UNIVERSITAT POMPEU FABRA

Tesi Doctoral

Història natural de les malalties genètiques mendelianes i complexes en poblacions humanes

Oscar Lao Grueso

Barcelona, Septembre del 2004

Departament de Ciències Experimentals i de la Salut
Universitat Pompeu Fabra (UPF)

Història natural de les malalties genètiques mendelianes i complexes en poblacions humanes

Memòria presentada per Oscar Lao Grueso per optar al grau de doctor en Ciències Biològiques. Aquesta tesi ha estat realitzada sota la direcció del Dr. Francesc Calafell i Majó, a la Unitat de Biologia Evolutiva del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra, dins del programa de doctorat en *Ciències de la Salut i de la Vida* (*bienni 2000-2002*).

Francesc Calafell i Majó **Oscar Lao Grueso**

Barcelona, Septembre del 2004

A mis padres, a mi abuela, a mi hermano.

A Lourdes.

HERM. – En verdad, Sócrates, has avanzado mucho.

SÓC. – Paréceme que estoy ya progresando en sabiduría.

Crátilo. Platón

AGRAÏMENTS

Ha estat molta la gent que m'ha acompanyat i ajudat al llarg d'aquests quatre anys de tesi doctoral. Més enllà del que pugui haver après a nivell científic, m'ho he passat molt bé amb tots ells i, sense ells, segurament no hauria estat pas capaç d'escriure aquesta tesi doctoral. És per això que voldria en aquest apartat fer un petit reconeixement (un reconeixement com cal ocuparia uns quants volums com el que presentem) a tota la gent que m'ha ajudat i que ha estat sempre allà quan ho he necessitat.

En primer lloc, voldria agrair profundament tot l'esforç, ajuda i paciència infinita que el meu director de tesi i amic Francesc Calafell ha tingut sempre amb mi. Moltes gràcies, Francesc, per haver-me introduït en el meravellós món de la ciència, i per haver-me ensenyat a ser crític amb els treballs dels altres però, sobretot, amb el meu propi treball.

També voldria agrair a en Jaume Bertranpetit, el "capo de i capi", per haver-me ensenyat a ser tafaner en el món de la ciència, per les seves xerrades i les seves idees (sempre uns quants de milions de neurones per sobre de la meva capacitat); ha estat un honor poder treballar amb tu, Jaume.

Als companys i companyes amb els que vaig començar; només dir-vos que ha estat una experiència genial, però genial de debò. A la Sephie, per haver aguantat les meves faltades sempre amb bon humor i per haver compartit amb mi totes les xerrades sobre el futur, sempre una mica incert, i que m'han ajudat tant per veure les coses d'una altra manera. A l'Aida, per totes les discussions sobre els pmicrosatel.lits, l'evolució, la selecció positiva (que no existeix, Aida, que no existeix), la política, i també per les classes de tango argentiiiiiino. No era molt bo ballant, però com em vaig arribar a divertir practicant els "sanwichitos" i "la media luna" amb tu i amb l'Ester. A en Jordi Clarimon, per tots els seus consells, les bromes, i els "ets tan abrupte" quan li deia "Jogdi Tontooooo". Sempre de bon rotllo. A l'Anna Perez, la Mònica i l'Anna González (que no AnnaG), per haver fet de germanes grans (que no pas velles!), per totes les històries "made in Anna Pérez" i per tots els consells tan bons que sempre m'han donat. A en David Comas, per ser un bon amic, pel seu riure, sempre contagiós, i per les seves faltades, sempre divertides. A la bona de l'Eva Mateu, per haver-me ajudat tant quan tot just començava. A la Marta Soldevila i al seu marit, en Carles, per tots els

cotis-cotis, per les tertúlies cinèfiles, i per la complicitat que dóna ser al patíbul dels que estan a punt d'acabar. A l'Arcadi Navarro, per totes les converses frikis (i algunes vegades també científiques), per tots els xistes i els emilios una mica picants i per aguantar totes les faltades que més d'un cop t'he fet, encara que fossin de bon rotllo. A en Tomàs Marqués, company aquests dos últims anys, per totes les converses que hem tingut, per tots els dubtes compartits i, com no, per les partides d'Unreal Tournament, la teràpia més bona contra l'estrès mai feta! A l'Anna Petita, l'Anna Ramirez, la Cristina Santa, la Cristine i en Ferran, sang nova que ha arribat fa poc al grup i que d'alguna manera em recorda com era quan jo vaig començar. Tan de bo us vagi tot molt bé. A la Gemma Berniell, que sempre m'ha aguantat amb bon humor les faltades "la tua lejenda personao" que com un lloro em dedico a repetir i a la Michelle Gardner, per totes les converses sobre SNPs, LD i jo que sé quantes coses més! Al pinche Andrés, per haver aguantat sempre amb bon humor els "güeiiiii" i coses per l'estil que no he parat de dir-li desde que va arribar. Als companys del CeGen, en Carlos Morcillo, en Pep i en Marc, perquè m'heu fet sentir com a casa aquest últim any. Friki power!

A tota la gent d'Itàlia. Especialment, a il caro capo Guido Barbujani, la cara capina Isabelle Dupanloup, il caro Giorgio Bertorelle i en Cristiano Venessi, a la Krisztina Vasarhelyi, a la cara Silvia Fusselli (cara Silvia-Bwana) i a la cara Francesca, així com a les estudiants que estaven a punt d'acabar la seva laurea. Grazie mille! Em vàreu fer sentir com a casa, vaig aprendre molt (parolace incloses!) i va ser una experiència genial.

A tota la gent del grup d'en Xavi Domingo, especialment a l'Ainhoa, en Flopez, en Josep Marmi, en Bruno, l'Olga Andres, la Montse i en Thomas. Sense vosaltres la pel.li "El senyor dels anells" no hauria estat pas la mateixa!

A mi amigo Jaime Maillo y su mujer Montse, que siempre han estado a mi lado, tanto en los buenos como en lo tan no buenos momentos. A sus padres, Pedro y Fina, a su hermana Cristina, a Danilo i a los dos pequeños, Marta y Alessio, que cuando estuve en Italia me adoptaron como tío, dándome una de las alegrías más grandes de mi vida y, de paso, que me salvaron de morir por inanición.

A tots els meus amics *BURNINGS*, per haver-me suportat durant tant de temps, per totes les tertúlies alcohòliques sobre temes científics i frikis (realment molt frikis) que fem els caps de setmana i pels bodrios de pel.lícules que ens hem acabat tragant. Gràcies a tots per haver estat sempre al meu costat.

A en Joan Rodon, amic i gran pintor, per totes les converses que hem tingut sobre la vida en general i que en moments difícils tant em varen ajudar. A la Lola Casas, mentora i amiga, que em va iniciar en el maravellós món de la ciència i sempre (però sempre sempre) ha estat allà per ajudar-me. Moltíssimes gràcies per tot, Lola!.

A tota la meva família, els meus pares, la meva àvia, els meus tiets i tietes, especialment la meva tieta Mercedes i el meu tiet Salvador, i els meus cosins, especialment en Javi, l'Aina i en Ferran, que ha suportat amb paciència estoica la meva fixació gairebé masoquista d'estudiar i que sempre m'ha animat a seguir endavant amb el seu amor. A mi hermano Sergio, que me enseñó a no rendirme nunca. Todavía te sigo echando de menos.

Finalment, encara que no en darrer lloc en el meu cor, voldria agrair especialment a la Lourdes i a la seva família per aquests dos últims anys. Gracias por haber estado siempre allí, preciosa, por toda tu comprensión y por todo tu cariño. Gracias por haberme enseñado que existe un mundo más allá de los libros, más maravilloso todavía, al que sólo se puede llegar con el corazón.

1 INTRODUCCIÓ 1

1.1 Variabilitat fenotípica i malaltia 3

1.2 Les forces evolutives 5
1.2.1 La mutació 6
1.2.2 La selecció natural 9
1.2.3 Factors demogràfics 10

1.3 L'origen de les poblacions humanes 13
1.3.1 El model multiregional 13
1.3.2 El model de l'origen recent africà 14
1.3.3 Les evidències genètiques 15
1.3.4 El poblament del continent europeu 17
1.3.5 Gens, *races* i malaltia 19

1.4 Les malalties genètiques 24
1.4.1 Malalties mendelianes 24
1.4.2 Malalties complexes 47
1.4.3 Determinació de variants associades a malalties complexes 51

2 MATERIALS I MÈTODES 67

2.1 Cerca bibliogràfica i construcció de la base de dades 69

2.2 Mapes d'isolínies 71

2.3 Anàlisi d'autocorrelació espacial 72

2.4 Test de Mantel 74

2.5 AMOVA i SAMOVA 74

2.6 MDS 75

3 RESULTATS 77

3.1 Capítol I: "Spatial patterns of cystic fibrosis mutation spectra in European populations." 79

3.2 Capítol II: "Mutation diversity, demographic history and selection in phenylketonuria" 91

3.3 Capítol III: "The spatial pattern of the β-thalassemia mutation spectrum is consistent with the Neolithic spread of malaria" 131

3.4 Capítol IV : "The European Paradox for risk factors in coronary heart disease extends to genetics" **169**

3.5 Capítol V: "Geographic structure of genes conferring risk for Coronary Heart Disease in European populations" **181**

4 DISCUSIÓ 211

4.1 Malalties mendelianes **213**
4.1.1 Anàlisi de la distribució espacial de les mutacions 215
4.1.2 Comparació de les diferents malalties i altres loci 217
4.1.3 La diversitat genètica de les malalties mendelianes 218

4.2 Malalties complexes: la malaltia coronària **225**
4.2.1 Distribució geogràfica de les diferents variants 225
4.2.2 Anàlisi de la covariació en l'espai de CHD i els polimorfismes de susceptibilitat 226

5 BIBLIOGRAFIA 229

1 Introducció

1.1 Variabilitat fenotípica i malaltia

El fenotip, entés com el conjunt de les característiques físiques i psíquiques observables, mesurables i quantificables de cadascun de nosaltres, és el resultat de l'expressió de les instruccions encriptades en el DNA dels nostres gens (el genotip) sota un determinat ambient. La diversitat que observem en els individus es deu tant a diferències genètiques com ambientals. Les actuals estimes de la diversitat genètica indiquen que si prenem dos humans no emparentats a l'atzar, 1 de cada ~1000 nucleòtids serà diferent (Reich et al. 2002).

Aquest és un número molt petit si el comparem amb el d'altres espècies, però en valor absolut és prou gran perquè en els diferents ~30,000 gens que s'estima hi ha en el genoma humà hi hagi canvis que afectin la funció de les proteïnes i, en combinació, confereixin la singularitat fenotípica de l'individu. Però fins i tot en el cas que el genoma és idèntic entre dos individus, com és el cas dels bessons univitelins, la presència de diferents factors ambientals, culturals i estocàstics en la història de cada persona és prou important per fer que cada individu sigui distingible: no hi ha dos fenotips humans idèntics. Malgrat que aquesta diversitat, a vegades és interessant i necessari classificar els individus d'acord amb el seu fenotip; aquest és el cas de les malalties.

Les malalties es poden definir com un tipus especial de fenotip en el qual l'expressió final és la patologia. Des d'Hipòcrates, ara fa més de 2000 anys, l'anàlisi sistemàtica i lògica per trobar les causes naturals implicades en el desenvolupament de la malaltia ha estat una pregunta constant per a la medicina occidental. Entendre en una determinada persona quins són els processos que han donat lloc a un fenotip patològic és entendre quins factors ambientals i quins de genòmics hi estan implicats (vegeu figura 1) i, per tant, veure on s'han d'aplicar les eines terapèutiques per revertir el fenotip. En determinades malalties, els factors genòmics seran tant importants que la patologia es transmetrà seguint els patrons de l'herència mendeliana i els factors ambientals només hi introduiran soroll de fons; en d'altres, però, l'elevat nombre de factors genètics i ambientals interactuant faran que separar-ne els uns dels altres sigui una feina feixuga i difícil.

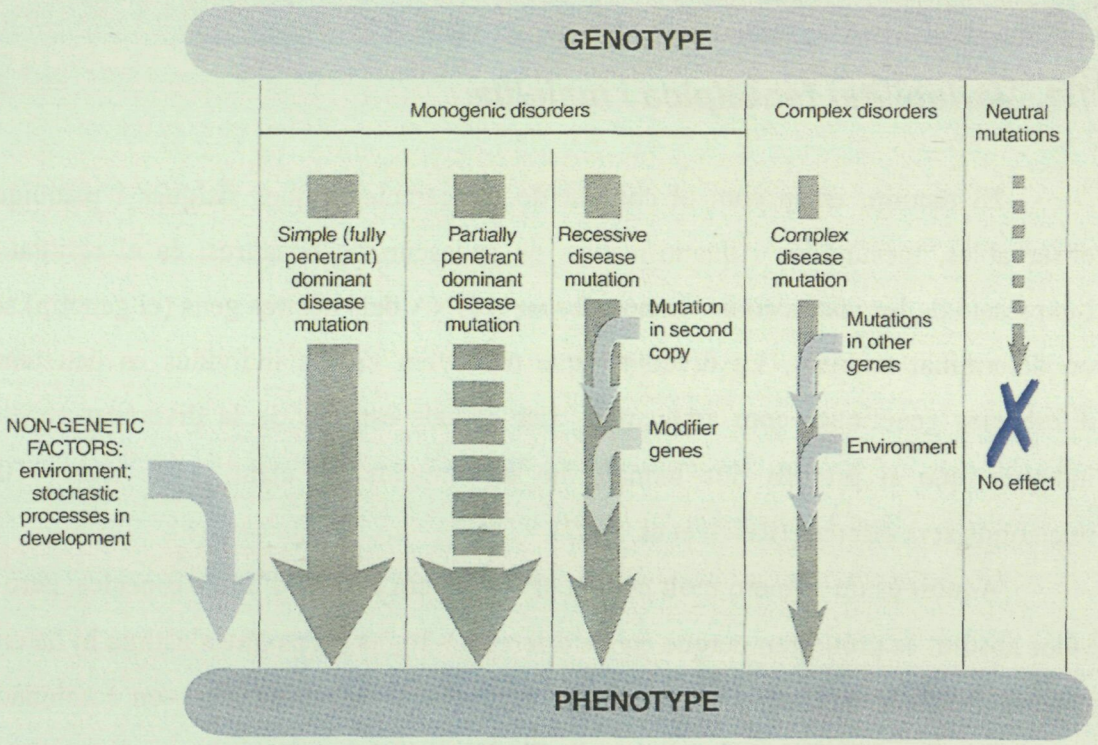


Figura 1. El fenotip és el producte de la interacció entre el genotip i l'ambient (Jobling et al. 2004)

La incidència de moltes malalties no es distribueix de forma homogènia a les poblacions humanes (vegeu figura 2). Entendre les diferències entre les incidències d'una determinada malaltia en diferents poblacions és comprendre quins han estat els factors històrics i evolutius que han conformat les freqüències de les variants causals o associades de la malaltia, així com sota quin context ecològic i socioeconòmic es desenvolupa la malaltia.

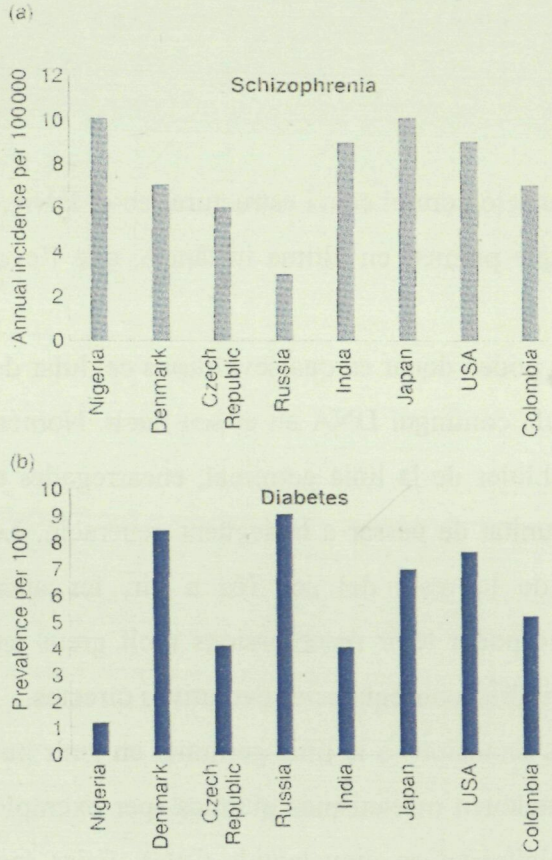


Figura 2. La prevalència de les malalties no és la mateixa a tots els països. (Jobling et al. 2004)

Aquesta informació no és pas trivial, ja que permet conèixer la història natural de la malaltia i aporta pistes sobre quins són els factors últims implicats en la patologia. En el nostre treball hem analitzat la distribució espacial de tres malalties mendelianes, la fibrosi quística, la fenilcetonúria i la β-talassèmia, així com la distribució espacial de set diferents polimorfismes associats a una malaltia complexa com és la malaltia coronària; sota una perspectiva evolutiva, aquest anàlisi ens ha permés entendre millor l'origen i distribució d'aquestes malalties a les poblacions humanes.

1.2 Les forces evolutives

La variació genètica que observem en les poblacions humanes actuals és el resultat d'una complexa interacció entre diferents forces evolutives: la mutació, la selecció i els processos demogràfics. En aquest capítol descrivim breument cadascuna d'aquestes forces fent referència al fenotip patològic.

1.2.1 La mutació

Hom defineix mutació com el canvi estructural en el DNA. És la font última de variació genètica, allò que permet, en última instància, que l'evolució sigui possible (Crow 1997).

Les mutacions es poden donar en qualsevol tipus cel.lular del nostre cos; l'única condició és que la cèl.lula contingui DNA en el seu nucli. Només aquelles mutacions que apareguin en les cèl.lules de la línia germinal, encarregades de la producció dels gàmetes, tindran l'oportunitat de passar a la següent generació. Les mutacions que es produeixin a cèl.lules de la resta del cos (és a dir, les anomenades mutacions somàtiques), malgrat que poder tenir repercussions molt greus en l'individu, com és l'aparició de càncer, no tindran conseqüències evolutives directes.

La majoria de les mutacions a la línia germinal en gens humans són el resultat d'errors endògens que inclouen mecanismes químics (per exemple, la deaminació del grup metil de la 5-metilcitosina en dinucleòtids CpG), físics (com el lliscament de DNA) o enzimàtics (com la reparació de nucleòtids mal aparellats). Segons la grandària del canvi estructural del DNA, les mutacions es classifiquen en: petits canvis estructurals que inclouen substitucions d'un únic nucleòtid, petites delecions i petites insercions, i grans canvis estructurals que inclouen expansions de tractes repetitius, grans insercions i duplicacions, reordenaments complexos i grans delecions (Human Gene Mutation Database ; HGMD (http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html)). Estudiant la freqüència de les mutacions que causen malaltia descrites en la base de dades del HGMD (vegeu figura 3) es pot observar que gairebé el 70% de les mutacions són causades per substitucions en un únic nucleòtid, seguides per petites delecions (el 16.5%), petites insercions (el 6.5%) i grans canvis (7.5%); ara bé, cal tenir en compte que donat que només un exemple de cada mutació és emmagatzemat a la base de dades, aquests resultats infravaloren el nombre total de mutacions, ja que esdevé impossible distingir aquelles mutacions que són idèntiques per descendència d'aquelles que són recurrents (Scriver et al. 2000). També cal tenir en compte que, donat que aquesta base de dades està formada amb canvis genètics causants de malalties, només inclou aquelles mutacions que produeixen un fenotip postnatal prou sever per cridar l'atenció dels

clínics (Krawczak et al. 1998) i, per tant, és poc probable que reflecteixi la taxa de mutació real (Scriver et al. 2000).

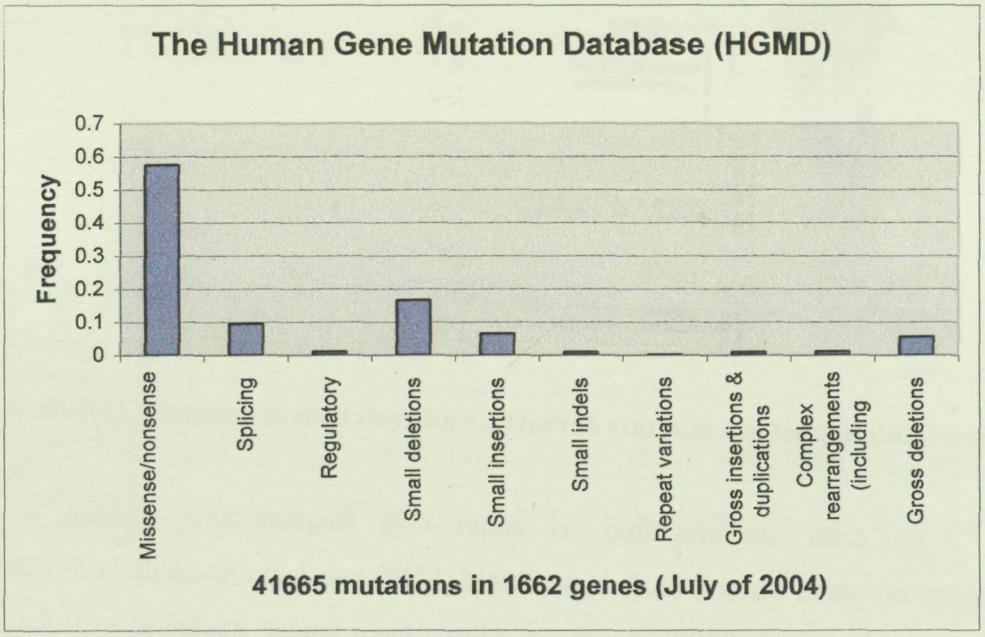


Figura 3. Espectre dels diferents tipus de mutacions en gens humans. Extret a partir de la base de dades HGMD

Així doncs, malgrat que la mutació és un procés estocàstic, no podem pas parlar d'una única taxa de mutació: la dinàmica mutacional depèn tant del tipus de canvi (sovint dependent de la seqüència) com del mecanisme implicat en el canvi, la qual cosa fa que no totes les mutacions ocorrin amb la mateixa freqüència (vegeu figura 4).

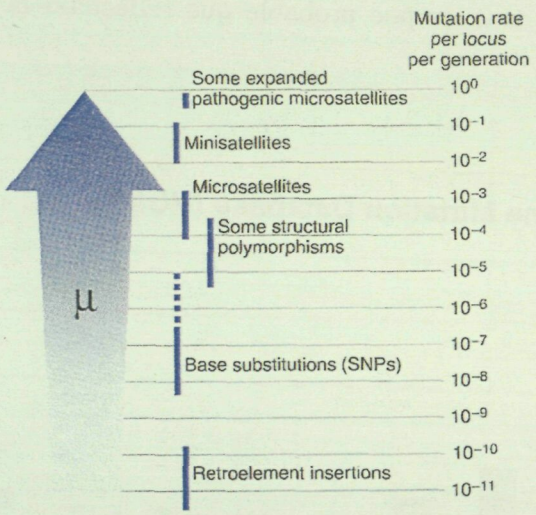


Figura 4. Diferents estimes de la taxa de canvi per diferents tipus de mutacions. (Jobling et al. 2004)

En ordre de magnitud, el canvi més freqüentment associat a malaltia monogènica és el canvi d'un únic nucleòtid (SNP) per deaminació de la 5-metilcitosina timina (Scriver et al. 2000); donat que normalment aquest nucleòtid es troba aparellat amb el nucleòtid guanina formant dinucleòtids CpG, el canvi que s'observa és cap a TG o CG, depenent de la cadena de DNA ("*sense*" o "*antisense*") en la que es produeixi el canvi de C cap a T (Antonarakis et al. 2000). La deaminació continuada de la 5-metilcitosina fa que la proporció de dinucleòtids CpG sigui un cinquè de l'esperada donada la fracció de C i G en el genoma humà; una excepció són les regions riques en dinucleòtids CpG (les anomenades "illes" CpG) que no estan metilades i ocorren amb la freqüència esperada donat el contingut local de nucleòtids C i G (Lander et al. 2001). La taxa de mutació d'aquests dinucleòtids no és la mateixa en homes que en dones; existeix una marcada asimetria cap a produir-se la mutació en la línea germinal masculina i, a més, a incrementar-se a mesura que l'edat de l'home augmenta. Aquest fenomen està relacionat amb el fet que el DNA dels espermatozoides està molt més metilat que el dels oòcits (Scriver et al. 2000) i que el nombre de divisions cel.lulars a la línia germinal dels espermatozoides és molt més elevat respecte al dels oòcits (Crow 2000; Lynn et al. 2004) (vegeu figura 5).

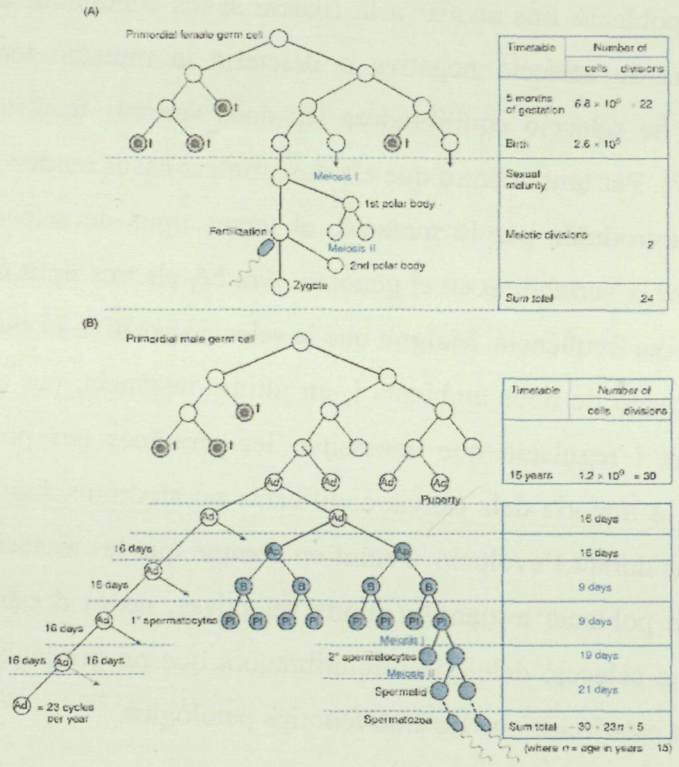


Figura 5. El nombre de divisions cel.lulars és molt més gran a la línia germinal masculina que a la femenina. A partir de (Strachan and Read 1999)

1.2.2 La selecció natural

Si bé la mutació és la font última de generació de diversitat, altres forces evolutives són les encarregades de modificar les freqüències de les noves mutacions en les poblacions. La selecció natural es defineix com el canvi direccional de la freqüència d'una determinada mutació en les següents generacions, conseqüència de l'efecte que aquesta produeix en l'organisme quan interacciona amb l'ambient. La selecció es pot donar al llarg del període que porta des de la formació de l'individu fins a la generació de progènie viable. Segons el fenotip que produeix la mutació, podem distingir tres tipus principals de selecció: (i) una mutació se selecciona *positivament* si el fenotip que produeix en un determinat ambient confereix algun avantatge a l'individu, de forma que deixa més descendència a la següent generació en comparació amb els altres individus de la població. (ii) Una mutació se selecciona *negativament o deletèriament* si fa que les probabilitats de deixar descendència a la següent generació siguin més petites en comparació amb els altres individus de la població. (iii) Una mutació es selecciona de forma *equilibradora* si la capacitat dels individus heterozigots per deixar descendents a la següent generació és superior a la dels individus homozigots. D'acord amb la genètica de poblacions clàssica, en el cas de selecció positiva la mutació incrementarà la

seva freqüència a la població fins arribar a la fixació si les condicions ambientals es mantenen i en el cas de selecció negativa o deletèria la mutació tendirà a ésser eliminada; en el cas de selecció equilibradora les dues variants tendiran a coexistir (Hartl and Clark 1997). Per tant, mentre que els dos primers casos tendeixen a reduir la variabilitat genètica introduïda per la mutació, el tercer tipus de selecció tendeix a mantenir i incrementar la variabilitat en el genoma. Ara bé, els tres tipus de selecció no es donen amb la mateixa freqüència. Malgrat que la selecció positiva és essencial per als processos de colonització de nous ambients i, en última instància, per a l'especiació, l'elevada especificitat i regulació que necessiten les proteïnes per poder funcionar correctament fa que la majoria dels canvis codificants no sinònims siguin severament penalitzats i eliminats durant l'evolució. Tenint en compte que les mutacions deletèries són eliminades de la població mitjançant morts selectives, no és d'estranyar que les mutacions sotmeses a selecció deletèria i equilibradora que no produeixin letalitat en l'embrió acostumin a estar relacionades amb fenotips patològics.

1.2.3 Factors demogràfics

La selecció només es podrà dur a terme en aquelles variants genètiques que, d'alguna forma, modifiquin la funcionalitat dels gens i facin que l'individu portador tingui un nombre de fills diferents de la resta de la població. Aquelles mutacions que produeixin fenotips patològics més enllà de l'edat reproductiva (com per exemple, les associades a la malaltia de Huntington (OMIM 143100)), no es veuran afectades per la selecció. A més, la redundància del codi genètic fa que no totes les mutacions produeixin un canvi aminoacídic i, per tant, la funció de la proteïna no es vegi afectada. Finalment, donat que només un ~1.5% del nostre genoma conté gens (uns 24.500 segons les últimes estimes, (Pennisi 2003)), un important nombre de mutacions es produeix en regions no codificants. Més enllà de l'efecte que aquestes mutacions puguin tenir sobre el control de l'expressió dels gens, sembla evident que moltes mutacions són neutres a nivell selectiu i que, per tant, el canvi de les seves freqüències a les poblacions, es degui principalment als factors demogràfics.

A diferència dels dos processos evolutius anteriors, els factors demogràfics no són específics de locus i afecten alhora tota la variabilitat genètica present en el genoma

humà. Hom distingeix dos tipus de factors demogràfics: la deriva genètica i els moviments poblacionals (migració i colonització).

1.2.3.1 La deriva genètica

La deriva genètica és un procés estocàstic que es defineix com la fluctuació a l'atzar de les freqüències gèniques d'una mutació d'una generació a la següent. Aquesta fluctuació és conseqüència directa del fet que tant el nombre d'individus com el de descendents a les poblacions és finit, el que fa que, pel simple fet de mostrejar els al.lels que passaran a la següent generació, les seves freqüències augmetnin o minvin a l'atzar. Una situació anàloga es produeix quan llancem a l'aire una moneda: la probabilitat que una de les cares surti un determinat nombre de vegades segueix una distribució binomial. Si el nombre de llançaments és prou gran, el 50% de les vegades observarem una de les dues cares, però si el nombre de llançaments és petit, llavors hi haurà més possibilitats d'obtenir un número de cares o de creus que s'aparti de la freqüència que hom esperaria. Igual que en l'exemple de la moneda, els canvis produïts per deriva es poden descriure matemàticament en termes de probabilitat, però no d'una manera determinista. En les successives generacions, les fluctuacions produïdes per deriva es poden anar acumulant, allunyant-se de l'estat inicial, fins que un dels al.lels s'acaba fixant amb una probabilitat de 1/2N (on N és la grandària poblacional). Per tant, la magnitud de la deriva genètica depèn directament de la grandària de la població.

1.2.3.1.1 L'efecte fundador

L'efecte fundador es produeix quan un petit nombre d'individus d'una població més gran funden una nova població, un procés relacionat amb la colonització. Donat que només un reduït nombre dels cromosomes presents a la població original es trobarà a la nova població, la conseqüència més immediata de l'efecte fundador és que a la nova població només es trobarà representada una fracció de la variabilitat genètica original; per atzar, variants que a la població original es troben a baixa freqüència podran incrementar-se a la població colonitzadora. A més, en una població petita es tendirà a incrementar la consanguinitat i, per tant, els nivells d'homozigositat. Malgrat que després la nova població pugui créixer ràpidament, com que la grandària poblacional efectiva a llarg termini és defineix més per la mitjana harmònica que per l'aritmètica (Wright 1938; Crow and Kimura 1970), la petjada sobre la variabilitat genètica serà

detectable moltes generacions després de la colonització inicial. Tot això fa que variants deletèries recessives, presents a baixa freqüència a la població original, puguin incrementar la seva freqüència a la població colonitzadora gràcies a la deriva genètica, causant malalties que a altres poblacions són inexistents o molt poc freqüents.

1.2.3.1.2 El coll d'ampolla

El coll d'ampolla es produeix quan hi ha una reducció dràstica del nombre d'individus en una població. Bàsicament, els efectes del coll d'ampolla són equivalents als de l'efecte fundador però menys marcats. Tot i que el nombre d'individus es pot incrementar fins arribar als nivells inicials, la diversitat genètica pateix una gran reducció.

1.2.3.2 La colonització i la migració

Hom denomina colonització a la fundació de noves poblacions en territoris prèviament inhabitats. Les colonitzacions inclouen des de la fundació de petites poblacions relativament aïllades per part de poblacions grans fins la conquesta de grans àrees per part d'una població petita. El posterior aïllament continuat de les poblacions portarà necessàriament a la fixació d'aquells al.lels que no es trobin sota selecció equilibradora per deriva genètica, selecció positiva i negativa, el que farà que les poblacions es diferenciïn genèticament i permetrà a nous processos d'especiació. La migració es defineix com l'intercanvi d'individus (migrants) entre diferents poblacions ja establertes. El flux genètic continuat entre poblacions (també anomenades dems) porta a homogeneitzar la variació genètica present a les poblacions abans d'iniciar l'intercanvi, donant lloc a una població global o metapoblació (Hey and Machado 2003). La petjada de la migració serà observable sempre i que la freqüència dels al.lels sigui diferent entre les poblacions i el procés migratori no hagi encara homogeneitzat les freqüències al.lèliques a les poblacions.

Les poblacions humanes actuals són el resultat de processos colonitzadors i migratoris extremadament complexos (Ray et al. 2003); entendre aquests processos i desenvolupar models realistes sobre les migracions humanes són essencials per poder

entendre i trobar les variants al.lèliques associades a processos patològics (Goldstein and Chikhi 2002).

1.3 L'origen de les poblacions humanes

Els humans som primats; els nostres parents més propers són els ximpanzés i els bonobos, dels quals vàrem divergir ara fa entre 4 i 6 milions d'anys i amb els quals tenim un 99.9% del nostre genoma idèntic (Svante Paabo, (Paabo 2003)). Malgrat que el registre fòssil indica que altres espècies d'homínids varen existir després de la separació del nostre llinatge amb el del ximpanzé, en l'actualitat som els únics representants no extints del gènere *Homo*.

Els fòssils més antics d'homínids s'han trobat a l'Àfrica, cosa que indica un origen africà del gènere *Homo*. La dispersió del gènere *Homo* per la resta del món no començà fins *Homo erectus/ergaster*, les restes dels quals s'han trobat a Àfrica, Europa, Orient Mitjà i Àsia en un període comprés entre els 2 i els 0.5 milions d'anys (Anton 2003); descendents d'aquesta primera diàspora són els conegut en conjunt com a "sàpiens arcàics", entre els quals s'inclou *Homo neanderthalensis*, que va habitar el continent europeu entre els ~250.000 i els ~28.000 anys.

L'origen dels humans anatòmicament moderns ha estat envoltat d'una forta polèmica entre els antropòlegs durant més de mig segle. Dos models extrems, el **model multiregional** i el **model d'origen recent africà**, intenten explicar la transició entre els "sàpiens arcaics" i els humans anatòmicament moderns, però partint de punts de vista totalment oposats.

1.3.1 El model multiregional

El model multiregional (o MRE de l'anglès MultiRegional Evolution), també anomenat en canelobre, postula que el llinatge humà ha evolucionat com un únic llinatge sense events d'especiació, existint una continuïtat genètica des de els primers habitants d'*Homo erectus/ergasters* fins els humans anatòmicament moderns actuals (Foley 1998)(vegeu figura 6). L'origen geogràfic dels humans anatòmicament moderns, per tant, no hauria estat únic (multiregionalitat). Donat que l'aïllament continuat de les poblacions de sapiens arcaiques hauria produït fenòmens d'especiació, aquest model

requereix que hagi existit un flux genètic continuat entre els grups continentals i, per tant, grandàries de població efectives molt grans.

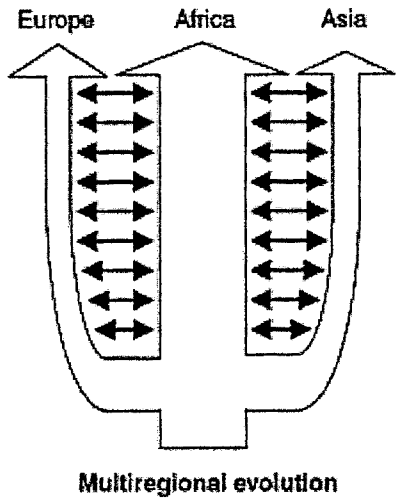


Figura 6. El model regional. Adaptat de Excoffier (Excoffier 2002)

1.3.2 El model de l'origen recent africà

El model de l'origen recent africà (o RAO de l'anglès Recent African Origin), també anomenat "fora d'Àfrica" (o OOA de l'anglès Out of Africa) o "jardí de l'Edèn" (Harpending and Rogers 2000), postula que els humans anatòmicament moderns varen evolucionar a partir d'una petita població africana aïllada i genèticament homogènia que posteriorment va colonitzar tot el món, reemplaçant les poblacions i espècies d'homínids anteriors, com els neandertals, ara fa uns ~120.000 – 200.000 anys (vegeu figura 7). Per tant, segons aquest model, les poblacions humanes actuals serien molt joves i genèticament molt properes.

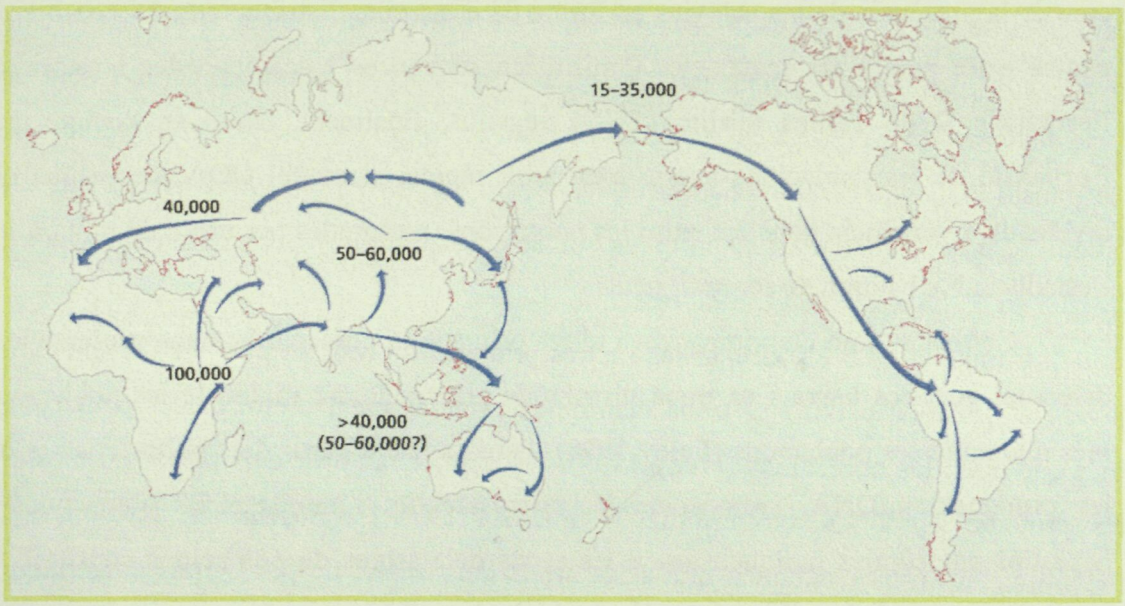


Figura 7. La dispersió de les poblacions humanes al món segons el model "fora d'Àfrica". Mapa modificat a partir de (Cavalli-Sforza and Feldman 2003)

1.3.3 Les evidències genètiques

El potencial de les dades gèniques per resoldre aquesta controvèrsia sembla clar, donat que els patrons de diversitat genètica s'han de preservar a la història demogràfica (Harpending and Rogers 2000). Com han apuntat Goldstein i Chikhi (Goldstein and Chikhi 2002), però, la combinació de diferents processos evolutius poden tenir efectes similars i donar lloc a petjades sobre la diversitat genètica humana molt semblants. Distingir l'efecte de cadascun és una feina sovint difícil i complicada, cosa que fa que les possibles inferències genètiques que puguem fer sense cap altra evidència exterior tendeixin a tenir poca potència i requereixin de la informació d'altres camps, com l'arqueològic o el lingüístic.

D'acord amb el model MRE, hom esperaria trobar uns nivells de variació genètica elevada a les poblacions humanes, producte dels més de dos milions d'anys d'evolució separada, així com temps de coalescència antics entre els llinatges genètics trobats a poblacions geogràficament disperses (Harpending and Rogers 2000), mentre que en el cas de OOA, la reducció de la diversitat genètica de tot el genoma produïda pel coll d'ampolla faria que la diversitat alèlica trobada fora d'Àfrica fos una submostra de la trobada dintre d'Àfrica i disminuís de forma gradual a mesura que ens anéssim allunyant del continent africà; la posterior expansió postulada pel model OOA també hauria d'haver deixat la seva petjada a la variabilitat genètica de les poblacions actuals,

com és la presència de genealogies en forma de d'estrella, distribucions del nombre de canvis entre parells de seqüències ("mismatch sequences") acampanades i valors de l'estadístic D de Tajima (Tajima 1989) negatius; finalment, tenint en compte que l'expansió de les poblacions hauria estat molt ràpida, el model OOA prediu que els nivells de divergència genètica entre les poblacions, mesurades per exemple mitjançant l'estadístic F_{st}, haurien de ser molt petits.

Comparat amb els ximpanzés i altres primats, la població humana presenta una diversitat genètica baixa i es troba distribuïda principalment dintre de les poblacions més que entre les poblacions (Foley 1998). Estudis fets a partir dels polimorfismes de les proteïnes, mtDNA, cromosoma Y, microsatèl.lits i minisatel.lits autosòmics i haplotips autosòmics indiquen que a les arrels dels arbres de poblacions construïts a partir d'aquestes dades s'hi troben principalment les poblacions africanes i/o els africans tenen els llinatges més divergents (Tishkoff and Verrelli 2003). Les distribucions de les diferències entre parells de seqüències, estimades a partir del DNA mitocondrial (mtDNA) en diferents poblacions, mostren patrons acampanats, compatibles amb una recent expansió de les poblacions que s'hauria produït entre 30.000 i 130.000 anys i, per tant, donaria suport al model OOA. La diversitat genètica de diferents marcadors genètics trobada fora d'Àfrica acostuma a ser una part de la trobada dintre d'Àfrica i presenta una reducció gradual a mesura que ens allunyem del continent africà (Excoffier 2002); també trobem uns nivells de desequilibri de lligament menors en poblacions africanes en comparació amb poblacions fora d'Àfrica, cosa que indica que la grandària poblacional de les poblacions humanes a Àfrica hauria estat tradicionalment més elevada que a la resta del món. Finalment, el càlcul de la D de Tajima en 313 gens és negatiu en 281 d'ells (el 90%) (Stephens et al. 2001), indicant de forma clara que el DNA nuclear mostra senyals d'expansions passades més que de selecció positiva (Excoffier 2002). Tots aquests resultats semblen evidenciar que l'origen dels humans anatòmicament moderns es va produir a Àfrica, ja que sota el model MRE la petjada d'expansions de la població només es podria obtenir a nivell mundial si s'haguessin produït colls d'ampolla simultanis i posteriors recolonitzacions als diferents continents (Excoffier 2002). Ara bé, la subestructuració geogràfica que observem a les poblacions humanes i que es tradueix en estimes de divergència genètica (calculades amb F_{st}) d'entre el 10% i el 15% no es poden explicar només amb el model OOA (Harpending and Rogers 2000), cosa que fa pensar que aquest model és una sobresimplificació del procés real. Si el flux genètic hagués estat prou gran entre les poblacions veïnes, per

exemple, expansions geogràfiques de poblacions petites també podrien haver produït genealogies en forma d'estrella i estadístics D de Tajima negatius (Ray et al. 2003).

És evident, per tant, que calen models més complexes per explicar de forma més realista el complicat tapís de la variabilitat genètica que observem en les poblacions humanes actuals. El model Jardí de l'Eden dèbil ("weak Garden of Eden"; representat a la figura 8), per exemple, modifica el model OOA suggerint que les poblacions podrien haver romàs petites i subdividides durant un temps després de la migració dels humans fora d'Àfrica per després produir una ràpida expansió ara fa uns 50.000 anys (Harpending and Rogers 2000).

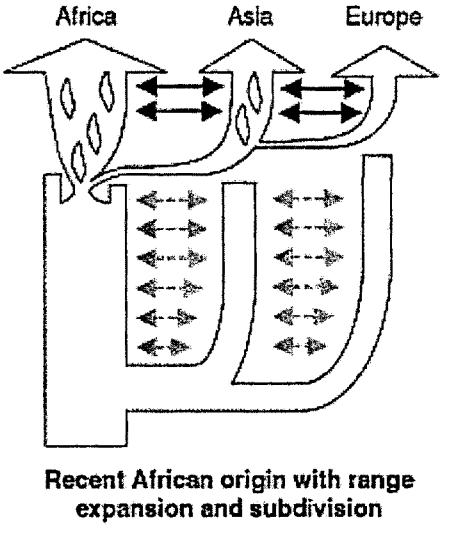


Figura 8. El model "weak garden of Eden". Adaptat de Excoffier (Excoffier 2002)

1.3.4 El poblament del continent europeu

El poblament del continent europeu és un complex fenomen de grans migracions (vegeu figura 9; (Dipple and McCabe 2000b)). Els humans anatòmicament moderns varen arribar al continent europeu fa uns 45.000 anys; les primeres poblacions d'*Homo sapiens* estaven constituïdes per petits grups de caçadors-recolectors associats a la cultura Gravetiana (Stiner 2001; Barbujani and Goldstein 2004) que, en colonitzar el continent, varen coexistir amb poblacions descendents de la primera diàspora fora d'Àfrica per part de *l'Homo erectus/ergasters*. Aquesta espècie homínida havia evolucionat per adaptar-se a les condicions climàtiques del continent europeu, regides per grans periodes glacials, donant lloc als neandertals (Jobling et al. 2004). Les dades

moleculars suggereixen que el flux genètic va ser molt petit o inexistent entre les dues espècies homínides (vegeu, per exemple, (Caramelli et al. 2003)), i que les poblacions d'humans anatòmicament modern varen reemplaçar les poblacions de neandertals. Durant l'últim període glacial, ara fa entre uns 23.000 i 14.000 anys, bona part del nord d'Europa es trobava coberta pel glaç i les poblacions humanes es varen veure obligades a migrar cap a latituds més càlides, en el que després s'han anomenat refugis glacials, a la Península Ibèrica, el nord de la Península Itàlica i els Balcans, des d'on posteriorment es varen començar a expandir, fa uns 13.000 anys (Pinhasi et al. 2000). Aproximadament fa 10.000 anys, a la zona del Pròxim Orient coneguda com Creixent Fèrtil, es produeix la denominada revolució neolítica que bàsicament consisití en el descobriment de noves tecnologies i, el que és més important, el desenvolupament de l'agricultura. Aquesta revolució va tenir conseqüències dràstiques, no només en l'estil de vida, sino també en els patrons demogràfics de les poblacions d'agricultrors, ja que la nova font estable d'aliments va propiciar un increment de la població sense precedents a les poblacions humanes (Jobling et al. 2004). La revolució tecnològica del neolític es va anar expandint per tota Europa, primer pel sud i després pel nord. Dues hipòtesis diferents s'han postulat per explicar com es va expandir la revolució neolítica a Europa. D'acord amb la **difusió cultural**, la tecnologia i l'agricultura s'haurien expandit entre els caçadors-recolectors que poblaven el continent, mentre que d'acord amb **la difusió dèmica**, la dispersió de l'agricultura hauria estat acompanyada per grans migracions i expansions de poblacions d'agricultors. La contribució neolítica en el genoma de les poblacions europees actuals és un debat de grans passions encara no completament resolt (vegeu, per exemple, les opinions de Guido Barbujani i Martin Richards a (Jobling et al. 2004)). La possibilitat de detectar la petjada genètica que haurien deixat aquest dos models a la població europea actual depèn en gran mesura de la diferenciació genètica preexistent entre les poblacions de caçadors-recolectors i les d'agricultors. En un model de difusió cultural s'esperaria trobar una diferenciació marcada a les poblacions europees, mentre que en el model de difusió dèmica esperaríem no trobar elevades afinitats genètiques entre les poblacions europees. Models intermedis, en els quals els agricultors migren i s'aparellen amb els caçadors-recolectors, produirien gradients en les freqüències gèniques (Jobling et al. 2004). L'anàlisi de components principals realitzat per Cavalli-Sforza i col.laboradors (Cavalli-Sforza et al. 1994) i posteriors anàlisis del patró de distribució espacial dels marcadors clàssics (Sokal et al. 1989) mostren que la distribució espacial de les freqüències

gèniques al continent europeu és clinal, cosa que s'ha interpretat com una prova a favor de la difusió dèmica. Les estimes de contribució neolítica a les poblacions europees actuals varien segons l'estudi, les dades genètiques i la metodologia emprada (vegeu, per exemple, (Chikhi et al. 2002) i (Semino et al. 2000)). Un recent estudi analitzant marcadors autosòmics, cromosoma Y i mtDNA (Dupanloup et al. 2004), assumint un model de mescla entre les poblacions paleolítiques i neolítiques i una posterior deriva genètica en les poblacions parentals i les produïdes pel contacte, troba que aproximadament un 50% del bagatge genètic podria ser d'origen neolític, amb percentatges més elevats a les poblacions properes a la regió del Pròxim Orient.

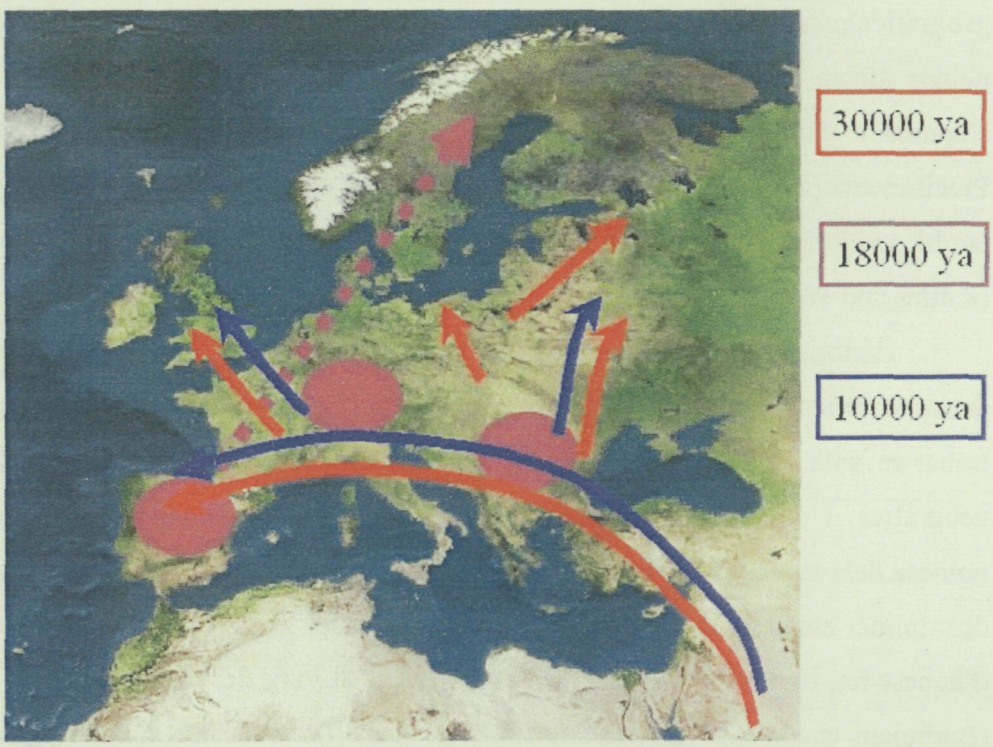


Figura 9. Principals migracions humanes al continent Europeu. Adaptat de (Simoni et al. 2000)

1.3.5 Gens, *races* i malaltia

Donat que a la majoria de malalties hi ha un component hereditari, poder definir grups de poblacions genèticament homogenis i determinar amb una certa probabilitat l'origen d'un individu a partir del seu genoma és molt important des d'un punt de vista biomèdic, ja que la utilització de poblacions genèticament homogènies eliminaria el problema de la subestructuració en els estudis epidemiològics i afavoriria la detecció de

grups de poblacions amb riscs genètics de patir determinades malalties o de tenir reaccions adverses front un determinat fàrmac. De fet, el concepte de *raça*, o grup d'individus genèticament homogenis en front d'altres grups (Kittles and Weiss 2003), està fortament arrelat a la literatura científica mèdica. Fent una cerca a la base de dades bibliogràfiques PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) per articles d'assaigs clínics en humans, en trobem 3741 (agost del 2004) on apareixen els termes "(race) OR (ethnicity)", un 42% (1561 articles) dels quals fa menys de cinc anys que han estat publicats. Igualment, la classificació de determinades malalties segons grups de població és una pràctica força estesa dintre de la literatura científica, especialment quan aquestes malalties són poc freqüents i estan ben definides geogràficament; així, per exemple, es diu que la fibrosi quística és una malaltia que només afecta europeus o que l'anèmia falciforme només afecta africans; però com s'apunta a (Barbujani et al. 1997), cal ser prudent amb aquest tipus d'afirmacions. Precisament perquè són poc freqüents, es troben ben delimitades espacialment, però també precisament perquè són poc freqüents, pocs individus de la població les pateixen (Kittles and Weiss 2003).

Històricament, les races s'han definit a partir de caràcters morfològics i antropomètrics, com el color de la pell o la forma del crani, els quals acostumen a trobar-se sota l'efecte de la selecció positiva; la llengua, la cultura, la localització geogràfica, i d'altres característiques més esperpèntiques, com la noblesa o falta de noblesa dels salvatges, són altres criteris, sovint massa subjectius, que s'han emprat per discriminar els individus en grups racials (Kittles and Weiss 2003); com a conseqüència d'aquest fet, el nombre de races ha anat variant al llarg del temps des de 3 fins a 200 (Barbujani et al. 1997) sense que importés gaire la validesa d'aplicar aquest terme a poblacions humanes. L'anàlisi dels gens, en principi, hauria de donar-nos una mesura molt més acurada de les diferències que observem entre les poblacions humanes i, arribat el cas, permetre'ns definir objectivament grups genèticament homogenis. Des d'un punt de vista genètic, la definició de races requereix que els individus de cada raça hagin divergit dels individus d'altres races prou temps perquè s'acumulin diferències genètiques, el que implica fenòmens d'aïllament entre les diferents races i mescla dintre de les poblacions de cada raça. El flux genètic continuat i fenòmens de selecció equilibradora tendiran a disminuir les diferències genètiques. Tradicionalment, un criteri biològic (malgrat que subjectiu) per definir subespècies és tenir estimes de diversitat genètica F_{st} superiors a 0.25 (Kittles and Weiss 2003). D'acord amb el model OOA, o

modificacions d'aquest, la població humana actual es va originar a Àfrica, d'on posteriorment es va expandir colonitzant la resta del món mitjançant successius events fundacionals. La dispersió de les poblacions humanes pel món va comportar que la variació genètica que s'observa en els gens majoritàriament mostri patrons espacials continus, de gradació (patrons clinals) seguint una distribució geogràfica d'Àfrica cap a altres continents (vegeu figura 10).

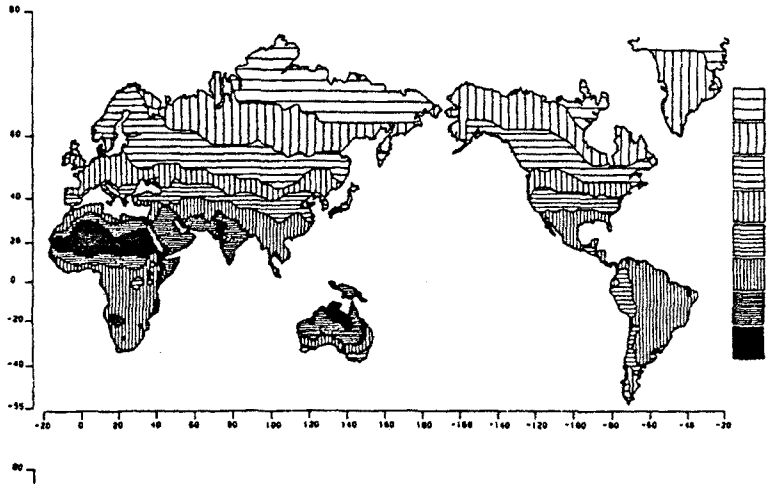


Figura 10. Mapa de components principals de variants genètiques electroforètiques on es pot observar que la variació dels marcadors clàssics és distribueix geogràficament en un patró clinal. A partir de Cavalli-Sforza el al

L'increment de la distància geogràfica entre les poblacions va propiciar l'aïllament, el qual es reflecteix en una correlació positiva entre les diferències genètiques i la distància geogràfica entre les poblacions (Bamshad et al. 2004); segons Neil Risch N i col.laboradors (Risch et al. 2002), aquesta subestructuració de les poblacions es podria traduir a nivell continental, i les poblacions es podrien classificar en cinc grups racials diferents: africans, caucasians, illencs del Pacífic, asiàtics de l'est i americans aborigens. Ara bé, la quantificació d'aquesta subestructuració geogràfica a nivell continental, calculada en diferents estudis (per exemple, (Romualdi et al. 2002)), indica que només entre un 10%-15% de la variació genètica total s'explica per les agrupacions continentals, mentre que més d'un 80% de la variació total es troba entre els individus dintre de cada població. Aquests resultats s'han interpretat tradicionalment com una prova de la continuïtat genètica a nivell espacial i, per tant, de la poca consistència del concepte de raça en les poblacions humanes (o, dit d'una altra manera, que les diferències que s'observen entre els diferents grups racials són bàsicament cosmètiques). Cal, però, no confondre una baixa diversitat genètica amb l'absència de

subestructuració de les poblacions: una estima de F_{st} més gran de 0 indica que realment existeix subestructuració a les poblacions (Edwards 2003). Els processos migratoris afecten per igual a tot el genoma i fan que la variació present a cada gen no sigui independent de la variació present a la resta; per tant, l'anàlisi de la variació d'un elevat nombre de gens ens obre la possibilitat de poder trobar l'afiliació poblacional dels individus (Edwards 2003). Estudis recents utilitzant l'algoritme implementat al programa STRUCTURE (http://pritch.bsd.uchicago.edu/), com els de Rosenberg i col.laboradors (Rosenberg et al. 2002) o Bamshad i col.laboradors (Bamshad et al. 2003), mostren que si s'agafa un nombre prou elevat de polimorfismes es pot arribar, amb una certa probabilitat, a classificar els individus d'acord amb el seu origen continental. És aquesta una prova que realment existeixen les races? La qüestió és molt més complexa del que sembla. Una crítica que es pot fer a aquests estudis és el fet que les poblacions analitzades no són representantives de tot el rang espacial continental (estan massa separades geogràficament) i per tant la gradació real que s'observa en la variació genètica en les poblacions s'estarà infravalorant; de fet, els individus de poblacions intermitges, com les del sud de la Índia, es distribueixen entre els europeus i els asiàtics quan es repeteix l'anàlisi ((Bamshad et al. 2004); vegeu figura 11).

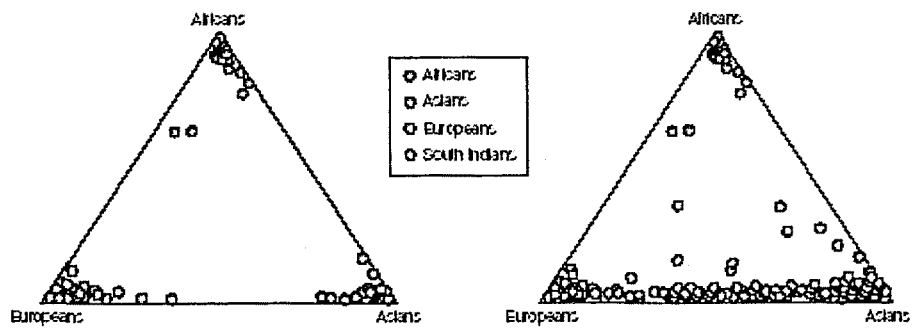


Figura 11. Classificació en tres grups dels individus. Quan només es consideren africans, asiàtics i europeus, els grups d'individus definits per STRUCTURE es corresponen als grups geogràfics. Quan s'inclou un grup d'indis, aquests es distribueixen entre els europeus i els asiàtics.

D'altra banda, malgrat que la diferenciació genètica pugui ser molt petita, quan aquests estudis varen analitzar la subestructuració dintre de cada continent, encara varen ser capaços de distingir subgrups de poblacions, indicant que els grups continentals no eren tan genèticament homogenis (Rosenberg et al. 2002). Des del punt de vista biomèdic, la classificació dels individus a nivell continental només podria tenir sentit si

les variants associades a la malaltia estiguessin delimitades continentalment (és a dir, si fossin producte de fenòmens selectius (Calafell 2003)) o si fossin prou comunes i mostressin un patró de variació genètica continental, una hipòtesi que ha estat demostrada teòricament, a l'igual que la contrària (vegeu més endavant). Les dades experimentals mostren una gran disparitat de resultats: mentre que alguns treballs semblen apuntar que són les variants comunes les que contribueixen a la susceptibilitat de les malalties comunes (Lohmueller et al. 2003), un article recent (Cohen et al. 2004) mostra que els nivells de colesterol associats a HDL estan controlats majoritàriament per al.lels rars amb un elevat efecte genotípic. Per una altra banda, l'estratificació podria arribar a tenir importants efectes sobre les conclusions d'estudis cas-control fins i tot entre les poblacions d'un mateix continent (Marchini et al. 2004), especialment perquè per trobar variants de baix risc el nombre d'individus que cal és molt gran ((Barbujani and Goldstein 2004), i dintre de les pròpies poblacions, ja que la selecció de la parella no acostuma a ser a l'atzar, i depèn tant de la localització geogràfica com de factors culturals, socioeconòmics i religiosos. Però fins i tot si moltes de les variants gèniques associades a malaltia fossin comunes a totes les poblacions mundials, factors socioculturals, com podria ser la construcció social de la raça, podrien estar correlacionant amb factors ambientals, com l'estil de vida, el que podria introduir factors confusors. Com apunten Kittles i Weiss, el concepte de raça és molt complex perquè, a part de les evidències biològiques, altres factors socioculturals i històrics influeixen de forma subjectiva sobre el criteri de l'investigador (Kittles and Weiss 2003). Un exemple flagrant d'aquest fet és la classificació com a hispans que normalment es fa a Nord Amèrica dels individus d'Amèrica del Sud hispanoparlants, sense tenir en compte per res el bagatge genètic i el nivell de mestissatge de les diferents poblacions d'Amèrica del Sud.

1.3.5.1 Bo abans, dolent ara: la hipòtesi de les variants estalviadores

Des de temps paleolítics els humans varen haver d'adaptar-se als diferents ambients amb què es varen trobar durant la diàspora africana. Aquesta adaptació afectà, per exemple, l'aprofitament metabòlic que el cos feia dels aliments ingerits en la dieta dels caçadors-recolectors, acostumats a recórrer grans distàncies buscant aliments. En aquest passat llunyà, variants al.lèliques que fossin capaces de metabolitzar, aprofitar i

enmagatzemar eficientment els recursos que el cos pogués aconseguir (és a dir, variants que podriem dir "estalviadores") haurien estat seleccionades a favor, mentre que variants al.lèliques "malbaratadores", que disminuïen l'efectivitat d'aprofitament dels recursos, devien ser severament penalitzades (Sharma 1998).

Però els temps han canviat i a les poblacions del primer món l'adquisició i consum d'aliments en quantitats industrials és relativament assequible i no cal fer grans esforços físics per aconseguir-los. En aquest nou ambient, ser estalviador ja no suposa un avantatge, si no tot el contrari, i caldria esperar que la selecció tendís a eliminar les variants "estalviadores" a favor de les "malbaratadores". La transició d'estil de vida, però, s'ha produït principalment després de la revolució industrial, ara fa uns 200 anys, un temps massa petit perquè la selecció hagi pogut actuar adequadament. Així, individus que són genèticament més susceptibles a tenir concentracions de metabòlits petites, a retenir el sodi o a mobilitzar ràpidament l'insulina front carbohidrats en aquest nou ambient ric estaran predisposats a l'obesitat, la hipertensió dependent de sal o a la diabetes no dependent d'insulina (NIDDM). Aquestes variants, a més, serien molt antigues i, per tant, es trobarien disperses per totes les poblacions humanes, així com a altres espècies de primats. Dos exemples de variants "estalviadores" són la variant ancestral de l'apoliproteïna E (APOE), l'al.lel 4 (Corbo and Scacchi 1999), o la variant ancestral treonina al codó 235 de l'angiotensinogen (Sharma 1998).

1.4 Les malalties genètiques

1.4.1 Malalties mendelianes

Les malalties mendelianes, també anomenades monogèniques o simples, deuen el seu fenotip patològic a mutacions produïdes en un únic gen que, en un fons genètic i ambiental normal, esdevenen necessàries i suficients per produir la malaltia; donat que només un gen està implicat en el fenotip, les malalties mendelianes s'hereten d'acord amb les lleis de l'herència de Mendel (Strachan and Read 1999) i presenten fenotips **discrets**. Les malalties mendelianes acostumen a tenir una incidència baixa i variable dependent tant de la malaltia com de la població (per exemple, la incidència de fibrosi quística a Europa és de 1 de cada ~2.500 naixements, però a Catalunya aquesta indicència és de 1 de cada ~4.500 naixements). El nombre de malalties mendelianes

descrites i el coneixement de la base biològica rera la patologia s'incrementa cada dia. La base de dades OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim/) recull informació sobre fenotips associats a caràcters mendelians, especialment aquells associats a processos patològics. Fent una cerca amb "(disease) OR (disorder) OR (syndrome) OR (mendelian)" limitada a aquelles entrades en les que es descriuen les variants al.lèliques, s'obtenen més de 1300 entrades (agost 2004), cosa que dóna una idea aproximada del coneixement actual de les malalties mendelianes.

L'acceptació científica de la causalitat d'una malaltia mendeliana per part d'una mutació ve determinada pels següents criteris (Inherited Metabolic Diseases; chapter 13: The nature and mechanisms of human gene mutation; Human Mutation 12:1-3 1998):

1) La mutació s'ha produït en una regió de funció o estructura coneguda
2) La mutació es troba en una regió conservada evolutivament
3) Es produeix en un pacient
4) No s'observa en una mostra gran d'individus sans (100 o més) o està present amb una freqüència per sota de 1%
5) La mutació i la malaltia cosegreguen en un llinatge
6) Es demostra que la proteïna sintetitzada *in vitro* presenta les mateixes característiques funcionals defectives que la proteïna sintetitzada *in vivo*.
7) Reversió del fenotip patològic en el pacient o en cèl.lules de cultiu per reemplaçament del gen amb la mutació o del seu producte proteic per la variant normal no mutada.

1.4.1.1 Gens i mutacions associades a malaltia mendeliana

Ni tots els canvis no sinònims en proteïnes tenen efectes patogènics ni tots els gens produeixen malaltia (vegeu figura 12). ¿Quines són les característiques, doncs, que fan que una mutació tingui més probabilitats de canviar la funció del gen i quines les característiques que fan que un gen sigui millor candidat a produir malaltia quan està mutat que un altre?

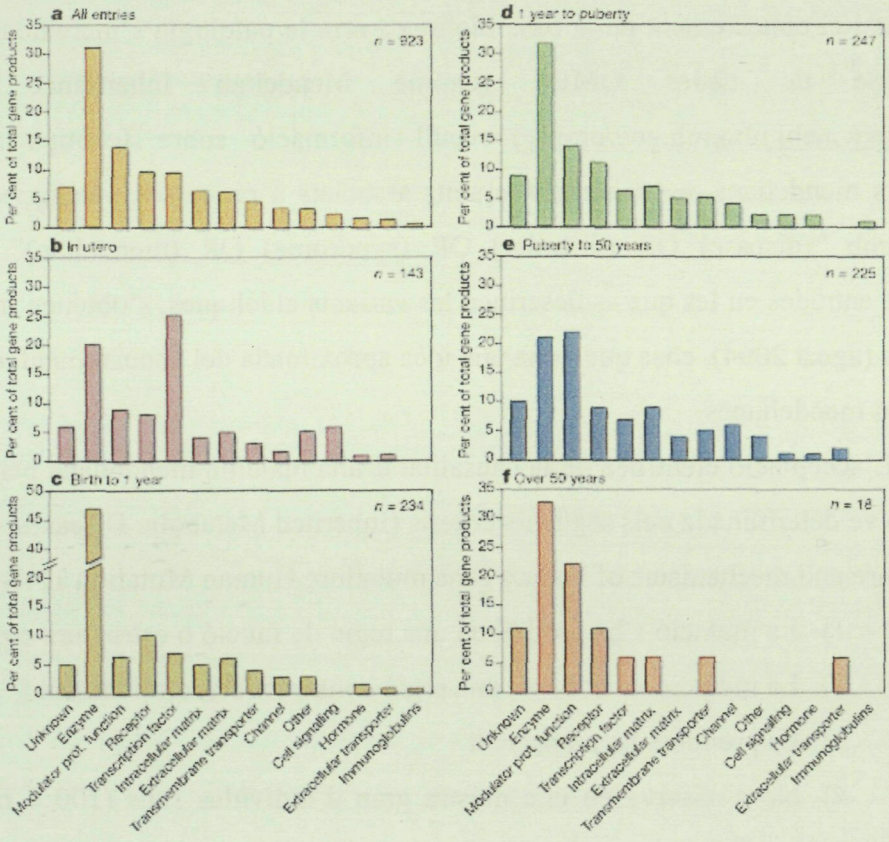


Figura 12. Gens associats a malalties mendelianes, classificats per funció i edat d'aparició. Un 31.2% de les proteïnes afectades són enzims. A partir de (Jimenez-Sanchez et al. 2001)

Les mutacions poden afectar la funció de la proteïna de maneres molt diverses. El mal plegament de la proteïna produït per mutacions en codons codificants, per exemple, és un procés que es creu força comú com a causant de malaltia genètica, ja que la maquinària cel.lular habitualment detecta i degrada les formes mal conformades (Waters 2001). D'altra banda, els canvis no sinònims que es produeixen en centres catalítics de la proteïna tenen més possibilitats de produir una disrupció en el funcionament de la proteïna, com s'observa comparant la freqüència de canvis no sinònims associats a malaltia respecte els no associats a malaltia en els centres catalítics (Stitziel et al. 2003) (vegeu figura 13).

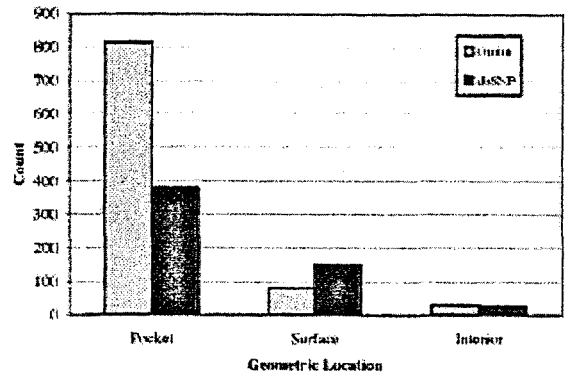


Figura 13. Distribució de les variants patogèniques respecte a les no patogèniques a diferent dominis proteics. A partir de (Stitziel et al. 2003)

Establir la proporció de nucleòtids que evolutivament han sofert canvis no sinònims (aquells en què el canvi de nucleòtid comporta canvi d'aminoàcid, Ka) respecte els sinònims (el canvi de nucleòtid no comporta canvi aminoacídic, Ks) és una forma de quantificar les constriccions evolutives que està patint una proteïna. Si les constriccions són fortes, de manera que petits canvis tenen efectes dràstics sobre la funcionalitat de la proteïna, llavors esperem que la proporció Ka/Ks sigui molt petita. Si, pel contrari, tot dos tipus són tolerats de la mateixa forma, aquesta proporció tindrà un valor proper a 1 i direm que la proteïna evoluciona de forma neutra. La fracció Ka/Ks acostuma a ser molt menor de 1 quan comparem, per exemple, proteïnes humanes amb les homòlogues de ximpanzé, on la mediana de la distribució és només de 0.11 (resultats no publicats de T. Marquès a partir de (Clark et al. 2003)); aquest valor, a més, depèn en gran mesura del teixit en el que s'està expressant la proteïna: gens que s'expressen a teixits com el cerebral o el muscular tenen uns quocients Ka/Ks més petits que els gens que s'expressen al teixit hepàtic o al pulmonar (Duret and Mouchiroud 2000), cosa que indica que les constriccions evolutives no són les mateixes en tots els teixits. D'acord amb aquesta hipòtesi, Winter i col.laboradors (Winter et al. 2004) han trobat una correlació positiva entre l'especificitat tissul.lar d'expressió i el percentatge de gens causants de malaltia genètica; també han trobat una correlació entre el percentatge de gens causants de malaltia i la fracció de gens que secreten el seu producte. Totes aquestes evidències semblen indicar que els gens causants de malalties estan més associats a determinats teixits (com l'hepàtic o el pulmonar) i que, principalment, són secretats. Els gens que són ubicus a tots els teixits ("housekeeping genes") i que evolucionen lentament també poden tenir variants no funcionals, però és

molt més probable que aquestes variants, degut a la seva funció essencial, produeixin letalitat embrionària i que, per tant, no observem cap fenotip patològic.

1.4.1.2 Patrons fenotípics de les malalties mendelianes

Depenent de si la mutació s'ha produït en un gen situat en un cromosoma autosòmic o sexual, les malalties mendelianes es classifiquen en autosòmiques o lligades als cromosomes sexuals (majoritàriament al X); segons si només cal una còpia d'un gen no funcional per produir-se la malaltia o si cal tenir les dues còpies, les malalties mendelianes es classifiquen en dominant o recessives. La combinació d'ambdós criteris de classificació dóna lloc a (Strachan and Read 1999):

1.4.1.2.1 Herència autosòmica dominant

Quan l'herència és autosòmica dominant, una persona afectada acostuma a tenir un dels progenitors afectats, els dos sexes tenen la mateixa probabilitat de patir la malaltia i el fill d'un individu afectat té un 50% de probabilitats de patir la malaltia; en el cas de malalties molt greus, que impedeixen que l'individu es pugui reproduir, la malaltia es produeix per mutacions *de novo*. Un exemple de malaltia mendeliana autosòmica dominant és la malaltia de Huntington (OMIM 143100), causada per explansions de triplets CAG en el gen huntingtina (que mapa a 4p16.3).

1.4.1.2.2 Herència autosòmica recessiva

En aquest cas, l'individu afectat, que pot ser de qualsevol sexe, acostuma a tenir pares no afectats, els quals acostumen a ser portadors en heterozigosi d'un al.lel mutat. La consanguinitat tendeix a incrementar la incidència de les malalties autosòmiques recessives. Els fills d'individus amb un fill afectat tenen el 25% de probabilitats de patir la malaltia. Malalties com la fibrosi quística, la fenilcetonúria o la β-talassèmia són malalties autosòmiques recessives (vegeu més endavant).

1.4.1.2.3 Herència recessiva lligada al cromosoma X

Els individus afectats acostumen a ser de sexe masculí. Els pares acostumen a ser asimptomàtics i la mare acostuma a ser portadora de l'al.lel causant de la malaltia;

no hi ha transmisió de la malaltia pare - fill masculí. Les filles poden estar afectades si el pare ja estava afectat i la mare era portadora o si, durant la inactivació del cromosoma X durant l'embriogènesi, s'inactiva principalment el cromosoma que no conté la mutació associada al fenotip patològic (donant lloc a fenòmens de mosaicisme). L'hemofília A (OMIM 306700), per exemple, és una malaltia recessiva lligada al cromosoma X deguda a mutacions en el gen que codifica pel factor de coagulació VIII (que mapa a la posició Xq28).

1.4.1.2.4 Herència dominant lligada al cromosoma X

La malaltia afecta tant a homes com a dones, però en les dones el fenotip acostuma a ser més suau que en els homes. La descendència d'una dona afectada té una probabilitat del 50% de patir la malaltia, independentment del seu sexe, mentre que en el cas d'un home afectat totes les filles estaran afectades però cap dels fills. L'hemofília B (306900), produïda per mutacions en el gen que codifica pel factor de coagulació IX (situat a Xq27.1-q27.2) és una malaltia mendeliana dominant lligada al cromosoma X.

Casos especials de malalties mendelianes no contemplats en la definició prèvia inclouen aquelles malalties que es produeixen per mutacions en gens associats al DNA mitocondrial o que impliquen fenòmens d'imprintació genètica (*imprinting*). En el cas de malalties associades a mtDNA, la transmissió de la malaltia es produeix només per línia materna, de manera que els pares afectats tindran fills no afectats; l'imprintació genètica és un procés que es produeix durant la gènesi dels gàmetes masculins i femenins i que consisteix en marcar epigenèticament un gen segons l'origen parental mitjançant la metilació, de forma que en l'embriogènesi de l'individu només s'expressa un únic al.lel (Falls et al. 1999). La manifestació fenotípica de mutacions en gens imprintats dependrà del patró d'imprintació que segueixi el gen (masculí o femení) i, per tant, de quin progenitor passi la mutació. Una característica de l'imprintació és que la malaltia "salta" entre generacions d'individus, depenent del sexe on vagi a parar la mutació, com en el cas del síndrome d'Angelman (OMIM 105830) o el síndrome de Prader-Willi (OMIM 176270).

1.4.1.3 Mutació, funció i malaltia mendeliana

Els canvis estructurals del DNA associats a malaltia mendeliana poden ser molt diversos, com es pot veure a la figura 3. Fins i tot mutacions en un únic nucleòtid poden tenir efectes dràstics sobre la funció de la proteïna: poden provocar l'activació de patrons de splicing alternatius, codificar per un aminoàcid diferent en un domini important per a la funció, o un codó STOP que trunqui prematurament la proteïna, afectar l'estabilitat del tràNscrit, tant a nivell de DNA missatger com de la proteïna (malgrat que el canvi semblés en principi inocu), o afectar la regulació de la proteïna. En el cas de petites delecions i insercions, la magnitud de l'efecte de la mutació dependrà en gran mesura de si generen canvis en la pauta de lectura o no: si les insercions i delecions són múltiples de 3, els efectes patològics tendiran a ser menys severs que si hi ha un canvi de la pauta de lectura. Delecions més grans poden incloure regions codificants per a un o més loci, cosa que fa que normalment els malalts presentin quadres clínics amb característiques de més d'una malaltia mendeliana.

Malgrat que aquesta disparitat estructural, les mutacions es poden classificar en dos grans grups depenent de com es produeix l'efecte fenotípic, per pèrdua de funció o per guany de funció. Els gens també poden perdre o guanyar funció per un funcionament inadequat dels diferents mecanismes que controlen l'expressió normal en la cèl.lula, com podrien ser canvis en la conformació estructural de la cromatina, i modificant els patrons d'imprintació.

1.4.1.3.1 Pèrdua de funció

Una mutació presenta pèrdua de funció quan la funció del producte gènic minva per efecte de la mutació; en el cas que aquesta estigui completament eliminada, llavors diem que l'al.lel és nul. El fenotip més habitual de les mutacions amb pèrdua de funció és el recessiu, ja que normalment amb la meitat de la dosi gènica la cèl.lula en té prou per poder funcionar. Ara bé, en casos en els que calgui més del 50% de la quantitat del producte original pel bon funcionament de la cèl.lula (efecte de dosi o haploinsuficiència) o en aquells casos on el producte de l'al.lel mutat interaccioni amb el producte de l'al.lel normal segrestant-lo (efectes dominants-negatius), també es podran obtenir fenotips dominants.

1.4.1.3.2 Guany de funció

Les mutacions de guany de funció són aquelles en les que el gen adquireix una funció nova. Les mutacions de guany de funció produeixen fenotips dominants mitjançant una elevada varietat de mecanismes: per l'adquisició de nous substrats, sobreexpressió, activació contínua del receptor, obertura inapropiada d'un canal iònic, multímers estructuralment anormals o gens quimèrics.

1.4.1.4 Estragègies generals per trobar gens causants de malalties mendelianes

La identificació d'un gen causant una malaltia genètica, en absència de coneixement del producte o la funció que el gen realitza, es fa mapant una regió que cosegregui amb la malaltia en els individus d'una mateixa família. Aquesta aproximació fa servir la covariació dels marcadors genètics informatius propers físicament a la variant causant de la malaltia a la descendència i la recombinació; com que la recombinació és un procés que té més possibilitats de donar-se entre marcadors que estan més allunyats que entre els que estan més propers, la variant causant de la malaltia estarà probablement més associada a un determinat context genètic. Per tant, analitzant com aquest context covaria amb la malaltia, hom pot delimitar regions concretes del genoma on hi ha més possibilitats que es trobi el gen causal. La probabilitat es mesura mitjançant l'estima de lod score, és adir, quantes vegades és més probable que el marcador genètic hagi cosegregat amb la variant patogènica amb un determinat nivell de recombinació respecte a la probabilitat que hagi covariat per atzar. Aquesta mesura es dóna en forma logarítmica i es considera que regions amb un lod score de 3 o superior (és a dir, que la probabilitat que la variant estigui associada sigui 1,000 vegades més gran que no ho estigui) són regions candidates per trobar el gen causal de la malaltia.

Una vegada es té acotada una determinada regió (vegeu figura 14) es poden utilitzar bases de dades per identificar gens candidats i passar a intentar trobar variants patogèniques que puguin estar causant la malaltia.

Aquesta aproximació ha estat espectacularment exitosa i ha permès identificar centenars de gens responsables de malalties mendelianes.

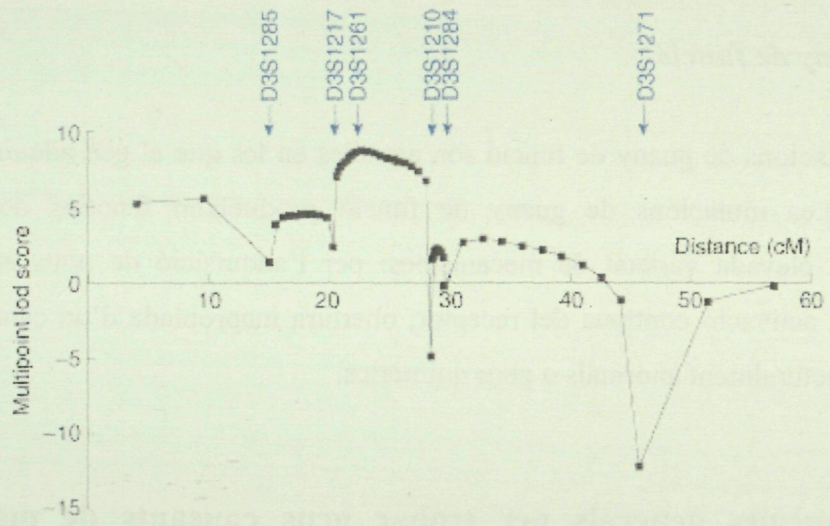


Figura 14. Lod score associat a marcadors distribuïts al llarg del cromosoma. El pic més gran mostra la regió més plaussible on mapa el gen. (Strachan and Read 1999)

1.4.1.5 Origen i manteniment de les malalties mendelianes a les poblacions humanes

1.4.1.5.1 Dinàmica de les malalties mendelianes

1.4.1.5.1.1 L'equilibri mutació-selecció

Un model relativament simple que permet explicar el manteniment de mutacions deletèries a les poblacions es troba representat a la figura 15. En aquest model, les mutacions deletèries es produeixen en una població infinita i panmíctica a partir del conjunt de cromosomes no mutats (sense possibilitat de reversió); els cromosomes deleteris s'eliminen de la població mitjançant l'acció de la selecció que, depenent de la intensitat, ho fa més o menys ràpidament (Hartl and Clark 1997).

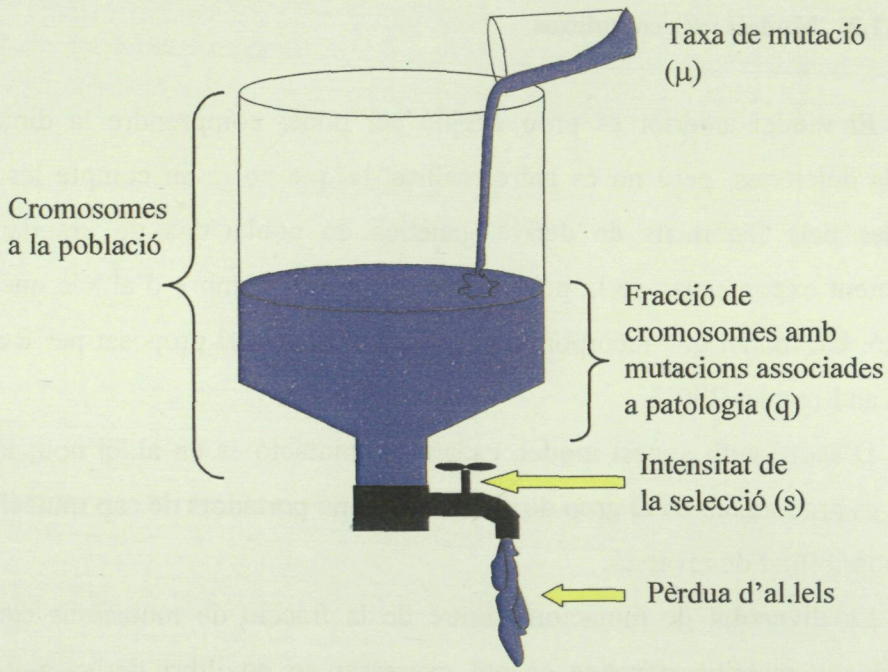


Figura 15. Esquema del model d'equilibri mutació-selecció

Aquest model prediu que la fracció d'al.lels mutats estarà en equilibri i dependrà de la taxa de mutació i de la intensitat de la selecció sobre els al.lels mutats. En situacions en les que l'al.lel tingui efecte dominant i impedeixi la reproducció de l'individu, la fracció d'al.lels mutats equivaldrà als al.lels introduïts cada generació per la taxa de mutació:

$$\hat{q} = \mu$$

Si, pel contrari, els al.lels són recessius i en homozigosi impedeixen la reproducció de l'individu, llavors el model prediu que, en equilibri, la fracció d'al.lels deleteris serà:

$$\hat{q} = \sqrt{\frac{\mu}{s}}$$

En situacions en les quals hi ha fenòmens de selecció a favor d'heterozigot (h>0), llavors:

$$\hat{q} = \frac{\mu}{hs}$$

1.4.1.5.1.2 Models més complexos

El model anterior és prou simple per poder comprendre la dinàmica de les variants deletèries, però no és gaire realista, ja que no té en compte les fluctuacions causades pels fenòmens de deriva genètica en poblacions de grandària finita, el creixement exponencial de la població humana o el nombre d'al.lels que produeix la mutació. Un model que incorpora aquestes variables és el proposat per Reich i Lander (Reich and Lander 2001).

D'acord amb aquest model, cada nova mutació és un al.lel nou; les mutacions només es produeixen en el grup de cromosomes no portadors de cap mutació prèvia i no hi ha possibilitat de reversió.

La diversitat de mutacions dintre de la fracció de mutacions causants d'una determinada malaltia genètica es pot expressar en equilibri deriva-mutació-selecció com:

$$\varphi = \frac{1}{1 + 4N\mu(1 - f_0)}$$

On φ és la probabilitat que prenent dos al.lels causants de malaltia, tots dos siguin iguals; φ depèn de la grandària de la població (N), de la taxa de mutació (μ) i de la fracció d'al.lels mutats (f). Un **patró de diversitat simple** és definit com la presència d'una mutació molt freqüent seguida de moltes de molt poc freqüents, i un **patró de diversitat complex** es defineix com la presència de moltes mutacions equifreqüents.

El model assumeix una població panmíctica prèvia a l'expansió dels humans de 10,000 individus i taxes de mutació de 3.2×10^{-6}. f0 és estimat a partir de les equacions clàssiques de l'equilibri mutació-selecció (apartat anterior).Amb aquests valors, $f_0 \approx 1$ i φ val aproximadament 1 abans de l'expansió. El posterior creixement exponencial de la població fa que el valor de φ s'allunyi de l'equilibri. Segons aquest model, la dinàmica que segueix φ depèn principalment de com varia en el temps la fracció d'al.lels causants de malaltia presents inicialment a la població i, per tant, de μ i de s. El model prediu que malalties amb una f inicial alta tindran un patró de diversitat simple, dominat per una única mutació molt freqüent, mentre que malalties amb una f petita (el que equival a tenir una taxa de reemplaçament molt alta) tindran un patró complex, amb moltes mutacions equifreqüents (vegeu figura 16).

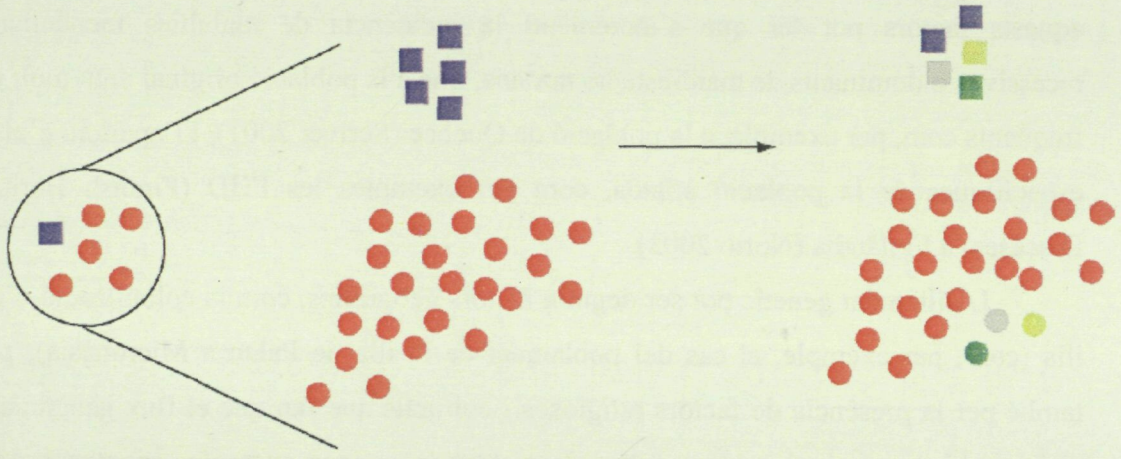


Figura 16. Dinàmica de les mutacions sota expansió. Els quadrats blaus representen al.lels altament deleteris, mentre que els cercles vermells representen al.lels no seleccionats deletèriament. Després de l'expansió, els quadrats blaus són més ràpidament reemplaçats per l'efecte de la selecció que els cercles vermells, que tenen una diversitat molt més petita (Smith and Lusis 2002)

Com els mateixos Reich i Lander apunten, malgrat que el model introdueix nous paràmetres i dóna una explicació plausible als patrons de diversitat genètica que observem a les malalties genètiques, es realitzen algunes assumpcions que probablement no són certes, com, per exemple: (i) absència d'estructuració a la població humana inicial i en les poblacions post-expansió, (ii) f només depèn de μ i s, sense tenir-se en compte la deriva (iii) la selecció roman constant en el temps, però sabem que per algunes malalties s'han produït fenòmens de selecció a favor de l'heterozigot (per exemple, moltes hemoglobinopaties), (iv) no totes les mutacions d'un mateix gen presenten el mateix coeficient de selecció: algunes es troben associades a un fenotip més sever mentre que altres es troben associades a un fenotip més benigne.

1.4.1.5.2 L'efecte fundador i les malalties mendelianes

Com ja hem vist, la creació de noves poblacions per part d'un petit nombre d'individus té com a conseqüència que variants que a la població general són rares, com les associades a les malalties mendelianes, puguin incrementar la seva freqüència i que d'altres que són freqüents a la població general no es trobin representades. Donat el reduït nombre d'individus, un posterior aïllament continuat afavorirà els processos de deriva genètica i tendirà a augmentar la consanguinitat (i per tant el nombre

d'homozigosts) malgrat que els aparellaments siguin a l'atzar. La combinació de tots aquests factors pot fer que s'incrementi la incidència de malalties mendelianes, recessives o dominants de manifestació tardana, que a la població original eren molt poc freqüents com, per exemple, a la població de Quebec (Scriver 2001), i l'aparició d'altres específiques de la població aïllada, com per exemples les FHD (Finnish Heritage Diseases) a Finlàndia (Norio 2003).

L'aïllament genètic pot ser degut a factors geogràfics, com la colonització d'una illa (com, per exemple, el cas del poblament de la illa de Palau a Micronèsia), però també per la presència de factors religiosos i culturals que fan que el flux genètic amb altres poblacions sigui molt reduït o inexistent (com, per exemple, en el cas de la població Amish). Aquest aïllament, fins i tot, es pot donar entre individus que tradicionalment són identificats com una única població, però que s'organitzen en grups jeràrquics, com el sistema de castes a la Índia.

A la taula 1 es numeren algunes de les poblacions considerades com a genèticament aïllades i els motius d'aïllament.

Taula 1. Algunes poblacions tradicionalment considerades com aïllats genètics i la causa de l'aïllament; a partir de (Arcos-Burgos and Muenke 2002); (Kalaydjieva et al. 2001); (Scriver 2001)

Població	Causa plausible d'aïllament
Finesos	Efecte fundador, aïllament cultural i geogràfic
Ordre Amish	Efecte fundador i aïllament cultural
Huterites	Efecte fundador i aïllament cultural
Sards	Efecte fundador, aïllament cultural i geogràfic
Jueus (diferents comunitats)	Patrons complexos d'efectes fundadors, aïllament geogràfic i cultural
Bascos	Efecte fundador, aïllament cultural i geogràfic
Gitanos	Patrons complexos d'efectes fundadors, aïllament geogràfic i cultural
Quebequencs	Efecte fundador continuat, aïllament cultural i geogràfic
Illa Palau	Efecte fundador, aïllament cultural i geogràfic

1.4.1.5.3 Mutació deletèria i selecció a favor d'heterozigot

Molt probablement, les malalties infeccioses han conformat la demografia de la població humana durant els últims milers d'anys. Epidèmies com la pesta, per exemple, varen assolar Europa durant l'edat mitjana, eliminant fins un 50% de la població total en alguns llocs. La malària (vegeu figura 17), per un altre costat, ja era coneguda pels antics metges grecs i romans i encara avui suposa un greu problema sanitari en molts països, ja que afecta uns 500 milions de persones i en mata uns 2 milions cada any.

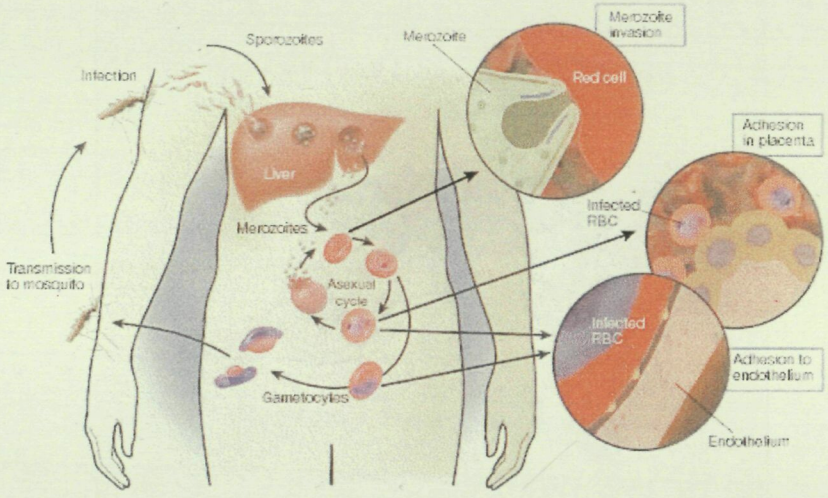


Figura 17. Cicle biològic del paràsit causant de malària. A partir de (Miller et al. 2002)

En front a aquesta alta taxa de mortalitat, és d'esperar que variants que hagin conferit algun tipus de resistència a la infecció o a la virulència del patogen hagin estat seleccionades a favor, malgrat que en homozigosi puguin ser deletèries. En moltes de les hemoglobinopaties com l'anèmia falciforme o l'α-talassèmia (molt probablement també la β-talassèmia) s'ha observat que la distribució geogràfica se solapa amb la distribució geogràfica que històricament ha tingut la malària (vegeu figura 18); en homozigosi aquestes variants patològiques tendeixen a ser letals per als individus portadors, però en heterozigosi confereixen resistència a la infecció per malària; aquest és un cas de selecció equilibradora i permet que variants que són deletèries, com les causants de les diverses hemoglobinopaties, es puguin mantenir a la població (Dean et al. 2002). En l'actualitat, la malària ha desaparegut de molts llocs on era endèmica com a conseqüència de les polítiques d'eradicació del mosquit vector de la malària i conversió del seu hàbitat. En aquestes regions encara es troben casos d'hemoglobinopaties, però és d'esperar que amb el temps la seva incidència disminuirà fins als nivells esperats per mutació-deriva-selecció.

Altres casos menys evidents de selecció equilibradora inclouen la fibrosi quística, que podria estar relacionada amb la resistència a patogens entèrics (vegeu més endavant) o la fenilcetonúria, que podria proporcionar resistència a les micotoxines (vegeu més endavant).

Figura 18. Solapament de les regions on hi ha malaria (en blau) i les regions on hi han hemoglobinopaties (zones en ombra) (Jobling et al. 2004)

1.4.1.6 Les malalties mendelianes no són tan simples

Si hom analitza la relació entre el genotip i el fenotip en una malaltia monogènica descubrirà que la definició de malaltia mendeliana que féiem a l'inici de l'apartat és, en el millor dels casos, una simplificació útil i, en el pitjor dels casos, una falàcia benintencionada. Primer, perquè no tots els casos clínics estan produïts per mutacions en el mateix gen (genocòpies) i segon perquè no tots els individus amb mutacions al mateix gen presenten el mateix fenotip (fenocòpies). A la taula 2 es poden observar alguns dels nombrosos exemples que es troben a la literatura científica (per exemple, a (Wolf 1997)) de discrepàncies entre un genotip donat i el fenotip esperat, que inclouen des de diferències en la gravetat de la malaltia fins a l'absència del fenotip patològic.

Taula 2. Alguns exemples de discordància entre el genotip i el fenotip (extrets a partir de (Wolf 1997)). A la taula s'indica el gen, el producte gènic, els fenotips discordants i on s'han trobat les discrepàncies

Gen (OMIM)	producte	fenotip	Parentesc
CFTR(602421)	Cystic fibrosis transmembrane conductance regulator	a) Fibrosi quística amb discordància de funció pancreàtica i pulmonar	Germans
		b) Home fèrtil o amb absència congènita de	Dos germans

		conducte deferent	
HBB(141900)	Hemoglobina β	Anèmia falciforme en diversos graus de gravetat	Diferents pacients
OTC(300461)	Ornitin transcarbamilasa	Variació en l'edat d'aparició i desenvolupament de la malaltia	5 famílies diferents
PAH(261600)	Fenilalanina hidroxilasa	Hiperfenilalaninèmia, inteligència normal o retard greu (independent de dieta)	3 germans
PRNP(176640)	Proteïna priònica	Insomni familiar fatal - malaltia Creutzfeld-Jakob	30 membres afectats de 11 famílies

Aquesta falta d'entesa entre el genotip i el fenotip ha provocat que s'hagin hagut de definir nous termes, com la penetrància o l'expressivitat, per englobar les cada cop més nombroses excepcions a la regla general. La penetrància es defineix com la probabilitat que un individu, essent portador del genotip, expressi el fenotip patològic. Es diu que un genotip que tant pot produir com no el fenotip patològic té penetrància incompleta. L'expressivitat fa referència al fet que el mateix genotip molecular pugui donar lloc a diferències tant en la gravetat del fenotip clínic com a la presentació clínica (Zlotogora 2003). Els motius que expliquen aquesta falta de correlació entre el genotip observat i el fenotip esperat poden ser molt variats, i inclouen (Zlotogora 2003):

(i) Penetrància pseudo-incompleta / pseudo expressivitat: l'absència del fenotip patològic es deu a la falta de precisió del clínic, que no detecta la malaltia quan, de fet hi és.

(ii) Premutació: La penetrància incompleta es pot observar en casos en els que hi hagi estadis intermedis de la mutació sense expressió clínica.

(iii) Anticipació: es produeix quan l'edat d'aparició de la malaltia es redueix dels pares afectats als fills afectats. Molecularment, aquest fenòmen acostuma a estar associat amb l'expansió anormal de triplets de nucleótids patogènics, com a les atàxies espinocerebel.lars (SCAs), on l'edat d'aparició de la malaltia es correlaciona amb el nombre de tractes repetitius (Schols et al. 2004); vegeu figura figura 19)

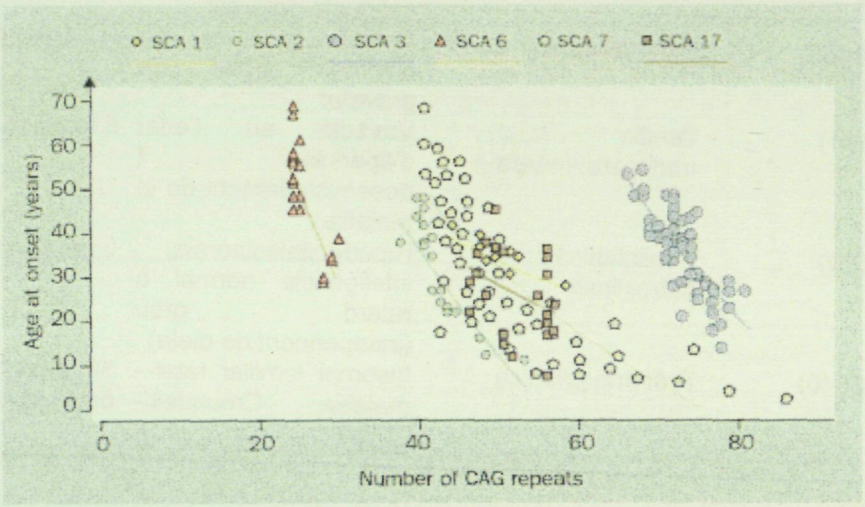


Figura 19. Efecte de la llargada del tracte repetitiu en l'edat d'aparició de la malaltia. A partir de (Schols et al. 2004)

(iv) Influència d'al.lels en *trans* en el mateix gen

(v) Influència d'al.lels en *cis* en el mateix gen

(vi) Imprintatge genètic

(vii) Herència digènica: quan per produir-se la malaltia calen mutacions a dos gens per separat

(viii) Gens modificadors: els gens modificadors són aquells que, degut a la presència d'una mutació/polimorfisme, modifiquen l'expressió clínica de la malaltia.

(ix) Mosaicisme somàtic: La presència d'una mutació a la línia somàtica i germinal acostuma a produir un fenotip més lleu en l'individu que en la seva descendència. El mosaicisme somàtic també es pot produir durant la inactivació d'un dels cromosomes X a les dones durant l'embriogènesi.

(x) Factors ambientals (vegeu figura 20): en moltes malalties genètiques metabòliques, la presentació clínica pot dependre de la quantitat del substrat de la dieta. En d'altres, l'estatus socioeconòmic de l'individu afectat pot influir sobre la gravetat de la malaltia i les probabilitats de supervivència.

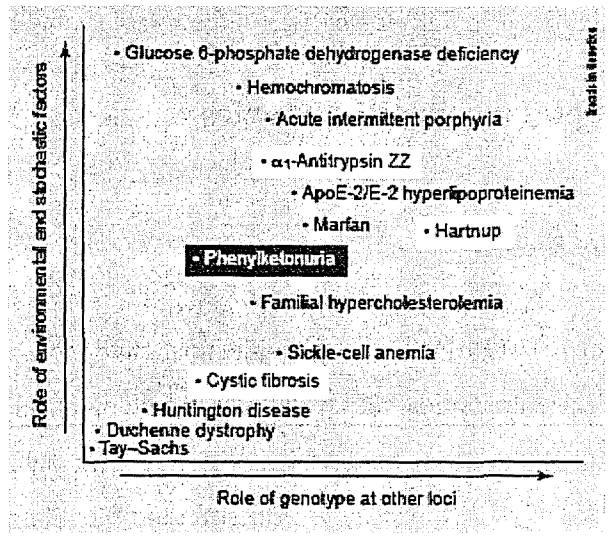


Figura 20. Relació entre el genotip i els factors ambientals en el fenotip. (Scriver and Waters 1999)

(xi) Efectes estocàstics durant l'embriogènesi

La distinció entre malaltia monogènica, mendeliana o simple, i malaltia complexa esdevé cada vegada més tènue a mesura que el coneixement sobre el fenotip i el genotip de les malalties mendelianes augmenta. Cada vegada sembla més clar que el context genètic (modificadors genètics, genocòpies...) i ambiental en el que apareixen les mutacions són fonamentals per a la definició precisa del fenotip (la malaltia). El descobriment que algunes malalties mendelianes són, de fet, malalties oligogèniques, causades per l'acció modulada d'un petit nombre de gens (Badano and Katsanis 2002), o la creació de models que assumeixen una continuïtat fenotípica en les malalties mendelianes (tradicionalment descrites com a discretes), amb uns llindars per sobre dels quals es desenvolupa la malaltia (Dipple and McCabe 2000b), ens donen una idea del canvi en el paradigma de l'estudi i comprensió de les malalties mendelianes i les malalties complexes. Ambdues representen els extrems d'un continu, on en un extrem trobem l'efecte d'un únic gen causal influenciat per modificadors i l'ambient, típic de les malalties mendelianes, i en l'altre múltiples gens de baix efecte i l'ambient, típic de les malalties complexes (Dipple and McCabe 2000a).

1.4.1.7 Tres exemples de malalties simples

1.4.1.7.1 La fibrosi quística

La fibrosi quística (OMIM 219700), amb una incidència de 1 de cada ~2.500 nounats, és la malaltia autosòmica recessiva més freqüent a les poblacions d'origen europeu. La fibrosi quística deu principalment el seu fenotip patològic a mutacions al gen CFTR (de l'anglés Cystic Fibrosis Transmembrane Regulator (OMIM 602421)), també conegut com ABCC7 (OMIM 602421), que es troba al braç llarg del cromosoma 7 (7q31.2), té unes 250 kb de llargada i comprèn 27 exons.

El gen CFTR és membre de la superfamília de transportadors de membrana ATP-binding cassette (ABC) i codifica per una proteïna de 1.480 aminoàcids amb cinc dominis diferents: dos dominis transmembrana, dos dominis intracel.lulars hidrofílics d'unió a nucleòtids i un domini regulador intracel.lular. El producte del gen CFTR funciona com un canal de clor, i també com a regulador d'altres canals iònics a la membrana apical de les cèl.lules epitelials.

La manifestació fenotípica clàssica de la fibrosi quística afecta el tracte respiratori, on es presenta com a obstrucció crònica pulmonar, i el pàncrees, on produeix una insuficiència pancreàtica exocrina. L'acumulació de moc viscós a les vies respiratòries afavoreix el creixement de microorganismes i fa que es produeixin infeccions recurrents, causades majoritàriament per *Pseudomonas aeruginosa* i *Staphylococcys aureus*; la insuficiència pancreàtica exocrina provoca una deficiència d'enzims pancreàtics que fa que la digestió de greixos i proteïnes sigui ineficient, i es produeixin quadres clínics de distensió abdominal i esteatorrea. Altres trets associats a la fibrosi quística inclouen la presència d'una elevada concentració de clor a la suor, així com una elevada incidència d'infertilitat masculina per absència o atròfia dels conductes deferents, l'epidídim i les vesícules seminals, i símptomes menys comuns com el *meconium ileus* (MI) del nounat o obstrucció intestinal que es dóna al naixement, diabetis o disfunció pancreàtica endocrina que cursa amb pancreatitits i alteracions al fetge que poden acabar en cirrosi. Malgrat que els avenços en el tractament de la fibrosi quística, els individus acostumen a morir a edats molt

primerenques. El tractament amb curcumina sembla revertir els efectes de la mutació F508del en ratolins i podria utilitzar-se com a teràpia en humans (Egan et al. 2004).

Es coneixen més de 1000 mutacions associades a la fibrosi quística (http://www.genet.sickkids.on.ca/cftr/), però en gairebé un 70% dels al.lels analitzats només s'ha trobat la mutació F508del (Estivill et al. 1997), una deleció de tres parells de bases que elimina una fenilalanina a l'exó 10 de la proteïna. Altres mutacions freqüents (entre 1 i 2.5%) són: G542X, G551D, N1303K i W1282X. Les mutacions es classifiquen en sis grups diferents, depenent del seu efecte molecular: classe I, síntesi proteica defectuosa; classe II, defecte en el processament proteic; classe III, alteracions en la regulació del canal de clor; classe IV, conductància del canal de clor disminuïda; classe V, reducció en la síntesi o el transport proteic; i classe VI, disminució en l'estabilitat proteica.

La fibrosi quística mostra una elevada heterogeneitat fenotípica (Nabholz and von Overbeck 2004), fins i tot entre germans amb mutacions al gen CFTR idèntiques, com ja s'apuntava a la taula 2. Diverses causes poden explicar aquesta disparitat, com es pot veure a la figura 21; el rerafons genètic i ambiental esdevé crucial per a la determinació del fenotip, malgrat que es consideri una malaltia monogènica.

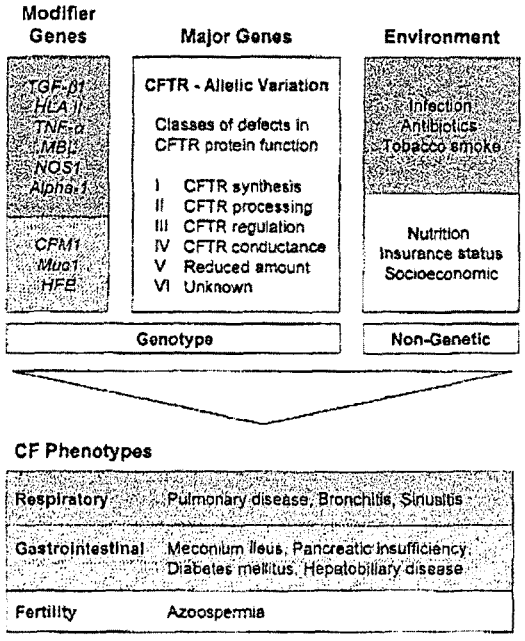


Figura 21. Malgrat que les mutacions al gen CFTR són essencials pel fenotip, aquest depèn també d'altres factors genètics i ambientals. A partir de (Nabholz and von Overbeck 2004)

Sembla poc probable que l'elevada incidència de fibrosi quística al continent europeu es pugui explicar només per fenòmens de deriva genètica (Bertranpetit and Calafell 1996). En els últims anys s'ha postulat un possible avantatge selectiu dels heterozigots relacionat amb la resistència a malalties infeccioses com el còlera (Gabriel et al. 1994), la febre tifoidea (Pier et al. 1996), la tuberculosi (Meindl 1987), la sífilis (Hollander 1982) o la grip (Shier 1979). De fet, Mateu i col.laboradors no varen trobar els haplotips associats a les mutacions causants de fibrosi quística en poblacions europees actuals, cosa que indicaria que l'origen de la fibrosi quística és molt antic (Mateu et al. 2002).

1.4.1.7.2 La fenilcetonúria

La fenilcetonúria (OMIM 261600), amb una incidència de 1 de cada ~10.000 nounats, és l'error del metabolisme autosòmic recessiu més freqüent en poblacions d'origen europeu. La fenilcetonúria és un tipus d'hiperfenilalaninèmia, però no totes les hiperfenilalaninèmies estan associades a fenotips patològics. Les causes genètiques que produeixen hiperfenilalaninèmies són variades i inclouen mutacions al gen PAH (de l'anglés Phenylalanine Hydroxylase), així com en els gens implicats en la síntesi i regeneració del cofactor tetrahidrobiopterina (BH_4). La detecció dels nivells de BH_4 ens permet distingir entre una hiperfenilalaninèmia produïda per mutacions a PAH d'aquella produïda per mutacions a la via metabòlica de BH_4. Aproximadament el 99% de les mutacions causants de fenilcetonúria es troben associades al gen PAH (Erlandsen and Stevens 1999). En alguns casos, mutacions al gen PAH poden produir uns nivells elevats de fenilalanina sense implicar els símptomes clàssics de la fenilcetonúria; en aquests casos es parla d'hiperfenilalaninèmia no-PKU. La distinció entre ambdos tipus es realitza mitjançant els nivells en sang de l'aminoàcid fenilalanina, considerant-se PKU en el cas de nivells més alts de 1000μM (els nivells de fenilalanina en sang sota condicions fisiològiques normals en adults acostumen a trobar-se en 58±15 μM) i una tolerància menor a la fenilalanina de la dieta (<500mg/dia).

El gen PAH es troba al cromosoma 12 (12q24.1), té unes 90kb de llargada, conté 13 exons i codifica per un enzim d'uns 50 KDa que està implicat en la conversió de l'aminoàcid L-fenilalanina cap a L-tirosina mitjançant l'ús del cofactor BH_4. PAH s'expressa únicament al fetge i es troba en un equilibri d'homotetràmers i homodímers.

S'han descrit més de 480 mutacions a la base de dades de la fenilcetonúria (http://www.pahdb.mcgill.ca/) en el gen PAH associades a fenilcetonúria i hiperfenilalaninèmia no-PKU. Algunes, com la mutació R408W, s'han trobat associades a diferents haplotips en diferents regions geogràfiques, cosa que indica fenòmens de mutació recurrent (Byck et al. 1994),(Eisensmith et al. 1995). Les mutacions es classifiquen estructuralment en cinc categories diferents: categoria I, residus actius, categoria II, residus estructurals, categoria III, interaccions entre dominis en el monòmer, categoria IV, residus que interactuen amb la seqüència N-terminal autoreguladora i, categoria V, residus al lloc d'interacció del tetràmer.

El fenotip clàssic de la fenilcetonúria és un retard mental profund, però presenta una elevada heterogeneitat fenotípica (mesurada en nivells de fenilalanina en sang o estimes del coeficient intel.lectual) fins i tot entre germans amb el mateix genotip (vegeu taula 2) (Scriver and Waters 1999) que evidencien la complexitat del coeficient intel.lectual i la seva dependència de molts factors genètics i ambientals; no totes les mutacions en el gen PAH tenen el mateix efecte deleteri sobre la proteïna. Depenent del fenotip al qual s'asocïin majoritàriament s'acostumen a classificar en greus, moderades/suaus i hiperfenilalaninèmiques sense efectes fenotipics observables; aquest fet pot explicar, per combinacions de les diferents variants, l'elevat espectre fenotípic que s'observa en els pacients. La dieta és un altre factor que també juga un paper important sobre els efectes neurològics, ja que variacions en el consum continuat de fenilalanina, especialment durant les primeres etapes de la vida, poden influir sobre els nivells de retard mental. De fet, una teràpia efectiva contra la fenilcetonúria clàssica ha estat la prescripció de dietes baixes en l'aminoàcid fenilalanina a individus que en el cribatge metabòlic donen positiu; una malaltia associada a la hiperfenilalaninèmia, s'hagi desenvolupat o no PKU és l'hiperfenilalaninèmia materna, que pot ser la causa d'embriopaties i fetopaties, incloent-hi microcefàlia i retard mental en un 80% dels fetus exposats.

Sembla que l'elevada presència de la fenilcetonúria a les poblacions europees no es pot explicar únicament per efecte de la deriva genètica (Krawczak and Zschocke 2003) i diverses hipòtesis han apuntat a favor d'una possible avantatge dels heterozigots front a micotoxines (Saugstad 1973; Woolf et al. 1975; Saugstad 1977).

1.4.1.7.3 β-Talassèmia

La β-globina és una proteïna formada per dos parells de cadenes peptídiques essencial per al transport de l'oxigen. Sis tipus diferents de cadenes peptídiques (α, β, γ, δ, ε i ζ) es troben en les hemoglobines normals expressades en diferent etapes del desenvolupament humà. En els adults, més del 97% de l'hemoglobina està formada per un parell de cadenes peptídiques α i un altre parell de cadenes peptídiques β, i s'anomena HbA. Mutacions a les diferents cadenes poden donar lloc a diferents tipus d'hemoglobinopaties, com l'anèmia falciforme. Les talassèmies són un conjunt heterogeni d'hemoglobinopaties caracteritzades per una producció reduïda o absent d'una o més de les cadenes de l'hemoglobina (Scriver et al. 2000). En el cas de la β-talassèmia, la cadena afectada és la de la β-globina, cosa que fa que la proporció de les cadenes de la molècula de l'hemoglobina estigui descompensada cap a un major nombre de cadenes α. La conseqüència directa d'aquesta hemoglobina defectiva és la presència d'anèmia en diversos graus, acompanyada per diferents processos fisiològics compensatoris com un desenvolupament de la melsa o una progressiva hepatoesplegnomegàlia. El fenotip de la β-talassèmia és molt heterogeni; depenent del grau d'afectació les β-talassèmies es classifiquen en **major**, si el fenotip és sever i l'individu requereix transfussions de sang i **minor**, si el fenotip és d'una anèmia lleu o asimptomàtica. Aquesta elevada variabilitat es troba relacionada amb el nivell residual de síntesi de la cadena β així com del tipus de mutació (Weatherall 2001). Les mutacions causants de β-talassèmia es classifiquen en β^0 i β^+, depenent si l'expressió de la cadena està completament abolida o si encara hi ha una expressió residual.

La distribució de la incidència de β-talassèmia no és homogènia a totes les poblacions humanes, trobant-se principalment restringida a poblacions mediterrànies, del Pròxim Orient i del sud i sud-est d'Àsia; l'anàlisi de la distribució de les mutacions causants de β-talassèmia revel.la que l'origen de la malaltia no ha estat únic a les poblacions humanes, ja que les mutacions més freqüents causants de β-talassèmia es troben delimitades en diferents regions (vegeu figura 22; (Weatherall 2001)). La distribució espacial de la malaltia s'ha relacionat amb una possible resistència dels

heterozigots a l'infecció de malària, cosa que suposaria una selecció equilibradora (Hartl and Clark 1997).

Figura 22. Distribució de les principals mutacions associades a β-talassèmia. A partir de (Weatherall 2001)

1.4.2 Malalties complexes

Moltes de les diferències fenotípiques que observem entre individus, incloent-hi la morfologia, la psicologia, el creixement i el comportament, no segueixen pas els patrons discrets de les característiques mendelianes. Aquests caràcters es defineixen mitjançant variables **quantitatives** que s'hereten sense seguir una herència típica mendeliana (tot i que els gens associats sí). Les malalties complexes es poden definir com a variables binàries (l'individu té o no té la malaltia) produïdes quan se supera un cert llindar d'un fenotip quantitatiu. Les malalties complexes són d'especial interés per a la biomedicina perquè inclouen moltes de les malalties més freqüents a les poblacions humanes, com poden ser les malalties psiquiàtriques, cardiovasculars, la diabetis, l'Alzheimer o diferents tipus de càncers. En alguns casos el fenotip pot estar produït per una única mutació (per exemple, mutacions en el gen presenilina (OMIM 104311) poden donar lloc a Alzheimer) però representen una petita fracció del total, i en més del 90% dels casos la malaltia acostuma a ser causada per un elevat nombre de factors, ambientals i genètics, que interaccionen donant lloc al fenotip patològic. Les malalties

complexes acostumen a presentar una elevada heterogeneitat genètica, cosa que implica que la malaltia pot estar relacionada amb un elevat nombre de gens diferents, així com per una elevada heterogeneitat fenotípica i, per tant, que en moltes malalties complexes no es pugui identificar inequívocament els individus en funció de la presència o absència de la malaltia (Hauser and Pericak-Vance 2000). Estimar la penetrància esdevé gaire bé impossible, pel fet que les malalties complexes depenen de l'efecte de gens potencials i de la presència o absència de factors ambientals. En aquest context no es pot parlar pas de "la mutació" causant, ja que cap mutació no acostuma a ser ni suficient ni necessària per produir la malaltia, i sí d'al.lels d'associació; tanmateix, la magnitud del seu efecte dependrà de l'ambient en el que es trobi i, per tant, caldrà tenir-lo en compte (Cooper 2003) a l'hora de definir els al.lels de susceptibilitat. Per regla general, l'edat de la presentació de la malaltia acostuma a estar relacionada amb una major preponderància dels factors genètics o dels factors ambientals, i edats de presentació més primerenques acostumen a associar-se amb un paper més important dels factors genètics. Les estimes del percentatge de variació explicada per factors genètics (l'heretabilitat) en bessons depén del tipus de malaltia, però suggereixen que una part important dels factors implicats en el desenvolupament de la malaltia són genètics (vegeu taula)

Taula 3. Heretabilitat de caràcters i malalties complexes a partir d'estudis de bessons (MacGregor et al. 2000)

Caràcter - Malaltia complexa	Heretabilitat (%)
Asma	60
Pressió sanguínia	40-70
Densitat mineral òssia	60-80
Degeneració dels discos cervical i lumbar	60-80
Diabetis depenent d'insulina	70
Obesitat	50-90
Osteoartritis	50-70
Artritis reumatoide	60
Colitis ulcerativa	50

Poc es coneix del model poblacional que regeix la freqüència de les variants genètiques que es troben associades a les malalties complexes i, en general, de les variants que determinen els caràcters complexos. Entre el conjunt de models possibles que defineixen l'espai al.lèlic de les variants associades a malalties complexes, la hipòtesi "variant comuna/malaltia comuna" (Common Variant/Common Disease; CV/CD) i la hipòtesi de "variants rares" (o model d'heterogeneitat) representen els dos possibles extrems.

La hipòtesi CV/CD prediu que les variants associades a les malalties complexes són antigues i freqüents a les poblacions humanes (>1%) (Chakravarti 1999). Aquesta hipòtesi assumeix un model additiu, on els polimorfismes tenen un baix pes en el desenvolupament de la malaltia, i el seu efecte sobre el risc de patir la malaltia es va acumulant fins a superar un cert llindar, moment en el que apareix el fenotip patològic (Harpending and Rogers 2000) (vegeu figura 23). Variants al.lèliques associades a malalties com l'al.lel 4 de l'APOE o la variant Pro12Ala del locus PPARγ serien exemples d'aquesta hipòtesi. Una variant d'aquesta hipòtesi es la hipòtesi "variant comuna/multiples malalties" que postula que, a més, aquestes variants comunes estaran implicades en un elevat nombre de malalties diferents (Becker 2004). La hipòtesi de les variants poc freqüents, per la seva banda, assumeix l'existència de polimorfismes a molt baixa freqüència (<1%), segurament sota dinàmica mutació-selecció, que tenen un efecte molt marcat sobre el desenvolupament de la malaltia (vegeu figura 23).

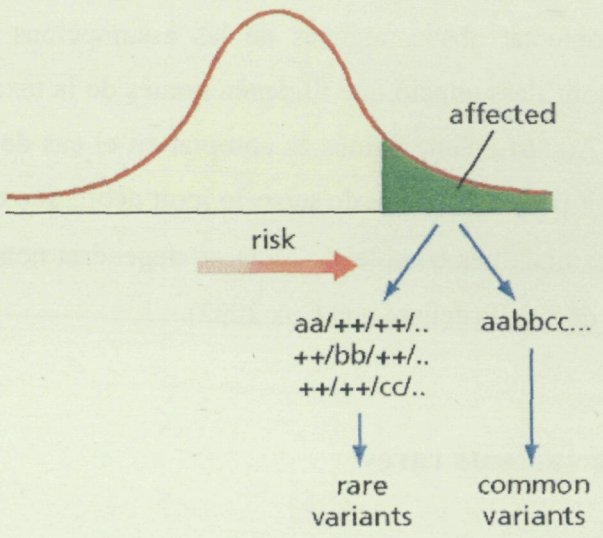


Figura 23. La hipòtesi CV/CD contra la hipòtesi de les variants rares (a partir de (Chakravarti 1999)

Saber a quina de les dues s'acosta més la realitat no és pas una pregunta trivial, perquè l'abordatge metodològic de les variants associades a malalties complexes implica estratègies molt diferents, ja que en el segon cas s'hauran d'aplicar tècniques més exaustives de seqüenciació per trobar els polimorfismes presents a cada individu.

Dos models en genètica de poblacions prediuen quina serà la distribució dels al.lels associats a malalties complexes. Paradoxalment, l'aplicació de cada model produeix conclusions completament diferents, cosa que indica que les conclusions depenen principalment de les assumpcions que es fan quan es crea el model (Goldstein and Chikhi 2002).

1.4.2.1 El model CV/CD

El model de Reich i Lander que hem comentat a l'apartat de malalties mendelianes també serveix per predir el patró al.lèlic que hom espera trobar a les malalties complexes. D'acord amb aquest model, i assumint que f_0 depèn de la taxa de mutació i del coeficient de selecció, es demostra que si el coeficient de selecció no és massa gran, com es podria esperar d'al.lels que afecten a l'individu a edats més grans de l'edat d'aparellament (però veieu més endavant), els patrons que s'observen són simples (una mutació molt freqüent i altres de molt poc freqüents). Dit d'una altra manera, els valors esperats de φ (és a dir, d'identitat al.lèlica) en el grup d'al.lels associats a la malaltia són elevats si la freqüència dels al.lels no és massa petita (f0>0.9%).

Com ja s'ha comentat abans, algunes de les assumpcions que el model fa podrien no ser certes, com l'assumpció que f0 depèn només de la taxa de mutació i del coeficient de selecció. Aquest balanç només és apropiat en el cas de gens en equilibri amb una freqüència molt petita. En el cas de selecció molt dèbil, que es podria donar en els al.lels associats a les malalties complexes, f0 ja no dependria només de μ i s, si no d'altres factors, com la deriva (Pritchard and Cox 2002).

1.4.2.2 El model de variants rares

Aquest model va ser desenvolupat per Jonathan K. Pritchard (Pritchard 2001). Assumeix una població panmíctica i constant, amb una grandària efectiva de ~10,000

individus. També assumeix dos tipus d'al.lels diferents, normals i associats a malaltia complexa, amb dues taxes diferents de mutació: d'al.lel normal cap a al.lel associat a malaltia complexa i a l'inversa, essent la primera taxa de mutació molt més elevada que la segona. Finalment, el model també assumeix que la variança genètica d'una malaltia complexa és additiva i que a nivell multilocus l'efecte de cada locus és multiplicatiu.

Segons aquest model, Pritchard troba que loci amb selecció purificadora dèbil i unes taxes de mutació elevades tenen més probabilitats de ser genèticament variables i, per tant, de contribuir més a la variància de la susceptibilitat de la malaltia i, per tant, donaria suport a les variants poc freqüents com agents associats a les malalties complexes.

Com el mateix Pritchard comenta, el seu model fa simplificacions i assumpcions que, si bé són plausibles, podrien molt bé no ser certes. La grandària efectiva de la població, per exemple, no ha estat constant a la població humana i les taxes de mutació poden arribar a ser molt variables entre els gens.

1.4.3 Determinació de variants associades a malalties complexes

La metodologia emprada per analitzar i trobar gens associats a fenotips complexes és molt àmplia (Khoury et al. 1993). A grans trets, els mètodes per identificar gens de susceptibilitat es poden agrupar en dos grups: els estudis de gens candidats i els estudis de cribatge genòmic (*genomic screens*) (Hauser and Pericak-Vance 2000).

En el primer cas, es parteix de variants presents en un gen candidat i s'estudia la seva associació amb la variació fenotípica. En el segon cas es parteix d'un nombre de marcadors polimòrfics (microsatèl.lits i SNPs) que es distribueixen per tot el genoma i s'intenta trobar associacions entre la variació d'un determinat marcador situat en una determinada localització amb la variació fenotípica. Una associació positiva pot indicar que el marcador està associat a la malaltia o bé que està proper a un locus implicat en el desenvolupament de la malaltia.

A nivell analític, dos grups generals de mètodes es poden aplicar tant a l'anàlisi dels gens candidats com dels cribatges genòmics: l'anàlisi d'associació i l'anàlisi de lligament.

L'anàlisi d'associació intenta identificar polimorfismes o al.lels que incrementin el risc a patir la malaltia mitjançat l'estudi de la covariància entre dos determinats al.lels o genotips i el fenotip de la malaltia. Mentre que l'aproximació directa estudia polimorfismes que puguin tenir efecte patogènic, l'aproximació indirecta cerca trobar variants que estiguin en desequilibri de lligament amb la variant associada a la malaltia (Brookes 1999). Les associacions positives poden interpretar-se com que la variant analitzada està implicada en el desenvolupament de la malaltia, o bé que està en desequilibri de lligament amb la variant associada a la malaltia. Els mètodes d'associació es poden fer servir tant amb variants qualitatives (presència o absència de la malaltia) com amb mesures quantitatives, i són especialment indicats amb mostres poblacionals (com és el cas dels estudis cas-control), cosa que els fa molt atractius per als investigadors, ja que es facilita l'obtenció de moltes mostres a l'hora que es disminueix el cost (Carlson et al. 2004).

En el cas de l'anàlisi de lligament (vegeu apartat malalties mendelianes), l'objectiu és trobar regions genòmiques més compartides entre els individus afectats que entre els individus no afectes de la mateixa família (Carlson et al. 2004). Les anàlisis de lligaments són molt útils en el cas de malalties causades per al.lels d'elevat efecte, però tenen menys potència per a detectar al.lels de baix risc (Risch and Merikangas 1996). Tanmateix, les anàlisis de lligament permeten trobar regions de lligament en un determinat cromosoma, però no acostumen a tenir gaire resolució a l'hora de trobar el locus o loci implicats.

1.4.3.1 Els estudis de casos i controls

Els estudis de casos i controls s'engloben dintre dels estudis d'associació. S'estima com de més freqüent és una determinada variant al.lèlica o genotípica entre dos grups d'individus que difereixen per un determinat fenotip, com podria ser tenir (casos) o no tenir (controls) la malaltia.

Normalment els estudis de casos i controls es poden representar mitjançant una taula de contingència com la que es presenta a la taula.

Taula 4. Taula de casos i controls. Els genotips es distribueixen a cada casella segons si són casos (individus afectes) o controls (individus no afectats per la malaltia)

	Genotip A	Genotip B
Casos	α	γ
Controls	δ	β

Diferents estimes es poden emprar per calcular el pes que té l'al.lel en el desenvolupament de la malaltia. Una de les més utilitzades és l'estima d'*odds ratio*, que és defineix com la raó de la proporció d'individus malalts respecte als indivius sans sota un determinat genotip respecte la mateixa proporció en l'altre genotip:

$$OR = \frac{\frac{\alpha}{\delta}}{\frac{\gamma}{\beta}} = \frac{\alpha \times \beta}{\delta \times \gamma}$$

El rang de valors que l'estadístic OR pot prendre va des de 0 fins a infinit; si l'estadístic OR és igual a 1 o aquest s'inclou a l'intèrval de confiança de l'estadístic, es diu que el genotip A no està associat a la malaltia. Si l'estadístic OR<1, llavors el genotip A és protector i si OR>1, llavors està associat a la malaltia.

1.4.3.2 SNPs, desequilibri de lligament i el projecte Hap Map

Els SNPs són marcadors bial.lèlics produïts per la substitució d'un únic nucleòtid; es troben densament distribuïts per tot el genoma (podria haver-hi més de 10.000.000 de SNPs en el genoma de tots els humans (Brookes 1999)), tant en regions codificants (cSNPs) com en regions intergèniques i en regions reguladores (rSNPs), cosa que fa que siguin idonis tant per realitzar estudis d'associació directa com indirecta. A diferència d'altres marcadors emprats en l'estudi de les malalties complexes, com els microsatèl.lits, els SNPs són molt freqüents, i el fet que només presentin dos estats els fa especialment interessants per a l'automatització del genotipat, per poder definir haplotips i per el càlcul de paràmetres estadístics com el desequilibri de lligament.

Els haplotips es defineixen com una seqüència de SNPs contigus. Cada nova mutació apareix en un determinada combinació al.lèlica. Posteriorment, a mesura que

passen les generacions, el nou SNP es pot associar a altres haplotips per successives recombinacions durant la meiosi; el nombre de recombinacions que es produeix depèn tant del sexe (les dones tenen en promig molts més events de recombinació que els homes), la regió cromosòmica (regions properes als centròmers són menys propenses a patir fenòmens de recombinació) i de l'individu (Lynn et al. 2004). Una mesura estadística de la covariació de SNPs veïns a nivell poblacional és el desequilibri de lligament (LD). Diferents paràmetres han estat definits per mesurar el LD (vegeu, per exemple, (Collins et al. 2001)) però, en qualsevol cas, com més elevat és el desequilibri de lligament, més covarien les variants al.lèliques dels dos SNPs i, per tant, més fàcil és predir l'estat al.lèlic d'un dels dos SNPs a partir de l'estat de l'altre; dit d'una altra manera: donat que les variants patològiques apareixen en un determinat haplotip, no caldria cercar el SNP putatiu associat a la malaltia, si no trobar SNPs en LD amb el SNP d'interès (per associació indirecta). Aquesta aproximació permetria una reducció substancial de la quantitat de genotipatge que caldria fer sense perdre informació: només caldria definir SNPs que abastessin la majoria de la variació genòmica (TagSNPs).

El desequilibri de lligament és un paràmetre poblacional i, com a tal, factors com la deriva genètica, el creixement de la població, la migració i la mescla, o la selecció natural poden influir sobre les estimes finals de LD. A aquests factors, a més, cal afegir-hi factors típicament genòmics, com diferents taxes de recombinació, taxes de mutació variables o conversió gènica. Tot això fa que, malgrat que existeix una certa tendència a la disminució del LD amb la distància física (la recombinació és més probable entre marcadors que estan físicament molt allunyats que entre marcadors que estan molt propers), es puguin trobar SNPs separats a una gran distància amb un alt desequilibri de lligament (Ardlie et al. 2002). Recentment, alguns autors han postulat la presència de blocs de LD (vegeu, per exemple, (Cardon and Abecasis 2003) per una revisió) distribuïts per tot el genoma. El LD seria elevat entre els marcadors d'un mateix bloc de desequilibri i baix entre marcadors de diferents blocs, cosa que indicaria la presència de punts calents de recombinació (o "hot spots"). Si realment existeixen els blocs de LD, el nombre de Tag SNPs necessaris per cobrir tot el genoma es reduiria considerablement. Ara bé, malgrat que diferents estudis han trobat blocs de LD, els llindars dels blocs depenen sovint de les freqüències dels SNPs i del mètode de definició (que és subjectiu) dels blocs. Més encara, malgrat que segurament existeixen punts calents en el genoma pels events de recombinació (Kauppi et al. 2004; Lynn et al. 2004), la presència d'altres

factors fa que en simulacions es puguin trobar blocs de LD fins i tot assumint una taxa de recombinació homogènia i constant per tot el genoma (Ardlie et al. 2002).

En aquest context apareix el projecte internacional HapMap, que té com objectiu trobar SNPs comuns (>1%) i caracteritzar les seves freqüències en poblacions d'Àfrica (individus yoruba de Nigèria), Àsia (individus han de Xina i japonesos) i Europa (individus d'Utah amb origen al nord i nord-oest d'Europa) , així com determinar els haplotips comuns que permetin seleccionar els tag SNPs (2003). El projecte HapMap porta implícita l'acceptació del model CV/CD, és a dir, que les variants associades a les malalties complexes són freqüents i comunes a totes les poblacions humanes, ja que si no fos així, la potència per trobar les variants rares associades als fenotips patològics serà petita. D'altra banda, el projecte HapMap també assumeix que les regions de desequilibri de lligament són comunes a totes les poblacions humanes i, per tant, és possible definir uns tag SNPs universals. Si aquest no fos el cas, caldria definir tag SNPs específics per cada població, cosa que incrementaria considerablement la quantitat de genotipatges que caldria fer.

1.4.3.3 La complexitat de les malalties complexes

L'ús dels mètodes de lligament, que en les malalties mendelianes ha estat molt exitós, no ha estat de gaire utilitat per detectar les variants associades a les malalties complexes, on és d'esperar que tinguin un baix pes en el desenvolupament de la malaltia i on els factors ambientals juguin un paper important. Això ha fet que molts investigadors hagin canviat d'enfoc cap als estudis d'associació, els quals són molt més idonis per trobar aquest tipus de variants (Risch and Merikangas 1996). En l'actualitat s'han publicats més de 600 estudis amb associacions positives entre un determinat al.lel i una determinada malaltia, però no sempre el resultat inicial positiu s'ha pogut replicar en posteriors estudis d'associació (Hirschhorn et al. 2002), malgrat que existeix un excés d'estudis replicats respecte el que hom esperaria per atzar (Lohmueller et al. 2003). Més important encara és el fet que quan s'incrementa el nombre d'individus analitzats en meta-anàlisis s'observa una caiguda progressiva de les primeres estimes de risc, realitzades amb pocs individus, cap a valors de risc molt més modestos, i a l'inrevés: obtenció d'associacions positives en estudis posteriors als primers estudis, que no trobaren cap associació (Ioannidis et al. 2001); vegeu figura 24).

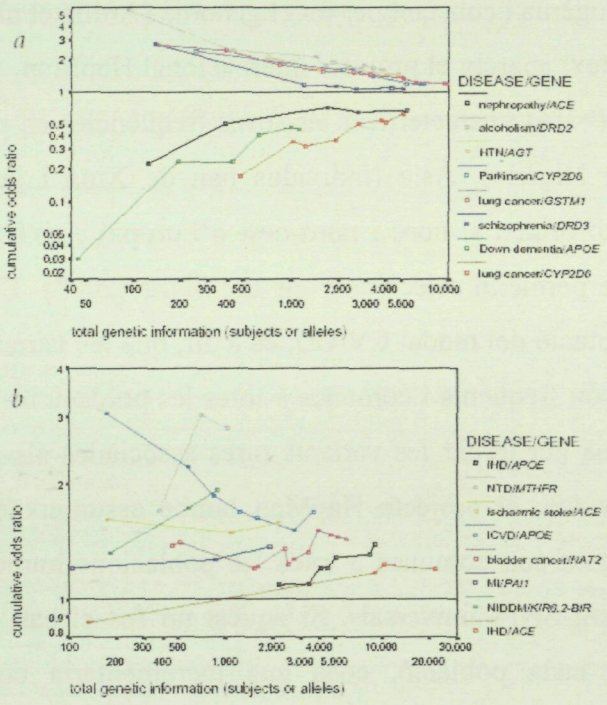


Figura 24. *Odds ratio* acumulatius en estudis en els quals es trobava una forta associació inicial (a) i en aquells on l' associació inicial no era estadísticament significativa (b) A partir de (Ioannidis et al. 2001)

Les raons per explicar aquesta disparitat són diverses i apunten tant a la natura i limitacions pròpies dels estudis d'associació com als processos editorials de publicació:

i. Biaix editorial a favor de publicar primer associacions positives i elevades. Estudis posteriors, normalment amb més individus i millor dissenyats, no són capaços de replicar els resultats inicials.

ii. Estratificació de la població d'estudi: en els estudis d'associació per casos i controls, es poden obtenir associacions espúries si els casos i els controls pertanyen a poblacions diferents que tenen freqüències gèniques del polimorfisme diferent.

iii. Desequilibri de lligament. Donat que el LD no depèn només de la taxa de recombinació, factors com la història de poblacions poden fer que en determinades poblacions un determinat marcador pugui estar en LD amb la variant associada a la malaltia (siguin o no contigus a nivell genòmic) mentre que en altres poblacions no.

iv. Interaccions entre gens i ambients, que poden ser diferents entre diferents poblacions.

v. Definició fenotípica dels individus afectats i sans
vi. Efecte dèbil de la variant i falta de poder estadístic: la capacitat per poder detectar una variant de baix efecte depèn del nombre d'individus analitzats.

Diferents recomanacions generals s'han proposat per tal de minimitzar l'efecte d'aquests factors i determinar aquelles variants al.lèliques que s'associïn a la malaltia de forma consistent (Lohmueller et al. 2003) i (Campbell and Rudan 2002):

i. Interpretació més conservadora dels resultats: les associacions positives s'han de veure com a indicatives fins que altres estudis hagin replicat de forma independent el mateix resultat
ii. Incrementar el nombre d'individus analitzats (>1.000 individus). Algunes simulacions demostren que la replicabilitat dels resultats és petita quan els estudis són de 100-500 casos.
iii. Incorporar en els estudis meta-anàlisis d'estudis anteriors per tal d'incrementar la potència de l'estudi (malgrat que això no exclou possibles biaixos (vegeu (Winkelmann et al. 2000))
iv. Consistència biològica: les variants estan associades a gens que estan implicats en processos fisiològics importants per al desenvolupament de la malaltia o bé s'expressen principalment en els teixits afectats

1.4.3.4 Un exemple de malaltia complexa : la malaltia coronària

La malaltia coronària (CHD), també anomenada malaltia de les artèries coronàries i malaltia ateroscleròtica del cor, es defineix com un tipus de malaltia del cor produïda per la reducció paulatina per acumulació lipídica del diàmetre de les artèries que l'envolten i li aporten oxigen i nutrients. L'arteriosclerosi és un procés progressiu que s'inicia a edats tempranes i/o es veu facilitada per la presència de lesions cròniques a l'endoteli de les artèries (Scheuner 2003). La ruptura o fissura de la placa fibrosa en estadis avançats de la malaltia pot provocar una hemorràgia a la placa, trombosi i, secundàriament, oclusió de l'artèria, cosa que produeix isquèmia del teixit irrigat i mort cel.lular i pot acabar en un atac de cor (vegeu figura 25).

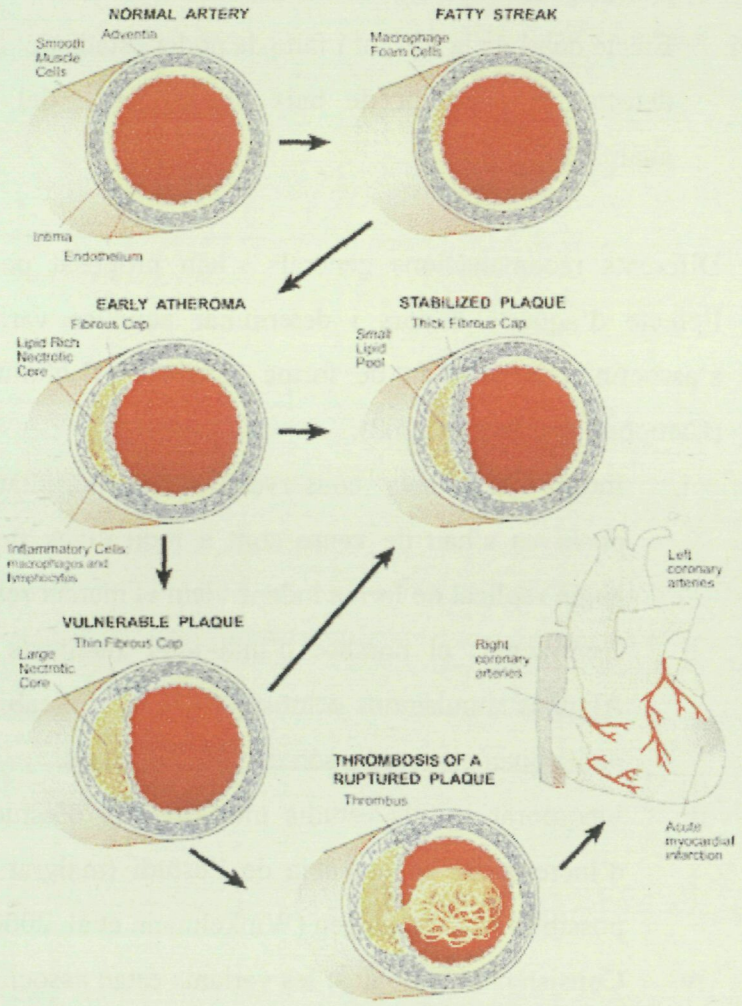
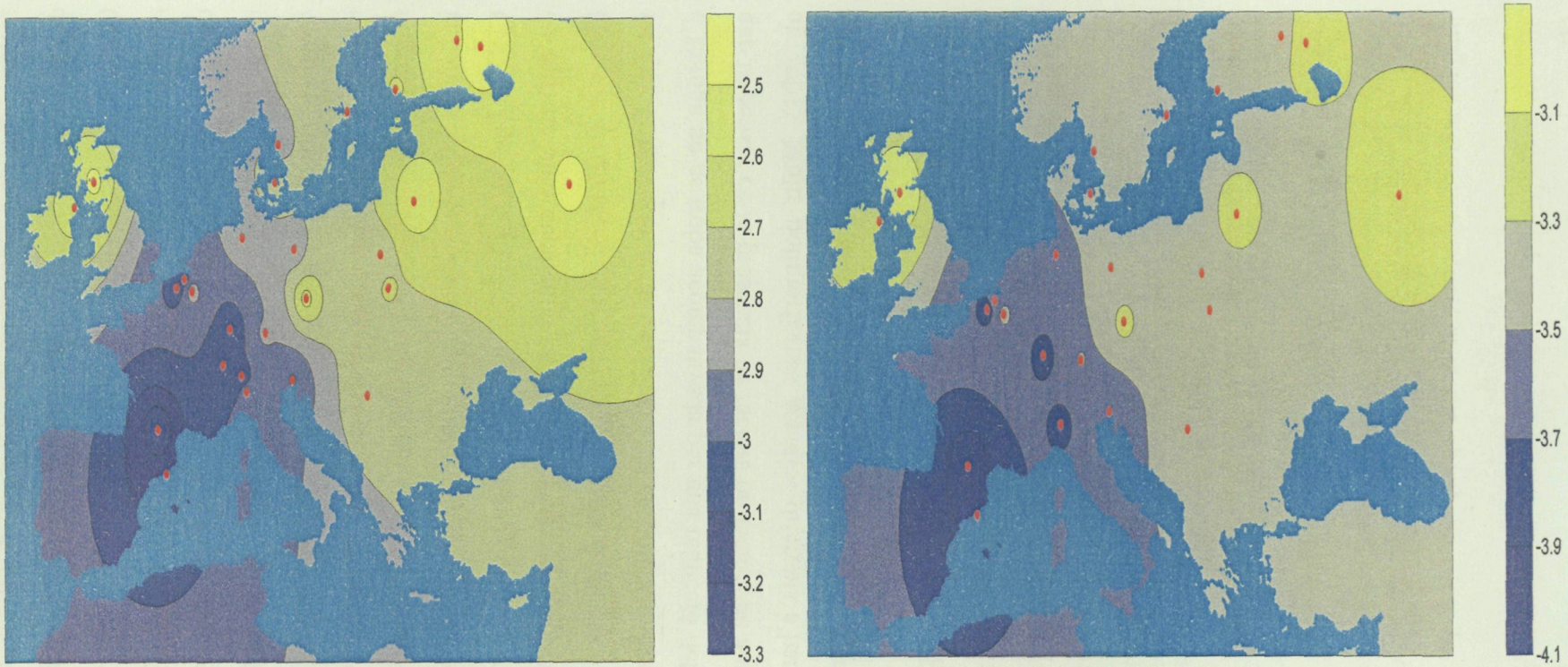


Figura 25. Progressió de les plaques ateroscleròtiques i obturació de les artèries coronàries. (Lusis et al. 2004)

Les malalties cardiovasculars, i en especial la malaltia coronària, són la primera causa de mortalitat en els països desenvolupats. Malgrat que afecta ambdos sexes, el risc de patir-la no és el mateix: el risc acumulatiu per a CHD en homes és del 35% a l'edat de 70 anys i puja fins al 49% a l'edat de 90 anys, mentre que en dones és del 24% i 32% a les edats de 70 i 90 anys respectivament (Scheuner 2003). La incidència de la malatia no és homogènia a les poblacions europees, essent molt més elevada en el nord d'Europa que en el sud (vegeu figura 26).



a) b)

Figura 26. Distribució geogràfica de la mortalitat per CHD en escala logarítmica a poblacions Europees en homes a) i dones b). Les poblacions amb una mortalitat deguda a CHD més baixa tenen un valor més negatiu del logaritme que les poblacions amb una mortalitat més elevada. A partir de (Tunstall-Pedoe et al. 1999)

Tot i que existeixen famílies on mutacions en un únic gen fan que el caràcter es transmeti de forma mendeliana amb una alta penetrança (Nabel 2003), aquestes només representen un petit percentatge del total. Com la malaltia complexa que és, molts factors genètics i ambientals intervenen en el desenvolupament de CHD. Per a una determinada població, l'heretabilitat de l'infart de miocardi s'estima entre ~25% i ~60% (Lusis et al. 2004). Normalment, aquells casos en els que la malaltia apareix abans acostumen a tenir un component genètic més penetrant que els casos en els que es desenvolupa de forma més tardana, ja que en el segon cas els factors ambientals han tingut més temps per actuar (Scheuner 2003). El consum de tabac, la dieta, l'activitat física, la diabetis o l'obesitat són alguns dels factors ambientals de risc que s'han associat al desenvolupament de la malaltia (Haapanen-Niemi et al. 1999). Els estudis de migracions de poblacions mostren que l'ambient explica una gran part de la variació de la incidència de la malaltia entre poblacions (per exemple, malgrat que la incidència de CHD al Japó és molt menor que als Estats Units, els japonesos americans que adopten un estil de vida occidental tenen una incidència semblant a la dels altres americans; (Lusis et al. 2004)). L'anomenada **paradoxa francesa o mediterrània** fa referència al fet que la reducció en la incidència en la població francesa i en poblacions del sud d'Europa no es troba acompanyada per una reducció dels factors de risc clàssics, cosa que fa pensar en la presència de factors protectors en aquestes poblacions (Jamrozik et al. 2001). Molts polimorfismes genètics diferents s'han associat a la malaltia, però, igual que amb d'altres malalties complexes, la reproducibilitat dels resultats en successius estudis ha estat molt controvertida. La cerca de gens s'ha centrat principalment en la formació i progressió de les plaques ateroscleròtiques. Aquest procés implica un elevat nombre de diferents mecanismes bioquímics (cadascun dels quals es troba, a la vegada, influït per diversos factors ambientals) que inclouen el metabolisme lipídic i de les apolipoproteïnes, l'oxidació lipídica, la resposta inflamatòria, funció endotelial, trombosi, fibrinòlisi, metabolisme de l'homocisteïna, sensibilitat a la insulina i regulació de la pressió sanguínea (Scheuner 2003). A continuació es fa un breu resum d'alguns d'aquests gens:

1.4.3.4.1 Metabolisme lipídic

Les lipoproteïnes juguen un important paper en el desenvolupament de les plaques ateroscleròtiques; diferents apolipoproteïnes presents en la seva superfície són

les encarregades de determinar el destí metabòlic de les lipoproteïnesn (vegeu figura 27).

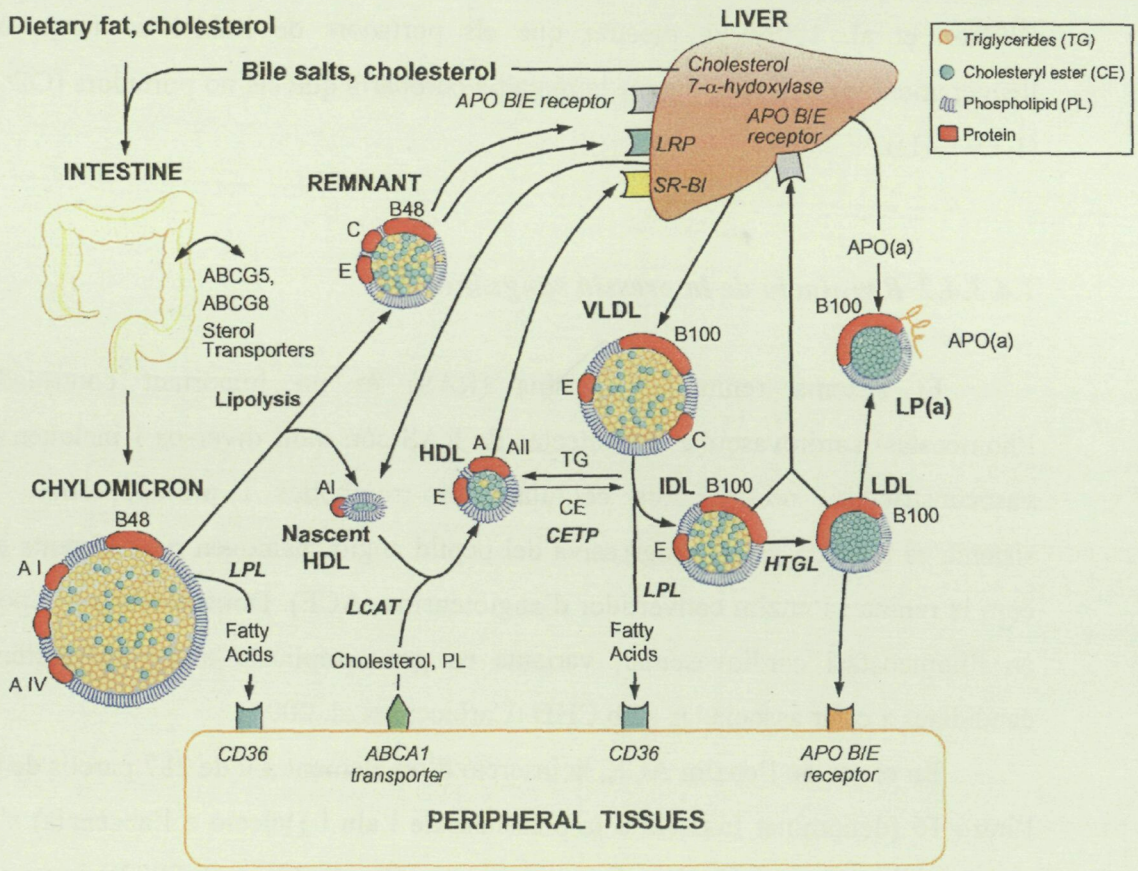


Figura 27. Metabolisme lipídic i paper de les lipoproteïnes (Lusis et al. 2004)

L'apolipoproteïna E (APOE) és una proteïna de 255 aminoàcids que se sintetitza en el fetge. A les poblacions humanes l'APOE és polimòrfica, amb tres al.lels isoelèctrics diferents anomenats ε2, ε3, ε4. La capacitat per relacionar-se amb diferents lipoproteïnes no és la mateixa depenent dels al.lels. Així, mentre que l'al.lel ε4 fa que l'APOE s'uneixi preferentment a lipoproteïnes riques en triglicèrids (com les de molt baix pes molecular o VLDL), les variants ε2 i ε3 s'uneixen més a HLD.

L'anàlisi comparativa entre humà i ximpanzé va demostrar que la variant ε4, que s'ha associat consistentment amb la predisposició a l'Alzheimer, és l'ancestral (Fullerton et al. 2000); aquest resultat, unit a la distribució geogràfica de la variant en poblacions de tot el món, ha fet postular a Corbo i col.laboradors (Corbo and Scacchi 1999) que la variant ε4 podria ser una variant "estalviadora" (vegeu apartat de la hipòtesi de les variants estalviadores). En el cas de CHD, diverses associacions i el fet

que la distribució de l'al.lel ε4 covaria amb la incidència de la malaltia en poblacions europees (Stengard et al. 1998) van suggerir que aquesta variant podria estar implicada amb CHD; una meta-anàlisi realitzada amb els estudis fets amb les variants d'APOE (Wilson et al. 1996) va mostrar que els portadors de l'al.lel ε4 tenen un risc lleugerament més elevat de patir la malaltia coronària que els no portadors (OD = 1.26 (1.13-1.41)).

1.4.3.4.2 Regulació de la pressió sanguínea

El sistema renina-angiotensina (RAS) és un important controlador de l'homeostasi cardiovascular. Els efectes de RAS són molt diversos i inclouen efectes vasoconstrictors, proliferadors cel.lulars, pro-trombòtics i pro-apoptòtics. Aquest sistema es basa en la lisi progressiva del pèptid angiotensinogen per diferents enzims, com la renina i l'enzim convertidor d'angiotensina (ACE). Donada la seva importància en l'homeostasi cardiovascular, variants en gens implicats en aquest sistema són candidates a estar associades amb CHD (Carluccio et al. 2001).

En el cas de l'enzim ACE, la inserció d'un element alu de 287 parells de bases a l'intró 16 (denominat Inserció a la presència de l'alu i Deleció a l'absència) s'ha vist que és polimòrfic a poblacions humanes (vegeu (Romualdi et al. 2002)) i que es troba associat als nivells de l'enzim en sang (explicant fins un 45% de la variabilitat total dels nivells d'ACE en sèrum), amb una concentració més elevada en el cas d'individus DD. Una meta-anàlisi realitzada amb diferents estudis publicats sobre el tema mostren que la presència de l'al.lel D en homozigosi incrementa lleugerament el risc a patir malaltia coronària (OD = 1.21) (Agerholm-Larsen et al. 2000).

El polimorfisme M235T del gen angiotensinogen (AGT) també s'ha associat a un major risc a patir malaltia coronària en el cas d'homozigots TT (OD = 1.22) (Kato et al. 1999); el polimorfisme de susceptibilitat podria ser també una variant estalviadora (vegeu apartat variants estalviadores), ja que és la variant ancestral i s'ha trobat associada a pressions selectives que podrien explicar-se amb la hipòtesi de la retenció de sodi, segons la qual les poblacions de regions tropicals i temperades tindrien una diferent disponibilitat de sodi (Nakajima et al. 2004).

1.4.3.4.3 Metabolisme de l'homocisteïna

Nivells alts de l'aminoàcid homocisteïna han estat identificats com a factor de risc per patir CHD. La concentració d'homocisteïna depèn de factors com l'edat i el sexe i s'incrementa per factors relacionats a l'estil de vida com el consum de cafè, de tabac i d'alcohol i disminueix amb suplementacions de folat. L'enzim 5,10-α–metiltetrahidrofolat (MTHFR) treballa en una de les vies metabòliques de l'homocisteïna, remetilant l'aminoàcid cap a metionina (De Bree et al. 2002). S'han descrit diversos polimorfismes codificants (com el polimorfisme 677 C>T que dóna lloc al canvi d'aminoàcid d'alanina a la possició 222 cap a valina) que són freqüents a la població i que podrien estar implicats en la regulació dels nivells d'homocisteïna en sang i, per tant, amb CHD i altres patologies associades a nivells alts d'homocisteïna (com, per exemple, defectes en el tancament del tub neural (Botto and Yang 2000)). De fet, els nivells d'homocisteïna en individus 677 TT són un 25% més elevats que en individus 677 CC, cosa que donaria suport a un paper del polimorfisme en el desenvolupament de CHD. Una meta-anàlisi (Kim and Becker 2003) realitzada amb diferents estudis d'associació comparant individus 677 TT amb 677 CC mostrà un OD de 1.21, indicant que aquesta variant confereix un petit increment en el risc de patir la malaltia.

1.4.3.4.4 Trombosi

L'aparició de ferides vasculars activa la coagulació, un procés dinàmic que es troba equilibrat per l'acció de factors pro-trombòtics i anti-trombòtics. Polimorfismes que afavoreixin l'activació dels primers o inhibeixin l'acció dels segons podrien ser factors de risc a patir trombosi i CHD. Aquest podria ser el cas del gen del factor V (FV). La inactivació del factor protrombòtic V es realitza mitjançant la proteòlisi per part de la proteïna C activada (APC) en tres posicions diferents (posicions Arg306, Arg506 i Arg679); el canvi d'aminoàcid arginina de la posició 506 cap a glutamina (també anomenat mutació FV_{Leiden}) fa que aquesta proteòlisi no sigui eficient i que el factor V estigui més activat de l'habitual (Endler and Mannhalter 2003). Aquest polimorfisme es troba únicament a la població europea, amb una freqüència que varia del ~15% fins el 0%, depenent de la població (article Lucotte), i té una edat estimada de

21.000-34.000 anys (Zivelin et al. 1997). La mutació FV_{Leiden} s'ha associat consistentment amb un increment en el risc de patir trombosi, però el paper del polimorfisme en CHD és més discutit. Dues meta-anàlisis diferents apunten que els portadors del polimorfisme R506Q tenen un OD de ~1.3, però només en un d'ells l'associació és estadísticament significativa (Juul et al. 2002).

Un altre gen implicat en la via de la coagulació en el que s'han trobat polimorfismes que podrien estar implicats amb CHD és el factor II o protrombina. La protrombina és una proteïna protrombòtica de ~70 kDa implicada en la conversió del fibrinogen cap a fibrina quan és activada per el factor X i el seu cofactor el FV. A la posició 20210, situat a l'extrem 3' del gen, presenta una posició polimòrfica d'adenina cap a guanina. La variant 20210 G>A es troba a 20 nucleòtids de la senyal poli A i incrementa la taxa de transcripció de la proteïna, associant-se amb un augment del ~25% de l'activitat de la trombina en plasma. (Endler and Mannhalter 2003). L'associació de la variant 20210 A amb la malaltia coronària i l'infart de miocardi no està gaire clara i diverses meta-anàlisis no han trobat una associació entre la variant i la malaltia (Boekholdt et al. 2001).

1.4.3.4.5 Fibrinòlisi

El coàgul creat com a conseqüència d'una ferida vascular s'ha d'eliminar una vegada aquesta ha estat reparada perquè la integritat del flux sanguini es pugui mantenir. La lisi del coàgul és portada a terme per la proteasa plasmina, que talla per múltiples llocs de la matriu de fibrina. La plasmina activa es forma a partir del seu zimogen inactiu, el plasminogen, per acció de l'activador de plasminogen específic de teixit (t-PA) i per l'activador de plasminogen de tipus urokinasa (u-PA); ambdos enzims es troben controlats per l'inhibidor de l'activador del plasminogen 1 (o PAI-1). A més, PAI-1 també pot inactivar APC cosa que s'afavoreix la formació de trombina (Horrevoets 2004). La completa deficiència de PAI-1 dóna lloc a un fenotip de diatesi sagnant lleu. Un elevat nombre d'articles han trobat associacions positives entre presentar uns nivells alts de PAI-1 en sang i el risc de patir diferents classes de malalties arteriotrombòtiques. Un polimorfisme (el 4G/5G, que consisteix en la inserció/deleció d'un nucleòtid) situat a la regió promotora del gen PAI-1 sembla controlar els nivells d'expressió de PAI-1 i, per tant, podria ser un polimorfisme de susceptibilitat per a

CHD. Una meta-anàlisi realitzada amb diferents estudis d'associació no ha trobat que les variants 4G/5G modifiquin el risc de patir CHD (Boekholdt et al. 2001), però diversos estudis apunten que aquest polimorfisme, junt amb la mutació FV_{Leiden} podrien estar implicades en un major risc de patir arteriosclerosi (Horrevoets 2004).

1.4.3.4.6 Oxidació

L'oxidació de les lipoproteïnes HDL i LDL és un procés fortament involucrat en el desenvolupament de l'arteriosclerosi. L'enzim paraoxonasa 1 (PON1) és un dels encarregats d'impedir la formació de peròxids lipídics. PON1 és un enzim de 43kDa que se sintetitza en el fetge, i en sèrum es troba unit a la lipoproteïna HDL; factors com la dieta o l'exercici físic poden modular els nivells d'expressió de l'enzim. Diverses variants polimòrfiques en regions codificants s'han relacionat amb la capacitat de protecció de PON1 a la presència de peròxids lipídics. L'al.loenzim Q de la variant Q192R, per exemple, sembla que confereix una major protecció contra l'oxidació de les LDL que la variant R. L'activitat paraoxonasa es troba modulada seguint un patró 50:40:10 en els individus QQ:QR:RR (La Du 1992). Els estudis d'associació de l'al.lel R amb la malaltia coronària, com en d'altres polimorfismes, no sempre arriben al mateix resultat; una meta-anàlisi (Wheeler et al. 2004) realitzada considerant els portadors de l'al.lel R mostra un OD = 1.44, indicant que aquest polimorfisme confereix un petit increment de risc.

INTRODUCCIÓ

2 Materials i mètodes

2.1 Cerca bibliogràfica i construcció de la base de dades

L'estudi poblacional de les diferents variants al.lèliques es va realitzar a partir de la recopilació bibliogràfica de dades prèviament publicades. Tres malalties mendelianes diferents varen ser considerades: fibrosi quística (gen CFTR), fenilcetonúria (gen PAH) i β-talassèmia (gen β-globina), i una malaltia complexa: la malaltia coronària.

Per a cada gen es va realitzar una cerca bibliogràfica a la base de dades PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). La cerca depengué de la malaltia considerada. En el cas de malalties mendelianes, per a cada malaltia i població es varen buscar articles publicats on es descrivís el patró de mutacions. Aquesta cerca es va realitzar de forma iterativa, iniciant-se amb paràmetres de cerca com :"(cystic fibrosis) AND (population)" o "(cystic fibrosis) AND (spectrum)"; aquestes cerques produeixen molt soroll (no són gaire específiques) però permeten trobar un elevat nombre d'articles candidats. Posteriorment es varen realitzar cerques molt més acurades per tal de trobar articles específics de poblacions que encara no s'haguessin trobat a les cerques generals, utilitzant paràmetres com: "(cystic fibrosis) AND (Spain)". La informació obtinguda mitjançant aquestes cerques es va complementar amb informació poblacional present a les bases de dades de les pàgines web específiques de cada malaltia i amb les referències presents als articles i al llibre "*The Metabolic and molecular bases of inherited Diseases*" (Scriver et al. 2000).

A la base de dades només es varen incloure aquells articles on es fes una descripció exacta del patró de mutacions en una població geogràficament ben definida (no immigrants recents), amb dades no publicades prèviament. Cada mutació es tractà com un únic event a no ser que hi haguessin dades externes que suggerissin molt clarament la possibilitat de fenòmens de recurrència (per exemple, en el cas de la mutació R408W de la fenilcetonúria). La incidència de la fenilcetonúria i de la fibrosi quística a cada població es determinà en un subgrup de poblacions a partir de la informació sobre la incidència present a cada article i a la base de dades "*The Frequency of Inherited Disorders Database*" (FIDD; http://archive.uwcm.ac.uk/uwcm/mg/fidd/). La base de dades es realitzà en fulls de Microsoft Excel 2000 i contingué: país i nom de la població, descripció bibliogràfica de l'article, coordenades geogràfiques, la incidència de la malaltia i el patró de les diferents mutacions. En el cas de la fibrosi quística i la fenilcetonúria, només es varen considerar

poblacions europees, del Nord d'Àfrica i de l'Oest d'Àsia. En el cas de la β-talassèmia es consideraren poblacions de tot el món. En el cas de les malalties fenilcetonúria i β-talassèmia es realitzà un posterior cribatge de les mutacions per gravetat fenotípica. Aquest cribatge es realitzà a partir de la informació present a les bases de dades de la malaltia i de la descripció fenotípica que feien alguns estudis. L'assignació a un determinat fenotip és complex i sovint polèmic, perquè inclou l'acceptació que el fenotip es deu únicament i exclusivament a les mutacions en el gen (cosa que com ja hem vist no és certa) però era condició *sine qua non* per poder treballar amb les dades.

La cerca dels polimorfismes associats a la malaltia coronària es portà a terme mitjançant una altra estratègia, ja que interessava trobar estudis cas-control realitzats amb el polimorfisme d'interès (sense tenir en compte la malaltia) amb la condició que els controls fossin individus sans (no controls hospitalaris) i amb una edat menor de 65 anys. Aquesta mostra es considerà una mostra representativa de les freqüències al.lèliques dels polimorfismes. A partir de la llista de gens associats a la malaltia coronària (Genet Med 2002 4(2):45-61) es va realitzar una cerca per trobar aquells polimorfismes (principalment bial.lèlics) que s'haguessin estudiat en un nombre prou elevat de poblacions (>20). Per a un gen determinat, la cerca inclogué tots els noms del gen i els termes polimorfisme o associació: *"((ANGIOTENSIN I-CONVERTING ENZYME) OR(ACE) OR (ACE1) OR (DCP1)) AND ((polymorphism) OR (association))"*. Aquesta cerca produí un elevat nombre de resultats que posteriorment es filtraren a mà per tal d'excloure aquells que no complissin les condicions dels controls (sans i menors de 65 anys). També s'utilitzà la base de dades ALFRED (http://alfred.med.yale.edu/alfred/) per trobar poblacions on es descrivís la freqüència gènica d'un determinat polimorfisme (per exemple, ACE), així com les referències dels diferents articles a altres estudis cas-control. L'assignació dels controls a una determinada població es va realitzar a partir de la descripció de la població que es feia en l'apartat de material i mètodes de cada article. En el cas que no s'especifiqués l'origen de les mostres, s'assumí que els individus eren autòctons de la regió on estava l'hospital en el que s'havia realitzat l'estudi. A continuació es va fer una selecció d'aquells gens dels quals es tingués informació per a més de 20 poblacions europees, cosa que va fer que finalment es consideressin per a estudis posteriors vuit gens diferents: ACE, APOE, MTHFR, F5, protrombina, AGT, PAI1, PON1. En el cas de ACE, APOE, MTHFR, F5 i protrombina es varen trobar prou poblacions a nivell mundial i aquestes es varen incloure a la base de dades. Per a cada població, quan hi

havia més d'un article per a un determinat polimorfisme es varen sumar el nombre de cromosomes si els autors eren diferents en els estudis o bé si el tipus de mostres utiltizades no eren les mateixes (per exemple, dones en un estudi i homes en un altre) i si les freqüències gèniques eren les mateixes en ambdos estudis (no diferent estadísticament significatiu amb un test de Khi-quadrat); en qualsevol altre cas es considerà l'estudi amb un major nombre de cromosomes.

La base de dades per a cada gen i polimorfisme es realitzà amb fulls de càlcul de Microsoft Excel i constà d'informació sobre el país i el nom de la població, la referència/es bibliogràfica/es, les coordenades geogràfiques i les freqüències al.lèliques de cada polimorfisme. Per terme mig, aproximadament més d'un 70% de les dades introduïdes varen ser contrastades per tal de detectar possibles errors produïts durant la introducció de les dades.

2.2 Mapes d'isolínies

Es varen construïr mapes d'isolínies amb el programa SURFER 7.0 (http://www.goldensoftware.com/) de les freqüències per a cada mutació freqüent (freqüència >1%) en el cas de les malalties mendelianes i per a cada polimorfisme de susceptibilitat en el cas de la malaltia complexa.

Els mapes d'isolínies són un mètode gràfic general de representar els valors d'un determinat paràmetre en diferents punts geogràfics; conecten mitjançant una línia aquells punts que tenen el mateix valor d'un paràmetre, com pot ser una determinada pressió baromètrica o la freqüència d'una variant gènica. En aquest mètode es construeix una matriu regular de punts que cobreix tot el terreny i després s'interpola a partir de les dades reals els valors en els punts interpolats. La interpolació pot dur-se a terme mitjançant un elevat nombre de mètodes; un dels més utilitzats és el mètode de l'invers de la distància al quadrat, on el valor del punt interpolat es calcula en funció de la seva distància geogràfica als punts reals. Així, la freqüència interpolada depèn principalment de la freqüència dels punts reals veïns. Cal tenir present que els mapes d'isolínies, degut a la interpolació (d'altra banda, totalment necessària) només es poden utilitzar de forma indicativa per conèixer el patró espacial de les dades i, en qualsevol cas, mai no s'haurien d'utilitzar els punts interpolats per a fer altres anàlisis espacials, ja

que es basen en l'assumpció d'un determinat model que no té perquè ser cert (Sokal et al. 1999).

2.3 Anàlisi d'autocorrelació espacial

La història de les poblacions humanes i la presència de factors selectius geogràficament ben delimitats fa que les freqüències gèniques no varïin independentment en l'espai; dit d'una altra manera, sabent el valor de la freqüència d'un polimorfisme és possible que puguem endevinar quina és la freqüència de les poblacions veïnes. Degut a aquest fet, aplicar tècniques estadístiques que assumeixin independència de les dades, com és el cas de la correlació, no és gaire adequat. L'anàlisi d'autocorrelació espacial permet veure com covaria una variable (per exemple, la freqüència gènica d'un polimorfisme) en funció de l'espai (Sokal and Oden 1978). L'estadístic de Gary i l'estadístic I de Moran permeten calcular el nivell d'autocorrelació entre un grup de punts, però normalment es prefereix l'estadístic I de Moran perquè els valors d'autocorrelació es troben compresos entre 1 i –1 i, per tant, es poden considerar una estima de r. La I de Moran es calcula com:

$$I = \frac{n\sum_{ij} w_{ij} z_i z_j}{W\sum_{i=1}^{n} z_i^2}$$

On n és el nombre de poblacions que s'estan utilitzant en el càlcul de la I, w_{ij} és el pes que s'assigna a la comparació de la població i amb la població j (per exemple w_{ij} val 1 si les dues poblacions es troben connectades espacialment i 0 si no ho estan), z_i és el valor normalitzat de la variable d'interès a la població i i z_j el valor normalitzat a la població j.

L'autocorrelograma és la representació gràfica del càlcul de la I de Moran a classes de distàncies en les quals la distància és cada vegada més gran entre els parells de punts. La forma que pren l'autocorrelograma permet descriure el patró espacial que tenen les dades (vegeu figura 28).

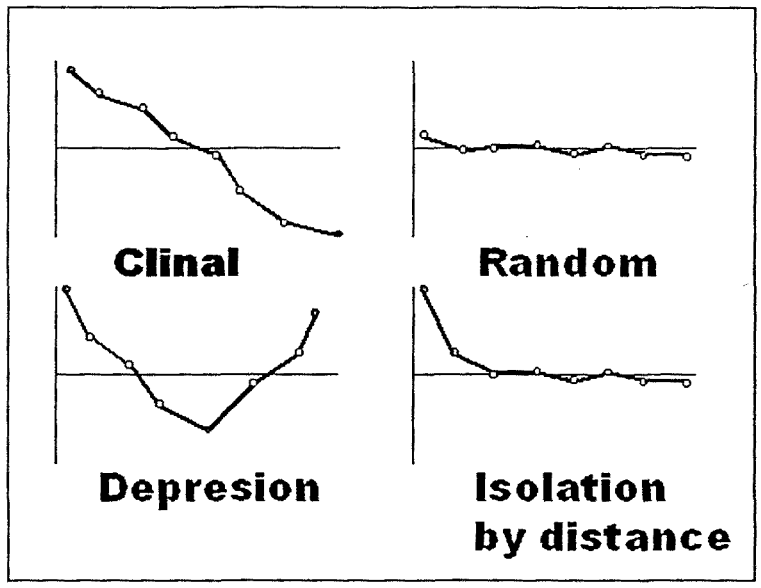


Figura 28. Diferents formes d'autocorrelograma i interpretació del patró. A les ordenades es distribueixen les classes de distància entre els parells de punts i a les abcises el nivell d'autocorrelació (I de Moran) A partir de (Barbujani 2000)

En el cas de gradació del valor de més a menys a l'espai (és a dir, un patró clinal) l'autocorrelograma prendria valors de I de Moran positius i propers a 1 en les classes de distàncies petites entre els parells de poblacions (és a dir, existirà una correlació positiva entre el valor observat en una població i l'observat a les veïnes) mentre que per classes de distàncies molt grans entre parells de poblacions , trobarem I de Moran negatives (és a dir, ara la correlació serà negativa). En absència d'un patró estructurat de les dades, el que observarem a l'autocorrelograma és que, tant per distàncies petites com grans el valor de I de Moran serà molt proper a 0. Altres situacions, com la presència de clines parcials o aïllament per distància, es troben representades a la figura 28. Cal dir que l'autocorrelograma permet conèixer si les dades es troben distribuïdes en forma clinal però no la direcció de la clina. Diferents mètodes s'han proposat per poder definir la direcció de la clina (vegeu (Rosenberg 2000)), però requereixen un elevat nombre de punts en l'espai per poder tenir prou potència. L'anàlisi d'autocorrelació es va dur a terme mitjançant el programa PASSAGE (http://lsweb.la.asu.edu/rosenberg/Passage/).

2.4 Test de Mantel

El test de Mantel (Mantel 1967) permet estimar la correlació lineal de dues matrius de distàncies com, per exemple, una matriu de distància genètica calculada entre parells de poblacions i una altra de la distància geogràfica entre els mateixos parells de poblacions. Aquest mètode és especialment indicat en aquelles variables on no hi ha independència dels valors, i permet realitzar aproximacions al càlcul d'autocorrelacions en tres dimensions, estimant la correlació entre una matriu de distàncies d'una determinada variable amb una matriu de distàncies geogràfiques i controlant per una tercera matriu de diferències en l'alçada.

L'estadístic de Mantel es calcula com:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(x_{ij} - \bar{x})}{s_x} \cdot \frac{(y_{ij} - \bar{y})}{s_y}$$

On r és el coeficient de correlació, n és el nombre de punts, x_{ij} i y_{ij} són els valors a la casella d'intersecció entre els punts i i j a les matrius x i y, i s_x i s_y són els errors estàndar de les matrius x i y.

Donat que els elements de les diferents matrius no són independents entre ells, l'estimació del nivell de significació es realitza mitjançant permutacions.

Cal notar que el test de Mantel equival a una correlació de Pearson i, per tant, no és capaç de detectar correlacions no lineals ni canvis en els patrons de correlació; aquestes anàlisis es poden dur a terme amb altres mètodes que incorporen el test de Mantel amb l'anàlisis d'autocorrelació (Rosenberg 2000). Els tests de Mantel es varen realitzar amb el programa PASSAGE.

2.5 AMOVA i SAMOVA

La presència de barreres, ja siguin físiques o culturals, i els processos selectius fan que la variació genètica tendeixi a disminuir dintre dels grups i a incrementar-se entre els grups. Una manera de testar la presència d'una determinada barrera entre grups

de poblacions per un determinat caràcter (com, per exemple, grups de poblacions per la llengua que parlen) és l'AMOVA o Anàlisi MOlecular de la VAriança (Excoffier et al. 1992). Aquest test permet estimar com es distribueix la variació genètica dintre de cada població, entre les poblacions d'un mateix grup i entre grups de diferents poblacions. L'anàlisi SAMOVA o Anàlisi MOlecular de la VAriança eSpacial (Dupanloup et al. 2004) es pot considerar una extensió de l'AMOVA en la qual, en comptes d'especificar a priori les poblacions de cada grup, només s'especifica el nombre de grups; a partir del nombre de grups indicats, SAMOVA maximitza la variació genètica entre grups (i per tant, maximitza l'homogeneització genètica dintre de cada grup) de poblacions geogràficament veïnes, cosa que es pot correlacionar amb la història comuna o/i a la presència de factors selectius.

2.6 MDS

El mètode de MultiDimensional Scaling (o MDS) és una tècnica multivariant que permet representar en un nombre determinat de dimensions els punts d'una matriu de distàncies (Kruskal and Wish 1990). MDS parteix d'una matriu de distàncies calculada entre parells de punts (com, per exemple, la distància genètica Fst entre parells de poblacions) i de l'especificació del nombre de dimensions en les quals es representaran els punts (o poblacions en el nostre cas). Mitjançant un procés iteratiu, el mètode de MDS posiciona els punts en el nombre de dimensions especificat de forma que la diferència entre la matriu de distàncies originals i la matriu de distàncies calculada amb les noves coordenades sigui la mínima. Aquesta estima es fa amb la variable del Stress (Phi), que es calcula com:

$$Phi = \sum \left[d_{ij} - f(\delta_{ij}) \right]^2$$

on d_{ij} és la distància entre el punt i i el punt j interpolada a partir del número de dimensions especificades i δ_{ij} és la distància observada a les dades. $f(\delta_{ij})$ indica una transformació monòtona no mètrica de les dades.

Una vegada que modificar les coordenades dels punts no disminueix el valor del Stress, el procés s'atura i s'obté una configuració final. Malgrat que no existeix cap

regla matemàtica que estipuli un límit d'acceptació de la configuració final, es considera acceptable un Stress inferior a 0.15 i molt bo per sota de 0.1.

3 Resultats

3.1 Capítol I: "Spatial patterns of cystic fibrosis mutation spectra in European populations."

Oscar Lao, Aida Andres, Eva Mateu, Jaume Bertranpetit, Francesc Calafell.

Eur J Hum Genet. 2003 May;11(5):385-94

Lao O, Andrés AM, Mateu E, Bertranpetit J, Calafell F. Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet. 2003; 11(5): 385-94.

3.2 *Capítol II: "Mutation diversity, demographic history and selection in phenylketonuria"*

Oscar Lao, Isabelle Dupanloup, Guido Barbujani, Jaume Bertranpetit, Francesc Calafell.

(manuscrit en preparació)

MUTATION DIVERSITY, DEMOGRAPHIC HISTORY AND SELECTION IN PHENYLKETONURIA

Oscar Lao[1], Isabelle Dupanloup[2,3], Guido Barbujani[2], Jaume Bertranpetit[1], Francesc Calafell[1]

1 Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

2 Dipartimento di Biologia, Università di Ferrara , Ferrara, Italy

3 Current affiliation: Center for Integrative Genomics, Faculté de Biologie et de Médecine, Université de Lausanne, Lausanne, Switzerland

Keywords: Phenylketonuria, spatial distribution, mutation spectra, effective population size, selective pressure, Europe

Correspondence:
Francesc Calafell
Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida
Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Catalonia
Spain
Tel:+34-93-542 28 41
Fax: +34-93-542 28 02
e-mail: francesc.calafell@upf.edu

ABSTRACT

Phenylketonuria (PKU) is one of the common inborn errors of metabolism in European populations. The mutation spectra of PKU mutations has been recently described by (Zschocke 2003), but the geographical description of the genetic diversity of PKU mutations still remains unknown; this knowledge could allow us to detect fingerprints of demographic and selective events. We have described the spatial pattern of the main PKU mutations and analyzed the genetic diversity of 36 European and SW Asian populations. Autocorrelogram analysis showed clinal patterns for most major mutations. Furthermore, the genetic diversity of PKU mutations also showed a clinal pattern being more diverse in SW and less in NE of Europe. The high genetic diversity of PKU and absence of correlation between the genetic diversity of PKU with other genetic markers, including cystic fibrosis, another common Mendelian disease in Europe, suggest that a recent balacing selection episode increased the incidence of PKU, creating localized, heterogeneous mutation spectra, which have not have sufficient time to diffuse across Europe. Thus, we have shown how selection and demography intertwine in creating the geographic patterns of PKU mutation diversity.

INTRODUCTION

Phenylketonuria (PKU), with a frequency of 1/10000 in Europe, is one of the most common inborn errors of metabolism (Scriver et al. 2000). Mutations at the PAH gene can cause classical, severe forms of PKU, as well as milder forms that can involve only higher blood levels of phenylalanine; over 450 such mutations have been described in the PKU-PAH database (http://www.pahdb.mcgill.ca). This disease can present a broad clinical spectrum that range from high levels of phenylalanine in blood with few other symptoms to a severe impairment of the neurological system (Scriver et al. 2000). Mutations at the PAH gene have been classified according to their most usual phenotype into hyperphenylalaninemia (HPA), mild/moderate PKU and severe/classical PKU (Pey et al. 2003). However, discordant phenotypes among siblings who share the same genotype at the PAH locus exist, implying complex genotype-phenotype correlations and a role for other genetic and environmental factors that influence clinical phenotype (Desviat et al. 1997; Guldberg et al. 1998; Kayaalp et al. 1997; Tyfield et al. 1995).

The geographical distribution of the main PKU mutations has been recently described (Zschocke 2003) in the European continent, but only the spatial pattern of the mutation R408W in the haplotype backgrounds 1 and 2 has been statistically analyzed (O'Donnell et al. 2002; Tighe et al. 2003). Geographic patterns of particular mutations have been traditionally interpreted as the consequence of population movements matching the distribution of the mutation, without factual consideration for the actual demographic strength and genetic consequences of those movements (Zschocke et al. 1997). However, as Lao et al (Lao et al. 2003) have shown for the CFTR locus, the study of the overall pool of the mutation diversity of a Mendelian genetic disease is an optimal way of detecting the influence of demographic history on the mutation pool of a Mendelian disease (Reich and Lander 2001). Moreover, comparison of different Mendelian diseases can allow us to better understand the natural processes that define not only the incidence but also the mutation pattern of each disease.

In order to analyze the genetic diversity of PKU mutations as well as the geographic distribution of the main PKU mutations in Europe, we have compiled and developed a mutation database. We have compared it with neutral regions of the

genome and with the mutation spectra of cystic fibrosis, another common Mendelian disease present in Europe. We discuss the results in terms of the forces that have shaped the mutation spectrum of PKU and reach the conclusion that a complex interplay of demographic history and selection may have operated.

MATERIALS AND METHODS

Database

The hyperphenylalaninemia-phenylketonuria mutation frequencies database was created with bibliographic data published previous to 2003 and considered only European and SW Asian populations. In contrast to a previous article that shows a comprehensive review of the PKU mutations in the same populations (Zschocke 2003), we collected data on the overall pool of mutations rather than on the most common ones. The database contained information about the geographical localization, the incidence of the disease, relative mutation frequencies and sample size was considered for each population. In the case of R408W, which is known to be recurrent (Byck et al. 1994), the two main background haplotypes (haplotype 1 and haplotype 2) were considered separately.

Since the level of disease severity depends on each mutation, we have classified the biological fitness of each mutation in two groups (hyperphenylalaninemia or PKU) according to the genotype/phenotype relationships described in PKU-PAH database (http://www.pahdb.mcgill.ca) as well as from bibliography (Desviat et al. 1997; Guldberg et al. 1998; Kayaalp et al. 1997; Mallolas et al. 1999). It should be noticed that, due to the complex relationships between the environment and the genotype, this classification must be taken as a rough guide and not as a precise description of the effect of each mutation in the phenotype (Pey et al. 2003; Waters 2003).

The fraction of chromosomes without a recognizable mutation were grouped as "unknown"; all unknown mutations were considered as PKU, since they were ascertained in severe cases. After mutation classification, population-specific mutations were gathered in a class called "other" but were taken into account separately when computing the genetic diversity of each population. For further analysis, we have only considered PKU mutations, as well as populations with sample sizes over 20 chromosomes and with a frequency of unknown mutations lower than 60%. PKU incidence data were obtained from articles (Kuzmin et al. 1995; Lillevali et al. 1996;

Mallolas et al. 1999; Rivera et al. 2000; Scriver et al. 2000; Zekanowski et al. 2001; Zschocke et al. 1997) and the Frequency of Inherited Disorders Database (FIDD; http://archive.uwcm.ac.uk/uwcm/mg/fidd/).

The final database contained 36 populations with incidence available for 24 of them, 82 shared PKU mutations and 118 private (population-specific) PKU mutations.

It should be noted that the same caveats stated by Zschocke (Zschocke 2003) apply to our database. Namely, mutation detection is not uniform across populations; samples may not be representative of the whole population; hyperphenylalaninemia and PKU mutations may not be readily distinguishable; patients are not selected according to authochtony; and haplotype data that could be useful in recognizing mutations with multiple origins are not always available.

Genetic diversity estimate

The genetic diversity of PKU mutations for each population was computed as an expected heterozygosity estimate by means of the Arlequin package (Schneider et al. 2000). Computation of the expected heterozygosity requires the complete specification of the frequencies of all alleles. However, an average 16.1% (ranging from 0% to 50%) of the PKU chromosomes carried unknown PKU mutations. Estimates of allele diversity are bracketed between two extremes: they would be minimal if all unknown mutations in a population were, in fact, the same allele, and they would be maximal if all unknown mutation were different from each other. The latter situation has been observed in the case of CF mutations (Le Marechal et al. 2001), and can be expected to be similar for PKU. Thus, it is likely that allele diversity estimates are much closer to their maximum possible values than to their minima; the estimates we present were computed under the assumption that all unknown mutations are different from each other.

Geographical patterns

Maps of gene frequencies for the PKU mutations with average frequencies >3% (namely R408W haplotype 1 and haplotype 2, IVS 10-11G>A, IVS12-1 G>A, R261Q and R158Q), as well as for genetic diversity estimates, were obtained using Surfer 7.0 (Golden Software Inc) with the inverse distance method for interpolation points. A regular grid covering Europe and the Middle East, and limited between 30 °N and 64 °N and between 10°W and 50.5°E was used. Interpolation points were spaced 0.1 degrees.

For each interpolation point, only data points within the same landmass (island or continent) were considered. It should be noted that interpolation was used only to map allele frequencies and diversities, and that interpolated values were not used in any other analysis.

The frequency of the most common mutations, maximum genetic diversity estimates, and countrywide PKU incidences were subjected to spatial autocorrelation analysis (Sokal and Oden 1978) by means of the PASSAGE program (Rosenberg 2001). Distance classes were defined as the equal geographical distance between pairs of populations. Due to the nature of the geographical distribution of the data, the most distant geographical distance class was excluded from the plot representation, since it often contained a very small number of pairs of points.

Genetic distances

Reynolds' genetic distances (Reynolds et al. 1983) based on PKU mutation relative frequencies were computed. The same measure of genetic distance was used by (Cavalli-Sforza et al. 1994) to estimate genetic distances among European populations based on *classical* genetic polymorphisms (i.e., blood groups, protein polymorphisms, and HLA). Both distance matrices were compared by means of a Mantel test (Mantel 1967). Reynolds' distances were also computed and compared to PKU mutation distances for Y chromosome haplogroup frequencies (Rosser et al. 2000). PKU mutation distances were also compared to corrected pairwise distances for hypervariable region I mitochondrial DNA sequences (Orekhov et al. 1999; Richards et al. 2000; Simoni et al. 2000); since some slightly negative values were obtained, a small positive quantity was added to all distance values so that all would be positive. Finally, PKU mutation distances were compared to CFTR mutation distances(Lao et al. 2003). Genetic distance calculations and Mantel tests were performed with Arlequin 2.000 (Schneider et al. 2000)

Multidimensional Scaling (MDS)

MDS was performed with the computed Reynolds' genetic distance matrix. This multivariate procedure plots in a number of dimensions specified a priori the data trying to reproduce the original distance matrix. The measure of how well the final configuration fits the original distance matrix is given by a parameter called stress.

Although there is not an established upper limit to the stress, a good representation of the data is accepted when the stress is under 0.15 and excellent when it does not reach 0.1 (referència).

SAMOVA

SAMOVA (Spatial Analysis of the Molecular Variance) was used to define homogeneous genetic groups of geographically neighbouring populations defined by the Delaunay triangulation. In contrast to AMOVA, which describes the percentage of variation between groups of populations defined a priori, SAMOVA only requires a preset number of groups of populations; the SAMOVA algorithm maximizes the fraction of genetic variation explained between groups of geographically connected populations by a recursive clustering of populations (Dupanloup et al. 2002).

RESULTS

PKU mutation spectra was analyzed in 36 populations from Europe and SW Asia. Information about populations, sample sizes, main mutation frequencies and mutation diversities can be found in Table 1. Frequency maps and spatial autocorrelograms of the six PKU mutations with an average frequency over 3% are displayed in Figs. 1-6. Each mutation showed a particular geographical distribution.

Mutation R408W is found in two different haplotype backgrounds probably as a result of independent mutation events (Byck et al. 1994), and each version of the mutation shows a different geographic distribution (O'Donnell et al. 2002; Tighe et al. 2003). Thus, R408W on haplotype 2 has an average frequency of 24.2% across the populations in the database and showed a statistically significant partially clinal pattern ($p<0.05$, Fig. 1), with high frequencies in NE Europe (up to 87.5% in Estonia) and much lower in SW Europe. This contrasts with the pattern observed for the same mutation on haplotype 1 (average frequency 4.8%; Fig. 2), which showed a pattern compatible with isolation by distance ($p<0.05$), and is almost restricted to NW Europe (the British Isles, where it constitutes ~1/3 of the PKU chromosomes, and Scandinavia).

Mutation R261Q has an average frequency of 6.8% ($p=0.032$; Fig. 3) showed a random distribution autocorrelogram, with high values in Switzerland, Calabria and Netherlands and low in the NE Europe. R158Q ($p=0.033$; average frequency 3.44%; Fig. 4) did not show any spatial distribution, with elevated values in Belgium, Tataria and South of Italy. Intron boundary mutation IVS 10-11G>A (Fig. 5) is more frequent in Southern (20-30% in S Italy and Spain) than in northern Europe. Actually, it shows a statistically significant autocorrelogram ($p<0.0005$) with a pattern indicative of a partial cline. IVS12+1G>A (Fig. 6) also showed a significant partially clinal autocorrelogram ($p<0.0005$) although in the opposite direction, with maxima in NW Europe (>25% in England and Scandinavia).

The richness of a mutation spectrum can be summarized by several genetic parameters, such as the expected heterozygosity and θ. The maximum allele diversity (see Methods) was computed for each population and subjected to spatial analysis. It ranges from less than 0.2 to over 0.9, and shows a statistically significant clinal autocorrelogram ($p<0.0005$, Fig 7), with maxima in S and SW Europe (Italy, Spain) and minima in N and NE Europe (Estonia). This predominant SW-NE orientation of the cline is also detected by geographical correlations; for latitude as well as longitude, PKU mutation diversity is correlated negatively with both ($r= -0.492$, $p=0.002$ and r=-

0,488, p= 0.003, respectively). Similar results are obtained for θ, with a three-fold decrease from SW to NE Europe, a significant clinal autocorrelogram ($p<0.001$), and a high negative correlation with latitude and longitude ($r= -0.487$, $p=0.003$ for latitude and $r=-0,432$, $p= 0.008$ for longitude).

All of the previous analyses were performed on the mutation frequencies relative to the total PKU chromosomes rather than to the total population (irrespective of whether it carries or not a PKU mutation). For instance, on average 29.14% of PKU chromosomes carry the mutation R408W in the haplotype 2 background, but, given the average PKU prevalence in Europe, 0.29% of all chromosomes carry R408W haplotype 2. It should be noted that the incidence of PKU in Europe is highly irregular and not correlated with geography ($r=0.089$, $p=0.678$ for latitude and $r=-0.094$, $p=0.661$ for longitude). Incidences are small and their estimate carries large standard errors. Actually, if incidence estimates are assumed to be derived from complete ascertainment of all births in the population over one year, then the incidences in 19 different European populations (all in Table 1 except for Ireland, Sicily, and Turkey) are homogeneous ($p=0.163$, Fisher's exact test). That is, PKU incidence in Europe can be considered statistically homogeneous in Europe except for those three outliers. Therefore, all analyses performed on mutation frequencies relative to the PKU chromosomes would translate linearly into the frequencies relative to the total population.

Genetic distances among PKU mutation pools (using all known mutations rather than just the common ones described above) in different European populations were computed with Reynolds' genetic distance. By a Mantel test, we found a significant correlation between PKU distances and geographic distances ($r=0.29$, $p=0.001$). The genetic distance matrix for PKU mutations was represented by means of a three-dimensional MDS. The stress was 0.047, thus suggesting a good reduction of the complexity of the original data. The plot of first and second dimensions are represented in figure 8. Populations can be clustered in three main geographical groups (Mediterranean populations, NW Europe and Central/NE Europe). The third dimension separates British and Irish populations from the Netherlands and Denmark.

A more specific analysis for detecting the geographical partitioning of PKU mutation variance (i.e., SAMOVA) was then performed. When the number of preset groups was larger than four, each new group defined by SAMOVA consisted of a single

population. Considering four groups the fraction of variance explained among groups was 13.33%; in rough agreement with MDS results, the four groups determined by SAMOVA were (figure 9) the Mediterranean and Switzerland, Scotland and Ireland, NW Europe (that is Scandinavia, England, W Germany, the Benelux and N Italy), and Central and NE Europe. Each group was defined by a subset of particularly frequent mutations (see table 2) that were absent or at low frequency in the other groups, such as IVS 10-11 G>A in the Mediterranean, R408W-haplotype 1 in Ireland and Scotland, IVS 12+1 G>A in NW Europe, and R408W-haplotype 2 in central/NE Europe.

For a subset of 17 populations, genetic distances based on classical polymorphisms (Cavalli-Sforza et al. 1994) were available. This was also the case for 18 populations and mtDNA control region sequences (Orekhov et al. 1999; Richards et al. 2000; Simoni et al. 2000) and 21 populations and Y chromosome haplogroups (Rosser et al. 2000). In addition to the analysis of neutral regions, we also studied the relationship between the PKU spectrum and the cystic fibrosis mutation spectrum (Lao et al. 2003)in 30 populations. Correlation coefficients of the Mantel test are presented in table 3. The correlation with PKU mutation spectra was high (~0.5) and significant only for Y-chromosome based distances, and remained significant after controlling for the geographic distance. We also analyzed the correlation between PKU mutation diversity (as measured by maximum expected heterozigosity) and genetic diversity in the same neutral regions. In both cases, correlation was not significantly different from 0 ($r = -0.336$, $p=0.136$ for PKU – Y chromosome and $r=-0.159$, $p=0.528$ for PKU-mtDNA HVRI). The correlation between CF heterozigosity and PKU was not significant either ($r=0.097$, $p=0.612$).

DISCUSSION

We have studied the geographical pattern of PKU mutations in Europe and NW Asia, and we have also analyzed the diversity of mutation spectra as well as their spatial pattern in this geographical frame.

Our results corroborate previous spatial studies based on specific mutations (O'Donnell et al. 2002; Tighe et al. 2003) and give statistical support to qualitative descriptions of the geographical patterns of mutation spectrum presented in (Zschocke 2003): the main PKU mutations have a particular geographical distribution and show spatial patterns in Europe (except for R261Q and R158Q, which are irregular); shape

and direction are specific to each one, ranging from depression to isolation by distance patterns.

Mutation diversity has shown a clinal pattern, with values ranging from 0.23 in NE Europe to 0.96 in SW Europe; however, it should be taken into account that 72% of countries analyzed have heterozygosities over 0.8, which implies that the determination of the causative mutations in an affected case is going to be more laborious in a large number of countries compared with other common Mendelian diseases present in Europe, such as cystic fibrosis, in which only 8.5% of populations present genetic diversities over 0.8 (Lao et al. 2003). It has been argued (Lao et al. 2003) that the characterization of mutation diversity by means of θ could explain this pattern. According to (Reich and Lander 2001), this pattern is an estimator of $4N_e\mu(1-f_0)$, where N_e is the effective population size, μ is the mutation rate and f_0 is the incidence of the overall disease allele class which depends on the mutation rate and the selection coefficient. Due to the fact that it is improbable that the mutation rate was higher in southern than in northern Europe, a higher θ between populations for the same Mendelian disease is likely to be the result of a lower f_0 and/or higher N_e in southern than in northern Europe. We will discuss first the possible impact of f_0 on θ and will deal with effective population size afterwards.

However, it should be noted first that (Lao et al. 2003) have shown that, even for the most frequent Mendelian disorders, a large increase in their incidence has a small impact in the general gene pool, so variation in incidence is unlikely to have a large effect on mutation diversity. This is confirmed by the lack of correlation between θ and the frequency of alleles carrying PKU mutations ($r=0.148$, $p=0.489$); according to Reich and Lander (Reich and Lander 2001), a negative correlation should have been observed. This result, however, should be taken with caution. As noted above, incidence figures, even in a disease with such methodic screenings as PKU, carry large sampling errors, thus preventing the detection of an expectedly weak signal.

The incidence of a Mendelian disease can be the equilibrium point of a selection-mutation equilibrium. Then, f_0 would depend on both the mutation rate and the fitness coefficients, and a non homogeneous selective pressure could have shaped the observed pattern of genetic diversity. In particular, a stronger clearance of the PKU mutations in southern Europe could explain the presence of milder mutations compared with other countries: the average frequency of mild mutations is 0.097 in southern

Europe and 0.060 in northern Europe, and the difference is statistically significant (χ^2=22.74, p=1.8×10^{-6}). This variable selective pressure could be due to dietary factors, such as the phenylalanine intake, as well as other unknown environmental and genetic factors. Nevertheless, prehistoric dietary habits have changed along time and among populations and it seems unlikely that a clinal pattern like that observed could be the product of these factors. Moreover, incidence in southern Europe is not significantly lower.

Heterozygote advantage is another factor that could change the pattern of incidence, and thus of genetic diversity. The elevated incidence of PKU in European populations, together with the fact that different mutations have risen to high frequencies in different regions of Europe, cannot be accounted for by genetic drift and mutation rate alone (Krawczak and Zschocke 2003) so a general selective advantage has been suggested for PKU carriers. A regression of heterozygosity against incidence for a subset of recessive Mendelian diseases (r =-0.39; p>0.05) shows that the diversity of PKU is higher than that predicted from its incidence, which would be compatible with balancing selection, although the difference between the observed and the expected diversities does not reach statistical significance. As with the case of negative selection, the selection coefficients should have been different across Europe to explain the observed pattern: heterozygotes should have had a larger advantage in N Europe. There is some external evidence both for and against selective advantage of PKU heterozygotes (Woolf et al. 1975); (Saugstad 1973, 1977) and, although it has been related to mycotoxin resistance in populations where the incidence is elevated such as Ireland, there is no contemporary evidence to sustain this hypothesis (Scriver et al. 2000). In summary, although negative selection against PKU homozygotes, and a possibly heterozygote advantage have taken place, they have had an undetectable role in shaping the geographical pattern of PKU mutation diversity in Europe.

As stated above, effective population size can be a main factor in explaining the clinal pattern of PKU diversity in Europe. This would imply a general, demographic effect, and we would expect that genetic distances based on PKU mutations would correlate with genetic distances computed from loci elsewhere in the genome, since demography acts on all of the genome simultaneously. However, comparison of the genetic distance matrices of the Y chromosome, mtDNA and CFTR with PKU mutation diversity has shown a statistical correlation only for the Y chromosome. Mutation diversity is high in S Europe both for CF and PKU mutations, but in NW Europe,

diversities are high for PKU but low for CF, thus contributing to an overall low correlation coefficient. Since we are comparing both Mendelian diseases in the same populations, we cannot postulate that exactly the same demographic histories have shaped CF and PKU diversities. This apparent discrepancy can be solved taking into the account the respective ages of CF and PKU mutations. PKU mutation spectra cluster in geographical groups that are mainly ruled by the presence of one mutation at a frequency below 0.5; in contrast, the CF mutation spectrum is widely dominated by 508Fdel, which is widespread around Europe and reaches an overall frequency of 0.6. The absence of the haplotypes associated with 508Fdel in the normal population has been interpreted as an ancient origin of this selection (Mateu et al. 2002), possibly predating the spread of anatomically modern humans into Europe, while the localized distribution of PKU mutants suggests a much more recent origin. Thus, the demographic processes that PKU and CF mutants have been subjected to would not have been equivalent. In this frame, the correlation of PKU mutation distribution with that of Y chromosome haplogroups is particularly relevant. Although some of them are quite ancient, the localized distribution of Y chromosome haplogroups has been attributed to the lower male- than female-mediated gene flow (Perez-Lezaun et al. 1999; Rosser et al. 2000; Seielstad et al. 1998) caused by marriage patrilocality. In the case of PAH mutations, their heterogeneity may be a reflection of a recent expansion linked to an episode of heterozygote advantage. Thus, many mutations may not have diffused far from the populations where they expanded. This diffusion may have been additionally hampered by the relatively low incidence of PKU, which is four times lower than that of CF.

The analysis of the mutation spectra of PKU and its comparison with other classical Mendelian diseases such as cystic fibrosis allows us to better understand the processes that are involved in the geographical distribution of mutations in its natural framework. We have shown how selection and demography are not independent processes, since selection can modulate the time of expansion of a mutation spectrum, and thus, the point in the population demographic history from which demographic processes can shape the mutation spectrum. Despite the importance of the demographic history on shaping the distribution of PKU mutations, other evolutionary forces are involved and must be considered when inferring the putative distribution of particular mutations in Mendelian diseases.

ACKNOWLEDGEMENTS

This work was supported by Dirección General de Investigación Científica y Técnica (Spanish Government) grants BMC2001-0772 and BOS2001-0794. O.L. was supported by a predoctoral fellowship from the Ministerio de Ciencia y Tecnología. We wish to thank Giorgio Bertorelle for his productive comments.

REFERENCES

Byck S, Morgan K, Tyfield L, Dworniczak B, Scriver CR (1994) Evidence for origin, by recurrent mutation, of the phenylalanine hydroxylase R408W mutation on two haplotypes in European and Quebec populations. Hum Mol Genet 3: 1675-7

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

Desviat LR, Perez B, Garcia MJ, Martinez-Pardo M, Baldellou A, Arena J, Sanjurjo P, Campistol J, Couce ML, Fernandez A, Cardesa J, Ugarte M (1997) Relationship between mutation genotype and biochemical phenotype in a heterogeneous Spanish phenylketonuria population. Eur J Hum Genet 5: 196-202

Dupanloup I, Schneider S, Excoffier L (2002) A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11: 2571-81

Guldberg P, Rey F, Zschocke J, Romano V, Francois B, Michiels L, Ullrich K, Hoffmann GF, Burgard P, Schmidt H, Meli C, Riva E, Dianzani I, Ponzone A, Rey J, Guttler F (1998) A European multicenter study of phenylalanine hydroxylase deficiency: classification of 105 mutations and a general system for genotype-based prediction of metabolic phenotype. Am J Hum Genet 63: 71-9

Kayaalp E, Treacy E, Waters PJ, Byck S, Nowacki P, Scriver CR (1997) Human phenylalanine hydroxylase mutations and hyperphenylalaninemia phenotypes: a metanalysis of genotype-phenotype correlations. Am J Hum Genet 61: 1309-17

Krawczak M, Zschocke J (2003) A role for overdominant selection in phenylketonuria? Evidence from molecular data. Hum Mutat 21: 394-7

Kuzmin AI, Eisensmith RC, Goltsov AA, Sergeeva NA, Schwartz EI, Woo SL (1995) Complete spectrum of PAH mutations in Tataria: presence of Slavic, Turkic and Scandinavian mutations. Eur J Hum Genet 3: 246-55

Lao O, Andres AM, Mateu E, Bertranpetit J, Calafell F (2003) Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet 11: 385-94

Le Marechal C, Audrezet MP, Quere I, Raguenes O, Langonne S, Ferec C (2001) Complete and rapid scanning of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by denaturing high-performance liquid chromatography (D-HPLC): major implications for genetic counselling. Hum Genet 108: 290-8

Lillevali H, Ounap K, Metspalu A (1996) Phenylalanine hydroxylase gene mutation R408W is present on 84% of Estonian phenylketonuria chromosomes. Eur J Hum Genet 4: 296-300

Mallolas J, Vilaseca MA, Campistol J, Lambruschini N, Cambra FJ, Estivill X, Mila M (1999) Mutational spectrum of phenylalanine hydroxylase deficiency in the population resident in Catalonia: genotype-phenotype correlation. Hum Genet 105: 468-73

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220

Mateu E, Calafell F, Ramos MD, Casals T, Bertranpetit J (2002) Can a place of origin of the main cystic fibrosis mutations be identified? Am.J Hum.Genet. 70: 257-264

O'Donnell KA, O'Neill C, Tighe O, Bertorelle G, Naughten E, Mayne PD, Croke DT (2002) The mutation spectrum of hyperphenylalaninaemia in the Republic of Ireland: the population history of the Irish revisited. Eur J Hum Genet 10: 530-8

O'Neill CA, Eisensmith RC, Croke DT, Naughten ER, Cahalane SF, Woo SL (1994) Molecular analysis of PKU in Ireland. Acta Paediatr Suppl 407: 43-4

Orekhov V, Poltoraus A, Zhivotovsky LA, Spitsyn V, Ivanov P, Yankovsky N (1999) Mitochondrial DNA sequence diversity in Russians. FEBS Lett. 445: 197-201

Ozguc M, Ozalp I, Coskun T, Yilmaz E, Erdem H, Ayter S (1993) Mutation analysis in Turkish phenylketonuria patients. J Med Genet 30: 129-30

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65: 208-19

Pey AL, Desviat LR, Gamez A, Ugarte M, Perez B (2003) Phenylketonuria: genotype-phenotype correlations based on expression analysis of structural and functional mutations in PAH. Hum Mutat 21: 370-8

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-10

Reynolds J, Weir B, Cockerham C (1983) Estimation for the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767-779

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am.J Hum.Genet. 67: 1251-1276

Rivera I, Cabral A, Almeida M, Leandro P, Carmona C, Eusebio F, Tasso T, Vilarinho L, Martins E, Lechner MC, de Almeida IT, Konecki DS, Lichter-Konecki U (2000) The correlation of genotype and phenotype in Portuguese hyperphenylalaninemic patients. Mol Genet Metab 69: 195-203

Rosenberg M (2001) Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 1.1., 1.1 edn. Department of Biology, Arizona State University, Tempe, AZ

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am.J Hum.Genet. 67: 1526-1543

Saugstad LF (1973) Increased "reproductive casualty" in heterozygotes for phenylketonuria. Clin Genet 4: 105-14

Saugstad LF (1977) Heterozygote advantage for the phenylketonuria allele. J Med Genet 14: 20-4

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

Scriver C, Sly W, Childs B, Beaudet A, Valle D, Kinzler K, Vogelstein B (2000) The Hyperphenylalaninemias

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

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am.J Hum.Genet. 66: 262-278

Sokal R, Oden N (1978) Spatial autocorrelation in biology 1. Methodology. Biol J Linn Soc 10: 199-228

Tighe O, Dunican D, O'Neill C, Bertorelle G, Beattie D, Graham C, Zschocke J, Cali F, Romano V, Hrabincova E, Kozak L, Nechyporenko M, Livshits L, Guldberg P, Jurkowska M, Zekanowski C, Perez B, Desviat LR, Ugarte M, Kucinskas V, Knappskog P, Treacy E, Naughten E, Tyfield L, Byck S, Scriver CR, Mayne PD, Croke DT (2003) Genetic diversity within the R408W phenylketonuria mutation lineages in Europe. Hum Mutat 21: 387-93

Tyfield LA, Zschocke J, Stephenson A, Cockburn F, Harvie A, Bidwell JL, Wood NA, Hunt LP (1995) Discordant phenylketonuria phenotypes in one family: the relationship between genotype and clinical outcome is a function of multiple effects. J Med Genet 32: 867-70

Waters PJ (2003) How PAH gene mutations cause hyper-phenylalaninemia and why mechanism matters: insights from in vitro expression. Hum Mutat 21: 357-69

Woolf LI, McBean MS, Woolf FM, Cahalane SF (1975) Phenylketonuria as a balanced polymorphism: the nature of the heterozygote advantage. Ann Hum Genet 38: 461-9

Zekanowski C, Jurkowska M, Bal J (2001) Association between minihaplotypes and mutations at the PAH locus in Polish hyperphenylalaninemic patients. Hum Hered 51: 117-20

Zschocke J (2003) Phenylketonuria mutations in Europe. Hum Mutat 21: 345-56

Zschocke J, Mallory JP, Eiken HG, Nevin NC (1997) Phenylketonuria and the peoples of Northern Ireland. Hum Genet 100: 189-94

Population	Latitud	Longitud	Incidence	ref	R408W(hap 2)	R408W(hap 1)	R261Q	R158Q	IVS10-11G>A	IVS12-1G>A	Other	Unknown	H_{max}
Belarus	53.50	28.00	-	(Zschocke et al. 1997)	0.679	0.000	0.014	0.017	0.000	0.010	0.000	0.280	0.538
Belgium	50.83	4.33	10000		0.079	0.000	0.070	0.128	0.070	0.161	0.285	0.207	0.933
Bulgaria	42.41	23.19	20000		0.500	0.000	0.033	0.033	0.167	0.000	0.033	0.233	0.736
Czech Republic	50.05	14.26	9000		0.560	0.004	0.019	0.050	0.031	0.042	0.201	0.058	0.679
Croatia	45.48	15.58	14500		0.380	0.000	0.060	0.020	0.040	0.020	0.140	0.320	0.845
Denmark	55.40	12.35	11764	(Zschocke et al. 1997)	0.179	0.010	0.017	0.030	0.054	0.389	0.264	0.010	0.802
Estonia	59.25	24.45	5236	(Lillevali et al. 1996)	0.875	0.000	0.025	0.000	0.000	0.025	0.000	0.075	0.237
Germany													
E Germany	52.33	13.30	-		0.422	0.000	0.045	0.037	0.031	0.086	0.257	0.061	0.807
Germany and W Germany	50.07	8.40	-		0.280	0.000	0.063	0.056	0.045	0.132	0.317	0.016	0.887
Great Britain													
London	51.30	-0.10	14285	(Zschocke et al. 1997)	0.048	0.127	0.000	0.000	0.000	0.190	0.365	0.190	0.936
SW England	50.23	-4.10	12000	(Zschocke et al. 1997)	0.036	0.073	0.009	0.000	0.045	0.300	0.427	0.073	0.893
Wales	51.29	-3.13	13000	(Zschocke et al. 1997)	0.036	0.125	0.036	0.000	0.000	0.196	0.482	0.125	0.940
W Scotland	55.53	-4.15	5263	(Zschocke et al. 1997)	0.064	0.316	0.011	0.005	0.000	0.059	0.385	0.123	0.879
Ireland													
Northern Ireland	54.35	-5.55	4500	(Zschocke et al. 1997)	0.034	0.349	0.017	0.000	0.011	0.046	0.509	0.006	0.846
Republic of Ireland	53.20	-6.15	4500	(O'Neill et al. 1994)	0.014	0.376	0.000	0.000	0.011	0.043	0.082	0.426	0.851
Italy													
Calabria	38.12	15.65	-		0.000	0.000	0.273	0.091	0.212	0.000	0.152	0.273	0.881
Campania	40.85	14.28	-		0.013	0.000	0.190	0.025	0.228	0.000	0.278	0.266	0.896
Piemonte	45.05	7.67	-		0.031	0.000	0.094	0.063	0.063	0.094	0.281	0.375	0.962
Puglia/Basilicata	41.12	16.87	-		0.020	0.000	0.000	0.000	0.143	0.000	0.469	0.367	0.948
Sicilia	37.30	14.00	2700	(Zschocke et al. 1997)	0.000	0.000	0.164	0.096	0.301	0.000	0.411	0.027	0.852
Lithuania	56.00	24.00	-		0.769	0.000	0.011	0.071	0.005	0.000	0.088	0.055	0.404
Netherlands	52.22	4.54	18000		0.000	0.000	0.214	0.024	0.024	0.286	0.286	0.071	0.872
Norway	59.55	10.45	14285		0.083	0.074	0.088	0.009	0.005	0.162	0.546	0.005	0.901
Palestina	31.30	34.28	-		0.000	0.013	0.079	0.026	0.184	0.000	0.158	0.461	0.939
Poland				(Zekanowski et al. 2001)									
Poland	52.15	21.00	8000		0.619	0.000	0.013	0.039	0.074	0.016	0.061	0.171	0.610
S Poland	50.03	19.58	5000		0.553	0.000	0.021	0.064	0.043	0.000	0.021	0.255	0.696
Portugal	38.43	-9.08	12500		0.022	0.000	0.159	0.043	0.181	0.000	0.239	0.297	0.933
Russia													
Moscow (+Rus in Tataria)	55.45	37.35	-		0.605	0.000	0.026	0.066	0.000	0.158	0.000	0.132	0.610
St Petersburg	59.55	30.15	-		0.707	0.000	0.043	0.014	0.007	0.021	0.093	0.100	0.497
Tataria	55.49	49.08	6000	(Kuzmin et al. 1995)	0.370	0.000	0.148	0.074	0.074	0.037	0.185	0.000	0.843
Slovakia	48.09	17.07	-		0.459	0.000	0.071	0.071	0.000	0.102	0.020	0.276	0.773
Spain													
Spain	40.24	-3.41	10000	FIDD	0.000	0.000	0.053	0.000	0.202	0.000	0.511	0.149	0.949
Catalonia	41.23	2.11	6600		0.000	0.000	0.078	0.031	0.172	0.000	0.461	0.156	0.953
Sweden	59.20	18.03	20000	(Zschocke et al. 1997)	0.213	0.012	0.018	0.018	0.000	0.166	0.278	0.284	0.890
Switzerland	46.95	7.43	16000	FIDD	0.060	0.000	0.320	0.040	0.040	0.040	0.000	0.500	0.897
Turkey	39.50	32.50	4370	(Ozguc et al. 1993)	0.026	0.000	0.072	0.013	0.346	0.000	0.098	0.412	0.874

Table 1. Database used in the analysis. N, sample size (in number of PKU chromosomes), relative frequencies of the main mutations; other, relative frequency of mutations <1%; private, relative frequency of those mutations present only in one population; unknown, fraction of chromosomes associated with disease bearing unidentified mutations. H_{max} allele diversity parameter (see text).

SAMOVA group	R408W hap 1	R408W hap 2	IVS 10-11G>A	IVS 12+1G>A	R261Q
Mediterranean	0.001	0.009	0.22	0	0.12
NW Europe	0.04	0.01	0.03	0.20	0.06
NE Europe	<0.001	0.58	0.04	0.04	0.04
Ireland & Great Britain	0.35	0.04	0.007	0.05	0.01

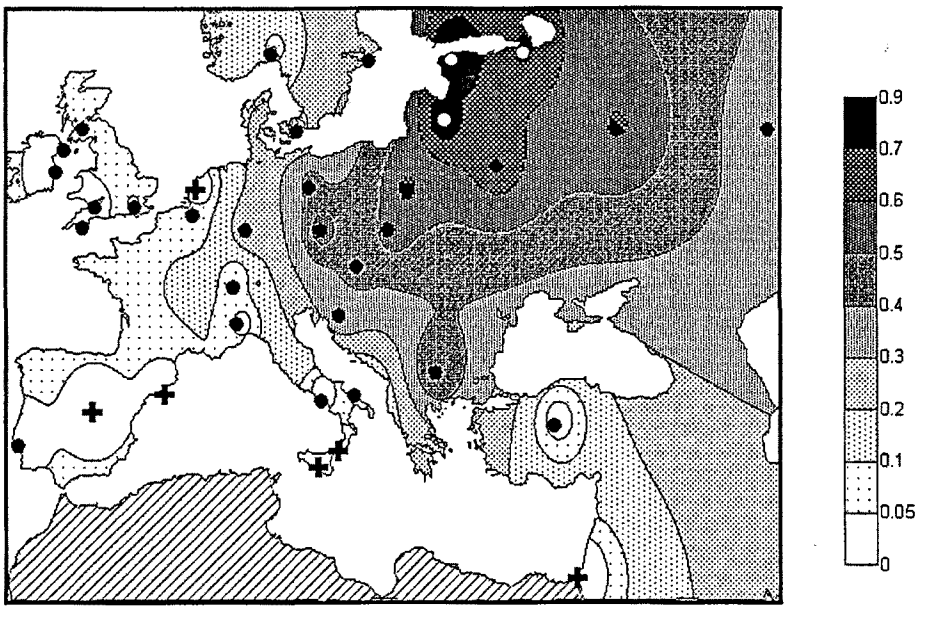
Table 2. Groups defined by SAMOVA and particularly frequent mutations in each group.

Locus	N	Mantel test with PKU mutation distances correlation	p	Mantel test with PKU mutation distances corrected by geographical distance correlation	p
Classical markers	17	0.26	0.056	0.12	0.19
MtDNA	18	-0.087	0.68	-0.081	0.661
Y chromosome	21	0.566	<0.0005	0.500	<0.0005
CFTR	30	0.018	0.338	-0.106	0.879

Table 3. Correlation observed between PKU genetic distance matrix and genetic distance matrix based on three neutral markers and one locus associated to a Mendelian disease. N = subset of shared populations in both PKU and locus genetic distance matrix.

Figure 1.

Geographical distribution (a) and spatial autocorrelogram (b) of R408W under the haplotype background 2 in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.9
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0

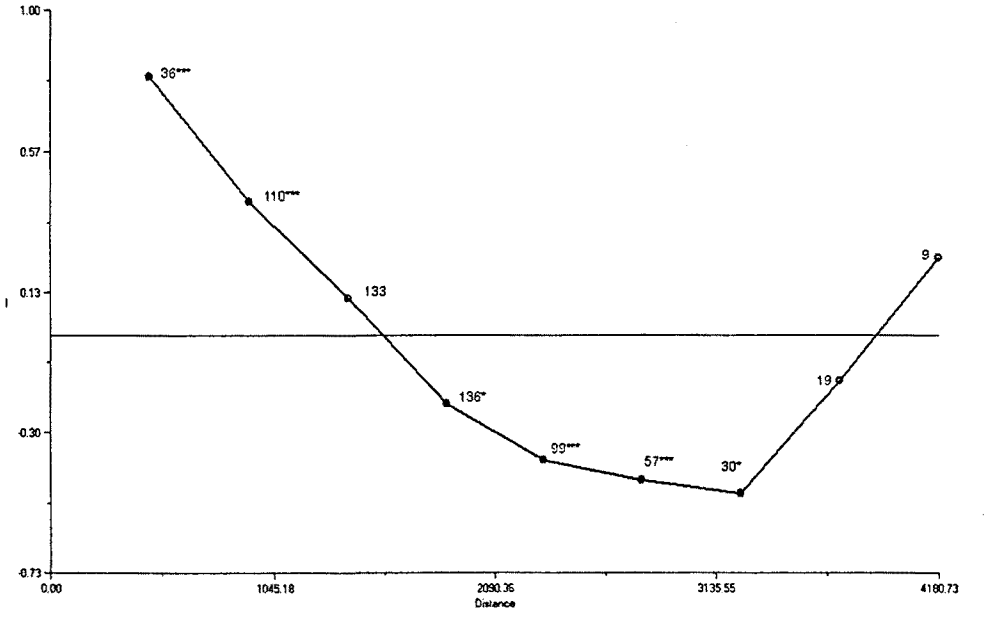
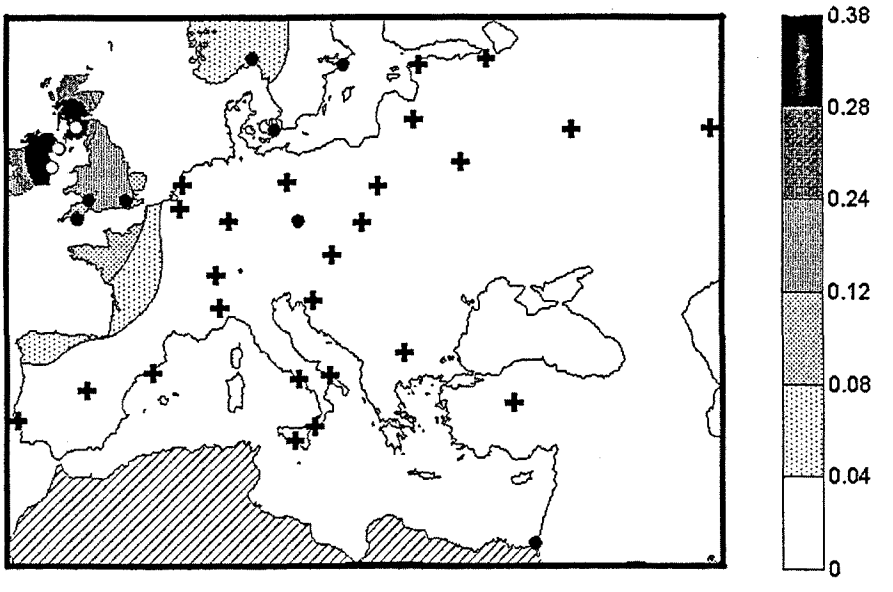

1.00
0.57
0.13
-0.30
-0.73
I
36***
110***
133
136*
99***
57***
30*
19
9
0.00
1045.18
2090.36
3135.55
4180.73
Distance

Figure 2.

Geographical distribution (a) and spatial autocorrelogram (b) of R408W under the haplotype background 1 in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.38
0.28
0.24
0.12
0.08
0.04
0

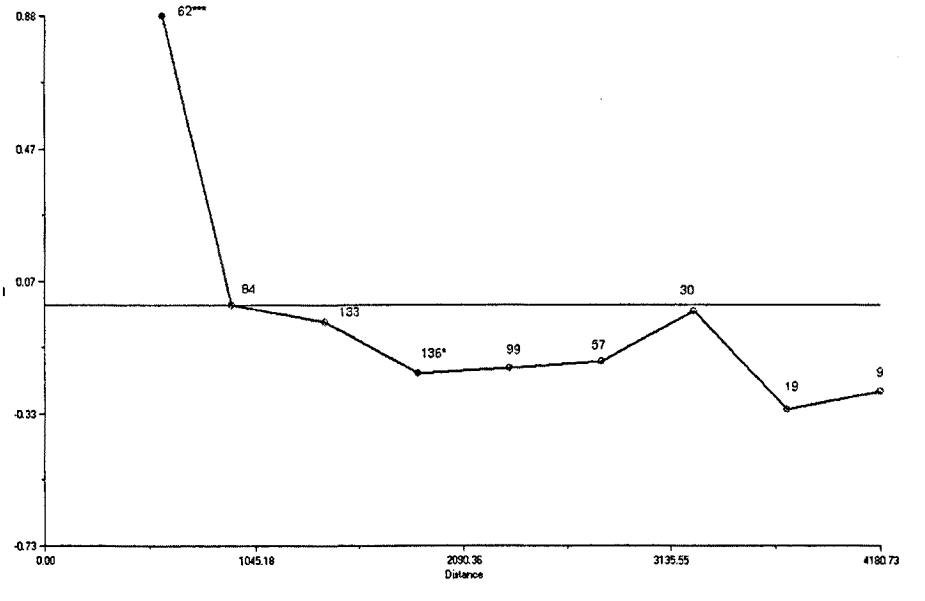
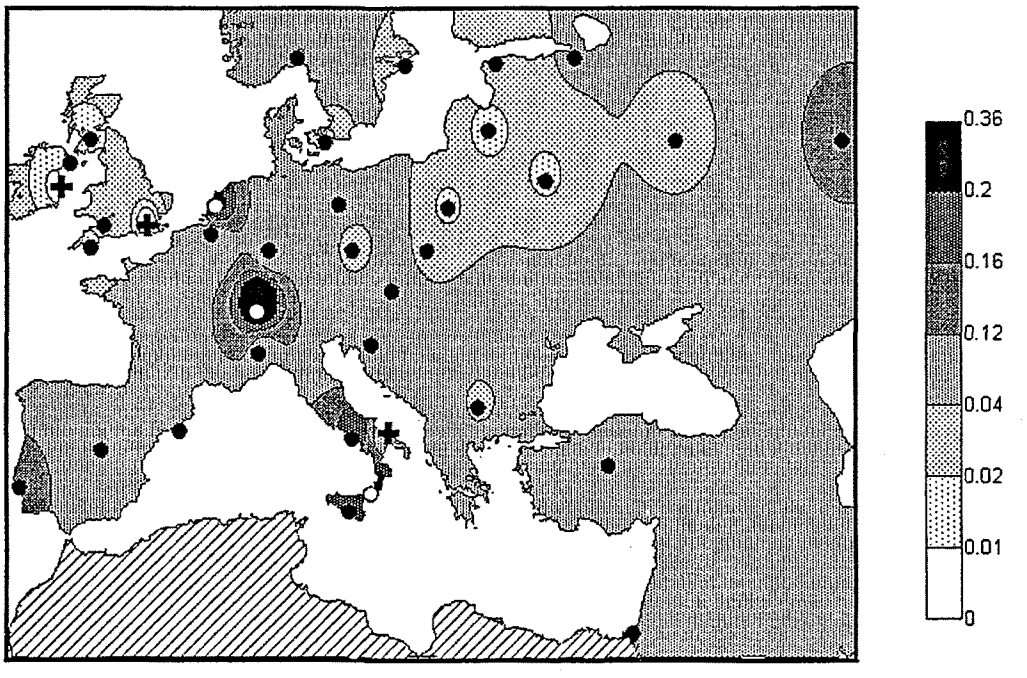

0.88
0.47
0.07
-0.33
-0.73
62***
84
133
136*
99
57
30
19
9
0.00
1045.18
2090.36
3135.55
4180.73
Distance

Figure 3.

Geographical distribution (a) and spatial autocorrelogram (b) of R261Q in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.36
0.2
0.16
0.12
0.04
0.02
0.01
0

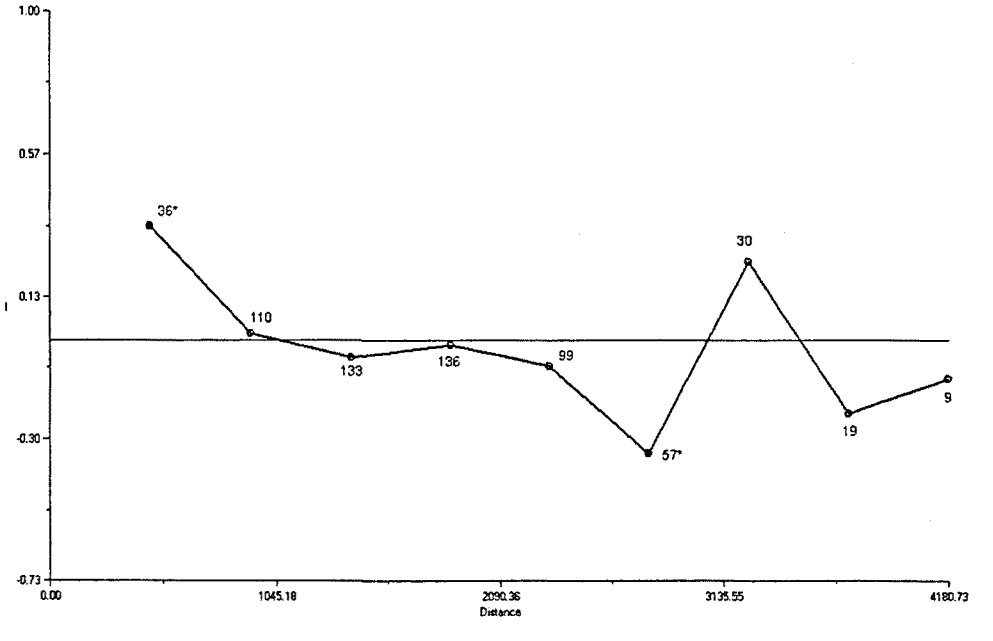
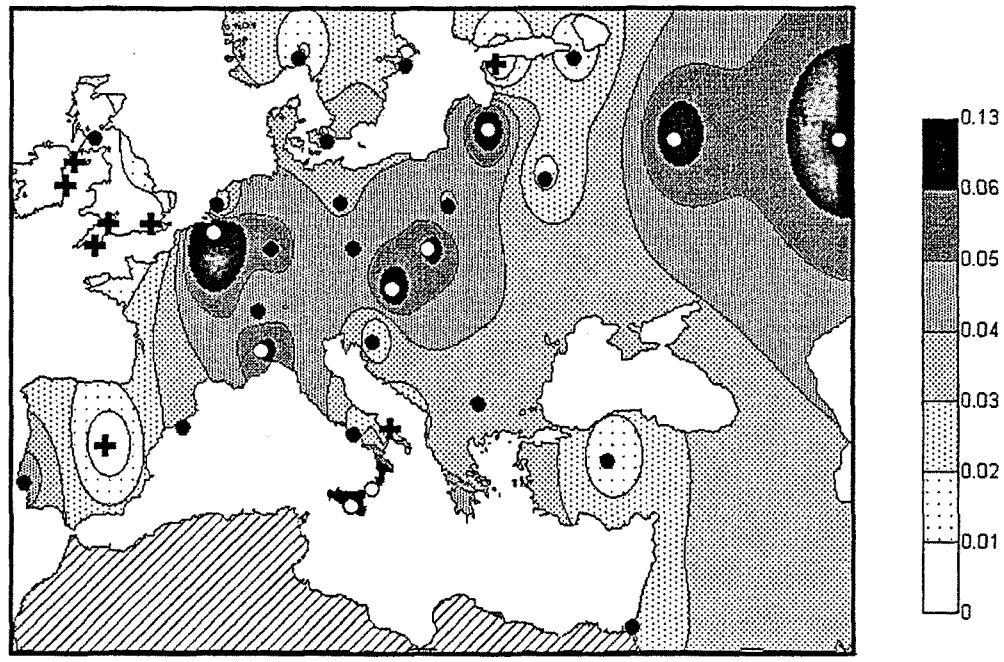

1.00
0.57
0.13
-0.30
-0.73
I
0.00
1045.18
2090.36
3135.55
4180.73
Distance
36*
110
133
136
99
57*
30
19
9

Figure 4.

Geographical distribution (a) and spatial autocorrelogram (b) of R158Q in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.13
0.06
0.05
0.04
0.03
0.02
0.01
0

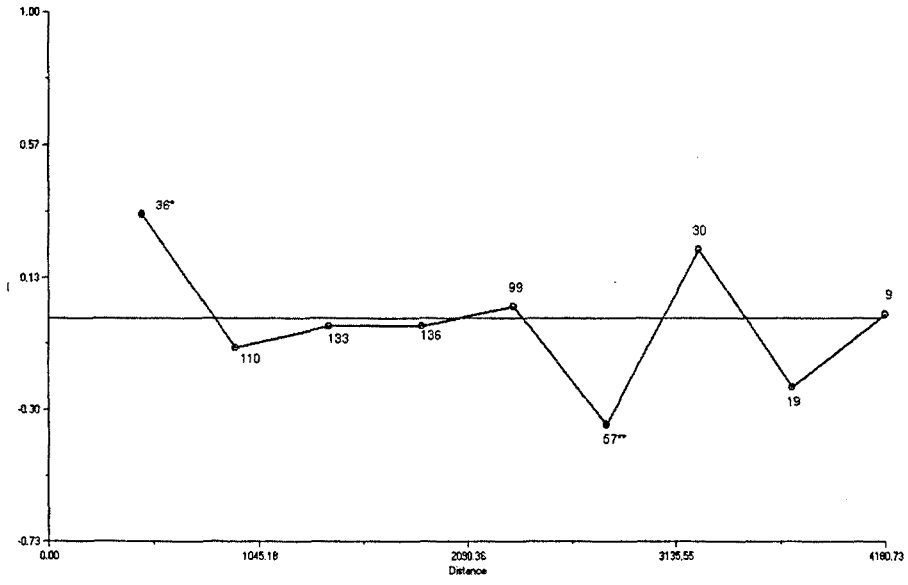
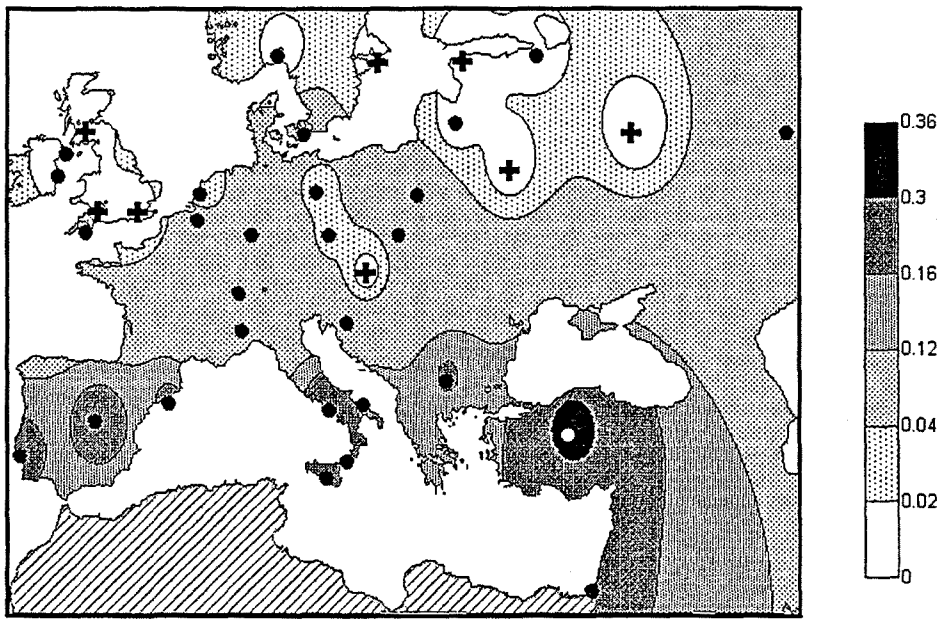

1.00
0.57
0.13
-0.30
-0.73
0.00
1045.18
2090.36
3135.55
4180.73
Distance
36*
110
133
136
99
57**
30
19
9

Figure 5.

Geographical distribution (a) and spatial autocorrelogram (b) of IVS 10-11G>A in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.36
0.3
0.16
0.12
0.04
0.02
0

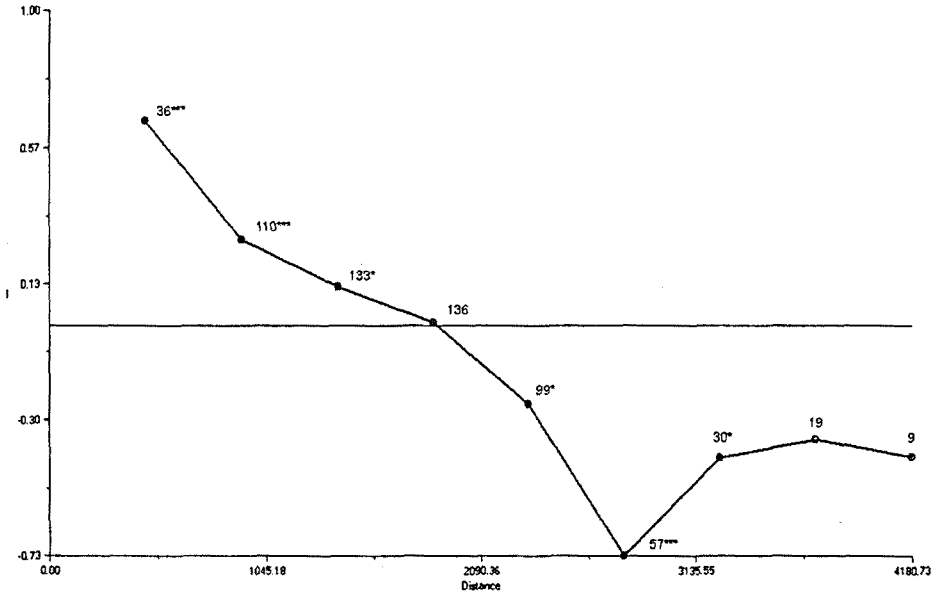
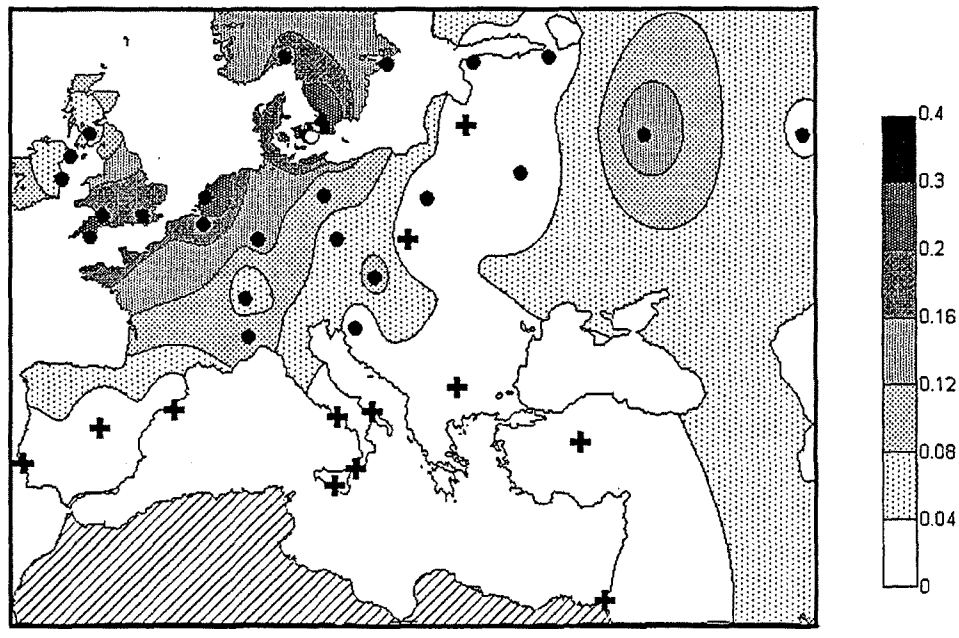

1.00
0.57
0.13
-0.30
-0.73
I
36***
110***
133*
136
99*
30*
19
9
57***
0.00
1045.18
2090.36
3135.55
4180.73
Distance

Figure 6.

Geographical distribution (a) and spatial autocorrelogram (b) of IVS12+1G>A in 36 middle Eastern and European populations. Populations where the mutation is absent are marked with a cross. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<0.05$; double asterisks (**) denote $P<0.01$


0.4
0.3
0.2
0.16
0.12
0.08
0.04
0

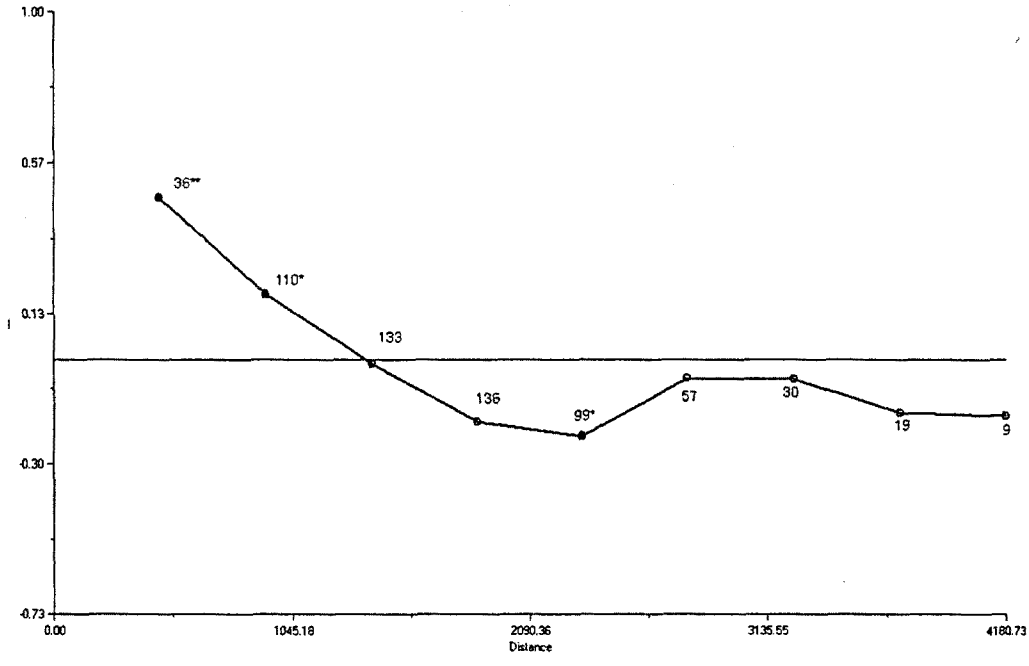
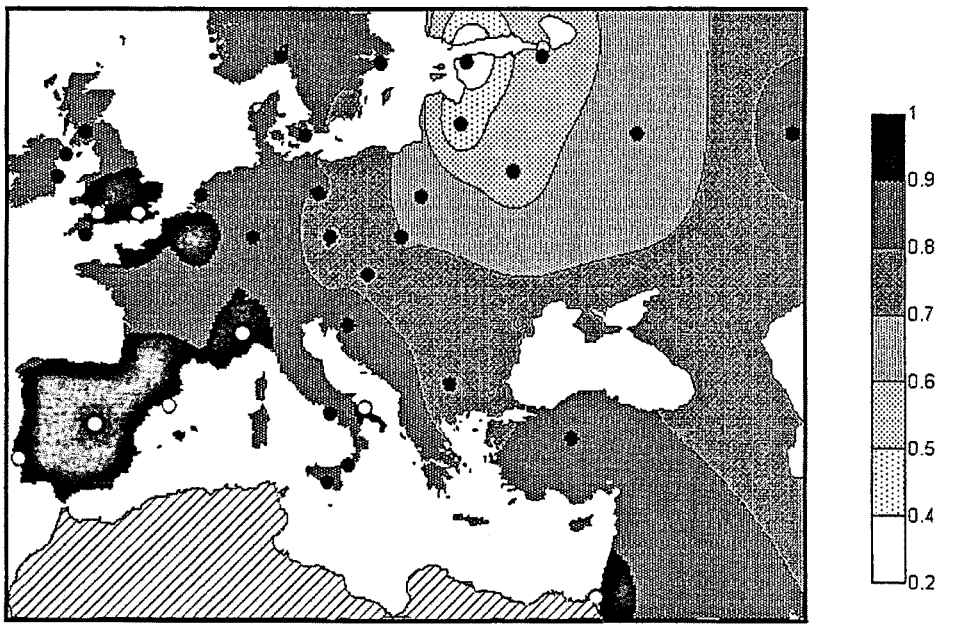

1.00
0.57
0.13
-0.30
-0.73
I
36**
110*
133
136
99*
57
30
19
9
0.00
1045.18
2090.36
3135.55
4180.73
Distance

Figure 7

Geographical distribution (a) and spatial autocorrelogram (b) of the maximum expected heterozygosity (see materials and methods) in 36 middle Eastern and European populations. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. The X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $P<.05$; double asterisks (**) denote $P<.01$


1
0.9
0.8
0.7
0.6
0.5
0.4
0.2

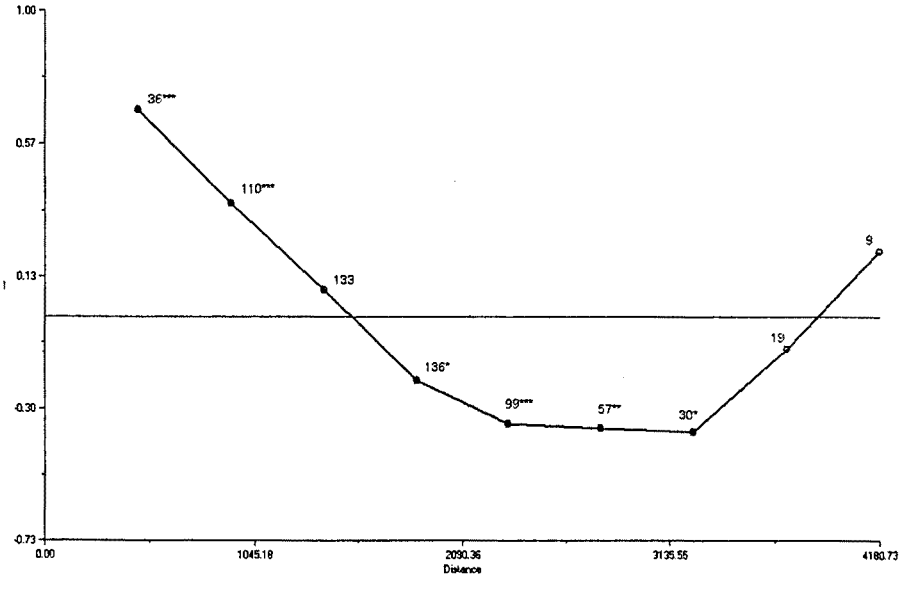

1.00
0.57
0.13
-0.30
-0.73
I
36***
110***
133
136*
99***
57**
30*
19
9
0.00
1045.18
2090.36
3135.55
4180.73
Distance

Figure 8 MDS plot based on a Reynolds' genetic distance matrix of PKU mutations. Populations have been clustered in three main geographical groups: Mediterranean, NW Europe and NE Europe.

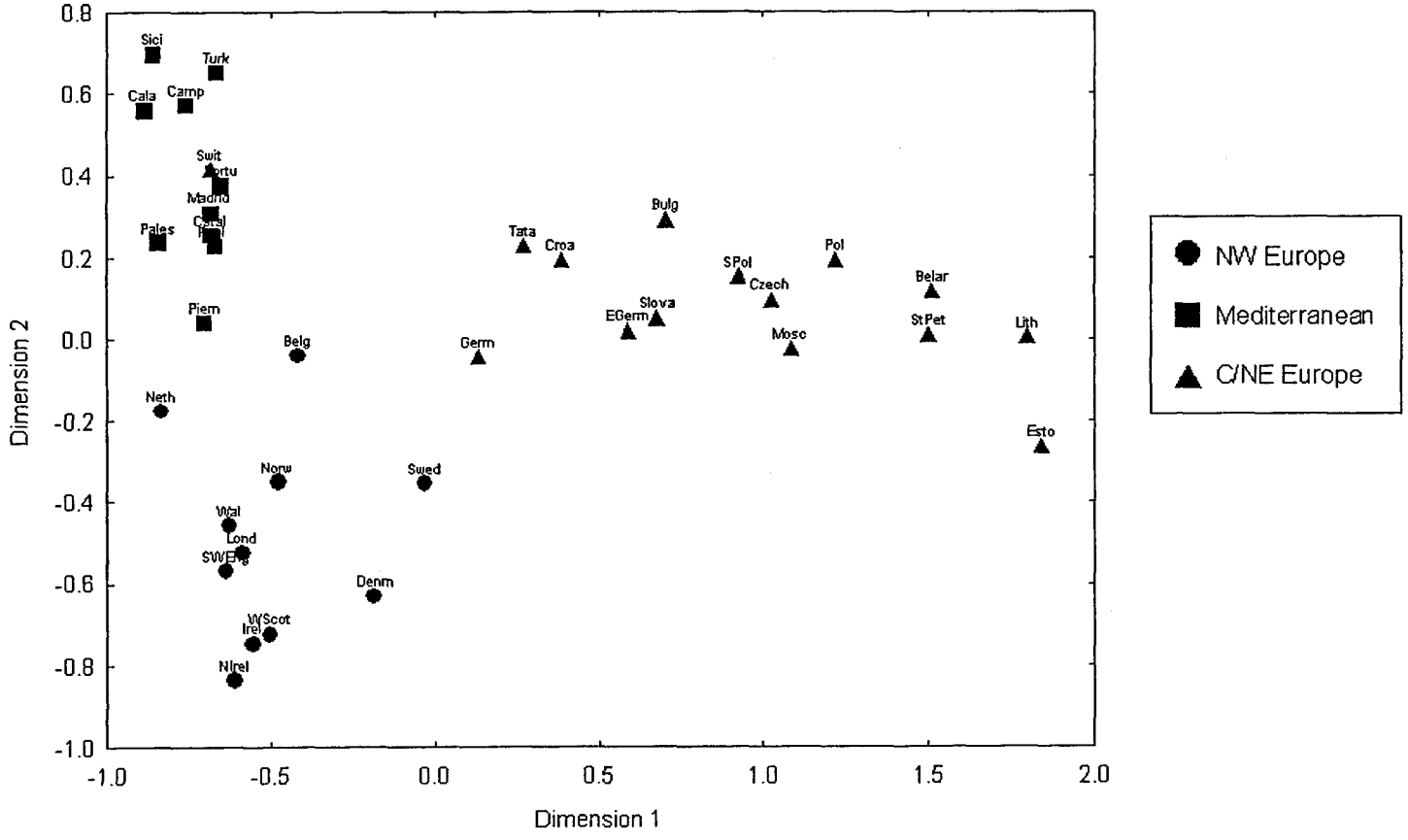
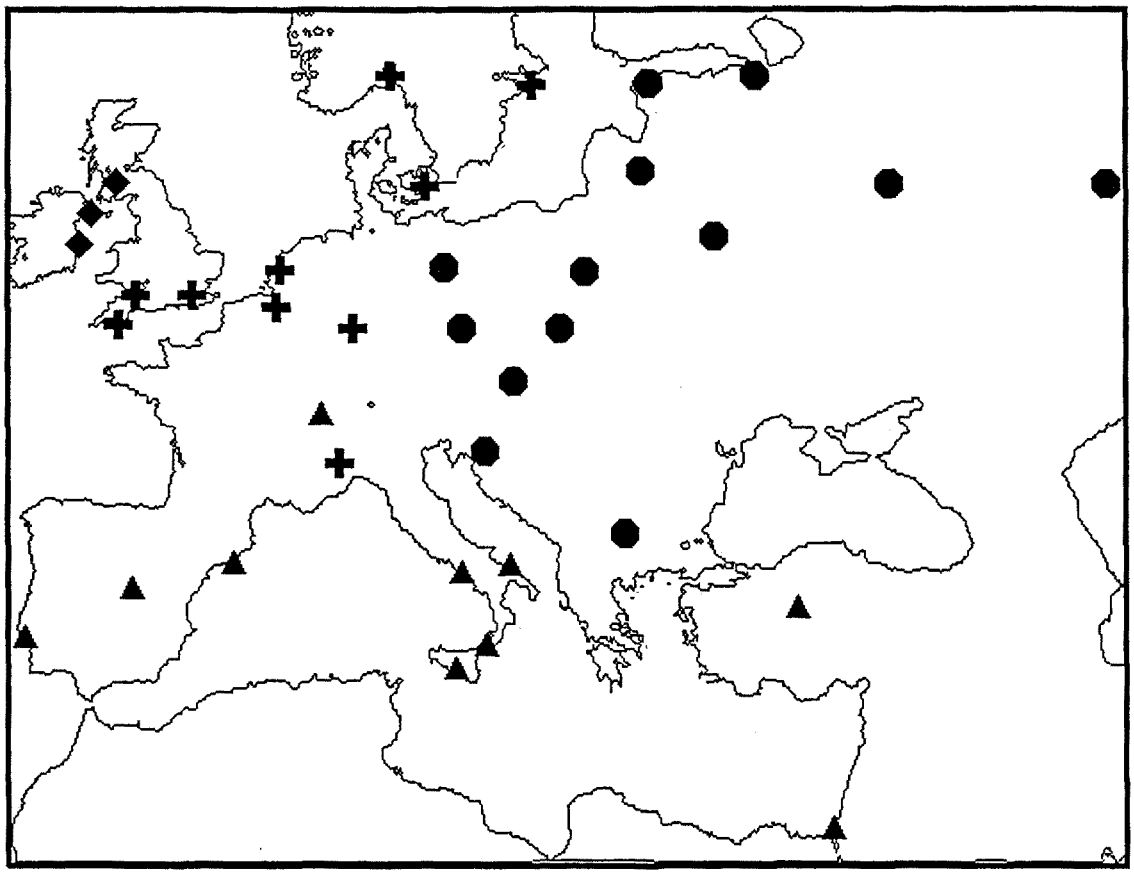

0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Dimension 2
Dimension 1
Sici
Turk
Cata
Camp
Swit
Madrd
Pales
Piem
Belg
Neth
Norw
Swed
Wal
Lond
Denm
WScot
Irel
NIrel
Germ
Tata
Croa
Bulg
EGerm
Slova
SPol
Czech
Pol
Belar
StPet
Mosc
Lith
Esto
NW Europe
Mediterranean
C/NE Europe

Figure 9. Clusters of populations according to the SAMOVA algorithm with 4 groups. Geographical clusters included: ▲ Mediterranean, ● Central-NE Europe, ◆ Ireland and ✚ NW Europe populations.



3.3 *Capítol III: "The spatial pattern of the β-thalassemia mutation spectrum is consistent with the Neolithic spread of malaria"*

Oscar Lao, Isabelle Dupanloup, Guido Barbujani, Jaume Bertranpetit, Francesc Calafell.

(manuscrit en preparació)

The spatial pattern of the β-thalassemia mutation spectrum is consistent with the Neolithic spread of malaria

Oscar Lao[1], Isabelle Dupanloup[3,2], Guido Barbujani[3], Jaume Bertranpetit[1], Francesc Calafell[1]

1 Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

2 Center for Integrative Genomics, Faculté de Biologie et de Médecine, Université de Lausanne, Lausanne, Switzerland

3 Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy

Running tittle :

Keywords: β-thalassemia, malaria, Neolithic spread, spatial distribution, genetic diversity

Correspondence:
Francesc Calafell
Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida
Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Catalonia, Spain
Tel:+34-93-542 28 41
Fax: +34-93-542 28 02
e-mail: francesc.calafell@upf.edu

ABSTRACT

β-thalassemia is produced by mutations at the β-globin gene and is one of the commonest mendelian disorders in world populations. Its geographical distribution is not homogeneous, but is mainly restricted to a strip overlapping with regions that have been traditionally associated with malaria; this has suggested that β-thalassemia, as well as happens with other hemoglobinopathies, could be subjected to processes of balancing selection. Although it is well known that the mutations that produce β-thalassemia have arisen independently, little is known about the geographical pattern of these mutations. Moreover, previous studies performed with other mendelian diseases such as cystic fibrosis or phenylketonuria have shown that the genetic diversity of the mutational spectrum could be used to improve the knowledge of the natural history of the disease. We have analyzed the spatial pattern of the main mutations associated to β-thalassemia as well as the genetic diversity of the mutational spectrum. Our results suggest that the main β-thalassemia mutations present clinal patterns and define three large genetically uniform regions: Mediterranean and W Asia, S Asian and SE Asian regions. The analysis of the genetic diversity of these geographical regions has shown that in case of the Mediterranean and W Asia region the genetic diversity of the alleles associated to β-thalassemia can be explained by demographic rather than selective factors and correlates with the recent Neolithic expansion.

INTRODUCTION

β–thalassemia (OMIM access number: 141900) is one of the commonest monogenic diseases in the world. Haemoglobin is a tetramer with two α and two β globin chains, although other, minor forms, do also exist. The imbalance produced by the reduction or absence of the β-chain protein causes β-thalassemia. Mutations in the HBB gene, which codes for β-globin and maps in 11p15, reducing or abolishing the production of the β chain can thus lead to β-thalassemia. This is a recessive disorder, although a few cases of dominance have been described (Scriver, et al., 2000). The principal phenotypic manifestation is anaemia and, related to the severity of anaemia, other important complications, such as splenomegaly, bone disease, and cardiac damage can also occur (Scriver, et al., 2000). The clinical diversity and the degree of afectation of β–thalassemia is not homogeneous and ranges from severe phenotypes, in which patients require regular blood transfusions, to mild and asymptomatic phenotypes. Over 150 mutations causing β–thalassemia have been described (http://globin.cse.psu.edu/hbvar/menu.html); they can be classified by expression level and by associated phenotype criteria. According to their genic expression, they have been classified as b0 (null expression), b+ and b++ (mild reduction in expression). According to the associated phenotype, mutations are classified as major (severe phenotype), intermedia (intermedium phenotype) and minor (mild and asymptomatic phenotype). Although there is a correlation between both classifications (i.e. b0 mutations are expected to be major, since there is no synthesis of β–chains), this is far from being perfect. The severity is related to the degree of globin-chain imbalance, and depends not only on the mutations the patient carries but also on other genetic and environmental factors (Weatherall, 2001).

The distribution of β–thalassemia is widespread around the world, mainly at high frequencies in North-West Africa, the Mediterranean region, the Middle East, South Asia and South-East Asia (Flint, et al., 1993). Haldane (Haldane, 1990) proposed an association between the high frequency of thalassemias and the presence of endemic malaria around the shores of the Mediterranean Sea. According to the "malarial hypothesis", thalassemias and other haemoglobinopathies (such as the sickle cell trait) provide protection against malaria and are elevated and maintained by natural selection (Carter and Mendis, 2002). Although there has been little firm evidence that this hypothesis applies to haemoglobinopathies other than the sickle cell trait (Flint, et al.,

1998), the presence of few frequent, regionally specific, mutations (Weatherall, 2001), the close association between mutations and haplotypes (Flint, et al., 1998), and results from immunological studies (Smith, et al., 2002) suggest that β–thalassemia mutations can be selected for malarial resistance. The study and analysis of the spatial distribution of mutations and their genetic diversity could help us to understand how the tempo and mode of the selection process act.

We have compiled β–thalassemia mutation frequencies for a wide set of populations from the Mediterranean area and Asia from the literature and have described their spatial patterns, both of a few single mutations and of mutation diversity.

MATERIALS AND METHODS

Database

A β–thalassemia mutation database was compiled from bibliographic data for European, North African, and Asian autochthonous populations. The latter comprised populations from West Asia (from the Levant to Iran), South Asia, South-East Asia, and East Asia (China and Japan), while no samples from Central or North Asia were available. European populations were mostly from Southern Europe, with some Central European samples. A previous compilation (Scriver, et al., 2000) was used as a starting point, which was completed with references published in journals indexed in Medline. The reference search was closed on early June 2002.

Location, mutation frequencies and sample size were recorded for each population sample, as well as additional linguistic information (language spoken, and the branch and family) classified in (Grimes, 1988; Ruhlen, 1987). When it was known, the severity of the disease caused by each mutation was noted, and only those mutations leading to non-minor disease were utilized for further analysis (Ho, et al., 1998; Weatherall, 2001) http://globin.cse.psu.edu/hbvar/menu.html. Frequently, population samples contained affected individuals for whom a mutation could not be found in one or both chromosomes; this fraction of the sample was labelled as "unknown". Since they were ascertained from patients, all unknown mutations were considered as non-minor. After discarding minor mutations, only samples with a number of chromosomes >20 and with a frequency of unknown mutations <60% were considered for study. Thus, the final database contains 89 different population samples adding up to 21,437

chromosomes carrying 65 different mutations. The fraction of unknown mutations is 13%, ranging from 0 in some populations to 29% in Macedonia.

Frequency Maps

Frequency maps were drawn for β–thalassemia mutations found at average frequencies over 4%. Maps were created with Surfer 7.0 (Golden Software Inc) using the inverse distance method for interpolating points. A regular grid covering S Europe, N Africa, and Asia, and limited between 10° W and 140° E and between 7° S and 51° N was used. The interpolated points were spaced 0.1°. No population samples were available for Africa below 20° N and this region was deleted from maps, since the values obtained would be completely artefactual. Interpolated points were only used to map allele frequencies and were not used in other analyses.

Genetic diversity

The diversity of β-thalassemia mutations within each population can be summarized with parameters that are relevant to population genetics models. We have computed the genetic diversity of β-thalassemia mutations as their expected heterozygosity from their relative frequencies in mutated chromosomes. This parameter can be used to estimate θ (Zouros, 1979), a compound parameter that, under certain condition, equals $4N_e\mu$, where N_e is the effective population size and μ is the mutation rate (Reich and Lander, 2001). The calculation of the genetic diversity and of θ depends on the exact specification of the frequencies of all mutations in the sample; then, the fraction of unknown mutations would prevent such calculations unless some assumptions are made. Although the actual genetic diversity would remain unknown, we can bracket it between two extremes. The minimum value that genetic diversity can take is obtained when all unknown mutations are considered as one single mutation, whereas the maximum value is obtained when all unknown mutations are private for each population and different between them. Although we cannot exclude the possibility that a particular unknown mutation has an elevated frequency, it seems more probable that a high fraction of unknown mutations would be rare and seems that this is the usual situation (Colosimo, et al., 2003; Su, et al., 2003). The estimates we present were computed under the maximum genetic diversity criterion, by means of the Arlequin 2.000 package (Schneider, et al., 2000).

Spatial autocorrelation analysis

The spatial patterns of the β–thalassemia mutation frequencies and of the measures of genetic diversity was described by means of spatial autocorrelogration analysis (Sokal and Oden, 1978) using the PASSAGE program (Rosenberg, 2001). In this kind of analysis, the level of autocorrelation (expressed as Moran's I index) between pairs of populations that are grouped in increasing geographic distance classes is calculated. Plotting autocorrelation values in relation to the geographic distance classes can show different shapes that describe the spatial relationship of the data. In the case of a clinal pattern, this plot shows a line decreasing from positive values for the closest distance classes to negative values for longest distance classes. If the clinal pattern radiates from the centre of the area considered, then the longest distance classes will show autocorrelations close to 0 or positive (Barbujani, 2000), since the peripheral values will tend to be similar.

Genetic distances

The pattern of genetic differentiation between populations show by their β-thalassemia mutation spectra was measured by means of Reynolds' genetic distances (Reynolds, et al., 1983) between populations. Only non-minor β–thalassemia mutations were considered. This matrix was compared with those obtained from other genome regions and types of polymorphisms: classical markers (Cavalli-Sforza, et al., 1994), Y chromosome haplogrup frequencies (Underhill, et al., 2000) and mitochondrial DNA (Bamshad, et al., 2001; Koyama, et al., 2002; Plaza, et al., 2003; Richards, et al., 2000; Yao, et al., 2002). It should be noted that this information was available for a subset of the population samples in the β-thalassemia database. Distance matrices obtained from the different genome regions were compared with Mantel tests (Mantel, 1967), and with partial Mantel tests subtracting the effect of the geographic distance among populations. Genetic distances and Mantel tests were calculated with Arlequin 2.000 (Schneider, et al., 2000).

Multidimensional scaling

Multidimensional scaling was used to represent the genetic distance matrix based on β–thalassemia mutations. The result of this analysis is a set of coordinates for each population so that the distances among them are as close as possible to the original genetic distances. Multidimensional scaling produces also a measure of the goodness-

of-fit between the original distance matrix and the multidimesional scaling results; this measure is called stress and it is actually smaller for better fits. The dimensions obtained were interpolated in space using the same methods described for mutation frequencies, and their spatial patterns were investigated by means of spatial autocorrelation.

AMOVA and SAMOVA

The apportionment of the molecular variance of non-minor β–thalassemia mutations between populations or groups of populations was analyzed by means of Analysis of Molecular Variance (AMOVA; (Excoffier, et al., 1992)) with the Arlequin 2.000 package. This approach allowed us to compute the fraction of molecular variance explained among population groups defined a priori. Populations were grouped by language, linguistic branch and linguistic family, as well as by epidemiological zones defined by the presence of different malaria vectors (Hume, et al., 2003). A different approach to the Analysis of Molecular Variance is provided by the Spatial Analysis of Molecular Variance (SAMOVA, (Dupanloup, et al., 2002)). In SAMOVA, populations are first linked by a Delaunay network, which defines which are the geographical neighbours of each population. Then, after specifying a given number of geographical groups, SAMOVA iteratively finds how the populations should be divided in geographically coherent groups in order to maximize the differentiation between groups and minimize it within groups. For instance, if the number of groups is set to three, SAMOVA would draw the barriers that divide the populations into the three groups that are most genetically different among them. We performed SAMOVA setting the number of groups to five and increasing it to ten in steps of one; one hundred iterations were used. In order to compare results between SAMOVA and the preset AMOVA groups SAMOVA was also performed with the number of groups specified by language, linguistic branch, linguistic family, and epidemiological zones of malaria that had been previously used for AMOVA.

RESULTS

The geographic distribution of main β–thalassemia mutations (that is, those found at an average frequency over 4%; see table 1) has been plotted and studied by means of Spatial Autocorrelogram Analysis.

The IVS-I-110 (G->A) mutation (HBB g.202G>A in the standard HUGO nomenclature; average frequency 13.5%) is found in Europe, N Africa and SW Asia as far as Iran, and it showed a clinal pattern (see figure 1) that seems to irradiate from the Eastern Mediterranean (maximum frequency in Cyprus, 83.5%). The codon 39 C->T mutation (HBB g.248C>T; average frequency 13.2%) presented a clinal pattern as shown by the spatial autocorrelogram (see figure 2). It peaks in the Western Mediterreanean (maximum frequency in Sardinia, 95.8%) but can be found at low frequencies in the Levant and Iran. A partial clinal pattern was obtained for the IVS-I-5 (G->C) mutation (HBB g.97G>C; average frequency, 14.1%; see figure 3). Although it is mainly present in South Asia (maximum frequency in Madya Pradesh, 94.7%); figure) it is also found at low frequencies in one Mediterranean population (Algiers), in Middle Eastern and South-East Asian populations. The mutation involving codons 41/42 (-TTCT) (HBB g.254 to 257 del TTCT; average frequency, 9.0%) showed a partial clinal pattern (with distances between pairs of populations up to 11,098 km). It is most frequent in South-East Asia (maximum frequency in Guangxi, 56.5%); figure 4) but it is also present in South Asia, Iran, and Germany. The IVS-I-1 (G->A) mutation (HBB g.93 G>A; average frequency, 8.6%) is mainly present in Mediterranean Europe (maximum frequency in the Spanish region of Alta Extremadura ((59.42%); figure 5) but it is also detected in other populations from the Middle East, South and South-East Asia, with a secondary maximum in Sri Lanka (28%). No spatial pattern has been obtained for the IVS-II-1 (G->A) mutation (HBB g.446G>A; average frequency, 4.2%); IVS-II-1 (G->A) is present in Iran region (with a maximum in North Iran, (52.6%); figure 6) but can also be found at high frequencies in the Middle East, the Mediterranean area, and Japan (with a frequency of 0.15).

Mutation heterozygosity ranged from 0 in Borneo to 0.91 in Southwest Iran; although it did not show any significant spatial pattern (figure 7a), high heterozygosities were observed in the Middle East and South-East Asia, and lower values in Borneo, Sardinia and populations close to them. A similar result was obtained for θ_{Hom} (figure 7b), with elevated values in the Middle Eastern and South-East Asian populations and low in Sardinia and populations from South Asia.

Population mutation spectra were compared by means of Reynolds genetic distances; the dimensionality of the genetic distance matrix obtained was reduced by means of multidimensional scaling analysis. With three dimensions, the stress reached 0.123. The first dimension showed a statistically significant clinal pattern in a West-

East direction (figure 8); the second dimension showed a partial clinal pattern for population distances up to 5220 km and was focused on South and South-East Asia. The third dimension separated Borneo from the rest of populations. A graphical representation of both first and second dimension clustered populations in three main groups in agreement with their geographical localization: Europe/ Middle East/Iran Region, South Asia, and South-East Asia (data not shown). Borneo and Sardinia appeared as separate groups.

This geographic grouping was also recovered with SAMOVA. If the number of groups was preset at five, the partitioning of genetic variance produced three large groups (see table 2): Europe/Middle East/Iran Region (plus three distant populations: Japan, Myanmar, and Java), South Asia and South-East Asia, and two smaller groups: Sardinia and Borneo. The fraction of the genetic variance explained by this grouping is 29.0%. The inclusion of distant populations, creating a topologically disjoint set may seem contrary to what is meant to be achieved with SAMOVA. However, it can be explained by the how a neighbour is defined with Delaunay's network makes and by the low number of groups specified to SAMOVA. Depending on the population topology, peripheral populations can be connected by the Delaunay network to each other even if they are unreasonably distant. If the number of groups is not large enough, and there are other populations with a higher genetic differentiation, then long distant populations may be clustered in the same group. This may explain how to Japan, Myanmar, and Java were grouped with European and W Asian populations; increasing the number of preset groups in SAMOVA led to these population forming each a single group (data not shown). These populations were excluded for further analysis from the Europe/Middle East/Iran Region group for the sake of geographic consistency.

For each SAMOVA group a few predominant, almost group specific, mutations were detected (see table). Thus, for the Europe/ W Asia group, the predominant mutation was IVS-I-110 (G->A) (23.8% of chromosomes in this group carry this mutation versus 6.7% in Sardinia and 0.8% in S Asia), for South Asia IVS-I-5 (G->C) (59.8% of chromosomes in this group versus 3.0% in the Europe/W Asia group and 2.5% in SE Asia), and for SE Asia mutations IVS-II-654 (C->T) (HBB g.100C>T; 19.4% versus 0.04% in Europe/W Asia and 0.7% in S Asia), Codon 17 (A->T) (HBB g.52 A>T; 20.9% in this group versus 0.5% in S Asia) and mutation 41/42 (-TTCT) (41.7% versus 0.4% in Europe/W Asia and 4.7% in S Asia). Mutation codon 39 C->T was at high frequency in the Sardinia group (83.6% versus 18.7% in the Europe/W Asia

and 0.4% in S Asia); finally, all chromosomes from Borneo carried a large deletion originating from position HBB g.–4279, which has not been described elsewhere.

To further describe variation within each SAMOVA group, a multidimensional scaling analysis was performed on the Reynolds' distance matrix computed from the mutation spectra of populations in each main; three dimensions were specified. The computed dimensions, as well as the θ_{Hom} estimates, were then geographically analyzed by spatial autocorrelation. In the case of the Europe/W Asia group the analysis comprised 49 populations (see table 2) and the stress obtained was 0.1. The first dimension showed a statistically significant ($p<0.05$) clinal pattern that was identified in an East-West direction; the second dimension had a statistically significant ($p<0.05$) clinal pattern focused on the middle East for population distances up to 4000 km and the third dimension showed a complex pattern. A clinal pattern ($p<0.05$) for distances between populations up to 6000 km was obtained for θ_{Hom}, which shows elevated values in the East (around the Fertile Crescent) and lower towards the West. The correlation between logarithmic normalized θ_{Hom} and the first dimension was 0.45 ($p<0.05$). No spatial pattern was detected neither for genetic diversity estimates nor for the dimensions computed by multidimensional scaling in the case of the S Asia (16 populations) and SE Asia (17 populations) groups.

Beyond a purely geographical criterion, we explored how other factors contributed to partition the genetic variance of β-thalassemia mutations. As a proxy for population history, we used a three-tiered hierarchical linguistic classification, in 39 languages, 14 linguistic branches, and seven families. The percentage of variation explained by language, branch, and family was respectively 22.9% ($p<0.05$), 18.8% ($p<0.05$), and 6.3% ($p<0.05$) respectively. Although significant, these figures are clearly lower than those obtained by using performing SAMOVA formed with 39 groups (31.79%, $p<0.05$), 14 groups (30.79%, $p<0.05$), and seven groups (30.06 $p<0.05$). Groups of populations clustered by epidemiological malaria zones (7 groups) explained 15.52% ($p<0.05$) of genetical variation.

To the extent that the general demographic history of the populations had a role in shaping the β-thalassemia mutation pools, this should be reflected in other loci. We have compared by means of Mantel tests the genetic distance matrix between populations based on β–thalassemia mutations and with those based on classic markers

(11 populations; data from (Cavalli-Sforza, et al., 1994)), mitochondrial DNA region I polymorphisms (24 populations; data from (Bamshad, et al., 2001; Koyama, et al., 2002; Plaza, et al., 2003; Richards, et al., 2000; Yao, et al., 2002)), and Y chromosome polymorphisms (9 populations; data from (Underhill, et al., 2000)). Correlations were close to zero with all three genome regions; however, scatter plots showed that Sardinia was an outlier, being different from other populations in β-thalassemia mutations but not for the other genome regions (Borneo may had been a similar case, but no matching data were found for the other genome regions). Then, when Sardinia was removed from the analysis, high levels of correlation were obtained for the Y chromosome ($r = 0.6$; $p<0.0005$, and with geographic correction $r= 0.56$; $p<0.0005$) , mtDNA region I ($r = 0.63$; $p<0.0005$, and with geographic correction $r=0.60$; $p<0.0005$), and classic markers ($r= 0.41$; $p<0.05$, and with geographic correction $r= -0.14$; $p>0.05$) after excluding this population.

DISCUSSION

We have constructed a β–thalassemia mutation frequency database from bibliography, taking into account information about the population and the non-minor β–thalassemia mutations. We have gathered information from populations in Europe (mainly Mediterranean), the Middle East, and S and SE Asia. With the exception of Germany, Czechoslovakia, Japan, and populations from the West and centre of the Iberian Peninsula, malaria may have been endemic to all these populations (Cavalli-Sforza, et al., 1994). We have confirmed clinal or partially clinal patterns for the relative frequencies of the main non-minor β–thalassemia mutations, which tend to have restricted but overlapping geographical ranges, thus defining geographical regions characterized by typical β-thalassemia mutation spectra. However, it should be noted that most mutations are also found at low frequencies in populations where they are not considered *typical*. The definition of these regions has been established with objective methods such as MDS and SAMOVA. The main groups found are Europe and W Asia (that is, Middle Eastern populations from Iran to the West), the S Asian subcontinent, and SE Asia. Two smaller regions were outliers, practically fixed for almost private mutations: Sardinia (with mutation codon 39 C->T) and a population from Sabah in the island of Borneo (with a large deletion originating from position HBB g.–4279).

Genetic distances based on β–thalassemia mutation frequencies were compared with distances based on polymorphisms from other genome regions, and large positive correlations were found once outliers such as Sardinia were dropped from the analysis. Thus, the difference in β-thalassemia mutation spectra between populations in Eurasia parallel those found elsewhere in the genome, in regions that are not subjected to the same selective pressures as HBB. All these results support independent origins of β–thalassemia (Flint, et al., 1998; Weatherall, 2001): mutations may have arisen independently in each population and have later spread to other neighbouring populations by migration. This may explain the presence of mutations in populations where malaria has never been endemic (Flint, et al., 1993).

Spatial analysis of the genetic diversity has not shown any clinal pattern when all populations were considered. However, if we restricted the analysis to the main groups, a cline was observed for the Europe/ W Asia group. Mutation diversity was higher in the Middle East and declined towards the West. Two factors can help us in explaining this pattern: selection and demographic history. According to Reich et al (Reich and Lander, 2001), if mutation diversity is measured as θ_{Hom}, it equals $4N_e\mu(1-f_0)$, where f_0 is the incidence of the disease, N_e is the effective population size, and μ is the mutation rate. Since there is no reason to assume that the mutation rate is not equivalent between populations, changes in θ_{Hom} (and hence in mutation diversity) should depend on the incidence of the disease and/or the effective population size. Mutation diversity decreases with incidence; in the case of β-thalassemia, incidence is a function of heterozygote advantage against malaria. Malaria tended to be more prevalent in the Middle East than in the Western Mediterranean: mortality rates in 1900 were one order of magnitude higher (Carter and Mendis, 2002) in the Middle East than in Europe. This would generate a diversity cline in the opposite direction than the one observed. However, Lao et al showed for cystic fibrosis mutations (Lao, et al., 2003) that the effect of each variable has not the same impact on shaping the genetic diversity. Large incidence changes are required to produce changes in θ_{Hom}, whereas population size changes are directly proportional to changes in θ_{Hom}. Moreover, the effect of incidence to lower the diversity of mutations may require a long time span. Given that the spread of malaria seems to coincide with that of farming (Tishkoff, et al., 2001), not sufficient time may have elapsed to deplete β–thalassemia mutation diversity. Thus, even if malaria could have played an important role in the maintenance of the

β–thalassemia incidence, it seems more likely that the genetic diversity of the disease has been mainly shaped by the historical events that have influenced the effective population size. This implies that the study of the genetic diversity from the mutation spectra can give us a way to analyze and interpret the common migratory processes that underlie the mutational distribution. It is going to depend on the particular history of each region and implies that global patterns are not going to be necessarily clinal patterns, but clinal patterns may be found at the regional level. In the case of Europe, the Neolithic expansion have shaped the genetic diversity of neutral markers (Barbujani and Goldstein, 2004) and may account for a higher Ne in the eastern populations and lower in western populations. It began on farming populations from the Fertile Crescent to the european populations of hunter-gatterers and had an important impact in the environment that has been related to the spread of malaria (Flint, et al., 1998). Neolithic transition allowed mosquito populations the possibility of obtain accessible sources of blood in the midst of abundant mosquito-breeding sites (Flint, et al., 1998) and could have helped to the diffusion of the β–thalassemia mutations present at low frequencies before the Neolithic. The Neolithic transition was not an exclusive process of the european continent but started independently in different places of the world (Jobling, et al., 2004); however, the patterns of population expansions and their complexity in each region are not equally known. The absence of clinal patterns in others regions could be explained by the low number of populations used in the analysis and later particular population migrations of each zone.

Thus, we have shown how the analysis of the genetic diversity can be used as a tool to make inferences about the demographic and selective events that have shaped the mutation spectrum of a mendelian disease such as β–thalassemia.

ACKNOWLEDGEMENTS

This work was supported by Dirección General de Investigación Científica y Técnica (Spanish Government) grants BMC2001-0772 and BOS2001-0794. O.L. was supported by a predoctoral fellowship from the Ministerio de Ciencia y Tecnología.

REFERENCES

Adekile AD, Gu LH, Baysal E, Haider MZ, al-Fuzae L, Aboobacker KC, al-Rashied A, Huisman TH. 1994. Molecular characterization of alpha-thalassemia determinants, beta-thalassemia alleles, and beta S haplotypes among Kuwaiti Arabs. Acta Haematol 92(4):176-81.

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A and others. 2001. Genetic evidence on the origins of Indian caste populations. Genome Res 11(6):994-1004.

Barbujani G. 2000. Geographic patterns: how to identify them and why. Hum Biol 72(1):133-53.

Barbujani G, Goldstein DB. 2004. Africans and asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119-50.

Benito A, Villegas A, Perez-Cano R, Bernal R. 1996. Beta-thalassaemia in south-western Spain: high frequency of G-->A (IVS I-1) mutation. Br J Haematol 92(2):336-8.

Bennani C, Bouhass R, Perrin-Pecontal P, Tamouza R, Malou M, Elion J, Trabuchet G, Beldjord C, Benabadji M, Labie D. 1994. Anthropological approach to the heterogeneity of beta-thalassemia mutations in northern Africa. Hum Biol 66(3):369-82.

Boletini E, Svobodova M, Divoky V, Baysal E, Curuk MA, Dimovski AJ, Liang R, Adekile AD, Huisman TH. 1994. Sickle cell anemia, sickle cell beta-thalassemia, and thalassemia major in Albania: characterization of mutations. Hum Genet 93(2):182-7.

Bolufer-Gilabert P, Perez-Sirvent M, Moreno-Miralles I. 1998. Molecular differences in beta-thalassemia between the Spanish Mediterranean area and inland populations. Hemoglobin 22(5-6):529-33.

Bouhass R, Perrin P, Trabuchet G. 1994. The spectrum of beta-thalassemia mutations in the Oran region of Algeria. Hemoglobin 18(3):211-9.

Brown JM, Thein SL, Weatherall DJ, Mar KM. 1992. The spectrum of beta thalassaemia in Burma. Br J Haematol 81(4):574-8.

Carter R, Mendis KN. 2002. Evolutionary and historical aspects of the burden of malaria. Clin Microbiol Rev 15(4):564-94.

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

Colosimo A, Guida V, Scolari A, De Luca A, Palka G, Rigoli L, Meo A, Salpietro DC, Dallapiccola B. 2003. Validation of dHPLC for molecular diagnosis of beta-thalassemia in Southern Italy. Genet Test 7(3):269-75.

Curuk MA, Arpaci A, Attila G, Tuli A, Kilinc Y, Aksoy K, Yuregir GT. 2001. Genetic heterogeneity of beta-thalassemia at Cukurova in southern Turkey. Hemoglobin 25(2):241-5.

Dupanloup I, Schneider S, Excoffier L. 2002. A simulated annealing approach to define the genetic structure of populations. Mol Ecol 11(12):2571-81.

Efremov GD. 1990. Beta-, delta beta-thalassemia and Hb lepore among Yugoslav, Bulgarian, Turkish and Albanian. Haematologica 75 Suppl 5:31-41.

el-Hazmi MA, Warsy AS, al-Swailem AR. 1995. The frequency of 14 beta-thalassemia mutations in the Arab populations. Hemoglobin 19(6):353-60.

el-Kalla S, Mathews AR. 1997. A significant beta-thalassemia heterogeneity in the United Arab Emirates. Hemoglobin 21(3):237-47.

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(2):479-91.

Faustino P, Pacheco P, Loureiro P, Nogueira PJ, Lavinha J. 1999. The geographic pattern of beta-thalassaemia mutations in the Portuguese population. Br J Haematol 107(4):903-4.

Filon D, Oppenheim A, Rachmilewitz EA, Kot R, Truc DB. 2000. Molecular analysis of beta-thalassemia in Vietnam. Hemoglobin 24(2):99-104.

Flatz G, Wilke K, Syagailo YV, Eigel A, Horst J. 1999. Beta-thalassemia in the German population: mediterranean, Asian and novel mutations. Mutations in brief no.228. Online. Hum Mutat 13(3):258.

Flint J, Harding RM, Boyce AJ, Clegg JB. 1998. The population genetics of the haemoglobinopathies. Baillieres Clin Haematol 11(1):1-51.

Flint J, Harding RM, Clegg JB, Boyce AJ. 1993. Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum Genet 91(2):91-117.

Fucharoen S, Winichagoon P. 1997. Hemoglobinopathies in Southeast Asia: molecular biology and clinical medicine. Hemoglobin 21(4):299-319.

Furuumi H, Firdous N, Inoue T, Ohta H, Winichagoon P, Fucharoen S, Fukumaki Y. 1998. Molecular basis of beta-thalassemia in the Maldives. Hemoglobin 22(2):141-51.

Giambona A, Lo Gioco P, Marino M, Abate I, Di Marzo R, Renda M, Di Trapani F, Messana F, Siciliano S, Rigano P. 1995. The great heterogeneity of thalassemia molecular defects in Sicily. Hum Genet 95(5):526-30.

Grimes BFE. 1988. Ethnologue. Dallas: Summer Institute of Linguistics.

Gulesken S, Oren H, Vergin C, Sanli N, Gulen H, Ucar C, Irken G. 2000. Mutational analysis of beta-thalassemia cases from the Aegean region of Turkey using an allele-specific oligonucleotide hybridization technique. Acta Haematol 104(4):181-4.

Haldane JBS. 1990. Disease and evolution. In: Dronamraju K, editor. Selected Genetic Papers of J.B.S. Haldane. New York/London: Garland Publishing.

Ho PJ, Hall GW, Luo LY, Weatherall DJ, Thein SL. 1998. Beta-thalassaemia intermedia: is it possible consistently to predict phenotype from genotype? Br J Haematol 100(1):70-8.

Hume JC, Lyons EJ, Day KP. 2003. Human migration, mosquitoes and the evolution of Plasmodium falciparum. Trends Parasitol 19(3):144-9.

Indrak K, Brabec V, Indrakova J, Chrobak L, Sakalova A, Jarosova M, Cermak J, Fei YJ, Kutlar F, Gu YC and others. 1992. Molecular characterization of beta-thalassemia in Czechoslovakia. Hum Genet 88(4):399-404.

Jobling MA, Hurles ME, Tyler-Smith C. 2004. Human evolutionary genetics. Origins, peoples&disease. New York: Garland Science.

Khan SN, Riazuddin S. 1998. Molecular characterization of beta-thalassemia in Pakistan. Hemoglobin 22(4):333-45.

Koyama H, Iwasa M, Maeno Y, Tsuchimochi T, Isobe I, Seko-Nakamura Y, Monma-Ohtaki J, Matsumoto T, Ogawa S, Sato B and others. 2002. Mitochondrial sequence haplotype in the Japanese population. Forensic Sci Int 125(1):93-6.

Kyriacou K, Al Quobaili F, Pavlou E, Christopoulos G, Ioannou P, Kleanthous M. 2000. Molecular characterization of beta-thalassemia in Syria. Hemoglobin 24(1):1-13.

Lao O, Andres AM, Mateu E, Bertranpetit J, Calafell F. 2003. Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet 11(5):385-94.

Lau YL, Chan LC, Chan YY, Ha SY, Yeung CY, Waye JS, Chui DH. 1997. Prevalence and genotypes of alpha- and beta-thalassemia carriers in Hong Kong -- implications for population screening. N Engl J Med 336(18):1298-301.

Le TH, Pissard S, Pham HV, Lacombe C, Truong DH, Goossens M, Truong DK. 2001. Molecular analysis of beta-thalassemia in South Vietnam. Hemoglobin 25(3):305-9.

Magro S, Santilli E, Mancuso R, Puzzonia P, Consarino C, Morgione S, Galati MC, Fersini G, Madonna G, Brancati C and others. 1995. Spectrum of beta-thalassemia mutations in Calabria: implications for prenatal diagnosis. Am J Hematol 48(2):128-9.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27(2):209-220.

Najmabadi H, Karimi-Nejad R, Sahebjam S, Pourfarzad F, Teimourian S, Sahebjam F, Amirizadeh N, Karimi-Nejad MH. 2001. The beta-thalassemia mutation spectrum in the Iranian population. Hemoglobin 25(3):285-96.

Ohba Y, Hattori Y, Harano T, Harano K, Fukumaki Y, Ideguchi H. 1997. beta-thalassemia mutations in Japanese and Koreans. Hemoglobin 21(2):191-200.

Old JM, Khan SN, Verma I, Fucharoen S, Kleanthous M, Ioannou P, Kotea N, Fisher C, Riazuddin S, Saxena R and others. 2001. A multi-center study in order to further define the molecular basis of beta-thalassemia in Thailand, Pakistan, Sri Lanka, Mauritius, Syria, and India, and to develop a simple molecular diagnostic strategy by amplification refractory mutation system-polymerase chain reaction. Hemoglobin 25(4):397-407.

Perrin P, Bouhassa R, Mselli L, Garguier N, Nigon VM, Bennani C, Labie D, Trabuchet G. 1998. Diversity of sequence haplotypes associated with beta-thalassaemia mutations in Algeria: implications for their origin. Gene 213(1-2):169-77.

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

Rady MS, Baffico M, Khalifa AS, Heshmat NM, el-Moselhy S, Sciarratta GV, Hussein IR, Temtamy SA, Romeo G. 1997. Identification of Mediterranean beta-thalassemia mutations by reverse dot-blot in Italians and Egyptians. Hemoglobin 21(1):59-69.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17(9):502-10.

Reynolds J, Weir B, Cockerham C. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767-779.

Ribeiro ML, Goncalves P, Cunha E, Bento C, Almeida H, Pereira J, Nunez GM, Tamagnini GP. 1997. Genetic heterogeneity of beta-thalassemia in populations of the Iberian Peninsula. Hemoglobin 21(3):261-9.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T and others. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67(5):1251-76.

Ringelhann B, Szelenyi JG, Horanyi M, Svobodova M, Divoky V, Indrak K, Hollan S, Marosi A, Laub M, Huisman TH. 1993. Molecular characterization of beta-thalassemia in Hungary. Hum Genet 92(4):385-7.

Rosenberg M. 2001. Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 1.1. Version 1.1. Tempe, AZ: Department of Biology, Arizona State University.

Ruhlen M. 1987. A Guide to the world's languages. Stanford: Stanford University Press.

Rund D, Cohen T, Filon D, Dowling CE, Warren TC, Barak I, Rachmilewitz E, Kazazian HH, Jr., Oppenheim A. 1991. Evolution of a genetic disease in an ethnic isolate: beta-thalassemia in the Jews of Kurdistan. Proc Natl Acad Sci U S A 88(1):310-4.

Sadiq MF, Eigel A, Horst J. 2001. Spectrum of beta-thalassemia in Jordan: identification of two novel mutations. Am J Hematol 68(1):16-22.

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

Scriver CR, Sly WS, Childs B, Beaudet AL, Valle D, Kinzler KW, Vogelstein B. 2000. The Metabolic and Molecular Bases of Inherited Diseases: McGraw-Hill Professional.

Setianingsih II, Williamson R, Marzuk S, Harahap A, Tamam M, Forrest S. 1998. Molecular Basis of beta-Thalassemia in Indonesia: Application to Prenatal Diagnosis. Mol Diagn 3(1):11-19.

Smith TG, Ayi K, Serghides L, McAllister CD, Kain KC. 2002. Innate immunity to malaria caused by Plasmodium falciparum. Clin Invest Med 25(6):262-72.

Sokal RR, Oden NL. 1978. Spatial autocorrelation in biology 1. Metodology. Biological Journal of the Linnean Society 10:199-228.

Su YN, Lee CN, Hung CC, Chen CA, Cheng WF, Tsao PN, Yu CL, Hsieh FJ. 2003. Rapid detection of beta-globin gene (HBB) mutations coupling heteroduplex and primer-extension analysis by DHPLC. Hum Mutat 22(4):326-36.

Thong MK, Rudzki Z, Hall J, Tan JA, Chan LL, Yap SF. 1999. A single, large deletion accounts for all the beta-globin gene mutations in twenty families from Sabah (North Borneo), Malaysia. Mutation in brief no. 240. Online. Hum Mutat 13(5):413.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J and others. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455-62.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P and others. 2000. Y chromosome sequence variation and the history of human populations. Nat Genet 26(3):358-61.

Valianou MG, Kouvatsi A, Hassapopoulou-Matamis H, Astrinidis A, Triantaphyllidis C. 1999. Heterogeneity of four beta-thalassemia mutations in Greece. Hemoglobin 23(1):79-82.

Varesi L, Vona G, Memmi M, Marongiu F, Ristaldi MS. 2000. Beta-thalassemia mutations in Corsica. Hemoglobin 24(3):239-44.

Vaz FE, Thakur CB, Banerjee MK, Gangal SG. 2000. Distribution of beta-thalassemia mutations in the Indian population referred to a diagnostic center. Hemoglobin 24(3):181-94.

Vetter B, Schwarz C, Kohne E, Kulozik AE. 1997. Beta-thalassaemia in the immigrant and non-immigrant German populations. Br J Haematol 97(2):266-72.

Waye JS, Borys S, Eng B, Patterson M, Chui DH, Badr El-Din OM, Aref MK, Afify Z. 1999. Spectrum of beta-thalassemia mutations in Egypt. Hemoglobin 23(3):255-61.

Weatherall DJ. 2001. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2(4):245-55.

Xu X, Liao C, Liu Z, Huang Y, Zhang J, Li J, Peng Z, Qiu L, Xu Q. 1996. Antenatal screening and fetal diagnosis of beta-thalassemia in a Chinese population: prevalence of the beta-thalassemia trait in the Guangzhou area of China. Hum Genet 98(2):199-202.

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

Yavarian M, Harteveld CL, Batelaan D, Bernini LF, Giordano PC. 2001. Molecular spectrum of beta-thalassemia in the Iranian Province of Hormozgan. Hemoglobin 25(1):35-43.

Zouros E. 1979. Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. Genetics 92(2):623-46.

Country	Population	latitude	longitude	Lenguage	branch	Family	IVS-I-110 (G->A)	codon 39 (C->T)	IVS-I-5 (G->C)	codons 41/42 (-CTTT)	IVS-I-1 (G->A)	IVS-II-1 (G->A)	Others	unknown	Total	Ref	HMax	θHom
Albania	Tirana	41.2	19.5	Albanian	Albanian	Indo-European	0.4783	0.2609	0.0000	0.0000	0.0580	0.0435	0.1159	0.0435	69	(Boletini, et al., 1994; Efremov, 1990)	0.705	1.830209
Algeria	Algiers	36.47	3.03	Arabic (Western Colloquial)	Semitic	Afro-Asiatic	0.3047	0.2813	0.0078	0.0000	0.1875	0.0000	0.2188	0.0000	128	(Bennani, et al., 1994)	0.766	2.555381
Algeria	Annaba	36.54	7.46	Arabic (Western Colloquial)	Semitic	Afro-Asiatic	0.3030	0.2424	0.0000	0.0000	0.0303	0.0000	0.3939	0.0303	33	(Bennani, et al., 1994)	0.7557	2.405108
Algeria	Oran	35.43	-0.43	Arabic (Western Colloquial)	Semitic	Afro-Asiatic	0.1930	0.3772	0.0000	0.0000	0.1404	0.0000	0.2895	0.0000	114	(Bouhass, et al., 1994; Perrin, et al., 1998)	0.7704	2.62369
Azerbaijan	Azerbaijan	40.3	47.3	Azerbaijani	Turkic	Altaic	0.2500	0.0238	0.0119	0.0000	0.0238	0.2500	0.4405	0.0000	84	(Giambona, et al., 1995)	0.8153	3.527731
Bulgaria	Bulgaria	42.41	23.19	Bulgarian	Balto-Slavic	Indo-European	0.2727	0.2545	0.0000	0.0000	0.0364	0.0182	0.3818	0.0364	55	(Efremov, 1990)	0.8411	4.301146
China	E. China	31.1	121.3	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0923	0.3385	0.0000	0.0000	0.4769	0.0923	65	(Xu, et al., 1996)	0.8072	3.331983
China	Fujian	28.01	116.2	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.4375	0.0000	0.0000	0.5625	0.0000	32	(Xu, et al., 1996)	0.6472	1.388297
China	Guangdon	23.3	111.27	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0364	0.5364	0.0000	0.0000	0.4273	0.0000	110	(Xu, et al., 1996)	0.6429	1.361496
China	Guangxi	22.48	108.2	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.5646	0.0000	0.0000	0.4286	0.0068	147	(Xu, et al., 1996)	0.6146	1.201302
China	Guangzhou	23.06	113.16	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.5055	0.0000	0.0000	0.4808	0.0137	364	(Xu, et al., 1996)	0.6555	1.442376
China	Guizhou	25.46	112.43	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.4250	0.0000	0.0000	0.5750	0.0000	40	(Xu, et al., 1996)	0.6449	1.373868
China	Hong Kong	22.15	114.09	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.5115	0.0000	0.0000	0.4046	0.0840	131	(Lau, et al., 1997; Xu, et al., 1996)	0.6499	1.405674
China	Hubei	30.36	114.17	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.3429	0.0000	0.0000	0.6571	0.0000	35	(Xu, et al., 1996)	0.6538	1.430765
China	Sichuan	30.39	104.04	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.2881	0.0000	0.0000	0.5932	0.1186	59	(Xu, et al., 1996)	0.7382	2.179065
Cyprus	Nicosia	35.1	33.22	Greek	Greek	Indo-European	0.8347	0.0260	0.0000	0.0000	0.0430	0.0000	0.0623	0.0340	883	(Kyriacou, et al., 2000; Varesi, et al., 2000)	0.2976	0.315617
Czechoslovakia	Praha	50.05	14.26	Czech	Balto-Slavic	Indo-European	0.0667	0.0267	0.0000	0.0000	0.5600	0.1733	0.1467	0.0267	75	(Indrak, et al., 1992)	0.6479	1.393038
Eqypt	Cairo	30.03	31.15	Arabic (Egyptian)	Semitic	Afro-Asiatic	0.4022	0.0181	0.0000	0.0000	0.1377	0.0362	0.2572	0.1486	276	(Waye, et al., 1999)	0.8027	3.229531
France	Corsica	42	9	Corsican	Italic	Indo-European	0.0940	0.7949	0.0000	0.0000	0.0000	0.0000	0.0342	0.0769	117	(Varesi, et al., 2000)	0.3606	0.419877
Germany	Germany	48.08	11.34	German	Germanic	Indo-European	0.2489	0.4346	0.0253	0.0042	0.0844	0.0127	0.1224	0.0675	237	(Flatz, et al., 1999; Vetter, et al., 1997)	0.7414	2.218325
Greece	Greece	37.58	23.43	Greek	Greek	Indo-European	0.4691	0.1960	0.0000	0.0000	0.1384	0.0038	0.0071	0.1857	1842	(Boletini, et al., 1994; Valianou, et al., 1999; Vetter, et al., 1997)	0.7227	2.004689
Hungary	Hungary	47.3	19.05	Hungarian	Finno-Ugric	Uralic	0.0000	0.3704	0.0000	0.0000	0.3333	0.0741	0.2222	0.0000	27	(Ringelhann, et al., 1993)	0.7493	2.318635
India	Gujrat	23.02	72.37	Gujarati	Indo-Iranian	Indo-European	0.0000	0.0000	0.4938	0.0848	0.0000	0.0025	0.4065	0.0125	401	(Vaz, et al., 2000)	0.7055	1.834904
India	Madhya Pradesh	23	79	Hindi	Indo-Iranian	Indo-European	0.0000	0.0000	0.9474	0.0263	0.0000	0.0000	0.0263	0.0000	38	(Vaz, et al., 2000)	0.1038	0.086639
India	Maharashtra	18.58	72.5	Marathi	Indo-Iranian	Indo-European	0.0000	0.0000	0.7629	0.0043	0.0043	0.0000	0.1681	0.0603	232	(Vaz, et al., 2000)	0.4072	0.511533
India	Punjab	30.44	76.55	Panjabi	Indo-Iranian	Indo-European	0.0000	0.0000	0.1618	0.0588	0.0000	0.0000	0.7500	0.0294	68	(Vaz, et al., 2000)	0.7906	2.979282

Country	Population	latitude	longitude	Lenguage	branch	Family	IVS-I-110 (G->A)	codon 39 (C->T)	IVS-I-5 (G->C)	codons 41/42 (-CTTT)	IVS-I-1 (G->A)	IVS-II-1 (G->A)	Others	unknown	Total	Ref	HMax	θHom
India	Rajasthan	27	74	Panjabi	Indo-Iranian	Indo-European	0.0000	0.0000	0.7021	0.1064	0.0213	0.0000	0.1489	0.0213	47	(Vaz, et al., 2000)	0.4958	0.734106
India	Sindh	25.3	69	Sindhi	Indo-Iranian	Indo-European	0.0000	0.0000	0.0554	0.0215	0.0000	0.0000	0.9138	0.0092	325	(Vaz, et al., 2000)	0.6679	1.52791
India	Uttar Pradesh	27	80	Hindi	Indo-Iranian	Indo-European	0.0000	0.0000	0.3721	0.1163	0.0233	0.0000	0.4651	0.0233	43	(Vaz, et al., 2000)	0.814	3.494687
India	West Bengal	22.32	88.22	Bengali	Indo-Iranian	Indo-European	0.0000	0.0000	0.5500	0.0500	0.0000	0.0000	0.1500	0.2500	20	(Vaz, et al., 2000)	0.7105	1.882128
Indonesia	Indonesia	-6.1	106.48	Javanese	Austronesian	Austric	0.0000	0.0000	0.5664	0.0177	0.0177	0.0000	0.3097	0.0885	113	(Fucharoen and Winichagoon, 1997; Setianingsih, et al., 1998)	0.6602	1.474007
Indonesia	Java	-6.58	110.25	Javanese	Austronesian	Austric	0.0000	0.0000	0.3051	0.0169	0.0000	0.0000	0.3898	0.2881	59	(Setianingsih, et al., 1998)	0.8796	6.128308
Iran	Boushehr	28.59	50.5	Farsi (Persian)	Indo-Iranian	Indo-European	0.0700	0.0406	0.0790	0.0023	0.0677	0.1625	0.4086	0.1693	443	(Najmabadi, et al., 2001; Yavarian, et al., 2001)	0.9188	9.91153
Iran	Central (Yazd)	31.53	54.25	Farsi (Persian)	Indo-Iranian	Indo-European	0.0520	0.0280	0.0960	0.0000	0.0480	0.2880	0.3520	0.1360	250	(Najmabadi, et al., 2001; Yavarian, et al., 2001)	0.8808	6.203485
Iran	Fars	29.36	52.32	Farsi (Persian)	Indo-Iranian	Indo-European	0.0952	0.0476	0.0000	0.0000	0.0476	0.3810	0.3333	0.0952	21	(Yavarian, et al., 2001)	0.8381	4.198205
Iran	Homozgan	27.5	56	Farsi (Persian)	Indo-Iranian	Indo-European	0.0060	0.0210	0.6637	0.0120	0.0210	0.0931	0.1291	0.0541	333	(Najmabadi, et al., 2001; Yavarian, et al., 2001)	0.5485	0.90962
Iran	Khouzestan	31.19	48.42	Farsi (Persian)	Indo-Iranian	Indo-European	0.2112	0.0000	0.0905	0.0000	0.2328	0.1853	0.2155	0.0647	232	(Yavarian, et al., 2001){Najmabadi, 2001 #57	0.8225	3.72049
Iran	North-Tehran	35.4	51.26	Farsi (Persian)	Indo-Iranian	Indo-European	0.0631	0.0017	0.0787	0.0017	0.0545	0.5264	0.1919	0.0821	1157	(Yavarian, et al., 2001){Najmabadi, 2001 #57	0.6984	1.770475
Iran	Sistan	30.3	62	Farsi (Persian)	Indo-Iranian	Indo-European	0.1056	0.0056	0.6444	0.0000	0.0778	0.0000	0.0778	0.0889	180	(Yavarian, et al., 2001){Najmabadi, 2001 #57	0.5678	0.985191
Israel	Gaza	31.3	34.28	Arabic	Semitic	Afro-Asiatic	0.4110	0.1233	0.0000	0.0000	0.2192	0.0137	0.1781	0.0548	73	(Kyriacou, et al., 2000)	0.7622	2.498106
Israel	Israel	31.46	35.14	Hebrew	Semitic	Afro-Asiatic	0.2684	0.1316	0.0177	0.0000	0.0633	0.0861	0.2734	0.1595	395	(Kyriacou, et al., 2000)	0.8833	6.373392
Italy	Calabria	39	16.3	Neapolitan-Calabrese	Italic	Indo-European	0.2406	0.4340	0.0000	0.0000	0.0849	0.0566	0.1085	0.0755	212	(Magro, et al., 1995)	0.7409	2.212749
Italy	Rome	41.58	12.4	Italian	Italic	Indo-European	0.2424	0.5030	0.0061	0.0000	0.0727	0.0303	0.1273	0.0182	165	(Rady, et al., 1997; Vetter, et al., 1997)	0.6805	1.621751
Italy	Sicily	37.3	14	Italian	Italic	Indo-European	0.3113	0.4511	0.0130	0.0000	0.1109	0.0186	0.0764	0.0186	1073	(Giambona, et al., 1995)	0.6865	1.669217
Japan	Tokyo	35.41	139.45	Japanese	Korean-Japanese	Altaic	0.0000	0.0053	0.0000	0.1005	0.0000	0.1534	0.7407	0.0000	189	(Ohba, et al., 1997)	0.866	5.356934
Jordania	Jordan	31.57	35.56	Arabic	Semitic	Afro-Asiatic	0.2827	0.0387	0.0238	0.0000	0.0952	0.1696	0.3185	0.0714	336	(el-Hazmi, et al., 1995; Kyriacou, et al., 2000; Sadiq, et al., 2001)	0.8563	4.896404
Kurdistan	Kurdistan (Jews)	37	45	Kurmanji	Indo-Iranian	Indo-European	0.1739	0.1739	0.0000	0.0000	0.0000	0.0435	0.5652	0.0435	46	(Rund, et al., 1991)	0.6773	1.597129

Country	Population	latitude	longitude	Lenguage	branch	Family	IVS-I-110 (G->A)	codon 39 (C->T)	IVS-I-5 (G->C)	codons 41/42 (-CTTT)	IVS-I-1 (G->A)	IVS-II-1 (G->A)	Others	unknown	Total	Ref	HMax	θHom
Kuwait	Kuwaiti (Saudi)	29.3	47.45	Arabic (Estándar)	Semitic	Afro-Asiatic	0.0000	0.0233	0.1628	0.0000	0.1163	0.4419	0.2558	0.0000	43	(Adekile, et al., 1994)	0.7453	2.266998
Lebanon	Lebanon	33.53	35.3	Arabic	Semitic	Afro-Asiatic	0.4706	0.0074	0.0221	0.0000	0.1397	0.0735	0.1324	0.1544	136	(el-Hazmi, et al., 1995; Kyriacou, et al., 2000)	0.7535	2.374862
Macedonia	Macedonia	41	23	Macedonian	Balto-Slavic	Indo-European	0.4929	0.0793	0.0000	0.0000	0.1105	0.0085	0.0142	0.2946	353	(Boletini, et al., 1994; Valianou, et al., 1999)	0.7395	2.195401
Malaysia	Malaysia	3.1	101.42	Malay	Austro-Tai	Austric	0.0000	0.0000	0.4839	0.0753	0.0000	0.0000	0.3763	0.0645	93	(Fucharoen and Winichagoon, 1997)	0.737	2.165502
Maldives	Maldives	3.15	73	Maldivian	Indo-Iranian	Indo-European	0.0000	0.0000	0.7436	0.0128	0.0897	0.0000	0.1538	0.0000	78	(Furuumi, et al., 1998)	0.4206	0.540665
Myanmar	Myanmar	16.46	96.15	Burmese	Tibeto-burman	Sino-tibetan	0.0000	0.0000	0.2462	0.2154	0.0000	0.0000	0.4154	0.1231	130	(Brown, et al., 1992; Fucharoen and Winichagoon, 1997)	0.7977	3.123014
Pakistan	Balochi	30.12	67	Baluchi	Indo-Iranian	Indo-European	0.0000	0.0000	0.7083	0.0104	0.0000	0.0000	0.2813	0.0000	96	(Khan and Riazuddin, 1998)	0.4886	0.712931
Pakistan	Gujrati and Memon	26.15	68.25	Gujarati	Indo-Iranian	Indo-European	0.0000	0.0000	0.1600	0.0100	0.0600	0.0000	0.7600	0.0100	100	(Khan and Riazuddin, 1998)	0.7251	2.030132
Pakistan	Pashtoon	35.26	72.36	Pashto	Indo-Iranian	Indo-European	0.0000	0.0147	0.1471	0.0441	0.0441	0.0000	0.7500	0.0000	68	(Khan and Riazuddin, 1998)	0.6721	1.55803
Pakistan	Punjabi	31	72	Panjabi	Indo-Iranian	Indo-European	0.0000	0.0000	0.2913	0.0291	0.0097	0.0194	0.6214	0.0291	103	(Khan and Riazuddin, 1998)	0.7432	2.240556
Pakistan	Sindhi	28.25	70.18	Sindhi	Indo-Iranian	Indo-European	0.0000	0.0000	0.4113	0.0323	0.0000	0.0000	0.5484	0.0081	124	(Khan and Riazuddin, 1998)	0.7771	2.733756
Pakistan	Urdu Speaking	27.48	66.37	Urdu	Indo-Iranian	Indo-European	0.0000	0.0000	0.5306	0.0408	0.0000	0.0204	0.4082	0.0000	98	(Khan and Riazuddin, 1998)	0.6789	1.609748
Portugal	Algarve	37.1	-8.15	Portuguese	Italic	Indo-European	0.0909	0.5152	0.0000	0.0000	0.3030	0.0000	0.0000	0.0909	33	(Benito, et al., 1996)	0.6515	1.415997
Portugal	Centre	38.43	-9.08	Portuguese	Italic	Indo-European	0.1032	0.4387	0.0000	0.0000	0.2452	0.0000	0.1742	0.0387	155	(Faustino, et al., 1999)	0.7108	1.884446
Portugal	North	41.11	-8.36	Portuguese	Italic	Indo-European	0.1392	0.2975	0.0000	0.0000	0.1899	0.0000	0.3291	0.0443	158	(Faustino, et al., 1999)	0.7745	2.690403
Portugal	South	38.01	-7.52	Portuguese	Italic	Indo-European	0.1192	0.4636	0.0000	0.0000	0.3775	0.0000	0.0199	0.0199	151	(Faustino, et al., 1999)	0.6321	1.297159
Sabah	Borneo	5.2	117.1	Javanese	Austronesian	Austric	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	80	(Thong, et al., 1999)	0	0
Sardinia	Sardinia	40	9	Sardinian	Italic	Indo-European	0.0049	0.9580	0.0000	0.0000	0.0000	0.0000	0.0250	0.0121	2881	(Varesi, et al., 2000)	0.0818	0.066627
Saudi Arabia	Saudi Arabia	24.38	46.43	Arabic (Estándar)	Semitic	Afro-Asiatic	0.3125	0.1500	0.0000	0.0000	0.0000	0.1500	0.2125	0.1750	80	(el-Hazmi, et al., 1995)	0.8405	4.281111
Singapore	Singapore (Malay)	1.16	103.51	Malay	Austro-Tai	Austric	0.0000	0.0000	0.4286	0.1905	0.0000	0.0000	0.1905	0.1905	21	(Fucharoen and Winichagoon, 1997)	0.8	3.17103
Spain	Alta Extremadura	38.53	-6.58	Spanish	Italic	Indo-European	0.1449	0.2029	0.0000	0.0000	0.5942	0.0000	0.0580	0.0000	69	(Ribeiro, et al., 1997)	0.5921	1.090691
Spain	Barcelona	41.23	2.11	Catalonian	Italic	Indo-European	0.1020	0.7551	0.0000	0.0000	0.0408	0.0000	0.0000	0.1020	49	(Benito, et al., 1996)	0.4243	0.54905
Spain	Cáceres	39.29	-6.22	Spanish	Italic	Indo-European	0.0959	0.1370	0.0000	0.0000	0.4932	0.0000	0.0000	0.2740	73	(Benito, et al., 1996)	0.7352	2.14353
Spain	Granada	37.13	-3.41	Spanish	Italic	Indo-European	0.2927	0.3415	0.0000	0.0000	0.2439	0.0000	0.0000	0.1220	41	(Benito, et al., 1996)	0.7537	2.37722
Spain	Madrid	40.24	-3.41	Spanish	Italic	Indo-European	0.1007	0.3381	0.0000	0.0000	0.3525	0.0000	0.0863	0.1223	139	(Benito, et al., 1996; Bolufer-Gilabert, et al., 1998)	0.7534	2.373893

Country	Population	latitude	longitude	Lenguage	branch	Family	IVS-I-110 (G->A)	codon 39 (C->T)	IVS-I-5 (G->C)	codons 41/42 (-CTTT)	IVS-I-1 (G->A)	IVS-II-1 (G->A)	Others	unknown	Total	Ref	HMax	θHom
Spain	Mallorca	39.34	2.39	Catalonian	Italic	Indo-European	0.2449	0.4592	0.0000	0.0000	0.0306	0.0000	0.2041	0.0612	98	(Bolufer-Gilabert, et al., 1998)	0.693	1.723854
Spain	Valencia	39.3	-0.45	Catalonian	Italic	Indo-European	0.0556	0.5278	0.0000	0.0000	0.1111	0.0000	0.2778	0.0278	36	(Bolufer-Gilabert, et al., 1998)	0.6714	1.553233
Sri Lanka	Sri Lanka	6.56	79.51	Sinhala	Indo-Iranian	Indo-European	0.0000	0.0000	0.5740	0.0201	0.2761	0.0000	0.0622	0.0676	547	(Old, et al., 2001)	0.594	1.099705
Syria	Syria	33.3	36.18	Arabic	Semitic	Afro-Asiatic	0.2994	0.0778	0.0000	0.0000	0.1796	0.0419	0.1976	0.2036	167	(el-Hazmi, et al., 1995; Kyriacou, et al., 2000)	0.8654	5.323684
Taiwan	Taiwan	25.03	121.3	Mandarin	Sinitic	Sino-tibetan	0.0000	0.0000	0.0000	0.3381	0.0000	0.0000	0.6619	0.0000	139	(Xu, et al., 1996)	0.6119	1.187585
Thailand	Central region	16.5	100.15	Tai	Austro-Tai	Austric	0.0000	0.0000	0.0482	0.4699	0.0000	0.0000	0.4247	0.0572	332	(Fucharoen and Winichagoon, 1997)	0.7326	2.114472
Thailand	North region	18.47	100.47	Tai	Austro-Tai	Austric	0.0000	0.0000	0.0183	0.4128	0.0000	0.0000	0.4312	0.1376	109	(Fucharoen and Winichagoon, 1997)	0.6634	1.49608
Thailand	Northeast region	17.52	102.44	Tai	Austro-Tai	Austric	0.0000	0.0000	0.0000	0.3800	0.0000	0.0000	0.5600	0.0600	50	(Fucharoen and Winichagoon, 1997)	0.7657	2.551115
Thailand	South region	13.45	100.31	Tai	Austro-Tai	Austric	0.0000	0.0000	0.2038	0.3269	0.0077	0.0000	0.4308	0.0308	260	(Fucharoen and Winichagoon, 1997)	0.805	3.281693
Tunisia	Tunisia	36.48	10.11	Arabic (Western Colloquial)	Semitic	Afro-Asiatic	0.1810	0.4095	0.0000	0.0000	0.0286	0.0000	0.1714	0.2095	105	(Giambona, et al., 1995)	0.7965	3.097768
Turkey	Aegean	41.01	28.58	Turkish	Turkic	Altaic	0.5054	0.0380	0.0000	0.0000	0.2283	0.0272	0.1196	0.0815	184	(Gulesken, et al., 2000)	0.6837	1.646694
Turkey	Çukurova	37.15	43.37	Turkish	Turkic	Altaic	0.6454	0.0722	0.0085	0.0000	0.0934	0.0382	0.1338	0.0085	471	(Curuk, et al., 2001)	0.5638	0.968584
Turkey	Istanbul	39.5	32.5	Turkish	Turkic	Altaic	0.5158	0.0532	0.0113	0.0000	0.0622	0.0628	0.1821	0.1126	1768	(Gulesken, et al., 2000; Kyriacou, et al., 2000; Vetter, et al., 1997)	0.7154	1.930136
UAE	UAE	24.28	54.23	Arabic (Estándar)	Semitic	Afro-Asiatic	0.0209	0.0418	0.5774	0.0000	0.0000	0.0377	0.3222	0.0000	239	(el-Kalla and Mathews, 1997)	0.6466	1.384897
Vietnam	North region	21.02	105.51	Vietnamese	Austroasiatic	Austric	0.0000	0.0000	0.0000	0.3448	0.0000	0.0000	0.6552	0.0000	29	(Filon, et al., 2000)	0.6502	1.407819
Vietnam	South region	10.02	105.47	Vietnamese	Austroasiatic	Austric	0.0000	0.0000	0.0000	0.4545	0.0000	0.0000	0.4091	0.1364	22	(Le, et al., 2001)	0.7749	2.697061
Yugoslav	Yugoslav	44	21	Servo-Croatian	Balto-Slavic	Indo-European	0.6000	0.0516	0.0000	0.0000	0.1484	0.0129	0.1226	0.0645	155	(Efremov, 1990)	0.6154	1.205708

Table 1. 89 populations used in the analysis. N, sample size (in number of non minor β-thalassemia chromosomes), relative frequencies of the main mutations; other, relative frequency of mutations <1%; unknown, fraction of chromosomes associated with disease bearing unidentified mutations. H_{max} and θ_{Hom} , allele diversity parameters (see text).

SAMOVA Group	N	IVS-I-110 (G->A)		Codon 39 (C->T)		IVS-I-1 (G->A)		IVS-I-5 (G->C)		codon 41/42(-CTTT)		Codon 17 (A->T)		Del -4279		IVS-II-654 (C->T)	
		Average	range	Average	range	Average	range	Average	Range	Average	Range	Average	Range	Average	Range	Average	range
Europe-Middle East-Iran region*	49	0.24	0-0.83	0.18	0-0.52	0.14	0-0.59	0.03	0-0.3	0.003	0-0.05	0	-	0	-	0.00047	0-0.018
Sardinian	3	0.06	0-0.1	0.83	0.75-0.95	0.01	0-0.04	0	-	0	-	0	-	0	-	0	0
South Asia	17	0.007	0-0.1	0.004	0-0.04	0.03	0-0.27	0.59	0.37-0.94	0.04	0-0.04	0.005	0-0.07	0	-	0.0072	0-0.079
South East Asia	16	0	-	0	-	0.0004	0-0.007	0.02	0-0.2	0.41	0.28-0.56	0.2	0.06-0.48	0	-	0.19	0-0.51
Borneo	1	0	-	0	-	0	-	0	-	0	-	0	-	1	-	0	0-0.51

Table 2. 5 groups of populations defined by the SAMOVA algorithm (see text), number of populations and common mutations found within each group

Figure 1.

Geographical distribution (a) and spatial autocorrelogram (b) of IVS-I-110 (G->A) in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

Figure 2.

Geographical distribution (a) and spatial autocorrelogram (b) of codon 39 C->T in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

Figure 3.

Geographical distribution (a) and spatial autocorrelogram (b) of IVS-I-5 (G->C) in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

Figure 4.

Geographical distribution (a) and spatial autocorrelogram (b) of codons 41/42 (-TTCT) in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

Figure 5.

Geographical distribution (a) and spatial autocorrelogram (b) of IVS-I-1 (G->A) in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

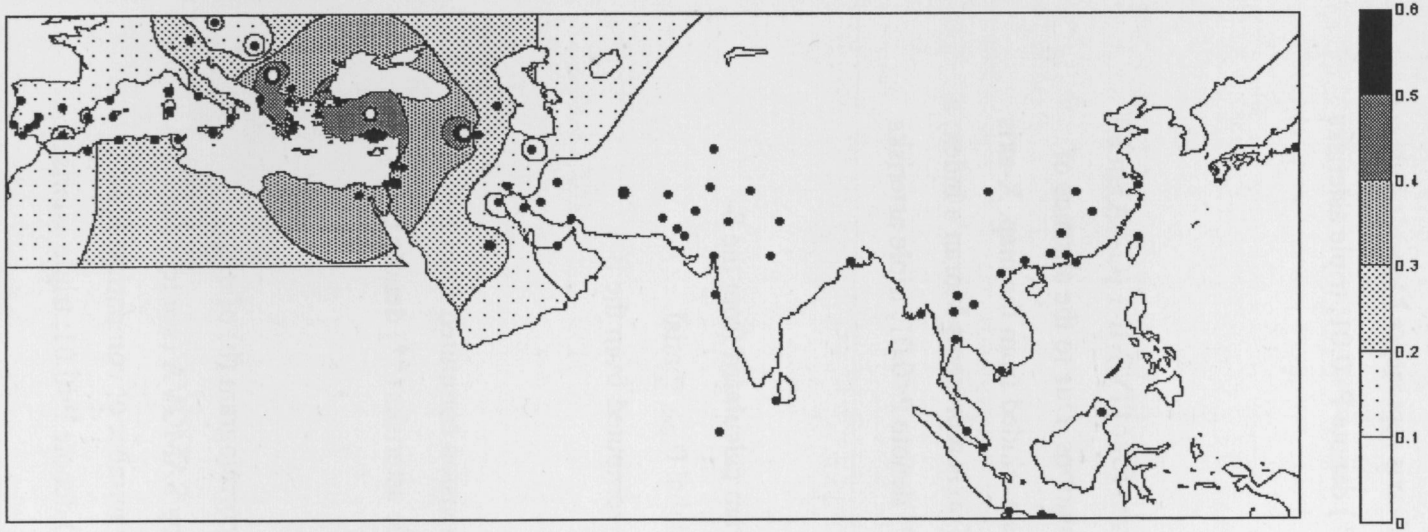
Figure 6.

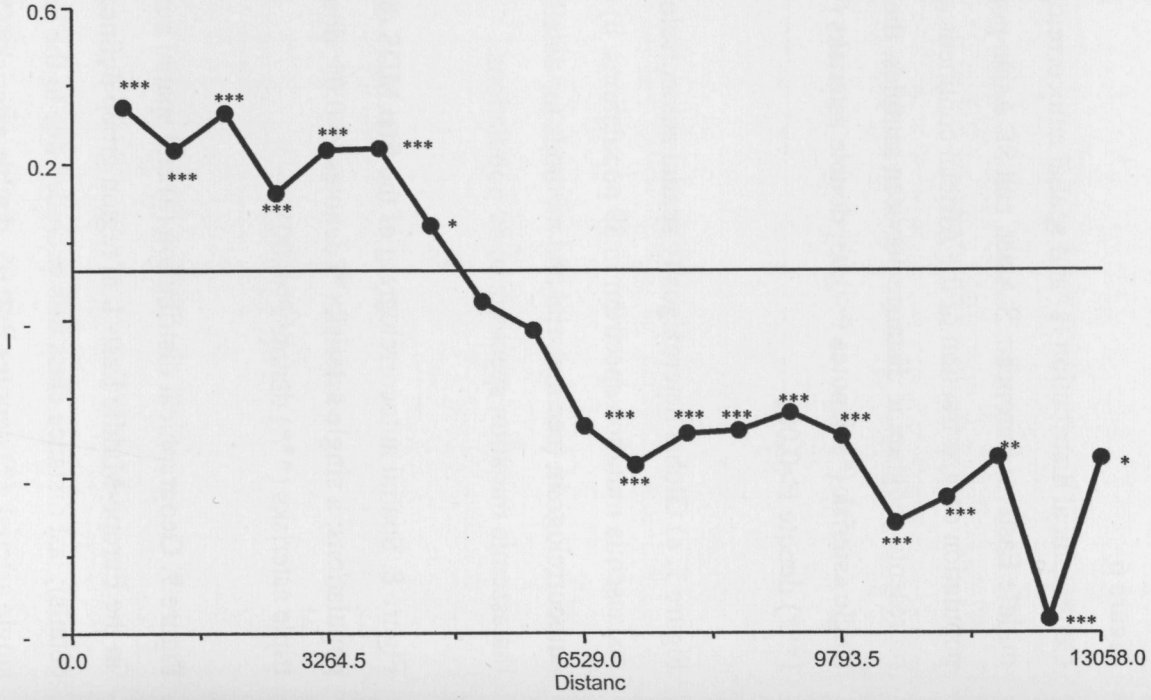
Geographical distribution (a) and spatial autocorrelogram (b) of IVS-II-1 (G->A) in 89 middle Eastern, European, S Asian and SE Asian populations. Due to the absence of population data, a fraction of the African continent was excluded from the map. X-axis represents geographic distance between samples; the Y-axis represents Moran's Index; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

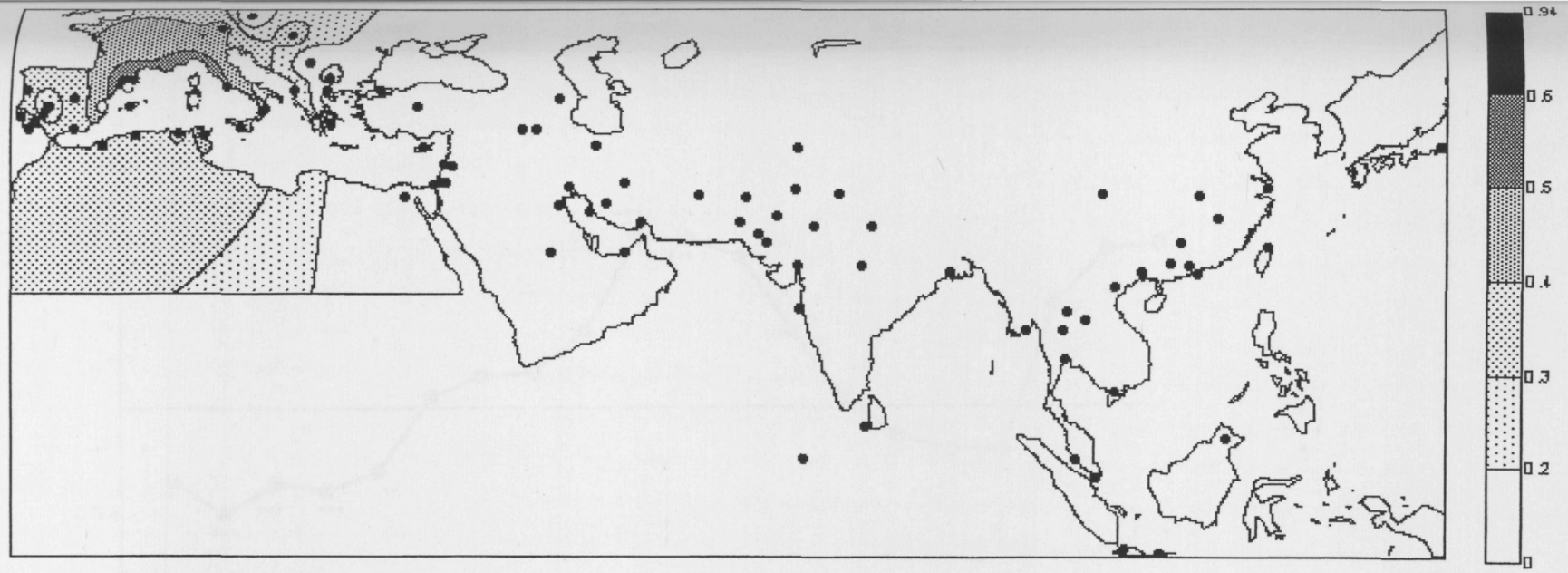
Figure 7. a) Global heterozigosity spatial autocorrelogram calculated from the β-thalassemia mutation spectrum in 89 populations. b) Global θ_{Hom} spatial autocorrelogram (see material and methods for details) computed from the β-thalassemia mutation spectrum in 89 populations.

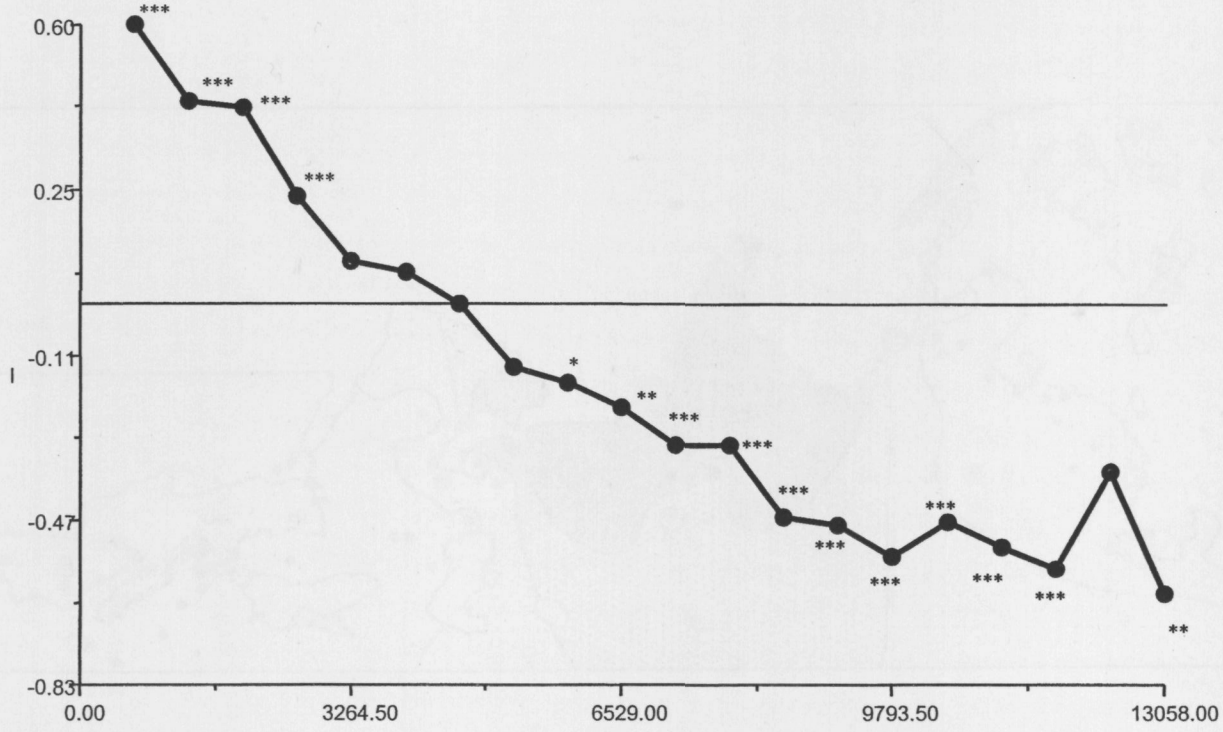
Figure 8. Spatial autocorrelogram of the first MDS dimension computed from 89 populations; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

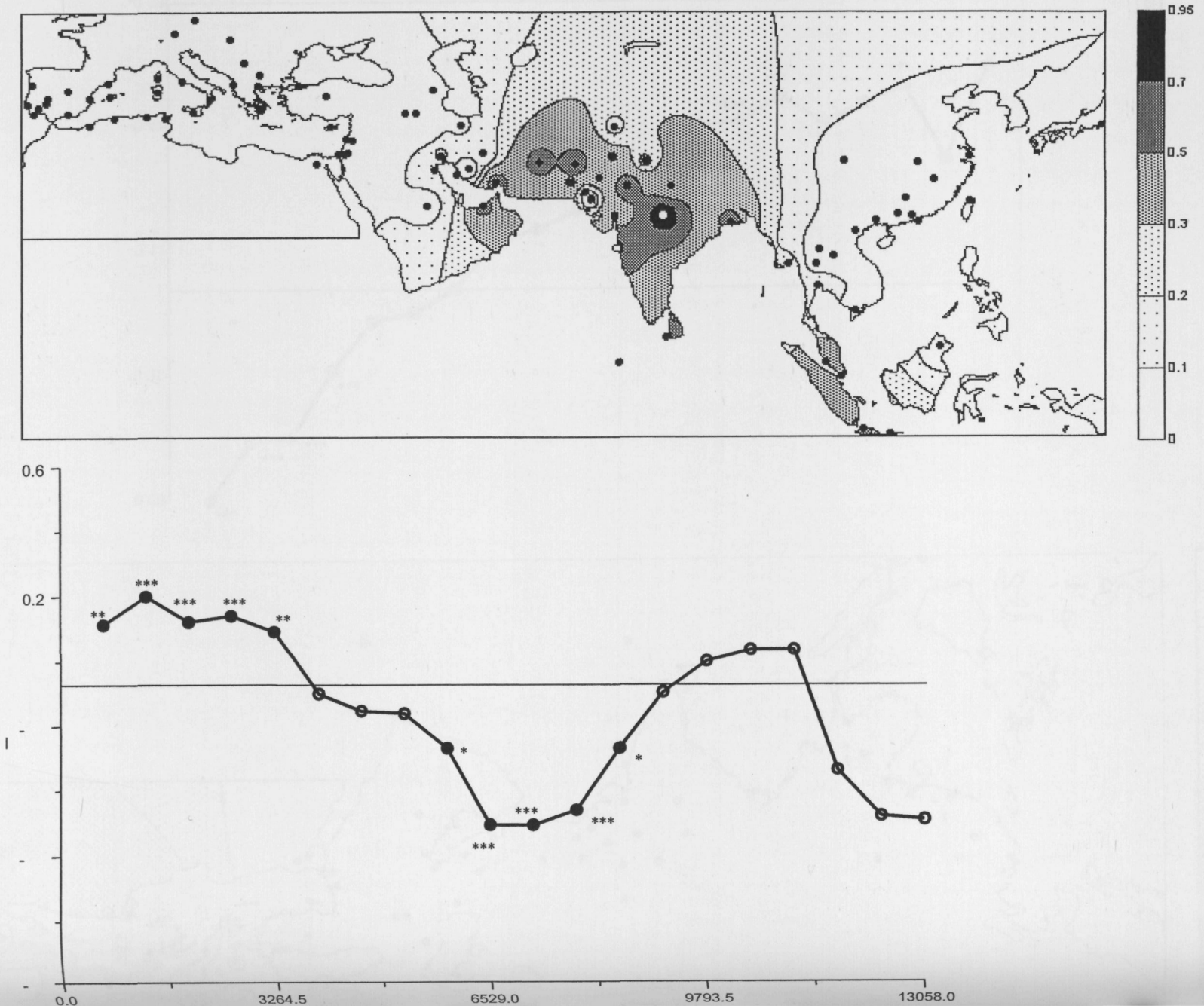
Figure 9. Geographical distribution (a) and spatial autocorrelogram (b) of θ_{Hom} in the Europe-Middle East- Iran region group defined by SAMOVA (see text for details); last distance class was excluded due to the low number of comparisons; a single asterisk (*) denotes P<0.05; double asterisks (**) denote P<0.01; triple asterisks (***) denote P>0.001

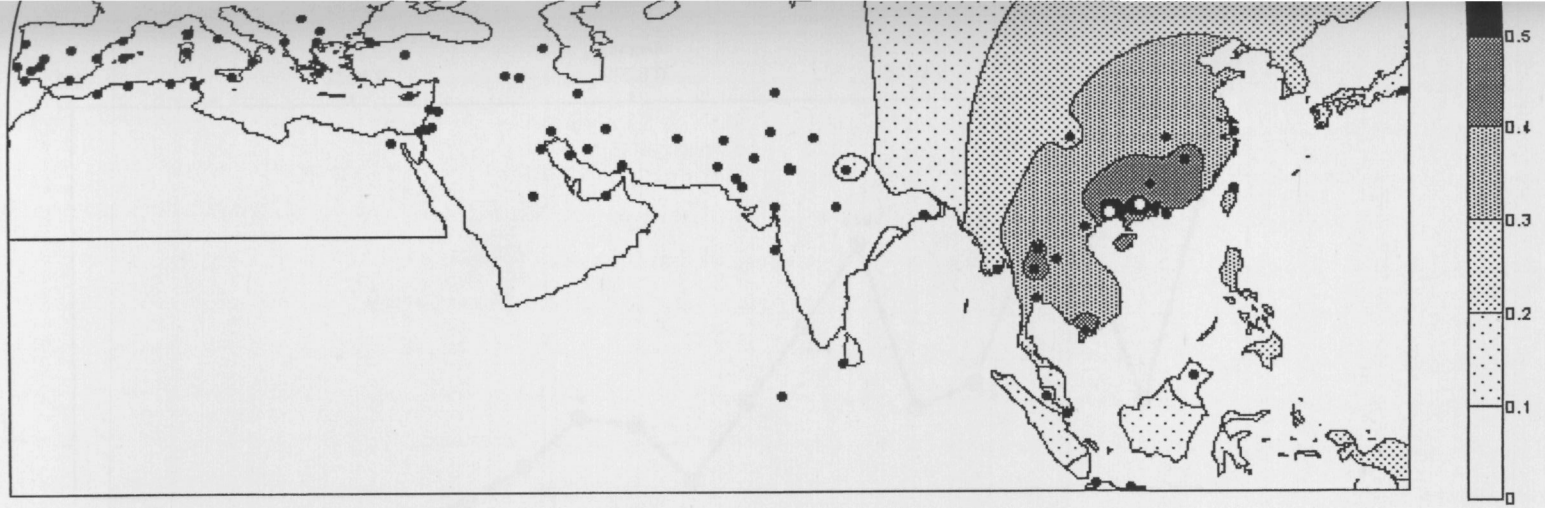

0.6
0.5
0.4
0.3
0.2
0.1
0

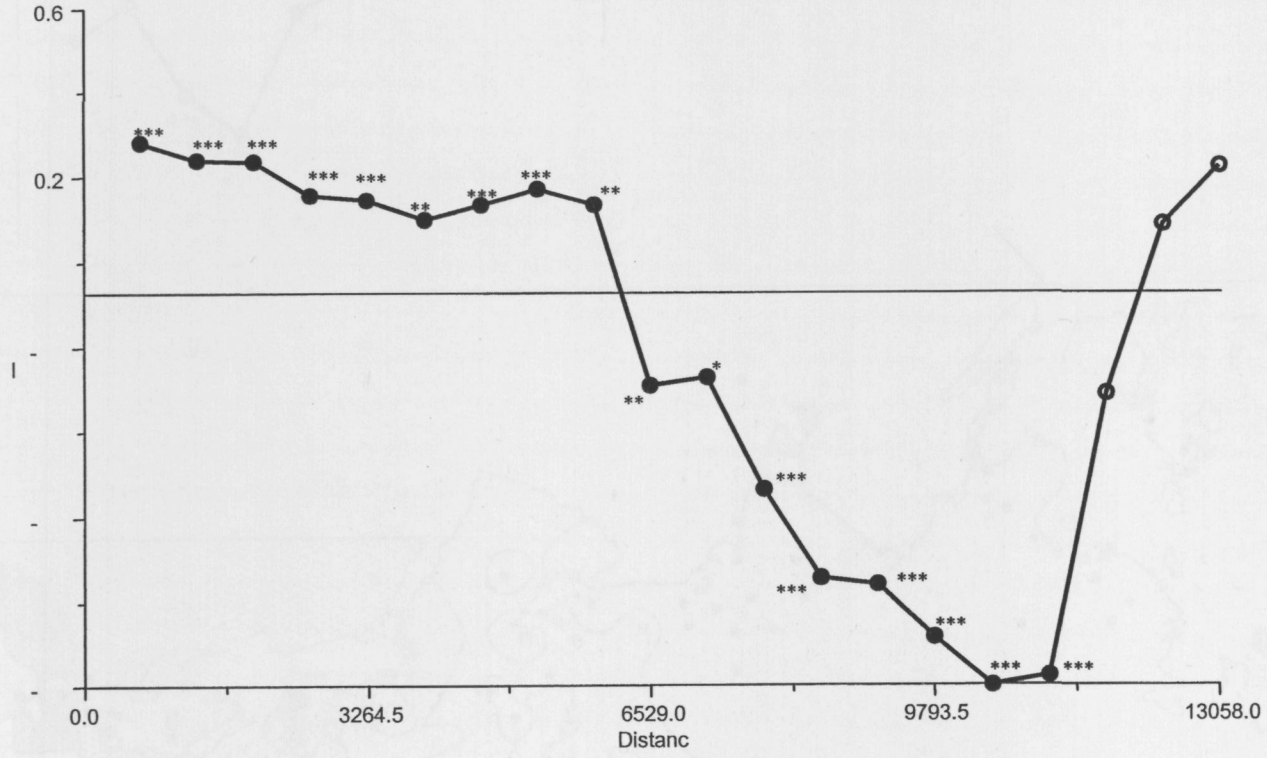

0.6
0.2
I
0.0
3264.5
6529.0
9793.5
13058.0
Distanc

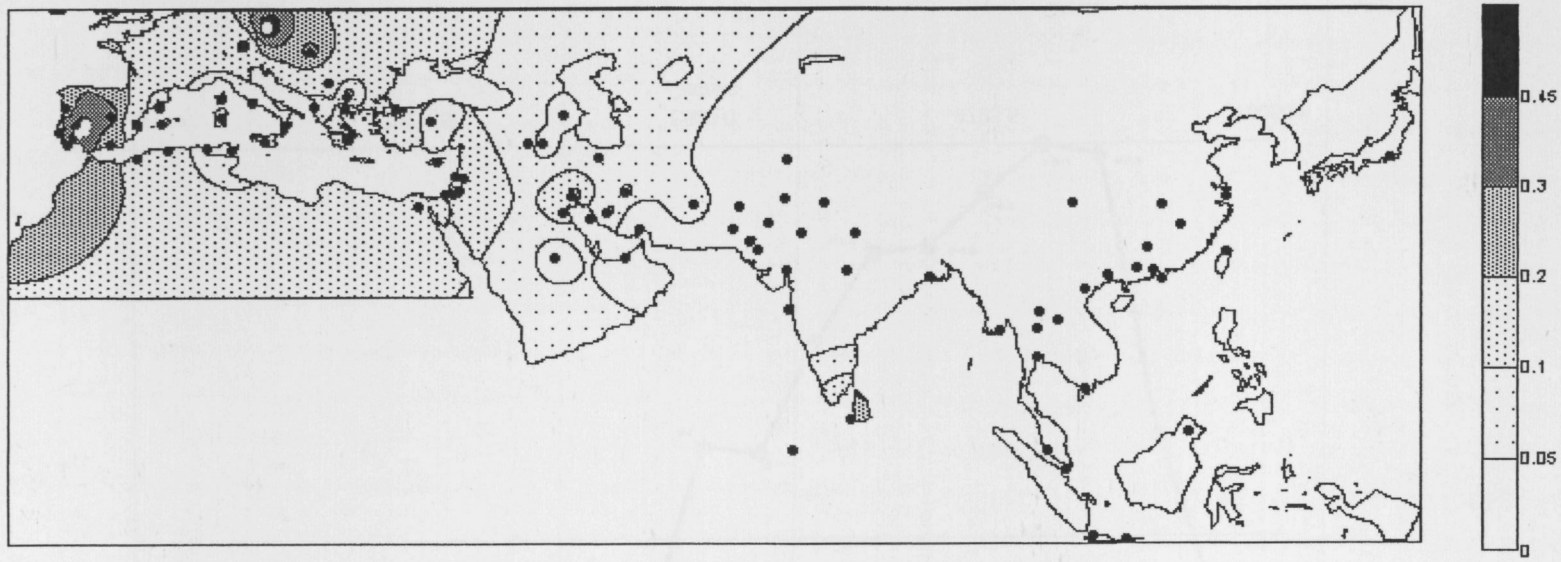

0.94
0.6
0.5
0.4
0.3
0.2
0

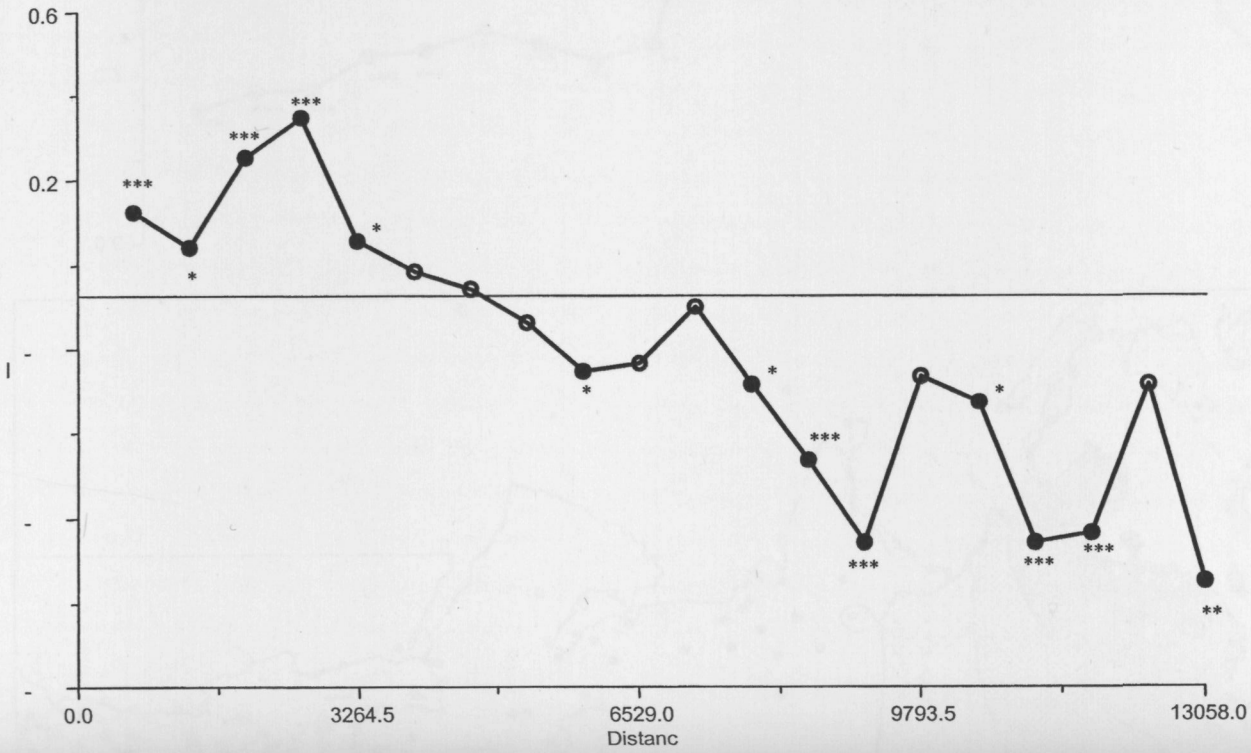

0.60
0.25
-0.11
-0.47
-0.83
0.00
3264.50
6529.00
9793.50
13058.00

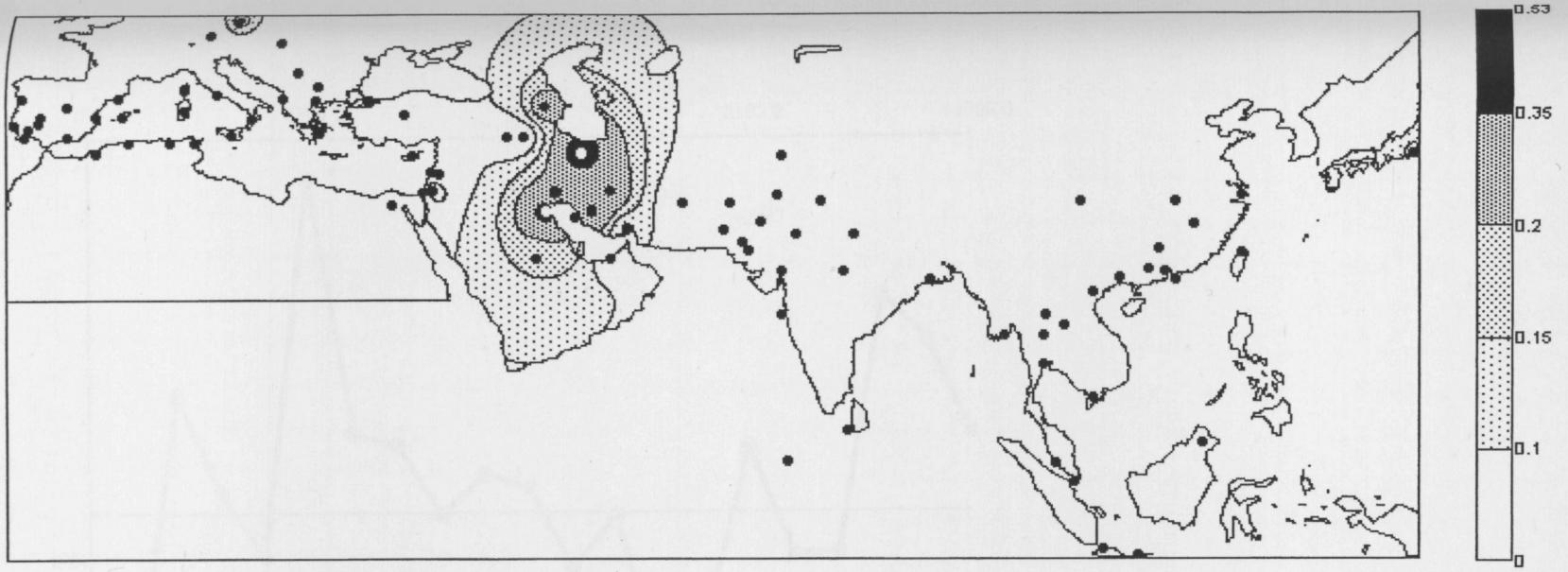

0.95
0.7
0.5
0.3
0.2
0.1
0
0.6
0.2
0.0
3264.5
6529.0
9793.5
13058.0

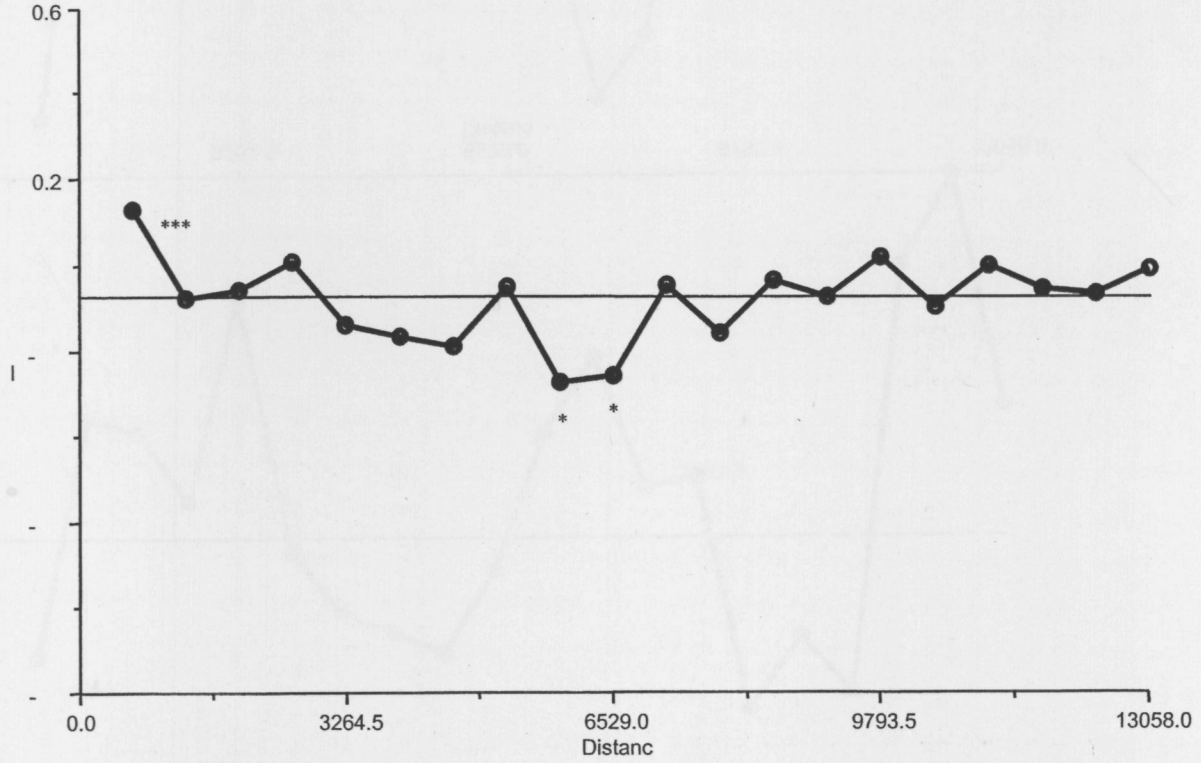

0.5
0.4
0.3
0.2
0.1
0

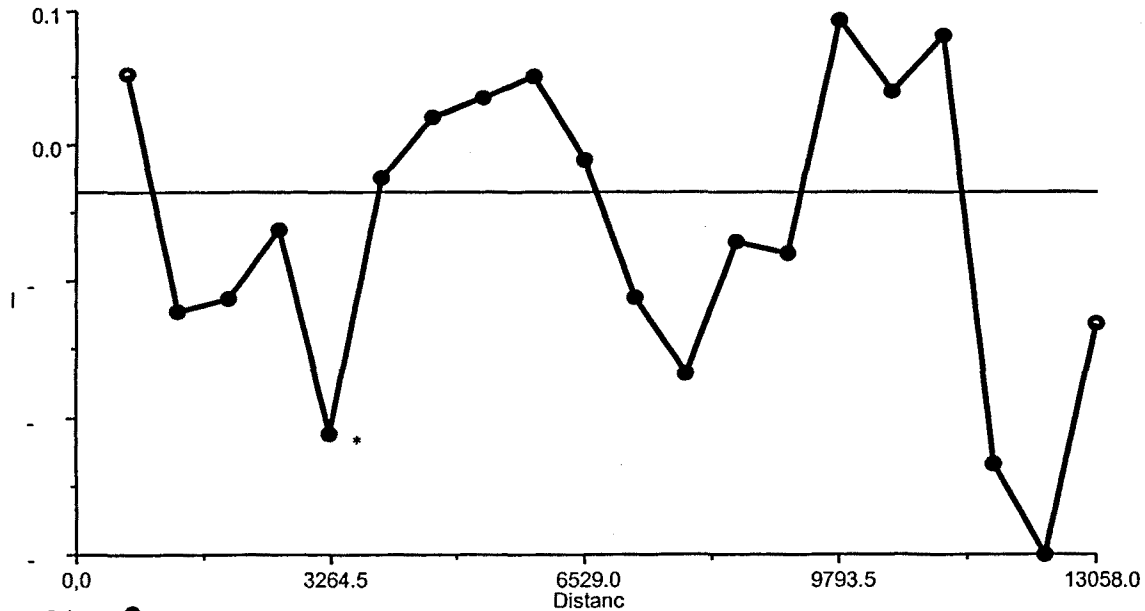

0.6
0.2
0.0
3264.5
6529.0
9793.5
13058.0
Distanc

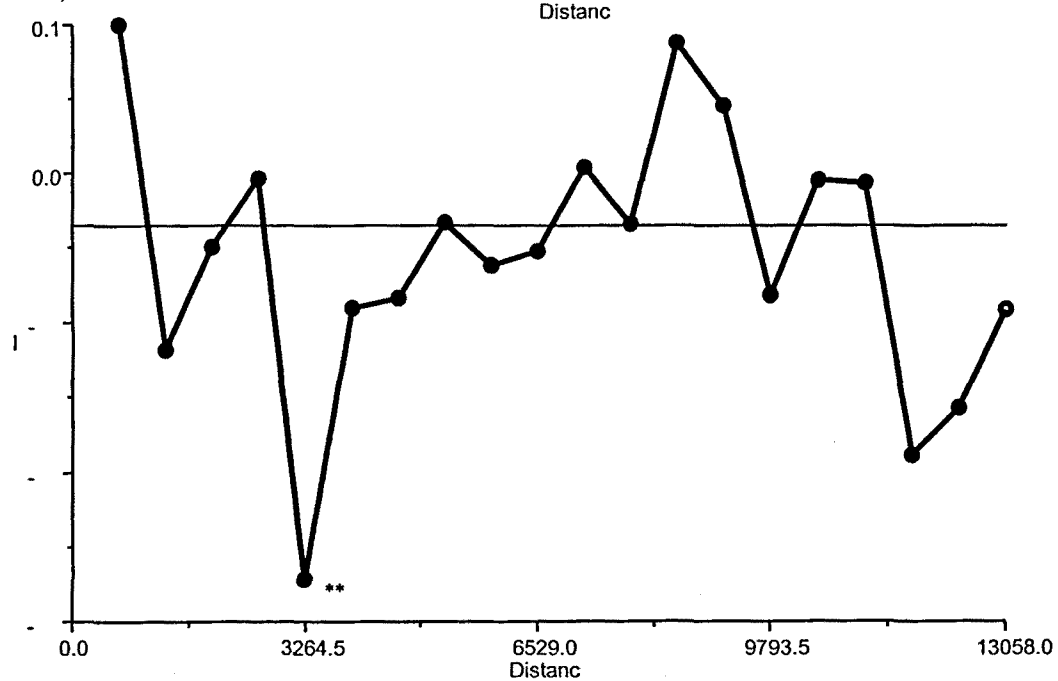

0.45
0.3
0.2
0.1
0.05
0

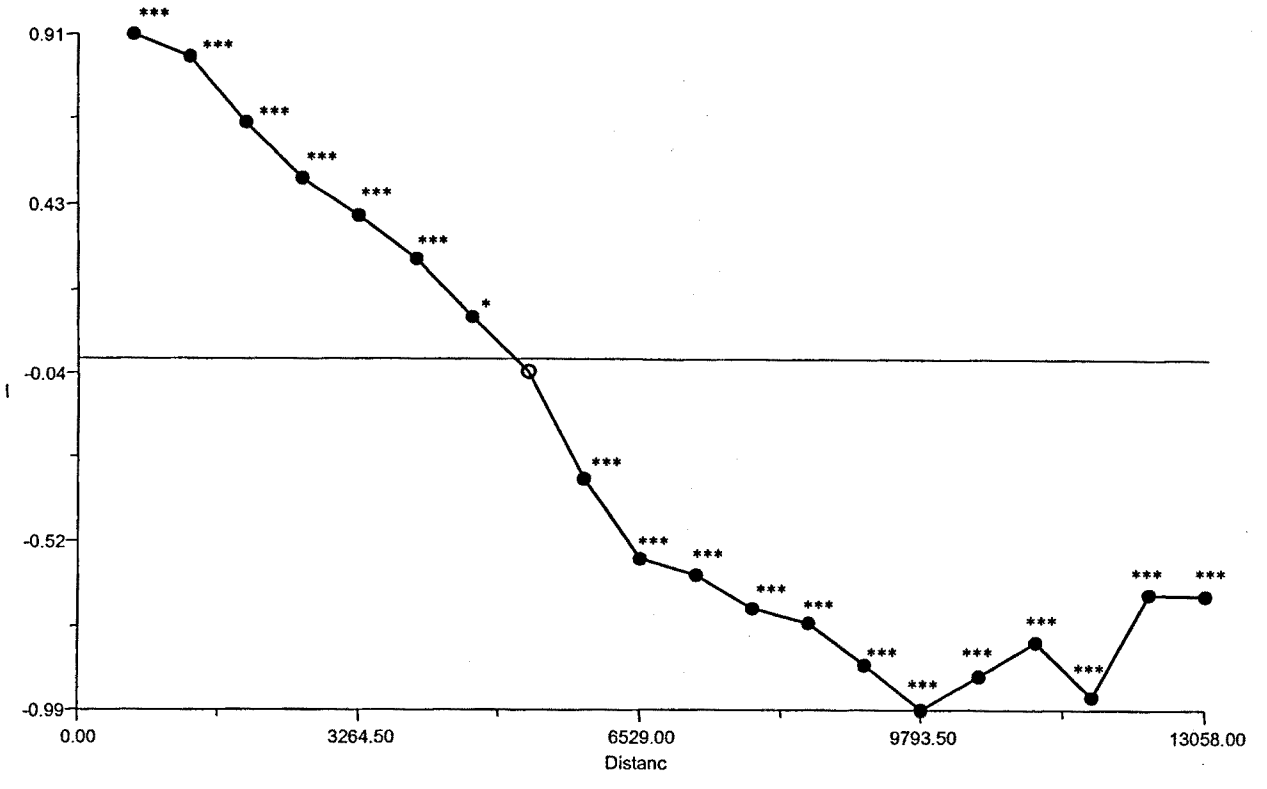

0.6
0.2
0.0
3264.5
6529.0
9793.5
13058.0
Distanc

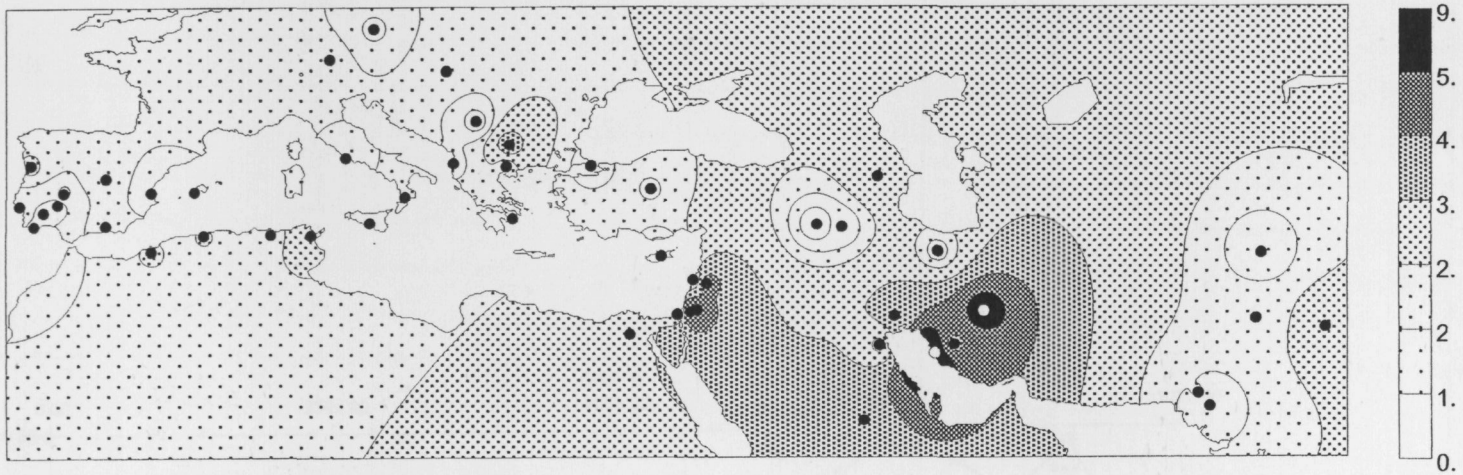

0.53
0.35
0.2
0.15
0.1
0

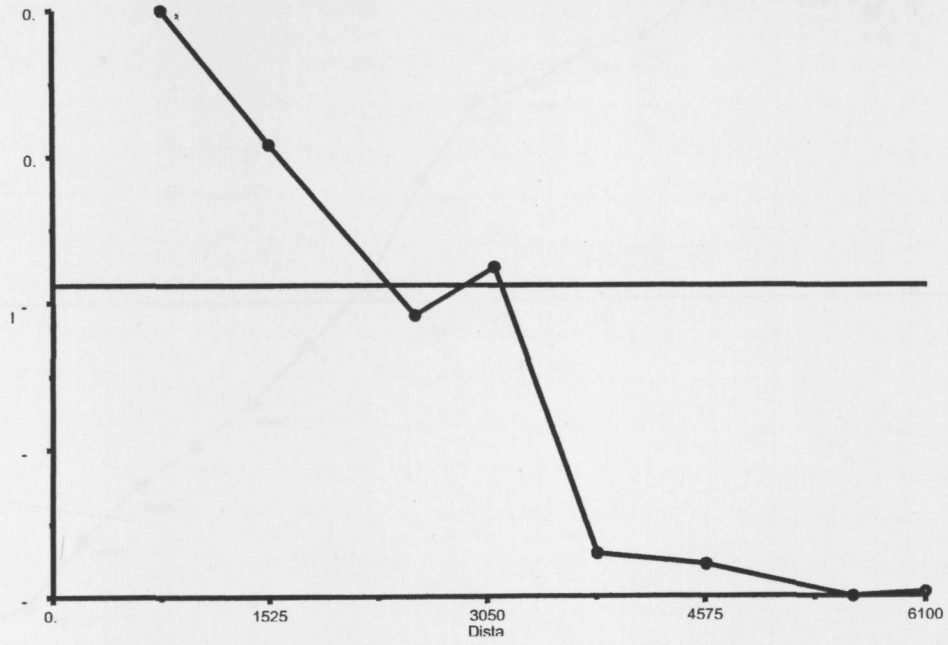

0.6
0.2
I
-
-
***
*
*
0.0
3264.5
6529.0
9793.5
13058.0
Distanc


0.1
0.0
I
-
-
-
*
0,0
3264.5
6529.0
9793.5
13058.0
Distanc


0.1
0.0
I
-
-
-
**
0.0
3264.5
6529.0
9793.5
13058.0
Distanc


0.91
0.43
-0.04
-0.52
-0.99
I
0.00
3264.50
6529.00
9793.50
13058.00
Distanc


9.
5.
4.
3.
2.
2
1.
0.


0.
0.
0.
1525
3050
4575
6100
Dista

Departament de Ciències Experimentals i de la Salut
Universitat Pompeu Fabra (UPF)

Història natural de les malalties genètiques mendelianes i complexes en poblacions humanes

Memòria presentada per Oscar Lao Grueso per optar al grau de doctor en Ciències Biològiques. Aquesta tesi ha estat realitzada sota la direcció del Dr. Francesc Calafell i Majó, a la Unitat de Biologia Evolutiva del Departament de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra, dins del programa de doctorat en *Ciències de la Salut i de la Vida* (*bienni 2000-2002*).

Francesc Calafell i Majó **Oscar Lao Grueso**

Barcelona, Septembre del 2004

3.4 *Capítol IV : "The European Paradox for risk factors in coronary heart disease extends to genetics"*

Oscar Lao, Isabelle Dupanloup, Guido Barbujani, Jaume Bertranpetit, Francesc Calafell.

(sotmès a consideració editorial)

The European Paradox for risk factors in coronary heart disease extends to genetics

Oscar Lao[1], Isabelle Dupanloup[2,3], Guido Barbujani[3], Jaume Bertranpetit[1], Francesc Calafell[1§]

[1]Unitat de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

[2]Center for Integrative Genomics, Faculté de Biologie et de Médecine, Université de Lausanne, Lausanne, Switzerland

[3]Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy

[§]Corresponding author

Email addresses:

OL: oscar.lao@upf.edu

ID: Isabelle.Dupanloup@cig.unil.ch

GB: g.barbujani@unife.it

JB: jaume.bertranpetit@upf.edu

FC: francesc.calafell@upf.edu

Abstract

Background

Coronary heart disease (CHD) is one of the commonest healthy problems in developed countries. The distribution of the incidence of CHD is not homogeneous in Europe, but shows a North to South clinal pattern that has been traditionally associated to the distribution of classical risk factors, such as diet. However, the reduction of CHD incidence in southern European countries is less than predicted by these classical risk factors, which has led to the known as "French or Mediterranean paradox", thus suggesting that other genetic and/or environmental factors are implied in the disease. We have conducted an ecological study trying to correlate the frequencies of susceptibility genotypes of well known genetic CHD risk factors with the incidence of the disease.

Methods

Gene frequencies in the European populations of the susceptibility polymorphisms were obtained from young healthy controls from case-control studies. Seven different genes were considered: ACE, AGT, APOE, F2, F4, MTHFR, PON1 and SERPINE1. CHD incidences in European populations were obtained from a previous study conducted by the WHO MONICA project [1] and were log transformed; matches included only populations sampled within a 300-Km radius of a European MONICA population, residing in the same country and speaking the same language. The correlation degree was computed by a Pearson r statistic.

Results

Surprisingly, in three of the seven polymorphisms analyzed we have obtained negative correlations, which would suggest that susceptibility factors are, in fact, protective and thus, a genetic component must be added to the "Mediterranean paradox".

Conclusions

This result can be understood from the history of populations, that have shaped the genetic diversity of the European populations in clines similar to the observed in case of CHD incidence; this will tend to create spurious correlations, both negative and positive, with polymorphisms not related to the disease or even with polymorphisms of low risk such as the considered in our study, despite they are the best known genetic risk factors for CHD.

Background

Coronary heart disease (CHD) is the leading cause of death in developed countries [2]. This disease is intimately associated to development and progression of atherosclerosis, that is, the progressive accumulation of lipids and fibrous elements in large arteries, which leads to the development of atheromatous plaques. In advanced stages, thrombi can occlude the arteries surrounding the heart, thus resulting in a myocardial infarction [3]. CHD is not homogeneously distributed in Europe, but shows a North to South gradient, with larger incidences in Finland and United Kingdom and lower in the Iberian peninsula [1]. Several environmental factors, such as diet, exercise, tobacco consumption, or obesity, have been classically described as a risk factors for CHD; however, the so-called "French" or "Mediterranean" paradox refers to a lower incidence than predicted by classical environmental factors in southern European countries compared to northern countries [4], which suggest that other factors, both environmental and genetic, are involved in the disease. Genetic risk factors have been well established for coronary heart disease [5], mainly by means of matched case-control association studies within a population. The frequencies of these risk factors (defined as the presence or absence of a given allele or genotype) vary among populations, but the joint effects of these factors have not yet been used systematically to understand the vast differences in CHD incidence among European populations. We have tried to evaluate the correlation between the frequency of genetic risk factors and the incidence of CHD in a broad set of European populations. In general, a positive correlation is expected between the frequency of the risk allele or genotype and the incidence of the disease if the weight of the genetic factors in the etiopathogenesis is important. On the contrary, factors with smaller weights compared to other known or unknown genetic and/or environmental variables may not show any correlation with the disease incidence. In fact, due to the odds ratio observed in the majority of polymorphisms associated to CHD (see table 1), this last situation seems more plausible than the first, but two previous studies have found a positive correlation between the incidence of the disease and the gene frequency [6, 7].

Methods

We checked for this correlation by drawing CHD incidence measures in European populations from the WHO MONICA Project ([1]; [8]), which was established in the early 1980s in many centres around the world to supervise trends in cardiovascular diseases. Independently, we surveyed the literature for risk allele

frequencies in matched healthy population samples of young individuals (<65 years old). Matches included only populations sampled within a 300-Km radius of a European MONICA population, residing in the same country and speaking the same language. The incidence measure taken was the logarithm of the number of CHD events per 100,000 individuals, and was considered separately for men and women. We used a subset of genes that have been previously associated to CHD in case-control studies [9], taking into account all known risk loci for which we found frequency data for 10 or more MONICA population matches. The risk loci considered were *ACE*, *AGT*, *APOE*, *F2* (coagulation factor II), *F5* (coagulation factor V), *MTHFR*, *PON1*, and *SERPINE1* (also known as PAI-1). It should be noted that, although their odd ratios appear modest, these are the polymorphisms which have consistently shown the strongest associations so far with CHD.

Results and Discussion

Along with their odds ratios with CHD and the correlation with incidence, the loci analyzed are listed in Table 1. Large positive correlations between incidence and frequency of the risk genotype(s) were found for the two loci with the highest odd ratios, *APOE* and *PON1*, and for *APOE* the correlation was statistically significant. *AGT*, *F5*, and *SERPINE1* showed low and non-significant correlations. Surprisingly, *negative* correlations were obtained for *ACE*, *MTHFR*, and *F2*(see figure 1), statistically significant in the two former cases. In fact, a previous study had found positive correlations for *ACE* in a small set of selected populations [7].

Positive correlations are consistent with a role of these genetic factors in the development of CHD, as stated above. However, the spread of the attributable fraction (AF), defined as the percentage of cases that is explained by the risk factor and calculated from the relative risk, given the relative risks conferred by alleles at these genes and the range of variation of the allele frequencies themselves, is modest across Europe for the genetic risk factors considered. In the case of *PON1*, the fraction of cases attributable to genotypes RR and RQ ranges from 0.9% in Southern Europe to 1.18% in Northern Europe; these figures are 1% and 2.58% for carriers of *APOE**ε4. This contrasts with the dramatic increase in the incidence of CHD, from around 200 cases per 100,000 individuals in Southern Europe to over 835 cases per 100,000 individuals in Northern Europe. Increasing the known genetic AF from Southern to Northern European levels would add just 3.7 cases per 100,000

individuals, far from the observed fourfold increase in incidence. In contrast, other authors [6] found that the variation in CHD mortality rates across nine populations (representing six European countries and China) could be explained in a greater proportion by the variation in the frequency of *APOE**ε4. Our data, in a wider set of populations and genes, show that the power of these genetic factors in predicting the geographical pattern of CHD in Europe is very limited.

As shown by *ACE*, *MTHFR*, and *F2*, European populations with a low frequency of well-established risk alleles have shown elevated incidences for CHD and vice versa, in a pattern paralleling the so-called "French paradox" or "Mediterranean paradox",. Thus, a genetic component may be added to the environmental component of the CHD paradox. Two considerations can help us to understand this genetic paradox. First, the odds ratios found for genetic risk factors are modest even in study designs that are optimized to find genetic (as opposed to environmental) risk factors; thus, their attributable fractions are low, and their contribution to shaping the geographical incidence pattern of the disease can be easily overrun by differences in known and unknown genetic and environmental factors. Second, the genetic variation of different loci is not independent between populations, because population history affects the whole genome through forces such as drift and gene flow. Thus, a number of genes covariate between European populations [10]. In particular, *ACE*, *APOE*, *MTHFR* and *F2* show European-wide clines that are similar to those presented by many other genes, which is a product of the continent's demographic history. And because these are mainly biallelic loci, their two alleles will show complementary spatial patterns (in the case of *APOE*, alleles ε4 and ε3 display complementary clinal patterns, while ε2 is less frequent and irregular). Since the geographic structure of the incidence of CHD does also show the same clinal pattern in Europe, spurious correlations between the incidence of the cardiovascular disease and some allelic frequencies are expected and would explain why CHD incidence appears to correlate with alleles that have been described as moderately protective.

Conclusions

Neither known environmental exposition differences nor the variation of known genetic factors explain the differences of CHD incidences in different European populations. The discovery of the "French paradox" for classical, environmental factors for CHD spurred research into the mechanisms, such as wine

consumption, that could solve the paradox. Similarly, the extension into genetics of the paradox implies that more research is needed to understand the complexity of the genetic architecture of CHD and of the genetic-environmental interactions.

Competing interests

None declared.

Authors' contributions

All authors assisted in designing the study and discussing the results and interpretations. Oscar Lao collected the data and carried out the statistical analyses, assisted by Isabelle Dupanloup, and under the supervision of Francesc Calafell, Guido Barbujani and Jaume Bertranpetit. Oscar Lao and Francesc Calafell wrote the first version of the manuscript, which was read and edited by all other authors.

Acknowledgements

This study received support from the Spanish Ministry for Science and Technology grants BMC2001-0772 and 2001SGR00285, and from the FISR fund of the Italian Ministry of Universities. The funding sources had no role in the study design, data collection, data analysis, data interepretation, or writing of the report.

References

1. Tunstall-Pedoe H, Kuulasmaa K, Mahonen M, Tolonen H, Ruokokoski E, Amouyel P: **Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease.** *Lancet* 1999, **353**(9164):1547-1557.

2. Nabel EG: **Cardiovascular disease.** *N Engl J Med* 2003, **349**(1):60-72.

3. Lusis AJ, Mar R, Pajukanta P: **Genetics of atherosclerosis.** *Annu Rev Genomics Hum Genet* 2004, **5**:189-218.

4. Jamrozik K, Spencer CA, Lawrence-Brown MM, Norman PEV: **Does the Mediterranean paradox extend to abdominal aortic aneurysm?** *Int J Epidemiol* 2001, **30**(5):1071-1075.

5. Winkelmann BR, Hager J: **Genetic variation in coronary heart disease and myocardial infarction: methodological overview and clinical evidence.** *Pharmacogenomics* 2000, **1**(1):73-94.

6. Stengard JH, Weiss KM, Sing CF: **An ecological study of association between coronary heart disease mortality rates in men and the relative**

frequencies of common allelic variations in the gene coding for apolipoprotein E. *Hum Genet* 1998, **103**(2):234-241.

7. Young RP, Thomas GN, Critchley JA, Tomlinson B, Woo KS, Sanderson JE: **Interethnic differences in coronary heart disease mortality in 25 populations: association with the angiotensin-converting enzyme DD genotype frequency.** *J Cardiovasc Risk* 1998, **5**(5-6):303-307.

8. **WHO MONICA project** [http://www.ktl.fi/monica/index.html]

9. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies.** *Genet Med* 2002, 4(2):45-61.

10. Cavalli-Sforza LL, Menozzi P, Piazza A: **The history and geography of human genes.** Princeton (NJ): Princeton University Press; 1994.

11. Agerholm-Larsen B, Nordestgaard BG, Tybjaerg-Hansen A: **ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites.** *Arterioscler Thromb Vasc Biol* 2000, **20**(2):484-492.

12. Wilson PW, Schaefer EJ, Larson MG, Ordovas JM: **Apolipoprotein E alleles and risk of coronary disease. A meta-analysis.** *Arterioscler Thromb Vasc Biol* 1996, **16**(10):1250-1255.

13. Kato N, Sugiyama T, Morita H, Kurihara H, Yamori Y, Yazaki Y: **Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies.** *J Hypertens* 1999, **17**(6):757-763.

14. Wald DS, Law M, Morris JK: **Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis.** *Bmj* 2002, **325**(7374):1202.

15. Juul K, Tybjaerg-Hansen A, Steffensen R, Kofoed S, Jensen G, Nordestgaard BG: **Factor V Leiden: The Copenhagen City Heart Study and 2 meta-analyses.** *Blood* 2002, **100**(1):3-10.

16. Boekholdt SM, Bijsterveld NR, Moons AH, Levi M, Buller HR, Peters RJ: **Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review.** *Circulation* 2001, **104**(25):3063-3068.

Figures

Figure 1 - Negative correlations observed between the logarithm of the incidence of CHD and the susceptibility genotypes.

a) correlation between CHD incidence in men and the DD genotype of ACE. b) correlation between CHD incidence in women and the DD genotype of ACE. c) correlation between CHD incidence in men and the VV genotype of MTHFR. d) correlation between CHD incidence in women and the VV genotype of MTHFR. e) correlation between CHD incidence in men and the AG genotype of F2. f) correlation between CHD incidence in women and the AG genotype of F2.

Tables

Table 1 - Genetic risk factors for CHD and their correlation with incidence in European populations

Correlations were computed separately for men and women given their different CHD incidences. The incidence measure taken was the logarithm of the number of CHD events per 100,000 individuals. OD pooled: odds ratio for CHD for the risk allele/genotype listed in large metaanalyses. N, number of data points (populations) considered; incidence in women was not available in some populations.

(a) OD pooled for hypertension

Locus	Polymorphism	Risk allele/ genotype	OD pooled	Reference	N	Men	N	Women
PON1	Q192R	QR/RR	1.44	[11]	10	r = 0.40 (p=0.245)	10	r = 0.34 (p=0.33)
APOE	ε2/ε3/ε4	ε4ε2/ε4ε3/ε4ε4	1.26	[12]	21	r = 0.59** (p=0.005)	20	r = 0.54* (p=0.015)
AGT	M235T	TT	1.22(a)	*[13]*	11	r = 0.16 (p=0.631)	11	r = 0.09 (p=0.798)
ACE	I/D	DD	1.21	*[11]*	23	r = -0.54** (p =0.007)	22	r =-0. 61** (p = 0.002)
MTHFR	A222V	VV	1.21	*[14]*	22	r =-0.51* (p=0.014)	20	r = -0.44* (p=0.049)
F5	R506Q	RQ	1.20	*[15]*	18	r = 0.05 (p=0.845)	17	r = 0.17 (p=0.51)
SERPINE1	4G/5G	4G4G	1.20	*[16]*	12	r = 0.19 (p=0.543)	12	r = 0.15 (p=0.637)
F2	G20210A	GA	1.11	*[16]*	15	r = -0.42 (p=0.12)	14	r = -0.43 (p=0.12)

3.5 *Capítol V: "Geographic structure of genes conferring risk for Coronary Heart Disease in European populations"*

Oscar Lao, Isabelle Dupanloup, Guido Barbujani, Jaume Bertranpetit, Francesc Calafell

(manuscrit en preparació)

Geographic structure of genes conferring risk for Coronary Heart Disease in European populations

Oscar Lao[1], Isabelle Dupanloup[3,2], Guido Barbujani[3], Jaume Bertranpetit[1], Francesc Calafell[1]

1 Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

2 Center for Integrative Genomics, Faculté de Biologie et de Médecine, Université de Lausanne, Lausanne, Switzerland

3 Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy

Running tittle :

Keywords: Coronary Heart Disease, ACE, APOE, F5, F2, MTHFR, PON1, AGT, spatial distribution, Europe

Correspondence:
Francesc Calafell
Unitat de Biologia Evolutiva
Facultat de Ciències de la Salut i de la Vida
Universitat Pompeu Fabra
Doctor Aiguader 80
08003 Barcelona, Catalonia
Spain
Tel:+34-93-542 28 41
Fax: +34-93-542 28 02
e-mail: francesc.calafell@upf.edu

ABSTRACT

Coronary Heart Disease (CHD) is one of the commonest healthy problems in developed countries. CHD is a complex disease and several environmental and genetic factors are involved in the progression and development of the disease. Environmental factors such as the diet, physical exercise or tobacco consumption have been traditionally associated to this disease. However, they fail to explain the "French paradox", that is that Southern european countries have a lower incidence than predicted by these classical environmental factors, thus suggesting that other environmental and genetic factors must be involved in the etiopathology of CHD. It is in this context that the knowledge of the spatial distribution of alleles associated with CHD could give us clues to a better understanding of the disease. In our study we have analyzed the geographical distribution of seven classical CHD susceptibility alleles of APOE, ACE, F5, F2, MTHFR, PON1, and AGT genes in Europe and have shown that their spatial distribution is similar to the distribution observed in case of the classical markers, thus suggesting that their spatial distribution could be shaped by the same demographic and migratory history that have shaped the genetic landscape in Europe in the case of neutral markers.

INTRODUCTION

Coronary Heart Disease (CHD) is the commonest cause of death in developed countries (Nabel 2003); epidemiological studies have shown that the incidence of the disease shows a North to South gradient in Europe (Tunstall-Pedoe et al. 1999). CHD is a complex disease where different environmental and genetic risk factors interact to produce the phenotypic disease trait. Tobacco consumption, the lack of physical exercise and a diet rich in fatty acids are classical environmental risk factors; however, the so-called Mediterranean diet has been postulated as a protective factor in CHD development. However, these well-known risk factors cannot explain the "French or Mediterranean paradox" (de Lorgeril et al. 2002): the north-south gradient in intake of fat in Europe is much less steep than that for heart disease. Furthermore, the elevated heritability in twin studies of physiological factors associated to CHD such as blood pressure (Evans et al. 2003) points to the importance of genetic factors in addition to environmental factors. Polymorphisms of susceptibility in CHD have been usually searched in genes involved in the control of blood pressure, lipid metabolism or thrombosis formation (Winkelmann and Hager 2000). Several genetic factors have been associated to CHD with a low risk (OR<2) (Hirschhorn et al. 2002). According to the Common Variant/Common Disease (CV/CD) hypothesis the genetic variants associated to common diseases are expected to be frequent in human populations (Chakravarti 1999), but even the neutral genetic variability observed has been modelled by migratory factors and is not equally distributed in human populations (Cavalli-Sforza et al. 1994). This opens the possibility of analyzing the correlation between the incidence of the disease and the genetic susceptibility factor which can lead to a better understanding of the ethiology of the disease (Stengard et al. 1998; Young et al. 1998). However, knowledge of the evolutionary history of the risk polymorphism in the populations is essential in order to distinguish putative correlations from spurious correlations product of common spatial patterns with causative risk factors (Lao et al. 2004), a usual situation in case of neutral polymorphisms. Previous works analyzing the spatial distribution of genes associated with CHD (i.e. the F5 Leiden Mutation (Lucotte and Mercier 2001)) suggest that this could be the case.

We have compiled allele frequencies for a subset of well known CHD risk factors and we have analyzed their geographical distribution. Results suggest that the

subset of genes associated with CHD show a spatial pattern similar to the observed in neutral regions of the genome which means that positive correlations between the incidence of the disease and the frequency of the genetic risk factor should be taken with caution.

MATERIALS AND METHODS

Databases

For each CHD susceptibility gene a frequency database was made considering European, North African and Middle Eastern populations. Risk allele frequencies for each population was obtained from healthy control population data described in case-control studies after a Medline search and from ALFRED database (http://alfred.med.yale.edu/alfred/index.asp). For some of the genes, information was sufficiently available for non-European populations, and the database was expanded to include data for Sub-Sahara, Africa and East Asia. Data was excluded when controls were hospitalised patients or older than 65 years. Populations of recent origin and admixed, such as European-Americans, were not considered. Data of the same population coming from different articles were pooled when none of the authors were present in both articles or controls used were of different sex or age. Genes were selected for further analyses when data was available for more than 25 populations in the European, North African and Middle Eastern populations . Each database contained information about the geographical localization, sample size and allele frequencies of each population.

Frequency maps

A frequency map was drawn with Surfer 8.0 (Golden Software; www.goldensoftware.com) for each risk allele. Each map only comprised Europe, the Middle East and North Africa and was limited between 10°W and 51°E and between 23°N and 64°N. Interpolation points were placed at a density of 0.1 degrees. Interpolation was based on inverse-squared weighing distance. A barrier between the European continent and North Africa was established in order to prevent interpolation from points in different landmasses. North Africa was dashed when no population data was available. Interpolated values were not used in any other analysis.

Spatial Autocorrelation Analysis

Spatial distribution of the risk allele frequencies were analysed by means of a Spatial Autocorrelation Analysis (Sokal and Oden 1978). PASSAGE (Rosenberg 2001) was used to compute and plot the autocorrelogram of each risk allele. Spatial Autocorrelation Analysis plots a measure of autocorrelation (i.e Moran's I) against classes of increasing point distances. The shape of the plot describes the spatial relation of the data; a change from positive autocorrelation values in the closest distances to negative autocorrelation values in longest distances is expected in the case of a clinal pattern (Barbujani 2000); other shapes, such as positive values for closest distances classes, negative values for intermediate classes and positive values for longest classes can be expected to find if the cline is focused on the middle of the map.

Genetic Distances

Reynolds' genetic distances (Reynolds et al. 1983) were computed for a joint group of loci: APOE, ACE, F5, F2 and MTHFR using the PHYLIP package (Felsenstein 1995). Populations without data in all the loci were matched with neighbouring populations that spoke the same language. This matrix was compared by means of Mantel test (Mantel 1967) with genetic distance matrices obtained for classical markers (Cavalli-Sforza et al. 1994), two different datasets of Y-chromosome ((Rosser et al. 2000; Semino et al. 2000); http://web.unife.it/progetti/genetica/pdata.htm) and mitochondrial DNA (Simoni et al. 2000) http://web.unife.it/progetti/genetica/pdata.htm). PASSAGE (Rosenberg 2001) was used to compute Mantel tests.

Multidimensional Scaling

Multidimensional Scaling (MDS) is a multidimensional technique that reduces the dimensionality of data in a distance matrix format (Kruskal and Wish 1990). It computes new values for each point from the number of dimensions specified a priori and produces a measure of good-of-fitness called stress. This parameter produces an estimation of how well a particular configuration reproduces the observed distance.

AMOVA

Analysis of Molecular Variance (Excoffier et al. 1992) was performed with APOE, ACE, F5, F2 and MTHFR genes. In order to compare the fraction of variation

explained between groups with the empirical Fst distribution of SNPs described in Akey et al (Akey et al. 2002) three groups of populations were defined: (i) Europe, North Africa, Middle East and South-West Asia, (ii) sub-Saharan Africa, (iii) East Asia. These groups trend to resemble the three groups used by Akey et al: Afroamericans, Caucasians and Asians.

RESULTS

Worldwide databases were constructed for ACE, APOE, MTHFR, F5 and F2, whereas in case of PON1 and AGT genes databases only considered European, North African and Middle Eastern populations (see materials and methods). Information about loci and polymorphisms analysed, their chromosomal localization, the number of populations studied, the number of chromosomes and average frequency of the susceptibility allele is presented in table 1. Briefly, for ACE we obtained information for 116 populations (67674 chromosomes), 74 populations in case of APOE (95520 chromosomes), 84 populations in case of MTHFR (49082 chromosomes), 71 populations in case of F5 (67288 chromosomes), 46 populations in case of F2 (28802 chromosomes), 27 populations in case of PON1 (19812 chromosomes) and 29 populations in case of AGT (12236 chromosomes). Frequency maps and spatial autocorrelation analysis were computed for the susceptibility alleles of each gene (figures 1-7). In case of APOE, ACE, MTHFR, F2 and F5 two different autocorrelograms were computed, one considering all populations and other excluding populations far away from the European continent or in its edges and considered as a potential outliers. Clinal patterns were obtained in case of the allele ε4 of APOE (after excluding populations with a longitude farther than to 45 E) with a positive and statistically significant correlation with the latitude ($r=0.64$, $p<0.05$) and longitude ($r=0.29$, $p<0.05$). The Leiden mutation of the F5 gene also show a clinal pattern (after excluding the Saami populations) with a positive and statistically significant correlation ($r=0.42$, $p<0.05$) between longitude and the frequency of the susceptibility allele of F5 (see figure) indicating an East to West gradient. A clinal pattern was observed in case of the susceptibility allele of F2 (after excluding Saudi Arabia) showing a negative and statistically significant correlation ($r=-0.52$, $p<0.05$) with latitude and positive and statistically significant with longitude ($p=0.33$, $p<0.05$), indicating a South-East to North-West distribution. Partially clinal patterns were obtained in case of ACE

(considering only European and Middle East populations) with a negative and statistically significant correlation with latitude ($r=-0.60$, $p<0.05$), and MTHFR (after excluding Azerbaijan) with a statistically significant correlation with latitude ($r=-0.31$, $p<0.05$). Random patterns were observed in the case of PON1 and AGT.

The graphical representation of the Reynolds' genetic distance matrix computed with a joint group of genes (APOE, ACE, MTHFR, F2, F5) was performed by means of MDS. Plot is represented in figure 8. The stress reached 0.049, indicating that the final configuration shows a low departure from the original distance matrix and an efficient reduction of the dimensionality of data. A latitudinal gradient was observed from Southern European populations to Northern European populations ($r=0.35$, $p<0.001$ in a Mantel test comparing the genetic distance matrix and the geographic distance matrix).

Results obtained in the comparison of the Reynolds' genetic distance matrix computed for the same joint group of genes and a subset of common populations for classical markers, Y-chromosome and mitochondrial DNA are in table 2. Positive and statistically significant correlations were observed in case of classical markers ($r=0.36$, $p<0.05$) and the data of Y-chromosome ($r=0.32$, $p<0.05$) described in Rosser et al (Rosser et al. 2000). Positive but not statistically significant correlation was obtained in case of the Y-chromosome data ($r=0.21$, $p>0.05$) described in Semino et al (Semino et al. 2000). No correlation was obtained for mitochondrial DNA ($r=0.088$, $p>0.05$).

AMOVA analysis was performed with three groups (see materials and methods) for the genes APOE, ACE, F5, F2 and MTHFR. The fraction of variation explained between groups was 0.79% ($p<0.05$) in case of F2, 1.01% in case of F5 ($p<0.05$), 1.53% ($p<0.05$) in case of APOE, 5.5% ($p<0.05$) in case of ACE and 2.89% ($p<0.05$) in case of MTHFR. Fst values were compared with the distribution of Fst obtained for more than 26,000 SNPs by Akey et al. The Fst value obtained for F2 fell in the 16 percentile of the empirical Fst distribution, the Fst of F5 in the 17 percentile, the Fst of ACE between the percentile 37 and 38, the Fst of APOE in the 20 percentile, and the Fst of MTHFR in the 26 percentile.

DISCUSSION

We have developed databases for seven genes that have been associated with CHD using frequency data of healthy individuals in case-control studies and we have described their spatial distribution in European, North African and Middle Eastern populations. Our results statistically confirms previous descriptions made in case of F5 (Lucotte and Mercier 2001) and APOE (Corbo and Scacchi 1999) genes. Furthermore, we have studied the fraction of variation explained between continental groups in the case of APOE, ACE, F5, F2 and MTHFR genes.

Although the development of CHD usually takes place after the reproductive period (Lusis et al. 2004), the physiologic role of genes associated with CHD make them good candidates to selective processes. Selective explanations have been postulated in case of ACE, where the excess of the presence of the Alu polymorphism (the insertion polymorphism) in elite athletes suggested that this polymorphism could be involved in the resistance to altitude (see (Woods and Montgomery 2001) for a review of the field) although it has been widely considered as a neutral marker (Romualdi et al. 2002); in case of APOE, it has been postulated that the replacement of the ancestral isoform ε4 by the isoform ε3 could be due its metabolic properties that could have been particularly advantageous in the transition from food collection to food production in human populations (Corbo and Scacchi 1999), although another study suggests beneficial effects during embryogenesis of the ε4 allele and a negative effect on foetal survival of the ε3 allele (Zetterberg et al. 2002). In case of AGT, Nakajima et al (Nakajima et al. 2004) found traces of a selective sweep in populations outside of Africa that could be driven by a differential necessity of sodium salt. In case of MTHFR, it has been postulated that homozygous foetuses for the 677T allele show an overall survival advantage and recurrent early and late foetal loss depending on mothers take sufficient folic acid in first case or inadequate folic acid intake in second during pregnancy (Rosenberg et al. 2002). Advantage in foetal implantation in heterozygote mothers for the F5 Leiden mutation has been suggested as balancing selection (Gopel et al. 2001) in front of the risk of abortion in the second trimester and a very preterm delivery, which is also observed in case of the 20210 polymorphism of F2 (Gopel et al. 1999). The fraction of variation explained between continental groups of populations for a subset of

the CHD genes analyzed has shown values within the normal range of variation observed for a large number of SNPs, indicating that if any of this selective forces has acted, it did uniformly across major human groups, and possibly operated at the species level.

Spatial autocorrelograms have shown clinal or partially clinal patterns in case of APOE, ACE, F5, F2 and MTHFR but not in case of PON1 and AGT whose autocorrelograms suggest a random distribution. The presence of clinal patterns has been traditionally associated with migratory processes but also to selective pressures. A random pattern can be obtained by chance or by the random distribution of a selective factor (Barbujani 2000), but it also can be observed when there are few points and their distribution is not homogeneous, which is the case of PON1 and AGT. However, whereas the spatial pattern shaped by selection depends on the particular spatial distribution of the selective factor and is locus specific, migratory processes affect the spatial distribution of the overall genome. A positive correlation was observed between the Reynolds' genetic distance matrix of the joint genes and a Reynolds' genetic distance matrix of neutral markers (Cavalli-Sforza et al. 1994) for the same populations as well as with one subset of Y-chromosome data, but not with mitochondrial DNA, which have a more homogeneous spatial distribution (Simoni et al. 2000). As these are neutral regions of the genome, their shared geographical distribution must be shaped by the same population history. Although it does not exclude the possibility that selective factors also show a similar spatial pattern (migration could be associated to the selective factor (Corbo and Scacchi 1999), this pattern suggests that the polymorphisms associated with CHD have a similar spatial distribution of neutral regions of the genome and thus have been modelled by historically large migrations into or within Europe. Europe was colonized by the anatomically modern humans from the Levant in the Paleolithic about 50,000 years ago and afterwards by farmers and their technological revolution from roughly the same region in the Neolithic about 10,000 years ago (Barbujani and Goldstein 2004). Between both, in the harsher phase of the last glaciation (the so-called Last Glacial Maximum, 18,000 years ago), it is supposed that human populations may have retreated to three glacial refugia in S Europe: Iberia, Italy, and the Balkans, from where they reexpanded when the ice shield melted (Torroni et al. 2001). These migrations had a deep impact in the geographical distribution of mostly all of the neutral genetic diversity, shaping it in a NW to SE gradient (Sokal et al. 1989; Cavalli-Sforza et al. 1994), which is similar to the observed in the graphical

representation of the Reynolds' genetic distance matrix computed with a joint group of genes (APOE, ACE, MTHFR, F2, F5).

Since the distribution of CHD mortality also shows a NW to SE gradient (Tunstall-Pedoe et al. 1999), it is tempting to establish correlations between the incidence of CHD and the history of populations (Lutz 1995) or with the geographical distribution of susceptibility alleles, such as APOE*ε4 or ACE*Del (Stengard et al. 1998; Young et al. 1998). However, as Lao et al have shown (Lao et al. 2004), positive correlations between the frequency of a polymorphism and the incidence of a disease should be taken with caution, especially when the spatial pattern of the polymorphism has been shaped by migratory processes because spurious, even negative, correlations can be easily obtained. In fact, although the susceptibility alleles of ACE*Del and MTHFR*677T are more frequent in Southern countries, CHD incidence is higher in Northern countries (Tunstall-Pedoe et al. 1999).

Thus, the analysis of the spatial distribution of genes associated to CHD gives us clues to a better comprehension of the evolutionary processes that have shaped their geographical distribution and is essential to better understanding the role of these polymorphisms in the ethiology of the disease. We have shown that the spatial distribution of this subset of genes is compatible with the demographic history of human populations in the European continent, a process that probably have shaped the distribution of other genes associated to complex diseases.

ACKNOWLEDGEMENTS

This work was supported by Dirección General de Investigación Científica y Técnica (Spanish Government) grants BMC2001-0772 and BOS2001-0794. O.L. was supported by a predoctoral fellowship from the Ministerio de Ciencia y Tecnología.

REFERENCES

Agerholm-Larsen B, Nordestgaard BG, Tybjaerg-Hansen A (2000) ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites. Arterioscler Thromb Vasc Biol 20:484-492

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805-1814

Barbujani G, Goldstein DB (2004) Africans and asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119-150

Barbujani GV (2000) Geographic patterns: how to identify them and why. Hum Biol 72:133-153

Boekholdt SM, Bijsterveld NR, Moons AH, Levi M, Buller HR, Peters RJ (2001) Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review. Circulation 104:3063-3068

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

Chakravarti AV (1999) Population genetics--making sense out of sequence. Nat Genet 21:56-60

Corbo RM, Scacchi R (1999) Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet 63 (Pt 4):301-310

de Lorgeril M, Salen P, Paillard F, Laporte F, Boucher F, de Leiris J (2002) Mediterranean diet and the French paradox: two distinct biogeographic concepts for one consolidated scientific theory on the role of nutrition in coronary heart disease. Cardiovasc Res 54:503-515

Evans A, Van Baal GC, McCarron P, DeLange M, Soerensen TI, De Geus EJ, Kyvik K, Pedersen NL, Spector TD, Andrew T, Patterson C, Whitfield JB, Zhu G,

Martin NG, Kaprio J, Boomsma DI (2003) The genetics of coronary heart disease: the contribution of twin studies. Twin Res 6:432-441

Excoffier L, Smouse PE, Quattro JMV (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491

Felsenstein J (1995) PHYLIP (phylogeny inference package) release 3.57, Seattle

Gopel W, Kim D, Gortner L (1999) Prothrombotic mutations as a risk factor for preterm birth. Lancet 353:1411-1412

Gopel W, Ludwig M, Junge AK, Kohlmann T, Diedrich K, Moller J (2001) Selection pressure for the factor-V-Leiden mutation and embryo implantation. Lancet 358:1238-1239

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45-61

Juul K, Tybjaerg-Hansen A, Steffensen R, Kofoed S, Jensen G, Nordestgaard BG (2002) Factor V Leiden: The Copenhagen City Heart Study and 2 meta-analyses. Blood 100:3-10

Kato N, Sugiyama T, Morita H, Kurihara H, Yamori Y, Yazaki Y (1999) Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies. J Hypertens 17:757-763

Kruskal JB, Wish M (1990) Multidimensional Scaling. Newbury Park, California

Lao O, Dupanloup I, Barbujani G, Bertranpetit J, Calafell F (2004) The European Paradox for risk factors in coronary heart disease extends to genetics. submited

Lucotte G, Mercier G (2001) Population genetics of factor V Leiden in Europe. Blood Cells Mol Dis 27:362-367

Lusis AJ, Mar R, Pajukanta P (2004) Genetics of atherosclerosis. Annu Rev Genomics Hum Genet 5:189-218

Lutz WJ (1995) The colonisation of Europe and our Western diseases. Med Hypotheses 45:115-120

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220

Nabel EG (2003) Cardiovascular disease. N Engl J Med 349:60-72

Nakajima T, Wooding S, Sakagami T, Emi M, Tokunaga K, Tamiya G, Ishigami T, Umemura S, Munkhbat B, Jin F, Guan-Jun J, Hayasaka I, Ishida T, Saitou N, Pavelka K, Lalouel JM, Jorde LB, Inoue I (2004) Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am J Hum Genet 74:898-916

Reynolds J, Weir B, Cockerham C (1983) Estimation for the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767-779

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani GV (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602-612

Rosenberg M (2001) Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 1.1. release 1.1, Tempe, AZ

Rosenberg N, Murata M, Ikeda Y, Opare-Sem O, Zivelin A, Geffen E, Seligsohn U (2002) The frequent 5,10-methylenetetrahydrofolate reductase C677T polymorphism is associated with a common haplotype in whites, Japanese, and Africans. Am J Hum Genet 70:758-762

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, et al. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. AmJ HumGenet 67:1526-1543

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PAV (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155-1159

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani GV (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262-278

Sokal RR, Harding RM, Oden NL (1989) Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267-294

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology 1. Metodology. Biological Journal of the Linnean Society 10:199-228

Stengard JH, Weiss KM, Sing CF (1998) An ecological study of association between coronary heart disease mortality rates in men and the relative frequencies of

common allelic variations in the gene coding for apolipoprotein E. Hum Genet 103:234-241

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, et al. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69:844-852

Tunstall-Pedoe H, Kuulasmaa K, Mahonen M, Tolonen H, Ruokokoski E, Amouyel P (1999) Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. Lancet 353:1547-1557

Wald DS, Law M, Morris JK (2002) Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. Bmj 325:1202

Wilson PW, Schaefer EJ, Larson MG, Ordovas JM (1996) Apolipoprotein E alleles and risk of coronary disease. A meta-analysis. Arterioscler Thromb Vasc Biol 16:1250-1255

Winkelmann BR, Hager J (2000) Genetic variation in coronary heart disease and myocardial infarction: methodological overview and clinical evidence. Pharmacogenomics 1:73-94

Woods DR, Montgomery HE (2001) Angiotensin-converting enzyme and genetics at high altitude. High Alt Med Biol 2:201-210

Young RP, Thomas GN, Critchley JA, Tomlinson B, Woo KS, Sanderson JE (1998) Interethnic differences in coronary heart disease mortality in 25 populations: association with the angiotensin-converting enzyme DD genotype frequency. J Cardiovasc Risk 5:303-307

Zetterberg H, Palmer M, Ricksten A, Poirier J, Palmqvist L, Rymo L, Zafiropoulos A, Arvanitis DA, Spandidos DA, Blennow K (2002) Influence of the apolipoprotein E epsilon4 allele on human embryonic development. Neurosci Lett 324:189-192

Figure 1.

Geographical distribution (a) and two different spatial autocorrelograms (b) and (c) of the ε4 allele of the APOE4 gene in 74 Middle Eastern and European populations. (b) was computed considering all populations and (c) without considering Saudi Arabia and the Volgo-Ural region. Due to the absence of population data the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 2.

Geographical distribution (a) and two different spatial autocorrelograms (b) and (c) of the absence of Alu insertion of the ACE gene in 116 Middle Eastern, North African and European populations. (b) was computed considering all populations and (c) only considering European and Middle Eastern populations. Due to the absence of population data the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 3.

Geographical distribution (a) and two different spatial autocorrelograms (b) and (c) of the A222V allele of MTHFR in 84 Middle Eastern and European populations. (b) was computed considering all populations and (c) without considering Azerbaijan. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 4.

Geographical distribution (a) and two different spatial autocorrelograms (b) and (c) of the R506Q (Leiden Mutation) polymorphism of the F5 gene in 71 Middle Eastern and European populations. Populations where the Leiden mutation is absent are marked with a cross. (b) was computed considering all populations and (c) without considering Saami populations. Due to the absence of population data, the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 5.

Geographical distribution (a) and two different spatial autocorrelograms (b) and (c) of the G20210A allele of F2 gene in 46 Middle Eastern and European populations. (b) was computed considering all populations and (c) without considering Saudi Arabia. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 6.

Geographical distribution (a) and spatial autocorrelogram (b) of the Q192R allele of PON1 gene in 27 Middle Eastern and European populations. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 7.

Geographical distribution (a) and spatial autocorrelogram (b) of the M235T allele of AGT gene in 29 Middle Eastern and European populations. Due to the absence of population data, the African continent is dashed and the Mediterranean islands have been excluded. X-axis represents geographic distance between samples; the Y-axis

represents Morans Index; numbers of the pairs of comparisons are indicated with the statistical significance; a single asterisk (*) denotes $p<.05$; double asterisks (**) denote $p<.01$; three asterisks (***) denote $p<.001$

Figure 8.

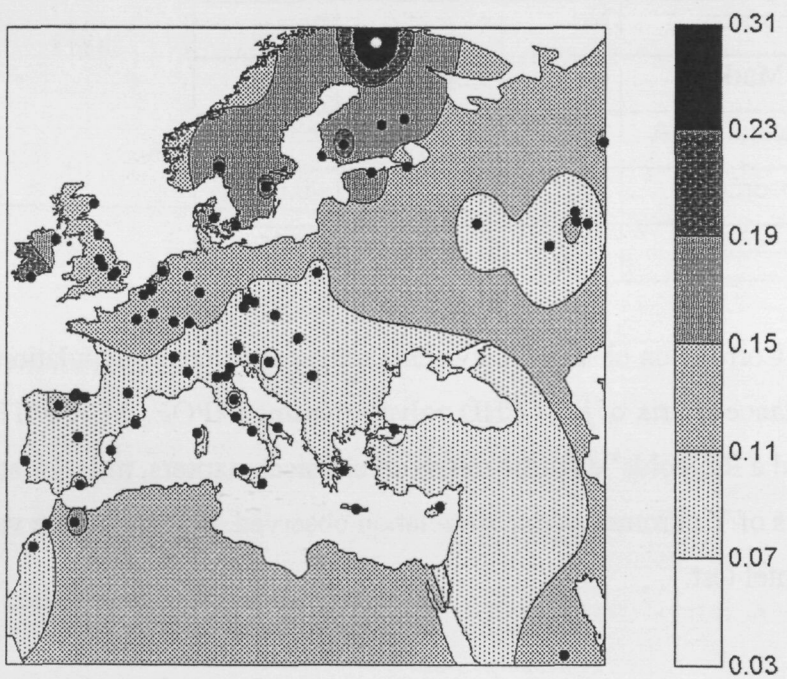
MDS plot based on a Reynolds' genetic distance matrix calculated from a joint subset of CHD genes: APOE, ACE, MTHFR, F5, F2. Populations have been clustered in three greographical groups: CW Europe, CE Europe and South Europe.

Loci	polymorphism	Chromosome localization	Pooled OR	reference	Average frequency of susceptibility allele	N populations	N chromosomes
ACE	I/D	17q23	1.21	*(Agerholm-Larsen et al. 2000)*	0.58	116	67674
APOE	ε2/ε3/ε4	19q13.2	1.26	(Wilson et al. 1996)	0.12	74	95520
MTHFR	A222V	1p36.3	1.21	*(Wald et al. 2002)*	0.33	84	47908
F5	R506Q	1q23	1.20	*(Juul et al. 2002)*	0.03	71	67288
F2	G20210A	11p11-q12	1.11	*(Boekholdt et al. 2001)*	0.01	46	28688
PON1	Q192R	7q21.3	1.44	(Agerholm-Larsen et al. 2000)	0.3	27	19812
AGT	M235T	1q42-q43	1.22	*(Kato et al. 1999)*	0.43	29	12236

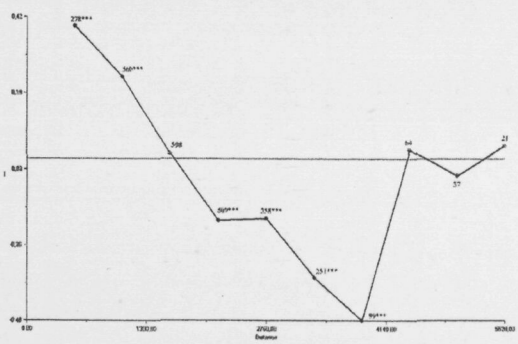
Table 1. Loci and CHD susceptibility alleles, chromosomal localization, pooled OR CHD ratios, the average frequency of the susceptibility alleles, the number of populations where frequency data was available and the number of chromosomes

Genetic markers	N	r	p
Classical Markers	15	0.36	<0.05
Mitochondrial DNA	18	0.088	0.313
Y chromosome1	15	0.32	<0.05
Y chromosome2	13	0.21	0.065

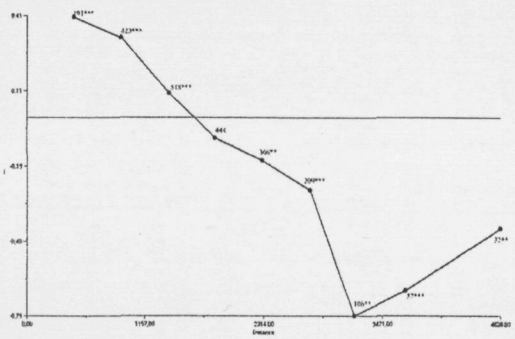
Table 2. Correlation observed between a subset of common populations between a Reynolds' distance matrix of joint CHD polymorphisms (APOE, ACE, MTHFR, Factor 5, F2) and a Reynolds' distance matrix of classical markers, mtDNA and two different datasets of Y chromosome. r, correlation observed. p, significance value obtained by Mantel test.



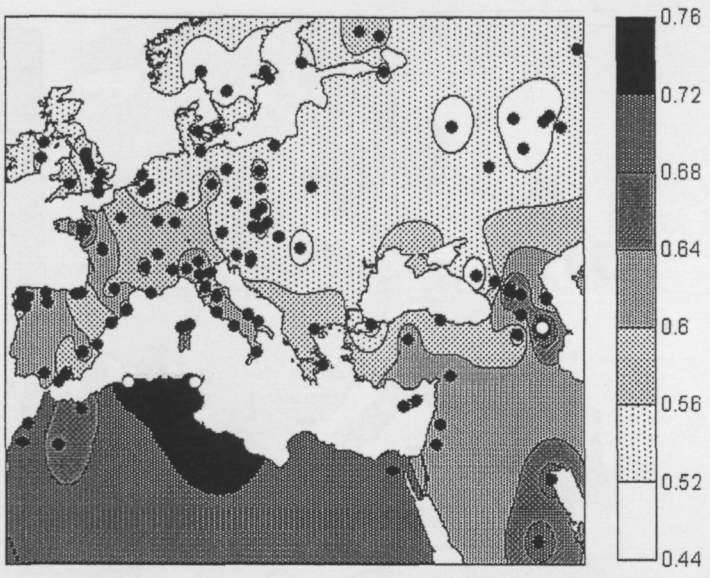
a)



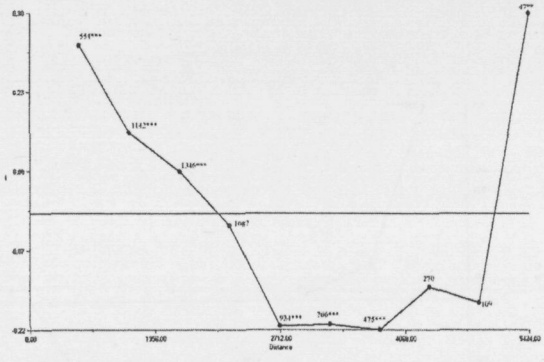
b)



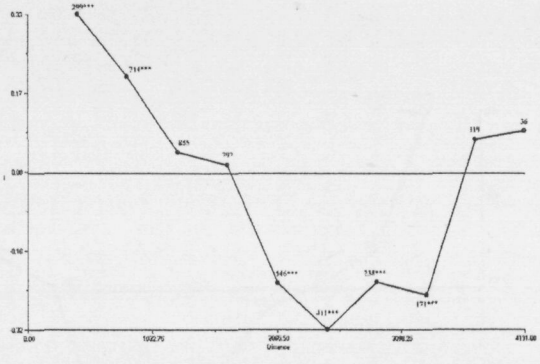
c)



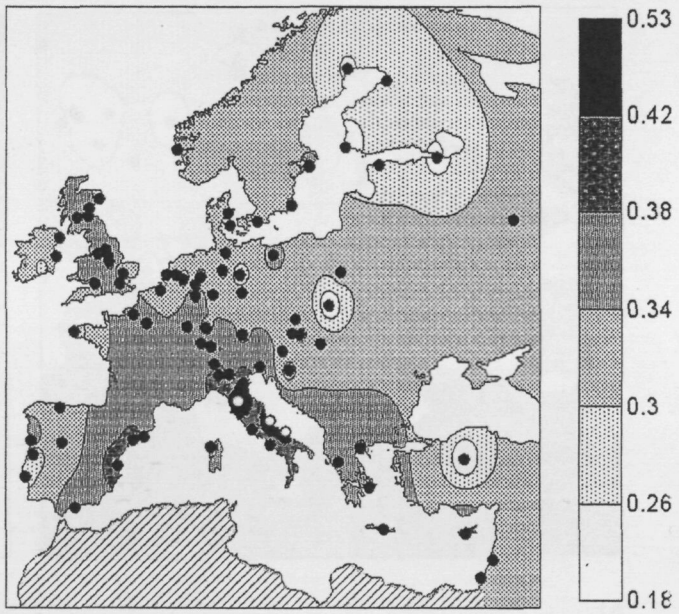
a)



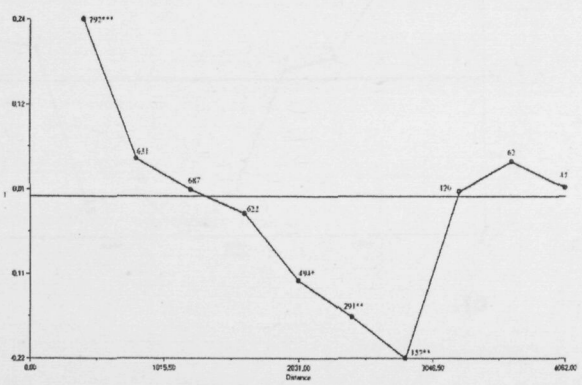
b)



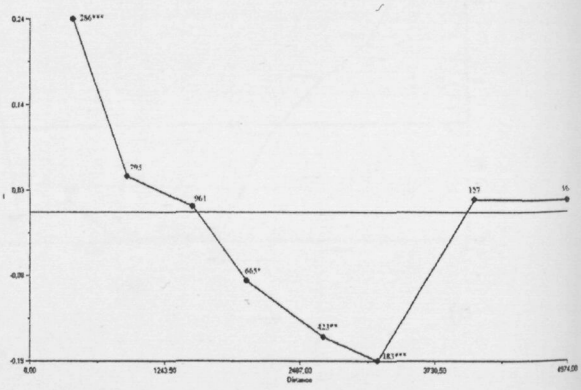
c)



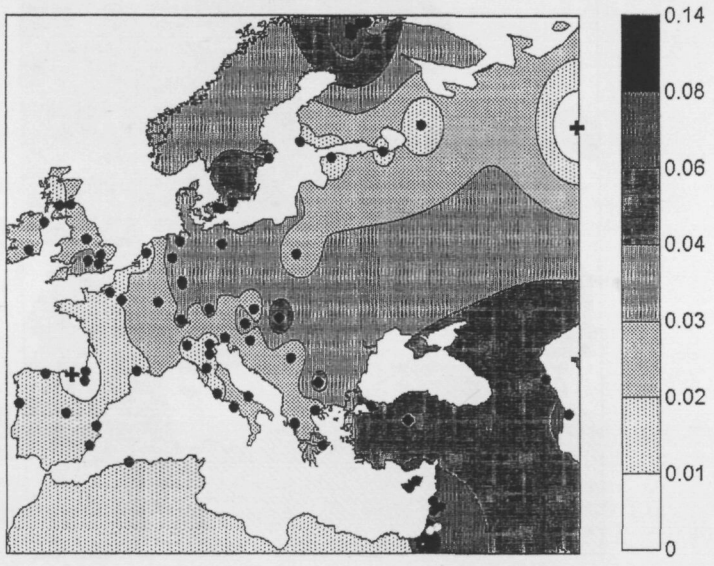
a)



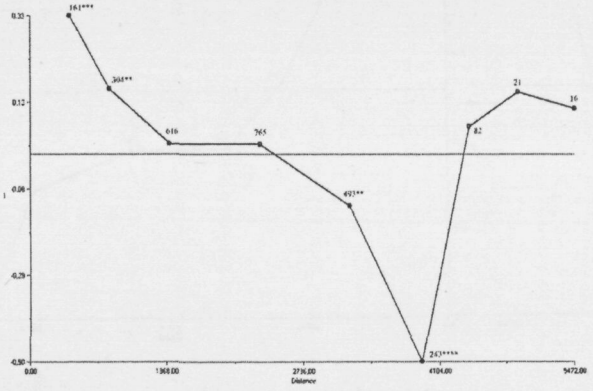
b)



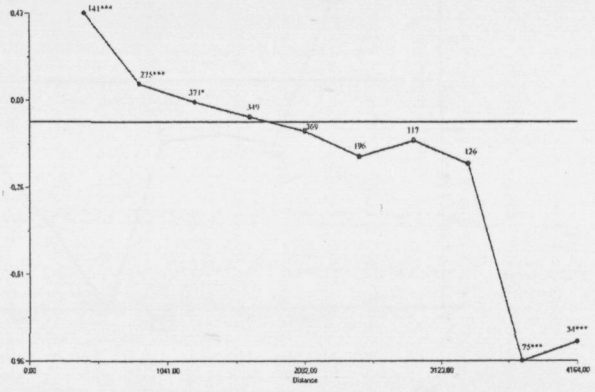
c)


0.14
0.08
0.06
0.04
0.03
0.02
0.01
0

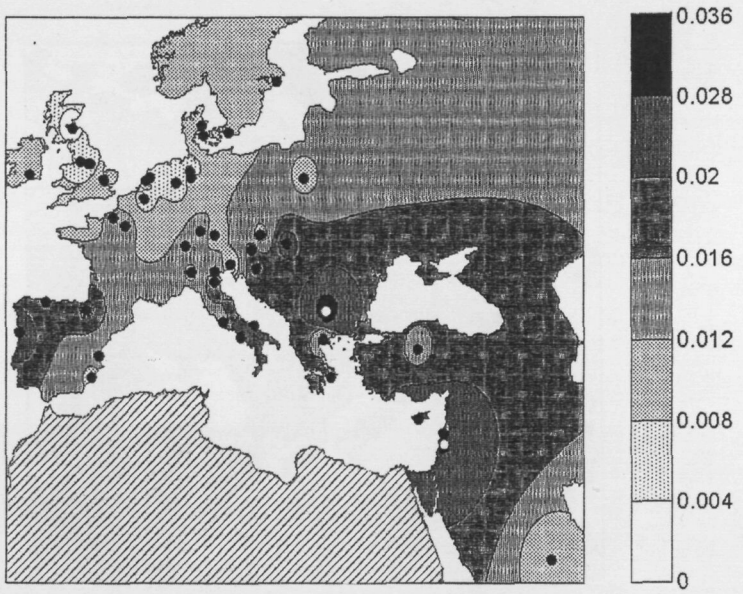
a)


161***
368**
616
765
493**
263****
82
21
16
0.30
0.12
-0.06
-0.29
-0.50
0.00
1368.00
2736.00
4104.00
5472.00
Distance

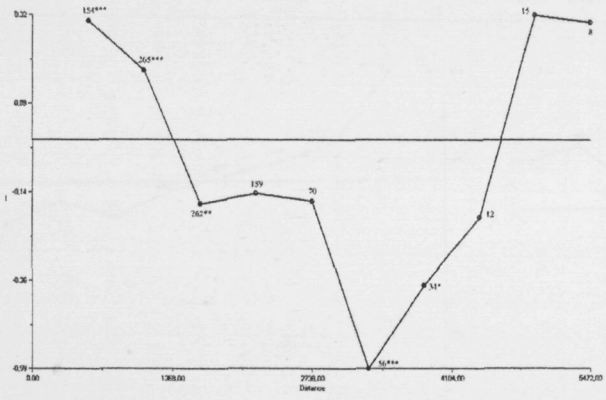
b)


141***
275***
371*
349
309
196
117
126
75***
34***
0.43
0.09
-0.25
-0.61
-0.95
0.00
1041.00
2082.00
3123.00
4164.00
Distance

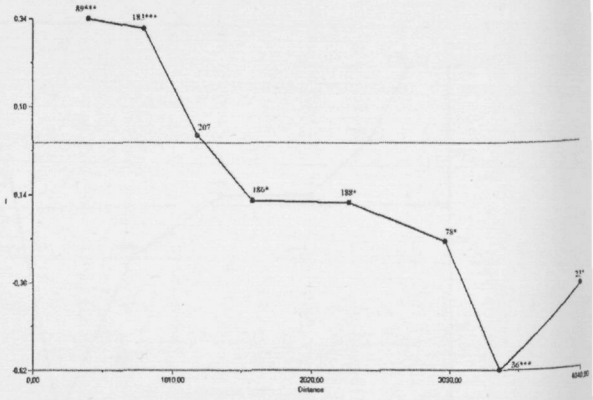
c)


0.036
0.028
0.02
0.016
0.012
0.008
0.004
0

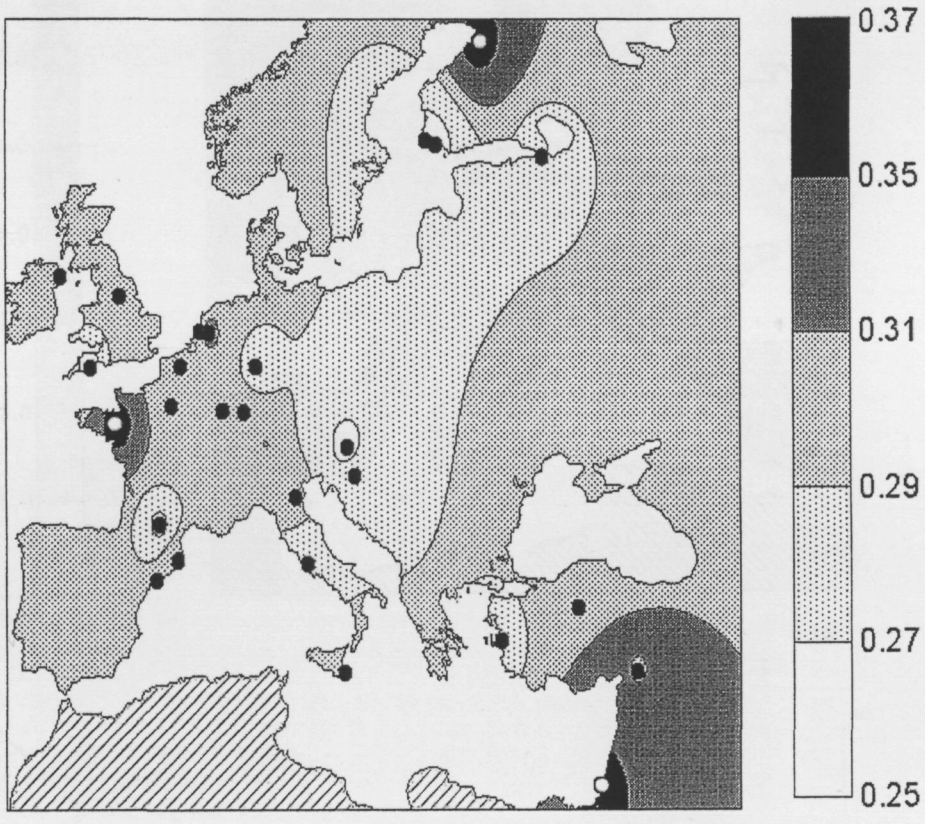
a)


154***
265***
282**
139
70
56***
31*
12
15
0.00
1368.00
2736.00
4104.00
5472.00
Distance

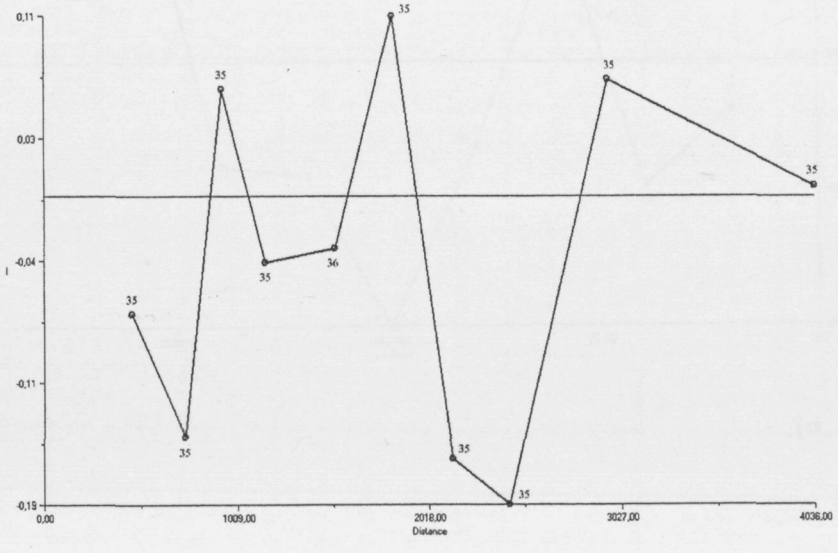
b)


89***
181***
207
160*
159*
78*
36***
0.00
1010.00
2020.00
3030.00
Distance

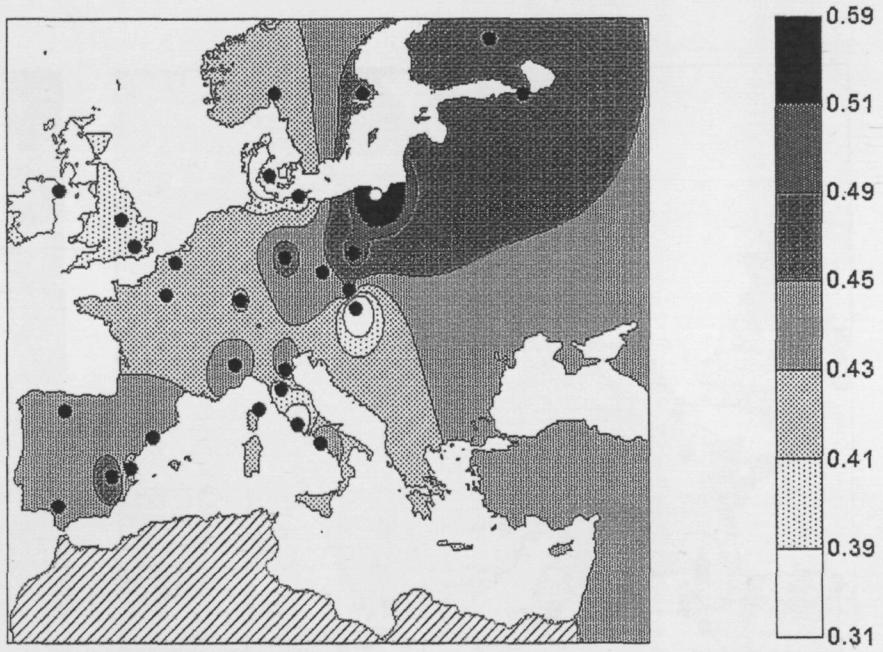
c)


0.37
0.35
0.31
0.29
0.27
0.25

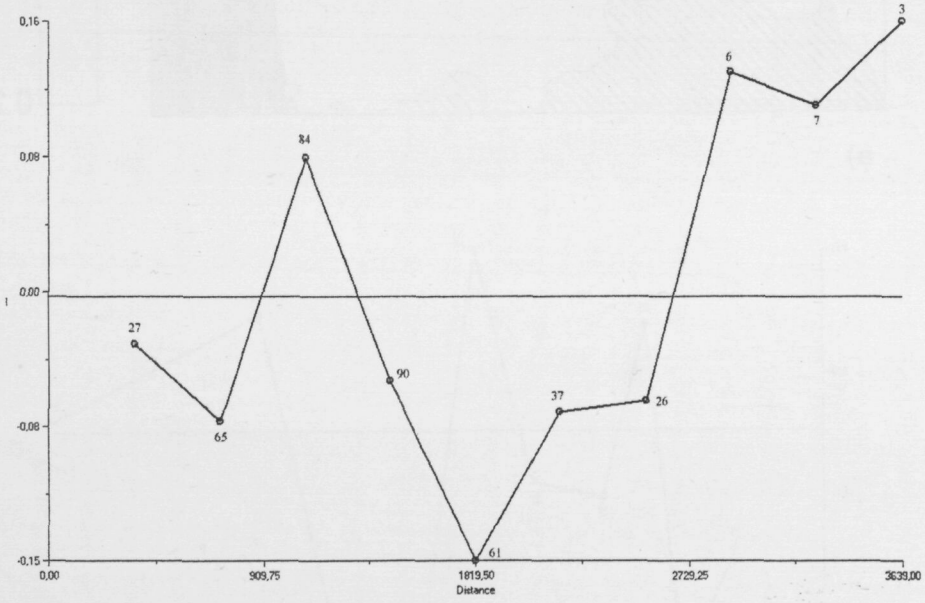
a)


0,11
0,03
-0,04
-0,11
-0,19
I
0,00
1009,00
2018,00
3027,00
4036,00
Distance
35
35
35
36
35
35
35
35
35

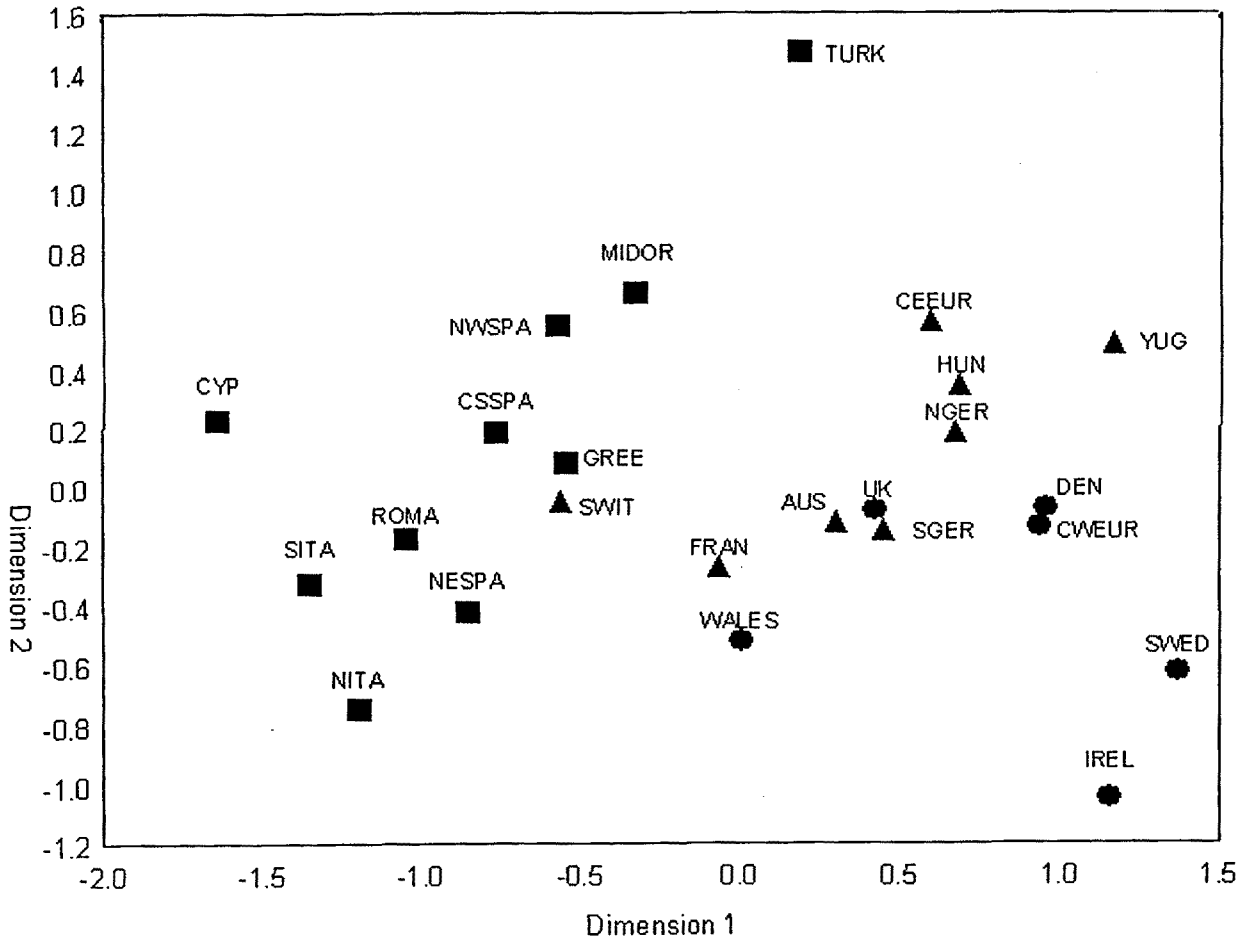
b)


0.59
0.51
0.49
0.45
0.43
0.41
0.39
0.31

a)


0,16
0,08
0,00
-0,08
-0,15
0,00
909,75
1819,50
2729,25
3639,00
Distance
27
65
84
90
61
37
26
6
7
3

b)



■ SEurope

▲ CEEurope

● CWEurope

4 Discussió

4.1.1 Anàlisi de la distribució espacial de les mutacions

L'anàlisi de la distribució espacial es va realitzar amb la freqüència relativa de les principals mutacions de cada malaltia. Per la fibrosi quística, cinc mutacions diferents es varen tenir en compte: 508Fdel, G542X, G551D, N1303K i W1285X; en el cas de la fenilcetonúria s'analitzà espacialment sis mutacions: R408W (recurrent a dos haplotips diferents; (Byck et al. 1994)), R261Q, IVS 10-11G>A, IVS12+1G>A i R158Q. Finalment, en el cas de la β-talassèmia s'estudià la distribució espacial de sis mutacions diferents: IVS 1-110G>A, IVS 11-1G>A, 41/42CTTT, IVS 1-5G>C, IVS 1-1G>A i CD39.

Un tret comú que es repeteix a les tres malalties és la presència de patrons clinals o parcialment clinals per a la majoria de les mutacions; aquesta característica és especialment marcada en aquelles mutacions amb una freqüència relativa superior al 7%, com és el cas de 508Fdel (amb una freqüència relativa mitjana del 60%) en la fibrosi quística, la mutació R408W associada a l'haplotip 2 (freqüència relativa mitjana del 24%), la mutació IVS 10-11G>A (freqüència relativa mitjana del 8%) i la mutació IVS12+1G>A (freqüència relativa mitjana del 8%) en el cas de la fenilcetonúria i la mutació IVS 1-110G>A (freqüència relativa mitjana del 14%), la mutació 41/42CTTT (freqüència relativa mitjana del 9%) i la mutació CD39 (freqüència relativa mitjana del 13%) en el cas de la β-talassèmia.

La presència de patrons clinals s'ha interpretat tradicionalment com el resultat de moviments migratoris i/o de la presència de fenòmens selectius que actuen diferencialment segons la localització geogràfica de les poblacions (Barbujani 2000); diferents factors poden explicar la falta de patrons espacials estructurats: la presència d'un paisatge genètic ja homogeni abans que es produïssin els moviments migratoris, la deriva i la mutació recurrent. En el cas de les variants associades a malalties mendelianes, tenint en compte que, a nivell absolut, presenten freqüències molt petites, sembla més probable que l'absència d'estructuració espacial en algunes mutacions es degui més a processos de deriva genètica i de mutació recurrent (vegeu, per exemple, (Byck et al. 1997)).

La distribució geogràfica de variants rares o poc comunes sovint s'ha associat amb determinats fenòmens migratoris; poblacions com celtes, vikings, fenicis, romans i

grecs (entre d'altres) han estat àmpliament utilitzades com explicació *ad hoc* de la distribució d'una determinada variant (preferentment associada a algun procés patològic) com, per exemple, la mutació C282Y causant d'hemocromatosi i el seu origen celta o viking (Distante et al. 2004) o el polimorfisme "viking" de la deleció de 32 parells de bases en el gen CCR5 (Lucotte and Dieterlen 2003). Aquest fenomen ha estat d'especial rellevància en les malalties mendelianes, on és un costum força típic etiquetar una determinada mutació amb el nom d'una població o civilització antiga quan ambdues distribucions geogràfiques se solapen. En trobem un elevat nombre d'exemples de mutacions amb "denominació d'origen" en les malalties mendelianes estudiades. Així, per exemple, en el cas de la fenilcetonúria s'ha dit que la mutació R408W associada a l'haplotip 1 que es troba present principalment a Irlanda i poblacions veïnes és d'origen "celta" o que les mutacions F299C, R408Q i Y414C són "vikingues" ja que es troben principalment a Noruega (Zschocke et al. 1997); d'altres, com les mutacions IVS 10-11G>A i L48S, freqüents a poblacions mediterrànies, tindrien el seu origen a l'orient mitjà (Perez et al. 1997). En el cas de la fibrosi quística, es diu que la mutació G542X és d'origen fenici donat que la seva freqüència relativa és més elevada en antics assentaments fenicis (Loirat et al. 1997).

Fer inferències sobre l'origen d'una mutació (és a dir, fer "*story telling*") sense cap altra evidència que la seva distribució geogràfica és una font potencial d'errors (Goldstein and Chikhi 2002) ja que donat l'elevat nombre de civilitzacions que han aparegut i declinat al llarg de la història i de l'elevada mobilitat geogràfica que han mostrat les poblacions humanes al continent Europeu, no resulta pas difícil trobar una civilització o població antiga que mostri un patró de distribució semblant al d'una variant determinada (Sokal 1991); vegeu http://life.bio.sunysb.edu/ee/msr/ethno.html). D'altra banda, l'origen de les variants al.lèliques acostuma a ser molt anterior a l'origen de les poblacions en les quals després s'hi troben (Barbujani and Goldstein 2004); així, per exemple, la mutació 508del causant de fibrosi quística es troba en un haplotip que no es troba a les poblacions actuals (Mateu et al. 2002) i ha estat datada entre els ~50,000 anys (Morral et al. 1994) i els ~10,000 anys (Slatkin and Bertorelle 2001). Altres mutacions com, per exemple, la mutació causant d'hiperfenilalaninèmia no PKU I65T i la causant de fenilcetonúria R408W associada a l'haplotip 1 podrien tenir edats també molt antigues (O'Donnell et al. 2002). A aquesta observació s'hi suma el fet que les anàlisis es realitzen amb la freqüència relativa de la mutació, calculada a partir del grup de cromosomes associats a la malaltia genètica i no a partir del grup total de

cromosomes. L'aparició d'una determinada mutació en una població on no hi havia cap (per tant, amb freqüència relativa de 1) i la posterior migració a poblacions veïnes on ja n'hi ha (per tant, tindrà una freqüència menor de 1) pot crear un patró clinal que serà exactament el mateix si una de les mutacions presents a una població (i per tant, amb una freqüència relativa menor de 1) migra cap a una població veïna on no hi havia cap (on tindrà una freqüència relativa propera a 1); una situació semblant s'observarà en el cas de migracions d'una població amb una freqüència absoluta de les mutacions elevada cap a poblacions amb una freqüència absoluta més petita. Finalment, donada la baixa freqüència absoluta de les mutacions associades a les malalties mendelianes, intuïtivament sembla més lògic pensar que els patrons geogràfics de moltes d'elles seran deguts principalment migracions de curt abast i al biaix a favor de la seva detecció que no pas a grans moviments de poblacions. En tot cas, si la distribució espacial d'una determinada mutació realment reflexa un fenomen migratori, la petjada d'aquesta s'ha de veure també en la variabilitat genètica de loci selectivament neutres.

4.1.2 Comparació de les diferents malalties i altres loci

En els tres estudis hem comparat mitjançant el test de Mantel (vegeu material i mètodes) matrius de distàncies genètiques calculades entre parells de poblacions per l'espectre de mutacions causant de la malaltia genètica i loci considerats neutres (i, per tant, deutors de la seva distribució espacial a processos demogràfics), com els marcadors autosòmics clàssics (Cavalli-Sforza et al. 1994), el cromosoma Y (Rosser et al. 2000) i el DNA mitocondrial (Simoni et al. 2000). Donat que molts dels polimorfismes clàssics comprenen diferents grups sanguinis (com el sistema ABO) i els grups HLA, s'ha postulat que molts d'ells podrien estar involucrats en resistència a patogens (Cavalli-Sforza et al. 1994; Jobling et al. 2004) i, per tant, no serien neutres. Per un altre costat, tant el cromosoma Y com el DNA mitocondrial són d'herència uniparental (el cromosoma Y s'hereta per la línia paterna i el DNA mitocondrial per la materna), cosa que implica que la seva grandària efectiva és molt menor que la dels marcadors autosòmics (per cada cromosoma Y i DNA mitocondrial que s'hereta n'hi ha quatre d'autosòmics) i l'efecte de la deriva és major (Perez-Lezaun et al. 1999); aquest efecte es trobaria encara més accentuat pel fet que, tradicionalment, els homes haurien estat polígams, el que hauria reduït encara més la grandària efectiva del cromosoma Y

(Dupanloup et al. 2003). A més, la comparació de la distribució geogràfica dels polimorfismes associats al cromosoma Y i el DNA mitocondrial sembla suggerir que les dones tradicionalment han mostrat una major mobilitat geogràfica que els homes (Perez-Lezaun et al. 1999; Oota et al. 2001). En el cas de la fibrosi quística, només vàrem obtenir una correlació positiva i estadísticament significativa quan es comparà amb el DNA mitocondrial, degut principalment que les diferents variants del DNA mitocondrial presenten una distribució força homogènia al continent europeu (Simoni et al. 2000). La mutació més freqüentment associada a la fibrosi quística, la mutació 508Fdel, és ubiqua a les poblacions europees i, malgrat que presenta un patró clinal, és també força homogènia (vegeu capítol I). En el cas de la fenilcetonúria es va obtenir una correlació positiva i estadísticament significativa quan es comparà amb el cromosoma Y; en aquest cas tant la variació del cromosoma Y com la de la fenilcetonúria a Europa estan espacialment ben delimitades i, a més, el patró de distribució espacial d'algunes mutacions es solapa amb el d'alguns haplogrups del cromosoma Y (com, per exemple, la mutació R408W associada a l'haplotip 2 i l'haplogrup HG16 (Rosser et al. 2000)). En el cas de la β-talassèmia es varen obtenir correlacions positives i estadísticament significatives entre les matrius calculades a partir de l'espectre de mutacions i els altres loci emprats. Aquest resultat es pot explicar perquè posteriorment a la diàspora de la població humana fora d'Àfrica (vegeu introducció) es donaren fenòmens d'aïllament geogràfic, cosa que propicià la diferenciació genètica dels marcadors neutres entre poblacions molt allunyades; aquest patró també s'observa en les mutacions associades a la β-talassèmia, probablement molt més joves que la majoria dels polimorfismes analitzats.

4.1.3 La diversitat genètica de les malalties mendelianes

Com prèviament ja hem comentat, la utilització de la distribució espacial d'una única mutació sense cap altra evidència no és una bona estratègia per fer inferències sobre la història poblacional de les mutacions. El model de Reich i Lander (vegeu introducció), d'altra banda, estableix una dependència entre la diversitat genètica del patró de mutacions causants de malaltia i paràmetres poblacionals (com la grandària efectiva de la població) i genòmics (com la taxa de mutació el coeficient de selecció en contra dels al.lels):

$$\varphi = \frac{1}{1+4N\mu(1-f_0)}$$

Ja que per una determinada malaltia mendeliana és poc probable que la taxa de mutació variï entre les poblacions, les diferències en el patró de les mutacions s'han de deure a diferències en la grandària efectiva entre les poblacions o a incidències diferents (i, en última instància, al coeficient de selecció; vegeu introducció). Ara bé, donat que les malalties mendelianes són molt poc freqüents, és fàcil veure que per obtenir un mateix canvi en la diversitat gènica caldran canvis molt més dràstics en la incidència que en la grandària efectiva de la població i, per tant, sembla més probable que siguin els canvis demogràfics els que modularan la diversitat del patró d'al.lels causants d'una determinada malaltia, malgrat que puguin haver-hi factors selectius diferencials a les poblacions. Cal tenir en compte que aquest model assumeix que tots els al.lels tenen el mateix coeficient de selecció; com ja hem comentat (vegeu l'apartat "la complexitat de les malalties mendelianes"), la gravetat fenotípica depèn d'un elevat nombre de factors ambientals i genètics; entre aquests últims s'inclou com afecta la mutació a l'expressió, estabilitat i funcionalitat de la proteïna i, per tant, sembla poc realista assumir un únic coeficient de selecció. En el cas de les malalties mendelianes estudiades es va intentar homogeneitzar aquest coeficient seleccionant aquelles mutacions associades a un fenotip més greu. En el cas de la fibrosi quística, es varen tenir en compte aquelles mutacions causants d'un fenotip greu, però també mutacions lleus, ja que els homes amb mutacions patològiques lleus al gen CFTR acostumen a tenir problemes de fertilitat i, per tant, evolutivament són equivalents a les que produeixen la mort de l'individu abans d'arribar a l'edat reproductiva. En el cas de la fenilcetonúria i la β-talassèmia, es va realitzar una selecció d'aquelles mutacions associades predominantment a un fenotip més greu (vegeu l'apartat de materials i mètodes). És evident que aquesta selecció, malgrat que totalment necessària, és força grollera perquè l'efecte fenotípic dels individus afectats heterozigots es deu a l'efecte dels dos al.lels, el qual, a més de la quantitat del producte produït per cada variant, també inclou fenòmens d'interacció entre els dímers de proteïnes. Una aproximació més òptima i acurada seria estudiar la quantitat de producte i la seva funcionalitat enzimàtica (Pey et al. 2003).

L'anàlisi de la distribució espacial de la diversitat de l'espectre de mutacions causants de fibrosi quística i fenilcetonúria a Europa mostrà que la diversitat d'aquestes

malalties és més elevada al sud d'Europa que al nord. D'acord amb el model de Reich i Lander, això voldria dir que, tradicionalment, la grandària efectiva hauria estat més gran al sud d'Europa que al nord. Diversos processos demogràfics podrien explicar aquesta conclusió; com ja hem comentat anteriorment (vegeu l'apartat "història del continent europeu" a la introducció), les poblacions del sud d'Europa històricament han estat sotmeses a processos que tendeixen a incrementar la diversitat genètica i la grandària efectiva, com són la mescla (en el neolític, per exemple) o unes condicions climàtiques més favorables (per exemple, durant l'últim període glacial) que haurien permès sostenir un major nombre d'individus. Ara bé, quan comparem el patró de la diversitat genètica de la fibrosi quística amb el de la fenilcetonúria, no obtenim pas una correlació positiva, que és el que hom esperaria tenint en compte que les poblacions estudiades són les mateixes i, per tant, els processos demogràfics són comuns; malgrat que la diversitat genètica és més elevada al sud que al nord d'Europa en ambdues malalties, en el cas de la fibrosi quística, la diversitat genètica és més petita en el nord-oest i en el cas de la fenilcetonúria en el nord-est d'Europa. Com es pot explicar aquesta discrepància? Primer de tot, cal tenir present que la incidència de les dues malalties no és la mateixa. La incidència de la fibrosi quística és ~4 vegades més elevada que la de la fenilcetonúria. Factors estocàstics o factors demogràfics, com pot ser el nivell de consanguinitat, poden tenir un impacte molt més gran a mesura que la incidència es fa menor (mireu figura 29)

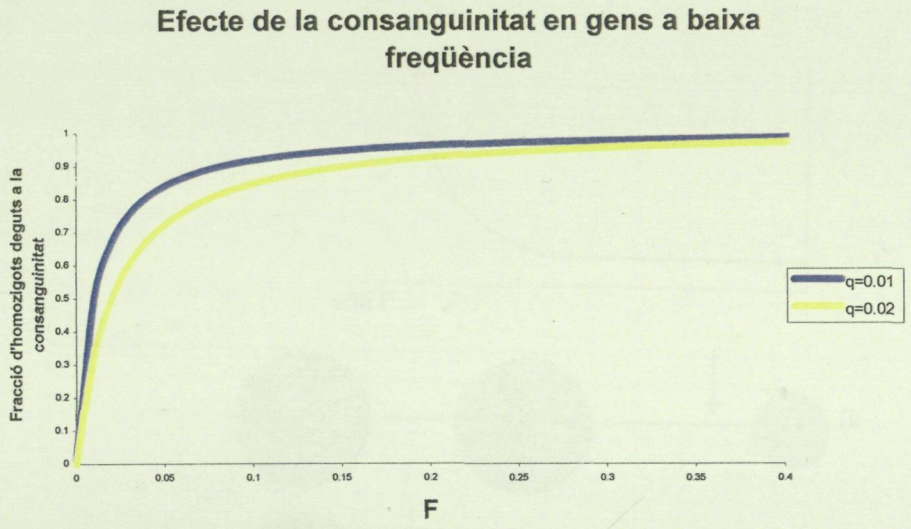


Figura 29. Fracció de la incidència deguda al coeficient de consanguinitat (F). Per valors de consanguinitat baixos (<0.1) típics de les poblacions humanes (Cavalli-Sforza and Bodmer 1981), la fracció de la incidència deguda a la consanguinitat és més elevada quan la freqüència gènica de la malaltia mendeliana és més baixa

Per un altre costat, d'acord amb el model de Reich i Lander, una incidència més petita voldria dir que la taxa de selecció és molt més elevada i/o una taxa de mutació més baixa a la fenilcetonúria que a la fibrosi quística; malgrat que no podem afirmar que les taxes de mutació siguin les mateixes en ambdós gens (degut a la presència diferencial de dinucleòtids CpG en ambdós gens molt probablement no ho seran), discursos selectius a favor de l'heterozigot s'han proposat tant per l'una com per l'altra, ja que sembla poc probable que les elevades incidències observades es puguin explicar simplement per fenòmens de deriva genètica (vegeu apartats fibrosi quística i fenilcetonúria). Donat que la població humana ha patit una recent expansió, el moment en el qual es va iniciar aquesta selecció equilibradora podria haver influït sobre el patró mutacional de la malaltia: si la selecció equilibradora hagués estat present abans de l'expansió poblacional, esperaríem trobar un patró de la diversitat genètica simple (és a dir, poques mutacions molt freqüents i moltes de molt poc freqüents), mentre que si la selecció a favor de l'heterozigot s'hagués donat durant l'expansió de les poblacions, llavors esperaríem trobar un patró de mutacions complex (moltes mutacions equifrequents), ja que la selecció equilibradora només hauria incrementat les freqüències absolutes però la diversitat genètica seria la prèviament modificada per la selecció purificadora (vegeu figura 30)

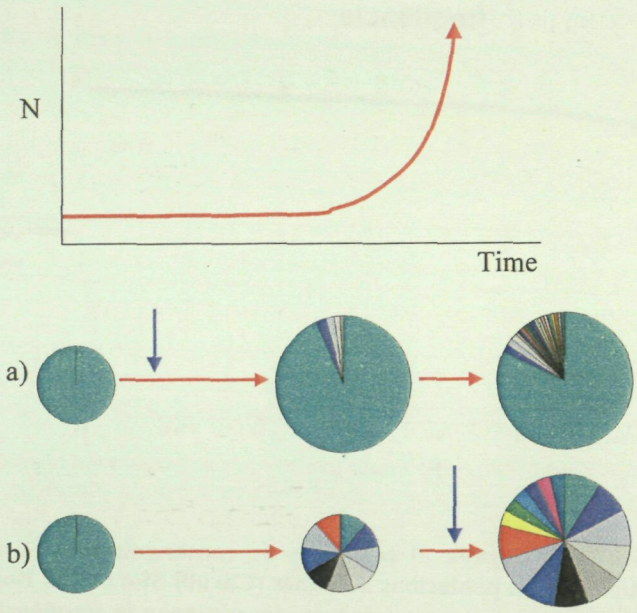


Figura 30. La selecció a favor de l'heterozigot (entesa com que els individus portadors d'un al.lel "sa" i un al.lel causant de malaltia són més resistents que els individus portadors de dos al.lels causants de malaltia o cap) en un principi només incrementa la freqüència d'aquells al.lels que en el moment de l'esdeveniment estan presents a la població (a les figures a i b l'increment de la freqüència es representa com un increment en la grandària dels pastissos), però depenent de quan es produeix pot donar patrons de mutacions molt diferents. Inicialment, com que la grandària demogràfica era petita, en equilibri deriva-mutació-selecció el nombre d'al.lels que es podrien mantenir eren pocs i, per això, tant en la situació a com b es parteix d'un únic al.lel (en verd) que copsa tota la diversitat genètica dintre de la malaltia. En a) l'event de selecció a favor de l'heterozigot (fletxa de color blau) es produeix abans de l'expansió de les poblacions i, per tant, s'incrementa la freqüència absoluta de l'únic al.lel que en aquell moment hi ha. Quan es produeix l'expansió de les poblacions apareixen noves mutacions, però respecte a l'antiga tenen una freqüència absoluta petita i el patró de mutació només en representen una petita fracció. El patró és simple. En b) el coeficient de selecció en contra de les variants associades a la malaltia fa que el reemplaçament sigui ràpid i la incidència molt petita. Quan es produeix l'expansió demogràfica apareixen noves variants. Com que les variants antigues ja estaven a una baixa freqüència absoluta, el resultat és un patró complex. Quan en aquesta situació es produeix el fenomen de selecció a favor de l'heterozigot s'incrementa la freqüència absoluta de les variants que ja hi eren present i, per tant, es manté el patró complex.

Una conseqüència d'aquesta situació és que en el cas de la fenilcetonúria hom esperaria uns nivells de diversitat genètica molt més grans dels que li correspondrien per la seva incidència. De fet, quan mirem els nivells de diversitat genètica amb la incidència d'altres malalties mendelianes (vegeu taula), l'heterozigositat de la fenilcetonúria és més elevada de l'esperat, però la diferència no és estadísticament significativa (vegeu capítol II); probablement si les estimes d'incidència de les malalties fossin més acurades i es realitzés l'anàlisi amb més malalties mendelianes, aquesta diferència seria estadísticament significativa. A aquest resultat s'hi afegeix el fet que no

hi ha cap mutació que domini el patró mutacional de la fenilcetonúria i que geogràficament es defineixen grups de poblacions genèticament homogènies mitjançant SAMOVA (vegeu material i mètodes), cosa que indica que les mutacions no han tingut prou temps per expandir-se (ja sigui per grans migracions o per petits moviments) i, per tant, que en conjunt no arriben al nivell d'antiguitat de la mutació 508Fdel que domina el patró de les mutacions de la fibrosi quística.

En el cas de la β-talassèmia, l'anàlisi espacial de la diversitat genètica es va dur a terme després de definir cinc grups de poblacions que maximitzesin la diferenciació genètica entre ells mitjançant l'algoritme de SAMOVA. Aquest primer pas va ser totalment necessari ja que la distribució de les mutacions a l'espai semblava indicar que l'origen de la β-talassèmia no havia estat pas únic a totes les poblacions del món d'acord amb estudis previs ((Weatherall 2001); vegeu figura 22). SAMOVA definí tres grans grups continentals (vegeu capítol III): Mediterrani i pròxim orient, sud d'Àsia i sud-est d'Àsia, més dos grups compresos per un petit nombre de poblacions (Sardenya i rodalies i Indonèsia); el percentatge de variació explicat entre els diferents grups fou de més del 30%, un valor molt gran si el comparem amb l'observat normalment en els marcadors neutres (vegeu apartat "races, gens i malaltia genètica"). Aquests tres grans grups continentals no s'han d'interpretar pas com una evidència a favor de la presència de barreres reproductores entre individus de diferents grups, si no més bé com una prova de l'origen recent de la β-talassèmia, posterior a la diàspora africana, cosa que faria que les mutacions estiguessin geogràficament ben localitzades i, per tant, que les distàncies genètiques entre poblacions allunyades fossin més grans que pels marcadors neutres, normalment amb un origen més antic.

L'anàlisi de la variabilitat genètica dels tres grups continentals revelà que, en el cas del grup que comprenia la regió del mediterrani i el pròxim orient la distribució geogràfica de la diversitat mostrà un patró parcialment clinal amb valors més elevats a la zona del Creixent Fèrtil i més petits a mesura que ens allunyàvem. A partir del model de Reich i Lander, aquesta diferència de diversitat genètica es podria explicar tant per factors demogràfics com selectius. La β-talassèmia, a l'igual que passa amb altres hemoglobinopaties (vegeu introducció "β-talassèmia"), podria estar conferint una major resistència en els individus heterozigots a la infecció per malària, una malaltia produïda per diverses espècies de protozous del gènere *Plasmodium* (la forma més greu està produïda per *Plasmodium falciparum*) que es transmet amb una major o menor eficàcia

depenent de l'espècie de mosquit vector (Miller et al. 2002). L'origen de la interacció entre els humans i el *Plasmodium falciparum* sembla ser molt antic, trobant-se en la separació del gènere *Homo* amb el de *Pan* (gènere del ximpanzé) (Rich and Ayala 2000). Les dades moleculars semblen apuntar que el protozou *Plasmodium falciparum* s'hauria expandit recentment a partir d'una única població africana del protozou (Conway et al. 2000); aquesta ràpida expansió s'hauria vist afavorida per un increment en la densitat de les poblacions humanes i un hàbitat més propici per al creixement dels mosquits, degut en gran part al descobriment i desenvolupament de l'agricultura durant el Neolític. Com ja hem comentat (vegeu l'apartat de història del continent europeu), l'agricultura comportà un creixement demogràfic sense precedents en la història de les poblacions humanes i la creació d'assentaments estables densament poblats, així com una modificació dràstica de l'habitat per tal d'adaptar-lo als conreus i als animals domesticats, cosa que resultà òptim per la propagació del paràsit i el creixement de les espècies antropofíliques dels mosquits vectors (Rich and Ayala 2000). Donat que l'agricultura s'estengué al continent europeu a partir de la regió del Creixent Fèrtil, és d'esperar que la malària es donés primer en aquesta regió i, per tant, també la selecció a favor de l'heterozigot de la β-talassèmia. Segons el model de Reich i Lander, hom esperaria trobar uns nivells de diversitat genètica més elevats en aquells llocs on la selecció deletèria hagués estat més forta (i/o on hagués durat més), precisament a les regions més allunyades del Creixent Fèrtil; és a dir, que si només hagués estat actuant la selecció, hom esperaria trobar un patró amb un sentit contrari al trobat. Donat que l'increment demogràfic primer es donà a la regió del Creixent Fèrtil, sembla més plausible que aquesta gradació de la diversitat genètica sigui deguda més a motius demogràfics que a motius selectius. L'absència de patrons clinals de la diversitat genètica en els altres dos grups podria deure's a una història demogràfica més complexa (vegeu, per exemple (Cavalli-Sforza et al. 1994)), però no podem pas descartar que els factors confusors com els que s'han apuntat abans també estiguin actuant en el cas de la β-talassèmia, malgrat que la seva elevada incidència fa que aquesta possibilitat sigui menys probable.

4.2 Malalties complexes: la malaltia coronària

4.2.1 Distribució geogràfica de les diferents variants

En el treball de la malaltia coronària (capítols IV i V) hem analitzat el patró de distribució geogràfic de variants associades a la malaltia coronària en gens que estan implicats en els diferents processos fisiopatològics involucrats en el desenvolupament de la malaltia (vegeu introducció). Dels set gens analitzats espacialment al continent europeu, nord d'Àfrica i Pròxim Orient, en cinc vàrem obtenir patrons clinals o parcialment clinals: la variant 4 de l'APOE, la variant D de l'ACE, la variant 222V de MTHFR, la variant Leiden del F5 i la variant 20210A de la protrombina, precisament aquells gens amb un major nombre de poblacions (>40; vegeu capítol IV). Aquest nombre elevat de poblacions ens va permetre calcular una matriu de distàncies genètiques entre els parells de poblacions comuns als cinc gens; la seva representació gràfica mostrà que les poblacions es distribuïren seguint un patró sud-est/nord-oest, semblant al que s'ha observat en molts altres loci nuclears (Barbujani and Goldstein 2004). Aquesta primera impressió es confirmà quan comparàrem la matriu de distàncies dels cinc loci conjunts amb una matriu de distàncies genètiques calculada per a les mateixes poblacions amb marcadors clàssics (vegeu capítol IV). Tots aquests resultats semblen indicar que, a nivell espacial, els polimorfismes estudiats es comporten com a marcadors clàssics, típicament considerats neutres (però vegeu apartats anteriors); malgrat que aquest resultat es pot interpretar pels mateixos processos demogràfics que han modelat la diversitat genètica a Europa, no exclou pas la possibilitat que altres processos selectius hagin donat lloc a les distribucions espacials observades; donat que els gens estudiats estan implicats en la regulació i homeostasi de processos fisiològics essencials per a la vida, no és d'estranyar que s'hagin proposat explicacions selectives per a molts d'ells (vegeu capítol IV). D'altra banda, alguns d'ells, com el polimorfisme I/D del gen ACE, s'han emprat tradicionalment com a marcadors neutres (vegeu, per exemple, Romualdi i col.laboradors (Romualdi et al. 2002) o la base de dades ALFRED) i d'altres, com la mutació Leiden del F5, s'han associat a determinats processos demogràfics com el neolític (Blood Cells Mol Dis. 2001 Mar-Apr;27(2):362-7). Per a aquests cinc gens vàrem obtenir dades a nivell mundial, cosa que ens permeté

estudiar la seva diferenciació a nivell continental; els nivells de diferenciació genètica observats són molt baixos i semblats als obtinguts per a un elevat nombre de SNPs (Akey et al. 2002), cosa que sembla indicar que d'haver-hi fenòmens selectius aquests no actuen diferencialment a nivell continental.

Donat que la distribució de la incidència de CHD a Europa és també clinal amb un patró nord-sud (article Lancet), es podria explicar la seva distribució a partir de la dels al.lels associats a CHD. Una hipòtesi proposada inicialment per Lutz (Lutz 1995) postula que la diferència d'incidència de CHD a Europa es deu principalment al canvi de dieta dels caçadors-recolectors cap a la dieta dels agricultors produïda durant el neolític. D'acord amb un model de difusió dèmica mixt, en el qual les poblacions d'agricultors del neolític s'haurien mesclat parcialment a mesura que ens allunyem del Creixent Fèrtil (vegeu introducció; (Dupanloup et al. 2004)), en el nord-oest d'Europa s'hi trobaria una fracció més elevada d'al.lels "preneolítics"; l'elevada incidència de CHD en aquesta regió es deuria, principalment, que aquests al.lels estarien adaptats a la dieta típica dels caçadors-recolectors (és a dir, serien al.lels estalviadors (vegeu introducció)) que, en trobar-se en un canvi brusc de dieta cap a la dels agricultors, conferirien una major susceptibilitat a patir CHD. Ara bé, cal tenir present que, a part de la presència de patrons clinals en la mateixa direcció espacial que la incidència de CHD, l'increment de la freqüència de l'al.lel de susceptibilitat també ha d'ésser en el mateix sentit que el de la incidència de CHD.

4.2.2 Anàlisi de la covariació en l'espai de CHD i els polimorfismes de susceptibilitat

Per tal de veure si la incidència de la malaltia coronària variava amb els canvis de la freqüència dels genotips de susceptibilitat, vàrem estudiar com correlacionaven ambdos factors en les mateixes poblacions. Hom esperaria que si els polimorfismes de susceptibilitat tenen un pes elevat en el desenvolupament de la malaltia es trobés una correlació positiva i estadísticament significativa, mentre que en el cas que el polimorfisme només confereixi un petit increment a l'hora de patir la malaltia coronària s'obtindria una correlació propera a 0 i no estadísticament significativa. Donat que els al.lels estudiats confereixen un risc modest a patir la malaltia, aquesta és la situació que hom esperaria més probable. En principi, no s'esperaria trobar correlacions negatives

5 BIBLIOGRAFIA

Agerholm-Larsen B, Nordestgaard BG, Tybjaerg-Hansen A (2000) ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites. Arterioscler Thromb Vasc Biol 20:484-492

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805-1814

Anton SC (2003) Natural history of Homo erectus. Am J Phys Anthropol Suppl 37:126-170

Antonarakis SE, Krawczak M, Cooper DN (2000) Disease-causing mutations in the human genome. Eur J Pediatr 159 Suppl 3:S173-178

Arcos-Burgos M, Muenke M (2002) Genetics of population isolates. Clin Genet 61:233-247

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309

Badano JL, Katsanis N (2002) Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 3:779-789

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598-609

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72:578-589

Barbujani G (2000) Geographic patterns: how to identify them and why. Hum Biol 72:133-153

Barbujani G, Goldstein DB (2004) Africans and asians abroad: genetic diversity in Europe. Annu Rev Genomics Hum Genet 5:119-150

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci U S A 94:4516-4519

Becker KG (2004) The common variants/multiple disease hypothesis of common complex genetic disorders. Med Hypotheses 62:309-317

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. Ciba FoundSymp 197:97-114

Boekholdt SM, Bijsterveld NR, Moons AH, Levi M, Buller HR, Peters RJ (2001) Genetic variation in coagulation and fibrinolytic proteins and their relation with acute myocardial infarction: a systematic review. Circulation 104:3063-3068

BIBLIOGRAFIA

Botto LD, Yang Q (2000) 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol 151:862-877

Brookes AJ (1999) The essence of SNPs. Gene 234:177-186

Byck S, Morgan K, Tyfield L, Dworniczak B, Scriver CR (1994) Evidence for origin, by recurrent mutation, of the phenylalanine hydroxylase R408W mutation on two haplotypes in European and Quebec populations. Hum Mol Genet 3:1675-1677

Byck S, Tyfield L, Carter K, Scriver CR (1997) Prediction of multiple hypermutable codons in the human PAH gene: codon 280 contains recurrent mutations in Quebec and other populations. Hum Mutat 9:316-321

Calafell F (2003) Classifying humans. Nat Genet 33:435-436

Campbell H, Rudan I (2002) Interpretation of genetic association studies in complex disease. Pharmacogenomics J 2:349-360

Caramelli D, Lalueza-Fox C, Vernesi C, Lari M, Casoli A, Mallegni F, Chiarelli B, Dupanloup I, Bertranpetit J, Barbujani G, Bertorelle G (2003) Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. Proc Natl Acad Sci U S A 100:6593-6597

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135-140

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446-452

Carluccio M, Soccio M, De Caterina R (2001) Aspects of gene polymorphisms in cardiovascular disease: the renin-angiotensin system. Eur J Clin Invest 31:476-488

Cavalli-Sforza LL, Bodmer WF (1981) Genética de las poblaciones humanas. Ediciones Omega

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl:266-275

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

Chakravarti A (1999) Population genetics--making sense out of sequence. Nat Genet 21:56-60

Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002) Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci U S A 99:11008-11013

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869-872

Collins A, Ennis S, Taillon-Miller P, Kwok PY, Morton NE (2001) Allelic association with SNPs: metrics, populations, and the linkage disequilibrium map. Hum Mutat 17:255-262

Conway DJ, Fanello C, Lloyd JM, Al-Joubori BM, Baloch AH, Somanath SD, Roper C, Oduola AM, Mulder B, Povoa MM, Singh B, Thomas AW (2000) Origin of Plasmodium falciparum malaria is traced by mitochondrial DNA. Mol Biochem Parasitol 111:163-171

Cooper RS (2003) Gene-environment interactions and the etiology of common complex disease. Ann Intern Med 139:437-440

Corbo RM, Scacchi R (1999) Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet 63 (Pt 4):301-310

Crow JF (1997) The high spontaneous mutation rate: is it a health risk? Proc Natl Acad Sci U S A 94:8380-8386

Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1:40-47

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York

De Bree A, Verschuren WM, Kromhout D, Kluijtmans LA, Blom HJ (2002) Homocysteine determinants and the evidence to what extent homocysteine determines the risk of coronary heart disease. Pharmacol Rev 54:599-618

Dean M, Carrington M, O'Brien SJ (2002) Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3:263-292

Dipple KM, McCabe ER (2000a) Modifier genes convert "simple" Mendelian disorders to complex traits. Mol Genet Metab 71:43-50

Dipple KM, McCabe ER (2000b) Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. Am J Hum Genet 66:1729-1735

BIBLIOGRAFIA

Distante S, Robson KJ, Graham-Campbell J, Arnaiz-Villena A, Brissot P, Worwood M (2004) The origin and spread of the HFE-C282Y haemochromatosis mutation. Hum Genet [Epub ahead of print]

Dupanloup I, Bertorelle G, Chikhi L, Barbujani G (2004) Estimating the impact of prehistoric admixture on the genome of europeans. Mol Biol Evol 21:1361-1372

Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, Amorim A, Barbujani G (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57:85-97

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17:68-74

Edwards AW (2003) Human genetic diversity: Lewontin's fallacy. Bioessays 25:798-801

Egan ME, Pearson M, Weiner SA, Rajendran V, Rubin D, Glockner-Pagel J, Canny S, Du K, Lukacs GL, Caplan MJ (2004) Curcumin, a major constituent of turmeric, corrects cystic fibrosis defects. Science 304:600-602

Eisensmith RC, Goltsov AA, O'Neill C, Tyfield LA, Schwartz EI, Kuzmin AI, Baranovskaya SS, Tsukerman GL, Treacy E, Scriver CR, et al. (1995) Recurrence of the R408W mutation in the phenylalanine hydroxylase locus in Europeans. Am J Hum Genet 56:278-286

Endler G, Mannhalter C (2003) Polymorphisms in coagulation factor genes and their impact on arterial and venous thrombosis. Clin Chim Acta 330:31-55

Erlandsen H, Stevens RC (1999) The structural basis of phenylketonuria. Mol Genet Metab 68:103-125

Estivill X, Bancells C, Ramos C (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. Hum Mutat 10:135-154

Excoffier L (2002) Human demographic history: refining the recent African origin model. Curr Opin Genet Dev 12:675-682

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

Falls JG, Pulford DJ, Wylie AA, Jirtle RL (1999) Genomic imprinting: implications for human disease. Am J Pathol 154:635-647

Foley R (1998) The context of human genetic evolution. Genome Res 8:339-347

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881-900

Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266:107-109

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3:129-152

Haapanen-Niemi N, Vuori I, Pasanen M (1999) Public health burden of coronary heart disease risk factors among middle-aged and elderly men. Prev Med 28:343-348

Harpending H, Rogers A (2000) Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361-385

Hartl DL, Clark AG (1997) Principles of Population Genetics. Sinauer Associates, Inc., Sunderland, Massachusetts

Hauser ER, Pericak-Vance MA (2000) Genetic analysis for common complex disease. Am Heart J 140:S36-44

Hey J, Machado CA (2003) The study of structured populations--new hope for a difficult and divided science. Nat Rev Genet 4:535-543

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45-61

Hollander DH (1982) Etiogenesis of the European cystic fibrosis polymorphism: heterozygote advantage against venereal syphilis? Med Hypotheses 8:191-197

Horrevoets AJ (2004) Plasminogen activator inhibitor 1 (PAI-1): in vitro activities and clinical relevance. Br J Haematol 125:12-23

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306-309

Jamrozik K, Spencer CA, Lawrence-Brown MM, Norman PE (2001) Does the Mediterranean paradox extend to abdominal aortic aneurysm? Int J Epidemiol 30:1071-1075

Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409:853-855

BIBLIOGRAFIA

Jobling MA, Hurles ME, Tyler-Smith C (2004) Human evolutionary genetics. Origins, peoples&disease. Garland Science, New York

Juul K, Tybjaerg-Hansen A, Steffensen R, Kofoed S, Jensen G, Nordestgaard BG (2002) Factor V Leiden: The Copenhagen City Heart Study and 2 meta-analyses. Blood 100:3-10

Kalaydjieva L, Gresham D, Calafell F (2001) Genetic studies of the Roma (Gypsies): a review. BMC Med Genet 2:5

Kato N, Sugiyama T, Morita H, Kurihara H, Yamori Y, Yazaki Y (1999) Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies. J Hypertens 17:757-763

Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5:413-424

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Vol 22. Oxford University Press, Oxford

Kim RJ, Becker RC (2003) Association between factor V Leiden, prothrombin G20210A, and methylenetetrahydrofolate reductase C677T mutations and events of the arterial circulatory system: a meta-analysis of published studies. Am Heart J 146:948-957

Kittles RA, Weiss KM (2003) Race, ancestry, and genes: implications for defining disease risk. Annu Rev Genomics Hum Genet 4:33-67

Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474-488

Krawczak M, Zschocke J (2003) A role for overdominant selection in phenylketonuria? Evidence from molecular data. Hum Mutat 21:394-397

Kruskal JB, Wish M (1990) Multidimensional scaling. Newbury Park, California

La Du BN (1992) Pharmacogenetics of drug metabolism. Pergamon Press, New York

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177-182

Loirat F, Hazout S, Lucotte G (1997) G542X as a probable Phoenician cystic fibrosis mutation. Hum Biol 69:419-425

Lucotte G, Dieterlen F (2003) More about the Viking hypothesis of origin of the delta32 mutation in the CCR5 gene conferring resistance to HIV-1 infection. Infect Genet Evol 3:293-295

Lusis AJ, Mar R, Pajukanta P (2004) Genetics of atherosclerosis. Annu Rev Genomics Hum Genet 5:189-218

Lutz WJ (1995) The colonisation of Europe and our Western diseases. Med Hypotheses 45:115-120

Lynn A, Ashley T, Hassold T (2004) Variation in human meiotic recombination. Annu Rev Genomics Hum Genet 5:317-349

MacGregor AJ, Snieder H, Schork NJ, Spector TD (2000) Twins. Novel uses to study complex traits and genetic diseases. Trends Genet 16:131-134

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512-517

Mateu E, Calafell F, Ramos MD, Casals T, Bertranpetit J (2002) Can a place of origin of the main cystic fibrosis mutations be identified? Am J Hum Genet 70:257-264

Meindl RS (1987) Hypothesis: a selective advantage for cystic fibrosis heterozygotes. Am J Phys Anthropol 74:39-45

Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415:673-679

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek M, Jr., Kalaydjieva L, et al. (1994) The origin of the major cystic fibrosis mutation (delta F508) in European populations. Nat Genet 7:169-175

Nabel EG (2003) Cardiovascular disease. N Engl J Med 349:60-72

Nabholz CE, von Overbeck J (2004) Gene-environment interactions and the complexity of human genetic diseases. J Insur Med 36:47-53

Nakajima T, Wooding S, Sakagami T, Emi M, Tokunaga K, Tamiya G, Ishigami T, Umemura S, Munkhbat B, Jin F, Guan-Jun J, Hayasaka I, Ishida T, Saitou N, Pavelka K, Lalouel JM, Jorde LB, Inoue I (2004) Natural selection and

population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am J Hum Genet 74:898-916

Norio R (2003) Finnish Disease Heritage I: characteristics, causes, background. Hum Genet 112:441-456

O'Donnell KA, O'Neill C, Tighe O, Bertorelle G, Naughten E, Mayne PD, Croke DT (2002) The mutation spectrum of hyperphenylalaninaemia in the Republic of Ireland: the population history of the Irish revisited. Eur J Hum Genet 10:530-538

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:20-21

Paabo S (2003) The mosaic that is our genome. Nature 421:409-412

Pennisi E (2003) Human genome. A low number wins the GeneSweep Pool. Science 300:1484

Perez B, Desviat LR, Ugarte M (1997) Analysis of the phenylalanine hydroxylase gene in the Spanish population: mutation profile and association with intragenic polymorphic markers. Am J Hum Genet 60:95-102

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208-219

Pey AL, Desviat LR, Gamez A, Ugarte M, Perez B (2003) Phenylketonuria: genotype-phenotype correlations based on expression analysis of structural and functional mutations in PAH. Hum Mutat 21:370-378

Pier GB, Grout M, Zaidi TS, Olsen JC, Johnson LG, Yankaskas JR, Goldberg JB (1996) Role of mutant CFTR in hypersusceptibility of cystic fibrosis patients to lung infections. Science 271:64-67

Pinhasi R, Foley RA, Lahr MM (2000) Spatial and temporal patterns in the mesolithic-neolithic archaeological record of Europe. In: Archaeogenetics: DNA and the population prehistory of Europe. McDonald Institute for archaeological research, Cambridge, pp 342

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11:2417-2423

Project TIH (2003) The International HapMap Project. Nature 426:789-796

Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol 20:76-86

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502-510

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32:135-142

Rich SM, Ayala FJ (2000) Population structure and recent evolution of Plasmodium falciparum. Proc Natl Acad Sci U S A 97:6994-7001

Risch N, Burchard E, Ziv E, Tang H (2002) Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3:comment2007

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Risch N, Tang H, Katzenstein H, Ekstein J (2003) Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. Am J Hum Genet 72:812-822

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602-612

Rosenberg MS (2000) The bearing correlogram: a new method of analyzing directional spatial autocorrelation. Geographical Analysis 32:267-278

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, et al. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. AmJ HumGenet 67:1526-1543

Saugstad LF (1973) Increased "reproductive casualty" in heterozygotes for phenylketonuria. Clin Genet 4:105-114

BIBLIOGRAFIA

Saugstad LF (1977) Heterozygote advantage for the phenylketonuria allele. J Med Genet 14:20-24

Scheuner MT (2003) Genetic evaluation for coronary artery disease. Genet Med 5:269-285

Schols L, Bauer P, Schmidt T, Schulte T, Riess O (2004) Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. Lancet Neurol 3:291-304

Scriver CR (2001) Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet 2:69-101

Scriver CR, Sly WS, Childs B, Beaudet AL, Valle D, Kinzler KW, Vogelstein B (2000) The Metabolic and Molecular Bases of Inherited Diseases. McGraw-Hill Professional

Scriver CR, Waters PJ (1999) Monogenic traits are not simple: lessons from phenylketonuria. Trends Genet 15:267-272

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155-1159

Sharma AM (1998) The thrifty-genotype hypothesis and its implications for the study of complex genetic disorders in man. J Mol Med 76:568-571

Shier WT (1979) Increased resistance to influenza as a possible source of heterozygote advantage in cystic fibrosis. Med Hypotheses 5:661-667

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262-278

Slatkin M, Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics 158:865-874

Smith DJ, Lusis AJ (2002) The allelic structure of common disease. Hum Mol Genet 11:2455-2461

Sokal RR (1991) Ancient movement patterns determine modern genetic variances in Europe. Hum Biol 63:589-606

Sokal RR, Harding RM, Oden NL (1989) Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267-294

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology 1. Metodology. Biological Journal of the Linnean Society 10:199-228

Sokal RR, Oden NL, Thomson BA (1999) A problem with synthetic maps. Hum Biol 71:1-13; discussion 15-25

Stengard JH, Weiss KM, Sing CF (1998) An ecological study of association between coronary heart disease mortality rates in men and the relative frequencies of common allelic variations in the gene coding for apolipoprotein E. Hum Genet 103:234-241

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-493

Stiner MC (2001) Thirty years on the "broad spectrum revolution" and paleolithic demography. Proc Natl Acad Sci U S A 98:6993-6996

Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J (2003) Structural location of disease-associated single-nucleotide polymorphisms. J Mol Biol 327:1021-1030

Strachan T, Read AP (1999) Human Molecular Genetics 2. John Wiley & Sons Inc, New York

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340

Tunstall-Pedoe H, Kuulasmaa K, Mahonen M, Tolonen H, Ruokokoski E, Amouyel P (1999) Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations. Monitoring trends and determinants in cardiovascular disease. Lancet 353:1547-1557

Waters PJ (2001) Degradation of mutant proteins, underlying "loss of function" phenotypes, plays a major role in genetic disease. Curr Issues Mol Biol 3:57-65

Weatherall DJ (2001) Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. Nat Rev Genet 2:245-255

BIBLIOGRAFIA

Wheeler JG, Keavney BD, Watkins H, Collins R, Danesh J (2004) Four paraoxonase gene polymorphisms in 11212 cases of coronary heart disease and 12786 controls: meta-analysis of 43 studies. Lancet 363:689-695

Wilson PW, Schaefer EJ, Larson MG, Ordovas JM (1996) Apolipoprotein E alleles and risk of coronary disease. A meta-analysis. Arterioscler Thromb Vasc Biol 16:1250-1255

Winkelmann BR, Hager J, Kraus WE, Merlini P, Keavney B, Grant PJ, Muhlestein JB, Granger CB (2000) Genetics of coronary heart disease: current knowledge and research principles. Am Heart J 140:S11-26

Winter EE, Goodstadt L, Ponting CP (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res 14:54-61

Wolf U (1997) Identical mutations and phenotypic variation. Hum Genet 100:305-321

Woolf LI, McBean MS, Woolf FM, Cahalane SF (1975) Phenylketonuria as a balanced polymorphism: the nature of the heterozygote advantage. Ann Hum Genet 38:461-469

Wright S (1938) Size of population and breeding structure in relation to evolution. Science 87:430-431

Zivelin A, Griffin JH, Xu X, Pabinger I, Samama M, Conard J, Brenner B, Eldor A, Seligsohn U (1997) A single genetic origin for a common Caucasian risk factor for venous thrombosis. Blood 89:397-402

Zlotogora J (2003) Penetrance and expressivity in the molecular age. Genet Med 5:347-352

Zschocke J, Mallory JP, Eiken HG, Nevin NC (1997) Phenylketonuria and the peoples of Northern Ireland. Hum Genet 100:189-194