimated after 1000 bootstrap iterations. All values × 10,000.

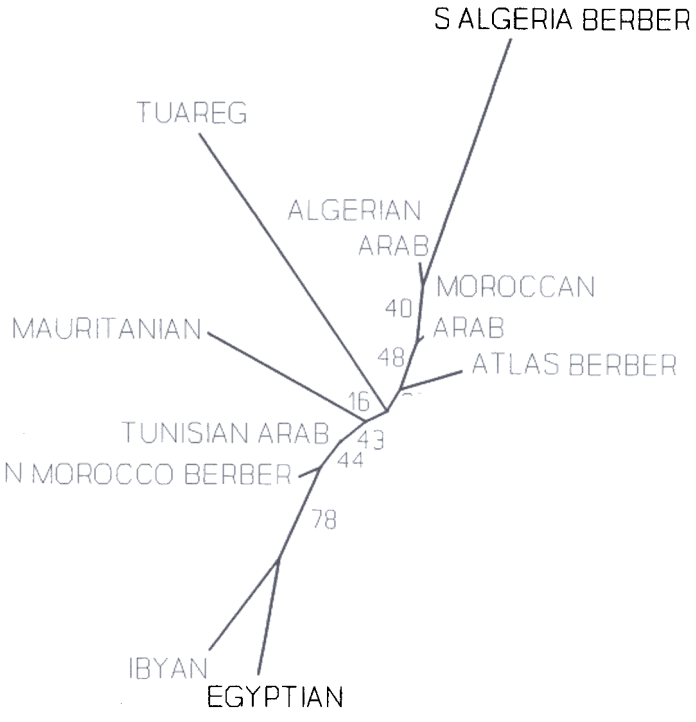


Figure 5. Neighbor-joining tree based on Reynolds genetic distances. Numbers in tree nodes represent the percentage of times that a certain node was found in 1000 bootstrapped trees. Abbreviations: S Algeria Berber, south Algerian Berbers; N Morocco Berber, north Moroccan Berbers.

The shape of the whole analyzed area is roughly a rectangle with its longest side oriented east-west; the area is bounded by the desert to the south. Thus a purely random isolation by distance process could have produced a longitudinal genetic cline. However, isolation by distance alone cannot account for the short genetic distance between Libya or Egypt and the European populations as far west as the Iberian Peninsula.

Directional migrations can produce genetic clines. In this case no west to east migrations are known, whereas a major demic diffusion process (Cavalli-Sforza et al. 1993, 1994) is known to have taken place in the opposite direction: the advance of Neolithic farmers. Demic diffusion of Neolithic populations from the Fertile Crescent is thought to have homogenized the genetic composition of the European populations (Cavalli-Sforza et al. 1993, 1994), and it created a major southeast to northwest gradient (Sokal et al. 1991). However, a steady progression of Neolithic populations, according to the wave of advance model (Ammerman and Cavalli-Sforza 1984), from the

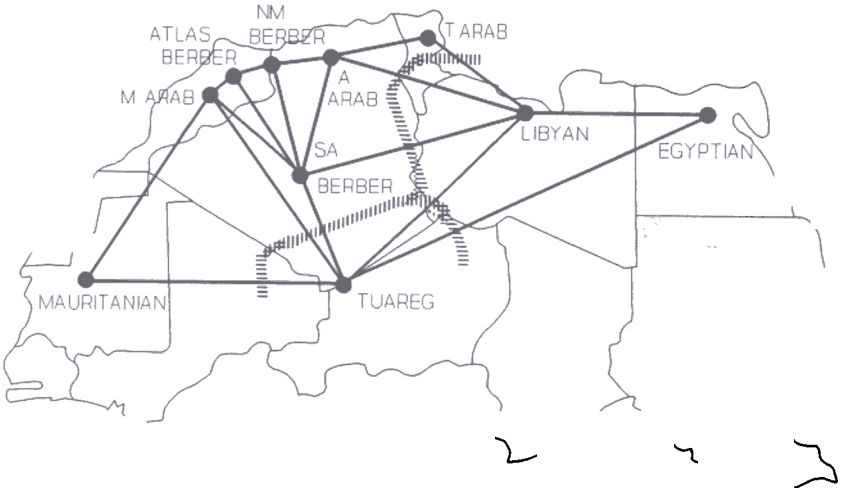


Figure 6. Delaunay triangulation between the estimated localities for the population groups used in Figure 5 (solid lines) and the two most significant genetic boundaries, recognized on the basis of the genetic distance approach (shaded lines). Abbreviations: M Arab, Moroccan Arabs; NM Berber, north Moroccan Berbers; A Arab, Algerian Arabs; T Arab, Tunisian Arabs; SA Berber, south Algerian Berbers.

Levant to the Atlantic would have produced a smooth cline rather than the abrupt separation between Egypt and Libya on the one hand and northwest Africa on the other. Differences in Mesolithic population size and local ecological factors may account for this pattern: Libya and Egypt are close to the original area of plant domestication; outside the Nile valley, food resources in a flat, desert land may have been scarce and Mesolithic population densities may have been low. Therefore population replacement during the Neolithic from the Levant could explain the genetic similarity between Libya, Egypt, and the European populations.

By contrast, the mountainous regions of northwest Africa may have had a higher carrying capacity for a hunter-gatherer population, which also may have delayed the adoption of early food production technologies. Thus the Neolithization of northwest Africa could follow an availability phase model (Zvelebil and Rowley-Conwy 1986), which would imply a limited or null gene flow and the preservation of the genetic differentiation generated before, be it during the Mesolithic or in older times. This model agrees with the archeological observations that the local pre-Neolithic Capsian culture was preserved well into the Neolithic, during which communities were both larger and more sedentary than before (Newman 1995).

Calafell and Bertranpetit (1993, 1994a,b) proposed a similar scenario to explain Basque differentiation. As seen in the neighbor-joining tree in

Table 4. Genetic Distance Matrix and Standard Errors for the Populations Used in Figure 7[a]

Population	*Berber*	*Northwest Arab*	*Egyptian*	*Libyan*	*Tuareg*	*Sardinian*	*Sicilian*	*Andalusian*	*Balearic Islands*	*Greek*	*West African*	*Saudi Arabian*	*Basque*
		38	78	75	106	171	71	58	70	92	133	142	107
	159		120	70	94	109	98	93	84	141	173	83	111
	265	234		33	114	76	32	49	40	23	241	123	161
	233	205	133		110	86	38	41	34	36	313	89	84
	284	301	399	410		183	168	188	172	167	150	74	214
	592	439	288	293	692		44	80	71	26	423	90	188
	331	292	138	102	508	139		12	14	6	329	87	90
	299	259	195	112	585	252	65		7	29	277	59	38
	21	264	169	78	574	220	53	31		29	315	73	61
	448	376	117	152	586	145	34	111	89		302	128	128
	385	496	840	854	587	1236	1012	933	1003	1155		208	323
	358	323	299	275	347	237	264	312	340	308	750		145
	420	410	470	257	728	499	235	93	145	332	1053	457	

a. Below diagonal, Reynolds distances; above diagonal, standard errors estimated after 1000 bootstrap iterations. All values × 10,000.

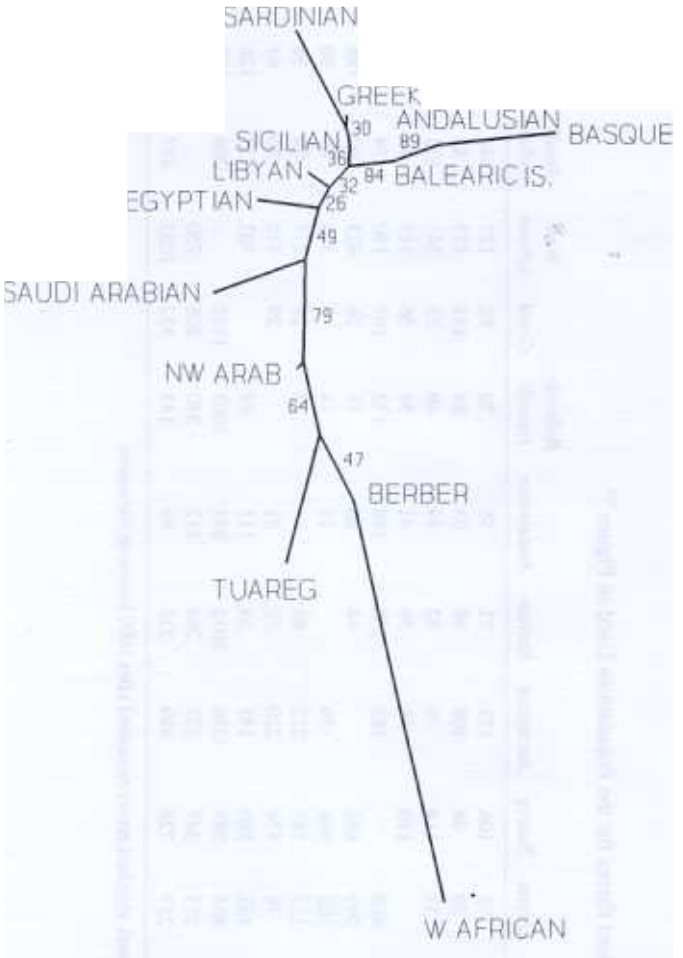


Figure 7. Neighbor-joining tree based on Reynolds genetic distances. Numbers in tree nodes represent the percentage of times that a certain node was found in 1000 bootstrapped trees. Abbreviations: NW Arab, northwest Arabs; Balearic Is., Balearic Islands; W African, west African.

Figure 7, the genetic distance between the northwest Arabs and Egyptians or Libyans is comparable to or slightly greater than that between the Basques and other European populations; it is much greater in the case of Berbers and Tuaregs, probably because of isolation.

Another migration could have contributed genes from the east: the Arab invasions of the seventh and eleventh centuries A.D. Although no clear distinction is seen between Arab and Berber speakers of northwest Africa in

Figure 5, Arab speakers appear closer to Libyans and Egyptians when a larger number of genes are incorporated into the analysis (see Figure 7). This can be interpreted as the result of genetic admixture between the local Berbers and the Arab migrants. A rough estimate of the genetic contribution of Arabs into the Maghreb can be produced with the triangle method proposed by Cavalli-Sforza et al. (1994). With 49 alleles (the same used in the second discrete approach), the proportion of Libyan genes in northwest Arab speakers would be $m = 0.346$, assuming that the present Berber population represents the genetic background into which the eastern genes were incorporated. This admixture estimate integrates the total hypothetical gene flow accumulated throughout history.

Unlike the Basques, the Berbers do not seem to have retained a local, isolated pre-Neolithic language; the Berber languages clearly belong to the Afro-Asiatic family. The most widely accepted hypothesis would place the diffusion of the Berber languages with the advance of the Neolithic populations, but our genetic analyses cannot determine the date and origin of the Berber languages. However, the present results indicate that the Neolithic diffusion did not involve massive gene flow, especially into northwest Africa, where most Berber languages are spoken today. This is in agreement with Barbujani et al. (1994), who showed that a coupled language-gene expansion for the Afro-Asiatic family from the Levant has little statistical support. Thus the Berbers could have an older origin.

The most widely differentiated populations in the Maghreb, that is, Mauritanians, Tuaregs, and south Algerian Berbers, are also the most isolated, with population sizes and mostly nomadic lifestyles that would allow drift to act deeply. Limited gene flow would not have been able to erase the differentiation generated by drift. It can be hypothesized that high levels of genetic heterogeneity exist among the numerous poorly studied Berber-speaking groups.

The genetic contribution of sub-Saharan Africa appears to be relatively small. A vague north-south pattern is seen only in the third principal component, which is compatible with limited gene flow across the Sahara. Moreover, this contribution should be studied through the analysis of regions of the genome for which a much clearer difference exists between sub-Saharan Africa and other populations; mtDNA (Vigilant et al. 1991) or the *MS205* minisatellite (Armour et al. 1996) could provide such a genetic tool.

According to historical sources, the Arab invasions of the Iberian Peninsula were carried out mostly by Berber men from north Morocco and Algeria under Arab leadership (Hitti 1990). Thus the northwest Arabs are the present population that best represents the invaders. Nevertheless, the genetic distance between Andalusia and the northwest Arabs (0.0259 ± 0.0058) is more than twice that between Andalusia and Libya (0.0112 ± 0.0041). The Arab invasion seems to have had a limited impact on the south Iberian gene pool, as recognized by historians.

Further information is needed to solve many of the problems that remain with the population history of northern Africa. The study of new genomic regions for well-defined populations may provide such new insight into old and intriguing questions.

Acknowledgments One anonymous reviewer and two other reviewers who waived anonymity provided meaningful insights that significantly improved this manuscript. Financial support for this research came from the Dirección General de Investigación Científico Técnica (Spain) through projects PB92-0722 and PB95-0267-CO2-01, from Human Capital and Mobility through contracts ERCHRXCT92-0032 and ERB-CHRX-CT920090, and from Grup de Recerca Consolidat through grants 95 SGR00205 and 96 SGR00041 (Comissionat per a Universitats i Recerca, Catalan Autonomous Government) awarded to Jaume Bertranpetit. Elena Bosch was awarded a Ph.D. fellowship (grant FI/96-1.153) from the Commisionat per a Universitats i Recerca (Catalan Autonomous Government). Anna Pérez-Lezaun received a Ph.D. fellowship from the Spanish Ministry of Education and Science (grant FP93-38110903). Eva Mateu was granted a Ph.D. fellowship from the Institut de Salut Pública de Catalunya (grant ISP11/95), and David Comas received a Ph.D. fellowship from the Comissionat per Universitats i Recerca (Catalan Autonomous Government) (grant FI/93-1.151). Robin Rycroft (Servei d'Assessorament Lingüístic, Universitat de Barcelona) helped with the English manuscript.

Received 10 June 1995; revision received 23 September 1996.

Literature Cited

Ammerman, A.J., and L.L. Cavalli-Sforza. 1984. *The Neolithic Transition and the Genetics of Populations in Europe.* Princeton, NJ: Princeton University Press.

Armour, J.A.L., T. Anttinen, C.A. May et al. 1996. Minisatellite diversity supports a recent African origin for modern humans. *Natur. Genet.* 13:154–160.

Barbujani, G., A. Pilastro, S. De Domenico et al. 1994. Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonization. *Am. J. Phys. Anthropol.* 95:137–154.

Brassel, K.E., and D. Reif. 1979. A procedure to generate Thiessen polygons. *Geogr. Anal.* 11:289–303.

Calafell, F. 1995. Anàlisi de la diversitat genètica de les poblacions humanes de la Península Ibèrica. Ph.D. dissertation, Universitat de Barcelona, Barcelona, Spain.

Calafell, F., and J. Bertranpetit. 1993. The genetic history of the Iberian Peninsula: A simulation. *Curr. Anthropol.* 34:735–745.

Calafell, F., and J. Bertranpetit. 1994a. Mountains and genes: Population history of the Pyrenees. *Hum. Biol.* 66(5):823–842.

Calafell, F., and J. Bertranpetit. 1994b. Principal component analysis of gene frequencies and the origin of Basques. *Am. J. Phys. Anthropol.* 93:201–215.

Cavalli-Sforza, L.L., P. Menozzi, and A. Piazza. 1993. Demic expansions and human evolution. *Science* 259:639–646.

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

Desanges, J. 1990. The proto-Berbers. In *General History of Africa,* v. 2, *Ancient Civilizations of Africa,* G. Mokhtar, ed. Paris, France: UNESCO, 236–245.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:785–791.

Felsenstein, J. 1989. PHYLIP: Phylogeny Inference Package (version 3.2). *Cladistics* 5:164–166.

Hitti, P.K. 1990. *The Arabs: A Short History.* Washington, DC: Gateway Editions.

McEvedy, C. 1980. *The Penguin Atlas of African History.* New York: Penguin Books.

Mourant, A.E., A.C. Kopeć, and K. Domaniewska-Sobczak. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms.* London, England: Oxford University Press.

Newman, J. 1995. *The Peopling of Africa: A Geographic Interpretation.* New Haven, CT: Yale University Press.

Renfrew, C. 1991. Before Babel: Speculations on the origins of linguistic diversity. *Cambridge Archaeol. J.* 1:3–23.

Reynolds, J., B.S. Weir, and C.C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767–779.

Roychoudhury, A.K., and M. Nei. 1988. *Human Polymorphic Genes: World Distribution.* New York: Oxford University Press.

Ruhlen, M. 1991. *A Guide to the World's Languages,* 2d ed. Stanford, CA: Stanford University Press.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molec. Biol. Evol.* 4:406–425.

Sokal, R.R. 1995. Genes, geography, and the human family. *Q. Rev. Biol.* 70:321–323.

Sokal, R.R., N.L. Oden, and L. Wilson. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* 351:143–145.

Starostin, S.A. 1990. A statistical evaluation of the time-depth and subgrouping of the Nostratic macrofamily. In *Evolution: From Molecules to Culture,* R. Dawkins and J. Diamond, eds. Cold Spring Harbor, NY: Cold Spring Harbor Press, 33.

Steinberg, A.G., and C.E. Cook. 1981. *The Distribution of the Human Immunoglobulin Allotypes.* Oxford, England: Oxford University Press.

Stringer, C., and C. Gamble. 1993. *In Search of the Neanderthals.* New York: Thames and Hudson.

Tills, D., A.C. Kopeć and R.E. Tills. 1983. *The Distribution of the Human Blood Groups and Other Polymorphisms, Supplement 1.* London, England: Oxford University Press.

Vigilant, L., M. Stoneking, H. Harpending et al. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507.

Zvelebil, M., and P.R. Rowley-Conwy. 1986. Foragers and farmers in Atlantic Europe. In *Hunters in Transition,* M. Zvelebil, ed. Cambridge, England: Cambridge University Press. 67–93.

CAPÍTOL II

Genomic structure of northwestern Africa revealed by STR analysis

Elena Bosch, Francesc Calafell, Anna Pérez-Lezaun, Jordi Clarimón, David Comas, Eva Mateu, Rosa Martínez-Rosa, Bernal Morera, Zahra Brakez, Omar Akhayat, Abdelaziz Sefiani, Ghania Hariti, Anne Cambon-Thomsen i Jaume Bertranpetit

European Journal of Human Genetics (en premsa)

GENETIC STRUCTURE OF NORTHWESTERN AFRICA REVEALED BY STR ANALYSIS

Elena Bosch[1], Francesc Calafell[1], Anna Pérez-Lezaun[1], Jordi Clarimón[1], David Comas[1], Eva Mateu[1], Rosa Martínez-Arias[1], Bernal Morera[1], Zahra Brakez[2],Omar Akhayat[2], Abdelaziz Sefiani[3], Ghania Hariti[4], Anne Cambon-Thomsen[5], and Jaume Bertranpetit[1].

(1) Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Spain.
(2) Laboratoire de Biologie Cellulaire et Moléculaire, Faculté des Sciences, Université Ibnou Zohr, Agadir, Morocco.
(3) Institut National d'Hygiène, Rabat, Morocco.
(4) Hôpital Mustapha, CHU Alger Centre, Algeria.
(5) INSERM U 518, Faculté de Medicine, Toulouse, France.

Address for correspondence:
Jaume Bertranpetit, 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 40
Fax: (+34 93) 542 28 02.
e-mail: jaume.bertranpetit@cexs.upf.es

RUNNING TITLE: STRs in Northwest African populations.

ABSTRACT

We have analysed a large set of autosomal short tandem repeat (STR) loci in several Arabic and Berber-speaking groups from Northwestern Africa (i.e., Moroccan Arabs, northern central and southern Moroccan Berbers, Saharawis, and Mozabites). Two levels of analysis have been devised using two sets of 12 (D3S1358, vWA, FGA, THO1, TPOX, CSF1PO, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820) and 21 (the former set plus D9S926, D11S2010, D13S767, D14S306, D18S848, D2S1328, D4S243, F13A1, and FES/FPS) STR loci. For each set, data for a number of external reference populations were gathered from the literature. Several methods of analysis based on genetic distances (neighbor-joining trees, principal coordinate analysis, boundary detection), as well as AMOVA, showed that genetic differentiation among NW African populations was very low and devoid of any spatial pattern. If the NW African populations were grouped according to cultural differences in Arabs vs Berber-speakers and Saharawis, this cultural partition was not associated to genetic differentiation. Thus, it is likely that the Arabization was mainly a cultural process. A clear genetic difference was found between NW African populations and Iberians, although some degree of gene flow into Southern Iberia may have existed. NW Africans were genetically closer to Iberians and to other Europeans than to African Americans.

KEYWORDS: STRs, Microsatellites, Population genetics, North Africa

INTRODUCTION

Microsatellites, also called short tandem repeats (STRs), are tandemly repeated DNA sequences of 2-5 bp in length, which are found highly widespread throughout the human genome. They are extraordinarily polymorphic and can be easily PCR-assayed. Because of these features, STRs are the markers of choice for most genetic mapping studies[1-3], forensic applications[4-5], plus evolutionary and populational studies[6-10]. Each unlinked autosomal STR behaves as an independent locus; then, the study of a number of unlinked autosomal STRs allows to peer at a number of independent realisations of the evolutionary process. Their joint analysis may reveal overall trends that may have been obscured by stochastic factors at any single locus. It is worth noting that mtDNA and the non-recombining portion of the Y chromosome, which have been largely used to unreavel human population history, contain very informative markers, but both genome regions behave each as a single locus.

We have analysed a large set of autosomal STR loci in several Arabic and Berber-speaking populations from NW Africa in an attempt of genetic characterisation of that region. Berber is a branch of the Afro-Asiatic language family; this branch embraces more than 30 distinct languages distributed from Egypt to Senegal and were the only languages spoken in the area until the Arab invasion. Arabic is a language classified within the Semitic branch of the Afro-Asiatic family and includes several regional dialects[11]. Part of these Berber languages and Arabic dialects are represented in our samples. Saharawis, who live in the Western Sahara, speak Hassani, an Arabic dialect distinct from the classical Arabic introduced into NW Africa from the 7th century by the Moslem invasions and spoken nowadays by the Moroccan Arab population. Tachelhit, spoken mainly in the Souss valley, is a Berber language confined to southern Morocco and distinct from other Berber languages spoken in the northern (Tarifit) and central (Tamazigh) areas of Morocco. The Mozabites are a very well defined Berber population in Algeria: they speak Mzab (a distinct Berber language) and originated from the Ibadite religious community who settled in this region in the 11th century, creating five neighbouring towns (known as the Pentapole), Ghardaia being one of them.

We are interested in elucidating the possible structure of STR variation in these NW African populations, but also in knowing about their relations with the neighbouring

populations to the north, that is, the Iberians, and to the south, the Sub-Saharan Africans.

The aims of this study are to approach:

i, the heterogeneity of speakers of different Berber languages. These are the North African human groups more likely to have preserved ancient charactheristics as they preserve the oldest known language and social structure. The extent of genetic differentiation and of the loss of diversity within these groups can be used to ascertain the historical patterns of gene flow between these populations.

ii, the genetic differentiation between Arabs and non-Arabs. The apportionment of genetic diversity by ethnicity in NW Africa (Arab vs. Berbers and Saharawis) can provide information about the demographic impact of the Arab expansions in the late 7th and the 11th centuries AD in the autochthonous populations of the region. Although there is a clear cultural and linguistic differentiation, it is not clear at all to which extent a main demographic migration caused the cultural change or it was acculturation without genetic change.

iii, the extent of population admixture associated with migrations between NW Africa and Iberia. No significant historical movements are known from the Iberian Peninsula to NW Africa, except for the expulsions of the Jews in the late 15th century and of the Moors in the 17th century. The Islamic invasion of the Iberian Peninsula in 711 AD (with subsequent invasions in the 11th century) appears to be the main migration event, although its actual demographic impact may have not been large[12-13]. In our study, we have compiled Iberian data from the literature that comprise areas with a null or short Arab occupation (Basques, Catalans and northern Portuguese) as well as samples from areas with a much longer Arab occupation (Andalusians). Allele and haplotype frequencies of some loci of a single genetic system (HLA) show certain similarities for Moroccan, Algerians, Basques and other Iberians[14], which according to some authors[15-16] suggests a common origin for the Iberian and NW African populations. Our data set, which contains data for 12 to 21 highly polymorphic loci, will also allow to test this hypothesis.

MATERIALS AND METHODS

Population samples analysed

The NW African samples analysed in the present study include a Moroccan Arab population (2N= 94-160), a pooled group of Berber speakers from northern and central Morocco (2N= 50-126), a Berber population from Southern Morocco (2N= 84-96), a Saharawi populational sample (2N= 104-118) from the Western Sahara plus an Algerian Berber group, the Mozabites, which were collected in the town of Ghardaia (2N= 88). Differences in number of chromosomes analysed within a population were due to DNA availability and to the different methods used for typing each STR. Geographic location is shown in Figure 4. Allele frequencies for 13 loci in a partial subsample of Arabs and northern Berbers are given in Pérez-Lezaun *et al*.[17]. DNA was extracted from fresh whole blood using standard phenol-chloroform methods. Appropriate informed consent was obtained from all participants in this study and, in most cases information about geographic origin of their four grandparents and maternal tongue was recorded.

STR typing

Most samples were PCR amplified using commercial kits (PE Applied Biosystems) AmpF*l*STR Profiler Plus (D3S1358, VWF/VWA, FGA, Amelogenin, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820) and AmpF*l*STR Cofiler (D3S1358, D16S539, Amelogenin, TH/THO1, TPOX, CSF1PO and D7S820) or AmpF*l*STR Green I (Amelogenin, TH/THO1, TPOX and CSF1PO) according to manufacters' recommendations. Additional samples were genotyped for some STR loci (VWF/VWA, D5S818, D7S820, TH/THO1, TPOX) overlapping the commercial kits using fluorescently labelled primers. Other nine STR loci (D11S2010, D13S767, D14S306, D18S848, D2S1328, D4S243, F13A1, FES/FPS, D9S926) were typed in all the NW African populations except for the Mozabites, as no sufficient DNA was available, also using fluorescently labelled primers. Almost all loci map in different chromosomes. Chromosomal locations for the loci without standard names are 12p12-pter (VWF/VWA), 4q28 (FGA), 11p15.5 (TH/THO1), 2p23-2p (TPOX), 5q33.3-34 (CSF1PO), 6p24.2-p23 (F13A1), and 15q25-qter (FES/FPS). Primer sequences can be found in Pérez-Lezaun *et al.*[17]. Single locus amplifications were carried out in a final

reaction volume of 10 µl according to the following PCR cycling conditions: 94 C for 1 min; 14 cycles at 94 C for 20 s, with annealing temperatures decreasing from 63 C by 0.5 C successively for 1 min, and 72 C for 1 min, 20 cycles at 94 C for 20 s, 56 C for 45 s, and 72 C for 1 min; plus an extension cycle at 72 C for 5 min. All PCR reactions were performed using a Perkin Elmer 9600 thermal cycler. Amplification products were run in an ABI 377™ sequencer using 36/48-cm well-to-read plates. ABI GS500 ROX and ABI GS500 TAMRA were used as internal lane standards. The GeneScan 672™ and Genotyper 2.1x3™ software packages were used to collect the data, analyse fragment sizes and to designate alleles by comparison to locus-specific ladders.

Reference populations

A variety of populational samples described in other genetic studies have been included as reference populations in the present analysis. Original references for Catalans (2N=100-176) and Basques (2N=98-200) can partly be found in Pérez-Lezaun *et al*.[17] and in Pérez-Lezaun *et al*.[18], where they were included in an overall European sample that comprised two other European samples. Original references for Portuguese (2N=72) and Andalusians (2N=68-72) can also be found in Pérez-Lezaun *et al.*[17]. Italians (2N=446) are described in Garofano *et al*.[19]. Populational data for African Americans (2N=390) and European Americans (2N=400; also called "Caucasians" in the original reference) is available in AmpF*l*STR Profiler Plus and AmpF*l*STR Cofiler (PE Applied Biosystems) User's Manuals.

Statistical Analysis

In a compromise between the number of populations and the number of STR loci that have been typed in them, we devised two levels of analysis using two different data sets, which we call basic and extended. In the basic data set, we use 12 STR loci (D3S1358, vWA, FGA, THO1, TPOX, CSF1PO, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820) typed in all the NW African populations analysed in this study (Moroccan Arabs, northern central and southern Moroccan Berbers, Saharawis and Mozabites) and that are also available in a number of populations compiled from the literature and used for external reference (Basques, Catalans, Italians, Portuguese, Andalusians, Italians, African American, and European American). In the extended data set, we use 21 STR loci (that consist of the 12 included in the basic data set plus D9S926, D11S2010, D13S767, D14S306, D18S848, D2S1328, D4S243, F13A1, and

FES/FPS) typed in Moroccan Arabs, northern central and southern Moroccan Berbers, and Saharawis, and that are also available in the literature for two populations from the Iberian Peninsula (Basques and Catalans).

We computed F_{st} distances[20] between each pair of populations from the basic and extended data sets and represented them as neighbor-joining trees[21]. Distance standard errors were approximated through bootstrap analysis[22-23] from the standard deviation of the bootstraped distances obtained from ten thousand resamples drawn at random with replacement from each locus set. Neighbor-joining tree robustness was assessed by bootstrap analysis[24]; every occurency of a particular cluster was recorded and given as a percentage of the 10,000 bootstrap trees drawn from the previously bootstrapped matrix distances. Principal coordinates analysis was performed on the F_{st} distance matrix by using NTSYS-pc version 1.70 (Applied Biostatistics, Inc.).

A Delaunay network[25] was used to define pairs of contiguous samples. In this way, the 9 estimated localities for the NW African and Iberian populations used in the basic data set were connected by 17 edges. We associated to each edge the F_{st} genetic distance between the pair of populations it links. To identify the zones of sharpest genetic change, or genetic boundaries, we initially traced a perpendicular line across the edge showing the highest genetic distance, which was the origin of the first boundary. The boundary was then extended across the adjacent edges showing the highest genetic distances until it reached the limits of the network. This procedure can be iterated to define a second, a third or any higher-order boundary.

Analysis of Molecular Variance or AMOVA[26] was performed using Arlequin v.11[27] independently for each locus; the resulting genetic variance apportionment fractions were averaged over loci and their associated p-values were combined by means of Fisher's technique[28].

RESULTS

We studied Northwest African populations by typing 21 autosomal STR loci in Moroccan Arabs, northern central and southern Moroccan Berbers, Saharawis, and a subset of 12 loci in Mozabites. Allele frequencies for each locus and population are presented in the Appendix. Hardy-Weinberg equilibrium was tested for all possible

locus-population combinations through two methods: χ^2 comparison of observed versus expected homozygotes, and through an exact test as implemented in the Arlequin package. Three out of 96 χ^2 tests were statistically significant ($p<0.05$): D18S51 in Moroccan Arabs, and D21S11 in Mozabites and in Saharawis. Seven out of 96 exact tests were statistically significant ($p<0.05$): D5S818 in Mozabites, D5S818 and TPOX in Moroccan Arabs, D21S11, FES and F13A1 in Saharawis, and D4S243 in northern central Moroccan Berbers. However, no consistently deviations were observed at all of the loci in a single population, not at a single locus in all the populations. Given the large number of locus-population combinations tested we applied Bonferroni correction to test whether chance departures from Hardy-Weinberg equilibrium due to multiple testing could explain the observed deviations. No significant deviations remained after correction for multiple testing; therefore, equilibrium may be assumed for all loci in all populations.

STR diversity within populations

The amount of internal genetic diversity in a population can be used as a measure of its present cultural and/or geographical isolation. Within each population of the basic and extended sets, STR variability has been explored by means of both mean expected heterozygosity and mean allele length variance per locus (Table 1). The latter when compared across populations may give an estimate of the relative effective population sizes.

In the basic set, we found levels of heterozygosity that were very similar across all the NW African populations (Kruskal-Wallis test, $p=0.828$). Only the Mozabite population seemed to present a slight reduction in heterozygosity with a mean value of 0.763 ± 0.016, although the differences with the other populations were not statistically significant (Mann Whitney's U, $p=0.237$). When compared to the external populations, heterozygosities in overall NW Africa (0.782 ± 0.016) were very similar to those in Iberians (0.780 ± 0.017) and other Europeans (0.784 ± 0.019) (NW African vs. rest of populations, Mann-Whitney's U, $p=0.672$). African Americans, with a heterozygoty of 0.793 ± 0.015, were only slightly more diverse. Estimates in the extended set yielded a very similar pattern.

Variance in allele length per locus displayed the same pattern than heterozygosity: slight reduction in the Mozabites (4.64 ± 1.37) with no statistically significant differences (Mann Whitney's U, $p = 0.605$) across the other NW Africans,

and similar levels of genetic variation (Kruskal-Wallis, p=0.929) in overall NW Africans (5.31 ± 1.62), Europeans (5.08 ± 1.41), and African Americans (5.67 ± 1.90).

STR diversity among populations

Genetic differentiation among the populations of both the extended and basic sets was analysed by computing F_{st} genetic distances and representing them by means of neighbor-joining trees. Genetic distances and their standards errors, which were estimated from 10,000 bootstrap iterations, are presented respectively below and above the diagonal in Table 2 for the basic set and in Table 3 for the extended set. The corresponding neighbor-joining trees are presented in Figures 1 and 2. Numbers along the branches represent the percentage of times that a certain node was found in 10,000 bootstrapped trees. In both trees, the NW African populations cluster together, although with short and not very statistically robust branches among them. Although Saharawis and southern Moroccan Berbers cluster together in the basic analysis (Figure 1) with moderate bootstrap support (55 %), in the extended analysis (Figure 2), Saharawis join the Moroccan Arabs first with a low bootstrap support (31 %). In the basic analysis, the Mozabites stand out from the rest of the NW Africans. As stated above, Mozabites have less internal diversity than other NW African populations. Both facts seem to indicate that they may have differentiated by drift. It should be noted that the most robust branch (77 %) in the tree is that separating NW Africans from Europeans. Within the Iberians, Andalusians and Portuguese are closest to NW Africa. African Americans appear linked to NW Africa through a long branch. In the extended data analysis, we find again little structure among NW Africans although with higher percentages of bootstrap support and a very robust separation from Iberians with a bootstrap value of 97 %. Clearly, the increase in number of loci considered in the analysis has added a significant statistically support to the topologies obtained.

A neighbor-joining tree imposes a bifurcating model onto a distance matrix, which may be inadequate for close populations with significant gene flow among them. Therefore, to assess further the pattern of STR diversity between populations, we also have represented the distance matrix through principal coordinate analysis (Figure 3). The first principal coordinate, which explains 53% of the total variation, separates African populations from non-Africans, and the extreme is represented by the Basques. The second principal coordinate explains 22% of the genetic variance and points to a pattern of genetic differentiation between Mozabites and African Americans. These two

populations present extreme and opposite second principal coordinate scores, whereas the rest of populations show intermediate scores. Together, the first two principal coordinates account for 75% of the variance in the distance matrix.

The genetic picture obtained in the two-dimensional representation (Figure 3) is similar to that displayed in the neighbor-joining topologies: all NW Africans, except for Mozabites, are close to each other and separated from Europeans and African Americans. The only discordance between the principal coordinate analysis and the neighbor-joining topologies is found for the Andalusians. Although they occupy an intermediate position between Africans and Europeans and appear linked to them through a relatively long branch in the neighbor-joining tree, in this new representation of the genetic distance matrix Andalusians hardly depart from the European genetic pool.

The two strongest genetic boundaries in the geographical area comprised by NW Africa and the Iberian Peninsula (see Figure 4) were found to encircle single populations, the Mozabites and the Basques. The third genetic boundary separated NW Africa from the Iberian Peninsula and finally, the fourth genetic boundary encircles the Saharawis.

Apportionment of genetic variance

We also explored how genetic variance is apportioned among different geographic, cultural or linguistic population groups by means of the Analysis of Molecular Variance (AMOVA). Within NW Africa, only 0.36% (p=0.0009) of the genetic variance in the basic data set is found among populations, while the rest is found within them. In the extended data set (which does not contain the Mozabites), the fraction of the genetic variance found among NW African populations dropped to 0.05% (p=0.147). When NW African populations are divided into Moroccan Arabs and non-Arabs (northern central and southern Moroccan Berbers, Saharawis and Mozabites), the genetic variance attributable to this partition is not significantly different from zero (-0.13%, p=0.590 for the basic data set and 0.19%, p=0.633 for the extended data set), which means that it has no genetic significance. When we compared the NW Africans to the Iberian populations, the fraction of genetic variance attributable to the difference between these two groups was 0.54% (p=0.0008) for the basic data set and 0.75% (p=0.001) for the extended data set. These values are comparable to those found

between NW Africans and non-Iberian Europeans (Italians and European Americans, 0.93%, $p<0.0001$) for the basic data set.

DISCUSSION

We have typed a large set of autosomal STR loci in several Berber and Arabic speaking population groups in order to obtain a genetic characterisation of Northwestern Africa. Different analyses have been performed to test several hypotheses about the population history in this geographical region and to compare its genetic composition to that of the surrounding populations.

The most distinctive feature in our results is the substantial genetic similarity found among most of the NW African populations studied. Although they represent different cultural and/or ethnical population groups, there is no clear pattern of genetic differentiation between cultural or ethnic groups. The three Berber-speaking groups studied here cannot be taken as a unique or homogenous cultural group. However, we find low levels of genetic differentiation among most of them. The Mozabites, which are geographically distant to all other populations considered, may be an exception. In this case, they seem to be the most distinctive and genetically isolated population within NW Africa. No significant genetic differences were found between Arabs and non-Arabs (i.e., Berbers and Saharawis). Another significant finding in this study concerns the comparison of the NW African autosomal gene pool to that of different external reference populations. The two neighbor-joining trees, the analysis of principal coordinates, and the analysis of the main genetic boundaries point to the grouping of all the NW African populations in a cluster quite separated and differentiated from those of the Iberian Peninsula and other populations within the European genetic variability. Nonetheless, the difference between the two continents is lower than the differentiation of some specific populations: Basques in Iberia and Mozabites in NW Africa.

The heterozygosities observed within the NW African populations studied are very similar among them and equivalent to those of other populations such as the Europeans and African American used as a reference. As discussed above, the Mozabites present slightly reduced heterozygosities, although no statistically significant differences were found. Heterozigosity in a population can be related to its effective population size and degree of genetic isolation. Given our results, we can conclude that

no reductions of the effective population size are apparent from the microsatellite diversity in the NW African populations and that most of them do not seem to be especially isolated, again with the Mozabite exception. This is in agreement with the history and way of life of the Mozabites who are a religious community that obey a strict rule of the Ibadite tradition and protect its identity with a socially prescribed strong endogamy[29].

The F_{st} values obtained among the NW African populations are low and without a clear structure, as shown by the star-like shape of the neighbor-joining tree, with short internal branches among the NW African populations. This observation agrees with the results derived from the analysis of the apportionment among and within populations of the genetic variance. Within NW Africa, only 0.36% of the genetic variance is found among populations. This is a very small figure, about one order of magnitude less of what was obtained in a previous study with classical polymorphisms[30].

A possible cause for the high observed homogeneity may lie in the specific properties of loci analyzed. We have used in this study several STR loci included in the commercial kits AmpF*l*STR Profiler Plus, AmpF*l*STR Cofiler and AmpF*l*STR Green I. STR loci tend to be included in forensic kits when they present some desirable properties, such as good amplification performance even on degraded and scarce DNA samples, and high heterozygosities. Besides, it may also be the case that some loci included in the commercial kits showed similar allele frequency distributions among "European" subpopulations. The forensic community has long debated how to select for a core set of STRs but low F_{st} values have not been a condition for inclusion. They have been selected by high heterozygosity, which has been shown to be correlated with low F_{st} (Calafell *et al*., in preparation). Thus, it is possible that some of our results (such as the absence of a clear genetic structure among the NW African populations) could be biased by the markers selected as a core set in US forensic genetics and that constitute the commercial kits used. In order to test this possible bias, we computed average F_{st}'s among the NW African populations considered in the extended data set, separately for the STR loci used in the commercial kits and for those STR not included in the kits (see Materials and Methods). The means obtained were, respectively, 0.0112 ± 0.0010 and 0.0104 ± 0.0013, which were not statistically significantly different (Mann-Whitney's U, $p = 0.345$). We can conclude that the lack of differentiation among the Berber and Arabic-speaking populations in NW Africa cannot be due to possible

biases in the commercial STR set used. Moreover, it has to be noted that it has been possible to distinguish clearly between Iberians and NW African populations.

Hence, we may reject a spurious cause for our observation of high genetic homogeneity in NW Africa. On the contrary, our results seem to indicate that there has been high levels of gene flow among the NW African populations, even across language barriers. The most plausible explanation for these observations seem to point that the cultural and ethnic differences among them have been recent and have not hindered gene flow among populations. Moreover, given that the linguistic and cultural differences among Arabic and Berber speakers is not reflected by their genes, it is plausible to consider that the Arabization of NW Africa was only a cultural phenomenon with subsequently little genetic impact.

We have found a quite clear genetic differentiation between Iberians and NW Africans in several of the analysis presented. Both groups of populations seem to belong to two different genetic pools. The strong bootstrap support obtained in the two neigbor-joining trees for the cluster that groups all the NW African populations and separates them from those of Iberian Peninsula adds substantial support to this hypothesis. Furthermore, the same pattern of differentiation is discerned in the third genetic barrier and repeated in the principal coordinate analysis, where the first coordinate separates African from non-African populations. It should be noted that although NW Africans and Iberians show a degree of genetic difference, they are closer to each other than to non-European groups, as depicted by the large genetic distance to African Americans. This observation is compatible with the hypothesis suggested by Bosch *et al*.[30]. NW Africa was peopled in the Upper Paleolithic with anatomically-modern humans roughly at the same time that the same human groups colonised Europe. Subsequently, the ancestors of the Berbers differentiated *in situ* before the Neolithic wave of advance[31], and the Neolithic demic diffusion, which seemed to have a large impact on the genetic make up of Europeans[32], may have had little impact in NW Africa.

We may also wonder to which extent this general genetic differentiation between NW Africa and the Iberian Peninsula is found for all the populations. While some Iberians have genetic distances to NW Africans that are as large as those between other Europeans and NW Africans, others, particularly the Andalusians, seem to be genetically closer and to have received some gene flow from NW Africa. In order to test whether this intermediate position could be due to admixture, we estimated the

possible genetic contribution of NW Africa into the Andalusian population using the triangle method[24]. Assuming that Portuguese, Catalans, and Italians represent the European gene pool into which the NW African genes (represented here by northern central and southern Moroccan Berbers, and Moroccan Arabs) were incorporated, the admixture proportion was estimated at m = 0.317, with 95% bootstrap confidence interval of –0.006 to 0.802. Very similar results were obtained when taking only Iberian populations or only Catalan and Portuguese as European references or a different method, the *my* statistic proposed by Bertorelle and Excoffier[33]. At face value, this result may indicate that NW Africa contributed up to a third of the Andalusian gene pool. However, this admixture estimate carries a large uncertainty and it is not significantly different from zero. Other genetic systems do not seem to show such gene flow. A group of mtDNA sequences (called U6) [34] seems to be exclusive of NW Africa, where it reaches frequencies up to 25%. U6 sequences were found in three out of 54 Portuguese[35] and in two out of 92 Galicians[36], but were absent in a sample of Andalusians[35] and in 162 other Iberians[35,37-38]. Thus, when we combine autosomal, Y chromosome (Bosch *et al*., in preparation) and mtDNA data[34], we may conclude that NW African gene flow into southern Iberian seems to be small.

As expected, Basques appeared as the most differentiated population within the Iberian Peninsula. This result is consistent with the postulated pre-Neolithic differentiation proposed for them in a wide variety of studies that present them as a genetic isolate within the European genetic variability[37, 39-41]. Some authors[15-16] studied HLA variation, and suggested, based on similar frequencies for some haplotypes, that Basques and Berbers had a common origin. This conclusion was challenged by new data and a proper numerical analysis[42]. The present results, based on up to 21 autosomal STRs, argue against this hypothetic genetic similarity; caution has to be taken when making population inferences based on a single locus.

The NW Africans have shorter distances to the African Americans than Europeans do, but it is unclear what we can infer from that about gene flow from Sub-Saharan Africa. The answer lies in typing the appropriate Western African samples. This work, plus typing additional NW African populations, is needed to achieve a more detailed reconstruction of the population history of NW Africa.

ACKNOWLEDGEMENTS

We thank all the blood donors that made this study possible. This research was supported by Dirección General de Investigación Científica y Técnica in Spain (PB95-0267-CO2-01) and by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009). Comissionat per a Universitats i Recerca, Generalitat de Catalunya supported E.B. (FI/96-1153) and Spanish Ministry of Education and Science supported R.M.-A. (AP96). The Mozabite samples from Ghardaia were obtained in the framework of an INSERM network: 490NS1 under the coordination of Pr. M.S. Isaad (Algiers) and with the help of the Medical team of Ghardaia. The help of J.M. Dugoujon and A. Sevin, CNR Toulouse is gratefully acknowledged. We especially thank all persons involved in reaching the Saharawi donors as well as Elisabeth Pintado (Sevilla), Josep Lluís Fernández Roure and Alba Bosch (Mataró) for their help in contacting Moroccan donors. We also thank the technical assistance offered by the Unitat de Seqüenciació, Servei Científico-Tècnic, Universitat de Barcelona. We appreciate some comments about our work by Michael Krawczak.

REFERENCES

1 Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet* 1991; **49**: 746-756.

2 Gyapay G, Morissette J, Vignal A, *et al*. The 1993-1994 Généthon human genetic linkage map. *Nature Genetics* 1994; **7**: 246-339.

3 Dib C, Faure S, Fizames C, *et al*. A comprehensive genetic map of the human genome based on 5,624 microsatellites. *Nature* 1996; **380**: 152-154.

4 Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R. Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet* 1994; **55**: 175-189.

5 Urquhart A, Kimpton CP, Downes TJ, Gill P. Variation in short tandem repeat sequences-a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med* 1994; **107**: 13-20.

6 Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 1994; **368**: 455-457.

7 Deka R, Shriver MD, Yu LM *et al*. Population genetics of dinucleotide (dC-dA)n (dG-dT)n polymorphisms in World populations. *Am J Hum Genet* 1995; **56**: 461-474.

8 Jorde LB, Rogers AR, Bamshad MJ, *et al*. Microsatellite diversity and the demographic history of modern humans. *Proc Natl Acad Sci* 1997; **94**: 3100-3103.

9 Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J. Microsatellite variation and the differentiation of modern humans. *Hum Genet* 1997; **99**: 1-7.

10 Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK. Short tandem repeat polymorphism in humans. *Eur J Hum Genet*. 1998; **6**: 38-49.

11 Ruhlen M.A guide to the world's languages. 2nd ed. Stanford, Stanford University Press, 1991.

12 Sánchez-Albornoz C. El Islam de España y el occidente. Spoleto, Centro italiano di studi sull'alto medioevo, 1965.

13 Guichard P. Structures sociales orientales et occidentales dans l'Espagne musulmane. Paris, Mouton, 1977.

14 Izaabel H, Garchon HJ, Caillat-Zucman S, *et al*. A. HLA class II DNA polymorphism in Moroccan population from Souss, Agadir area. *Tissue Antigens* 1998; **51:** 106-110.

15 Arnáiz-Villena A, Benmamar D, Alvarez M, *et al*. HLA allele and haplotype frequencies in Algerians. Relatedness to Spaniards and Basques. *Hum Immunol* 1995; **43**: 259-268.

16 Arnaiz-Villena A, Martinez-Laso J, Gomez-Casado E, *et al*. Relatedness among Basques, Portuguese, Spaniards, and Algerians studied by HLA allelic frequencies and haplotypes. *Immunogenetics* 1997; **47:** 37-43.

17 Pérez-Lezaun A, Calafell F, Clarimón J, *et al.* Allele Frequencies of 13 Short Tandem Repeats in Population Samples of the Iberian Peninsula and Northern Africa. *Int J Legal Med* In press.

18 Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Bertranpetit J. Allele frequencies for 20 microsatellite loci in a worldwide population survey. *Hum Hered* 1997; **47:** 189-196

19 Garofano L, Pizzamiglio M, Vecchio C, *et al.* Italian population data in thirteen short tandem repeat loci: HUMTHO1, D21S11, D18S51, HUMVWFA31, HUMFIBRA, D8S1179, HUMTPOX, HUMCSF1PO, D16S539, D7S820, D13S317, D5S818, D3S1358. *For Sci Int* 1998; **97:** 53-60.

20 Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short term genetic distance. *Genetics* 1983; **105:** 767-779.

21 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4:** 406-425.

22 Efron B. The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA, Society for Industrial and Applied Mathematics, 1982.

23 Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985; **39:** 783-791.

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

25 Brassel KE, Reif D. A procedure to generate Thiessen polygons. *Geogr Anal* 1979; **11:** 289-303.

26 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.

27 Schneider S, Kueffer J-M, Roessli D, Excoffier L. Arlequin ver 1.1: A software for population genetic data analysis. Switzerland, Genetics and Biometry Laboratory, University of Geneva, 1997

28 Sokal RR, Rohlf FJ. Biometry. New York, Freeman and co., 1995.

29 Khaldoun I. Histoire des Berbères et des dynasties musulmanes de l'Afrique septentrionale. Alger, Librairie orientaliste Paul Geuthner Ed., 1956.

30 Bosch E, Calafell F, Pérez-Lezaun A, Comas D, Mateu E, Bertranpetit J. A population history of Northern Africa: evidence from classical genetic markers. *Hum Biol* 1997; **69:** 295-311.

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

32 Menozzi P, Piazza A, Cavalli-Sforza LL. Synthetic maps of human gene frequencies in Europeans. *Science* 1978; **201:** 786-792.

33 Bertorelle G, Excoffier L. Inferring Admixture Proportions from Molecular Data. *Mol Biol Evol* 1998; **15:** 1298-1311.

34 Rando JC, Pinto F, Gonzalez AM, Hernández M, *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.

35 Côrte-Real HBSM, Macaulay V, Richards MB, *et al.* Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 1996; **60:** 331-350.

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

37 Bertranpetit J, Sala J, Calafell F, Underhill P, Moral P, Comas D. Human mitochondrial DNA variation and the origin of the Basques. *Ann Hum Genet* 1995; **59:** 63-81.

38 Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM. Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrialDNA sequences. *Ann Hum Genet* 1996; **60:** 321-330.

39 Bertranpetit J, Cavalli-Sforza LL. A genetic reconstruction of the history of the population of the Iberian Peninsula. *Ann Hum Genet* 1991; **55:** 51-67.

40 Calafell F, Bertranpetit J. Principal component analysis of gene frequencies and the origin of Basques. *Am J Phys Anthrop* 1994; **93:** 201-215.

41 Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bertranpetit J. HLA evidence for the lack of genetic heterogeneity in Basques. *Ann Hum Genet* 1998; **62:** 123-132.

42 Comas D, Mateu E, Calafell F, *et al*. HLA class I and class II DNA typing and the origin of Basques. *Tissue Antigens* 1998; **51:** 30-40.

Table 1. Mean expected heterozygosity and allele length variance per locus (± standard error) for the populations in the basic and extended data sets.

	Mean expected heterozygosity		Mean allele length variance	
	basic	extended	basic	extended
ARA	0.790 ± 0.016	0.758 ± 0.015	5.74 ± 1.76	4.20 ± 1.10
SAH	0.782 ± 0.019	0.753 ± 0.015	5.30 ± 1.59	3.87 ± 0.99
SOB	0.788 ± 0.017	0.761 ± 0.013	5.44 ± 1.63	3.97 ± 1.01
NCB	0.786 ± 0.017	0.745 ± 0.019	5.42 ± 1.73	3.90 ± 1.08
MZA	0.763 ± 0.016	_	4.64 ± 1.37	_
CAT	0.782 ± 0.017	0.746 ± 0.016	5.01 ± 1.43	3.58 ± 0.90
BAS	0.775 ± 0.018	0.736 ± 0.016	5.72 ± 1.62	3.96 ± 1.02
POR	0.782 ± 0.019	_	5.05 ± 1.45	_
AND	0.779 ± 0.015	_	4.54 ± 1.21	_
ITA	0.784 ± 0.019	_	5.07 ± 1.37	_
EAM	0.779 ± 0.022	_	5.04 ± 1.62	_
AAM	0.793 ± 0.015	_	5.64 ± 1.90	_

Abbreviations: ARA, Moroccan Arabs; SAH, Saharawis; SOB, southern Moroccan Berbers, NCB, northern central Moroccan Berbers, MZA, Mozabites; CAT, Catalans; BAS, Basques; POR, Portuguese; AND, Andalusians; AAM, African Americans; EAM, European Americans; and ITA, Italians.

Table 2. Genetic distance matrix and standard errors for the populations used in the basic data set.

	ARA	SAH	SOB	NCB	MZA	CAT	BAS	POR	AND	AAM	EAM	ITA
ARA		.0028	.0012	.0012	.0022	.0034	.0052	.0032	.0043	.0031	.0060	.0037
SAH	.0118		.0018	.0025	.0055	.0033	.0049	.0030	.0037	.0021	.0050	.0034
SOB	.0093	.0105		.0010	.0040	.0037	.0041	.0036	.0037	.0023	.0054	.0031
NCB	.0102	.0136	.0067		.0027	.0023	.0048	.0021	.0022	.0031	.0038	.0018
MZA	.0170	.0219	.0233	.0145		.0056	.0059	.0031	.0026	.0064	.0053	.0036
CAT	.0180	.0216	.0163	.0132	.0270		.0036	.0017	.0018	.0045	.0008	.0008
BAS	.0190	.0282	.0211	.0207	.0309	.0160		.0034	.0042	.0083	.0026	.0028
POR	.0146	.0178	.0150	.0131	.0221	.0096	.0156		.0027	.0032	.0013	.0012
AND	.0156	.0187	.0160	.0128	.0225	.0127	.0214	.0121		.0058	.0015	.0014
AAM	.0176	.0217	.0184	.0150	.0259	.0245	.0367	.0210	.0243		.0067	.0061
EAM	.0201	.0212	.0187	.0169	.0286	.0054	.0163	.0074	.0129	.0292		.0005
ITA	.0142	.0171	.0133	.0110	.0200	.0050	.0124	.0069	.0112	.0240	.0029	

Below diagonal, F_{st} distances; above diagonal standards errors estimated after 10,000 bootstraps iterations. Abbreviations as in Table 1

Table 3. Genetic distance matrix and standard errors for the populations used in the extended data set.

	ARA	SAH	SOB	NCB	CAT	BAS
ARA		.0018	.0011	.0018	.0024	.0038
SAH	.0102		.0012	.0019	.0025	.0038
SOB	.0087	.0091		.0019	.0025	.0034
NCB	.0135	.0123	.0109		.0028	.0053
CAT	.0163	.0177	.0145	.0151		.0023
BAS	.0186	.0231	.0208	.0228	.0129	

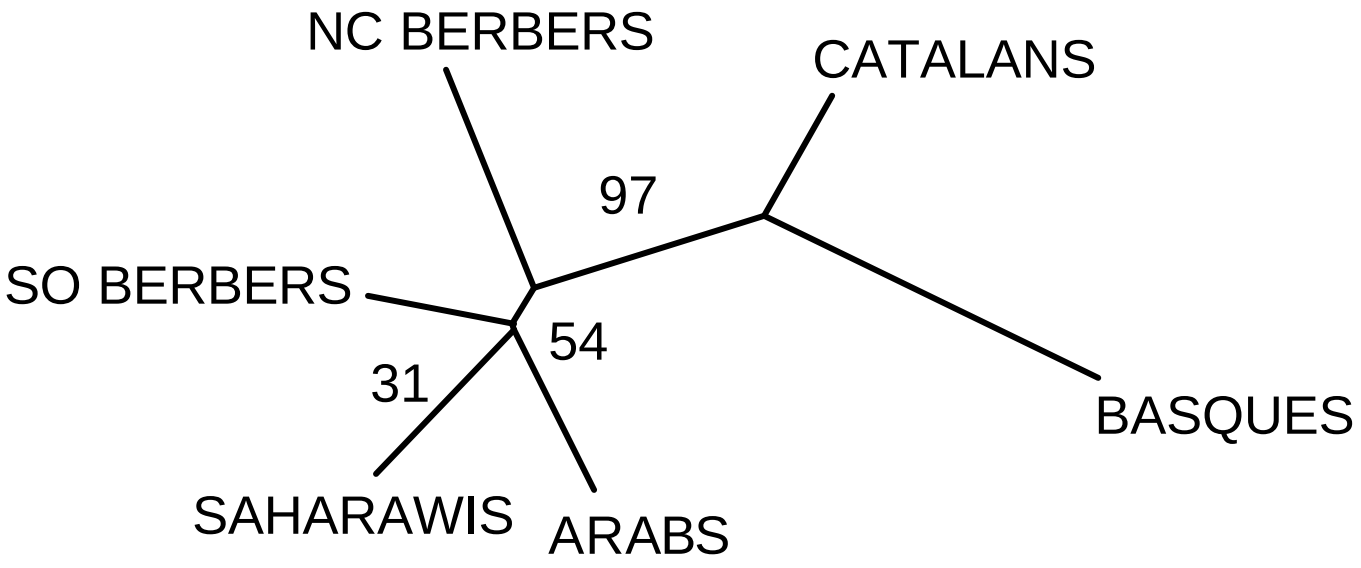
Below diagonal, F_{st} distances; above diagonal standards errors estimated after 10,000 bootstraps iterations. Abbreviations as in Table 1.

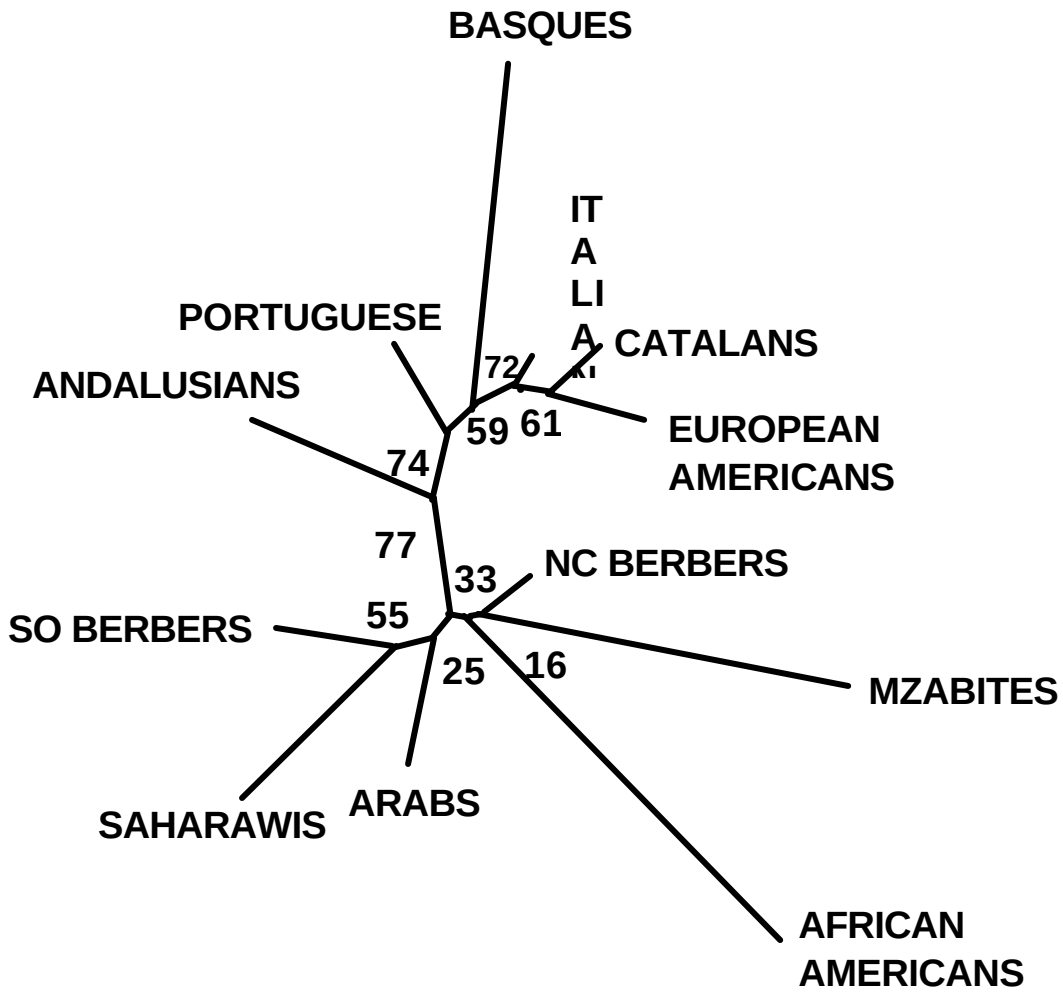
Figure 1. Neighbor-joining tree based on F_{st} genetic distance for the populations and loci included in the basic data set. The figures along the tree branches represent the percentage of times that a certain branch is found in 10,000 bootstrapped trees. Abbreviations: Arabs, Moroccan Arabs; NC Berbers, northern central Moroccan Berbers and SO Berbers, southern Moroccan Berbers.

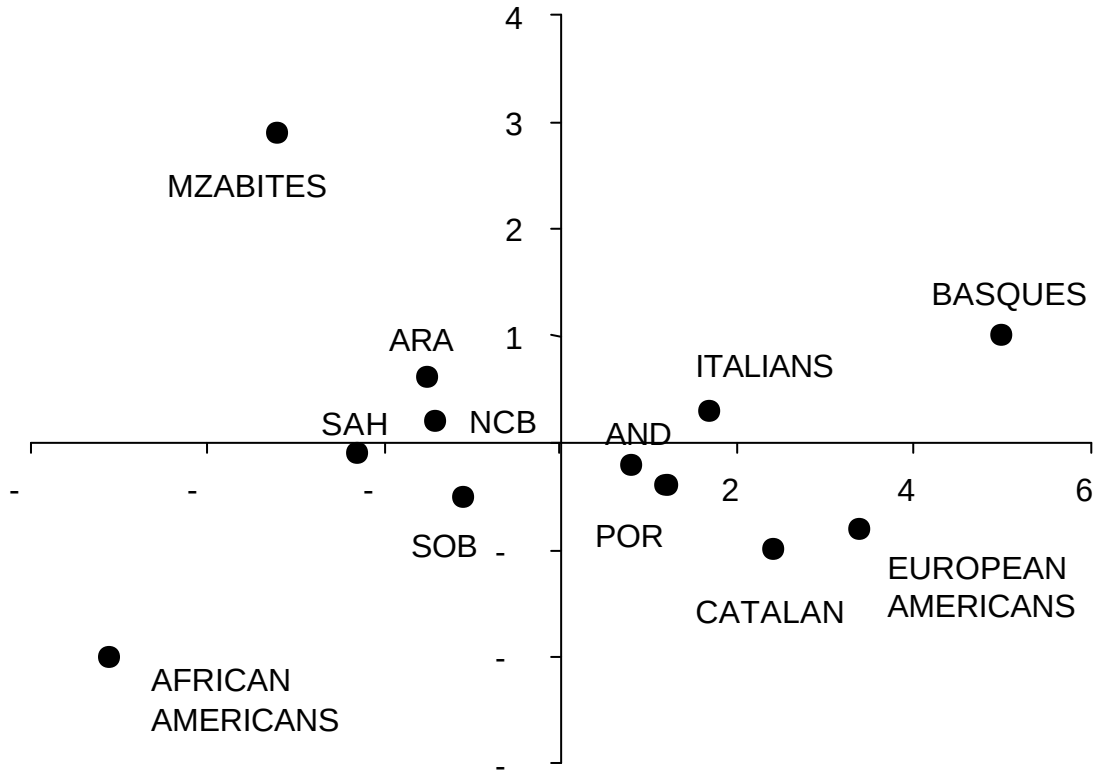
Figure 2. Neighbor-joining tree based on F_{st} genetic distance for the populations and loci included in the extended data set. The figures along the tree branches represent the percentage of times that a certain branch is found in 10,000 bootstrapped trees. Abbreviations as in Figure 1.

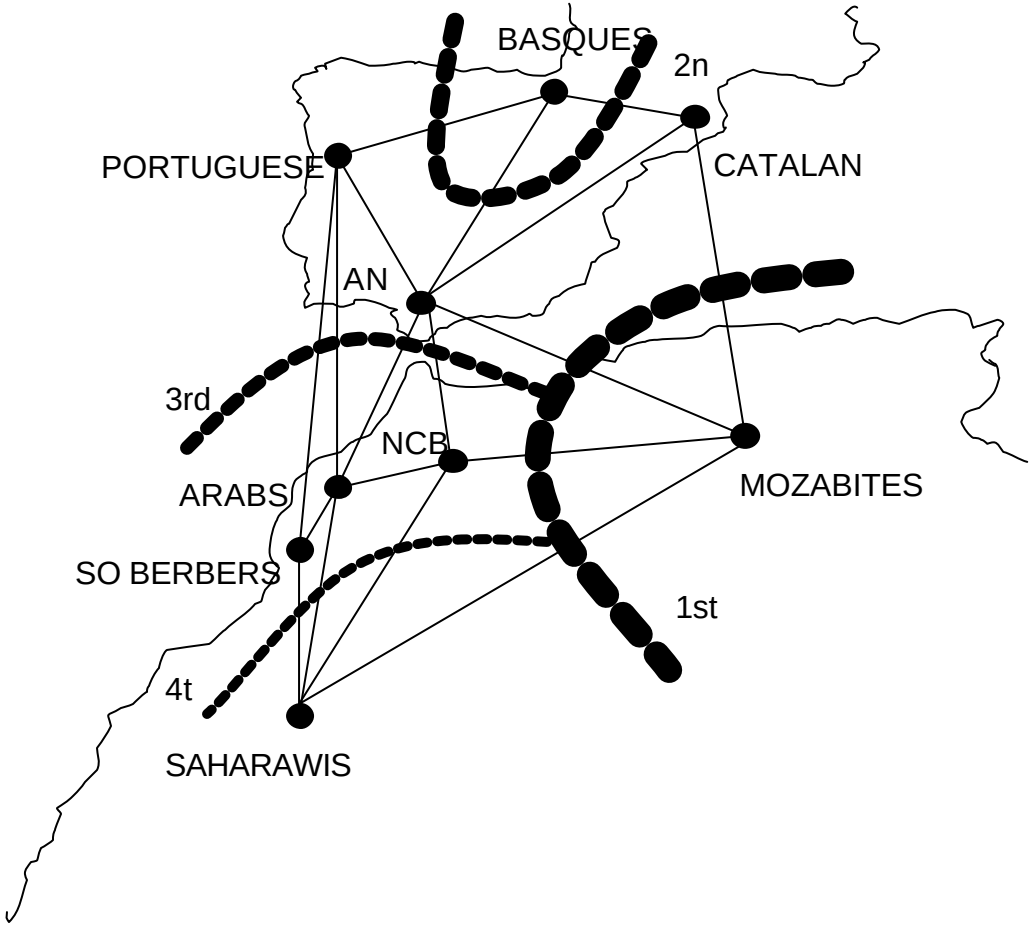
Figure 3. Representation of the two first principal coordinate scores derived from the analysis of the extended F_{st} genetic distance matrix. All score values on the coordinates have been multiplied by a factor of 10^5. Abbreviations: ARA, Moroccan Arabs; NCB, northern central Moroccan Berbers; SOB, southern Moroccan Berbers; SAH, Saharawis; AND, Andalusians; POR, Portuguese.

Figure 4. Genetic boundaries found by superimposing an F_{st} distance matrix on a Delaunay triangulation. Abbreviations: AND, Andalusians; NCB, northern central Moroccan Berbers and SO Berbers, southern Moroccan Berbers.


NC BERBERS
CATALANS
97
SO BERBERS
54
31
BASQUES
SAHARAWIS
ARABS


BASQUES
IT
A
LI
A
CATALANS
PORTUGUESE
ANDALUSIANS
72
59
61
EUROPEAN
AMERICANS
74
77
NC BERBERS
33
55
SO BERBERS
25
16
MZABITES
SAHARAWIS
ARABS
AFRICAN
AMERICANS


4
3
2
1
-
-
-
2
4
6
-
-
-
MZABITES
BASQUES
ARA
ITALIANS
SAH
NCB
AND
POR
SOB
EUROPEAN
AMERICANS
CATALAN
AFRICAN
AMERICANS


BASQUES
2n
PORTUGUESE
CATALAN
AN
3rd
NCB
ARABS
MOZABITES
SO BERBERS
1st
4t
SAHARAWIS

APPENDIX

Allele frequencies for 21 STR loci in 4-5 NW African populations.

Abbreviations: ARA, Arabs; SAH, Saharawis; SOB, southern Moroccan Berbers; NCB, northern central Moroccan Berbers, and MZA, Mozabites. Alleles present in AmpF*l*STR Profiler Plus™ and AmpF*l*STR Cofiler™ allelic ladders or described in other populations but absent in our samples are not indicated.

D3S1358

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 84	MZA 2N = 88
12	0	0.019	0	0	0
13	0	0	0	0.023	0
14	0.037	0.038	0.043	0.048	0.023
15	0.343	0.269	0.298	0.286	0.284
16	0.241	0.279	0.330	0.286	0.250
17	0.176	0.260	0.170	0.214	0.273
18	0.176	0.125	0.160	0.143	0.170
19	0.019	0.010	0	0	0
20	0.009	0	0	0	0

VWA

Allele	ARA 2N = 160	SAH 2N = 118	SOB 2N = 96	NCB 2N = 122	MZA 2N = 88
14	0.131	0.076	0.167	0.115	0.114
15	0.150	0.203	0.219	0.131	0.102
16	0.319	0.127	0.208	0.279	0.420
17	0.181	0.288	0.240	0.270	0.125
18	0.163	0.186	0.104	0.107	0.182
19	0.044	0.119	0.031	0.090	0.045
20	0.006	0	0.031	0.008	0.011
21	0.006	0	0	0	0

FGA

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 84	MZA 2N = 88
16	0	0	0	0.012	0
17	0	0	0.011	0.012	0
18	0	0.019	0	0.012	0
19	0.065	0.058	0.064	0.035	0.011
19.2	0	0	0	0	0.011
20	0.102	0.087	0.181	0.131	0.045
21	0.176	0.231	0.160	0.143	0.318
22	0.111	0.173	0.160	0.143	0.182
23	0.185	0.212	0.170	0.226	0.159
24	0.176	0.106	0.106	0.155	0.170
25	0.120	0.029	0.064	0.095	0.080
26	0.046	0.019	0.064	0.024	0
27	0.009	0.029	0.011	0.012	0.011
28	0.009	0.019	0	0	0
29	0	0.019	0.011	0	0.011

D8S1179

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 82	MZA 2N = 88
8	0.028	0.029	0.011	0.012	0
9	0.009	0	0	0.012	0
10	0.074	0.096	0.117	0.110	0.205
11	0.093	0.096	0.160	0.159	0.034
12	0.065	0.115	0.074	0.122	0.080
13	0.231	0.202	0.191	0.171	0.148
14	0.259	0.221	0.181	0.195	0.250
15	0.194	0.192	0.202	0.183	0.250
16	0.037	0.048	0.064	0.012	0.034
17	0.009	0	0	0.024	0

D21S11

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 82	MZA 2N = 88
26	0.009	0.010	0	0	0
27	0.056	0.019	0.021	0.012	0.034
28	0.083	0.087	0.085	0.183	0.159
28.2	0	0	0	0.012	0
29	0.176	0.231	0.234	0.245	0.227
29.2	0.009	0	0	0	0
30	0.231	0.231	0.234	0.232	0.159
30.2	0.028	0.001	0.032	0.012	0.045
31	0.056	0.048	0.074	0.049	0.045
31.2	0.102	0.096	0.128	0.085	0.148
32	0	0.019	0.011	0.012	0
32.2	0.120	0.163	0.085	0.049	0.125
33	0	0	0	0	0
33.2	0.083	0.087	0.053	0.073	0.045
34	0.009	0	0	0	0
34.2	0.019	0	0.021	0.012	0
35	0.019	0	0.011	0.012	0
35.2	0	0	0.011	0	0
36	0	0	0	0.012	0.011

D18S51

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 82	MZA 2N = 88
11	0.019	0.010	0.011	0	0
12	0.111	0.087	0.149	0.146	0.125
13	0.157	0.144	0.128	0.110	0.091
13.2	0	0	0	0.012	0
14	0.083	0.154	0.160	0.171	0.193
15	0.130	0.106	0.106	0.085	0.125
16	0.231	0.154	0.170	0.159	0.148
17	0.102	0.144	0.096	0.134	0.205
18	0.046	0.038	0.096	0.110	0.080
19	0.046	0.048	0.021	0.037	0.023
20	0.028	0.077	0.053	0.024	0
21	0.037	0	0	0.012	0.011
22	0	0.019	0.011	0	0
23	0.009	0.019	0	0	0

D5S818

Allele	ARA 2N = 158	SAH 2N = 118	SOB 2N = 96	NCB 2N = 124	MZA 2N = 88
8	0.082	0.059	0.021	0.024	0.023
9	0.051	0	0.052	0.016	0
10	0.114	0.059	0.052	0.065	0.125
11	0.259	0.263	0.313	0.282	0.159
12	0.329	0.492	0.344	0.363	0.489
13	0.152	0.119	0.208	0.226	0.205
14	0.013	0.008	0.010	0.024	0

D13S317

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 82	MZA 2N = 88
8	0.111	0.058	0.085	0.073	0.080
9	0.028	0.048	0.011	0.061	0.034
10	0	0.077	0.032	0.024	0.045
11	0.380	0.288	0.330	0.317	0.273
12	0.278	0.250	0.298	0.403	0.443
13	0.139	0.221	0.138	0.073	0.080
14	0.065	0.048	0.106	0.049	0.045
15	0	0.010	0	0	0

D7S820

Allele	ARA 2N = 158	SAH 2N = 118	SOB 2N = 96	NCB 2N = 124	MZA 2N = 88
7	0.025	0	0.010	0	0.011
8	0.146	0.051	0.083	0.113	0.091
9	0.133	0.127	0.063	0.145	0.148
10	0.380	0.373	0.354	0.323	0.364
11	0.133	0.195	0.229	0.226	0.227
12	0.146	0.237	0.240	0.177	0.080
13	0.025	0.008	0.010	0.016	0.080
14	0.013	0.008	0	0	0
15	0	0	0.010	0	0

D16S539

Allele	ARA 2N = 94	SAH 2N = 104	SOB 2N = 92	MZA 2N = 88
8	0.032	0	0.054	0
9	0.138	0.058	0.141	0.159
10	0.043	0.125	0.065	0.091
11	0.309	0.279	0.348	0.205
12	0.191	0.250	0.174	0.239
13	0.234	0.240	0.185	0.273
14	0.053	0.048	0.033	0.034

THO1

Allele	ARA 2N = 160	SAH 2N = 118	SOB 2N = 96	NCB 2N = 128	MZA 2N = 88
5	0.006	0	0	0	0
6	0.163	0.153	0.240	0.227	0.159
7	0.244	0.263	0.219	0.203	0.136
8	0.156	0.161	0.219	0.164	0.125
9	0.325	0.297	0.229	0.266	0.375
9.3	0.063	0.102	0.052	0.125	0.193
10	0.044	0.025	0.031	0.015	0.011
11	0	0	0.010	0	0

TPOX

Allele	ARA 2N = 160	SAH 2N = 118	SOB 2N = 96	NCB 2N = 126	MZA 2N = 88
6	0	0.008	0.001	0.016	0
7	0.019	0.025	0.021	0	0
8	0.450	0.441	0.438	0.397	0.443
9	0.206	0.237	0.177	0.151	0.159
10	0.113	0.110	0.073	0.135	0.114
11	0.181	0.153	0.260	0.301	0.273
12	0.031	0.025	0.021	0	0.011

CSF1PO

Allele	ARA 2N = 108	SAH 2N = 104	SOB 2N = 94	NCB 2N = 84	MZA 2N = 88
7	0.009	0	0.021	0	0
8	0.019	0.010	0	0.024	0.068
9	0	0.019	0.021	0.024	0
10	0.306	0.327	0.340	0.345	0.227
11	0.324	0.279	0.266	0.321	0.364
12	0.287	0.337	0.298	0.238	0.284
13	0.056	0.029	0.043	0.048	0.034
14	0	0	0.011	0	0.023

D11S2010

Allele	ARA 2N = 152	SAH 2N = 118	SOB 2N = 94	NCB 2N = 54
9	0.013	0.017	0.021	0
10	0.007	0.034	0.064	0.055
11	0.599	0.492	0.511	0.463
12	0.158	0.186	0.181	0.185
13	0.125	0.195	0.149	0.204
14	0.099	0.068	0.074	0.093
15	0	0.008	0	0

D13S767

Allele	ARA 2N = 148	SAH 2N = 118	SOB 2N = 84	NCB 2N = 50
9	0.014	0	0	0
10	0.264	0.220	0.214	0.320
11	0.453	0.500	0.452	0.520
12	0.236	0.220	0.274	0.160
13	0.034	0.034	0.048	0
14	0	0.026	0	0
15	0	0	0.012	0

D14S306

Allele	ARA 2N = 152	SAH 2N = 118	SOB 2N = 94	NCB 2N = 52
9	0.013	0	0.011	0.019
10	0.007	0.034	0.011	0.019
11	0.072	0.051	0.096	0.115
12	0.283	0.339	0.298	0.385
13	0.178	0.161	0.181	0.154
14	0.309	0.271	0.330	0.173
15	0.112	0.136	0.064	0.135
16	0.026	0.008	0.011	0

D18S848

Allele	ARA 2N = 152	SAH 2N = 116	SOB 2N = 92	NCB 2N = 54
6	0.066	0.017	0.076	0
7	0.007	0	0.022	0
8	0.086	0.069	0.043	0.037
9	0.237	0.259	0.228	0.185
10	0.500	0.526	0.500	0.685
11	0.105	0.112	0.130	0.093
12	0	0.009	0	0
13	0	0.009	0	0

D2S1328

Allele	ARA 2N = 150	SAH 2N = 118	SOB 2N = 94	NCB 2N = 50
8	0.153	0.178	0.106	0.180
9	0.073	0.025	0.021	0.020
10	0.087	0.093	0.117	0.100
11	0.313	0.347	0.245	0.360
12	0.247	0.271	0.351	0.240
13	0.100	0.068	0.106	0.080
14	0.020	0.017	0.053	0.020
15	0.007	0	0	0

D4S243

Allele	ARA 2N = 148	SAH 2N = 116	SOB 2N = 94	NCB 2N = 50
10	0.020	0.043	0.053	0.020
11	0.324	0.397	0.383	0.420
12	0.338	0.207	0.223	0.260
13	0.209	0.147	0.149	0.160
14	0.041	0.060	0.096	0.020
15	0.027	0.017	0.032	0.040
16	0.020	0.086	0.032	0.080
17	0.020	0.009	0.011	0
18	0	0.026	0.021	0
19	0	0.009	0	0

F13A1

Allele	ARA 2N = 150	SAH 2N = 118	SOB 2N = 94	NCB 2N = 52
3.2	0.153	0.169	0.117	0.173
4	0.093	0.068	0.096	0.058
5	0.260	0.271	0.309	0.365
6	0.153	0.110	0.138	0.135
7	0.260	0.347	0.277	0.192
8	0.040	0.025	0.032	0.058
9	0	0	0	0
10	0.007	0	0.011	0
11	0.007	0	0.011	0
12	0	0	0	0
13	0	0	0	0
14	0.007	0	0	0
15	0	0.008	0.011	0
16	0.013	0	0	0.019
17	0.007	0	0	0

FES/FPS

Allele	ARA 2N = 152	SAH 2N = 118	SOB 2N = 96	NCB 2N = 54
8	0.026	0.042	0.021	0
9	0.007	0.008	0	0
10	0.408	0.314	0.292	0.241
11	0.303	0.331	0.281	0.370
12	0.237	0.263	0.323	0.315
13	0.020	0.034	0.063	0.074
14	0	0.008	0.021	0

D9S926

Allele	ARA 2N = 152	SAH 2N = 118	SOB 2N = 94	NCB 2N = 54
7	0.020	0	0.011	0
8	0	0	0	0
9	0	0	0	0
10	0.026	0.034	0.011	0.018
11	0.322	0.280	0.362	0.278
12	0.336	0.424	0.351	0.426
13	0.230	0.229	0.191	0.185
14	0.053	0.034	0.064	0.093
15	0.007	0	0.011	0
16	0.007	0	0	0

CAPÍTOL III

Y chromosome STR haplotypes in four populations from northwestern Africa

Elena Bosch, Francesc Calafell, Anna Pérez-Lezaun, David Comas, Hassan Izaabel, Omar Akhayat, Abdelaziz Sefiani, Ghania Hariti, Jean-Michel Dugoujon i Jaume Bertranpetit

International Journal of Legal Medicine (en premsa)

Y CHROMOSOME STR HAPLOTYPES IN FOUR POPULATIONS FROM NORTHWESTERN AFRICA

E. Bosch • F. Calafell • A. Pérez-Lezaun • D. Comas • H. Izaabel • O. Akhayat • A. Sefiani • G. Hariti • J.M. Dugoujon • J. Bertranpetit

E. Bosch • F. Calafell • A. Pérez-Lezaun • D. Comas • J. Bertranpetit ($^{H}_{T}$)

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

e-mail: jaume.bertranpetit@cexs.upf.es

Tel: (+34) 93 542 28 40 Fax: (+34) 93 542 28 02

H. Izaabel • O. Akhayat

Laboratoire de Biologie Cellulaire et Moléculaire, Faculté des Sciences, Université Ibnou Zohr, Agadir, Morocco.

A. Sefiani

Institut National d'Hygiène, Rabat, Morocco.

G. Hariti

Hôpital Mustapha, CHU Alger Centre, Alger, Algeria.

J.M. Dugoujon

CNRS ERS 1590, CHU Purpan, Toulouse, France.

ABSTRACT

Eight short tandem repeat (STR) polymorphisms mapping on the male-specific region of the human Y chromosome (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) have been typed in four populations from Northwestern Africa (Moroccan Arabs, Southern Moroccan Berbers, Saharawis and Mozabites). Allele frequency distributions showed statistically significant differences for all loci among all the populations except for DYS19. Complete typings were obtained for 185 chromosomes, which showed 74 different haplotypes. The two most frequent haplotypes were found in 16.2 % and 15.1 % of the individuals, although the latter was almost exclusively found in the Mozabites. Locus and haplotype informativeness were measured by means of the gene diversity (*D*). Haplotype diversity ranged from 0.856 (Mozabites) to 0.967 (Southern Moroccan Berbers). For some loci, allele frequencies in NW Africans were clearly different from those in Europeans. The most common NW African haplotype was found only in one individual out of a total of 494 Europeans typed for the whole STR set. Thus, NW African and European Y chromosomes are clearly differentiated.

KEY WORDS: Y chromosome • Short tandem repeats (STRs) • haplotypes • Northwest African populations

INTRODUCTION

Because of its sex-determining function, the human Y chromosome is paternally inherited and does not recombine on most of its length. Its genetic content is not homologous to any other part of the genome except for a few housekeeping genes, that are homologous but highly divergent from genes in the X chromosome (Lahn and Page, 1997). These special genetic features make the Y chromosome an important genetic tool for the study of male aspects of human evolution and population diversity, as well as in forensic practice and paternity testing (Jobling and Tyler-Smith 1995; Jobling *et al*. 1997). Different types of genetic markers such as biallellic polymorphisms, short tandem repeat (STR) polymorphisms (also called microsatellites) and the MSY1 minisatellite (Jobling *et al*. 1998; Bouzekri *et al*. 1998) are now available on the non-recombining (or male-specific) region of the human Y chromosome. Within them, the PCR-amplificable Y-STR loci have become some of the most useful markers to identify male individuals in forensic applications. In a multicenter study, a basic set of seven STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) has been recommended for standard Y-haplotyping in forensic and paternity casework (Kayser *et al*. 1997; de Knijff *et al*. 1997). Currently, one of the major goals of the forensic geneticist community is the compilation of data for this selected basic set in a number of different reference populations. This will allow to obtain a well-established ethnic and geographic stratification of Y chromosome STR allele and haplotype frequencies as is needed in the routine forensic casework.

Previous genetic studies of forensic applicability in NW Africa include a set of 13 autosomal STRs in Moroccan Arabs and Berbers (Pérez-Lezaun *et al*. 1999a), HLA class II in Southern Moroccan Berbers (Izaabel *et al*. 1998), plus mtDNA sequences in Mozabites (Côrte-Real *et al*. 1996) and in a number of NW African populations (Rando *et al*. 1999). Three autosomal STRs (Meyer *et al*. 1995) and four Y chromosome STRs (DYS19, DYS389I, DYS389II and DYS390) were typed in a sample of Moroccans living in Brussels, Belgium (Kayser *et al*. 1997); in the same sample a detailed study of DYS389 haplotypes was also performed (Rolf *et al*. 1998).

This study presents allele and haplotype frequencies for eight Y specific STRs in four populations from Northwestern Africa (Moroccan Arabs, Southern Moroccan Berbers, Saharawis, and Mozabites), which represent four distinct ethnical, linguistic and cultural groups in the area. These populations are of interest not only to the local forensic geneticists, but to those working in Western Europe, where many immigrants from NW Africa have settled recently.

MATERIAL AND METHODS

Samples

Genetic analyses were performed in a sample of Y chromosomes from 185 unrelated healthy men from Northwestern Africa. Appropriate informed consent was obtained from all participants in this study and, in most cases, information about geographic origin of their four grandparents and maternal tongue was recorded. Samples include 44 Moroccan Arabs, 44 Southern Moroccan Berbers, 29 Saharawis (Western Sahara), and 68 Mozabites from the town of Ghardaia (Algeria). DNA was extracted from fresh blood by standard phenol-chloroform protocols.

STR polymorphism typing

PCR amplifications for all loci (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) were carried out in a 10 µl final reaction volume containing 100 ng of genomic DNA, 50 mM KCl, 10 mM tris-HCl (pH 8.3), 1.5 mM MgCl2 (2.5 mM for DYS19), 250 µM dNTP, 0.2 µM each primer and 1 U Taq DNA Polymerase (GIBCOBRL) using a Perkin Elmer 9600 thermal cycler. Forward primers were fluorescently labeled. PCR cycling conditions were as described in Pérez-Lezaun *et al*. (1999b). PCR products were run in an ABI 377™ sequencer. ABI GS500 TAMRA was used as internal lane standard. The GeneScan 672™ and Genotyper 1.1™ software packages were used to collect the data and analyse fragment sizes. Y chromosome STR alleles were named according to the number of repeat units they contain, which was established through the use of sequenced allele ladders and reference samples kindly provided by P. de Knijff.

Statistical analysis

Population differentiation was tested through an exact test (Raymond and Rousset 1995) as implemented in the Arlequin package (Schneider *et al*. 1997). Gene diversity (*D,* which corresponds to expected heterozygosity for autosomal loci) was computed for each locus as $D = 1 - S\ p_i^2$, where p_i are the allelic frequencies. Haplotype diversity was computed with the same equation, using haplotype frequencies instead of allele frequencies. In the Y chromosome, haplotype diversity, power of discrimination and change of paternity exclusion are numerically identical, and this parameter is sufficient to adequately describe the informativeness of Y chromosome haplotypes in a population.

RESULTS AND DISCUSSION

Allele frequency distributions for loci DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392 and DYS393 in the four NW African populations analysed are shown in Table 1. An exact test of population differentiation (Raymond and Rousset 1995) showed statistically significant differences ($p < 0.0001$ except $p = 0.008$ for DYS388 and $p = 0.020$ for DYS390) for all loci among all the populations except for DYS19 ($p = 0.15$).

Several loci showed allele frequencies that were very different from those found in European populations (Kayser *et al*. 1997). For instance, the most frequent allele for DYS19 is 13 in NW Africa but 14 in Europe. For DYS389I/II, the most frequent alleles in NW Africa are 11/27, whereas in Europe are allele repeats 10/26. It is evident that European Y chromosome STR allele frequencies should not be used in forensic casework involving NW African subjects.

Gene diversity (D) by population for each Y chromosome STR and for the whole haplotype are presented in Table 2. The most informative loci, as measured by their average gene diversities, were DYS390 ($D = 0.637$), DYS389II ($D = 0.527$), DYS389I ($D = 0.496$) and DYS391 ($D = 0.492$). The populations with the highest and lowest haplotypic diversities were the Southern Moroccan Berbers (0.967) and the Mozabites (0.856), respectively.

Y haplotype frequency distributions constructed considering the eight Y STR loci studied (DYS19-DYS388-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393) are shown in Table 3. From a total of 185 chromosomes with complete typings, 74 distinct haplotypes were obtained. The two most frequent haplotypes, 19 (13-12-11-27-24-9-11-13) and 21 (13-12-11-27-24-10-11-13), were found in 30 (12 Moroccan Arabs, 4 Mozabites, 7 Saharawis and 7 Southern Moroccan Berbers) and in 28 individuals (3 Moroccan Arabs, 24 Mozabites, 1 Southern Moroccan Berber) respectively. Fifty-one haplotypes were observed in unique copies. Out of 74 different haplotypes, 61 (82.4%) were found only in one population.

Haplotype 19, which is the most frequent in NW Africa (16.2 %), was found in one of 56 Basques (Pérez-Lezaun *et al*. 1997), but was not found in 33 Catalans (Pérez-Lezaun *et al*. 1997), 125 Central and Southern Italians (Caglià *et al*. 1998) or 280 Finns (Kittles *et al*. 1998). In a sample of 93 Galicians (Pestoni *et al*. 1988) typed for DYS19, DYS389I, DYS389II, DYS390 and DYS393, two individuals matched partially haplotype 19. This is evidence for the specificity of NW African Y chromosome haplotypes.

Since the Y chromosome has one quarter of the effective population size of any autosome, genetic drift acts more strongly on the Y chromosome than on the autosomes (Pérez-Lezaun *et al*. 1997), which explains the high levels of between population differentiation observed in our samples and the reduction of the internal diversity in the Mozabites. This means that i) gene diversities in the Y chromosome STRs are lower than those found in autosomal STRs and it is not likely that other Y chromosome STRs with higher gene diversities will be found and ii) more differentiation is found among populations for Y chromosome STRs than for autosomal STRs, and thus more subpopulations need to be typed to obtain reliable population data for Y chromosome markers. The appropriate reference database must be used for each case.

ACKNOWLEDGEMENTS

We thank all the blood donors that made this study possible. This research was supported by Dirección General de Investigación Científica y Técnica in Spain (PB95-0267-CO2-01) and by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009). Comissionat per a Universitats i Recerca, Generalitat de Catalunya supported E.B. (FI/96-1153). The Mozabite samples from Ghardaia were obtained in the framework of an INSERM network: 490NS1 under the coordination of Dr. A. Cambon-Thomsen (Toulouse) and with the help of the Medical team of Ghardaia and Pr. M.S. Isaad (Alger). The help of A. Cambon-Thomsen and A. Sevin (CNRS Toulouse) is gratefully acknowledged. We also thank all persons involved in reaching the Saharawi donors as well as Elisabeth Pintado (Sevilla), Josep Lluís Fernández Roure and Alba Bosch (Mataró) for their help in contacting Moroccan donors.

REFERENCES

Bouzekri N, Taylor PG, Hammer MF, Jobling MA (1998) Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. Hum Mol Genet 7:655-659

Caglià A, Dobosz M, Boschi I, d'Aloja E, Pascali VL (1998) Increased forensic efficiency of a STR-based Y-specific haplotype by addition of the highly polymorphic DYS385 locus. Int J Legal Med 111:142-146

Còrte-Real HBSM, Macaulay V, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha SS, Bertranpetit J, Sykes B (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331-350

Izaabel H, Garchon HJ, Caillat-Zucman S, Akhayat O, Bach J-F, Sanchez-Mazas A (1998) HLA class II DNA polymorphism in Moroccan population from Souss, Agadir area. Tissue Antigens 51:106-110

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and the human evolution. Trends Genet 11:449-456

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110:118-124

Jobling MA, Bouzekri N, Taylor PG (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet 7:643-653

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Pérez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, de Knijff P, Roewer L (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125-133

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171-1179

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling MA, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Pérez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, Roewer L (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110:134-140

Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. Science 278:675-680

Meyer E, Wiegand P, Brinkmann B (1995) Phenotype differences of STRs in 7 human populations. Int J Legal Med 107:314-322

Pérez-Lezaun A, Calafell F, Seielstad MT, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y chromosome short tandem repeats in humans. J Mol Evol 45:265-270

Pérez-Lezaun A, Calafell F, Clarimón J, Bosch E, Mateu E, Gusmao L, Amorim A, Benchemsi N, Bertranpetit J (1999a) Allele frequencies of 13 short tandem repeats in population samples of the Iberian Peninsula and Northern Africa. Int J Legal Med (in press)

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimón J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999b) Gender-specific migration in Central Asian populations revealed by the analysis of Y-chromosome STRs and mtDNA. Am J Hum Genet 65: 208-219

Pestoni C, Cal ML, Lareu MV, Rodríguez-Calvo MS, Carracedo A (1998) Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). Int J Legal Med 112:15-21

Rando JC, Pinto F, Gonzalez 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

Raymond M, Rousset F (1995) An exact test of differentiation. Evolution 49:1280-1283

Rolf B, Meyer E, Brinkmann B, Knijff de P (1998) Polymorphism at the tetranucleotide repeat locus DYS389 in 10 populations reveals strong geographic clustering. Eur J Hum Genet 6:583-588

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Table 1. Allele frequencies for eight Y chromosome STRs (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) in four NW African populations.

LOCUS	Allele (repeat)	ARA n = 44	SOB n = 44	SAH n = 29	MZA n = 68
DYS19	13	0.592	0.682	0.760	0.822
	14	0.295	0.205	0.172	0.074
	15	0.045	0.045	0.034	0.059
	16	0.045	0.068	0	0.015
	17	0.023	0	0.034	0.015
	18	0	0	0	0.015
DYS388	12	0.795	0.841	0.828	0.926
	13	0.045	0.045	0	0.059
	14	0.023	0	0	0
	15	0	0.045	0	0
	16	0	0.023	0.034	0.015
	17	0.114	0.023	0.138	0
	18	0.023	0.023	0	0
DYS389I	9	0.205	0.114	0.103	0.044
	10	0.205	0.409	0.138	0.162
	11	0.590	0.431	0.759	0.794
	12	0	0.023	0	0
	13	0	0.023	0	0
DYS389II	23	0	0	0	0.015
	24	0.023	0	0	0
	25	0.045	0	0	0.015
	26	0.136	0.295	0.070	0.118
	27	0.728	0.523	0.448	0.764
	28	0.045	0.136	0.448	0.059
	29	0.023	0.023	0.034	0.029
	30	0	0.023	0	0

Table 1. (continued)

LOCUS	Allele (repeat)	ARA n = 44	SOB n = 44	SAH n = 29	MZA n = 68
DYS390	20	0	0	0	0.015
	21	0.068	0.023	0.034	0.029
	22	0	0.068	0.034	0.029
	23	0.318	0.250	0.207	0.132
	24	0.546	0.545	0.310	0.604
	25	0.068	0.114	0.415	0.176
	26	0	0	0	0.015
DYS391	8	0	0.023	0	0
	9	0.409	0.613	0.794	0.147
	10	0.386	0.250	0.034	0.765
	11	0.182	0.114	0.172	0.088
	12	0.023	0	0	0
DYS392	10	0	0.023	0	0
	11	0.932	0.954	0.759	0.956
	12	0	0	0.241	0
	13	0.068	0.023	0	0.044
DYS393	10	0.023	0	0	0
	11	0	0	0	0
	12	0.159	0.114	0.172	0.015
	13	0.750	0.818	0.828	0.941
	14	0.068	0.068	0	0.015
	15	0	0	0	0.029

Abbreviations: ARA, Moroccan Arabs; SOB, Southern Moroccan Berbers; SAH, Saharawis; MZA, Mo

Table 2. Gene diversity (*D*) by population for the eight Y chromosome STRs analysed, as well as for the whole haplotype. The mean gene diversity for each locus is also given. Abbreviations as in Table 1.

Locus	ARA	SOB	SAH	MZA	Mean
DYS19	0.572	0.498	0.406	0.317	0.448
DYS388	0.359	0.294	0.305	0.140	0.275
DYS389I	0.580	0.647	0.409	0.346	0.496
DYS389II	0.458	0.634	0.613	0.403	0.527
DYS390	0.606	0.636	0.712	0.594	0.637
DYS391	0.665	0.560	0.352	0.392	0.492
DYS392	0.130	0.090	0.379	0.086	0.171
DYS393	0.417	0.320	0.296	0.115	0.287
Haplotypic diversity	0.921	0.967	0.867	0.856	

Table 3. Y chromosome STR haplotypes for the four NW African populations analysed. Abbreviations as in Table 1.

	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	ARA	SOB	SAH	MZA	TOTAL
1	13	12	9	25	24	10	11	13				1	1
2	13	12	9	26	23	10	11	10	1				1
3	13	12	9	26	23	10	11	13	2				2
4	13	12	9	26	23	10	11	14		1			1
5	13	12	9	26	24	10	10	13		1			1
6	13	12	9	26	24	10	11	13	1				1
7	13	12	9	26	25	10	11	13		1			1
8	13	12	10	26	23	9	11	13		2			2
9	13	12	10	26	24	9	11	13		2	1	3	6
10	13	12	10	26	24	10	11	13				3	3
11	13	12	10	27	24	9	11	13	1	1			2
12	13	12	10	28	22	9	11	13		2			2
13	13	12	10	28	24	10	11	13		1			1
14	13	12	10	30	22	9	11	13		1			1
15	13	12	11	27	23	8	11	13		1			1
16	13	12	11	27	23	9	11	13		2	1		3
17	13	12	11	27	23	10	11	13				5	5
18	13	12	11	27	23	11	11	13	1				1
19	13	12	11	27	24	9	11	13	12	7	7	4	30
20	13	12	11	27	24	9	11	14		1			1
21	13	12	11	27	24	10	11	13	3	1		24	28
22	13	12	11	27	24	11	11	13	1	1			2
23	13	12	11	27	25	9	11	13	2	1			3
24	13	12	11	27	25	10	11	13				7	7
25	13	12	11	27	25	10	11	15				1	1
26	13	12	11	28	24	9	11	13	1		1		2
27	13	12	11	28	24	10	11	13				4	4
28	13	12	11	28	25	9	11	13			6		6

Table 3. (continued)

	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	ARA	SOB	SAH	MZA	TOTAL
29	13	12	11	28	25	9	12	13			6		6
30	13	12	12	28	23	9	11	13		1			1
31	13	12	13	29	23	9	11	13		1			1
32	13	13	10	26	24	9	11	13		1			1
33	13	13	10	28	25	10	11	13		1			1
34	13	13	11	27	23	9	11	13	1			3	4
35	13	13	11	27	25	10	11	13				1	1
36	14	12	9	24	23	10	11	14	1				1
37	14	12	9	25	24	10	13	13	1				1
38	14	12	9	26	23	10	11	13	1				1
39	14	12	9	26	25	10	11	13				1	1
40	14	12	10	27	24	11	13	13	1				1
41	14	12	10	29	25	10	11	15				1	1
42	14	12	11	27	24	9	11	13	1	4			5
43	14	12	11	27	24	10	11	13	1			1	2
44	14	12	11	29	20	10	11	14				1	1
45	14	14	11	29	25	10	11	12	1				1
46	14	15	9	26	23	10	11	12		1			1
47	14	15	10	26	24	10	11	12		1			1
48	14	16	10	27	22	11	11	12				1	1
49	14	16	10	27	23	11	11	12		1			1
50	14	16	11	29	23	11	11	12			1		1
51	14	17	9	26	23	11	11	12			1		1
52	14	17	10	26	23	11	11	12	1	1			2
53	14	17	10	27	23	10	11	12	1				1
54	14	17	10	27	23	11	11	12	3		3		6
55	14	18	10	27	23	12	11	12	1				1
56	14	18	10	27	24	11	11	12		1			1

Table 3.(continued)

	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	ARA	SOB	SAH	MZA	TOTAL
57	15	12	9	26	22	11	11	13				1	1
58	15	12	9	27	21	10	11	13			1		1
59	15	12	9	27	21	10	11	14	2				2
60	15	12	10	23	23	10	13	13				1	1
61	15	12	10	27	21	11	11	13				1	1
62	15	12	10	27	25	10	11	13		1			1
63	15	12	10	28	24	10	11	13		1			1
64	15	12	11	27	26	11	13	13				1	1
65	16	12	9	26	21	10	11	14		1			1
66	16	12	10	26	25	11	13	13		1			1
67	16	12	10	27	21	11	11	13				1	1
68	16	12	11	27	24	9	11	13		1			1
69	16	12	11	27	24	11	13	13	1				1
70	16	12	11	28	21	10	11	13	1				1
71	17	12	9	27	22	9	12	13			1		1
72	17	12	11	27	24	11	13	13				1	1
73	17	13	10	25	23	10	11	13	1				1
74	18	12	11	27	25	10	11	13				1	1

CAPÍTOL IV

Variation in Short Tandem Repeats is deeply structured by genetic background on the human Y chromosome

Elena Bosch, Francesc Calafell, Fabrício R. Santos, Anna Pérez-Lezaun, David Comas, Noufissa Benchemsi, Chris Tyler-Smith i Jaume Bertranpetit

American Journal of Human Genetics (1999) 65:1623-1638

Am. J. Hum. Genet. 65:1623–1638, 1999

Variation in Short Tandem Repeats Is Deeply Structured by Genetic Background on the Human Y Chromosome

Elena Bosch,[1] Francesc Calafell,[1] Fabrício R. Santos,[2,3] Anna Pérez-Lezaun,[1] David Comas,[1] Noufissa Benchemsi,[4] Chris Tyler-Smith,[2] and Jaume Bertranpetit[1]

[1]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona; [2]Cancer Research Campaign Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, Oxford; [3]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil; and [4]Centre National de Transfusion Sanguine, Rabat, Morocco

Summary

Eleven biallelic polymorphisms and seven short-tandem-repeat (STR) loci mapping on the nonrecombining portion of the human Y chromosome have been typed in men from northwestern Africa. Analysis of the biallelic markers, which represent probable unique events in human evolution, allowed us to characterize the stable backgrounds or haplogroups of Y chromosomes that prevail in this geographic region. Variation in the more rapidly mutating genetic markers (STRs) has been used both to estimate the time to the most recent common ancestor for STR variability within these stable backgrounds and to explore whether STR differentiation among haplogroups still retains information about their phylogeny. When analysis of molecular variance was used to study the apportionment of STR variation among both genetic backgrounds (i.e., those defined by haplogroups) and population backgrounds, we found STR variability to be clearly structured by haplogroups. More than 80% of the genetic variance was found among haplogroups, whereas only 3.72% of the genetic variation could be attributed to differences among populations—that is, genetic variability appears to be much more structured by lineage than by population. This was confirmed when two population samples from the Iberian Peninsula were added to the analysis. The deep structure of the genetic variation in old genealogical units (haplogroups) challenges a population-based perspective in the comprehension of human genome diversity. A population may be better understood as an association of lineages from a deep and population-independent gene genealogy, rather than as a complete evolutionary unit.

Received July 21, 1999; accepted September 8, 1999; electronically published November 23, 1999.

Address for correspondence and reprints: Dr. Jaume Bertranpetit, Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain. E-mail: jaume.bertranpetit@cexs.upf.es

0002-9297/1999/6506-0018$02.00

Introduction

Human evolution and population studies based on the Y chromosome have increased notably during recent years (Hammer 1995; Jobling and Tyler-Smith 1995; Santos et al. 1995*b*; Whitfield et al. 1995; Cooper et al. 1996; Deka et al. 1996; Roewer et al. 1996; de Knijff et al. 1997; Hammer et al. 1997, 1998; Pérez-Lezaun et al. 1997, 1999; Hurles et al. 1998). The singular characteristics of the Y chromosome, which include paternal inheritance and absence of recombination through most of its length, make this chromosome a powerful tool for tracing and comparing paternal lineages of human populations in a way similar to the use of mtDNA to study maternal lineages. In spite of the slow initial discovery of polymorphic markers on the Y chromosome, which started in the mid 1980s (Casanova et al. 1985; Lucotte and Ngo 1985), the detection of variation on this chromosome has grown dramatically during recent years (Jobling and Tyler-Smith 1995; Hammer and Zegura 1996; Underhill et al. 1997). The variety of polymorphic markers now available on the nonrecombining portion of the Y chromosome ranges from base substitutions and deletion/insertion polymorphisms, which are rare (probably even unique) events in evolution and which tend to be biallelic, to faster-mutating polymorphisms such as microsatellites—also known as short tandem repeats (STRs)—and the MSY1 minisatellite (Bouzekri et al. 1998; Jobling et al. 1998*a*).

The presence of different types of polymorphisms with different mutational mechanisms and rates strengthens even more the applicability of analysis of the nonrecombining portion of the Y chromosome to the study of human evolution at different geographic and time scales. Some examples of the wide range of subjects to which the study of Y-chromosome polymorphisms has been applied include investigations of both the origin and the dispersal of anatomically modern humans (Hammer et al. 1997, 1998), the first settlement of the Americas (Pena et al. 1995; Santos et al. 1995*a*, 1996, 1999*a*; Underhill et al. 1996; Karafet et al. 1997), the Asian

paternal contribution to northern European populations (Zerjal et al. 1997), the colonization of Polynesia as well as later European admixture (Hurles et al. 1998), the colonization of mountain habitats in central Asia (Pérez-Lezaun et al. 1999), and the study of both the Cohen (Thomas et al. 1998) and the Jefferson (Foster et al. 1998) lineages.

The evolution of each of these markers on the Y chromosome is not independent, since changes to a particular locus always happen on a well-defined background for all the other polymorphisms to which it remains linked because of the absence of recombination. In particular, the absence of recombination has facilitated the inference of the genealogy of haplogroups—that is, groups of chromosomes defined by the combination of alleles at different biallelic polymorphisms. The combination of slow- and fast-mutating systems has added values. The typing of STRs within haplogroups has allowed the investigation of the origin and dispersal of certain haplogroups (Zerjal et al. 1997; Hurles et al. 1998), as well as examination of the population movements associated with those dispersals.

Genetic diversity is often thought of as being structured by population, to the extent that a population perspective has been thought of as a complete description of the genetic diversity. Classic population-genetics theory modeled the dynamics of allele-frequency change among populations, in terms of forces such as drift, selection, and migration. This approach implies that a mere description of variation is fully informative on genetic grounds. However, genetic diversity may be more deeply structured by gene genealogy than by the ethnogenesis process that gave rise to the population. The gene genealogy of a particular genome region can be recognized from the inferred genealogy of its slowly mutating polymorphisms, which constitute the stable background on which the variation of faster-mutating systems took place. This rapidly produced variation may contain information on the evolution of each deep branch of the gene genealogy and may even include a detailed footprint of the full gene geneology. The population perspective may thus be out of the scale of the genetic processes. For instance, Estivill et al. (1994) found that haplotypes of three STRs in the cystic fibrosis transmembrane conductance regulator (CFTR) gene were clearly different in chromosomes bearing the ΔF508 mutation (which causes cystic fibrosis) compared with those found in nonaffected chromosomes. Moreover, STR haplotypes from any of a number of nonaffected European populations were much closer to each other than were haplotypes from nonaffected and ΔF508-bearing chromosomes belonging to the same population. Therefore, genetic diversity within the CFTR gene in Europe is more deeply structured by genetic background than by population.

Since the production of biallelic variation is continuous in time, some of that variation—the most ancient—will define haplogroups with wide geographic distributions; however, most of it, being more recent, will be found in more-restricted areas and may have occurred after certain population splits. In fact, there are population-specific Y-chromosome biallelic polymorphisms, such as Tat in northwestern Asians (Zerjal et al. 1997), SRY-2627 in north Iberians (Hurles et al. 1999), and DYS199 in Native Americans (Underhill et al. 1996). The most recent biallelic markers may be population-specific and even family-specific. However, a characteristic distribution of haplogroups will be found in each population, and some of those haplogroups may be much older than the populations in which they are found.

In this study, we have scored 11 Y-specific biallelic markers in men from northwestern Africa; this has allowed the characterization, for the first time, of Y-chromosome haplogroup distribution in this area. We have also explored Y-chromosome variation at seven STRs in the same individuals, to characterize genetic variation by stable genetic background (i.e., by biallelic polymorphism haplogroup) and to compare the apportionment of genetic variation by haplogroup with that apportioned by population. This will allow an alternative perspective to the population approach for the comprehension of human genetic diversity. The time to the most recent common ancestor (TMRCA) for the STR variability within haplogroups and the microsatellite diversity within them were analyzed as well. These analyses were also done in two population samples from the northern Iberian Peninsula. Finally, we explored whether STR differentiation by haplogroup retains information about the haplogroup phylogeny.

Material and Methods

Samples

Genetic analyses were performed on a sample of Y chromosomes from 129 unrelated healthy men from northwestern Africa. Appropriate informed consent was obtained from all participants in this study, and, in most cases, information about both the geographic origin and native language of each man's four grandparents was recorded. The samples obtained were from 44 Arabs, 42 Tahelhits, and 14 other Moroccan Berbers, as well as from 29 Saharawis. DNA was extracted from fresh blood by use of standard phenol-chloroform protocols. For parts of the data analysis, additional samples (from 51 Basques and 27 Catalans) from northern Iberia were included.

Biallelic Polymorphism Typing

We typed eight base substitutions, an Alu insertion, the polymorphic number of adenine residues in its 3′ end, and a duplication/deletion polymorphism (see table 1). Polymorphism 92R7, which was originally described as an RFLP by Mathias et al. (1994), was converted to a PCR format by M. E. Hurles, F. R. Santos, and C. Tyler Smith (unpublished data), by amplification of a 55-bp fragment containing the polymorphic site in the 92R7 system. *Hin*dIII digestion was used to detect the C→T base substitution that destroys a *Hin*dIII site equivalent to the presence of the 6.7-kb band in the 92R7 Southern blots (Mathias et al. 1994). SRY-2627 (also known as pSRY-373) was analyzed by PCR amplification, as reported elsewhere (Bianchi et al. 1997), by use of the pSRY244 (forward) and pSRY634 (reverse) primers. *Ban*I digestion was employed to detect the C→T transition at base-pair position 373 (Santos et al. 1999*a*). SRY-1532 screening was performed with primers SRY-1 and SRY-2 (Santos et al. 1999*b*), which amplify a 167-bp male-specific fragment spanning the polymorphic position 10,831 of region SRY (Whitfield et al. 1995; Kwok et al. 1996). PCR and cycling conditions were performed as described elsewhere (Santos et al. 1999*a*). Amplified products were digested with *Dra*III and were incubated with Pronase (Boehringer Mannheim). The Y-chromosome Alu insertion polymorphism (YAP) element at the DYS287 locus was analyzed by PCR amplification, as

Table 1

Y-Chromosome Biallelic Polymorphisms Analyzed and Their Allele Frequencies in Northwestern Africa

Polymorphism[a] and Allele	Frequency	Type	Method	References
92R7:				
C	124 (.961)	Base substitution	*Hin*dIII site loss, PCR	Mathias et al. (1994)
T	5 (.039)			
SRY-2627:				
C	129 (1)	Base substitution	*Ban*I site loss, PCR	Bianchi et al. (1997); Santos et al. (1999*a*)
T	0 (0)			
SRY-1532:				
A	0 (0)	Base substitution	*Dra*III site gain, PCR	Whitfield et al. (1995); Kwok et al. (1996); Santos et al. (1999*a*)
G	129 (1)			
YAP:[b]				
0	24 (.186)	Alu insertion	PCR	Hammer (1994); Hammer and Horai (1995)
1	105 (.814)			
Poly (A) tail:[c]				
L	1 (.010)	Poly (A) tail–length		
S	104 (.990)	polymorphism	PCR	Hammer (1995); Hammer et al. (1997)
SRY-8299:				
G	24 (.186)	Base substitution	*Bsr*BI site loss, PCR	Whitfield et al. (1995); Santos et al. (1999*a*)
A	105 (.814)			
sY81:				
A	123 (.953)	Base substitution	*Nla*III site loss, PCR	Seielstad et al. (1994)
G	6 (.047)			
M9:[d]				
C	123 (.953)	Base substitution	PCR, DHPLC[e]	Underhill et al. (1997)
G	6 (.047)			
12f2:[d]				
10 kb	114 (.884)	Duplication/deletion	*Taq*I (*Eco*RI), filter hybridization	Casanova et al. (1985)
8 kb	15 (.116)			
50f2 P:[d]				
8.5 kb	129 (1)	Base substitution	*Taq*I, filter hybridization	Guellaen et al. (1984)
3.1 kb	0 (0)			
50f2 I:[d]				
8.5 kb	129 (1)	Base substitution	*Taq*I, filter hybridization	Guellaen et al. (1984)
4 kb	0 (0)			

[a] For each biallelic polymorphism, the ancestral state is presented above the derived state.

[b] For the YAP polymorphism, "0" denotes absence of the Alu sequence and "1" denotes presence of the *Alu* sequence.

[c] The poly (A) tail–length polymorphism is found within the YAP element, and four different alleles have been described so far. Of those, we found two—S (small, 26 bp) and L (large, 46 bp)—in our sample. Their frequencies are given with respect to the total number of YAP-positive chromosomes.

[d] Only YAP-negative individuals were tested for these polymorphisms. YAP-positive individuals were presumed to have polymorphisms M9 C, 12f2 10 kb, 50f2 P 8.5 kb, and 50f2 I 8.5 kb.

[e] DHPLC = denaturing high-performance liquid chromatography.

described by Hammer and Horai (1995). Variation in the number of adenine residues at the 3′ end of the YAP Alu sequence, also known as the poly (A) tail–length polymorphism (Hammer 1995; Hammer et al. 1997), was typed by resolution of the amplified products of YAP-positive individuals on 20 × 20–cm (1 mm–thick) 6% polyacrylamide gels in 1 × Tris-borate EDTA at 40 mA for 5 h and by visualization with silver staining. The SRY-8299 system was genotyped with the primers and PCR conditions described by Santos et al. (1999*a*). The amplified fragments, which contained the G→A polymorphic site at position 4,064 of the SRY region (Whitfield et al. 1995), were then digested with *Bsr*BI and were analyzed by electrophoresis (Santos et al. 1999*a*). Polymorphism sY81 (DYS271) was amplified as described elsewhere (Seielstad et al. 1994). Amplified products were digested with *Hsp*92II (isoschizomer of *Nla*III). The M9 C→G base substitution was PCR amplified and was typed by denaturing high-performance liquid chromatography, as described elsewhere (Underhill et al. 1997). A total of 1 μg of genomic DNA from each YAP-negative individual was digested to completion by use of 20 U of *Taq*I (Boehringer Mannheim); it was then electrophoresed on a 1% agarose gel in 0.5 × Tris-acetate EDTA for 14 h at 25 V and was transferred to Hybond™ -N+ nylon membranes, by use of standard procedures (Southern 1975). DNA probes 50f2 (DYS7) and 12f2 (DYS11) for Southern blot analysis were radioactively labeled by random priming (Feinberg and Vogelstein 1983, 1984). After prehybridization with salmon sperm DNA, filters were hybridized overnight with probe 12f2 (Casanova et al. 1985) at 68°C in 1% SDS 5 × Denhardt's solution, 10% dextran sulphate, and 0.5 M sodium phosphate; they were then washed, at 65°C, to a stringency of 0.1 × SSC plus 1% SDS. The filter hybridization and washing conditions used, plus further details for probe 50f2, can be found in a report by Jobling (1994). In both cases, bands were visualized by autoradiography done at −70°C with Fuji Rx film. Biallelic polymorphism data for Basques and Catalans were obtained from Semino et al. (1996), Underhill et al. (1997) and Underhill (unpublished results), and Hurles et al. (1999).

STR Polymorphism Typing

Two trinucleotide repeat polymorphisms—DYS388 and DYS392—and six tetranucleotide repeat polymorphisms—DYS19, DYS389I, DYS389II, DYS390, DYS391, and DYS393—were typed in all Y chromosomes. A PE Biosytems 9600 thermal cycler was used. PCR reactions were performed in a 10-μl final reaction volume containing 100 ng genomic DNA, 50 mM KCl, 10 mM tris-HCl (pH 8.3), 1.5 mM MgCl2 (2.5 mM for DYS19), 250 μM each dNTP, 0.2 μM each primer, and 1 U *Taq* DNA Polymerase (Gibco BRL). Forward primers were fluorescently labeled. The PCR cycling conditions used were those described by Pérez-Lezaun et al. (1999). PCR products were run in an ABI 377™ sequencer. ABI GS500 TAMRA was used as internal lane standard. The GENESCAN 672™ and GENOTYPER 1.1™ software packages were used to collect the data and to analyze fragment sizes. Y-STR alleles were named according to the number of repeat units they contain. The number of repeats was established through the use of sequenced allele ladders and reference samples kindly provided by P. de Knijff. Genome Database primers for the DYS389 locus amplify a partially duplicated region and generate two PCR products, which are referred to as DYS389I (239–263 bp) and DYS389II (353–385 bp). Both fragments are variable in length, and the study of their sequence structure has shown that DYS389II contains DYS389I plus two additional stretches of tetranucleotide repeats (Rolf et al. 1998; Pestoni et al. 1998). Therefore, we have used only the length variability of the shorter fragment in the numerical analysis. The STR haplotypes for northern Iberians were those described by Pérez-Lezaun et al. (1997).

Data Analysis

Analysis of molecular variance (AMOVA) was performed for STR allele frequencies among haplogroups, by use of the Arlequin package (Schneider et al. 1997). A simple hierarchical partitioning of haplotypes in haplogroups was tested, without further pooling haplogroups. Genetic dissimilarity among STR haplotypes was weighed by the difference in allele length, which is equivalent to the R_{ST} measure (Slatkin 1995; Schneider et al. 1997). AMOVA was also performed directly on STR haplotype frequencies; since the results were very similar to those obtained with R_{ST}, only R_{ST} results are given. Genetic diversity within each haplogroup was measured by different parameters, such as the mean number of allele differences between pairs of Y-chromosome haplotypes and the mean number of differences in repeat sizes between pairs of Y-chromosome haplotypes. To test whether these parameters were statistically significantly different between haplogroups, we performed a permutation procedure similar to those described in Graven et al. (1995) and in Mateu et al. (1997). In each iteration, chromosomes are permuted across haplogroups, the relevant parameter is recomputed in both haplogroups, and the difference is recorded. In this way, after 10,000 iterations, a distribution of the difference in the parameter between haplogroups is obtained under the null hypothesis of no difference. The probability of obtaining a difference in

Table 2

Y-Chromosome Haplogroups Studied and Their Frequencies in Northwestern Africa

Biallelic Polymorphism	HG 7	HG 3	HG 2+	HG 26	HG 1	HG 22	HG 15	HG 6	HG 9	HG 4	HG 21 / 21L	HG 8
	Status of Allele											
92R7	0	1	0	0	1	1	0	0	0	0	0	0
SRY-2627	0	0	0	0	0	1	0	0	0	0	0	0
SRY-1532	0	0	1	1	1	1	1	1	1	1	1	1
YAP (poly [A] tail)[a]	−	−	−	−	−	−	−	−	−	+ (S)	+ (S) / + (L)	+ (S)
SRY-8299	0	0	0	0	0	0	0	0	0	0	1	1
sY81	0	0	0	0	0	0	0	0	0	0	0	1
M9	0	1	0	1	1	1	0	0	0	0	0[b]	0[b]
12f2	0	0	0	0	0	0	0	0	1	0	0[b]	0[b]
50f2 P	0	0	0	0	0	0	0	1	0	0	0[b]	0[b]
50f2 I	0	0	0	0	0	0	1	0	0	0	0[b]	0[b]
	No. (Frequency [n = 129])											
	0 (0)	0 (0)	3 (.023)	1 (.008)	5 (.039)	0 (0)	0 (0)	0 (0)	15 (.116)	0 (0)	98 (.759)/1 (.008)	6 (.047)

NOTE.—For each biallelic polymorphism, alleles are represented as "0" or "1," according to the presence of an ancestral or derived state, respectively (see table 1).

[a] Minus sign (−) = YAP-negative; plus sign (+) = YAP-postive; (S) = short poly (A) tail; and (L) = long poly (A) tail.

[b] Assumed but not typed.

the parameter between two haplogroups that was larger than the observed difference was recorded.

TMRCA of the STR variability generated within the chromosomes bearing the derived allele of each biallelic polymorphism was estimated from the mean allele-size variance of the seven STRs within all chromosomes bearing that derived allele by use of equation 5 in Di Rienzo et al. (1998): $V = T\mu\eta_2$, where V is the variance in repeat size, T stands for time in generations after a population expansion, μ is the mutation rate, and η_2 is the variance in the number of repeats gained or lost at each mutation event. The average mutation rate used was 1.2×10^{-3}, with a 95% confidence interval (CI) of 4.6×10^{-4} to 2.8×10^{-3} (Bianchi et al. 1998). This estimate comprises data from deep-rooting pedigrees (Heyer et al. 1997) and for father-son transmissions (Kayser et al. 1997), as well as from family cell lines from the CEPH. It has been reported that microsatellites tend to accumulate mutations in lymphoblastoid cell lines, which would result in an overestimation of germline mutation rates (Banchs et al. 1994). However, Bianchi et al. (1998) did not find any mutations in the CEPH cell lines they typed, and, therefore, this method does not bias their germline-mutation-rate estimate upward. Since all mutations observed (Heyer et al. 1997; Kayser et al. 1997) consist of the gain or loss of one repeat, mutation-size variance (η_2) was set to 1; the generation time used was 20 years. The 95% CIs for TMRCAs were estimated by taking into account both the interlocus sampling variance and the 95% CI of the mutation rate estimate.

Net genetic distances between haplogroups (D_{ij}) were computed, according to the Jensen Difference (Rao 1982), as $D_{ij} = d_{ij} - ([d_{ii} + d_{ij}]/2)$, where d_{ij} is the mean number of absolute differences in allele length between pairs of chromosomes of haplogroups i and j and where d_{ii} and d_{jj} are the mean pairwise allele length differences within haplogroups i and j, respectively. In its version for DNA-sequence data, D_{ij} is the most common measure of genetic distance among population samples of mtDNA sequences (see, for example, Bertranpetit et al. [1995]). It represents the average genetic difference between haplotypes belonging to different groups, corrected for the average variability within those groups.

Results

Y-Chromosome Biallelic Polymorphisms

Allele frequencies of the 11 Y-chromosome biallelic polymorphisms in the total sample of 129 northwestern African men studied are shown in table 1. The combination of their allelic states allowed us to identify 6 of the 12 different haplogroups of Y chromosomes that have been described to date with the same polymorphisms (see table 2). Ancestral and derived alleles for each biallelic polymorphism (92R7 in Jobling 1994 and in Mathias et al. 1994) were inferred or were established by typing of nonhuman primates. These polymorphisms include SRY-2627 (Bianchi et al. 1997; Hurles et al. 1999), SRY-1532 (Whitfield et al. 1995), YAP (Hammer and Horai 1995), poly (A) tail (Hammer et al. 1997), SRY-8299 (Whitfield et al. 1995), sY81 (Seielstad et al. 1994), M9 (Underhill et al. 1997), 12f2 (Casanova et al. 1985), and 50f2 P and 50f2 I (Guellaen et al. 1984). Haplogroup distribution within the northwestern African population

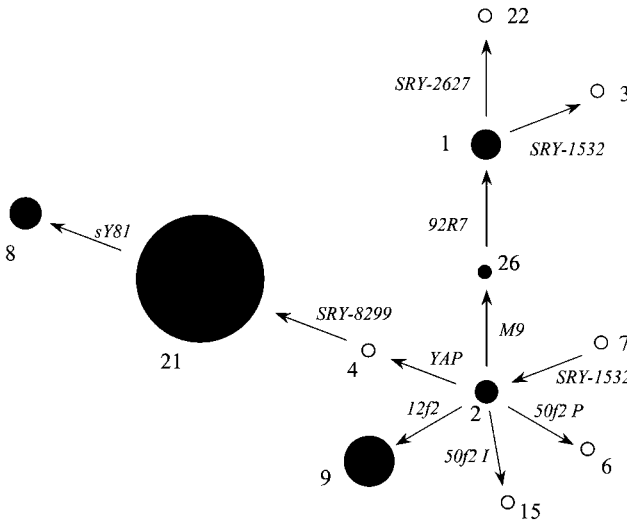


Figure 1 Parsimony network relating the haplogroups defined by 10 of the unique-event polymorphisms analyzed. Arrowheads indicate the derived states, as inferred or established by typing nonhuman primates (Jobling and Tyler-Smith 1995; Jobling et al. 1996, 1997; Hurles et al. 1998). Blackened circles are proportional to haplogroup frequency, and unblackened circles indicate haplogroups detectable but not found in our set of 129 Y chromosomes from northwestern Africa.

studied is also shown in the parsimony network of figure 1, on the basis of other networks (Jobling and Tyler-Smith 1995; Jobling et al. 1996, 1997; Hurles et al. 1998), except for the consideration of the SRY-8299 polymorphism defining haplogroup (HG) 21, which will be described elsewhere (Vogt et al. 1997). All but one of the Y chromosomes carrying the YAP insertion presented the small allele with respect to the poly (A) tail–length polymorphism of the YAP element (Hammer et al. 1997). The exception carried the large allele and belonged to HG 21. Except where indicated, this chromosome has been included in HG 21 in numerical analyses. The main feature of Y-chromosome–haplogroup distribution within northwestern Africa is the high frequency (76.7%) of HG 21. This haplogroup, which is a subset of YAP groups 3 and 4 (Hammer et al. 1997) and which is equal to the YAP groups later called "3A and 4A" (Altheide and Hammer 1997), has previously been found in Europe (14%), Egypt (47%), and sub-Saharan Africa (12%), but it has never been found at such high frequencies (Altheide and Hammer 1997; Hammer et al. 1997). No Y chromosome belonging to HG 4 (which is characterized by the fact that it carries the ancestral allele of SRY-8299 with respect to HG 21) was found; however, this haplogroup, which, according to the nomenclature used by Altheide and Hammer (1997), is equivalent to YAP haplotype 3G, has so far been found to be restricted to central and east Asian populations (Hammer et al. 1998). Six (4.7%) of the 129 Y chromosomes analyzed contained the A→G transition that defines HG 8, which has been found at high frequencies in sub-Saharan African populations, in 2% of Egyptians, and in 5% of west Asians. This transition has not been found in 983 chromosomes from Europe, the rest of Asia, Australasia, and the Americas (Seielstad et al. 1994; Hammer et al. 1997). A sub-Saharan African origin for this haplogroup was suggested (Hammer et al. 1998), and, thus, its presence in northwestern Africa may indicate sub-Saharan admixture in northwestern Africa. No other sub-Saharan–specific haplogroups—such as HG 6, which is present in Pygmies and in San (Jobling 1994; Jobling et al. 1997), or HG 7 (Jobling et al. 1997)—were found in our samples. HG 9, seen here with a frequency of 11.6%, has a Mediterranean distribution, with its highest frequencies occurring in the Near East (Casanova et al. 1985; Brega et al. 1987; Mitchell et al. 1993, 1997; Semino et al. 1995). This pattern was interpreted as being produced either by such colonizing seafaring peoples as the Phoenicians (Mitchell and Hammer 1996; Mitchell et al. 1997) or by neolithic farmers (Gonçalves and Lavinha 1994). On the other hand, HG 1 (3.9%) has been found predominantly in Europe (Mathias et al. 1994; Jobling et al. 1997; Mitchell et al. 1997), and HG 2 (2.3%), in Europe and Asia (Jobling and Tyler-Smith 1995; Jobling et al. 1997). HG 26 was found at a frequency of .8%. No Y chromosome was found to belong either to HG 3, which is present in European populations (Jobling et al. 1997), or to HG 15, which seems to be specific to Indian populations (Pandya et al., in press).

Y-Chromosome STR Polymorphisms

Allele-frequency distributions for eight Y-chromosome STRs are given in the last column (designated as "Overall") of table 3; gene diversities and allele-size variances can also be found in the same table. Of a total of 129 complete haplotypes constructed, considering seven of the Y-chromosome STR polymorphisms studied, 56 distinct Y-chromosome–haplotype configurations were obtained. The most frequent Y-chromosome haplotype, 13-12-11-24-9-11-13 (DYS19-DYS388-DYS389I-DYS390-DYS391-DYS392-DYS393), was found in 33 individuals, whereas 42 haplotypes were observed in unique copies. Haplotype diversity was estimated at .93 ± .02.

Y-Chromosome STR Polymorphism within Haplogroups

Y-chromosome STR allele frequency distributions, by haplogroup, are shown in table 3. Clearly, allelic variation at each STR locus shows striking differences among haplogroups (see also table 4, in which the

Table 3

STR Allele Frequencies by Haplogroup and by Total in Northwestern Africa

Marker and Allele	HG 9 (n = 15)	HG 1 (n = 5)	HG 26 (n = 1)	HG 2+ (n = 3)	HG 21 (n = 99)	HG 8 (n = 6)	Overall (n = 129)
DYS19:[a]							
13	0	0	0	0	.889	0	.681
14	1	.400	1	.667	.071	0	.209
15	0	.200	0	0	.020	.500	.047
16	0	.400	0	0	.010	.500	.047
17	0	0	0	.333	.010	0	.016
DYS388:[b]							
12	0	1	1	0	.960	1	.827
13	0	0	0	.333	.040	0	.039
14	.067	0	0	0	0	0	.008
15	.067	0	0	.333	0	0	.016
16	.133	0	0	0	0	0	.016
17	.600	0	0	.333	0	0	.078
18	.133	0	0	0	0	0	.016
DYS389I:[c]							
9	.067	.200	1	.333	.091	.666	.132
10	.800	.600	0	.667	.172	.167	.271
11	.133	.200	0	0	.717	.167	.581
12	0	0	0	0	.010	0	.008
13	0	0	0	0	.010	0	.008
DYS389II:[d]							
24	0	0	1	0	0	0	.008
25	0	.200	0	.333	0	0	.016
26	.200	.400	0	.667	.152	.167	.178
27	.667	.400	0	0	.626	.500	.596
28	0	0	0	0	.202	.333	.171
29	.133	0	0	0	.010	0	.023
30	0	0	0	0	.010	0	.008
DYS390:[e]							
21	0	0	0	0	0	1	.047
22	0	0	0	0	.040	0	.031
23	.800	0	1	1	.172	0	.256
24	.133	.600	0	0	.606	0	.503
25	.067	.400	0	0	.182	0	.163
DYS391:[f]							
8	0	0	0	0	.010	0	.008
9	0	0	0	0	.778	0	.597
10	.200	.200	1	.667	.182	1	.240
11	.733	.800	0	.333	.030	0	.147
12	.067	0	0	0	0	0	.008
DYS392:[g]							
10	0	0	0	0	.010	0	.008
11	1	0	1	1	.909	1	.891
12	0	0	0	0	.081	0	.062
13	0	1	0	0	0	0	.039
DYS393:[h]							
10	0	0	0	0	.010	0	.008
11	0	0	0	0	0	0	0
12	1	0	0	.667	0	0	.132
13	0	.800	0	.333	.970	.333	.798
14	0	.200	1	0	.020	.667	.062

NOTE.—Both gene diversity (D) and allele-size variance (V) were computed from the total sample.

[a] D = .487; V = .814.
[b] D = .304; V = 2.559.
[c] D = .573; V = .580.
[d] D = .584; V = .643.
[e] D = .654; V = .881.
[f] D = .566; V = .609.
[g] D = .200; V = .209.
[h] D = .342; V = .25.

Table 4

Most-Frequent STR Alleles within Each Haplogroup Found in Northwestern Africa

Haplogroup	STR Alleles[a]						
	DYS19	DYS388	DYS389I	DYS390	DYS391	DYS392	DYS393
HG 9 (n = 15)	14 (100%)	17 (60%)	10 (80%)	23 (80%)	11 (73%)	11 (100%)	12 (100%)
HG 1 (n = 5)	14/16 (40%)	12 (100%)	9 (60%)	24 (60%)	11 (80%)	13 (100%)	13 (80%)
HG 6 (n = 1)	14	12	9	23	10	11	14
HG 2+ (n = 3)	14 (67%)	13-15-17 (33%)	10 (67%)	23 (100%)	10 (67%)	11 (100%)	12 (67%)
HG 21 (n = 99)	13 (89%)	12 (96%)	11 (72%)	24 (61%)	9 (78%)	11 (91%)	13 (97%)
HG S8 (n = 6)	15-16 (50%)	12 (100%)	9 (67%)	21 (100%)	10 (100%)	11 (100%)	14 (67%)

[a] Unique alleles are underlined.

most frequent alleles are shown for each haplogroup). Given the variability of a particular STR locus, some haplogroups display several alleles, but, in others, the number of alleles is highly restricted and differentiated. When variation at locus DYS390 is taken as an example, it can easily be seen that allele 21 is present exclusively in HG 8 (where it is the only allele observed), that allele 23 is the only allele found in HG 2+ and is the most frequent allele in HG 9, and that several alleles at intermediate frequencies are present in HG 21 and HG 1. Even when the small number of individuals in some of the haplogroups is taken into account, this obvious microsatellite differentiation among the haplogroups seems to indicate that Y-chromosome genetic variation is strongly structured by haplogroup background. Of the 56 STR-distinct haplotypes found, only one was shared by two different haplogroup backgrounds (HG 2+ and HG 9). The apportionment of Y-chromosome STR-haplotype diversity among and within haplogroups was assessed by AMOVA (table 5): 83.5% ($P < .0001$) of the total genetic variation was attributable to differences between haplogroups. Compared with between-population differentiation, this is an extremely high value. Among the four linguistic subpopulations represented in our sample, the proportion of STR genetic variance explained by between-population difference was 3.72% ($P = .0088$). Of the genetic variability for Y-chromosome STRs, 3.5% could be apportioned to between-population differences among four European samples (de Knijff et al. 1997), and that fraction was 20.6% among four central Asian samples (Pérez-Lezaun et al. 1999). With the use of 10 Y-chromosome STRs, the apportionment of diversity between three African groups of populations was estimated to be 2.52% (Seielstad et al. 1998).

A number of diversity parameters were computed for each haplogroup, to characterize more extensively the Y-chromosome STR allelic variation within them (see table 6). HG 2+, despite having a low frequency, was found to be the most diverse, as is shown by the average gene diversity, the mean number of different alleles, the mean number of differences in repeat size, and the mean repeat-size variance. However, given the small sample size for HG 2+, the difference, in those parameters, between HG 2+ and any other haplogroup was statistically significant only for the difference in repeat length among haplotypes against HG 21 ($P = .042$, by permutation test). As discussed below, HG 2+ is probably the oldest haplogroup among those found in this study, and, thus, it may contain distinct, old lineages with heterogeneous STR-haplotype variation. On the other hand, HG 1 seems to occupy an intermediate position between HG 2+ and the rest of haplogroups, which were found to be more compact. Again, these differences were not statisti-

Table 5

Percent Fractions of the Genetic Variation That Can Be Attributed to Differences among Haplogroups and to Differences among Populations in Two Different Geographic Areas

Difference in Genetic Variation	Percent Fractions[a] of Genetic Variation in Two Different Geographic Areas Studied		
	Northwestern Africa	Iberian Peninsula	Northwestern Africa and the Iberian Peninsula
Among haplogroups	83.5 ($P < .0001$)	23.4 ($P = .0029$)	66.4 ($P < .0001$)
Among populations	3.7 ($P = .0088$)	2.2 ($P = .0890$)	19.2 ($P < .0001$)

[a] As measured by AMOVA.

Table 6

Diversity Parameters in Northwestern Africa by Haplogroup, Except for HG 26, Which Was Found in Only One Individual

Diversity Parameters	HG 9 (n = 15)	HG 1 (n = 5)	HG 2+ (n = 3)	HG 21 (n = 99)	HG 8 (n = 6)
No. of polymorphic sites[a]	4	5	5	7	3
No. of different STR haplotypes	9	5	3	33	5
Haplotype diversity[b]	.80 ± .11	1.00 ± .13	1.00 ± .27	.88 ± .03	.93 ± .12
Mean gene diversity[c]	.26 ± .18	.41 ± .30	.52 ± .45	.27 ± .17	.25 ± .19
Mean no. of allele differences[d]	1.81 ± 1.10	2.90 ± 1.82	3.67 ± 2.52	1.90 ± 1.10	1.73 ± 1.17
Mean no. of differences in repeat size[e]	2.38 ± 1.85	3.40 ± 3.06	6.67 ± 4.24	2.44 (2.31)[f] ± 1.85	2.00 ± 1.97

[a] No. of STR loci found to be variable within each haplogroup in this sample.
[b] Gene diversity computed from STR haplotype frequencies.
[c] Mean gene diversity at each STR locus within each haplogroup.
[d] Pairwise average of the number of different STR alleles for all pairs of chromosomes within each haplogroup.
[e] Pairwise average of the cumulative absolute difference in STR allele length for all pairs of chromosomes within each haplogroup.
[f] Without large poly (A) tail individual.

cally significant. In table 6, for HG 21, the number shown in parentheses corresponds to the mean number of differences in repeat size obtained without taking into account the individual with the large poly (A) tail, who seems to contribute disproportionately to the parameter. In fact, this individual would be classified as having YAP haplotype 3A, according to Hammer et al. (1998), whereas all other individuals in HG 21 would be classified as having YAP haplotype 4. This individual may belong to a related, but different, evolutionary branch in the Y-chromosome genealogy.

For the STR variability generated within the chromosomes bearing derived states in the biallelic markers that characterize the haplogroups found in northwestern Africa, TMRCAs were estimated from the mean allele length variance of the seven STRs considered (table 7). Mean allele length variances ranged from .181 to .844. The most ancient TMRCA, estimated at 14,067 years ago (ya) (95% CI 757–68,782 ya), was for SRY-1532 A→G, which is found in all 129 northwestern African Y chromosomes analyzed. The TMRCA for STR variability within substitution SRY-8299 G→A, which defines haplogroups 21 and 8, was estimated to be 6,483 ya (95% CI 493–30,782 ya). The TMRCA for mutation sY81 A→G, which is found solely on chromosomes bearing allele SRY-8299 A and which defines HG 8, was estimated to be younger (3,017 ya, with a 95% CI 0–18,565 ya), and that for the 8-kb allele at the 12f2 polymorphism was estimated to be 4,583 ya (95% CI 0–27,609 ya). Finally, the TMRCAs for mutations M9 C→G and 92R7 C→T were found to be ~7,867 ya (95% CI 1,264–33,347 ya) and 5,233 ya (95% CI 0–27,696 ya), respectively.

The mean repeat-number size difference was also computed for pairs of haplogroups, and it was found to be much larger than the within-haplogroup means (see the numbers below the diagonal in table 8). With the use of a stepwise mutation model, the difference in repeat size between a pair of alleles is expected to grow with time (Goldstein et al. 1995*a*, 1995*b*). The mean number of differences between pairs of haplogroups was converted to genetic distances (which are indicated by numbers above the diagonal in table 8), by means of the Jensen Difference (Rao 1982), which takes into account variability within each haplogroup. Thus, a measure of differentiation between haplogroups, which is closely related to D_{SW} (Shriver et al. 1995), a measure of genetic distance for independent STRs, was obtained. HG 26 was not included either in the calculation of the mean repeat-number size difference for pairs of haplogroups or in the reconstruction of the haplogroup genealogy, since it contained only one chromosome. In accordance with AMOVA results, the average difference in repeat lengths between pairs of haplotypes belonging to different haplogroups was always larger than the average differences within the haplogroups being compared. With a few exceptions (e.g., the average difference between HG 1 and HG 2+ compared with the internal differences in either haplogroup or the average difference between HG 2+ and HG 9 compared with the internal average difference in HG 2+), this pattern was statistically significant by permutation test (P values .039–.001). The largest distances were found between HG 8 and HG 9 (9.68) and between HG 1 and HG 9 (8.10), a finding that is in accordance with the parsimony network (see fig. 1), where these haplogroups occupy peripheral positions. On the contrary, the shortest distance was found between HG 2+ and HG 9, which are separated by only one unique-mutation event. The discordant genetic distances obtained between HG 8 and HG 21 (a large genetic distance for a single point mutation) and between HG 8 and HG 2+ (a rather low value for haplogroups three mutational events apart), in relation to those that were expected, could be attributed

Table 7

TMRCAs Computed from STR Allele-Size Variance in Northwestern Africa

	Mutation					
	SRY-1532 (A→G)	12f2 (10Kb→8Kb)	M9 (C→G)	92R7 (C→T)	SRY-8299 (G→A)	sY81 (A→G)
Haplogroups containing the derived allele	All (n = 129)	9 (n = 15)	1 + 26 (n = 6)	1 (n = 5)	21 + 8 (n = 105)	8 (n = 6)
Mean STR allele-length variance[a]	.844	.275	.472	.314	.389	.181
Expansion time (ya)[b]	14,067	4,583	7,867	5,233	6,483	3,017
(95% CI[c])	757–68,782	0–27,609	1,264–33,347	0–27,696	493–30,782	0–18,565

[a] Mean STR allele length variance is the average over STRs of allele length variance.
[b] Times estimates were computed on the basis of equation 5 in Di Rienzo et al. (1998; see also the Material and Methods section).
[c] CI was computed by taking into account variance interlocus sampling error and mutation rate estimate error.

to the small sample size of HG 8 (n = 6) and HG 2+ (n = 3), to the high internal diversity within HG 2+, or to stochastic processes in the generation of STR variation within haplogroups, which may have played an important role, as discussed below. A minimum spanning tree constructed from the distance matrix showed all other haplogroups stemming from HG 2+, which matches the known haplogroup genealogy (fig. 1), with the exception of the position of HG 8, which derives from HG 21. The different levels of internal diversity within haplogroups could bias the genetic distance estimates among them (Charlesworth 1998); however, if the correction for internal diversity was not applied, the haplogroup phylogeny reconstructed had little resemblance to the actual known phylogeny.

Replication Analysis

Given the small sample sizes in some of the haplogroups, we sought to confirm our results by adding Y-chromosome data for northern Iberians (51 Basques and 27 Catalans). The same STRs had been typed (Pérez-Lezaun et al. 1997) in those samples, and it was possible to allocate the chromosomes to the haplogroups defined by biallelic polymorphisms (Semino et al. 1996; Underhill et al. 1997; Hurles et al. 1999; P. Underhill, unpublished data). This sample had a different, partly overlapping haplogroup composition with respect to northwestern Africa: 55 (70.5%) chromosomes belonged to HG 1, eight (10.3%) to HG 2+, one (1.3%) to HG 9, two (2.6%) to HG 21, and 12 (15.4%) to HG 22, which has a geographic distribution that is almost restricted to northern Iberia and that is suggested to have sprung quite recently from HG 1 (Hurles et al. 1999).

AMOVA performed on STR-haplotype variability among haplogroups in northern Iberia showed that 23.4% (P = .0029) of the genetic variation could be apportioned among haplogroups. This result contrasts with the 83.5% result found among haplogroups in northwestern Africa (table 5). This may be the result of the presence, in northern Iberia, of HG 22, which is suggested to have recently originated from HG 1 in northern Iberia (Hurles et al. 1999); in fact, all but one of the Y-chromosome STR haplotypes in HG 22 were found in HG 1. The fraction of Y-chromosome STR genetic variation among Basques and Catalans is 2.2%, which is not significantly different from zero (P = .0890) and which is still much smaller than the fraction of genetic variation among haplogroups. When we pooled the samples from northwestern Africa and northern Iberia, the fraction of genetic variation attributable to haplogroups was 66.4%, whereas it was 19.2% among populations (table 5). Thus, although the results in the pooled sample were not as extreme as those seen when samples from northern Africa alone were considered, it still holds that genetic variation is more deeply structured by lineage than by population.

If we pooled the samples from northwestern Africa and northern Iberia, we reduced the number of haplogroups with small sample sizes and achieved a greater statistical power. The pattern of genetic diversity within haplogroups was confirmed in the pooled sample: HG 2+ (n = 11) had the largest in-

Table 8

Genetic Distances among Haplogroups (with One Exception[a]) in Northwestern Africa

	Genetic Distances among HGs				
Haplogroup	1	2+	21	8	9
HG 1	3.40	4.77	4.44	5.9	8.1
HG 2+	9.80	6.67	4.40	4.56	.25
HG 21	7.29	8.95	2.31	6.13	7.76
HG 8	8.60	8.89	8.29	2	9.68
HG 9	10.99	4.78	10.1	11.87	2.38

Note.—Distances below the diagonal of underlined numbers represent the mean number of repeat differences among all pairs of STR haplotypes between two haplogroups. Distances along the diagonal of underlined numbers denote the mean number of repeat differences within haplogroups. Distances above the diagonal of underlined numbers denote genetic distances among hap

[a] HG 26, which was found in only one chromosome, was not included in the determination of genetic distances.

ternal diversity (average difference in repeat length among STR haplotypes 6.67), when compared with HG 1 (n = 60; 3.22), HG 9 (n = 16; 2.85), HG 21 (n = 101; 2.42), HG 22 (n = 12; 2.41), and HG 8 (n = 6; 2.00). Using a permutation procedure, we tested whether this parameter was significantly different between pairs of haplogroups, and we found that it was significantly different between HG 2+ and any other haplogroup (P values between .0001 and .0112), and between HG 1 and HG 21 (P = .0021).

Mutation ages were also estimated from the pooled sample. Although sample sizes were greatly increased for the sets of chromosomes bearing some mutations, mutation ages did not change noticeably. The age of SRY-1532 A→G was estimated, in the pooled sample (n = 207), to be 13,550 ya (95% CI 2,279–56,826 ya), whereas its estimated age in northwestern Africa was 14,067 ya (table 7). The age of 12f2 10 kb→8 kb (n = 16) was estimated to be 5,516 ya (95% CI 0–31, 130 ya). The age estimates for M9 C→G (n = 73) and for 92R7 C→T (n = 72) were 6,617 (95% CI 1,236–27,000) and 6,300 (95% CI 1,079–26,304) ya, respectively. On the YAP-positive branch, the age estimate for SRY-8299 (n = 107) was 6,383 ya (95% CI 493–30,304 ya), and that for sY81 remains unchanged, since no northern Iberian Y chromosome was found in HG 8. The haplogroup phylogeny reconstructed from STR haplotypes was the same with the pooled samples as it was with northwestern African chromosomes alone, with the addition of the correct link between HG 1 and HG 22.

Discussion

Although biallelic polymorphisms can be regarded as unique mutational events (or as events of very low probability) and can allow us to identify deep lineages of Y chromosomes (called, in the present study, "haplogroups"), STR polymorphisms exhibit a faster mutation rate (0.0012 per locus per generation) (Heyer et al. 1997; Kayser et al. 1997; Bianchi et al. 1998), producing highly informative markers for studies of recent evolutionary (or historical) events (Roewer et al. 1996; Pérez-Lezaun et al. 1997). As a result of their differential rate of evolution, the combination of the data obtained from both types of markers on the nonrecombining portion of the Y chromosome provides an interesting perspective on different levels of resolution in the phylogeny of Y chromosomes.

The six different haplogroups found in northwestern Africa may offer clues as to which groups of Y chromosomes contributed to the composition of the present-day population. As we have described, HG 21 is the major haplogroup that characterizes Y-chromosome diversity in northwestern Africa. Its high frequency—the highest ever reported for this haplogroup—can be interpreted as the result of a specific founder effect or drift process in this geographic region. On the contrary, HG 8 and HG 9 may have been introduced by gene flow from sub-Saharan Africa and from the eastern Mediterranean basin, respectively (Jobling and Tyler-Smith 1995; Jobling et al. 1996, 1997).

Compartmentalization of Genetic Variance

Since no recombination occurs between haplogroups, we can consider that they behave like independent units (or subpopulations) without migration. Therefore, it would be expected that the compartmentalization of the genetic information they carry would be complete. AMOVA showed that>80% of the genetic variance was found among haplogroups in northwestern Africa and that >65% was found in northwestern Africa plus northern Iberia; these levels of compartmentalization are very high, but they are not complete. There are two complementary explanations for this result.

First, the origins of haplogroups are not independent. Each haplogroup arose when a rare mutation occurred on a given Y chromosome that belonged to a certain haplogroup and that carried a microsatellite haplotype. Genetic variation in the new haplogroup was reset to zero, but the founding haplotype was either similar or altogether identical to other haplotypes in the parental haplogroups. Thus, immediately after the origin of a haplogroup, its genetic background was likely to be closely related to that of the parental haplogroup, until mutation and drift differentiated them. Sometimes, it may be possible to identify this founder haplotype from the most common haplotype of the new haplogroup. The signal will generally decrease with time. This is clearly exemplified by HG 22 chromosomes in northern Iberia, which contain STR haplotypes that are also found in its parental HG 1.

Second, the recurrent nature of microsatellite mutation implies that haplotypes that are identical by state may not be identical by descent. In the same way that microsatellite haplotypes in each haplogroup will differentiate with time from the common ancestor from which they derive, haplotypes belonging to different haplogroups could occasionally converge with time as well. In fact, in our northwestern African sample, two Y chromosomes belonging to HG 2+ and HG 9 shared an STR haplotype.

We have analyzed 11 biallelic polymorphisms that define haplogroups with known phylogeny (Jobling and Tyler-Smith 1995; Jobling et al. 1996, 1997; Hurles et al. 1998, 1999) and with only one instance of backmutation (SRY-1532; see fig. 1), which does not affect the present analysis, since neither HG 3 nor HG 7 has

been found. On that well-defined background, we have considered seven tri- and tetranucleotide STRs, and, as noted above, we have observed an almost complete compartmentalization. Malaspina et al. (1998) typed four dinucleotide STRs and found that 36 of 179 STR haplotypes were shared among the four haplogroups ("frames" in Malaspina et al. [1998]) defined by two biallelic polymorphisms. The much smaller number of STR haplotypes shared among haplogroups in our study may reflect the larger number of biallelic and STR loci typed.

It could be argued that, if more polymorphic sites were included and if the haplogroups were further subdivided, the apportionment of genetic variation among haplogroups would be lower. Preliminary results (E. Bosch, unpublished data) show that our sample of HG 21 Y chromosomes from northwestern Africa can be subdivided by typing three additional biallelic polymorphisms, thereby resulting in three different subhaplogroups. When AMOVA was repeated in the whole set of Y chromosomes from northwestern Africa, and when the three subhaplogroups of HG 21 were considered, the fraction of genetic variation among haplogroups was 84.3%, which is very similar to (and is, in fact, slightly higher than) that found without splitting HG 21.

We have also shown that genomic and population perspectives on STR genetic variation give very different results. When we defined groups of chromosomes according to the biallelic variants they carry, the fraction of STR genetic variation found among them was much higher than that found among groups of chromosomes defined by their population origin. This may happen if the origin of the populations (the ethnogenesis process) is more recent than that of most of the biallelic polymorphisms. The compartmentalization of STR genetic variation is expected to be tighter for haplogroups than it is for populations, since a haplogroup goes through a strict bottleneck with a size of one chromosome, whereas population bottlenecks have not been so narrow, and since there is gene flow between populations but there is no gene exchange between the nonrecombining portion of the Y chromosome. In summary, genetic background determines the STR genetic variation to a much greater extent than does the population background.

Variance in Repeat Number: STR Variation and Mutation Age

The analysis of STR repeat-size variance within lineages of Y chromosomes bearing a particular derived allele at a biallelic locus has been used to estimate TMRCAs (table 7). That is, we have tried to estimate the time to which the observed STR variation within each Y-chromosome lineage coalesces. As we will discuss, this time is expected to be correlated to the actual age of the mutation that defines each lineage. Thus, we found both an old mutation (SYR-1532), which was borne by all the chromosomes in our sample, and five younger mutations. These estimated TMRCAs (see table 7) match the known haplogroup genealogy (fig. 1), in which SRY-8299 precedes sY81 and M9 precedes 92R7.

Hammer et al. (1998) typed a worldwide sample of Y chromosomes for a set of biallelic markers that partly overlaps with our set. The authors estimated mutation ages by use of coalescence analysis (Griffiths and Tavaré 1994), which is done on the basis of both the shape of the gene tree and the number of chromosomes bearing each combination of mutations. The ages for those biallelic mutations that overlap with our study were 110,000 ± 36,000 ya for the polymorphic site at position 10,831.1 of the SRY region (synonymous to SRY-1532), ~31,000 ya for the polymorphic site at position 4,064 of the SRY region (synonymous to SRY-8299), ~30,000 ya for DYS257 (which seems to have happened in the same phylogenetic branch as 92R7, as they appear to be phylogenetically equivalent; Jobling et al. 1998*b*), and ~11,000 ya for PN1 (which cosegregates with sY81; Hammer et al. [1997]). All of them are clearly older than our estimates for the TMRCAs through STR variation, although both age sets are correlated and have broad, overlapping 95% CIs. The largest discrepancy is found for SRY-1532, a deep mutation in the phylogeny, from which many different haplogroups—most of which have not been found in this study—derive. Since most (>80%) of the chromosomes in our sample belong to a single recent branch of the genealogy (that bearing SRY-8299 and sY81) and since other branches bearing SRY-1532 are found at low frequency or are altogether absent, this could lead to an underestimation of the STR allele length variance within SRY-1532 and, therefore, of its TMRCA. However, when we added Y chromosomes from northern Iberia, which contributed other branches derived from SRY-1532, the age estimate through STR variation did not increase. In a study of Y chromosomes from Polynesia and Melanesia (Hurles et al. 1998), the authors also refrained from dating the TMRCA for SRY-1532 from STR data, on the basis that mutation/drift equilibrium renders such dating methods unable to resolve suitably far back in time. Apart from this particular case, several reasons could account for the general lag between our TMRCA estimates through STR variation and mutation ages obtained by coalescence analysis in Hammer et al. (1998):

1. *Mutation Age Precedes TMRCA by Definition.* We have estimated the time to the current observed STR variation. The observed STR haplotypes may coalesce to a time that is more recent than the actual mutation age, because of the extinction of lineages that appeared

in the first generations after the mutation arose and that we are not able to detect. Moreover, the variance in repeat size, which we have used to estimate the TMRCA, is zero until the first mutation in a STR. The method we used (Di Rienzo et al. 1998) dates, in fact, an expansion that should have happened, obviously, at some unknown time after the mutation arose. The distributions of pairwise differences in repeat number found within the most frequent haplogroups (not shown) are smooth and bell-shaped, as is expected for a lineage that underwent an expansion (Shriver et al. 1997). This justifies the use of the equation given by Di Rienzo et al. (1998). Coalescence prior to the actual mutation, the onset of STR mutations, and the expansion detected by the method suggested in Di Rienzo et al. (1998) all may contribute to the lag between mutation age and our estimate of TMRCA.

2. *Population Biases.* We have northwestern African Y chromosomes, which contain a particular set of haplogroups as a result of their population history (founder events, gene flow, and admixture), and this fact could lead to an underestimation of haplogroup variability and mutation ages. To test for this possible bias, we used coalescence analysis on northwestern Africa haplogroup phylogeny and frequencies, according to the same methods and parameters used in Hammer et al. (1998), and we were able to reproduce their mutation age estimates, with minor discrepancies. Another test for this possible bias results from the addition of chromosomes from other populations. When we added two population samples from northern Iberia, the age estimates through STR variation of the mutations present in NW Africa did not change. Therefore, it does not seem that population bias plays a major role in the discrepant differences between mutation ages, as seen elsewhere (Hammer et al. (1998), and our STR-based TMRCA estimates.

3. *STR Saturation.* Given the fast mutation rate of STRs, it is possible that they have reached mutation-drift equilibrium and that, therefore, genetic variation in STRs would remain constant in time. However, if STRs have reached a complete saturation, we would not have observed a correlation between TMRCA and the relative mutation ages from the gene genealogy.

4. *STR Mutation Rate Overestimation.* If STR mutation rates had been overestimated by a factor of 5–6, our TMRCAs and the age estimates of Hammer et al. (1998) would match. Note that mutation rates are not homogeneous across STR loci (Carvalho-Silva et al. 1999). Then, mutation in genealogies will tend to be observed preferentially in the STRs with the fastest mutation rates, and the mean mutation rate will be overestimated. A similar situation is found in the mtDNA control region, where the mutation rate estimates per nucleotide are roughly 20 times higher when derived from mother-child transmission studies than when estimated from phylogenetic comparisons with nonhuman primate sequences (Jazin et al. 1998).

In sum, the discrepancy between the mutation age estimates provided elsewhere (Hammer et al. [1998]) and our estimates of the TMRCA may be mainly the result of the actual lag between the mutation event and the lineage expansion we dated and of possible overestimation of STR mutation rate, which is impossible to check, given current knowledge.

Haplogroup Phylogeny and STRs

In the two previous sections of this Discussion, we have seen that STR haplotypes in a haplogroup can preserve an amount of phylogenetic information about the parental haplogroup that declines with time and that the relative time of origin of a haplogroup can be estimated from STR allele-size variance. We can combine both types of information, reconstruct a haplogroup phylogeny, and compare it with the actual phylogeny constructed from the biallelic polymorphisms that define the haplogroups. The results of this exercise have shown that STRs within each haplogroup do trace most of the haplogroup phylogeny. In spite of the small sample sizes of some of the haplogroups, the phylogeny obtained from the STRs matches the known haplogroup phylogeny, with only one discrepancy: the position of HG 8, which stems from HG 2+ in our reconstruction but which is actually derived from HG 21. A similar analysis was performed (Rocha et al. 1997) with the allele frequencies of an STR located at the 5′ end of the autosomal α1-antitrypsin gene within the electrophoretic variants of protein α1-antitrypsin.

We have shown that genetic background prevails over population background in the determination of STR genetic variation on the human Y chromosome. We have also used STR variation within lineages to infer the TMRCAs and have shown that the genetic differentiation that is left among haplogroups contains still phylogenetic information but that the overall STR variation is mainly driven by the haplogroup composition of a specific population. The genetic composition of humans may thus be better understood in terms of evolutionarily related genealogical units rather than in demographic terms.

Acknowledgments

We express our appreciation to the original DNA donors who made this study possible. We especially thank Arpita Pandya for her help and advice concerning the typing of some of the biallelic polymorphisms. The warm reception E.B. received from Cancer Research Campaign Chromosome Molecular Biology Group staff during her stay in the Department of Biochemistry, Oxford University, is also acknowledged with grat-

itude. We appreciate the technical assistance offered by the Unitat de Seqüenciació, Servei Científico-Tècnic, Universitat de Barcelona. We also thank Peter de Knijff, for providing us with allelic ladders, and Peter A. Underhill, for typing YAP-negative individuals for M9 and for sharing with us unpublished individual data on Basques and Catalans. Two anonymous reviewers made suggestions that improved significantly this manuscript. This research was supported by Dirección General de Investigación Científica y Técnica in Spain (PB95-0267-CO2-01), by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009), and by Institut d'Estudis Catalans. Comissionat per a Universitats i Recerca, Generalitat de Catalunya supported E.B. (FI/96-1153); F.R.S. was supported by the Leverhulme Trust; and C.T.-S. was supported by the Cancer Research Campaign.

Electronic-Database Information

URLs for data in this article are as follows:

CEPH, http://www.cephb.fr/ (for data from CEPH-family cell lines)

Genome Database, The, http://gdbwww.gdb.org/ (for primers for the DYS389 locus)

Arlequin: A Software for Population Genetic Data Analysis, http://anthropologie.unige.ch/arlequin/

References

Altheide TK, Hammer MF (1997) Evidence for a possible Asian origin of YAP+ Y chromosomes. Am J Hum Genet 61:462–466

Banchs I, Bosch A, Guimerà J, Lázaro C, Puig A, Estivill X (1994) New alleles at microsatellite loci in CEPH families mainly arise from somatic mutations in the lymphoblastoid cell lines. Hum Mutat 3:365–372

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 59:63–81

Bianchi NO, Bailliet G, Bravi CM, Carnese RF, Rothhammer F, Martínez-Marignac VL, Pena SDJ (1997) Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am J Phys Anthropol 102:79–89

Bianchi NO, Catanesi CI, Bailliet G, Martínez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ, et al (1998) Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet 63:1862–1871

Bouzekri N, Taylor PG, Hammer MF, Jobling MA (1998) Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. Hum Mol Genet 7:655–659

Brega A, Torroni A, Semino O, Maccionni L, Casanova M, Scozzari R, Fellous M, et al (1987) The p12f2/*Taq*I Y-specific polymorphism in three groups of Italians and in a sample of Senegalese. Gene Geogr 1:201–206

Carvalho-Silva DR, Santos FR, Hutz MH, Salzano FM, Pena SD (1999) Divergent human Y-chromosome microsatellite evolution rates. J Mol Evol 49:204–214

Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, Purrello M, et al (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science 230:1403–1406

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol 15:538–543

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

Deka R, Jin L, Shriver MD, Yu LM, Saha N, Barrantes R, Chakraborty R, et al (1996) Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Genome Res 6:1177–1184

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110:134–140

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, et al (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148:1269–1284

Estivill X, Morral N, Bertranpetit J (1994) Reply to Kaplan et al. Nat Genet 8:216–218

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6–13

——— (1984) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 137:266–267

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, et al (1998) Jefferson fathered slave's last child. Nature 396:27–28

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995*a*) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

——— (1995*b*) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727

Gonçalves J, Lavinha J (1994) The Y-associated XY275G (low) allele is common among the Portuguese. Am J Hum Genet 55:583–584

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

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307–319

Guellaen G, Casanova M, Bishop C, Geldwerth D, Andre G, Fellous M, Weissenbach J (1984) Human XX males with Y single-copy DNA fragments. Nature 307:172–173

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

——— (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, et al (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427–441

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, et al (1997) The geographic distribution of human Y chromosome variation. Genetics 145: 785–805

Hammer MF, Zegura SL (1996) The role of the Y chromosome in human evolutionary studies. Evol Anthropol 5:116–134

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, et al (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63:1793–1806

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Pérez-Lezaun E, Bosch E, et al (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphism. Am J Hum Genet 65:1437–1448

Jazin E, Soodyall H, Jalonen P, Lindholm P, Stoneking M, Gyllensten U (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. Nat Genet 18:109–110

Jobling MA (1994) A survey of long-range DNA polymorphisms on the human Y chromosome. Hum Mol Genet 3: 107–114

Jobling MA, Bouzekri N, Taylor PG (1998*a*) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet 7:643–653

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110:118–124

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767–1775

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11:449–456

Jobling MA, Williams G, Schiebel K, Pandya A, McElreavey K, Salas L, Rappold GA, et al (1998*b*)A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol 8:1391–1394

Karafet T, Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, Romero F, et al (1997) Y chromosome markers and trans–Bering Strait dispersals. Am J Phys Anthropol 102:301–314

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133

Kwok C, Tyler-Smith C, Mendoza BB, Hughes I, Berkovitz GD, Goodfellow PN, Hawkins JR (1996) Mutation analysis of the 2Kb 5′ to SRY in XY females and XY intersex subjects. J Med Genet 33:465–468

Lucotte G, Ngo NY (1985) p49F, a highly polymorphic probe, that detects *Taq*I RFLPs on the human Y chromosome. Nucleic Acids Res 13:8285

Malaspina P, Cruciani F, Ciminelli BM, Terrenato L, Santolamazza P, Alonso A, Banyko J, et al. (1998) Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. Am J Hum Genet 63:847–860

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

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Mitchell RJ, Earl L, Fricke B (1997) Y-chromosome specific alleles and haplotypes in European and Asians populations: linkage disequilibrium and geographic diversity. Am J Phys Anthropol 104:167–176

Mitchell RJ, Earl L, Williams J (1993) Two Y-chromosome–specific restriction fragment length polymorphisms (DYS11 and DYZ8) in Italian and Greek migrants to Australia. Hum Biol 65:387–399

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. Curr Opin Genet Dev 6:737–742

Pandya A, King TE, Santos FR, Taylor PG, Thangaraj K, Singh L, Jobling MA, et al

Pena SDJ, Santos FR, Bianchi NO, Bravi CM, Carnese FR, Rothhammer F, Gerelsaiklan T, et al (1995) A major founder Y-chromosome haplotype in Amerindians. Nat Genet 11: 15–16

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimón J, et al (1999) Gender-specific migration in Central Asian populations revealed by the analysis of Y-chromosome STRs and mtDNA. Am J Hum Genet 65: 208–219

Pérez-Lezaun A, Calafell F, Seielstad MT, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y chromosome short tandem repeats in humans. J Mol Evol 45: 265–270

Pestoni C, Cal ML, Lareu MV, Rodríguez-Calvo MS, Carracedo A (1998) Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). Int J Legal Med 112:15–21

Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theor Pop Biol 21:24–43

Rocha J, Pinto D, Santos MT, Amorim A, Amil-Dias J, Cardoso-Rodrigues F, Aguiar A (1997) Analysis of the allelic diversity of a (CA)n repeat polymorphism among α-1-antitrypsin gene products from northern Portugal. Hum Genet 99:194–198

Roewer L, Kayser M, Dieltjes P, Nagy M, Bakker E, Krawczak M, de Knijff P (1996) Analysis of molecular variance (AMOVA) of Y-chromosome–specific microsatellites in two closely related human populations. Hum Mol Genet 5:1029–1033

Rolf B, Meyer E, Brinkmann B, de Knijff P (1998) Polymorphism at the tetranucleotide repeat locus DYS389 in 10 populations reveals strong geographical clustering. Eur J Hum Genet 6:583–588

Santos FR, Carvalho-Silva DR, Pena SDJ (1999*a*) PCR-based DNA profiling of human Y chromosomes. In: Epplen JT, Lubjuhn T (eds) Methods and tools in biosciences and medicine. Birkhauser Verlag, Basel, Switzerland, pp 133–152

Santos FR, Hutz M, Coimbra CEA, Santos RV, Salzano FM, Pena SDJ (1995*a*) Further evidence for the existence of a

major founder Y chromosome haplotype in Amerindians. Braz J Genet 18:669–672

Santos FR, Pandya A, Tyler-Smith C, Pena SDJ, Schanfield M, Leonard WR, Osipova L, et al (1999*b*) The central Siberian origin for Native Americans' Y chromosomes. Am J Hum Genet 64:619–628

Santos FR, Pena SDJ, Tyler-Smith C (1995*b*) PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. Gene 165:191–198

Santos FR, Rodriguez-Delfin L, Pena SDJ, Moore J, Weiss KM (1996) North and South Amerindians may have the same major founder Y chromosome haplotype. Am J Hum Genet 58:1369–1370

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Seielstad MT, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D, Cavalli-Sforza LL (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet 3:2159–2161

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti A (1996) A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. Am J Hum Genet 59:964–968

Semino O, Passarino G, Liu A, Brega A, Fellous M, Santachiara-Benerecetti AS (1995) Three Y-specific polymorphisms in populations of different ethnic and geographic origin. Y Chromosome Newsletter 2:5–6

Shriver MD, Jin L, Boerwinkle E, Deka R, Ferrell RE, Chakraborty R (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol 12: 914–920

Shriver MD, Jin L, Ferrell RE, Deka R (1997) Microsatellite data support an early population expansion in Africa. Genome Res 7:586–591

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–140

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, et al (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005

Underhill P, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196–200

Vogt PH, Affara N, Davey P, Hammer M, Jobling MA, Lau YF-C, Mitchell M, et al (1997) Report of the Third International Workshop on Y Chromosome Mapping 1997: Heidelberg, Germany, April 13–16, 1997. Cytogenet Cell Genet 79:2–16

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379–380

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

CAPÍTOL V

Y chromosome lineages and northwestern African populations

Elena Bosch, Francesc Calafell, David Comas, Peter J. Oefner, Peter A. Underhill i Jaume Bertranpetit

(en preparació)

Y-CHROMOSOME LINEAGES AND NORTHWESTERN AFRICAN POPULATIONS

[1]Elena Bosch, [1]Francesc Calafell, [1]David Comas, [2]Peter J. Oefner, [3]Peter A. Underhill and [1]Jaume Bertranpetit

[1]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, 08003 Barcelona, Spain.
[2]Stanford DNA Sequencing and Technology Center, 855 California Ave., Palo Alto, CA 94304.
[3]Department of Genetics, Stanford University, 300 Pasteur Dr., Stanford, CA 94305-5120

Correspondence should be addressed to:
Jaume Bertranpetit
Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain
Tel: +34-93-542.28.40
Fax: +34-93-542.28.02
e-mail. jaume.bertranpetit@cexs.upf.es

INTRODUCTION

The application of the DHPLC technique is demonstrating to provide a main step forward the understanding of human Y chromosome biallelic variation. The DHPLC set of markers contains 166 polymorphisms that have allowed to describe the most detailed phylogeny known for the Y chromosome (Underhill et al. 1999). Given the lack of recombination on most of its lenght, the rare event nature of the biallelic polymorphisms used and the large number of Y chromosomes typed, its structure appears clearly well defined and all possible homoplasies resolved. While most markers are located toward the tips of the phylogenetic tree and define haplotypes with a limited distribution, a few markers are found deeper in the tree. These bifurcating deep markers or nodes define distinctive major groups of haplotypes with much wider geographical distributions, which are related to different layers in the history of populations. This fine level of genetic dissection allows to pinpoint an origin for each Y chromosome, not only in space but also hints on their origin in time.

We have investigated Y chromosome biallelic variation in populations from northwestern Africa. Their Y chromosome lineages have been compared to those found in Iberian populations. We have identified distinctive layers in the history of the NW African populations: a first colonisation probably from eastern Africa; an Upper Paleolithic expansion than brought haplotype H38 to a very high frequency; the influence of the Neolithic from the Middle East, and subsequent gene flow from Europe and Sub-Saharan Africa. However, Y chromosomes in Iberia display a quite different picture with an European Paleolithic background with some local derivatives and the influence of the Neolithic, which probably came from a different route that the one that reached NW Africa.

RESULTS

Male lineage structure of Northwestern African populations

Haplotype frequencies for Moroccan Arabs, North Central Moroccan Berbers, Southern Moroccan Berbers and Saharawis are given in table 1. Differences in haplotype frequencies among those populations were tested through Analysis of Molecular Variance (AMOVA). Only 0.84% of the genetic variance was found to be due to the haplotype frequency differences among them (not statistically significantly different from zero, p=0.169). H38, which belongs to group III of Y chromosome haplotypes according to Underhill et al. 1999, is the most common haplotype in NW Africa (64%). Its higher frequency is found within the Saharawis (76%). H71, which belongs to group VI, is the second most frequent haplotype found in the region (10.8%). Other haplotypes found in decreasing frequencies are H22, H35 (6.25% each) and H36 (5.1%), all belonging to group III. The rest of haplotypes, which represent 8% of the NW African Y chromosomes, are found at frequencies below 3%.

Male lineage structure of Iberian populations

Haplotype frequencies for Basques, Catalans and Andalusians are also given in table 1. AMOVA showed that 2.48% of the genetic variance was attributable to the difference in haplotype frequencies among them (non statistically significant differences from zero, p=0.08). The most frequent haplotype in those populations is H103 (55.7%), which belongs to group IX. Haplotypes H101 and H102, which also belong to group IX, are found, respectively, at frequencies around 10%. The frequency of H71 (8.2%) is similar to that found in NW Africa. The total fraction of Y chromosomes belonging to group VI (which includes H71) shows a slightly higher frequency in Iberia (16.5%) when compared to NW Africa (13.7%). H35, H36 and H38, the only haplotypes found to belong to group III, comprise 4.1% of the Iberian Y chromosomes.

Geographical origins of Y chromosome haplotypes found in NW Africa and Iberian Peninsula

The worldwide distribution of the Y chromosome haplotypes may help to establish the putative origins of the haplotypes that make up the present NW African and Iberian populations. Fig. 2 show the detailed frequencies of haplotypes H22, H35,

H36, H38, H58, H71, H101, H102 and H103 for the populations we have studied, as well as their worldwide distribution. This type of descriptive analysis will allow to recognize the haplotypes found as autochthonous or of external origin: Sub-Saharan, European, or Mediterranean.

Group III is the most frequent in North Africa, representing 83% of all chromosomes analyzed. Its haplotypes are prevalent in Africa: Haplotype H22 (Fig. 2*a*) is found in the Middle East (7%) and in Africa, where it shows its higher frequencies (around 50%) in South and Central Africa, and intermediate frequencies in Ethiopia (9%), Mali (16%) and in the Khoisans (18%). As discussed below, its presence in NW Africa (6.25%) is likely to be due to a Trans-Saharan origin. Haplotype H28 is mainly found in Mali (29.5%) although it has also been described at very low frequency in the Middle East (3.5%). H28 Y chromosomes in NW Africa (1.71%) could be easily explained by gene flow from the Sahel. H35 (Fig. 2*b*) highest frequencies are found in Ethiopia (22.7%) and Sudan (17.5%). Its geographical distribution includes Sardinia (18.2%), as well as the Middle East and Central Asia where it is found at very low frequencies.The highest frequency for H36 (Fig. 2*c*) has been described in the Khoisans (10.3%), which is followed by those described in Ethiopia (6.8%) and NW Africa (5.1%). It is also found in South Africa at a very low frequency (1.8%). Its presence in Iberia (as that of H35 and H38) can be explained by gene flow from NW Africa. H38 (Fig. 2*d*) is the haplotype that clearly charactherises the NW African male lineages. Although it has also been described in Mali (27.3%) and Sudan (5%), a specific founder effect is the most likely explanation for its high frequency in NW Africa.

On the contrary, group VI haplotypes are preferently distributed across Europe and W Asia: H50 highest frequency is found in Sardinia (50%) and its geographical distribution is mainly limited to Europe. Within NW Africa, it is only found in Moroccan Arabs. H52 has only been found in Europeans, including Basques. H58 (Fig. 2*e*) has been described at low frequencies across Europe and Western and Central Asia. H71 (Fig. 2*f*) is found all around the Mediterranean basin as well as in Ethiopia, Sudan, Central Asia, Pakistan and India. H87, which belongs to group VIII, is represented by only three chromosomes (found in Morocco, the Indian subcontinent, and in a native American; Underhill et al., 1999) in a worldwide sample of 1,062 individuals.

Haplotypes belonging to group IX are most prevalent in Europe: H100, H101 (Fig. 2*g*) and H102 (Fig. 2*h*) have only been found so far in the Iberian Peninsula. H103 (Fig. 2*i*) is mainly present in Europe and Central Asia but it is also found in Northern

Africa. H107 is most frequent in Central and Western Asia although it has also been found at a low frequency in Europe.

Regarding the specific relation NW Africa-Iberian Peninsula, we find a clear differentiation of their male lineage content.While group III haplotypes prevail in NW Africa, Iberian haplotypes belong mostly to group IX. Moreover, in most cases of shared haplotypes between NW Africa and Iberia, those seem to show a larger geographical distribution encompassing Europe or the whole Mediterranean basin. We performed AMOVA in order to assess the fraction of Y chromosome genetic variance that could be attributed to the difference between Iberia and NW Africa. Whereas the fraction of genetic variance found among populations of the same region (NW Africa or Iberia) was 0.93% (non statistically significantly different from zero, p=0.065), the fraction of genetic variance attributed to the difference between the NW African and Iberian populations was 35.17% (statistically significantly different from zero, p=0.024). This result clearly confirms the large genetic difference between NW Africa and Iberia.

DISCUSSION

Specific founder effect for some NW African haplotypes: an Upper Palaeolithic differentiation?

H36 and its two directly derived haplotypes, H35 and H38, comprise 75% of the NW African Y chromosomes. Although these group III haplotypes are found elsewhere, their global frequency in NW Africa is by far the highest reported so far. In particular, H38 stands out, and clearly constitutes the male population core of this area. Their geographical distributions seem to suggest that they could had been introduced into NW Africa from Eastern Africa during a population expansion, probably ancient, such as the expansion of anatomically modern humans at the Upper Palaeolithic. Using *classical* genetic markers, Bosch et al. (1997) suggested that the NW African populations may have an important Upper Paleolithic background. This hypothesised Upper Palaeolithic expansion could have involved the descendants of the haplotypes differentiated from M35 that already presented two specific, derived mutations (M78, originating H35; and M81, originating H38). Alternatively, the expansion could have been brought Y chromosomes carrying an ancestor of both H35 and H38, although a

mutation that links both haplotypes and excludes the rest of group III is yet to be discovered. We estimated the age of M35 at 58,000 ± 22,000 years ago (ya), that of M78 at 15,000 ± 10,500 ya, and M81 at 30,200 ±11,500 ya. Then, the expansion that brought the ancestors of H35 and H38 into NW Africa can be bracketed between 30 and 58 Kya, that is, at the putative expansion of anatomically modern humans (or Upper Paleolithic).

Male Neolithic traces in NW Africa and Iberia

According to Underhill et al (1999), group VI haplotypes could had been spread with the Neolithic wave of expansion. This possible Neolithic component appears mainly represented in Iberia and NW Africa by haplotypes H58 and H71. Both haplotypes include 12f2*8Kb Y-chromosomes (see Appendix), which have been found all around the Mediterranean basin with their higher frequencies in the Middle East. We find similar frequencies for group VI haplotypes in Iberia (16.4%) and NW Africa (13.6%). The presence of this group of chromosomes in both regions could be due to two non-exclusive historical processes: i) the parallel, independent expansion of the Neolithic wave of advance along the northern and southern shores of the Mediterranean and ii) the early arrival of the Neolithic into one of the two regions and the subsequent crossing of the Gibraltar Straits. In order to assess the amount of genetic exchange across the Gibraltar Straits, we compared within each SNP-defined haplotype Y chromosome STR haplotypes constructed with eight markers that were available for a subset of the chromosomes currently analyzed (Pérez-Lezaun et al. 1997; Bosch et al. 1999). No STR haplotypes were shared among Iberian (n=4) and NW African (n=16) H71 chromosomes. The differences in number of repeat units among Iberian and NW African H71 chromosomes ranged from 4 to 13, with a mean of 10.11. Within NW African H71 chromosomes, the mean difference was 2.40. Thus, both sets of chromosomes were clearly distinct, as confirmed also by a median-joining network (Bandelt, 1995). This lends support for independent origins of H71 (and probably of the whole haplogroup VI) chromosomes in Iberia and NW Africa as they have been accumulating STR mutations independently. The average square distance (ASD; Goldstein et al. 1995) between NW Africa and Iberian H71 chromosomes is 1.99, which translates, for a mutation rate of 2.1 10^{-3} (Heyer et al. 1997) into a separation time of 474 generations or 9,480 ya, which is compatible with a split in the Middle East and subsequent parallel Neolithic expansions along both Mediterranean shores.

The European Paleolithic background

Group IX haplotypes show a geographical distribution compatible with a proto-European, probably Paleolithic, origin (Underhill et al. 1999) well differentiated from Africa. Group IX encompasses three local haplotypes (H100, H101, and H102) in Iberia. However, if a wider sample of the European diversity had been analysed they might have been found within other populations, but at a much lower frequency. All those Iberian-specific haplotypes have been found in Basques, although H102 is very frequent in Catalans. Only 2.8% of the NW African chromosomes belong to group IX and, in this case, all chromosomes present H103, the ancestral haplotype for all group IX. On the contrary, 55.7% of the Iberian Y-chromosomes present haplotype H103. The comparison of the STR haplotypes within H103 chromosomes was used to assess whether that hipothesized proto-European chromosomes found in NW Africa may have been carried across the Gibraltar Straits. STR haplotypes in four out of five H103 NW African chromosomes were one mutation step away from H103 Iberian chromosomes, while the fifth was two mutation steps away. Moreover, the mean repeat size difference within 31 H103 Iberian STR haplotypes was 3.29 (0-11). This STR haplotype similarity seems to indicate that H103 chromosomes found in NW Africa are a subset of the Iberian gene pool and may have been introduced in recent times.

Sub-Saharan gene flow into NW Africa

H28 and H22, which belong to group III, show a Sub-Saharan distribution pattern (Underhill et al. 1999) with their highest frequencies, respectively, in Mali (29.5%), and in South Africa (51%) and Central Africa (57%). While H28 probably originated in the Sahel, the derived haplotype H22 may have arisen in central or southern Africa. Both haplotypes comprise 8% of the NW African Y chromosomes and, given their geographical distribution, their presence in NW Africa can be interpreted as resulting from Sub-Saharan gene flow. The NW African contact with the Southern peoples was specially important during the Almoravid Berber expansion (1056-1147), and it has been mainly maintained till recently by the Trans-Saharan commercial routes (Kasule, 1998).

NW African homogeneity

Once the major historical components of the NW African Y chromosomes were identified (i.e., North African, Neolithic and Sub-Saharan), we explored whether there were differences among the NW African populations. Non statistically significantly differences were found (χ^2=4.85, p=0.563). Thus, NW Africa can be considered homogenous in its male lineage content. Although non-statistically significantly differences were found (χ^2=3.197, p=0.202), the Moroccan Arabs seem to show an increase of what has been considered as the possible Neolithic component and a slight reduction of the NW African autochthonous component. This pattern could be consequence of the Arabization, the process of cultural change that brought the Arabic language and the Islamic religion from the Middle East in the 7th – 11th centuries AD. This process could have brought Middle Eastern Y chromosomes into NW Africa, enriching the local pool with chromosomes from the center of the Neolithic wave of advance. However, it must be noted that haplotype frequencies in Moroccan Arabs are essentially the same as in the other NW African populations, and, thus, the number of chromosomes brought in with the Arabization may have been low. Thus, the Arabization could be regarded mainly as a cultural replacement phenomenon, in an élite dominance process (Renfrew, 1987).

NW Africa and the Iberian Peninsula

The historical origins of the Iberian Y chromosome pool may be summarised as 5.2% NW African, 16.5% Neolithic and 78.4% proto-European or Iberian-specific. No sub-Saharan African-specific haplotype was found in the Iberian Peninsula. H35, H36 and H38, which have been considered as the possible NW African component in Iberia, comprise 75% of the Y chromosomes in NW Africa. Thus, the maximum NW African contribution to the Iberian Y chromosome pool can be estimated at 7%. In particular, this NW African contribution in Iberia is found in four out of 37 Andalusians (10.8%) and in one out of 44 (2.3%) Basques. Although the possible difference observed between northern and southern Iberia, we found non-statistically significant differences among the Iberian populations (χ^2=4.038, p=0.133). In fact, the Basque individual presumed to present a NW African Y chromosome has the same STR haplotype that is most frequent in H38 Y chromosomes in NW Africa and that is not found in other haplotype backgrounds. Thus, it can be concluded that the NW African contribution to the Iberian Y chromosome pool has been very low (7%, maximum value), even in Andalusia

(14.4%, maximum value), the southernmost Iberian population. From 711 AD, the Islamic invaders from NW Africa conquered Spain. They left a rich cultural heritage, from language and religion to agriculture and architecture. Their occupation was longest in the South, where they were ousted from political power in 1492. However, our results show that their demographic contribution in Iberia must have been small, which runs contrary to the historical perception.

The typing of the DHPLC set of markers has allowed a clear dissection of the NW African male mediated gene pool into several layers of historic origins. Future work may extend such analysis to other regions, in which the effects of some historical or demographic processes, as well as, the relation with neighbouring areas are poorly understood.

METHODS

Samples

Different autochthonous samples from NW Africa and the Iberian Peninsula were typed. Northwestern African samples comprised 29 Saharawis, 40 southern Moroccan Berbers, 44 Morroccan Berbers and 63 northern central Moroccan Berbers; and the Iberian Peninsula was represented by 37 Andalusians and 16 Catalans, and by 44 Basques typed in Underhill et al. 1999. Appropriate informed consent was obtained from all participants in this study and information about geographic origin of their four grandparents and maternal tongue was recorded. DNA was extracted from fresh blood by standard phenol-chloroform protocols.

Biallelic polymorphism typing

All samples in this study were characterised by means of a top-down approach in which markers were successively typed in hierarchical order according to their position in the genealogy given in Underhill et al. 1999. With the exception of M1 (also known as YAP polymorphism) denaturing high performance liquid chromatography (DHPLC) was used to genotype the biallelic markers studied. Marker information such as ancestral and derived alleles, primer sequences and PCR conditions for their amplification, as well as additional details for their typing conditions by DHPLC can be found in Underhill et al 1997, 1999. YAP polymorphism was assayed as described in Hammer and Horai (1995).

Data analysis

Differences in haplotype frequencies among populations from NW Africa and the Iberian Peninsula were tested through Analysis of Molecular Variance (AMOVA) by using the Arlequin package (Schneider et al. 1997). We performed two levels of analysis, that is, we explored haplotype frequencies among populations within each region separately, and by pooling them in two hierarchical partitions, corresponding to NW Africa and Iberian Peninsula. In order to estimate mutation ages for M35, M78 and M81, we applied coalescence analysis on NW Africa haplotype frequencies using the same methods and parameters described by Griffiths and Tavaré (1994) and Hammer et al (1998).

ACKNOWLEDGEMENTS

We express our appreciation to the original blood donors who made this study possible. We thank all persons involved in reaching the Saharawi donors as well as Elisabeth Pintado (Sevilla), Josep Lluís Fernández Roure and Alba Bosch (Mataró) for their help in contacting Moroccan donors. We especially thank A.A. Lin and P. Shen for their technical support. This research was possible thanks to E.B stay in L.L. Cavalli-Sforza's laboratory at Stanford University. This work was supported by Dirección General de Investigación Científica y Técnica in Spain (PB95-0267-CO2-01 and PB98-1064) and by Direcció General de Recerca, Generalitat de Catalunya (1998SGR00009). Comissionat per a Universitats i Recerca, Generalitat de Catalunya supported E.B. (FI/96-1153).

REFERENCES

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

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

Bosch E, Calafell F, Santos FR, Pérez-Lezaun A, Comas D, Benchemsi N, Tyler-Smith C, et al. (1999) Variation in Short Tandem Repeats is deeply structured by genetic background on the human Y chromosome. Am J Hum Genet 65:1623-1638

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 USA 92:6723-6727

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307-319

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951-962

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, et al. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427-441

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799-803

Hurles,M.E., Irven,C., Nicholson,J., Taylor,P.G., Santos,F.R., Loughlin,J., Jobling,M.A. and Sykes,B.C. (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63: 1793-1806.

Jobling,M.A. and Tyler-Smith,C. (1995) Fathers and sons: the Y chromosome and the human evolution. Trends Genet 11: 449-456.

Jobling,M.A., Pandya,A. and Tyler-Smith,C. (1997) The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 110: 118-124.

Kasule S (1998) The history Atlas of Africa. Macmillan, New York

Pérez-Lezaun A, Calafell F, Seielstad MT, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y chromosome short tandem repeats in humans. J Mol Evol 45:265-270

Renfrew C (1987) Archaeology and Language. The Puzzle of Indoeuropean origins. Jonathan Cape, London (UK)

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, et al. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Gen Res 7:996-1005

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, et al. (1999) The architecture of the Y-chromosome biallelic haplotype diversity: an emerging portrait of mankind. Submitted.

TABLES & FIGURES

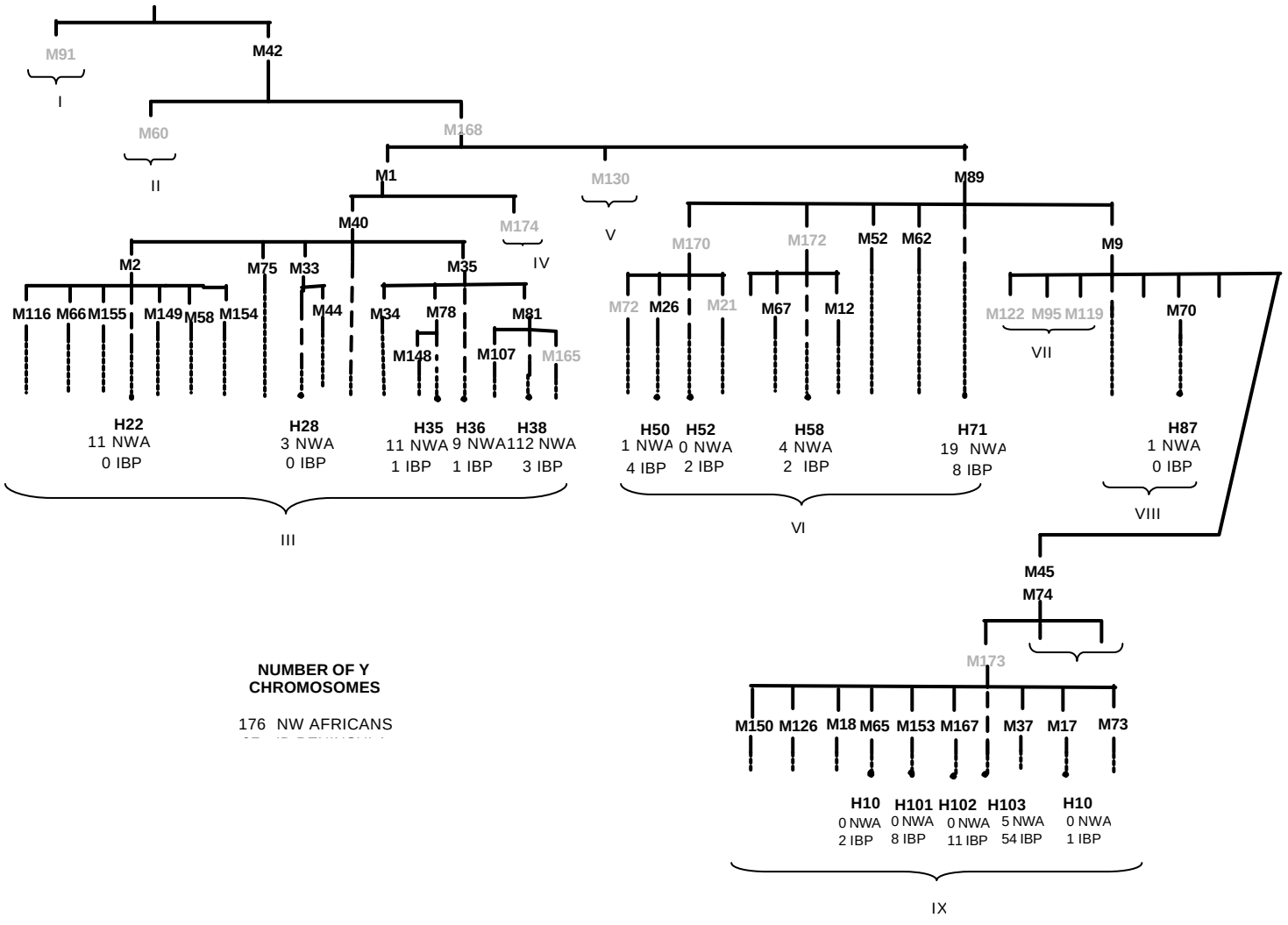
Fig. 1: phylogenetic tree of Y chromosome haplotypes and their absolute frequencies in NW Africa and Iberian Peninsula.

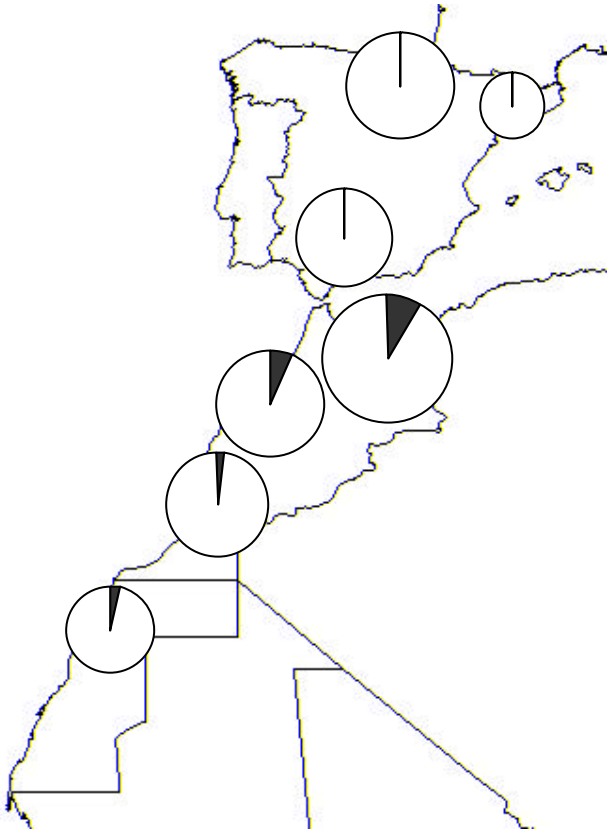
Fig. 2: frequency of Y chromosome haplotypes in NW Africa and Iberian Peninsula. Upper right: phylogenetic position of the haplotype. Lower right: worldwide distribution. Circle areas are proportional to sample size. *a*, H22; *b*, H35; *c*, H36; *d*, H38; *e*, H58; *f*, H71; *g*, H101; *h*, H102 and *i*, H103.

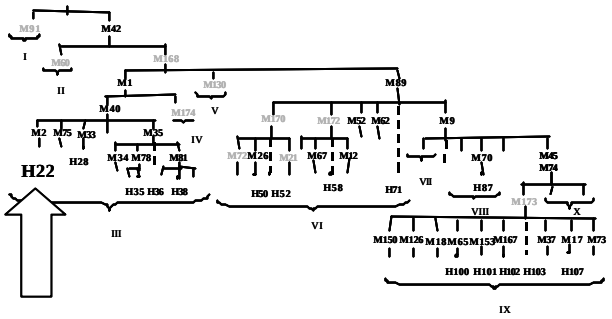
Table 1: haplotype frequencies in NW African and Iberian populations. Groups and haplotypes appear as in Underhill et al. 1999.

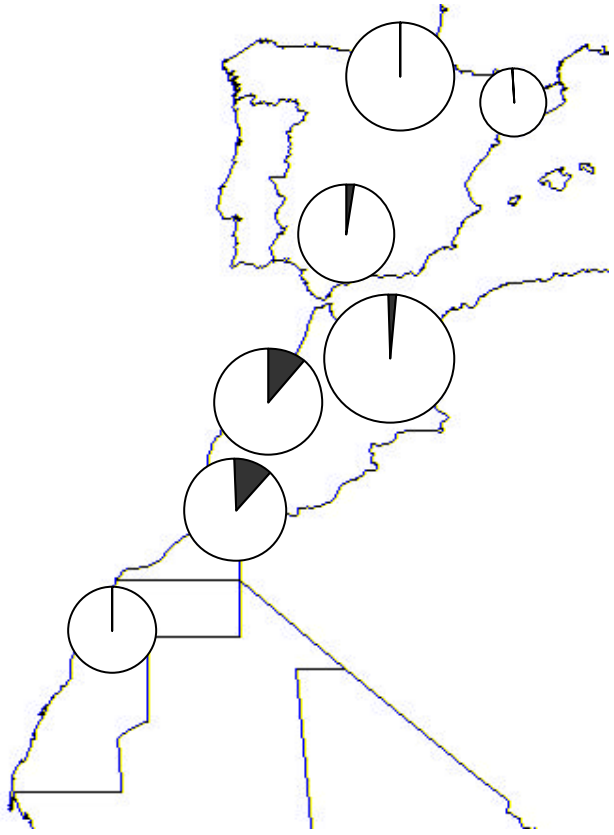
Groups	**III**						**VI**					**VIII**	**IX**					
Haplotypes	**H22**	**H28**	**H35**	**H36**	**H38**	**Total**	**H50**	**H52**	**H58**	**H71**	**Total**	**H87**	**H100**	**H101**	**H102**	**H103**	**H107**	**Total**
NW Africa																		
Saharawis																		
N=29	1	1	-	-	22	24	-	-	-	5	5	-	-	-	-	-	-	0
SM Berbers																		
N=40	1	-	5	3	26	35	-	-	1	3	4	-	-	-	-	1	-	1
M Arabs																		
N=44	3	-	5	1	23	32	1	-	1	6	8	1	-	-	-	3	-	3
NCM Berbers																		
N=63	6	2	1	5	41	55	-	-	2	5	7	-	-	-	-	1	-	1
Total																		
N=176	11	3	11	9	112	146	1	0	4	19	24	1	0	0	0	5	0	5
Iberian Peninsula																		
Andalusians																		
N=37	-	-	1	1	2	4	2	-	2	4	8	-	-	1	1	22	1	25
Catalans																		
N=16	-	-	-	-	-	0	-	1	-	3	4	-	-	-	5	7	-	12
Basques*																		
N=44	-	-	-	-	1	1	2	1	-	1	4	-	2	7	5	25	-	39
Total																		
N=97	0	0	1	1	3	5	4	2	2	8	16	-	2	8	11	54	1	76

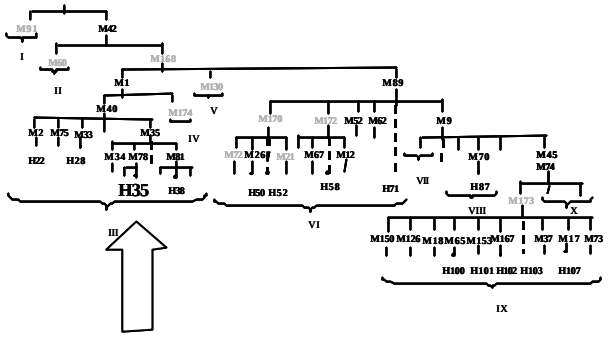
Abbreviations: SM Berbers, southern Moroccan Berbers; M Arabs, Moroccan Arabs and NCM Berbers, northern central Moroccan Berbers. Note: (*) Data from Underhill et al. 1999.

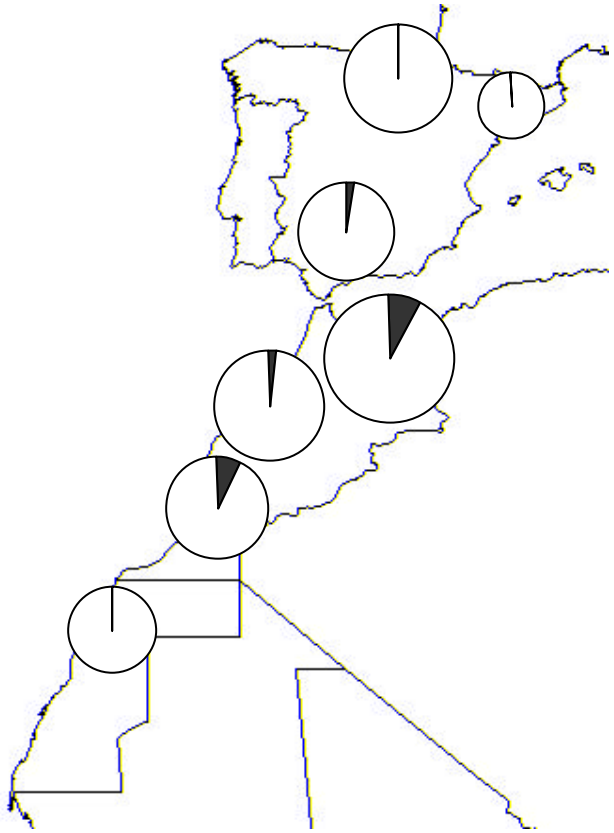


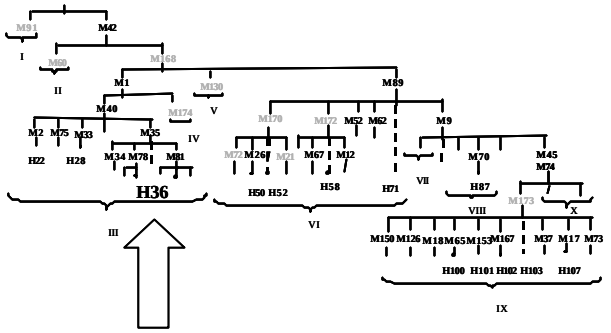


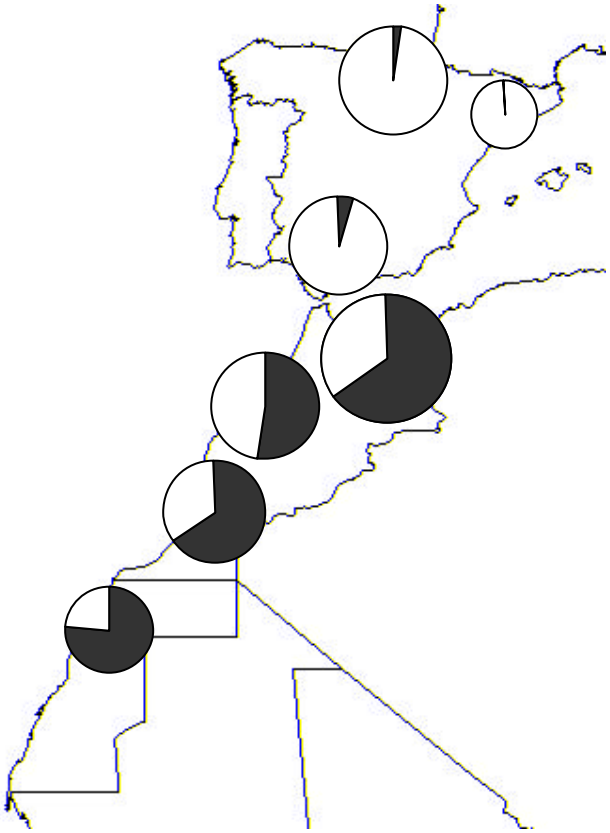

M91
M42
I
M168
M69
II
M1
M130
M40
M174
V
M170
M172
M52
M62
M9
M2
M75
M33
M35
IV
M72
M26
M21
M67
M12
M70
M45
M74
H28
M34
M78
M81
H22
H87
M173
H35
H36
H38
H50
H52
H58
H71
VII
VIII
X
III
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
IX

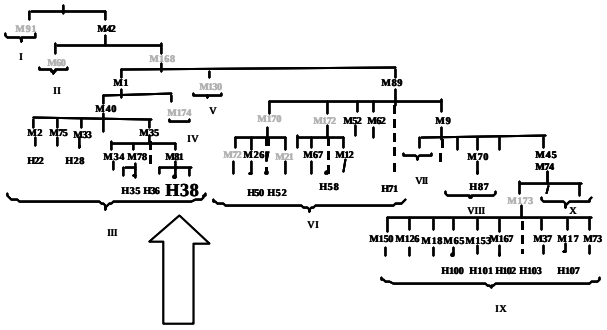


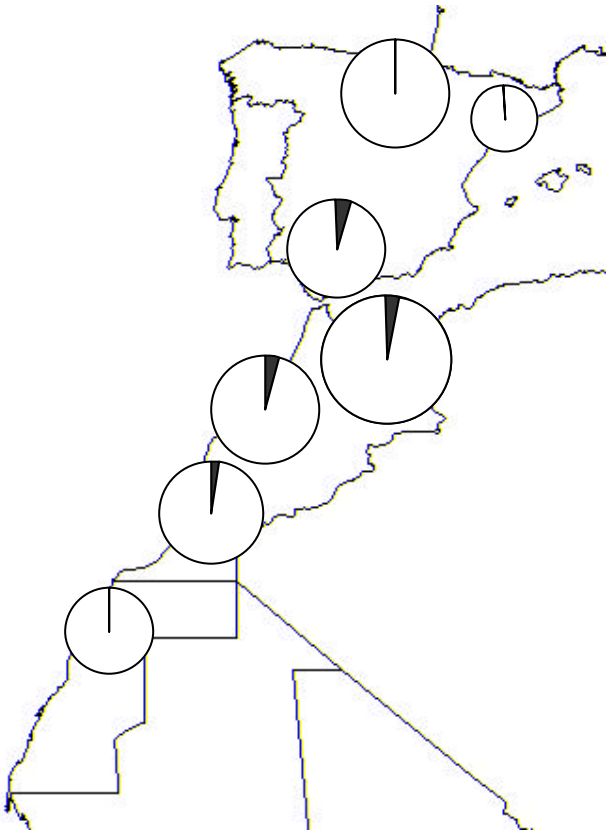

M91
M42
I
M168
M69
II
M1
M130
M89
V
M40
M174
M170
M172
M52
M62
M9
M2
M75
M33
M35
IV
M72
M26
M21
M67
M12
M70
M45
M24
H22
H28
M34
M78
M81
H35
H38
H50 H52
H58
H71
VII
H87
M173
III
VI
VIII
X
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
IX

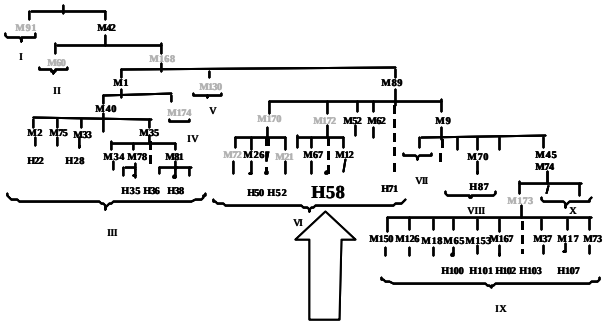


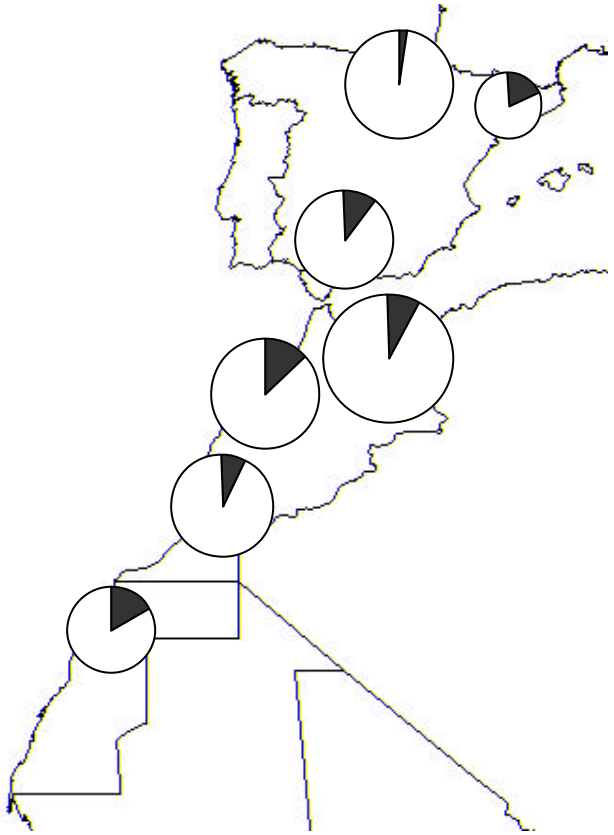

M91
M42
I
M68
M168
M69
II
M1
M130
M89
M40
M174
V
M170
M172
M52
M62
M9
M2
M75
M33
M35
IV
M22
H28
M34
M78
M81
M72
M267
M21
M67
M12
M70
M45
M24
VII
H87
M173
VIII
X
H36
H50 H52
H58
H71
III
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
IX

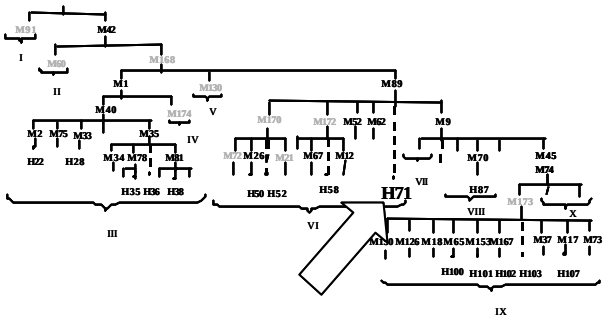


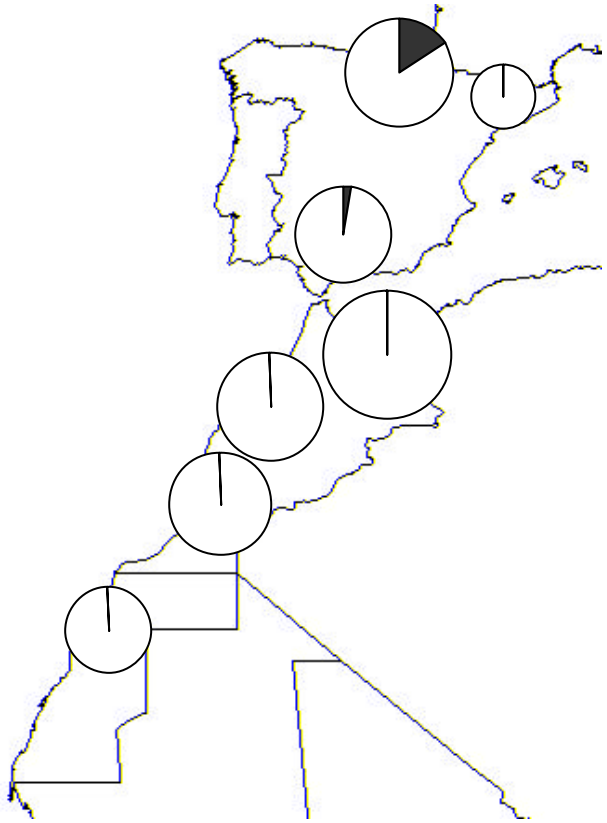

M91
M42
I
M69
M168
II
M1
M130
M89
M40
M174
V
M2
M75
M33
M35
IV
H22
H28
M34
M78
M31
H35
H36
H38
III
M170
M172
M52
M62
M9
M72
M26
M21
M67
M12
H50
H52
H58
H71
VII
M70
M45
M74
H87
VIII
M173
X
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
IX

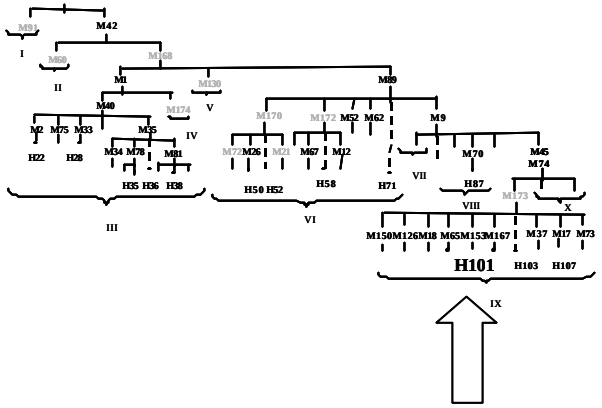


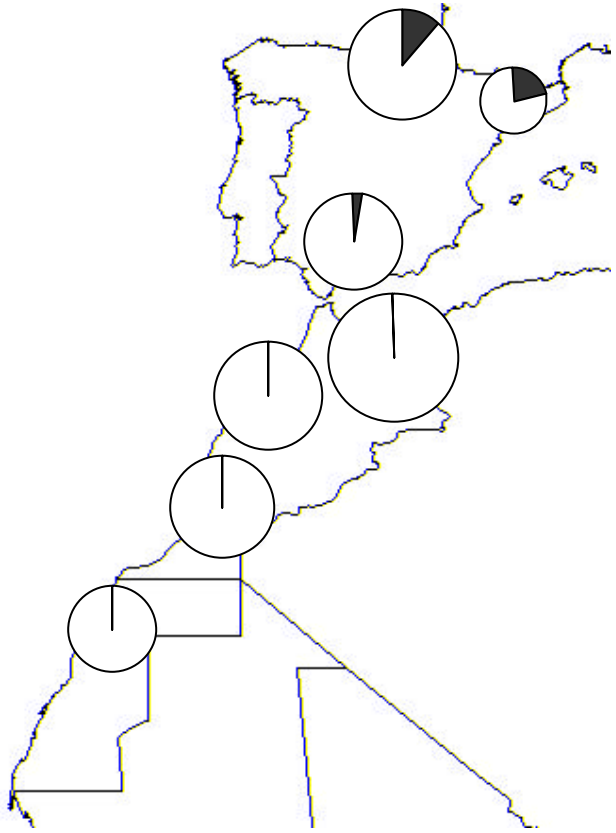

M91
M42
I
M69
M168
II
M1
M130
M89
V
M40
M174
M270
M272
M52
M62
M9
M2
M75
M33
M35
IV
M72
M267
M21
M67
M12
M70
M45
M74
H22
H28
M34
M78
M81
H87
M173
VII
H35 H36 H38
H50 H52
H58
H71
VIII
X
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
III
IX

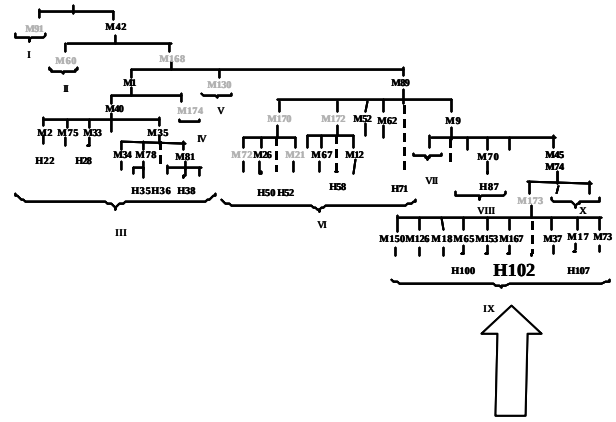


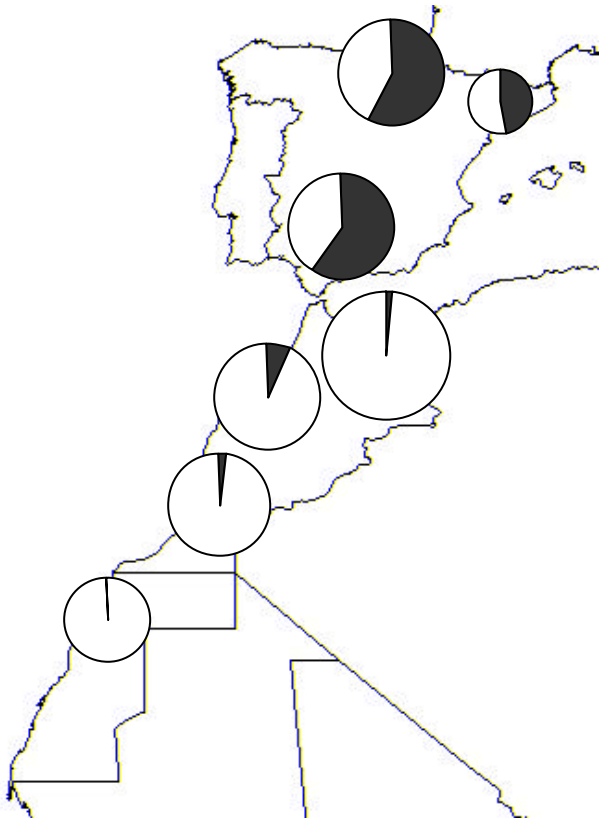

M91
M42
I
M68
M168
II
M1
M130
V
M89
M40
M174
M2
M75
M33
M35
IV
M170
M172
M52
M62
M9
H22
H28
M34
M78
M31
M70
M45
M24
M72
M26
M21
M67
M12
VII
H35
H36
H38
H50
H52
H58
H71
H87
M173
VIII
X
III
VI
M130
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H102
H103
H107
IX

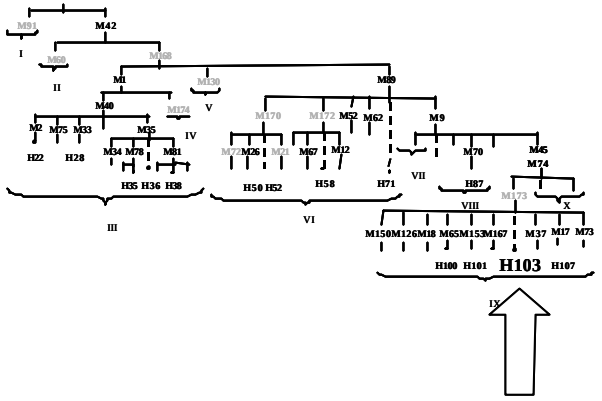



M91
M42
I
M60
M168
II
M1
M130
M89
V
M40
M174
M170
M172
M52
M62
M9
M2
M75
M33
M35
IV
H22
H28
M34
M78
M81
M72
M26
M21
M67
M12
M70
M45
M74
VII
H35
H36
H38
H50
H52
H58
H71
H87
M173
VIII
X
III
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H101
H103
H107
IX




M91
M42
I
M60
M168
II
M1
M130
M89
V
M40
M174
M170
M172
M52
M62
M9
M2
M75
M33
M35
IV
M34
M78
M81
M72
M26
M21
M67
M12
M70
M45
M74
H22
H28
H35
H36
H38
H50
H52
H58
H71
VII
H87
M173
VIII
X
III
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H102
H107
IX




M91
M42
I
M89
M168
M1
M130
II
M89
M40
M174
V
M2
M75
M33
M35
IV
H22
H28
M34
M78
M81
H35
H36
H38
III
M170
M172
M52
M62
M9
M23
M26
M21
M67
M12
H50
H52
H58
H71
VII
M70
M45
M74
H87
VIII
M173
X
VI
M150
M126
M18
M65
M153
M167
M37
M17
M73
H100
H101
H103
H107
IX

APPENDIX

A different classification of Y-chromosome haplotypes into haplogroups by using biallelic polymorphisms was described by Jobling and Tyler-Smith 1995; Jobling et al. 1997; Hurles et al. 1998 among others. Since both sets of markers map onto the non-recombining portion of the Y chromosome, they have evolved along, a single, common phylogeny. A subset of the chromosomes analysed in this study had previously been typed for that set of markers (Bosch et al. 1999). The correspondences between DHPLC-based haplotypes and haplogroups defined by that other set of markers are presented in Table a.

Table a: relation haplotypes-haplogroups. Notes: (*) In these cases, the same polymorphism defines the last derived marker in both sets. ([1]) Last derived marker that defines the haplotypes in the DHPLC set as described in Underhill et al. (1999) and ([2]), last derived marker that defines the haplogroups as described in Jobling and Tyler-Smith 1995; Jobling et al. 1996, 1997; Hurles et al. 1998 among others

GROUPS	*HAPLO TYPES*	**LAST DERIVED MARKER[1]**	**HAPLO-GROUPS**	**LAST DERIVED MARKER[2]**	**Number of Y chromosomes typed**
	H22	***M2****	***HG8***	***sY81****	6
	H28	M33	HG21 (L)	SRY-8299 (L)	1
III	H35	M78	HG21	SRY-8299	11
	H36	M35	HG21	SRY-8299	4
	H38	***M81***	HG21	***SRY-8299***	83
	H50	***M26***	HG2	***SRY-1532***	3
VI	H52	M170	HG2	SRY-1532	1
	H58	M172	***HG9***	12f2	2
	H71	**M89**	HG2 / HG9	**SRY-1532 / 12f2**	5 / 13
VIII	H87	M70	HG26	M9	1
	H100	M65	HG1	92R7	2
IX	H101	M153	HG1	92R7	7
	H102	M167*	HG22	SRY-2627*	11
	H103	***M173***	HG1	92R7	37

DISCUSSIÓ

La recopilació de dades en la literatura sobre marcadors *clàssics* en poblacions humanes del nord d'Àfrica ha permès d'establir una primera anàlisi de referència per emmarcar la variabilitat genètica de les poblacions d'aquesta regió. Clarament, la principal característica del paisatge genètic obtingut a partir d'aquests marcadors és un patró de variació est-oest. En l'anàlisi de detecció de fronteres genètiques, en els diferents arbres *neighbor-joining* calculats i en la primera component principal, observem una clara diferenciació entre el nord-est (Líbia més Egipte) i les poblacions àrabs i berbers del nord-oest, sent el primer grup el que mostra major semblança amb les poblacions europees considerades. Cal assenyalar també que les distàncies genètiques calculades mostren una clara separació entre les poblacions ibèriques i les poblacions nordafricanes.

El patró observat ens portà a suggerir una primera hipòtesi sobre el poblament del nord d'Àfrica basada en una onada d'expansió poblacional durant el neolític, la qual hauria tingut escàs impacte en les regions del nord-oest. La similitud entre el nord-est d'Àfrica i Europa seria compatible amb un procés de flux gènic des de l'Orient Mitjà, que de manera paral·lela avançaria cap a l'oest al llarg de les dues ribes de la Mediterrània. Per altra banda, una diferenciació mesolítica (o anterior, probablement al paleolític superior) de les poblacions del nord-oest d'Àfrica amb un flux gènic limitat posterior podria explicar el paisatge genètic observat. Detectem, per tant, que una part important del rerafons genètic de les poblacions del nord-oest podria correspondre al romanent d'una diferenciació pre-neolítica, de manera similar com s'ha descrit per a la població basca (Calafell i Bertranpetit, 1994).

A partir de la informació aportada pels marcadors *clàssics*, no podem discernir cap estructura clara entre els diferents grups àrabo- o berber- parlants. Tanmateix, en conjunt, els àrabs sembla que es trobin genèticament més propers a Egipte i Líbia. Per altra banda, tot i que la tercera component principal al nord d'Àfrica sembla mostrar un patró de diferenciació nord-sud, l'abast d'una possible contribució genètica sud-sahariana sembla ser petit. Cal assenyalar, però, que les anàlisis realitzades a partir d'aquests marcadors no són les més precises per a quantificar-la.

L'anàlisi de polimorfismes clàssics ens ha servit per a esbossar les principals qüestions relatives a la genètica de les poblacions del nord-oest d'Àfrica. A continuació, discutiré l'evidència genètica pertinent a les següents hipòtesis: el poblament paleolític al nord-oest d'Àfrica; l'expansió del neolític a ambdues ribes de la Mediterrània i el seu impacte genètic al nord-oest d'Àfrica; l'abast de l'aportació genètica sud-sahariana; les possibles conseqüències genètiques de l'arabització al nord-oest d'Àfrica i, finalment, el flux gènic que pot haver travessat l'estret de Gibraltar en ambdues direccions.

El poblament paleolític al nord-oest d'Àfrica

Quan estudiem la composició dels haplotips construïts a partir de marcadors bial·lèlics específics del cromosoma Y tipificats mitjançant DHPLC (capítol V), trobem que l'haplotip H36 i dos haplotips directament derivats d'ell, H35 i H38, constitueixen el 75% dels cromosomes del nord-oest d'Àfrica. Aquest conjunt d'haplotips comprenen tots els cromosomes nordafricans que pertanyen a l'haplogrup 21 (capítol IV), el qual suggerim que podria haver-se diferenciat al nord-oest d'Àfrica per deriva i/o efecte fundador. La major resolució que ofereix el nou ventall de marcadors específics del cromosoma Y descrits per Underhill et al. (1999) amplia clarament les possibilitats d'ànalisi a l'hora d'establir un origen i situar en el temps els haplogrups descrits inicialment. Al nord-oest d'Àfrica, destaca especialment l'elevada freqüència de l'haplotip H38 (63.4% del total), la més alta mai descrita, el qual clarament constiuteix una gran part dels llinatges masculins d'aquesta regió. Les freqüències més altes per a H36 han estat descrites a Etiòpia (6.8%) i entre la població khoisànida (10.3%) mentre que H35 presenta les seves freqüències més altes a Etiòpia (22.7%), Sardenya (18.2%) i Sudan (17.5%). Hem calculat l'edat de la mutació M35 (que defineix H36) en 58.000 anys, mentre que les dues mutacions (M78 i M81) presents només en cromosomes derivats per M35 i que defineixen, respectivament, H35 i H38 han estat datades a 15.000 i 30.000 anys. Per tant, la distribució geogràfica de H36 i dels seus dos haplotips derivats, juntament amb les edats que hem estimat per aquests polimorfismes i les hipòtesis sobre moviments de població que coneixem a partir de l'arqueologia ens porten a hipotetitzar un origen a l'est d'Àfrica per a M35, la qual fou portada, potser ja amb els haplotips derivats H35 i H38 cap al nord-oest d'Àfrica amb

el poblament associat a l'expansió del paleolític superior, fa entre trenta i quaranta mil anys. Un cop al nord-oest d'Àfrica, potser per un efecte fundador o per l'acció de la deriva en un temps més perllongat, H38 assolí les elevades freqüències a les quals es troba actualment. Cal assenyalar que H36 i els seus derivats es troben a freqüències més baixes (inclús en algun cas, són absents) a l'Orient Mitjà que no pas a l'est d'Àfrica. Aquesta distribució podria suggerir que el poblament del nord-oest d'Àfrica fou independent del de l'Orient Mitjà. Hom suposa que els humans anatòmicament moderns arribaren a Europa durant el paleolític superior procedents de l'Orient Mitjà; aquest no sembla ser el cas per el nord-oest d'Àfrica. És possible que els humans anatòmicament moderns es desplacessin independentment des de l'est d'Àfrica cap a l'Orient Mitjà i cap al nord-oest d'Àfrica. La persistència d'un possible substrat paleolític suggerida a partir de la compilació de polimorfismes *clàssics* en el nostre primer treball (capítol I) sembla doncs que caracteritza el llinatge masculí d'aquestes poblacions.

Cal assenyalar que els haplotips H35, H36 i H38 engloben tots els cromosomes nordafricans que pertanyen a l'haplogrup 21 (capítol IV) i, en el seu conjunt, són equivalents a l'haplotip 4 descrit a Hammer et al. (1997; 1998). Aquests autors van descriure la variabilitat en el cromosoma Y a partir d'un conjunt de marcadors bial·lèlics específics del cromosoma Y que se solapen parcialment amb els nostres, la qual cosa permet d'establir equivalències entre els haplotips definits a partir d'ambdós grups de marcadors. En aquest cas, Hammer i col·laboradors (1997) en analitzar 15 poblacions d'arreu del món (sense incloure cap població del nord-oest d'Àfrica) van trobar que la freqüència més elevada per l'haplotip 4 (aproximadament del 50%) corresponia a una mostra de 93 egipcis. Posteriorment, Scozzari et al. (1999) analitzaren els mateixos marcadors que Hammer et al. (1997; 1998) en 25 poblacions africanes que incloïen una mostra de 56 àrabs marroquins. En aquesta mostra, la freqüència de l'haplotip 4 era del 70%, altre cop, la més elevada respecte la resta de poblacions analitzades. Ambdós grups d'investigadors, utilitzant diferents metodologies que comentaré més endavant, estimen l'edat de l'avantpassat comú més recent per l'haplotip 4 a 38.800 anys (Hammer et al. 1998) i 20.000 anys (Scozzari et al. 1999). Aquestes estimes són compatibles amb una diferenciació en el paleolític superior. Per tant, l'anàlisi de marcadors genètics al cromosoma Y per altres autors confirma, per una banda, la diferenciació dels llinatges masculins de les poblacions del nord-oest d'Àfrica tal com descrivim als capítols IV i V, i per l'altra, apunta també cap a un origen paleolític per al principal component d'aquests llinatges masculins.

Per tant, els diferents marcadors bial·lèlics específics del cromosoma Y emprats en aquest treball, i, com s'ha comentat anteriorment, els marcadors *clàssics* semblen revelar un important component paleolític en el *pool* genètic de les poblacions del nord-oest d'Àfrica. Per altra banda, la persistència d'un substrat paleolític al nord-oest d'Àfrica podria explicar parcialment la diferenciació observada en l'anàlisi dels microsatèl·lits autosòmics entre la Península Ibèrica i el nord-oest d'Àfrica.

Un cop diferenciat el substrat paleolític de la població del nord-oest d'Àfrica, cal intentar entendre com aquest substrat no ha estat diluït per posteriors aportacions poblacionals, com l'onada d'expansió del neolític, de la qual tractarà la següent secció. Tal i com suggeríem inicialment en el treball del capítol I, la persistència genètica paleolítica sembla tenir un reflex cultural. En efecte, la cultura capsiana (7.000 – 5.000 aC) s'inicia en el paleolític i molts dels seus elements perviuen en el neolític tot i el canvi en el mode d'obtenció de l'aliment que suposà el neolític (Desanges, 1990; Newman 1995). Per tant, sembla que la difusió del neolític al nord-oest d'Àfrica fou un procés cultural amb una expansió poblacional limitada, la qual cosa ens ajuda a entendre la persistència d'un substrat genètic paleolític tan important.

L'expansió del neolític a ambdues ribes de la Mediterrània i el seu impacte genètic al nord-oest d'Àfrica

Segons Underhill i col·laboradors (1999), els haplotips del cromosoma Y englobats dins el grup IV podrien ser contemporanis a l'onada d'expansió del neolític. Aquest possible component neolític l'hem trobat representat a la Península Ibèrica (16.5% del total) i al nord-oest d'Àfrica (13.7% del total) majoritàriament pels haplotips de marcadors bial·lèlics específics del cromosoma Y H58 i H71. Cal assenyalar que ambdós haplotips inclouen cromosomes Y amb l'al·lel derivat per a la mutació 12f2 (capítol IV). La distribució geogràfica d'aquest al·lel ha estat abastament estudiada per diversos autors, els quals el troben al llarg de tota la costa mediterrània amb freqüències especialment elevades a l'Orient Mitjà (Brega et al. 1987, Mitchell et al. 1993, 1997; Semino et al. 1995). Mentre que alguns autors suggereixen que aquest patró de distribució podria haver estat provocat pels colonitzadors fenicis (Mitchell i Hammer 1996) altres suggereixen que podria haver resultat de l'onada d'expansió dèmica del neolític (Gonçalves i Lavinha 1994; Semino et al. 1996). En el present

treball, la comparació d'haplotips de microsatèl·lits construïts amb vuit d'aquests marcadors en un total de vint cromosomes H71 procedents del nord-oest d'Àfrica i de la Península Ibèrica mostrà que ambdós grups de cromosomes H71, considerats per d'altres autors com a neolítics, estaven clarament diferenciats. Aquesta darrera observació és compatible amb la hipòtesi suggerida a partir de l'anàlisi realitzada amb la informació aportada pels marcadors *clàssics* on, com s'ha comentat anteriorment, proposem una expansió paral·lela i independent del neolític al llarg de les dues costes de la Mediterrània. L'anàlisi de microsatèl·lits autosòmics en poblacions d'ambdues regions mostra igualment una important diferenciació entre els seus *pools* genètics tant en els arbres genètics obtinguts a partir de distàncies genètiques, com en la anàlisi de coordenades principals i de detecció de les principals fronteres genètiques.

Simoni i col·laboradors (1999) analitzen deu freqüències al·lèliques de marcadors *clàssics* en 39 poblacions de la regió mediterrània i identifiquen quines són les zones de major canvi genètic. En aquest cas, les principals fronteres genètiques clarament separen les costes nord i sud, especialment en les seves regions més occidentals. Aquesta observació ha estat replicada per Kandil et al. (1999) utilitzant mètodes i marcadors *clàssics* diferents. En aquest cas, es tracta de set sistemes genètics estudiats en 22 poblacions d'Europa, nord d'Àfrica i Orient Mitjà que s'analitzen mitjançant components principals, i de nou sistemes genètics estudiats en deu poblacions per a calcular distàncies genètiques. Simoni i col·laboradors (1999) atribueixen la gran diferència genètica observada a la possibilitat que la Península Ibèrica i el nord-oest d'Àfrica siguin els extrems d'una expansió poblacional paral·lela al llarg de les dues ribes de la Mediterrània. Donada la natura de les seves dades, no poden situar amb precisió en el temps aquesta expansió i suggereixen que es podria haver donat en el paleolític superior i/o en el neolític. En aquest sentit, la seva interpretació és compatible amb la nostra hipòtesi referent a l'expansió del neolític per camins paral·lels a ambdues ribes de la Mediterrània. En el nostre cas, però, les dades són més detallades, permetent no solament diferenciar els haplotips que s'expandiren en el neolític, sinó també la importància relativa respecte a cromosomes Y d'altres origens, ja sigui en el temps (com els paleolítics) o en l'espai (com els d'origen sud-saharià).

Com s'ha comentat en a la introducció, l'arqueòleg Colin Renfrew va proposar una hipòtesi que explicava alhora l'expansió de quatre famílies lingüístiques, que, segons d'altres autors, estan relacionades entre elles: l'afroasiàtica cap al nord

d'Àfrica, l'altaica cap a l'Àsia central, la dravidiana cap a l'Índia i la indoeuropea cap a Europa. El nexe que uniria totes aquestes famílies, segons Renfrew (1991), és l'expansió del neolític des de l'Orient Mitjà, fa entre 12.000 i 10.000 anys, la qual desplaçaria la caça i la recol·lecció per l'agricultura i la ramaderia com a fonts principals d'aliment. Aquest canvi resultà en un increment de la població i la necessitat d'ocupar nou territoris, provocant una expansió poblacional. La hipòtesi d'una expansió poblacional i lingüística conjunta des de l'Orient Mitjà té conseqüències que podem verificar. Un exemple seria l'existència de gradients o clines genètiques. És a dir, els parlants d'aquestes famílies lingüístiques s'assemblarien tant menys als habitants putatius de l'Orient Mitjà com més lluny en visquessin. Tanmateix, Barbujani i col·laboradors (1994), a partir d'un estudi amb informació sobre polimorfismes *clàssics,* van comprovar que aquests gradients genètics es donaven per a totes les famílies excepte per a l'afroasiàtica, i que aquests gradients desapareixien si no es tenia en compte la família lingüística a la qual pertanyien les poblacions. L'excepció afroàsiatica sembla indicar que l'expansió neolítica al nord-oest d'Àfrica devia constituir una aportació genètica relativament petita, tot i que va comportar un canvi dràstic en els mecanismes d'obtenció d'aliments, i, potser, una substitució lingüística amb l'arribada de les llengües berbers (pertanyents a la família afroasiàtica). A la vegada, és compatible amb el fet de que bona part dels gens dels pobladors paleolítics del Magreb persisteix en les poblacions actuals tal i com proposàvem en la secció anterior.

Homogeneïtat gènetica en les poblacions del nord-oest d'Africa

Al llarg de tot el treball, hem trobat reiteradament un resultat recorrent: la manca d'un patró clar i consistent de diferenciació i/o d'estructuració genètica de les diferents poblacions del nord-oest d'Àfrica. L'excepció potser es troba en la població berber mozabita de la ciutat de Ghardaïa, a Algèria. Tot i no presentar diferències estadísticament significatives respecte la resta de poblacions, paràmetres com l'heterozigositat i la variància en la longitud al·lèlica per locus en l'estudi de microsatèl·lits autosòmics ens assenyalen una menor diversitat genètica interna en la població mozabita. L'anàlisi de les distàncies genètiques sembla mostrar la població mozabita força separada de la resta de poblacions del nord-oest d'Àfrica. Ho detectem tant en l'arbre *neighbor-joining*, en la detecció de fronteres genètiques, com en l'anàlisi

de coordenades principals. Els nivells d'heterozigositat poden informar respecte la grandària efectiva de la població. La menor diversitat interna observada a partir de l'anàlisi de microsatèl·lits autosòmics sembla indicar, per tant, que la població mozabita es podria haver diferenciat més extensament per deriva respecte la resta de poblacions nordafricanes. La construcció i anàlisi d'haplotips de microsatèl·lits específics del cromosoma Y sembla confirmar aquesta observació. De fet, en analitzar la diversitat genètica tant per *loci* com per haplotips de microsatèl·lits específics del cromosoma Y, trobem que és clarament menor en la població mozabita que en les poblacions circumdants. Cal assenyalar que el cromosoma Y té una grandària efectiva quatre vegades més petita respecte qualsevol cromosoma autosòmic, fet que implica que la deriva hi actuï més fortament que en els autosomes (Pérez-Lezaun et al. 1997).

Malgrat l'excepció de la població mozabita, els diferents mètodes d'anàlisi aplicats fins a 21 microsatèl·lits, basats tant en distàncies genètiques (arbres, detecció de fronteres i coordenades principals) com en l'anàlisi de la variància molecular (AMOVA) ens mostra que la diferenciació genètica entre les poblacions del nord-oest d'Àfrica és molt petita i sense un patró espacial clar. En concret, en agrupar les diferents poblacions nordafricanes segons característiques culturals en àrabs marroquins per una banda més berber-parlants i sahrauís per l'altra, la partició no es trobà associada a cap diferenciació genètica significativa.

Tal i com hem comentat en la introducció, l'àrab no s'introduí al nord d'Àfrica fins a partir del segle VII. A partir d'aquest moment és quan el poble àrab, procedent de l'Orient Mitjà, realitza les primeres incursions i comença a imposar la seva llengua i religió sobre la població berber. La diferenciació entre àrabs i berbers al nord d'Àfrica no sempre és evident. Es basa essencialment en la llengua, en el sentit que és berber aquell que en conserva la llengua i els costums. En molts casos, la població àrab actual pot haver sorgit d'una població inicialment berber que ha perdut la seva llengua i s'ha arabitzat. Aquest efecte és més comú a les grans ciutats on la influència àrab arribà, sens dubte, més fàcilment que a les regions muntanyoses.

Els nostres resultats semblen indicar que l'expansió de la llengua àrab i la religió islàmica no van comportar una aportació genètica important de població. Aquesta observació queda palesa tant en l'anàlisi de microsatèl·lits autosòmics com en la comparació de freqüències haplotípiques per marcadors bial·lèlics específics del cromosoma Y. Cal assenyalar que, en aquest últim cas, hom disposa d'un important conjunt d'haplotips específics de població que permetrien de detectar aquesta possible

aportació genètica àrab. De fet, sabem que la composició d'haplotips específics del cromosoma Y de les poblacions actuals de l'Orient Mitjà, putatives representants del substrat del qual van sorgir els àrabs invasors, és completament diferent a la dels berbers. Tan sols si observem en detall les freqüències haplotípiques del cromosoma Y en la població àrab del Marroc detectem un lleuger increment del que considerem com a component neolític acompanyat d'una lleugera disminució del substrat autòcton paleolític. Aquesta petita diferència (estadísticament no significativa) podria haver estat conseqüència de l'efecte de l'arabització. Podríem pensar que aquest procés de canvi cultural hauria enriquit el *pool* local de cromosomes Y nordafricans amb cromosomes de l'Orient Mitjà on el component neolític s'espera que sigui clarament més important. Tanmateix, no trobem diferències significatives, fet que ens confirma de nou que el nombre de cromosomes Y portats amb l'arabització deu haver estat petit, fins i tot en la població actual de parla àrab; i que, per tant, l'arabització molt probablement fou només un reemplaçament cultural per una classe dominant.

L'abast de l'aportació genètica sud-sahariana

Tot i que l'anàlisi de microsatèl·lits autosòmics no ens permet de quantificar amb precisió la possible aportació genètica sud-sahariana al nord-oest d'Àfrica, trobem que les distàncies genètiques calculades a partir d'aquests marcadors entre nordafricans i afroamericans (única mostra d'ascendència sud-sahariana de la que disposàvem) són més petites que les que hi ha entre europeus i afroamericans. Per altra banda, la tercera component principal en l'anàlisi de marcadors *clàssics* sembla mostrar un patró irregular de diferenciació nord-sud, que podria ser conseqüència d'un flux gènic amb poblacions sud-saharianes més intens a la meitat meridional del Magreb. La informació aportada per aquests marcadors és poc precisa per la seva mateixa naturalesa: els diferents al·lels es troben en totes les poblacions sense poder-ne d'estriar l'origen ni l'evolució, i és la seva freqüència la que ens pot indicar cer grau de barreja entre poblacions. Per altra banda, s'assumeix que les poblacions actuals són un reflex de les del passat cosa que pot induir a importants errors en les estimacions d'*admixture*.

És a partir de la informació aportada pels marcadors bial·lèlics específics del cromosoma Y que podem detectar i quantificar certa contribució genètica sud-

sahariana al nord-oest d'Àfrica de manera inequívoca. Com veurem a continuació, tot el possible flux gènic sud-saharià en el llinatge masculí del nord-oest d'Àfrica sembla estar representat pels haplotips H22 i H28, els quals comprenen el 8% de cromosomes d'aquesta regió.

La presència de cromosomes Y amb l'al·lel derivat per a la mutació sY81 (també anomenada M2 en la notació d'Underhill), el qual es troba en freqüències especialment elevades a Àfrica Central (57%) i Sud Àfrica (51%) permet detectar part d'aquest component sud-saharià. Cal assenyalar que l'al·lel derivat per sY81 caracteritza l'haplotip H22 (capítol V), el qual conté tots els cromosomes nordafricans pertanyents a l'haplogroup 8 (capítol IV). Aquest darrer haplogrup havia estat descrit, en estudis anteriors a Underhill et al. (1999), a elevades freqüències en poblacions sud-saharianes, i trobat també a molt baixes freqüències a Egipte i a l'oest d'Àsia (Seielstad et al. 1994; Hammer et al. 1997). Underhill et al. (1999) descriu l'haplotip H22 també entre la població khoisànida (18%), a Mali (16%) i a Etiòpia (9%), i a un 7% a l'Orient Mitjà. En conjunt, aquest patró de distribució sembla indicar que H22 té un origen al sud d'Àfrica i que la seva presència al nord-oest d'Àfrica s'explica per contactes trans-saharians. Aquests contactes també haurien portat l'haplotip H28 (caracteritzat per la mutació M33) al nord-oest d'Àfrica, on es troba a baixa freqüència (1.7% del total). Aquest darrer haplotip sembla tenir, però, una distribució (i probablement un origen) restringida al Sahel. Tot i que s'ha descrit a baixes freqüències a l'Orient mitjà, majoritàriament el trobem localitzat a Mali on mostra la freqüència més elevada (un 29.5%). Tanmateix, cal assenyalar que en estudiar en detall la distribució geogràfica d'ambdós haplotips considerats com a sud-saharians en les diferents poblacions del nord-oest d'Àfrica no trobem cap patró regular d'influència sud-nord en el llinatge masculí. És més, tot i que les diferències no són estadísticament significatives, sembla que detectem major influència sud-sahariana en la població berber del nord i centre del Marroc (12.7%) que no pas en la població berber del sud del Marroc (2.5%) o entre els sahrauis (6.9%).

Rando i col·laboradors (1998), en estudiar el llinatge matern, detecten també contribució sud-sahariana al nord-oest d'Àfrica. En aquest cas, però, la influència sud-sahariana que ens mostra la distribució d'haplogrups definits a partir d'RFLPs en la molècula del DNA mitocondrial sembla ser més gran que en el cromosoma Y. En concret, en analitzar diferents poblacions del nord-oest d'Àfrica detecten un 21.5% pel conjunt d'haplogrups L1, L2 i L3A, considerats d'origen sud-saharià per l'elevada

freqüència en què es troben en poblacions sud-saharianes (82% tuaregs, 94% wolof, 95% serer, 99% mandenka 96% senegalesos). Cal assenyalar que, en aquest cas, el possible component sud-sahari à en el llinatge matern mostra cert gradient nord-sud si hom considera la posició geogràfica de les diferents poblacions analitzades: berbers marroquins (4%), berbers d'Algèria (17%), àrabs marroquins (21%), i sahrauis (44%). És més, Brakez i col·laboradors (en preparació), en un estudi preliminar portat a terme al nostre laboratori, detecten un 20% d'haplogrups mitocondrials sud-saharians en una mostra de berbers tahelhits (berbers del sud del Marroc). L'excepció d'aquest gradient nord-sud d'influència sud-sahariana potser es troba en la població d'àrabs marroquins, que semblen mostrar major influència sud-sahariana del que hom esperaria donada la seva posició geogràfica. En aquest cas, potser podríem pensar que les ciutats, de component majoritàriament àrab respecte les regions més muntanyoses, atreuen més migrants sud-saharians.

Cal considerar que la influència sud-sahariana que detectem actualment en les poblacions del nord-oest d'Àfrica ha estat el resultat de la suma de tots els contactes ocorreguts al llarg de la història. Com que ambdues regions geogràfiques, nord-oest d'Àfrica i regió sud-sahariana, han estat en contacte més o menys permanent, podria tractar-se d'un flux gènic acumulat al llarg de moltes generacions. Cal assenyalar que el contacte del nord-oest d'Àfrica amb els pobles sud-saharians fou especialment important durant l'expansió cap al sud del l'imperi berber dels *almoràvids* (1056-1147). Tot i que en menor intensitat, de ben segur, les rutes comercials trans-saharianes poden haver ajudat a mantenir-lo.

El fet que només es trobi gradient nord-sud d'influència sud-sahariana en el llinatge femení podria indicar certa migració preferencial de les dones de la regió del sud i a les ciutats (on els seus descendents serien àrabs) respecte els homes sud-saharians, i/o una més fàcil acceptació d'una dona sud-sahariana com a cònjuge que no pas d'un home. Seielstad et al. (1998), en comparar la diversitat genètica per polimorfismes de substitució nucleotídica (SNPs) entre el DNA mitocondrial, autosomes i el cromosoma Y, van trobar que les diferents variants en el cromosoma Y tendeixen a trobar-se més localitzades geogràficament que les del DNA mitocondrial i autosomes. Els autors suggereixen que una major migració femenina via matrimoni patrilocal (on la dona emigra al poble del seu marit en casar-se) podria ser la causa més probable d'aquests resultats. Aquest autors també consideren els possibles efectes de la poligínia, que redueix la grandària efectiva dels cromosomes Y respecte

el mt DNA, però troben que els efectes que se li podrien atribuir són molt menors que les diferències observades entre variabilitat en cromosoma Y i en DNA mitocondrial. Pérez-Lezaun i col·laboradors (1999) estudien quatre poblacions de l'Àsia Central i observen també una gran diferència entre els patrons de diferenciació aportats pel cromosoma Y i el DNA mitocondrial. Mentre que una part força significativa de la variabilitat en el cromosoma Y estudiada a partir d'haplotips de microsatèl·lits era deguda a diferències entre poblacions, els resultats a partir del DNA mitocondrial en les mateixes poblacions no palesaven aquesta diferència. En aquest cas, els autors proposen també que gran part dels resultats observats podrien ser conseqüència de fluxos migratoris diferencials entre homes i dones.

Per tant, la diferent aportació de llinatges femenins i masculins sud-saharians al nord-oest d'Àfrica podria ser el reflex d'un fenomen d'abast molt general: la migració diferencial de dones i homes.

Detecció del flux gènic a l'estret de Gibraltar

En diverses de les anàlisis realitzades trobem una clara diferenciació entre els *pools* genètics de la Península Ibèrica i el nord-oest d'Àfrica. En l'estudi de fins a 21 microsatèl·lits autosòmics, les poblacions analitzades d'ambdues regions apareixen clarament separades, ja sigui en l'anàlisi de coordenades principals, en la detecció de fronteres genètiques o en els arbres genètics on es troben unides per una branca llarga i molt robusta, tal com indiquen els valors de *bootstrap* obtinguts.

Aquesta diferenciació és també especialment patent en la composició d'haplotips de marcadors bial·lèlics del cromosoma Y que trobem en ambdues regions. Mentre el 83% de cromosomes Y al nord-oest d'Àfrica pertanyen al grup III, el 78% de cromosomes a la Península Ibèrica pertànyen al grup IX (capítol V). Per altra banda, la presència d'haplotips considerats específics del nord-oest d'Àfrica (especialment H38, però també H35 i H36) a la Península Ibèrica permet d'estimar la contribució nordafricana mínima al *pool* genètic del cromosoma Y ibèric en un 7%. En aquest cas, s'ha tingut en compte la freqüència de cada haplotip considerat nordafricà a la Península Ibèrica i s'ha corregit per la freqüència en que cada un d'aquests haplotips es troba actualment al nord-oest d'Àfrica. En concret, aquesta contribució l'estimem en un 14% en la mostra d'andalusos, i en un 3% en la de bascos, fet que potser podria

ser el reflex d'una variació clinal en l'aportació genètica nordafricana en el llinatge masculí ibèric.

Com s'ha comentat anteriorment, tant Kandil i col·laboradors (1999) com Simoni i col·laboradors (1999), a partir de l'anàlisi de polimorfismes *clàssics* i emprant mètodes diferents, troben que la diferenciació genètica més important que hi ha entre les poblacions mediterrànies és la que travessa l'estret de Gibraltar. Aquests resultats confirmen la diferenciació trobada entre poblacions d'ambdues ribes en la nostra anàlisi inicial, també realitzada a partir de marcadors *clàssics*.

Altres sistemes genètics tampoc semblen mostrar un flux gènic apreciable del nord d'Àfrica cap a la Península Ibèrica. Rando i col·laboradors (1998) descriuen l'haplogrup U6 del DNA mitocondrial i el troben a freqüències de l'ordre del 10-20% en nordafricans, mentre que es troba absent o a molt baixes freqüències en europeus i en altres africans. Per aquest motiu, Rando i col·laboradors (1998) proposen que l'haplogrup U6 va sorgir al nord d'Àfrica. Altres estudis sobre DNA mitocondrial mostren que les seqüències U6 es troben a molt baixa freqüència a la Península Ibèrica: 3 de 54 portuguesos, 2 de 96 gallecs, absent en andalusos i en 162 altres ibèrics (Bertranpetit et al. 1995, Côrte-Real et al. 1996, Pinto et al. 1996 i Salas et al. 1998). La migració femenina del nord d'Àfrica a la Península Ibèrica sembla ser també petita. Per tant, quan combinem la informació aportada pels microsatèl·lits autosòmics, pel cromosoma Y i pel DNA mitocondrial podem concloure que el flux gènic nordafricà a la Península Ibèrica ha estat petit, tot i que es detecta clarament.

Cal assenyalar que Kandil i col·laboradors (1999), a partir de marcadors de *clàssics*, troben que la major aportació genètica nordafricana en la Península Ibèrica, correspon a una mostra de las Alpujarras (Granada). Aquest resultat sembla coincidir amb la posició intermèdia en la que trobem la població andalusa en un arbre *neighbor-joining* construït a partir de 21 microsatèl·lits autosòmics, tot i que aquesta posició no es retroba en l'anàlisi de coordenades principals i que el coeficient de mescla (*admixture*) nordafricana en andalusos no és significativament diferent de zero. Aquestes dificultats en escatir l'aportació nordafricana a les poblacions del sud de la Península Ibèrica pot ser un reflex de la migradesa d'aquesta aportació, tal com apareix en els llinatges del cromosoma Y i en les seqüències de DNA mitocondrial. Cal assenyalar que els marcadors *clàssics* no són especialment apropiats per fer aquestes aproximacions ja que donen uns resultats amb uns errors enormes de l'estimació de les aportacions de les diferents poblacions, fins al punt que gairebé qualsevol

interpretació és possible. Els marcadors emprats del cromosoma Y aporten, en canvi, unes possibilitats molt més acurades degut a la gran variació observada, a la seva interpretació filogenètica i a la distribució restringida dels haplotips.

De la mateixa manera que estudiem quina ha estat l'aportació nordafricana a la Península Ibèrica, podem preguntar-nos si detectem aportació genètica ibèrica al nord-oest d'Àfrica. Hom disposa de tres haplotips específics de la Península Ibèrica: H100, H101 i H102, caracteritzats respectivament per les mutacions M65, M153 i M167 (equivalent a SRY-2627 en la notació del capítol IV) i els quals representen el 21.6% dels cromosomes Y ibèrics. Tanmateix, no hem identificat cap d'aquests haplotips a les poblacions nordafricanes. H103, tot i no ser específic de les poblacions de la Península Ibèrica (el seu patró de distribució inclou la resta d'Europa i a freqüències més baixes també Àsia Central, Orient Mitjà, el Pakistan i l'Índia), sembla ser l'únic haplotip que ens permet detectar un possible flux gènic des de la Península (on es troba a una freqüència del 56%) al nord-oest d'Àfrica (on compren el 2.8% de cromosomes Y estudiats). La comparació dels haplotips de microsatèl·lits en cromosomes H103 d'ambdues regions mostrà que els haplotips ibèrics i nordafricans són molt semblants entre ells, fet que és compatible amb una introducció de cromosomes H103 des de la Península Ibèrica. Considerant les freqüències d'aquest haplotip, estimem que la contribució màxima d'aquesta aportació ibèrica al nord-oest d'Àfrica seria d'un 5%. Entre els moviments de població des de la Península Ibèrica al nord-oest d'Àfrica trobem possibles contactes fa 16.000 anys, en el moment en què les cultures paleolítiques íbero-marusiana al nord d'Àfrica i magdaleniana a la Península Ibèrica mostren una gran similitud. Durant l'antiguitat clàssica, pobles com els fenicis, els cartaginesos i els romans van ocupar ambdues ribes de la Mediterrània i podien haver vehiculat el flux gènic del nord cap al sud. Al segle V aC, foren els vàndals qui travessaren la Península Ibèrica en direcció al nord d'Àfrica, i podien també haver aportat cromosomes Y europeus. Tanmateix, el moviment humà millor documentat de la Península Ibèrica cap al Magreb fou l'expulsió dels moriscos el segle XVII. Tot i així, desconeixem el substrat poblacional d'aquesta població i ignorem si es tractava de descendents dels primers invasors musulmans de la Península Ibèrica o si incloïen descendents d'hispano-romans convertits a la fe islàmica. En aquest darrer cas, l'expulsió dels moriscos podia haver portat cromosoma Y de la Península Ibèrica al nord d'Àfrica.

Arnáiz-Villena i col·laboradors (1995; 1997) estudien la variació en el sistema HLA i troben que les freqüències al·léliques i haplotípiques d'alguns *loci* d'aquest sistema mostren certa similitud entre el nord d'Àfrica (marroquins, algerians), la Península Ibèrica (bascos inclosos) i Sardenya. Aquests resultats els portaren a suggerir un origen comú anterior al neolític per a les poblacions de la Península Ibèrica, especialment la basca, i el nord-oest d'Àfrica. Tanmateix, Comas et al. (1998), en considerar tota la variació en HLA i no només alguns haplotips en el càlcul de distàncies genètiques i l'anàlisi de components principals, no detecten cap relació especial entre la població basca i la nordafricana. És més, en un arbre *neigbor-joining* construït a partir de deu poblacions europees i nordafricanes considerant la informació de set *loci* HLA de les classes I i II, Comas et al. (1998) troben que les poblacions basca i sarda són les més diferenciades respecte els algerians. Cal assenyalar que la coincidència d'alguns al·lels en un sistema pot informar de migracions individuals, però no pressuposa *per se* un origen comú. Els nostres resultats, basats en 21 *loci* autosòmics independents i en els llinatges paterns, contradiuen completament la hipòtesi d'un origen comú per a les poblacions ibèriques i nordafricanes.

En resum, la presència àrabo-berber a la Península Ibèrica va aportar un devesall d'innovacions culturals en camps tan diversos com l'arquitectura, l'agricultura i l'astronomia però no sembla haver afaiçonat la genètica de les poblacions peninsulars.

Aplicacions en genètica forense

Com hem vist, els haplotips de marcadors bial·lèlics al cromosoma Y mostren que el contingut de cromosomes Y al nord d'Àfrica és diferent al de la Península Ibèrica, i al d'Europa en general. Per tant, no seria correcte emprar dades que no siguin específiques del nord d'Àfrica en casos criminals i/o d'identificació individual en genètica forense que impliquin membres d'aquestes poblacions. A part de l'òbvia necessitat de la pròpia genètica forense en els països del Magreb, donada la creixent inmigració de la població nordafricana a Europa en els darrers anys, hom preveu una gran necessitat de disposar d'aquest tipus de caracterització genètica en bases de dades de referència. Si es dóna una coincidència en les mostres analitzades (preses del lloc del crim i d'un sospitós, per exemple), la probabilitat que aquesta coincidència

sigui deguda a l'atzar es calcula a partir de les freqüències al·lèliques en la població. Per tant, la probabilitat de coincidència calculada per a un individu nordafricà amb una base de dades europea pot ser del tot errònia.

Com s'ha comentat en la introducció, els microsatèl·lits específics del cromosoma Y són fàcilment tipificables per tècniques senzilles com l'amplificació per PCR a partir de petites quantitats de DNA. Això els confereix un avantatge indiscutible sobre d'altres tipus de marcadors a l'hora de ser emprats en genètica forense. Tanmateix, cal assenyalar una limitació de la informació que ens aporta el cromosoma Y en els casos d'inclusió, donat que, en no presentar recombinació, tots els barons d'una família emparentats per línia masculina compartiran el mateix cromosoma tret que s'hi hagin produït mutacions *de novo*.

En el capítol II s'han tipificat fins a 21 microsatèl·lits autosòmics, entre els quals s'inclouen els 13 marcadors STRs més emprats en genètica forense, establerts com a estàndards en la base de dades nordamericana de genotips de convictes, anomenada CODIS. Les dades de freqüències al·lèliques que presentem per aquests marcadors autosòmics en poblacions del nord-oest d'Àfrica són també una significativa contribució a la genètica forense. Cal assenyalar en aquest cas que, tot i no permetre determinar directament l'aportació masculina en un cas d'agressió sexual, són dades idònies per el seu ús en paternitat i altres casos criminals.

Estructura de la variació genètica en el cromosoma Y

Fins ara hem considerat la variació genètica estructurada per població, però, sens dubte, la podem considerar des d'altres perspectives. Sovint s'entén la diversitat genètica com a estructurada per poblacions, fins a l'extrem que una perspectiva poblacional s'accepta com a única i completa descripció de la diversitat genètica. La teoria de la genètica de poblacions *clàssica* modela la dinàmica del canvi en les freqüències al·lèliques en termes de les forces evolutives que afecten les poblacions com la deriva gènica, la selecció i la migració. Aquest tipus d'aproximació implica que una simple descripció de la variació per població és totalment informativa en el terreny genètic. Tanmateix, la diversitat genètica pot presentar una estructura més profunda causada per la genealogia en la que es troba, més que no pas pel procés d'etnogènesi

que donà origen a la població. La genealogia de qualsevol regió genòmica es pot reconstruir a partir de la genealogia dels polimorfismes d'evolució lenta, els quals permetran d'identificar un rerafons genètic estable on d'altres polimorfismes amb major taxa de canvi evolucionen.

En el treball presentat al capítol IV, mitjançant la comparació de dos sistemes genètics amb diferents taxes d'evolució en la regió no recombinant més gran del nostre genoma, el cromosoma Y, intentem comprendre com s'estructura la variació genètica des d'una perspectiva innovadora: segons el rerafons genètic en comptes de segons l'estructura poblacional.

Tipificàrem dos tipus de sistemes genètics en 129 cromosomes Y de quatre poblacions del nord-oest d'Àfrica: onze polimorfismes bial·lèlics estables i set microsatèl·lits multial·lèlics. El conjunt de polimorfismes bial·lèlics va permetre de definir fins a dotze haplogrups, els quals molt probablement sorgiren un sol cop al llarg de l'evolució humana. La variació en els sistemes de majors taxa evolutiva (STRs o microsatèl·lits) fou examinada per a caracteritzar la variació genètica segons els rerafons genètics anteriorment identificats, i per a qüestionar si la variació en els STRs retenia informació en quan a la filogènia i el temps de coalescència dels haplogrups.

Trobàrem una elevada freqüència de l'haplogrup 21 (76%) i freqüències substancialment menors per cinc altres haplogrups. Aplicàrem l'anàlisi de la variància molecular (AMOVA) a l'estudi de la repartició de la variació dels microsatèl·lits entre els diferents rerafons genètics (en el nostre cas definits per haplogrups i amb una clara estructura filogenètica) i trobàrem que la variabilitat dels microsatèl·lits clarament estava estructurada per haplogrups, els quals expliquen més del 80% de la variància genètica total al nord-oest d'Àfrica. En canvi, en analitzar la variància genètica dels STRs entre poblacions nordafricanes, el percentatge de la variància atribuïble a les diferències entre poblacions fou del 4%. Podem preguntar-nos si aquest resultat és específic de les poblacions analitzades, les quals presenten una composició d'haplogrups molt particular. Per tal de comprovar si aquest resultat és més general, van afegir dues poblacions ibèriques a l'anàlisi. En aquest cas, tot i que l'estructuració de la variació dels microsatèl·lits per haplogrups es mantenia, el percentatge de variabilitat genètica atribuïble als haplogrups era inferior (20%), tot i que continua sent molt superior a la variància explicable per diferències entre poblacions (2%). Molt probablement, aquesta disminució en el percentatge de variació explicat per haplogrup

és causada per la presència de l'haplogrup 22, específic de les poblacions ibèriques, i el qual s'ha originat molt recentment a partir de l'haplogrup 1 (Hurles et al. 1999).

El fet de trobar una elevada variància entre haplogrups respecte una baixa variància entre poblacions molt probablement indica un origen molt més recent de les poblacions respecte la majoria de polimorfismes bial·lèlics. La profunda estructuració de la variació genètica dels microsatèl·lits en unitats genealògiques antigues o haplogrups permet fins i tot qüestionar una perspectiva poblacional en la comprensió de la diversitat del genoma humà. Igualment, ens porta a repensar la població com a constituïda per un conjunt de llinatges independents que poden ser molt anteriors a la pròpia etnogènesi.

Trobem dos exemples en la literatura on es mostra que la diversitat genètica en d'altres regions del nostre genoma presenta una clara estructuració per rerafons genètic. Estivill i col·laboradors (1994) trobaren que els haplotips de tres microsatèl·lits en el gen de la fibrosi quística (CFTR) eren clarament diferents en cromosomes portadors de la mutació ΔF508 (la més freqüent entre les més de 800 que causen la fibrosi quística) respecte aquells trobats en cromosomes no afectats. És més, els haplotips d'STRs dels cromosomes no afectats de qualsevol població europea eren molt més propers entre ells que no pas als haplotips de cromosomes ΔF508 de la mateixa població. Per tant, la diversitat genètica en el gen CFTR a Europa té una major estructura per rerafons genètic que no pas per població. Rocha i col·laboradors (1997), a partir de les freqüències al·lèliques d'un STR localitzat a l'extrem 5' del gen α1-antitripsina, aconsegueixen reconstruir la filogènia entre diferents variants electroforètiques del gen. Sembla, doncs, que aquest microsatèl·lit conté informació relativa al rerafons genètic en el que es troba.

Com s'ha demostrat en els reordenaments cromosòmics de *Drosophila subobscura,* si hi ha un intercanvi restringit d'informació degut a la manca de recombinació, la variabilitat genètica (en aquest cas de seqüència nucleotídica en el gen rp49) tendeix a compartimentar-se en blocs estancs (Rozas et al. 1999). Tal com hem presentat, això arriba al seu extrem en el cas dels humans en les 30 Mb de la regió no recombinant del cromosoma Y.

A part d'analitzar l'estructuració de la variacio dels microsatèl·lits per rerafons genètic, vàrem utilitzar la variància del nombre de repeticions en els STRs per a datar el temps a l'avantpassat comú més recent (TMRCA) en cada branca de la filogènia del cromosoma Y. Les edats estimades comprenien un rang entre 3.000 i 14.000 anys,

totes amb amplis intervals de confiança. Les edats estimades es corresponien amb la posició de les mutacions datades en l'arbre filogenètic, de manera que obtinguerem edats més antigues per a mutacions que es troben en les regions més profundes i ancestrals. Per tant, aquests resultats mostren clarament que la variació en marcadors genètics amb elevada taxa d'evolució encara reté informació respecte la seva filogènia i no han assolit una saturació completa.

Com hem explicat anteriorment, Hammer i col·laboradors (1998) analitzen polimorfismes bial·lèlics que parcialment se solapen amb el conjunt d'onze marcadors bial·lèlics discutits en aquesta secció. A partir de les freqüències haplotípiques trobades en poblacions representatives de la variabilitat genètica mundial i de la pròpia estructura de la filogènia dels marcadors del cromosoma Y, Hammer i col·laboradors (1998) apliquen els mètodes de coalescència de Griffiths i Tavaré (1994) per a obtenir estimes de les edats d'aquests polimorfismes. En els casos en què podem comparar els resultats de Hammer i col·laboradors (1998) amb els nostres, observem que les nostres estimes d'edat són sempre força més recents que les d'aquests altres autors, tot i que els intervals de confiança se solapen.

Scozzari i col·laboradors (1999) utilitzen els mateixos marcadors bial·lèlics que Hammer i col·laboradors (1998) en un estudi més detallat de les poblacions africanes. A més, analitzen quatre STRs dinucleòtids (diferents dels microsatèl·lits analitzats per nosaltres) en els mateixos cromosomes. Scozzari i col·laboradors (1999) també estimen dates per els polimorfismes bial·lèlics, però amb un mètode diferent: el proposat per Goldstein i col·laboradors (1996), basat en el model de mutació *stepwise* dels microsatèl·lits i que requereix conèixer tant la grandària efectiva de la població com la taxa de mutació. Els valors que Scozzari i col·laboradors (1999) fan servir per aquests paràmetres són, respectivament, N_e = 4.500 cromosomes i $m = 5.6 \times 10^{-4}$ mutacions/ gàmeta/generació. Cal assenyalar que aquest darrer valor és una estima mitjana obtinguda a partir de quinze microsatèl·lits del cromosoma 19 (Weber i Wong, 1993), mentre que l'estima emprada en el nostre treball és aproximadament el doble i correspon directament a la taxa de mutació dels STRs del cromosoma Y que hem emprat. Malgrat el diferent mètode utilitzat, Scozzari i col·laboradors (1999) troben edats per els polimorfismes bial·lèlics del mateix ordre de magnitud que les estimades per Hammer i col·laboradors (1998).

Hem suggerit diferents explicacions possibles per a la discrepància entre les edats estimades per nosaltres i els valors obtinguts per Hammer i col·laboradors (1998) i per Scozzari i col·laboradors (1999):

i) El mètode que hem emprat intenta datar el TMRCA de la variació en els microsatèl·lits, que és, per definició, posterior a l'edat de la mutació.

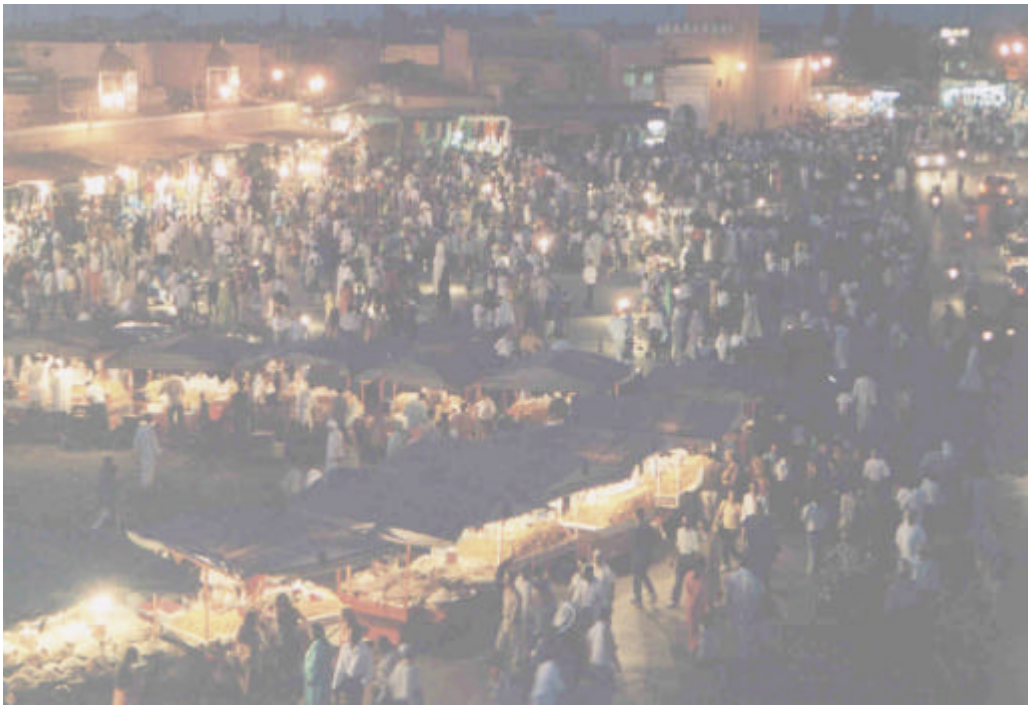
ii) Biaix poblacional. Com que hem analitzat cromosomes d'un grup molt definit de poblacions, podriem haver subestimat la quantitat de variació generada dins els haplogrups. De totes maneres, l'aplicació del mètode de Griffiths i Tavaré (1994) a les nostres dades de freqüències d'haplogrups aconsegueix reproduir força fidelment les estimes d'edats proposades per Hammer et al. (1998)

iii) Possible saturació dels STRs. Si aquests *loci* haguessin assolit l'equilibri mutació-deriva, un punt en el qual la variació s'ha saturat i tota aquella variació adquirida per mutació es perd per deriva, no resultarien informatius per a estimes d'edats. Tanmateix, hem mostrat que aquests *loci* contenen encara informació filogenètica, i que les edats estimades, per bé que en termes absoluts puguin resultar massa recents, el seu ordre relatiu és correcte.

iv) Sobreestimació de la taxa de mutació. Les edats estimades són una funció lineal de la taxa de mutació. Si aquesta hagués estat sobreestimada en 5-6 vegades, les nostres estimes coincidirien amb les d' altres autors. En el cas del DNA mitocondrial, s'ha observat una àmplia discrepància entre les estimes genealògiques i filogenètiques de la taxa de mutació (Jazin et al. 1998).

Hem mostrat que el *background* genètic predomina sobre el *background* poblacional en l'estructura de la variació genètica dels STRs en la regió no recombinant del cromosoma Y humà. Hem emprat la variació dels STRs dins els diferents llinatges per inferir-ne els seus TMRCAs i hem mostrat que la diferenciació genètica que roman entre haplogroups conté encara informació filogenètica. Per altra banda, trobem que la variació dels STRs en una població ve determinada per la composició d'haplogroups en aquella població específica. Així, doncs, la composició genètica dels humans molt probablement es pot entendre millor en termes d'unitats genealògiques relacionades evolutivament més que no pas en termes demogràfics.

Hem estudiat la variació genètica de les poblacions del nord d'Àfrica a partir de l'anàlisi de diverses regions genòmiques per tal de reconstruir les diferents capes sobreposades que constitueixen la seva història. Talment com arqueòlegs del nostre genoma, hem pogut excavar en el jaciment de la història nordafricana i n'hem identificat, estrat a estrat, els seus elements constituents. Hem vist però que els llinatges nordafricans són un feix de vímets que conflueixen en un passat molt més remot que la història de les poblacions. I potser sense proposar-nos-ho, hem contribuït a crear eines per a la justícia. Arribem ara al final del nostre viatge que ens ha dut més enllà d'exotismes, a un íntim coneixement de la història d'unes poblacions veïnes cap a les quals hem de construir ponts de mútua comprensió.



BIBLIOGRAFIA

BIBLIOGRAFIA

Amos W, Rubinsztein D (1996) Microsatellites show mutational bias and heterozygote instability. Nature Genetics 13:390-391

Arnáiz-Villena A, Benmamar D, Alvarez M, Díaz-Campos N, Varela P, Gómez-Casado E, Martínez-Laso J (1995) HLA allele and haplotype frequencies in Algerians. Relatedness to Spaniards and Basques. Hum Immunol 43:259-268

Arnáiz-Villena A, Martínez-Laso J, Gómez-Casado E, Díaz-Campos N, Santos P, Martinho A, Breda-Coimbra H (1997) Relatedness among Basques, Portuguese, Spaniards, and Algerians studied by HLA allelic frequencies and haplotypes. Immunogenetics 47:37-43

Ashley CT, Warren ST (1995) Trinucleotide repeat expansion and human disease. Ann Rev Gene 29:703-728

Barbujani G, Pilastro A, De Dominico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonization. Am J Phys Anthrop 95:137-154

Bertranpetit J, Sala J, Calafell F, Underhill P, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of the Basques. Ann Hum Genet 59:63-81

Bianchi NO, Catanesi CI, Bailliet G, Martínez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera R, et al. (1998) Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet 63:1862-1871

Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 5:393-401

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

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455-457

Brassel KE, Reif D (1979) A procedure to generate Thiessen polygons. Geogr Anal 11:289-303

Brega A, Torroni A, Semino O, Maccionni L, Casanova M, Scozzari R, Fellous M, et al. (1987) The p12f2/TaqI Y-specific polymorphism in three groups of Italians and in a sample of Senegalese. Gene Geog 1:201-206

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the sctructure and length of the tandem repeat. Am J Hum Genet 62:1408-1415

Calafell F, Bertranpetit J (1994) Principal component analysis of gene frequencies and the origin of Basques. Am J Phys Anthrop 93:201-215

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism in humans. Eur J Hum Genet 6:38-49

Camps G (1974) Les civilisations préhistoriques de l'Afrique du Nord et du Sahara. Doin, Paris

Camps G (1994) Els berbers, mite o realitat?. In: Maria-Àngels Roque (ed) Les cultures del Magreb. Enciclopèdia Catalana, Barcelona pp 75-96

Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, Purrello M, et al. (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science 230:1403-1406

Cavalli-Sforza LL, Cavalli-Sforza F (1994) Qui som. Història de la diversitat humana. Enciclopèdia Catalana, Barcelona

Cavalli-Sforza LL, Minch E (1997) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 61:247-251

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. Science 259:639-646

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

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 USA 94:1041-1046

Comas D, Mateu E, Calafell F, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Bertranpetit J (1998) HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51:30-40

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Molec Genet 5:1759-1766

Côrte-Real HBSM, Macaulay V, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha SS, et al. (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331-350

Deka R, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, Ferrell RE, et al. (1995) Population genetics of dinucleotide (dC-dA)n (dG-dT)n polymorphisms in World populations. Am J Hum Genet 56:461-474

Deka R, Jin L, Shriver MD, Yu LM, Saha N, Barrantes R, Chakraborty R, et al. (1996) Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Gen Res 6:1177-1184

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110:134-140

Desanges J (1990) The proto-Berbers. In: Mokhtar G (ed) General History of Africa. Unesco, Paris, France, pp 236-245

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al. (1996) A comprehensive genetic map of the human genome based on 5,624 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 USA 91:3166-3170

Dorit R, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183-1185

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet 49:746-756

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241-253

Efron B (1982) The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA

Estivill X, Morral N, Bertranpetit J (1994) Reply to Kaplan et al. Nature Genetics 8:216-218

Estoup A, Tailliez C, Cornuet JM, Solignac M (1995) Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). Mol Biol Evol 12:1074-1084

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

Foster EA, Jobling MA, Taylor PG, Donnely P, de Knijff P, Mieremet R, Zerjal T, et al. (1998) Jefferson fathered slave's last child. Nature 396:27-28

Freimer NB, Slatkin M (1996) Microsatellites: evolution and mutational processes. In: Chadwick D and Cardew G (eds) Variation in the human genome. Ciba Foundation Symposium, Wiley, Chichester, pp 51-72

Gill P, Kimpton CP, Urquhart A, Oldroyd NJ, Millican ES, Watson SK, Downes TJ (1995) Automated short tandem repeat (STR) analysis in forensic casework--a strategy for the future. Electrophoresis 16:1543-1552

Goldstein DB, Clark AG (1995) Microsatellite variation in North American populations of Drosophila melanogaster. Nuc Acids Res 23:3882-3886

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 USA 92:6723-6727

Goldstein DB, Zhivotovsky LA, Nayar K, Ruiz-Linares A, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13:1213-1218

Gonçalves J, Lavinha J (1994) The Y-associated XY275G (Low) allele is common among the Portuguese. Am J Hum Genet 55:583-585

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307-319

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376-378

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951-962

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, et al. (1997) The geographic distribution of human Y chromosome variation. Genetics 145:785-805

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, et al. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427-441

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

Henke J, Henke L (1999) Mutation rate in human microsatellites. Am J Hum Genet 64:1473-1473

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799-803

Hitti PK (1990) The Arabs: A Short History. Gateway Editions, Washington, DC

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, et al. (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63:1793-1806

Hurles ME, Veita R, Arroyo E, Armenteros M, Bertranpetit J, Pérez-Lezaun A, Bosch E, et al. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphism. Am J Hum Genet 65: 1793-1806

Jazin E, Soodyall H, Jalonen P, Lindholm P, Stoneking M, Gyllensten U (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. Nature Genetics 18:109-110

Jobling MA (1994) A survey of long range DNA polymorphisms on the human Y chromosome. Hum Mol Genet 3:107-114

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and the human evolution. Trends Genet 11:449-456

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al. (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523-538

Kandil M, Moral P, Esteban E, Autori L, Mameli GE, Zaoui D, Calò C, et al. (1999) Red cell enzyme polymorphisms in Moroccans and Southern Spaniards: new data for the genetic history of the Western Mediterranean. Hum Biol 71:791-802

Karafet T, Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, Romero F, et al. (1997) Y chromosome markers and trans-Bering strait dispersals. Am J Phys Anthrop 102:301-314

Kasule S (1998) The history Atlas of Africa. Macmillan, New York

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125-133

Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M (1993) Automated DNA profiling employing multiplex amplification of short tandem repeat loci. PCR Meth Appl 3:13-22

Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. Science 278:675-680

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203-221

McEvedy (1980) The Penguin Atlas of African History. Penguin Books, New York

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. Curr Opin Genet Dev 6:737-742

Mitchell RJ, Earl L, Williams J (1993) Two Y-chromosome-specific restriction fragment length polymorphisms (DYS11 and DYZ8) in Italian and Greek migrants to Australia. Hum Biol 65:387-399

Mitchell RJ, Earl L, Fricke B (1997) Y-chromosome specific alleles and haplotypes in European and Asians populations: linkage disequilibrium and geographic diversity. Am J Phys Anthrop 104:167-176

Mourant AE, Kopec AC, Domaniewska-Sobczak K (1976) The distribution of the human blood groups and other polymorphisms. Oxford University Press, London (UK)

Moxon ER, Wills C (1999) Microsatélites de ADN. Investigación y Ciencia 270: 68-74

Newman J (1995) The peopling of Africa: A geographic interpretation. University Press, New Haven, CT

Page RDM, Holmes EC (1998) Molecular Evolution. A Phylogenetic Approach. Blackwell Science Ltd, Oxford, England, p. 184

Pena SD, Santos FR, Bianchi NO, Bravi CM, Carnese FR, Rothhammer F, Gerelsaikhan T, et al. (1995) A major founder Y-chromosome haplotype in Amerindians. Nature Genetics 11:15-16

Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997a) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1-7

Pérez-Lezaun A, Calafell F, Seielstad MT, Mateu E, Comas D, Bosch E, Bertranpetit J (1997b) Population genetics of Y chromosome short tandem repeats in humans. J Mol Evol 45:265-270

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martínez-Arias R, Clarimon J, et al. (1999) Sex-specific migration patterns in Central Asian populations, revealed by the analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208-219

Perucho M (1998) Cáncer del fenotipo mutador de microsatélites. Investigación y Ciencia 261:46-55

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 mitochondrialDNA sequences. Ann Hum Genet 60:321-330

Rampino N, Yamamoto H, Ionov Y, Li Y, Saway H, Reed J, Perucho M (1997) Somatic frameshift mutations in the Bax gene in colon cancers of the microsatellite mutator phenotype. Science 275:967-969

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 Europeans, Near-Eastern, and Sub-Saharan populations. Ann Hum Genet 62:531-550

Renfrew C (1987) Archaeology and Language. The puzzle of Indoeuropean origins. Jonathan Cape, London (UK)

Renfrew C (1991) Before Babel: speculations on the origins of linguistic diversity. Camb Archaeol J 1:3-23

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics 105:767-779

Richards RI, Sutherland GR (1994) Simple repeat DNA is not replicated simply. Nature Genetics 6:114-115

Rocha J, Pinto D, Santos MT, Amorim A, Amil-Dias J, Cardoso-Rodrigues F, Aguiar A (1997) Analysis of the allelic diversity of a (CA)n repeat polymorphism among α1-antitrypsin gene products from northern Portugal. Hum Genet 99:194-198

Roychoudhury AK, Nei M (1988) Human polymorphic genes world distribution. Oxford University Press, Oxford

Rozas J, Segarra C, Ribó G, Aguadé M (1999) Molecular population genetics of the rp49 gene region in different chromosomal inversions of *Drosophila subobscura*. Genetics 151:189-202

Said R, Faure H (1990) Chronological framework: African pluvial and glacial epochs. In: Ki-Zerbo J (ed) General History of Africa. Unesco, Paris, France, pp 146-166

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425

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

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning. A laboratory manual. Cold Spring Harbour Laboratory Press, New York

Santos FR, Hutz M, Coimbra CEA, Santos RV, Salzano FM, Pena SDJ (1995) Further evidence for the existence of a major founder Y chromosome haplotype in Amerindians. Braz J Genet 18:669-672

Santos FR, Rodríguez-Delfín L, Pena SDJ, Moore J, Weiss KM (1996) North and South Amerindians may have the same major founder Y chromosome haplotype. Am J Hum Genet 58:1369-1370

Santos FR, Pandya A, Tyler-Smith C, Pena SDJ, Schanfield M, Leonard WR, Osipova L, et al. (1999) The Central Siberian origin for native Americans Y chromosomes. Am J Hum Genet 64:619-628

Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. Nuc Acids Res 20:211-215

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellito D, Arredi B, et al. (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, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D, Cavalli-Sforza LL (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet 3:2159-2161

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nature Genetics 20:278-280

Semino O, Passarino G, Liu A, Brega A, Fellous M, Santachiara-Benerecetti AS (1995) Three Y-specific polymorphisms in populations of different ethnic and geographic origin. Y Chromosome Newsletter 2:5-6

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti S (1996) A view of the Neolithic diffusion in Europe through two Y chromosome-specific markers. Am J Hum Genet 59:964-968

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

Smeets HJM, Ropers HH, Wieringa B (1989) Use of variable simple sequence motifs as genetic markers: Application study of myotonic dystrophy. Hum Genet 83:245-251

Steinberg AG, Cook CE (1981) The distribution of the human immunoglobulin allotypes. Oxford University Press, Oxford, England

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origen of Old Testament priests. Nature 394:138-139

Tills D, Kopec AC, Tills RE (1983) The distribution of the human blood groups and other polymorphisms. Supplement 1. Oxford University Press, London (UK)

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196-200

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, et al. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Gen Res 7:996-1005

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, et al. (1999) The architecture of the Y-chromosome biallelic haplotype diversity: an emerging portrait of mankind. Sota consideració editorial.

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

Weber JL (1990) Human DNA polymorphisms based on length variation in simple-sequence tandem repeats. In: Genome analysis. Cold Spring Harbor Laboratory Press, New York, pp 159-177

Weber J, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123-1128

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379-380

Wright S (1951) The genetical structure of populations. Annals of Eugenetics 15:323-354

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al. (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174-1183

FE D'ERRATES

• Capítol III: pag. 160

Since the Y chromosome has one quarter of the effective population size of any autosome, genetic drift acts more strongly on the Y chromosome than on the autosomes. As demonstrated by Pérez-Lezaun *et al.* (1997) this means that:

1. Gene diversities in the Y chromosome STRs are lower than those found in autosomal STRs. Recently, White *et al.* (1999) selected, out of 185 possible Y chromosome $(GATA)_n$ STRs, six loci with an average D = 0.663 in an unidentified sample, which is still far from the usual levels of polymorphism in the autosomal STRs used in forensic casework (>0.75, Pérez-Lezaun *et al.* 1999a).
2. More differentiation is found among populations for Y chromosome STRs than for autosomal STRs, and thus more subpopulations need to be typed to obtain reliable population data for Y chromosome markers.

These basic genetic properties explain the high levels of between population differentiation observed in NW Africa for the Y chromosome markers and the reduction of the internal diversity in one of the populations, the Mozabites.

White PS, Tatum OL, Deaven LL, Longmire JL (1999) New, male-specific microsatellite markers from the human Y chromosome. Genomics 57: 433-437.

• Capítol IV: pag. 1630, Table 4.

HG S8 → **HG 8**

• Capítol IV: pag. 1637.

Pandya A., King TE, Santos FR, Taylor PG, Thangaraj K, Singh L, Jobling MA, Tyler-Smith C. **A polymorphic human Y-chromosomal G to A transition found in India. Ind J Hum Genet, in press.**

• BIBLIOGRAFIA: pag. 258.

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Pérez-Lezaun A, Bosch E, et al. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphisms. Am J Hum Genet. 65: **1437-1448**.

• Capítol V i DISCUSSIÓ.

Haplotypes 87-115 → **88-116**.