LUÍSA MARIA de SOUSA MESQUITA PEREIRA

A Genetic Portrait of Portugal in Iberian, European and African Frameworks as Drawn from the Study of Y-Chromosome and Mitochondrial DNA Polymorphisms

Porto
2001

LUÍSA MARIA de SOUSA MESQUITA PEREIRA

A Genetic Portrait of Portugal in Iberian, European and African Frameworks as Drawn from the Study of Y-Chromosome and Mitochondrial DNA Polymorphisms

Dissertation presented to the Faculty of Sciences from the University of Porto to obtain the PhD in Biology

Dissertação apresentada à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Doutor em Biologia

Porto

2001

AGRADECIMENTOS - ACKNOWLEDGEMENTS 1

RESUMO 5

SUMMARY 9

RÉSUMÉ 13

INTRODUCTION

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS 19
1- The Y-chromosome 19
2- The mtDNA 21
3- Y-chromosome and mtDNA insights into Population Genetics 24
4- Comparing results under the same statistic method – mismatch distribution analysis 26

II- POPULATION GENETICS' APPLICATIONS 28
1- Population Genetics - some insights into its potential contribution to the Anthropological debate 28
2- Genetic and historic backgrounds of Portugal and Mozambique 31
2.1- Portugal 32
2.1.1- The country context 32
2.1.2- The Iberian context 35
2.1.3- The European context 38
2.2- Portugal – as contributor for... 38
2.2.1- Mozambique 40
2.2.1.2- Before the Portuguese arrival – the Bantu expansion 41
2.2.1.2- After the Portuguese arrival – Mozambican contribution to the slave trade 42

MATERIAL AND METHODS

1- Population samples 45
2- DNA extraction 45
3- Y-BMs screening 45
4- MtDNA screening (HVRI, HVRII and RFLPs) 47

RESULTS

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS 49

ARTICLE 1 53

PEREIRA, L., DUPANLOUP, I., ROSSER, Z.H., JOBLING, M.A., BARBUJANI, G. (2001) Y-chromosome mismatch distributions in Europe. *Mol. Biol. Evol.* **18**:1259-1271.

ARTICLE 2 67

DUPANLOUP, I., PEREIRA, L., BERTORELLE, G., CALAFELL, F., PRATA, M.J., AMORIM, A., BARBUJANI, G. No evidence of demographic expansions in human Y-chromosome biallelic variation. (in preparation).

ARTICLE 3 87

PEREIRA, L., PRATA, M.J., AMORIM, A. (2002) An evaluation of the proportion of identical Y-STR haplotypes due to recurrent mutation. In: Sensabaugh G.F., Lincoln, P.J., Olaisen, B. (eds) *Progress in Forensic Genetics* 9 (in press). Elsevier Science, Amsterdam.

II- Population Genetics' Applications 91

A- Portugal – the country context 91

ARTICLE 4 95

PEREIRA, L., PRATA M.J., JOBLING M.A., AMORIM, A. (2000) Analysis of the Y-chromosome and Mitochondrial DNA pools in Portugal. In Renfrew C., Boyle K. (eds) Archaeogenetics: DNA and the population prehistory of Europe. Chapter **20**:191-195. McDonald Institute Monographs. Oxbow Books, Cambridge.

ARTICLE 5 101

PEREIRA, L., PRATA, M.J., AMORIM, A. (2000) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann. Hum. Genet.* **64**:491 -506.

B- Portugal – the Iberian context 121

ARTICLE 6 125

PEREIRA, L., PRATA, M.J., BRIÓN, M., JOBLING, M.A., CARRACEDO, A., AMORIM, A. (2000) Clinal variation of the YAP^+ Y chromosome frequencies in Western Iberia. *Hum. Biol.* **72**: 937-944.

ARTICLE 7 133

PEREIRA, L., MACAULAY, V., PRATA, M.J., AMORIM, A. (2002) Phylogeny of the mtDNA haplogroup U6. Analysis of the sequences observed in North Africa and Iberia. In: Sensabaugh G.F., Lincoln, P.J., Olaisen, B. (eds) *Progress in Forensic Genetics* 9 (in press). Elsevier Science, Amsterdam.

C- Portugal – the European context 137

ARTICLE 8 141

ROSSER, Z.H., ZERJAL, T., HURLES, M.E., ADOJAAN, M.A., ALAVANTIC, D., AMORIM, A., AMOS, W., ARMENTEROS, M., ARROYO, E., BARBUJANI, G., BECKMAN, L., BERTRANPETIT, J., BOSCH, E., BRADLEY, D.G., BREDE, G., COOPER, G.C., CÔRTE-REAL, H.B.S.M., DE KNIJFF, P., DECORTE, R., DUBROVA, Y.E., EVGRAFOV, O., GILISSEN, A., GLISIC, S., GÖLGE, M., HILL, E.W., JEZIOROWSKA, A., KALAYDJIEVA, L., KAYSER, M., KRAVCHENCO, S.A., LAVINHA, J., LIVSHITS, L.A., MARIA, S., MCELREAVEY, K., MEITINGER,T.A., BELA MELEGH, B., MITCHELL, R.J., NICHOLSON, J., NØRBY, S., NOVELLETTO, A., PANDYA, A., PARIK, J., PATSALIS, P.C., PEREIRA, L., PETERLIN, B., PIELBERG, G., PRATA, M.J., PREVIDERÉ, C., RAJCZY, K., ROEWER, L., ROOTSI, S., RUBINSZTEIN, D.C., SAILLARD, J., SANTOS, F.R., SHLUMUKOVA, M., STEFANESCU, G., SYKES, B.C., TOLUN, A., VILLEMS, R., TYLER-SMITH, C., JOBLING, M.A. (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. *Am. J. Hum. Genet.* **67**:1526-1543.

D- Portugal – as contributor for Mozambique 159

ARTICLE 9 163

PEREIRA, L., MACAULAY, V., TORRONI, A., SCOZZARI, R., PRATA, M.J., AMORIM, A. (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann. Hum. Genet.* **65** (in press).

ARTICLE 10 201

PEREIRA, L., GUSMÃO, L., ALVES, C., AMORIM, A., PRATA, M.J. Y-chromosome pool in the southeastern African population of Mozambique: the small European influence and the Bantu diversity reduction. (in preparation).

CONCLUSIONS

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS 225

1- The Y-BM mismatch distribution 225

2- The mismatch distribution as an evaluation method for the proportion of Y-STR haplotypes that can become identical-by-state 227

II- POPULATION GENETICS' APPLICATIONS 228

A- Portugal – the country context 228

B- Portugal – the Iberian context 229

C- Portugal – the European context 230

D- Portugal – as contributor for Mozambique 231

REFERENCES 235

AGRADECIMENTOS - ACKNOWLEDGEMENTS

Se condensar os resultados acumulados ao longo de quatro anos não é uma tarefa fácil, menos o é a tentativa de concentrar sentimentos, emoções e convivências experimentadas ao longo desse período. E como só disponho de estas breves páginas para o fazer, perdoem-me a "explosão" que se segue.

À minha orientadora, Doutora Maria João Prata, agradeço ter-me facultado um doutoramento nos moldes que eu ambicionava e cujo melhor resumo penso ser o seguinte: apoio na independência. Ficou-me uma sensação excitante e vertiginosa de crescimento pessoal e científico. Espero, sinceramente, que a partilhe.

Ao Professor Doutor António Amorim, responsável pela Unidade de Genética Populacional e vice-presidente do IPATIMUP, local onde desenvolvi a totalidade do trabalho apresentado nesta tese, agradeço o seguinte facto, por si só esclarecedor: apesar de não ter sido meu co-orientador formalmente, foi-o realmente. Agradeço a paciência e a confiança incutidas. É que há chefes que por acaso são amigos e há amigos que por acaso são chefes.

Ao Professor Doutor Manuel Sobrinho Simões, na qualidade de presidente do IPATIMUP, agradeço a oportunidade que me proporcionou de fazer investigação num meio em que tal surge de uma forma saudável e livre dos inúmeros, enfadonhos e limitantes aspectos burocráticos que, de outro modo, teria que enfrentar. Poder usufruir de um ambiente de camaradagem, cooperação e boa disposição é, sem dúvida, um privilégio, e isso agradeço a todos os amigos e colegas do IPATIMUP.

Não posso deixar, no entanto, de personalizar o meu agradecimento a todos os elementos do Grupo da Genética Populacional, ou mais afectuosamente, GEPO. Começo pela Cíntia Alves e Leonor Gusmão: agradeço toda a colaboração (e à Cíntia a preciosa ajuda nas dúvidas de inglês) e espírito de equipa, mas deixo de lado a amizade, porque essa não se agradece e sim retribui-se. A convivência saudável e enriquecedora com: Sandra Alves, Susana Seixas, Luísa Azevedo, Sandra Martins, Gil Tomás, Solange Costa e Alexandra Lopes. Perdoem-me as cantorias!

Aos meus amigos Maria do Céu Moreia (fisicamente longe e, todavia, tão perto; merci aussi pour le français!), Carla Afonso, Edite Barbosa e Jorge Magalhães agradeço o incentivo e o desempenho do papel tão essencial de "muro das lamentações".

E o inevitável agradecimento, por último, porque é especial e abrangedor, à minha família: mãe, irmã, cunhado, sobrinhas, avó e tia-avó (que me ensinou as primeiras letras, atirando-me desse modo para um longo e desafiador caminho de aprendizagem).

Dedico esta tese à memória da minha irmã Rosário e do meu pai... na vã tentativa de diminuir um vazio, irremediavelmente abismal, devido à ausência e à impossibilidade de partilha de momentos bons.

A instituições

Agradeço à Fundação para a Ciência e a Tecnologia a atribuição de uma bolsa de doutoramento (PRAXIS XXI/BD/13632/97), essencial à realização deste projecto.

Agradeço à Faculdade de Ciências da Universidade do Porto (particularmente ao Departamento de Zoologia e Antropologia) a aceitação da minha candidatura como sua aluna de doutoramento.

Acknowledgements

The best way that I found to express my feelings and to thank all the collaboration was to abuse a little from what could be called "Population Genetics Language".

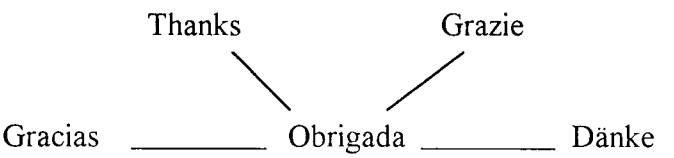
"In a time being of four years of generation, that conducted to the present thesis offspring, we moved from a situation of bottleneck, centred around our once small group, to a situation of significant information flow.

An information flow that, despite some linguistic barriers, conducted to an appreciable enrichment and a considerable reduction on the distance, measured in

friendship and scientific/personal improvement, between my group and the following groups:

- Mark Jobling – Department of Genetics, University of Leicester, England.
- Guido Barbujani – Dipartimento di Biologia, Università degli Studi di Ferrara, Italia.
- Vincent Macaulay – Department of Statistics, University of Oxford, England.
- Angel Carracedo y Maria José Brión– Instituto de Medicina Legal, Universitad de Santiago de Compostela, España.
- Hans-Jürgen Bandelt - Fachbereich Mathematik, Universität Hamburg, Germany.
- António Brehm – Universidade da Madeira.

In an acknowledgement network



RESUMO

Esta tese está dividida em duas secções, correspondentes aos seus dois objectivos principais, que foram, utilizando marcadores do cromossoma Y e do DNA mitocondrial (mtDNA):

(1) comparar, do ponto de vista da Genética Populacional, as inferências que podem ser obtidas a partir da análise de ambos os conjuntos de marcadores, usando as mesmas metodologias estatísticas; e

(2) reconstituir, a partir dos dados obtidos pela análise destes marcadores, aspectos histórico-demográficos das populações de Portugal e Moçambique.

O trabalho desenvolvido na primeira secção justifica-se por um facto que domina actualmente o campo da Genética Populacional – a comparação de extensas bases de dados por diferentes procedimentos analíticos.

A abordagem pela *mismatch distribution* foi aplicada à análise das três bases de dados mais extensas actualmente disponíveis para marcadores bialélicos do cromossoma Y (Y-BMs): duas a uma escala europeia (Rosser et al., 2000, trabalho também incluído nesta tese; Semino et al., 2000); e uma à escala mundial (Underhill et al., 2000). A *mismatch distribution* tem sido a metodologia estatística mais usada para análise dos dados de mtDNA, daí o interesse na comparação dos resultados obtidos, para ambos os *pools* genéticos, com a mesma metodologia. Verificou-se que o padrão de *mismatch* é muito diferente para o mtDNA (unimodal) e para os Y-BMs (bi- ou multimodal), o que, associado ao facto do primeiro evidenciar um desvio significativo (negativo) relativamente ao equilíbrio neutro, enquanto os segundos não, parece apontar para uma panorâmica demográfica, à escala mundial, muito distinta para cada um deles. Os possíveis cenários demográficos associados a esta diferença (tais como selecção, migração ou capacidade reprodutiva diferencial), não sendo mutuamente exclusivos, continuam objecto de debate. Foi também avaliada a possível influência de artefactos analíticos (diferente número de polimorfismos e distintas taxas de mutação em cada um dos conjuntos), tendo-se verificado não constituírem factores significativos de enviesavamento dos resultados, o que indicará que as diferenças observadas nos dois *gene pools* são de natureza genuinamente demográfica.

A metodologia da *mismatch distribution* foi também aplicada à avaliação da distância genética entre pares de haplótipos do cromossoma Y, definidos por STRs. A possibilidade de quantificar as diferenças, no número de unidades repetitivas, entre todos os pares de haplótipos numa amostra, permitiu a determinação da proporção daqueles que, devido a mutação recorrente, se podem tornar idênticos na geração seguinte. Esta avaliação é de extrema importância no campo da Genética Forense, onde a utilização das bases de dados se tem reduzido à simples quantificação de *matches* ou à sua ausência. A aplicação desta metodologia à maior base europeia de dados genéticos forenses (*Y-STR Haplotype Reference Database*) revelou uma elevada estruturação populacional quanto à proporção dos haplótipos que se podem tornar idênticos por estado, em cada geração. Esta estruturação justifica um maior cuidado aquando a procura de *matches* a uma escala europeia.

O objectivo principal deste trabalho foi, contudo, contribuir para a caracterização da história genética de Portugal, ou seja, inferir a partir das suas características no presente, quais os grupos populacionais que no passado a terão modulado. O estudo complementar de marcadores do cromossoma Y e do mtDNA pretendeu comparar as histórias, não necessariamente coincidentes, dos componentes populacionais masculino e feminino, contextualizando-as através da documentação histórica disponível.

A análise dos resultados obtidos para Portugal foi feita em três contextos: a nível de país, Ibérico e Europeu.

Quanto ao primeiro, foi observada uma estruturação populacional significativa no sentido norte-sul para o cromossoma Y. Esta estruturação deve-se, principalmente, ao gradiente crescente da frequência do haplogrupo 21, tipicamente Norte Africano, em consonância com a historicamente bem documentada maior influência Islâmica no sul de Portugal. Pelo contrário, não foi observada estruturação populacional para o mtDNA, ainda que, curiosamente, o haplogrupo U6, tipicamente Berber, se tenha detectado apenas no Norte de Portugal. Outra diferença importante entre os dois *gene pools* foi a detecção de linhagens sub-Sarianas unicamente no componente feminino, sugerindo um enviesamento no tipo de cruzamento entre portugueses e sub-Sarianos, sendo o dominante aquele entre homens portugueses e mulheres escravas. Num horizonte histórico mais distante, as influências Neolíticas foram baixas nos dois *gene pools*, como seria de esperar pela localização de Portugal no extremo oeste do continente europeu, encontrando-se em ambos os haplogrupos tipicamente europeus.

No contexto ibérico, os padrões observados em Portugal, quanto à influência Norte Africana, mostraram-se coincidentes com os de outras regiões da península. Assim, o gradiente crescente norte-sul para o haplogrupo 21 do cromosoma Y foi observado em toda a faixa oeste da Ibéria (sendo a Galiza, tal como o Norte de Portugal, estatisticamente diferentes do Sul de Portugal). Quanto ao mtDNA, o haplogrupo U6 foi também apenas detectado no norte da Ibéria.

No contexto europeu, a análise de 11 Y-BMs evidenciou uma elevada estruturação populacional das 47 populações europeias estudadas, que se organiza na forma de gradiente para 5 marcadores: 2 com orientação oeste-este (em direcções opostas para os haplogrupos 1 e 9, representando os marcadores Paleolítico e Neolítico, respectivamente); 1 norte-sul (para o haplogrupo 21, restrito à área Mediterrânea); e 2 regionais (para os haplogrupos 3 e 16 no nordeste da Europa). Esta estruturação populacional, à escala continental, revelou-se estar significativamente correlacionada com a geografia mas não com a linguística.

O impacto genético dos portugueses em populações não europeias, decorrente do período dos Descobrimentos, foi investigado através do estudo da ex-colónia portuguesa de Moçambique (sudeste de África). Este impacto foi nulo a nível do mtDNA e muito baixo no cromossoma Y, revelando, também em Moçambique, um predomínio no cruzamento homem europeu/mulher sub-Sariana. A pesquisa de *match* para haplótipos do mtDNA entre Moçambique, Europa e América evidenciou outro facto já historicamente documentado: uma maior exportação de escravos do este de África em direcção às América do que para a Europa. Foi impossível fazer a contraprova deste enviesamento para a componente masculina, uma vez que não foram detectados haplótipos sub-Sarianos em Portugal e na maioria dos países europeus. Dado localizar-se na rota este da expansão Bantu em direcção ao sul de África, o estudo de Moçambique, quando enquadrado num contexto sub-Sariano, permitiu revelar que a redução da diversidade, em direcção ao sul, devida a esse movimento, foi mais acentuada no *pool* do cromossoma Y (e neste, mais acentuada na costa oeste) do que no mtDNA. Pode, portanto, inferir-se ter existido um cruzamento mais frequente entre os recém-chegados homens Bantu e as mulheres nativas, como parece indicar a presença de 7% de sequências L1d (típicas de povos Khoisan) no *pool* actual de mtDNA Moçambicano.

This thesis is divided in two sections corresponding to its two broad aims, which were, using Y-chromosome and mitochondrial DNA markers (mtDNA):

(1) to compare the inferences that can be drawn from the analysis of each set of markers, applying the same statistic methodologies; and

(2) to reconstruct, using these markers, historic-demographic features of Portugal and Mozambique populations.

The work developed in the first section is justified by a fact that dominates current Population Genetics work – large data set comparisons by different analytical procedures.

The mismatch distribution approach was applied to the analysis of the three largest datasets available for Y-biallelic markers (Y-BMs): two at a European scale (Rosser et al., 2000, also included in this thesis; Semino et al., 2000); and one worldwide (Underhill et al., 2000). Mismatch distribution has been the main statistical methodology applied to the analysis of mtDNA data and, hence, the interest in the comparison between the results obtained, under the same methodology, for both gene pools. It was found that the mismatch pattern is very different for mtDNA (unimodal) and for Y-BMs (bi- or multimodal), in addition to the fact that the first shows (negative) significant departure from the neutral equilibrium while the seconds do not, seems to point to a very distinct demographic picture, at a worldwide scale, for each of them. The possible demographic scenarios associated to this difference (such as, differential selection, migration or reproductive capacity), are not mutually exclusive, and remain under debate. The possible influence of analytical artefacts (different number of polymorphic sites and heterogeneity of mutation rates) was also checked, and it was verified that they are not constitute significant biasing factors, which suggests that the differences observed for both gene pools have a true demographic cause.

The mismatch distribution methodology was also applied to the evaluation of the genetic distance between pairs of haplotypes of the Y-chromosome, defined by STRs. The possibility of quantifying the differences in repeat units between all the pairs of haplotypes in a sample, allowed the determination of the proportion of those that, due to recurrent mutation, are prone to become identical in the next generation. This evaluation

is of extreme importance in the forensic field, where the use of databases has been purely restricted to the quantification of the number of matches or their absence. Applying this methodology to the largest European forensic genetic data set (Y-STR Haplotype Reference Database), a high degree of population structuring for the proportion of haplotypes that can become identical by state in each generation has been shown. This significant structuration justifies a higher caution when searching matches at a European scale.

The main purpose of this work was, however, to contribute to characterise the genetic history of Portugal, that is, to infer from the present characteristics, which population groups have modulated it in the past. The simultaneous study of the Y-chromosome and mtDNA markers intended to compare the male and female histories, which are not necessarily identical, framing them into the available historical documentation.

The analysis of the results obtained for Portugal was conducted in three contexts: the country, the Iberian, and the European.

With respect to the first, a north-to-south significant structuring for the Y-chromosome was observed. This structuring is mainly due to the presence of an increasing gradient for the frequency of haplogroup 21, typically North African, in accordance with the historically well-documented Muslim influence, predominant in South Portugal. In opposition, no population structuring was observed for the mtDNA, although the typical Berber haplogroup U6 was, surprisingly, detected only in North Portugal. Another important difference between gene pools was the detection of sub-Saharan lineages only in the female component, suggesting a bias in the mating pattern between Portuguese and sub-Saharans, favouring the one between Portuguese males and slave females. Neolithic influences were low in both gene pools, which show typical European haplogroups, as expected from the westernmost location of Portugal in the continent.

In the Iberian context, the patterns observed in Portugal with respect to the North African influence showed to be coincident with the ones for other regions of the Peninsula. Thus, the increasing north-south gradient for haplogroup 21 in the Y-chromosome was observed in all the western fringe of Iberia (being Galicia, as North Portugal, statistically differentiated from South Portugal). With respect to mtDNA, haplogroup U6 was also only observed in the north of Iberia.

In the European context, the analysis of 11 Y-BMs revealed a high degree of population structurintg in the 47 European populations studied, clinally organised as clines for 5 markers: 2 were west-east (in opposite directions for haplogroups 1 and 9, representing the Palaeolithic and the Neolithic markers, respectively); one north-south (for haplogroup 21, restricted to the Mediterranean area); and 2 regional (for haplogroups 3 and 16 in northeastern Europe). This population structuring, at a continental scale, was significantly correlated with geography but not with language.

The genetic impact of the Portuguese in non-European populations, since the "Discoveries" period, was investigated by studying the former Portuguese colony of Mozambique (Southeast Africa). This impact was nil at the mtDNA level and very low for the Y-chromosome, revealing, also in Mozambique, a predominance of the mating European male/sub-Saharan female. Search for mtDNA haplotype matches between Mozambique, Europe and America showed, as historically documented, that eastern African slaves were mainly directed towards America rather than to Europe. It was impossible to evaluate this bias in the male counterpart because no sub-Saharan haplotypes were detected in Portugal as in the majority of European countries. The study of Mozambique, located in the eastern route of Bantu expansion towards the south of Africa, allowed to reveal that the reduction of diversity, towards south, due to that movement, was stronger in the Y-chromosome pool (and stronger in the western coast) than in mtDNA counterpart. It is possible, therefore, to infer that the most frequent mating was between the arriving Bantu males and the native females, as shown by the presence of 7% of L1d sequences (typical of Khoisan populations) in nowadays mtDNA Mozambican pool.

RÉSUMÉ

Cette thèse est divisée en deux sections, dont les deux objectifs principaux, sont basés sur l'utilisation des marqueurs du chromosome Y et du DNA mitochondrial. Ils ont consisté à:

(1) comparer, du point de vue de la Génétique des Populations, les conclusions qui peuvent être obtenues à partir de l'analyse des deux ensembles de marqueurs, en utilisant les mêmes méthodologies statistiques; et

(2) reconstruire, depuis les données obtenues à travers l'analyse de ces marqueurs, les aspects à la fois historiques et démographiques des populations du Portugal et du Mozambique.

Le travail développé dans la première section est un sujet qui domine actuellement le domaine de la Génétique des Populations – la comparaison des vastes bases de données par différentes procédures analytiques.

L'étude par *mismatch distribution* a été appliquée à l'analyse des trois bases de données les plus vastes actuellement disponibles pour les marqueurs biallélíques du chromosome Y (Y-BMs): deux à l'échelle européenne (Rosser et al., 2000, travail aussi inclus dans cette thèse; Semino et al., 2000); et une l'échelle mondiale (Underhill et al., 2000). Le *mismatch distribution* a été la méthodologie statistique la plus utilisée pour l'analyse des données de mtDNA, justifiant l'intérêt de faire la comparaison des résultats obtenus, pour les deux *pools* génétiques, avec la même méthodologie. A été vérifié que le modèle de *mismatch* est très différent pour le mtDNA (unimodal) et pour les Y-BMs (bi- ou multimodal), tout ça, associé au fait que le premier rend évident une déviation significative (négative) relative à l'équilibre neutre, mais pas les seconds. Cela semble pointer vers une panoramique démographique, à l'échelle mondiale, très différente pour chacun des modèles. Les cénarios démographiques possibles associés à cette différence (comme la sélection, migration ou la capacité reproductive différentielle), n'étant pas mutuellement exclusifs, ils restent l'objet du débat. A été aussi évaluée la possible influence des artefacts analytiques (différent nombre de polymorphismes et différents taux de mutation dans chaque ensemble), ils ne constituent pas des facteurs significatifs de biais des résultats, ce que peut indiquer que les différences observées dans les deux *gene pools* sont de nature purement démographiques.

La méthodologie de la *mismatch distribution* a été aussi appliquée à l'évaluation de la distance génétique entre paires d'haplotypes du chromosome Y, définis par STRs. La possibilité de quantifier les différences, dans le nombre d'unités répétées, parmi toutes les paires d'haplotypes d'un échantillon, a permis la détermination de la proportion dans la génération suivante. Cet évaluation est d'extrême importance dans le domaine de la Génétique Criminelle, où l'utilisation des bases de données a toujours été faite dans le sens d'une simple quantification ou absence de *matches*. L'application de cette méthodologie à la plus grande base européenne de données génétiques criminelles (*Y-STR Haplotype Reference Database*) a montré une stratification élévée des populations concernant la proportion des haplotypes pouvant devenir identiques par état, à chaque génération. Cette stratification nécessite donc un très grand soin quand on recherche les *matches* à une échelle européenne.

L'objectif principal de ce projet a été, aussi, de contribuer à la caractérisation de l'histoire génétique du Portugal, et conclure des groupes de populations qui l'ont modulée dans le passé. L'étude complémentaire des marqueurs du chromosome Y et du mtDNA a consisté à comparer les histoires, pas nécessairement coïncidentes, des populations masculine et féminine, en les contextualisant par la documentation historique disponible.

L'analyse des résultats obtenus pour le Portugal a été faite à trois niveaux: au niveau du pays, au niveau ibérique et au niveau européen.

Au niveau du pays, a éte observée une stratification significative des populations dans le sens nord-sud pour le chromosome Y. Cette stratification est due, principalement, au gradient croissant de la fréquence de l'haplogroupe 21, typiquement Nord Africain, en concordance avec la notion bien documentée d'une plus grande influence Islamique dans le sud du Portugal. A l'opposé, aucune stratification des populations n'a été observée pour le mtDNA dans le nord du Portugal, bien que, curieusement, l'haplogroupe U6, typiquement berbère, n'ait été détecté que dans cette région. Une autre différence importante entre les deux *gene pools* a été la détection de lignages sub-saharien uniquement dans la population féminine, suggérant un biais dans les mariages entre portugais et africains d'origine sub-saharienne, dont le plus important est le mariage entre les hommes portugais et les femmes esclaves. Si on prend un horizon historique plus ancien, on observe que les influences néolithiques ont été peu importantes dans les deux *gene pools*, comme le laisse présupposer la localization du

Portugal dans l'extrémité occidentale du continent européen, en se trouvant dans les deux *gene pools* les haplogroupes typiquement européens.

Dans un contexte ibérique, l'influence nord-africainne s'est traduite au Portugal par des modèles coïncidents avec ceux qui sont présent dans d'autres régions de la péninsule. Ainsi, le gradient croissant nord-sud pour l'haplogroupe 21 du chromosome Y a été observé sur toute la façade ouest de la Péninsule Ibérique (c'est à dire: la Galize, et le Nord du Portugal, statistiquement différents du sud du Portugal). Pour le mtDNA, l'haplogroupe U6 n'a été detecté que dans le nord de la Péninsule Ibérique.

Dans un contexte européen, l'analyse de 11 Y-BMs, a montré une stratification élevée des 47 populations européennes étudiées, stratification qui s'organise sous une forme de gradient pour 5 marqueurs: 2 marqueurs avec une orientation ouest-est (pour les haplogroupes 1 et 9, représentant les marqueurs Paléolithique et Néolithique, respectivement); 1 marqueur nord-sud (pour l'haplogroupe 21, limité à la région méditerranée); et 2 marqueurs régionaux (concernant les haplogroupes 3 et 16 localisés dans le nord-est de l'Europe). Cette stratification des populations, à l'échelle continental, est significativement correlée avec la géographie et non pas avec la linguistique.

L'impact génétique des portugais sur les populations non-européennes, provenant de la période des Découvertes, a été évalué à travers l'étude de l'ancienne colonie portugaise du Mozambique (sud-est africain). Cet impact est nul au niveau du mtDNA et très bas au niveau du chromosome Y, montrant qu'au Mozambique il y a eu un biais dû aux mariages entre hommes européens et femmes d'origine sub-saharienne. La recherche de *match* pour les haplotypes du mtDNA entre le Mozambique, l'Europe et l'Amérique a montré un autre évènement déjà historiquement documenté: une plus grande circulation d'esclaves de l'Afrique de l'est vers les Amériques que vers l'Europe. A été impossible de faire la contre-épreuve de ce biais pour la population masculine, vu qu'aucun haplotype sub-saharien n'a éte détecté au Portugal et dans la plupart des pays européens. L'expansion des Bantu vers le sud de l'Afrique a traversé le Mozambique (situé sur la route est de cet expansion). Par l'étude de ce pays dans un contexte sub-saharien, on a montré que la réduction de la variété dans la direction sud (et plus encore dans la côte ouest), est plus accentuée dans le *pool* du chromosome Y que pour le mtDNA. Ainsi, est possible de conclure de l'existence d'un mariage plus fréquent entre les nouveaux venus hommes Bantu et les femmes indigènes, comme semble le montrer la présence de 7% de séquences L1d (typiques des peuples Khoisan) dans le *pool* actuel du mtDNA des habitants du Mozambique.

Introduction

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS

1- The Y-chromosome

The Y-chromosome is the second smallest human chromosome, having an estimated average size of 60 million base pairs (Mb). It presents, cytologically, two regions (Foote et al., 1992): (1) a heterochromatic region on the distal long arm (Yq), which can vary in size from virtually undetectable to a length more than half the entire chromosome; and (2) an euchromatic region, with a minimum size estimate of 28Mb (encompassing the rest of the chromosome), where: (2.1) blocks of sequences that are homologous to the X-chromosome, (2.2) repetitive sequences comprising and surrounding the centromere, (2.3) families of Y-specific repetitive sequences, and (2.4) Y-specific single-copy sequences are located.

The Y-chromosome is paternally inherited, that is, it is only transmitted from fathers to sons. Nearly all the Y-chromosome consists on the non-recombining portion (NRPY), which, as the name specifies, does not recombine with the X-chromosome, being haploid. The only exceptions to this non-recombining pattern are the pseudoautosomal regions (PARs), located at both telomeres of the Y-chromosome, which undergo recombination with the X-chromosome: PAR1 with 2.7 Mb at Yp; and PAR2 with 0.32 Mb at Yq.

So, the most important genetic property of the NRPY is the paternal transmission in block through generations, in an unchanged way, except when a mutation occurs. In other words, this region consists in a macro haplotype, where the absence of recombination maintains the record of the mutational events that occurred through a certain male lineage. Further references to the Y-chromosome markers in this thesis will refer to markers located in this region.

During many years the idea that there were very few Y-chromosome genes (including the male-determining gene, SRY, sex-determining region Y) reigned. However, in the 90´s, a large number of additional genes, including some that participate in fundamental cellular processes (Fisher et al., 1990; Lahn and Page, 1997) was revealed.

The knowledge on Y-chromosome polymorphisms has also deeply changed. Initial reports, using probes for conventional RFLPs, showed a low level of DNA polymorphisms on the Y-chromosome (Malaspina et al., 1990; Spurdle and Jenkins, 1992). More recently, the employment of new technologies, such as the denaturing high-performance liquid chromatography (DHPLC) (Underhill et al., 1997), pulsed-field gel electrophoresis (PFGE) (Jobling, 1994) and large-scale sequencing projects, are revealing an increasing number of Y-chromosome polymorphisms of several kinds (Ayub et al., 2000; Underhill et al., 2000).

The Y-chromosome markers can be classified in two categories: biallelic and multiallelic markers. Each one of these categories includes two broad types of polymorphisms: (1) the biallelic markers (BMs) comprise single nucleotide polymorphisms (SNPs), which are base substitutions, and insertion/deletions (indels), like the *Alu* element; (2) the multiallelic markers (MMs), which are tandem repetitive polymorphisms, can be classified as microsatellites (also called STRs, Short Tandem Repeats) if the tandem motif varies from 2 to 6bp, or as minisatellites if the tandem motif is larger. While BMs are characterised by a low mutation rate, around $5x10^{-7}$ per site per generation (Hammer, 1995; Jobling et al., 1997), MMs display higher mutation rates, estimated around $3.17x10^{-3}$ per generation for microsatellites (Kayser et al., 2000), and $6\text{-}11x10^{-3}$ per generation for the minisatellite MSY1 (Jobling et al., 1998). So, BMs are unique or near-unique events, while recurrent mutation is a common phenomenon in the case of MMs. From this point on, we will use the notations *haplotype*, for a pattern defined by Y-STRs, and *haplogroup*, for one characterised by Y-BMs, as been used by other authors (e.g. De Knijff, 2000a).

Following the initial studies of characterisation of some populations based on particularly informative markers (Zerjal et al., 1997; Hurles et al., 1998, 1999), three large data sets were published for Y-BM, in the last year: at the European scale, the studies of Rosser et al. (2000), employing 11 polymorphic markers (which is incorporated in this thesis), and Semino et al. (2000), surveying 22 markers; and the worldwide survey of Underhill et al. (2000), where 166 biallelic (and 1 triallelic) markers were analysed in 1062 individuals.

With respect to the Y-chromosome microsatellites (Y-STRs), as the STRs are the markers of choice in forensic genetics, many of the published results have forensic purposes (Jobling et al., 1997; Kayser et al., 1997). The International Forensic Y-User group constructed a Y-STR Haplotype Reference Database, available in

http://ystr.charite.de, which is mainly directed for searching haplotype matches in European populations. This Database is presently being enlarged to American and Asian populations. Although the search for haplotype matching is the main tool used in forensics, the evaluation of how significant is a match remains a controversial issue. A frequently neglected phenomenon, that can introduce a bias in this kind of analysis is, precisely, the possibility of recurrence, due to the high mutation rate of STRs. Recurrence gives rise to haplotypes identical-by-state (IBS) rather than identical-by-descent (IBD). Till now, the only approach that has been used to evaluate the proportion of these classes among identical haplotypes was to assess haplotype identity by the combined information from SNPs, considering that the same haplotype in two different haplogroups points to IBS and not IBD. De Knijff (2000a) estimated a proportion of 0.8% of IBS in a sample of 275 Dutch screened for 6 Y-STRs and 4-SNPs, which decreased to 0.2% when analysing 2 more Y-STRs. As the author underlines, these values must be kept in mind for the evaluation of a haplotype match in forensics.

Bosch et al. (1999) also compared the haplotype diversity between haplogroups for 4 North African and 2 Iberian populations, but in a different perspective. Their main conclusion was that Y-STRs are mainly structured by haplogroup rather than by geography, putting forward an alternative way to understand a population "... as an association of lineages from a deep and population-independent gene genealogy, rather than as a complete evolutionary unit".

However, the main use of the combined information for both slow- and fast-mutating Y-markers has been the age estimation of a specific Y-BM, by analysing haplotype diversity within that specific haplogroup (e.g. Hurles et al., 1999).

Unfortunately, combined information for both Y-STRs and Y-BMs in the same samples is still very rare, and, at least in the forensic field, it will remain so as long as more advanced techniques allowing an easy and fast Y-SNP typing are not available.

2- The mtDNA

Each one of the hundreds of mitochondria in a cell presents thousands of copies of a closed circular molecule of DNA, the mtDNA, containing 16,569bp. The complete sequence of the mtDNA was described in 1981 by Anderson et al. (known as the Cambridge reference sequence, or CRS) and recently revised by Andrews et al. (1999).

The distribution of nucleotides is different on both strands of the mtDNA molecule, purines being predominant in one strand - heavy strand (H) - and pyrimidines in the other - light strand (L). The terms heavy and light result from the fact that both strands can be separated in a density gradient.

The coding portion of the mtDNA encompasses the majority of the molecule, and codes for 13 polypeptides essential to the enzymes of the energy-generating pathway oxidative phosphorylation (OXPHOS), and for the small (12S) and the large (16S) rRNAs and 22 tRNAs necessary for protein synthesis (Anderson et al., 1981). This region displays a mutation rate approximately ten times higher than the nuclear DNA mutation rate (Brown et al., 1979).

The remaining (approximately) 1.2Kb non-coding region, also known as control region or D-loop, displays a mutation rate ten times higher than the rest of the molecule (Vigilant et al., 1991). Although non-coding, this region is functionally very important because it contains the main regulatory elements for transcription and replication and presents binding sites for numerous molecules such as DNA and RNA polymerases and other transcription and regulatory factors, which makes at least probable that this region is under selective pressure (see e.g. Meyer et al., 1999). In 1989, Vigilant and collaborators defined two regions in the D-loop, which present an even higher level of variation than the rest of the control region, therefore called hypervariable regions, HVRI and HVRII, usually considered to include nucleotide positions 16,024-16,383 and 73-340, respectively.

The first analyses of the mtDNA were based on scoring RFLPs throughout the entire molecule (Cann et al., 1987; Chen et al., 1995). Very soon they were replaced, at least in population and forensic genetics, by HVRs' sequencing (Vigilant et al., 1991; Soodyall, 1993; Krings et al., 1999). Currently, a combination of both methods (Graven et al., 1995; Watson et al., 1997; Chen et al., 2000) is gaining wider use, since the typing of more stable positions outside the control region makes safer the classification into haplogroups, in a similar fashion to Y-chromosome haplogrouping with SNPs and STRs. Just by the end of last year, there was the first study, in 53 individuals from all over the world, where all the mtDNA molecule was sequenced (Ingman et al., 2000), representing, still nowadays, a great technical effort.

MtDNA transmission is believed to be exclusively maternal. Indeed for many years it was thought that during fecundation, only the head of the sperm entered the oocyte, and that the intermediate piece, containing the male mitochondria, remained

outside. Recently, it was discovered that the intermediate piece does enter the oocyte, but due to an unknown process, the father's mitochondria are "not observed" in the child (Giles et al., 1980).

Furthermore, the transmission of the mtDNA is strictly non-Mendelian in the sense that the number of mitochondria that is transmitted to the new individual is not fixed. An additional factor of unpredictability is the occurrence of heteroplasmy, that is, the presence of more than one type of mtDNA molecule in the female gamete, allowing a wide spectrum of transmission outcomes. Thus, mithocondrial transmission is a population phenomena rather than a gene transmission at individual level. So, in each cell division, meiotic or mitotic, demographic events such as bottlenecks and "molecular drifts" are acting upon the mtDNA pool.

Besides heteroplasmy, other properties of the mtDNA have been widely discussed in the last years, namely the mutation rate and the occurrence or not of recombination.

With respect to the question of how fast is the HVR mutation rate, the two estimation methods, by pedigree or by phylogeny analyses, produced extremely different results, considerably higher for the first, but varying extensively even inside the same type of analysis. In fact, the first pedigree studies pointed to a control region mutation rate of 0.75/site/Myr (Howell et al., 1996) and 1.5-2.5/site/Myr (Parsons et al., 1997), but later, values of <0.46/site/Myr (Jazin et al., 1998) and 0.32/site/Myr (Sigurðardóttir et al., 2000) were obtained. Ward et al. (1991), applying a phylogenetic strategy, obtained a divergence rate (twice the substitution rate) of ~0.30/Myr or 1 transition/10,000 years across HVRI. These discrepancies can be at least partially explained by the non-homogeneous mutation rate across the various nucleotide positions in the HVRs, especially for HVRII (Aris-Brosou and Excoffier, 1996; Meyer et al., 1999), which is not taken into consideration in the pedigree approach. Many authors are trying to evaluate this heterogeneity in the mutation rate, estimating which nucleotides are slow-, intermediate- and fast-evolving (e.g. Meyer et al., 1999). A deep discussion has been raised about (a) as to weight differentially each nucleotide position and (b) the consequences of heterogeneous mutation rates on the demographic parameter estimations according to some statistics (Rogers et al., 1996 versus Bertorelle and Slatkin, 1995; Aris-Brosou and Excoffier, 1996). There is also heterogeneity in the type of substitution, with a clear bias towards transitions in animal mtDNA (Wakeley, 1994). Estimates obtained for the transition-transversion ratio vary from between 12 and

37 for HVRI to between 7.1 and 12.2 for HVRII (Meyer et al., 1999), figures depending on the methodology and on the incorporation of mutation rate heterogeneity for its calculation.

Recombination is also a hot issue of debate. In fact, the use of the same kind of analyses (either by linkage disequilibrium or phylogenetic) to the same data sets has lead to opposite results and conclusions (Awadalla et al., 1999 versus Kivisild and Villems, 2000; Elson et al., 2001). It is known since 1996 (Thyagarajan et al., 1996) that mitochondria possess the machinery for recombination. However, even data using physical approaches (rather than genetic or molecular analysis of marker exchange) indicate the extreme rarity (or possibly absence) of intergenomic or reciprocal mtDNA recombination in human cells (Howell, 1997).

3- Y-chromosome and mtDNA insights into Population Genetics

The shared characteristic of non-recombination between the Y-chromosome and the mtDNA, with absence of reshuffling of the mutation events, enables the easy establishment of each system phylogeny. In other words, going backwards, it is possible to relate all the present male and female lineages and infer their ancestors.

Furthermore, their effective population size is one quarter of the autosomal counterparts, rendering them more sensitive to founder effects and genetic drift, very convenient properties for detection of geographical substructuring in human populations.

Their simultaneous study allows inferences about the male and female contribution to genetic makeup of a population. In accordance, sex-specific contributions, migration, selection and gene flow are now open to analysis.

So far, the comparison between the data obtained for Y-BM and mtDNA has shown that male- and female-transmitted gene pools have evolved under different conditions. Y-BMs are characterised by lower diversity within populations, and more significant spatial structure when compared with mtDNA (see e.g. Jorde et al., 2000). At least in Europe, some Y-BMs are clinally distributed (Rosser et al., 2000; Semino et al., 2000), in a similar pattern to the one observed for some autosomal markers (Chikhi et al., 1998a).

It is unclear why there is this difference between both gene pools, but the most common explanations are: (1) selection (Excoffier 1990; Wise et al. 1998; Wyckoff et al. 2000), being negative for some Y-chromosome lineages and/or positive for rare lineages in mtDNA; and (2) different demographic histories for females and males, such as higher migration rate for females, due to patrilocality, or the practice of male polygamy (Seielstad et al., 1998). Another possible cause of differentiation could be the occurrence of expansion for mtDNA lineages, whose signals were reported to occur in all populations, except present hunter-gatherers (Excoffier and Schneider, 1999), and constancy for the Y-BM lineages, although no direct test was applied till now to the published Y-BM datasets. The presence/absence of signals of lineage expansion does not necessarily imply presence/absence of male/female effective population size expansions, since a high heterogeneity in lineages transmission inside one gender (few lineages very successful in opposition to others unsuccessful) would hide signals of effective population size expansion (Crow, 1958). In fact, there seems to exist a higher heterogeneity in passing their lineages among males than among females, due not only to negative selection, as already shown for some lineages associated with higher infertility (Jobling and Tyler-Smith, 2000), but mainly to a neutral phenomenon. Males are not biologically constrained in bearing children and are fertile since puberty, while female fertility is cyclic at roughly a monthly basis and biologically constrained during long periods (pregnancy, breast feeding and menopause). Moreover this discrepancy between both sex reproductive periods is expected to be stronger in recent times than when the lifespan was considerably shorter.

The contrasting picture displayed by mtDNA and Y-BMs (both sequence polymorphisms) is not shared by mtDNA versus Y-STRs. Although the type of polymorphism is very different in Y-STRs, mutation rate is of the same order of magnitude as sequence variation at mtDNA (Ward et al., 1991; Weber and Wong, 1993; Chakraborty et al., 1997). This fact was acknowledged by Belledi et al. (2000), showing that values for female and male migration rates were similar for HVRI and Y-STRs, instead of a 4.13 ratio obtained when comparing HVRI and Y-SNPs (in agreement with the estimate of Seielstad et al., 1998). In opposition to what was found for Y-SNPs, no significant correlation was observed between geography and HVRI or Y-STRs.

When testing the demographic models, expansion versus constancy, Pritchard et al. (1999) have demonstrated that for Y-STRs (as for mtDNA), in most populations (except for West Africa and Oceania), the constancy model could be rejected.

From all these results, Y-BMs and Y-STRs seem to show very different evolving scenarios. The only exception is the work of Shen et al. (2000), who studied four Y-chromosome genes in 53-72 males from the five continents and have found a sign of expansion inferred from: (1) a better fit to the expectations of a Luria-Delbrück distribution (rather than of a constant population size model); (2) the unimodal shape of mismatch distributions; and (3) statistically significant Tajima's D negative values.

4- Comparing results under the same statistic method – mismatch distribution analysis

Although in recent years there has been an enormous growth of Y-chromosome data, mtDNA was for many years the only non-recombining system for which there was sufficient data at a worldwide scale. From the beginning, there was an effort to develop methods suitable to interpret demographic signs displayed by this kind of data. Mismatch distribution analysis was then developed (Rogers and Harpending, 1992), and continues to be the most used analytic method.

The mismatch distribution is the distribution of pairwise sequence differences, in the number of nucleotide or restriction sites, between each pair of sequences that can be drawn randomly from a sample. Since changes in population size tend to leave recognizable signatures in the patterns of nucleotide diversity, the mismatch distribution contains information on the population's history (Rogers and Harpending, 1992). The genealogy of a population of constant size is expected to have long deep branches (Donnelly, 1996); mutations occurring along these branches will be shared by several lineages, resulting into an irregular or ragged distribution of pairwise sequence differences. Conversely, the genealogy of a population that has substantially grown in size exhibits long terminal branches, and the mutations that have occurred along those branches, i.e., most mutations, will be specific to a single lineage (Donnelly, 1996). Under these latter conditions, unimodal mismatch distributions are expected (Harpending et al., 1993; Harpending, 1994; Marjoram and Donnelly, 1994), whose means, under an infinite-site mutation model, increase as a function of the time elapsed since the population expansion (Sherry et al., 1994).

Mitochondrial mismatch distributions, based on both RFLPs (Rogers and Harpending, 1992; Harpending, 1994) and HVR sequences (Excoffier and Schneider, 1999), are, with very few exceptions, unimodal, and so have been interpreted as representing population expansions (Rogers and Harpending, 1992; Rogers and Jorde, 1995). In general, both the mean and the variance are highest (and therefore the curve is smoothest) in African populations, less so among Asian and lowest among European ones. A tentative dating of that expansion has been made, by relating the mean number of differences and the mutation rate. The estimates for these major demographic expansions were around 110 KY ago in East Africa, 70 KY ago in the rest of Africa and Asia, 55 KY ago in America and 40 KY ago in Europe and in the Middle East (Excoffier and Schneider, 1999). Populations of hunter-gatherers are the only exceptions, probably because they underwent recent bottlenecks, erasing the typical genetic characteristics of an expanding population (Excoffier and Schneider, 1999).

Mismatch distribution analyses, for the reasons explained above, were not applied to the large data sets accumulated for Y-BMs. In fact, except for the four genes studied by Shen et al. (2000), it is not known yet if the same unimodal pattern is obtained for the male component. The basic assumption of the mismatch distribution - the infinite-sites mutation model - is not violated by Y-BMs, which are equivalent to the first mtDNA RFLPs databases, to which this analysis was applied for the first time (Rogers and Harpending, 1992).

With respect to Y-STRs, they do not seem very prone to mismatch distribution analysis since they mutate by a different mode, although not well established. Some have been advanced, namely according to expectations of the infinite-allele (IAM), stepwise (SMM) or the two-phase mutation models, with the additional noise of constrains in the allelic size and the differential mutation rate between alleles of different sizes and sequences (Zhivotovsky et al., 1997). In any case, mismatch distributions can be used at least to estimate molecular distances (in loci or repeat units) between all pairs of haplotypes in a sample.

II- POPULATION GENETICS' APPLICATIONS

1- Population Genetics - some insights into its potential contribution to the Anthropological debate

Reconstructing the past has always been a major issue for humanity, especially when it relates directly to the origin of modern humans. This has been the object of study of Anthropology, a branch science of Archaeology (Renfrew and Bahn, 2000), devoted to the analysis and interpretation of the remaining evidences (bones, artistic objects, wall paintings, constructions) of those remote times.

Population Genetics contributes with another kind of evidence: the features of contemporary populations are used to reconstruct the past. This approach is based on the basic fact that present genomes have derived from past ones by mutation. For some portions of the genome, namely mtDNA and NRPY, there is no recombination mixing up of the historical events, and it is possible to reconstruct and to date phylogenies, provided the mutation rate is known. Although some issues remain controversial, it has been applied extensively for the last years, as is reflected by the coining of the specific word of "Archaeogenetics" for this concept (Amorim, 1999; Renfrew, 2000).

Obviously, this way of recovering past history is restricted to the reconstruction of the genetic characteristics of those that were reproductively successful. Since many lineages have not contributed to the present gene pool, they cannot be inferred in this way. However, recent technical developments have allowed the study of fossil DNA from human remains, such as in bones, and then, enabling the characterisation of ancient genomes, no matter they have contributed or not to present gene pools.

This is the case for the controversial issue of the hybridisation between modern humans and Neanderthals, who co-inhabited Europe and some Asian regions, for several millennia. If this gene flow had occurred, Neanderthal contribution to modern humans would have a major impact in Europeans. Classical Anthropology has not been able to answer to this question, in part because no clear-cut transitional fossil forms have been found. Even the putative exception of the recent Lapedo finding (Duarte et al., 1999) remains controversial (Tattersall and Schwartz, 1999). The genetic evidence collected from mtDNA from three geographically very distant Neanderthal specimens (Krings et al., 1997; Ovchinnikov et al., 2000; Krings et al., 2000), suggests that the

Neanderthal signature in present-day modern human mtDNA pool is insignificant. In fact, the differences between Neanderthal and Europeans were as high as between Neanderthals-Africans and Neanderthals-Asians and the age of the most recent common ancestor of both lineages was estimated around 317,000-741,000 years ago.

If there is no discussion about the Eastern African origin of the genus *Homo* and that the first specie spreading to Europe and Asia was *Homo erectus*, the debate is still hot about the place of origin of modern humans and the way they evolved from *Homo erectus*, condensed in two opposite models commonly named "multiregional" and "Out of Africa".

According to the multiregional hypothesis, modern humans arose in several parts of the Old World from local populations of *Homo erectus* and the genetic continuity between those populations was maintained by gene flow (Wolpoff et al., 1984; Harris and Hey, 1999). This would imply a substantial number of individuals constituting the *Homo erectus* populations and considerable geographic proximity between populations to enable enough interbreeding. A necessary consequence of this model is that contemporary human genetic diversity would be much greater and geographically more structured than it is observed.

The Out of Africa hypothesis, on the contrary, assumes a unique event that occurred in Africa, around 200-100 thousand years ago, from a reduced number of ancestors (Cavalli-Sforza et al., 1993), in the order of 10,000 breeding adults (Harpending et al., 1998). Several migration waves would then colonise the rest of the world. The modern human settlement was estimated as 90,000 years ago for Near East, 40,000 for Europe and Asia, 50,000-60,000 for Australia, and finally, around 14,000 years ago (although older times have been claimed), America was first touched, via the Bering Strait (Renfrew and Bahn, 2000).

Although the "Out of Africa" hypothesis has been clearly favoured, particularly in the last years and among geneticists, the alternative model still collects a reasonable number of supporters.

The main lines of the genetic evidence can be summarised as follows. First, African populations present on average the highest levels of diversity (Vigilant et al., 1991; Bowcock et al., 1994; Jorde et al., 1997; Tishkoff et al., 1996), which can be interpreted as being the oldest ones with more time to accumulate diversity. However, it must be said that the alternative explanation of a bigger long-term effective population

size for Africans is possible (arguments favouring older age rather than bigger size have been reviewed by Seielstad et al., 1999). Second, phylogenetic analyses on several markers conduct most frequently to trees whose first branch separates African populations (Cann et al., 1987; Ingman et al., 2000; Mountain and Cavalli-Sforza, 1994; Bowcock et al., 1994; Underhill et al., 2000). Third, coalescence analyses have shown consistently recent times (incompatible with "multiregional" expectations) for the most recent common ancestor: within the last 250,000 years for mtDNA (Cann et al., 1987; Vigilant et al., 1991), 188,000 years for 3 Y-polymorphisms (Hammer, 1995), and ~800,000 years for the autosomal β-globin locus (note that for autosomal markers, coalescence times are expected to be 4 times greater than uniparental ones; Harding et al., 1997).

The archaeological evidence is much more difficult to assess and interpret, although most authors claim that no transitional skeletal remains were found in other parts of the Old World besides East Africa. Furthermore, the oldest skeletal and associated typical industries are found in Africa, and successively in Near East, Europe, Asia, Australia and America, with dates matching the previously referred.

Finally, we will refer the question of agriculture dispersal, which is particularly relevant in Europe.

The first undisputed remains of agriculture were found in the Near East around 10,000 years ago. Whether or not independent "inventions" or contributions to agriculture and domestication were made elsewhere, the main discussion has been centred on the way of dispersion from that focus. Traditionally, two main hypothesis polarise the discussion: (1) the demic diffusion model (Ammerman and Cavalli-Sforza, 1984) assumes that agricultural spreading was due to a movement of people, and therefore the genetic composition of European populations has changed substantially; and (2) the cultural diffusion model (Dennell, 1983; Zvelebil and Zvelebil, 1988) holds that the technological transition was performed without substantial population movement, by cultural transmission, suggesting that current patterns of genetic diversity should have their roots in the Palaeolithic.

Other models have been advanced, mainly mixed scenarios of the opposite models (Zvelebil, 2000) or suggesting a patchy dispersion of agriculture (different models and influences) to the several European regions (Zvelebil, 2000; Pinhasi et al., 2000; Lahr et al., 2000).

Presently, genetic studies have failed to clarify the issue, not only due to current computational limitations, but mainly because these models are over-simplistic in the sense that produce predictions for patterns of diversity that must be easily recognisable (e.g. Barbujani et al. 1994; Cavalli-Sforza et al. 1993; Chikhi et al. 1998a,b). It is not surprising the popularity of the demic diffusion model among geneticists, since it predicts clinal variations of gene frequencies centred on the place of origin.

In Europe, the presence of frequency gradients was detected for various markers: (1) some classical loci (indeed it was the 1st principal component with focus in the Near East, that led to the formulation of the demic diffusion hypothesis; Menozzi et al., 1978; Cavalli-Sforza et al., 1993); (2) autosomal microsatellites (Chikhi et al., 1998a); (3) Y-STRs (Casalotti et al., 1999); and especially (4) some Y-SNP haplogroups (Semino et al., 1996; Rosser et al., 2000; Semino et al., 2000). In opposition, mtDNA describes a European homogeneous landscape (Comas et al., 1997), with almost 85% of the lineages being dated as Palaeolithic (Richards et al., 1996), favouring the cultural diffusion hypothesis. But even for mtDNA, an east-west gradient of pairwise differences has been discerned, and claimed to be compatible with expansion from the Middle East (Comas et al., 1997). Spatial autocorrelation has also revealed clines in the south (Simoni et al., 2000), and founder analysis applied to an enlarged European and Near Eastern database showed lesser Neolithic lineages with increasing distances from the Near East (Richards et al., 2000).

Unfortunately, the interpretation of the data in terms of decision between the main contributions to the genetic shaping of Europe remains speculative since not only the time estimates for population movements are still insufficiently accurate, but also they refer to geographically concordant axis of diffusion.

2- Genetic and historic backgrounds of Portugal and Mozambique

Traditional concept of population used for diploid markers has to be adapted for the use of uniparentally-transmitted markers, since no individual admixture is involved and the deme is just a group of lineages.

We have tried to study the Portuguese genetic background, in this context, by investigating what could be called the "ins" and "outs" to and from Portugal, in terms of lineages. In other words, we aimed to infer contributions to the nowadays Portuguese haploid diversity and, at the same time, to analyse the same kind of genetic impact made by the Portuguese through the large migrations that occurred at recent historic times - those resulting from the enterprises taken by the Portuguese (and other European maritime powers) in Africa, Asia and Americas.

So, while most studies in this field deal with phenomena that have occurred long ago, namely during Palaeolithic and Neolithic, our approach was focused on recent historic movements that were also able to redraw, at least partially, the genetic composition of populations at a worldwide scale.

These recent migrations had expectedly very different consequences across world regions and also considering male and female contributions. These expectations were a departure point for our research.

2.1- Portugal

2.1.1- The country context:

Portugal occupies the westernmost coast of Europe, extending to 90,000 square kilometres, in a North-South axis, and being only accompanied at its north by the Spanish province of Galicia.

Portuguese Iberian political borders were established as early as 1249 (Saraiva, 1993), and kept almost immutable since then, which makes Portugal one of the first European countries recognisable in maps, contrasting with the characteristic changes observed in most European countries.

Before the establishment of Portugal as a political entity, several people of European origin (e.g. Iberians and Celts) settled in the region, as they were moving westwards. Apart from terrestrial migrations, it is well documented that Mediterranean Sea trading routes were established mainly by Phoenicians and Greeks, although it is not known to what extent those people really settled in Portugal. This Mediterranean

influence was surely much more important in the south of Portugal (and Spain) than in the north. Around 200 B.C., Iberia became a province of the Roman Empire. But this was just a geographic concept since the Portuguese territory was divided in two administrative areas, *Gallaecia*, corresponding to North Portugal (and Galicia) and *Lusitania*.

In the fifth century A.D., the Peninsula witnessed the invasion of various Barbarian tribes, some of which settled permanently and acquired political unity under the form usually known as Barbarian kingdoms. Continuous struggles for dominance between Alans (of Iranian ancestry), Suebi, Vandals (Germanic) and the Visigoths (from Sweden) lasted till the Islamic conquest, but their demographic influx is considered to have been very low, around 5% (Serrão and Marques, 1993).

The Islamic rule, which begun in the seventh century (711 A.D.), and extended to all Iberia except the very northern region (North Portugal, Galicia, Asturias and Leon), brought a significant influence of North African origin. At that time, there were around 500,000 inhabitants in the territory that is today Portugal, a number that was enlarged by the successive migrations of Arabs and Berbers (Medina, 1998). The first military contingent was made up mainly of Berbers, many of which returned to North Africa, but some settled in the new territory, and those who have not brought their family married locally. The Arab rulers installed in rich or fertile regions, while to Berbers (as "second class" Muslims), poorer and unstable northern regions were given (Medina, 1998). A second Arab contingent, composed of Syrians and Egyptians, arrived in Iberia after 742, to help fight a Berber insubordination; around a 1,000 Egyptians from this contingent is known to have settled in Baixo Alentejo and Algarve (South Portugal) (Serrão and Marques, 1993). These migrations were not restricted to military contingents, and from the eighth century till the tenth, many civil groups belonging to several tribes entered Iberia. Their presence mostly influenced South and Central Portugal, as judged from toponymical, architectural and cultural traits. By the time the most southern Portuguese region was conquered by the Christians (1249), a significant ethnic-religious Muslim minority, including both slaves and free people, was present. Especially in South Portugal, but also in Central Portugal, those Moors lived in communities called "Mourarias", amounting to at least 16 around the year 1300. The peninsular northwest, almost untouched during the Muslim rule, received also large groups of Moors, as servers-settlers, in the twelfth and thirteenth centuries (Serrão and Marques, 1996).

The available census of the Portuguese testify that population density was kept low over many centuries: one million at the beginning of the fifteenth century; between one million and 1,5 million from 1527 to 1532; 2 million in 1640; 2,5 million by 1758; and 3 million at the turn to the nineteenth century (Russell-Wood, 1998).

Also documented is the presence of sub-Saharan individuals that were enslaved and brought to Portugal since as early as 1440 up to 1750 (when the entrance of slaves was forbidden, although many were still coming, either furtively or accompanying their owners). The overall European intake of slaves, during this period, was estimated as 200,000 (Thomas, 1998), the biggest fraction corresponding to Portugal, so that in 1550, the country as a whole had probably 40,000 slaves and Lisbon boasted 10,000 resident slaves in a population of 100,000 (Thomas, 1998). In the mid-sixteenth century, the birth of slaves' children was stimulated in Portugal for internal traffic purposes. Inter-breeding between autochthonous individuals and African slaves certainly occurred and the predominant mating must have been between slave African females and Portuguese males, due to social pressures as well as to legal constraints: offspring of slave females would be slaves (economic profit), whereas offspring of slave males would not.

The main growth of Portuguese population was only observed by the turn of the nineteenth century and, at present, it amounts to 10 million, with a clear trend – that started in the 60's of last century - for higher population densities in the coastal regions, and a general depopulation of the hinterland particularly in the south.

Published genetic studies on Portuguese populations are scarce and mainly based on autosomal markers (e.g., Amorim et al., 1996; Espinheira et al., 1996; Gusmão et al., 1997; Souto et al., 1998), and have not addressed regional structuring.

The same happens with the only study for HVRI mtDNA, in a small sample of 56 Portuguese (Côrte-Real et al., 1996) that is analysed in an Iberian-North African context. So, it was not possible to get information on possible population structure of Portugal at the mtDNA level. But important observations were done, namely the detection of 3 North African sequences and 1 typical of West Africa.

Concerning Y-chromosome, the situation is the same as described for autosomal markers: a single report on STR defined haplotype data from North Portugal (online at http://ystr.charite.de).

We aimed therefore to study the Portuguese mainland, using the non-recombining markers, mtDNA (HVRI and HVRII) and Y-BMs, taking into

consideration the potential of these markers in detecting population structure. We have divided the country into three main regions, North, Central and South, whose borders were defined by the two main rivers, Douro and Tagus. As we have seen above, this division along a North-South axis is supported by geographic and historic facts.

2.1.2- The Iberian context

The Iberian Peninsula, with an area of nearly 600,000 square kilometres, is defined by three natural borders: the Pyrenees, the Mediterranean Sea and the Atlantic Ocean.

As Straus et al. (2000) pointed out, the study of the Upper Palaeolithic in Iberia, although the scarcity of data gathered so far, seems very promising for several reasons. First, Iberia was one of the richest macro-regions of continuous human settlement in Pleistocene Europe, with modern humans reaching the Ebro River around 33,750-32,500 years before present, and colonising all the Peninsula between 32,500-27,500 years, although much more slowly than the rest of Europe (Bocquet-Appel and Demars, 2000). Second, in Iberia, modern humans and Neanderthals coexisted for 2,500 years (from 30,000 to 27,500, just before Neanderthals have disappeared; Bocquet-Appel and Demars, 2000). Third, Iberia was a major refugium for human populations during the Last Glacial Maximum (LGM), with an apparent explosion of archaeological sites, especially in Andalucia (Straus et al., 2000). Fourth, one of the wealthiest collections of the Palaeolithic rupestral art can be found in Iberia, both in caves and in the open air.

It is likely that the Upper Palaeolithic sites in Iberia were denser in the peri-coastal regions (many on the now-underwater continental shelf), since, with the exception of the major river basins such as those of the Ebro, Douro, Tagus, Guadiana and Guadalquivir, high tablelands and mountains, often with poor soils, dominate the interior of the Peninsula. The oceanic limestone regions would have been favoured in terms of more moderate temperatures, more abundant precipitation, shelter and diverse resources. And only in the LGM, the interior of Iberia was settled.

Neolithic is documented in Iberia between 7,000-6,000 years, as a result of one of the two waves of agriculture diffusion in Europe. This wave moved along the western Mediterranean coast to southern France and finally the Iberian Peninsula, while the other, originated in the Hungarian Plains, advanced northwards towards the Czech

Republic and Slovakia, and so forth (Pinhasi et al., 2000). Some archaeologists (Zvelebil, 2000; Zilhão, 1997) are advancing a model of "Leapfrog colonisation" or "enclave migration" for Iberia, which consists in a "selective colonisation of an area by small groups, who target optimal areas for settlement, thus forming an enclave, or colony, among native inhabitants" (Zvelebil et al., 2000).

Although east-west clines dominate the overall European picture, a north-south axis played an important role in modelling the population dynamics in this Peninsula. In fact, since the Neolithic, a Mediterranean influence, mediated by more advanced cultures can be traced till historical times. The Islamic occupation of Iberia is the best example: in opposition to an almost untouched Northern region (North Portugal, Galicia, Asturias and Leon), there was an increasing influence on the regions towards south (Algarve in South Portugal remained 5 centuries under the Islamic rule, and Granada, South Spain, 8 centuries, capitulating to Christian government only in 1492, the same year Colombo reached America).

Both Iberian countries were the key players of the "Discoveries Period". Both were actively involved in the Atlantic slave trade, although the sub-Saharan slave intake to Spain was more restricted than the one to Portugal. Slaves traded by the Spanish (but also by the Portuguese) were mainly directed to the Spanish colonies in South and Central America.

Many genetic studies focusing on Iberia have already been published, and special attention has been devoted to the issues of: (1) the ancestry of the Basque population, said to be representative of the Palaeo-Iberian population, and considered untouched (genetically) by the Neolithic movement; (2) the North-African genetic influence in Iberia.

The group of Arnaiz-Villena (Arnaiz-Villena et al., 1997, 1999) is the main defender of a clear proximity between Iberians (including Basques) and North Africans based upon HLA and linguistics. They claim that Palaeo-North Africans were forced to move northwards, reaching Iberia, Canary Islands, Sardinia, Crete and Etruria, when the area of present Sahara got dried around 10,000-6,000 years before present. A restricted Neolithic genetic influence in Iberia is also advanced. However, Comas et al. (1998), also based upon HLA data, found evidence for a different scenario.

Simoni et al. (1999) have found that the main boundary separating Mediterranean populations is the one between northern and southern coasts, even at the

Gibraltar strait. Only minor genetic boundaries were observed around linguistically (such as the Basque Country) or geographically (Sardinia, Corsica, Balearic Islands and Cyprus) isolated groups. Evidence for major Neolithic population replacements is also presented.

Bosch et al (2001) reached a similar conclusion on the Mediterranean Sea as a strong barrier to gene flow, by analysing 44 Y-BMs and 8 Y-STRs. A major Upper Palaeolithic influence was detected in both coasts, in opposition to a minor Neolithic contribution. Bi-directional gene flow was estimated in mean as 4% from Europe to North Africa and 7% in the opposite direction (in this last case, a double value of 14% was detected in Andalusians, in southern Iberia). Hurles et al. (1999) reported that haplogroup 22 is found at highest frequencies in Basques (11%) and Catalans (22%), being almost sporadic in the rest of Iberia, taken as evidence for gene flow across the linguistic barrier.

MtDNA analyses showed signals of expansion in the Upper Palaeolithic (Côrte-Real et al., 1996; Salas et al., 1998), and one of the lowest levels of intake of Neolithic sequences in Europe (Richards et al., 2000). The east-west decreasing gradient of mean pairwise sequence differences (Comas et al., 1997) attains a minimum in Galicia (besides Basques; Salas et al., 1998), as an edge of European genetic variation. Some North-African lineages (haplogroup U6) were however detected in Iberia (Côrte-Real et al., 1996; Salas et al., 1998), although this influence is far less important than in the Canary Islands (Rando et al., 1999). Typical sub-Saharan L lineages were also found, in low frequencies but higher than in other European countries (Côrte-Real et al, 1996; Salas et al., 1998), representing maybe a signature of the slave intake. Finally, concerning the role of Iberia as a refuge during the LGM, with subsequent expansion, around 10,000-15,000 years ago, conflicting results were reported: Torroni et al. (1998), describing the autochthonous haplogroup V (southwestern France/Cantabria) as the marker for that expansion towards northeastern Europe (later rechecked by applying a more precise haplogroup classification; Torroni et al., 2001); Izagirre and de la Rúa (1999), failing to detect V sequences in Neolithic Basque populations; and Simoni et al. (2000), finding no clinal pattern for haplogroup V between Iberia and Saami.

2.1.3- The European context

Archaeology and Genetics agree with respect to the range of time for the earliest date for the occupation of Europe by anatomically modern humans: sometime around 40,000 years before present. The main proportion of present European genetic diversity was established during the Upper Palaeolithic (Richards et al., 1996, 2000; Semino et al., 2000). The classical scenario of stable and small Palaeolithic population size, in the context of the limited resources of a hunting-gathering economy (Landers 1992), followed by expansions after the development of agriculture (the Neolithic transition; Hassan 1973), has been reviewed recently due to new data on the density of Mesolithic sites (extremely dense in the north and scarce in central and eastern Europe; Pinhasi et al., 2000).

Whatever the relative importance of Upper Palaeolithic settlement and Neolithic replacement, both had a geographically similar origin (around the Near East) and followed an identical (southeastern-northwestern) axis of dispersion. The consequent decreasing diversity gradient was recognised for proteins and nuclear DNA (Chikhi et al., 1998a,b), as well as for Y-BMs (Rosser et al., 2000; Semino et al., 2000) and Y-STRs (Casalotti et al., 1999). In contrast, no population structuring at a European scale is found for mtDNA, although Simoni et al. (2000) have observed a cline around the Mediterranean Sea, but without significant differentiation between southern and northern European populations.

2.2- Portugal – as contributor for...

As early as the beginning of the fifteenth century, Portugal began its maritime expansion. The start of this period can be set by the conquest of some North African coastal towns (the first was Ceuta, in 1415). Some Portuguese settlements (called *Praças*) were established in the nowadays Morocco, but the number of Portuguese was always low, and composed mainly by deported convicts and soldiers (Farinha, 1999). From the Moroccan southern settlements, Portuguese contacted the established commercial routes between Moors and sub-Saharans, joining the trade for gold and slaves.

With the main goal of reaching India by sea, as an alternative commercial route, the exploration towards south started in 1419 (Russell-Wood, 1998). This strategy implied the establishment of bases (forts and factories) along the African coast.

In 1500, Pedro Álvares Cabral, in another voyage to India, made landfall on the coast of Brazil, in South America, which would become, till its independence in 1822, the largest and richest Portuguese colony.

Such a small and under-populated nation could only afford to be present in Africa, Asia and America, for such a long period, in extraordinarily low numbers, not exceeding 10,000 Portuguese overseas by the end of the sixteenth century (Russell-Wood, 1998). Those migrants were predominantly males, and many of them made permanent residence overseas, marrying local females, and bringing up the progeny as Portuguese and Catholics. The paradox of the apparent overwhelming Portuguese presence around the world is thus explained by the constant movement of people inter and intra-continentally, specially bureaucrats, soldiers and missionaries. Thus, with the exception of Azores and Madeira archipelagos, Brazil was the only overseas possession where the Portuguese demographic impact was significant (Russell-Wood, 1998).

If the movement of Portuguese people was not substantial during this long period, the same cannot be said about the forced relocation of local populations between continents. Among those, Indians, Amerindians and coolie Chinese deserve to be mentioned, but the most important, both numerically and also in terms of geographical dispersion, was the Atlantic slave trade of sub-Saharan Africans.

Sub-Saharan and North African slaves are known to had been introduced during the Roman Period in Europe and also under the Muslim rule in Iberia. In fact, the trans-Saharan slave trade between West and North Africa probably began as early as 1,000 B.C. The southern shore of the Mediterranean was the main market place, but the presence of black slaves is also reported in Iberia, Italy, Sardinia and France (Thomas, 1998). Nevertheless, the numbers involved are insignificant in comparison with those after the "Discoveries".

The first disembark of black slaves (235 individuals from Arguin) traded by Portuguese occurred in Lagos, south Portugal, in 1444. By the end of the nineteenth century the total number of slaves rose to around 13,000,000, from which about 15.4% were from the Slave Coast (Dahomey, Adra, Oyo), 15.4% from Benin to Calabar, 15.4% from Senegambia (in Arguin) and Sierra Leone, 11.5% from Gold Coast (Ashanti) and 23.1% from Congo/Angola (Thomas, 1998). In the eighteenth century,

due to the loss of Portuguese control over some of the western possessions, Mozambique and Madagascar became the major source of the slaves shipped, around 7.7% of the total trade.

The slaves were taken first to Portugal and the Atlantic islands (Madeira, Azores, Canary Islands, Cape Verde and São Tomé and Príncipe), but soon also to India and Macao, and in more massive numbers to Brazil. The slave intake was first forbidden to Europe, around mid-eighteenth century, a time when slave trade was still expanding to the American colonies, where it ended officially in 1808, but continued at a lower rate for several more decades.

It is therefore not surprising that Americas have been the stage for most genetic studies on admixture. In all the cases, the reported data show a bias in the mating pattern, with predominance of interbreeding between European males and African or/and Native American females. For Brazil, mtDNA data allowed the estimation of the proportions: 33% Amerindian, 28% African and 49% European (Alves-Silva et al., 2000); while for Y-chromosome, only 2.5% were attributable to sub-Saharan Africans and none to Amerindians (Carvalho-Silva et al., 2000). In Colombia, European Y-chromosome proportions are variable across populations, attaining a maximum (~94%) in Antioquia; while mtDNA Amerindian lineages amount to 90%-95% (Mesa et al., 2000; Carvajal-Carmona et al., 2000). Similar proportions were described for North America (Parra et al., 1998).

2.2.1- Mozambique

Mozambique is located in southeast Africa, and according to a 1985 census, the population size is around 17,913,000 individuals, including 7,000 Chinese and 15,000 Indians. The number of languages listed for Mozambique is 33, all belonging to the Bantu supra-family language.

2.2.1.1- Before the Portuguese arrival – the Bantu expansion

The first inhabitants of Mozambique were hunter-gatherers, Khoisan nomadic tribes that explored the savannah resources southeastern the equatorial forest. These populations were substantially replaced by Bantu farmers, as linguistic (Renfrew, 1987) and mtDNA (Excoffier and Schneider, 1999) data seem to point out.

The Bantu expansion was a major population movement in the African continent, occurring in several waves and directions, and being responsible for the dispersal of farming to southern and central Africa. The linguistic evidence points to a Bantu origin in the vicinity of the Cross River valley near the present-day border between Nigeria and Cameroon (Newman, 1995). Around 5,000 years ago, the Bantu expansion began in two directions: southwestern, attaining the equatorial rain forest by 3,500 years ago and eastern, entering the fringes of the interlacustrine region in what is now Uganda, by 3,000 years ago, forming a new core, the eastern Bantu core area. From this new core, two new expansions moved towards South Africa: one group along the Ruvuma River toward the coast, reaching present-day Natal by the end of the third century A.D. and the other along the shores of Lake Malawi, through what is now eastern Zimbabwe reaching the northern Transvaal around A.D. 500.

In the last decade, data on African mtDNA has been accumulated with the aim of unravelling some of the demographic phenomena that have contributed to the settlement of the continent. This task is still at a primitive stage and is more difficult in Africa than in the rest of the world because the characteristic demographic phenomena of recurrent migration, population expansion and contraction including bottlenecks, population sub-structure generated by limits on gene flow, and more recent admixture effects have occurred over a longer time depth. In addition, there is a rather poor archaeological context in which to set the genetics. Several studies of restriction-fragment length polymorphisms (RFLPs) (Cann et al., 1987; Chen et al., 1995), of control region sequences (Vigilant et al., 1991; Soodyall, 1993; Krings et al., 1999) or a combination of both (Watson et al., 1997; Graven et al., 1995; Chen et al., 2000) have involved African populations. However, the sampling is still surprisingly patchy, and quite poor in the southeast.

Nonetheless, at the mitochondrial level, some Bantu expansion markers have been proposed: Watson et al. (1997) pointed to a 9-bp deleted subset of haplogroup L1a

(and also Soodyall et al., 1996) and to the haplogroup L3b, while Bandelt et al. (2001) proposed that haplogroup L3e1a must had been prominent in the southern Bantu expansion. How informative these markers are in following the several Bantu waves of migration towards south remains to be clarified.

The most recent data on Bantu Y-chromosome (Thomas et al., 2000) indicated a Bantu expansion in southeastern Africa occurring around 3,000-5,000 years before present, although a substantial reduction in diversity (2 Y-STR haplotypes, one-step neighbours, comprise almost half the Bantu Y chromosomes) was also detected. This diversity reduction has no parallel at the mtDNA level (Bandelt and Forster, 1997).

2.2.1.2- After the Portuguese arrival – Mozambican contribution to the slave trade

Portuguese first touched the Mozambican coast in 1498, during the voyage of Vasco da Gama towards India. After a period of trading, the first Portuguese settlement was established in Sofala in 1505, and subsequently many forts were constructed. In 1752, Mozambique was proclaimed a colony of Portugal and it remained so till the independence in 1975.

South-east Africa was an important source of slaves, from 1643 onwards, when individuals from Mozambique and Madagascar constituted a major portion of the slaves shipped by the Portuguese to the former European colonies in America, e.g. Brazil and the Caribbean, in such a way that "by the eighteenth century this commerce, directed to the Americas, was more important on that coast than anywhere else" (Thomas, 1998). Records point to ~1,000,000 slaves originating from Mozambique/Madagascar in a total of ~13,000,000 leaving African ports (Thomas, 1998). This shift in the place of origin of the slaves was accompanied by a change in the place of export: European ports were closed to the slave trade, being by that time all conducted to the former European colonies in America, e.g., Brazil. So the eastern sub-Saharan contribution to the European African sequences, sporadically detected in some countries of this continent (Côrte-Real et al., 1996; Salas et al., 1998) is expected to have been reduced, compared to its influence in America.

Material and Methods

This section is a brief summary describing the population samples analysed and the broad methodology used in this thesis. Further details can be found in the papers included in this work.

1- Population samples

We studied the populations described in table 1. In all cases, the individuals inhabited and were born in the region under consideration and were unrelated.

Portugal was divided into three regions, North (NP), Central (CP) and South (SP), defined by the rivers Douro and Tagus. Mozambicans belonged to different ethnic groups (mainly to Changana, Ronga, Chope, Bitonga and Matsua), but all were Bantu speakers (http://www.sil.org/ethnologue/ countries/Moza.html).

Table 1: Populations and sample sizes studied for the different markers.

	Y-BMs	mtDNA (HVRI and HVRII)
Portugal		
North	328	100
Central	118	82
South	49	59
Mozambique	68	109

2- DNA extraction

DNA was extracted by the resin Chelex-100 method (Lareu et al., 1994) from total blood (15µl if liquid or 1cm if spot).

3- Y-BMs screening

A total of ten biallelic markers were typed. Their selection was the result of an international effort intended to standardise the detection of the same polymorphisms in European populations under the "Y chromosome European Diversity Project" co-ordinated by Dr. Mark A. Jobling. There were previous indications (Hammer et al. 1998; Hurles et al. 1999; Santos and Tyler-Smith 1996; Semino et al. 1996; Underhill et al. 1997; Zerjal et al. 1997) that the haplogroups (Hap) they define were likely to be

found within European or circum-European populations. The Y-BM marker DYS257, was not analysed in this thesis because it is phylogenetically equivalent to 92R7.

Primers, PCR (Polymerase Chain Reaction) amplification conditions and, when necessary, restriction enzyme treatment were as specified in Table 2. Amplicons for YAP and 12f2 and restricted fragments for the other markers were run on polyacrylamide gels (T9%; C5%) and visualised by silver staining (Budowle et al., 1991).

Table2: Methodological conditions for the Y-BMs studied.

Y-BM	Primers	Amplification conditions	Restriction enzime	Allele sizes (bp)
SRY-1532	tcc tta gca acc att aat ctg g aaa tag caa aaa atg aca caa ggc	94°C 30s / 59°C 30s / 72°C 30s (34x)	DraIII	0 (A)– 167 1 (G)- 112+55
SRY-8299	aca gca cat tag ctg gta tga c tct ctt tat ggc aag act tac g	94°C 30s / 62°C 30s / 72°C 60s (33x)	BsrBI	0 (G)- 362+147 1 (A)- 509
SRY-2627	agg tct ttt ttg cct tct ta atg cac ggt ttc ttt tga	94°C 30s / 54°C 30s / 72°C 120s (33x)	BsiHKAI	0 (C)- 1242 1 (T)- 298+944
92R7	gac ccg ctg tag acc tga ct gcc tat cta ctt cag tga ttt ct	94°C 30s / 62°C 30s / 72°C 60s (33x)	HindIII	0 (C)- 197+512 1 (T) – 709
Tat	gac tct gag tgt aga ctt gtg a gaa ggt gcc gta aaa gtg tga a	94°C 30s / 60°C 30s / 72°C 30s (33x)	NlaIII	0 (T)- 85+27 1 (C)- 112
YAP	cag ggg aag ata aag aaa ta act gct aaa agg gga tgg at	94°C 30s / 54°C 30s / 72°C 30s (33x)	-	0 (YAP^-)- 150 1 (YAP^+)- 455
sY81	agg cac tgg tca gaa tga ag aat gga aaa tac agc tcc cc	94°C 30s / 60°C 60s / 72°C 60s (32x)	NlaIII	0 (A)- 102+65+42 1 (G)- 144+65
LLY22g	cca ccc agt ttt atg cat ttg ata gat ggc gtc ttc atg agt	94°C 30s / 55°C 60s / 72°C 60s (33x)	HindIII	0 (C)- 500+230+120 1 (A)- 650+500+230+120
M9	gca gca tat aaa act ttc agg aaa acc taa ctt tgc tca agc	94°C 30s / 58°C 30s / 72°C 30s (33x)	HinfI	0 (C)- 182+93+66 1 (G)- 248+93
12f2 *+amplicon* *3Sry15* *3Sry16*	tct tct aga att tct tca cag aat tg ctg act gat caa aat gct tac aga tc *ctt gat ttt ctg cta gaa caa g* *tgt cgt tac ata aat ggg cac*	94°C 30s / 59°C 30s / 72°C 40s (34x)	-	0 (presence) 1 (absence)

The nomenclature used for the Y-BM haplogroups is specified in table 3.

Table 3: Y-BM haplogroups definition and nomenclature.

Hap	YAP	SRY-8299	92R7	SRY-1532	SRY-2627	Tat	sY81	M9	LLY22g	12f2
2	0	0	0	1	0	0	0	0	0	0
16	0	0	0	1	0	1	0	1	1	0
21	1	1	0	1	0	0	0	0	0	0
8	1	1	0	1	0	0	1	0	0	0
3	0	0	1	0	0	0	0	1	0	0
1	0	0	1	1	0	0	0	1	0	0
22	0	0	1	1	1	0	0	1	0	0
9	0	0	0	1	0	0	0	0	0	1
4	1	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
12	0	0	0	1	0	0	0	1	1	0
26	0	0	0	1	0	0	0	1	0	0

4- MtDNA screening (HVRI, HVRII and RFLPs)

MtDNA was amplified using the primers L15997 (5'-CAC CAT TAG CAC CCA AAG CT-3') and H16401 (5'-TGA TTT CAC GGA GGA TGG TG-3') for HVRI and L48 (5'-CTC ACG GGA GCT CTC CAT GC-3') and H408 (5'-CTG TTA AAA GTG CAT ACC GCC A-3') for HVRII. The temperature profile was 95°C for 10 sec., 60°C for 30 sec. and 72°C for 30 sec., for 35 cycles of amplification, following an initial denaturation step at 95°C for 4 min. and ending with a final extension at 15°C for 10 min.

The amplified products were purified with Microspin^{TM} S-300 HR columns (AB Applied Biosystems), according to the manufacturer's specifications. The sequence reactions were carried out using the kit Big-Dye ^{TM} Terminator Cycle Sequencing Ready Reaction (AB Applied Biosystems), with one of the primers above described, in both forward and reverse directions. The temperature profile was 96°C for 15 sec., 50°C for 9 sec. and 60°C for 2 min., for 30 cycles of amplification, with an initial denaturation at 96°C for 4 min. and a final extension at 60°C for 10 min.

A protocol based on $MgCl_2$/ethanol precipitation was used for post-sequencing reaction purification of samples, which were then applied in a 6% PAGE gel and run in an automatic sequencer ABI 377 (AB Applied Biosystems).

The nucleotide positions considered for analysis were between bp 16024 and 16383 for HVRI and between 73 and 340 for HVRII (in the numbering system of Anderson et al., 1981). Sequence classification into haplogroups is indicated in the publications for the European and African populations studied here.

To check the assignment to haplogroup L3 and its subclusters, in the Mozambican sample, the following RFLPs were checked in putative members of L3: 2349*Mbo*I (present in L3e), 3592*Hpa*I (absent in L3 in general), 8616*Mbo*I (absent in L3d) and 10084*Taq*I (present in L3b). PCR amplifications were performed using primers and conditions described by Torroni et al. (1992). Digestions were carried out according to the manufacturer's specifications and the resulting fragments were run in 9% polyacrylamide gels and visualised by silver staining (Budowle et al., 1991).

RESULTS

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS

Mismatch distributions, a statistical tool described for mtDNA analysis, is here applied for the first time to Y-BMs. Demographic inferences, at a worldwide scale, could be therefore undertaken in an approach both for mtDNA and Y-BMs. The evaluation of possible artefacts due to non-demographic parameters (such as the number of polymorphic sites and the mutation rate) was performed. This is essentially dealt in the papers:

ARTICLE 1

PEREIRA, L., DUPANLOUP, I., ROSSER, Z.H., JOBLING, M.A., BARBUJANI, G. (2001) Y-chromosome mismatch distributions in Europe. *Mol. Biol. Evol.* **18**:1259-1271.

ARTICLE 2

DUPANLOUP, I., PEREIRA, L., BERTORELLE, G., CALAFELL, F., PRATA, M.J., AMORIM, A., BARBUJANI, G. No evidence of demographic expansions in human Y-chromosome biallelic variation. (in preparation).

Mismatch analysis was also applied as a tool to evaluate the distance (per locus or per number of repeat units) between all the pairs of Y-STR haplotypes in a sample. This approach intends to improve the information content of database matching. In fact, due to the high mutation rate in Y-STRs, the probability of two Y-haplotypes becoming identical by state is high, and hence cannot be disregarded. We discuss the theoretical assumptions of this approach and its application at an European scale, showing its relevance for forensics in the paper:

ARTICLE 3

PEREIRA, L., PRATA, M.J., AMORIM, A. (2002) An evaluation of the proportion of identical Y-STR haplotypes due to recurrent mutation. In: Sensabaugh G.F., Lincoln, P.J., Olaisen, B. (eds) *Progress in Forensic Genetics* 9 (in press). Elsevier Science, Amsterdam.

Y-Chromosome Mismatch Distributions in Europe

Luísa Pereira,† Isabelle Dupanloup,† Zoë H. Rosser,‡ Mark A. Jobling,‡ and Guido Barbujani†*

*Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP) and Faculdade de Ciências da Universidade do Porto, Porto, Portugal; †Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy; and ‡Department of Genetics, University of Leicester, Leicester, England

Ancient demographic events can be inferred from the distribution of pairwise sequence differences (or mismatches) among individuals. We analyzed a database of 3,677 Y chromosomes typed for 11 biallelic markers in 48 human populations from Europe and the Mediterranean area. Contrary to what is observed in the analysis of mitochondrial polymorphisms, Tajima's test was insignificant for most Y-chromosome samples, and in 47 populations the mismatch distributions had multiple peaks. Taken at face value, these results would suggest either (1) that the size of the male population stayed essentially constant over time, while the female population size increased, or (2) that different selective regimes have shaped mitochondrial and Y-chromosome diversity, leading to an excess of rare alleles only in the mitochondrial genome. An alternative explanation would be that the 11 variable sites of the Y chromosome do not provide sufficient statistical power, so a comparison with mitochondrial data (where more than 200 variable sites are studied in Europe) is impossible at present. To discriminate between these possibilities, we repeatedly analyzed a European mitochondrial database, each time considering only 11 variable sites, and we estimated mismatch distributions in stable and growing populations, generated by simulating coalescent processes. Along with theoretical considerations, these tests suggest that the difference between the mismatch distributions inferred from mitochondrial and Y-chromosome data are not a statistical artifact. Therefore, the observed mismatch distributions appear to reflect different underlying demographic histories and/or selective pressures for maternally and paternally transmitted loci.

Introduction

Analyses of mitochondrial and Y-chromosome polymorphisms in humans tend to suggest that the female- and the male-transmitted gene pools evolved under somewhat different conditions. Although some studies (notably, Poloni et al. 1997; see Bertranpetit 2000) found congruent results, Y-chromosome data seem to be characterized by lower diversity within populations and more significant spatial structure than mitochondrial data (see, e.g., Jorde et al. 2000). It is unclear why this is so. Selection (Excoffier 1990; Wise et al. 1998; Wyckoff, Wang, and Wo 2000) and different demographic histories for females and males (Sajantila et al. 1996; Seielstad, Minch, and Cavalli-Sforza 1998; Perez-Lezaun et al. 1999) are two popular types of explanations.

Changes in population size tend to leave recognizable signatures in the patterns of nucleotide diversity. Therefore, the distribution of pairwise sequence differences in a sample (or, simply, the mismatch distribution) contains information on the population's history (Rogers and Harpending 1992). The genealogy of a population of constant size is expected to have long deep branches (Donnelly 1996); mutations occurring along these branches will be shared by several lineages, which will result in an irregular or ragged distribution of pairwise sequence differences. Conversely, the genealogy of a population that has substantially grown in size has long terminal branches, and the mutations that have occurred along these branches, i.e., most mutations, will be specific to a single lineage (Donnelly 1996). Under these conditions, one expects unimodal mismatch distributions (Harpending et al. 1993; Harpending 1994; Marjoram and Donnelly 1994), whose means, under an infinite-sites mutation model, increase as a function of the time elapsed after population growth (Sherry et al. 1994). However, different selective regimes may mimic the effects of changes in population size.

In addition, recombination acts as a confounding factor, for it brings together chromosome regions that evolved independently. For that reason, with only one exception (Alonso and Armour 2001), human mismatch distributions have only been studied at the mitochondrial level so far, based on both restriction fragment length polymorphisms (RFLPs) (Rogers and Harpending 1992; Harpending 1994) and hypervariable region I (HVRI) sequences (Excoffier and Schneider 1999). Almost all of these distributions are unimodal. In general, both the mean and the variance are highest (and therefore the curve is smoothest) in African populations, lower among Asians and Americans, and lowest among Europeans. The mean mismatch is related to the time of the expansion through the mutation rate (Rogers and Harpending 1992; Rogers and Jorde 1995), so the dates of the main demographic expansions are estimated at around 110,000 years ago in East Africa, 70,000 years ago in the rest of Africa and Asia, 55,000 years ago in America, and 40,000 years ago in Europe and in the Middle East (Excoffier and Schneider 1999). The few exceptions are represented by populations which may have undergone recent bottlenecks, thus presumably losing the typical genetic features of expanding populations (Excoffier and Schneider 1999).

Abbreviations: HVRI, hypervariable region I; RFLP, restriction fragment length polymorphism.

Key words: Y chromosome, polymorphism, human, mismatch distribution, population expansion.

Address for correspondence and reprints: Guido Barbujani, Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, I-44100 Ferrara, Italy. E-mail: bjg@unife.it.

Mol. Biol. Evol. 18(7):1259–1271. 2001

Because only mitochondrial mismatch distributions have been studied so far, it is not yet known whether the inferred expansions affected the entire populations or only their female components. In this study, we calculated the mismatch distributions in a data set of Y-chromosome biallelic polymorphisms in 48 population samples from Europe and the Mediterranean area (data from Rosser et al. 2000). The results obtained differ sharply from those observed for mitochondrial data. To understand the causes of that discrepancy, a database of European mitochondrial sequences was reanalyzed and patterns in the mismatch distributions of computer-simulated populations were studied with the aim of determining the effects of expansions versus constant population sizes.

Materials and Methods

Databases

The database of Y-chromosome biallelic markers we considered (Rosser et al. 2000) comprised data on 48 populations (listed in table 1), for a total of 3,677 individuals. Most were European, but populations from the eastern and southern shores of the Mediterranean sea, the Caucasus region, and Greenland were also included.

The 11 biallelic markers considered, namely, 2 insertion/deletion and 9 single-nucleotide (SNP) polymorphisms, defined 10 alleles, 6 of them polymorphic or subpolymorphic (frequency > 0.01) and 4 of them rare in Europe. For the sake of consistency with other studies, and because micro- and minisatellite variation has been observed within most such alleles (Jobling and Tyler-Smith 1995, 2000; Karafet et al. 1999), we refer to each of them as a "haplogroup" (De Knijff 2000). A single minimum-spanning tree could be constructed on the basis of those 10 haplogroups (fig. 1). Therefore, there was reason to believe that each of the 11 polymorphisms of interest represented the effect of a unique mutational event, without ambiguities. Full haplogroup descriptions and frequencies are in Rosser et al. (2000).

A database of mitochondrial DNA HVRI sequences in Europe (updated from Simoni et al. 2000) was used for comparisons of patterns of genetic diversity in maternally transmitted genes. Of the 48 population samples available there, 21 were selected that approximately matched the geographic location of the samples considered in this analysis of Y-chromosome diversity.

Mismatch Distributions and Neutrality Tests

Mismatch distributions and gene diversity, i.e., the probability that two randomly sampled chromosomes differ from each other (Nei 1987), were estimated for both databases by ARLEQUIN, version 2.0 (Schneider, Roessli, and Excoffier 2000). The fit of the observed distribution of mismatches to a model of population expansion was tested by a bootstrap approach, also implemented in ARLEQUIN. Note that for this method, the null hypothesis is one of population expansion, due to the fact that there is no quantitative expectation as for the shape of the mismatch distribution in a stationary population (Harpending 1994), whereas under the hypothesis of expansion, a parameter τ estimated from the data allows one to predict the average mismatch.

Within each population sample (including simulated samples; see below), departures from mutation-drift or mutation-selection equilibrium were tested by means of Tajima's D and Fu's F_S. In Tajima's (1989*a*, 1989*b*) test, the parameter $\theta = 2N\mu$ (where N is the population size and μ is the mutation rate) is independently estimated twice, once from the number of polymorphic sites and once from the average mismatch in the sample. Differences between the two estimates are then attributed to selection or to the demographic history of the population studied. Similarly, Fu's (1997) F_S statistic compares the observed number of alleles in a sample with the number of alleles expected if the population has kept a constant size.

The significance of D and F_S was tested by randomization. By the coalescent simulation program implemented in the ARLEQUIN package (Schneider, Roessli, and Excoffier 2000), separately for each sample studied, we generated random samples from a hypothetical stationary population whose parameter $\theta = 2N\mu$ was equal to the average number of observed pairwise differences. For each sample, the procedure was repeated 1,000 times, in every case recomputing the D and F_S statistics so as to obtain empirical null distributions of these statistics and hence the probability of the observed D and F_S values under the hypothesis of demographic stationarity.

Reanalysis of Reduced Mitochondrial Data Sets

To understand whether the Y-chromosome mismatch distributions could reflect an insufficient number of sites considered, we reanalyzed the mtDNA database in two ways.

1. From one randomly selected European population sample, we repeatedly calculated the mismatch distribution, each time removing 10 nucleotide sites from the initial 360 (starting from positions 16024–16033 of the Cambridge reference sequence; Anderson et al. 1981), until 10 sites were left.
2. In the 21 European samples, four sets of 11 polymorphic sites were selected so as to analyze the same number of sites for mtDNA and for the Y chromosome. The criteria of selection were as follows. In two runs of the analysis (data sets A and B), 11 sites were selected at random; in one run (data set C), 10 highly variable sites were used; and in one run (data set D), selection was among poorly polymorphic sites.

Coalescent Simulations

We generated samples from stationary and expanding populations by Monte Carlo simulation to see whether the shapes of the Y-chromosome mismatch distributions described in this study were compatible with some form of demographic expansion. The simulation algorithm was based on the coalescent process with su-

Table 1
Measures of Genetic Diversity Estimated from Y-Chromosome Data

Population	N	Hp	Average Mismatch	P(exp)	H	D	$P(D)$	F_S	$P(F_S)$
Northern Central Europe	939	9	1.48 ± 0.90	0.100	0.514	−0.003	0.505	0.863	0.682
Bavaria	80	6	2.07 ± 1.17	0.460	0.701	−0.187	0.442	−1.684	0.263
Belgium	92	7	1.60 ± 0.96	0.009	0.551	−0.688	0.244	−2.661	0.120
Cornwall	51	2	0.89 ± 0.63	0.030	0.297	−1.846	0.014	−7.573	0.000
Holland	84	6	2.02 ± 1.15	0.210	0.698	−0.224	0.380	−1.715	0.253
East Anglia	172	7	1.65 ± 0.98	0.110	0.585	−0.340	0.358	−1.550	0.314
France	40	7	2.19 ± 1.24	0.280	0.691	−0.467	0.281	−2.908	0.103
Ireland	257	6	1.05 ± 0.70	0.030	0.329	−0.966	0.152	−3.306	0.110
Scotland	43	4	0.84 ± 0.61	0.090	0.364	−2.010	0.004	−8.576	0.000
West Scotland	120	4	1.19 ± 0.77	0.130	0.437	−1.069	0.111	−3.754	0.072
Iberia	537	7	1.87 ± 1.07	0.210	0.562	1.198	0.875	2.993	0.879
Basque	26	3	0.77 ± 0.58	0.210	0.440	−2.440	0.000	−11.317	0.000
North Portugal	328	6	1.91 ± 1.09	0.130	0.575	0.244	0.598	−0.134	0.580
South Portugal	57	6	2.27 ± 1.27	0.140	0.637	−0.138	0.422	−1.906	0.238
Spain	126	7	1.69 ± 0.10	0.310	0.509	−0.430	0.328	−1.889	0.255
South-Central Europe	384	9	2.44 ± 1.33	0.600	0.795	1.288	0.876	2.530	0.881
Bulgaria	24	5	2.33 ± 1.32	0.900	0.772	−0.708	0.217	−4.325	0.011
Greece	36	6	2.40 ± 1.33	0.680	0.798	−0.301	0.376	−2.768	0.092
Italy	99	6	2.34 ± 1.29	0.300	0.728	0.254	0.560	−0.799	0.447
Romania	45	7	2.49 ± 1.37	0.320	0.810	−0.036	0.458	−2.022	0.216
Sardinia	10	4	2.53 ± 1.49	0.130	0.778	−1.560	0.040	−27.76	0.000
Slovenia	70	6	2.34 ± 1.29	0.280	0.745	0.071	0.508	−1.370	0.357
Yugoslavia	100	7	2.14 ± 1.20	0.580	0.706	0.016	0.481	−1.169	0.364
Eastern Europe	835	8	2.73 ± 1.45	0.100	0.773	2.384	0.976	5.556	0.991
Belarus	41	7	2.53 ± 1.39	0.200	0.728	−0.049	0.475	−2.170	0.172
Chuvash	17	7	2.65 ± 1.49	0.670	0.882	−0.692	0.244	−5.486	0.001
Czech Republic	53	6	2.56 ± 1.40	0.300	0.779	0.159	0.521	−1.532	0.317
Estonia	207	8	2.71 ± 1.44	0.080	0.762	1.076	0.816	0.923	0.760
Germany	30	6	1.87 ± 1.10	0.320	0.731	−1.047	0.120	−4.621	0.009
Hungary	36	5	2.52 ± 1.39	0.680	0.773	−0.154	0.423	−2.530	0.123
Latvia	34	4	2.63 ± 1.44	0.030	0.711	−0.069	0.407	−2.497	0.134
Lithuania	38	4	2.70 ± 1.47	0.020	0.656	0.099	0.537	−2.070	0.200
Mari	48	6	2.67 ± 1.45	0.050	0.776	0.231	0.574	−1.555	0.293
Poland	112	7	1.99 ± 1.13	0.330	0.645	−0.114	0.437	−1.318	0.315
Russia	122	8	2.78 ± 1.48	0.050	0.727	0.908	0.776	0.274	0.678
Slovakia	70	8	2.42 ± 1.33	0.330	0.717	0.171	0.560	−1.220	0.376
Ukraine	27	5	2.55 ± 1.41	0.030	0.681	−0.354	0.332	−3.402	0.050
Scandinavia and Finland	325	8	2.45 ± 1.33	0.010	0.744	1.575	0.916	3.110	0.930
Denmark	56	6	1.90 ± 1.10	0.060	0.647	−0.586	0.267	−2.769	0.103
Finland	57	6	2.03 ± 1.16	0.030	0.569	−0.420	0.315	−2.411	0.142
Gotland	64	5	1.89 ± 1.09	0.050	0.599	−0.522	0.299	−2.515	0.144
North Sweden	48	6	2.25 ± 1.26	0.050	0.709	−0.268	0.372	−2.318	0.155
Norway	52	6	2.16 ± 1.22	0.110	0.727	−0.331	0.369	−2.339	0.153
Saami	48	4	2.58 ± 1.41	0.000	0.696	0.115	0.529	−1.716	0.257
Southern Mediterranean	156	5	1.27 ± 0.80	0.160	0.446	0.041	0.566	1.663	0.793
Algeria	27	4	1.60 ± 0.98	0.000	0.584	−1.446	0.038	−5.959	0.000
North Africa	129	5	1.13 ± 0.74	0.150	0.396	−1.111	0.112	−3.909	0.066
Turkey	212	8	2.18 ± 1.21	0.640	0.782	0.978	0.823	1.841	0.824
North Cyprus	45	7	2.05 ± 1.17	0.430	0.775	−0.554	0.284	−2.929	0.086
Anatolia	167	8	2.17 ± 1.21	0.490	0.779	0.303	0.620	−0.370	0.495
Caucasus	200	7	2.04 ± 1.15	0.230	0.743	0.750	0.784	2.228	0.865
Armenia	89	7	2.11 ± 1.19	0.240	0.757	−0.078	0.479	−1.413	0.318
Georgia	64	6	1.70 ± 1.01	0.390	0.682	−0.747	0.200	−3.031	0.086
Ossetia	47	6	2.27 ± 1.27	0.120	0.700	−0.265	0.370	−2.337	0.180
Northwest Atlantic	89	5	1.50 ± 0.91	0.010	0.534	0.207	0.630	1.734	0.819
Greenland	61	4	1.41 ± 0.87	0.020	0.455	−1.130	0.108	−4.170	0.027
Iceland	28	3	1.71 ± 1.03	0.130	0.659	−1.295	0.068	−5.425	0.002

NOTE.—N = sample size; Hp = number of different haplogroups observed; average mismatch = mean and standard deviation of mismatch distribution; P(exp) = probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time estimated from the data; H = gene diversity; D = Tajima's D; $P(D)$ = P value for D; F_S = Fu's F_S; $P(F_S)$ = P value for F_S.

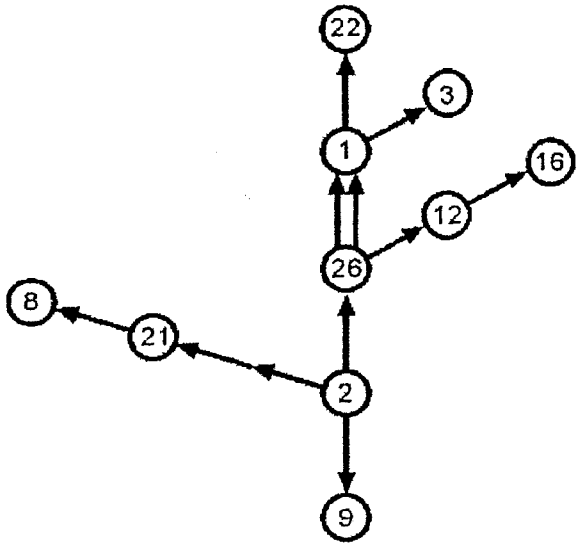


FIG. 1.—Network summarizing the evolutionary relationships among the 10 haplogroups observed in Europe. Each arrow represents one mutational event, whose probable direction is indicated by the arrow (from Rosser et al. 2000).

perimposed mutations, as described by Hudson (1990). Each sample was obtained by first generating its genealogy. Mutations were then randomly placed on the genealogy assuming they occurred according to a uniform and constant Poisson process.

First, we simulated 1,000 samples of genes under the assumption of a large and constant population size from a single panmictic deme. Each sample was composed of 80 individuals, i.e., 80 sets of 300 potentially variable sites. The size of the deme was 5,000 haploid individuals, and the mutation rate was 2×10^{-4} per generation for the whole sequence, i.e., 6.7×10^{-7} for each site. Although a plausible mutation rate for Y-chromosome biallelic polymorphisms seems to be around 5×10^{-7} (Hammer 1995; Jobling, Pandya, and Tyler-Smith 1997) or less (Thomson et al. 2000), we chose this higher rate so as to obtain similar mean pairwise differences in the simulated and in the real samples. The initial number of sites, 300, was also chosen because with those mutation rates, most sites (>97% in stationary populations) were monomorphic at the end of each simulation.

Second, four processes of exponential population expansion were simulated using the same coalescent approach, namely, (1) expansion from 50,000 years ago until now, with a 100-fold increase in population size, final effective population size N_0 = 50,000; (2) expansion from 50,000 years ago until now, with a 100-fold increase in population size, N_0 = 100,000; (3) expansion from 50,000 years ago until now, with a 100-fold increase in population size, N_0 = 200,000; and (4) expansion from 100,000 years ago until now, with a 100-fold increase in population size, N_0 = 100,000. The mutation rate was the same as that considered for the stationary populations. One thousand samples of 80 individuals were generated in this way for each of the four processes.

To more easily compare the simulation results, we defined three basic shapes of the mismatch distribution, namely, unimodal with a maximum at 0 (type 0), unimodal with a maximum >0 (type 1), and bimodal (type 2); examples are shown in figure 5.

Results

Mismatch Distributions and Neutrality Tests

Mismatch distributions obtained for Y chromosome biallelic markers were multimodal with one exception: the Chuvash population from Russia (fig. 2). Each distribution had at least two peaks, one at 0 and the other at a number of differences that varied among populations. These shapes reflect the fact that in many populations, most Y chromosomes belong to two frequent haplogroups, whereas other haplogroups occur at lower frequencies. Therefore, the peak at 0 differences corresponded to the comparisons between individuals that share the same allele, and the second peak was located at the mismatch representing the number of mutational steps separating the most frequent haplogroups. Where more than two haplogroups occurred at intermediate or high frequencies, there was a third, and sometimes a fourth, peak.

For 13 populations, the hypothesis of expansion could be rejected at the $P < 0.05$ level (fifth column of table 1). Although these probabilities were only nominal, Bonferroni's correction for multiple tests (Sokal and Rohlf 1995) confirmed significant overall departure from expansion expectations for the samples in this study. Conversely, all 21 mismatch distributions of the mtDNA samples appeared compatible with the effects of a population increase (fifth column of table 2). One parameter of the mismatch distribution, τ, estimates the time elapsed since population expansion (Rogers and Jorde 1995; Rogers 1995). This parameter is not reported in table 1 because we found little or no evidence for expansions in the shape of the Y-chromosome mismatch distributions.

It is possible to lump together Y-chromosome distributions based on their shapes; the clusters obtained in this way corresponded to sets of geographically near populations (fig. 3). This seems to be a consequence of the clinal variation shown by most nuclear markers in Europe (Chikhi et al. 1998; Casalotti et al. 1999; Quintana-Murci et al. 1999; Rosser et al. 2000; Barbujani and Bertorelle 2001). Most Western and Central European populations (British Isles, France, Belgium, and the Netherlands) showed a peak at three differences, i.e., the mutational distance between haplogroups 1 and 2, which represented the largest fraction of haplogroups there. Iberian populations (except Basques) showed an additional peak at five differences, resulting from the presence of a substantial number of haplogroup 21 chromosomes, which differ from those of haplogroup 1 by five mutational steps. The Southern-Central European and Turkish samples showed smoother distributions, reflecting larger numbers of different haplogroups. The

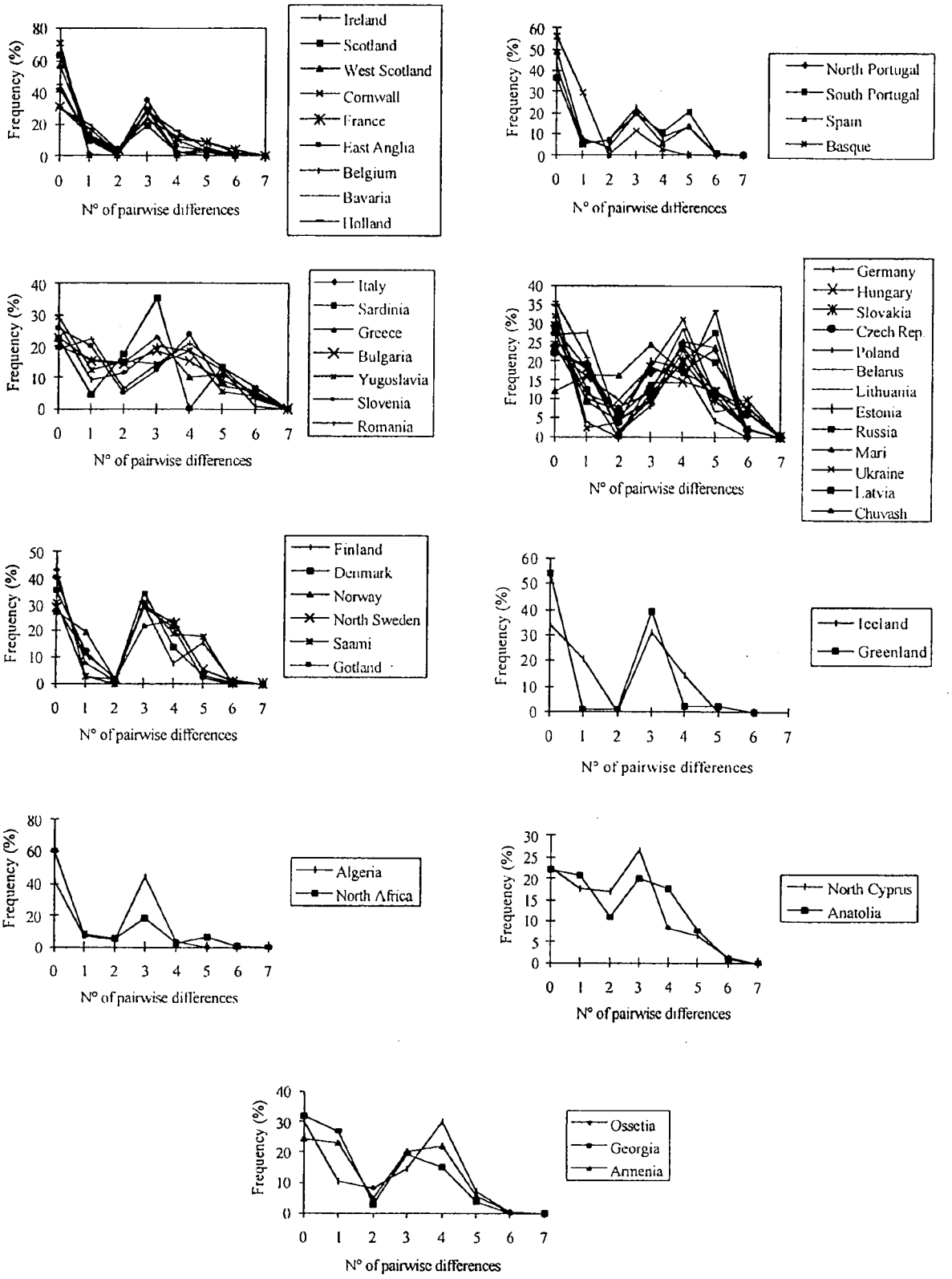


FIG. 2.—Mismatch distributions for Y-chromosome biallelic markers in 48 populations.

Table 2
Measures of Genetic Diversity Estimated from mtDNA Data

Population	N	Hp	Mean Mismatch	P(exp)	H	D	$P(D)$	F_S	$P(F_S)$
Basque	106	52	2.95 ± 1.56	0.430	0.936	−2.226	0.001	−26.498	0.000
Cornwall.....	69	45	3.89 ± 1.89	0.540	0.965	−2.127	0.002	−25.967	0.000
Sardinia	73	50	4.25 ± 2.13	0.790	0.956	−2.116	0.002	−25.824	0.000
Turkey	96	79	5.45 ± 2.65	0.850	0.988	−2.159	0.000	−25.345	0.000
Greece	48	37	4.67 ± 2.33	0.700	0.991	−2.001	0.005	−25.647	0.000
Bulgaria	30	22	4.55 ± 2.30	0.430	0.977	−1.878	0.016	−14.397	0.000
Denmark.....	32	20	3.41 ± 1.79	0.810	0.934	−1.586	0.041	−12.986	0.000
Sweden.....	32	27	4.58 ± 2.31	0.840	0.988	−1.871	0.014	−24.394	0.000
Georgia.....	45	28	4.57 ± 2.29	0.930	0.964	−1.774	0.013	−18.602	0.000
Germany.....	108	70	3.92 ± 1.98	0.630	0.973	−2.115	0.000	−25.912	0.000
Iceland	53	38	4.83 ± 2.40	0.300	0.979	−1.574	0.031	−25.597	0.000
Italy.....	115	95	6.14 ± 2.94	0.910	0.993	−2.139	0.000	−25.085	0.000
Norway.....	30	20	3.26 ± 1.73	0.860	0.954	−1.798	0.017	−14.418	0.000
France.....	111	73	3.83 ± 1.94	0.820	0.961	−2.217	0.000	−25.955	0.000
Belgium	33	25	3.35 ± 1.77	—[a]	0.974	−2.196	0.003	−26.498	0.000
Saami	240	37	3.60 ± 1.83	0.300	0.799	−1.172	0.107	−16.668	0.000
Estonia	28	23	4.36 ± 2.22	0.790	0.979	−1.728	0.018	−18.773	0.000
Portugal	54	38	3.60 ± 1.85	1.000	0.934	−1.988	0.007	−26.077	0.000
Finland.....	79	46	3.74 ± 1.91	0.720	0.970	−1.909	0.005	−26.048	0.000
Spain.....	74	61	5.25 ± 2.57	0.680	0.987	−2.043	0.001	−25.454	0.000
Russia.....	103	64	4.22 ± 2.11	0.960	0.965	−2.010	0.002	−25.784	0.000

NOTE.—N = sample size; Hp = number of different haplogroups observed; mean mismatch = mean and standard deviation of mismatch distribution; P(exp) = probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time estimated from the data; H = gene diversity; D = Tajima's D; $P(D)$ = P value for D; F_S = Fu's F_S; $P(F_S)$ = P value for F_S.

[a] This value is missing because the last-squares procedure to fit model mismatch distribution and observed distribution did not converge after 1,800 steps.

increased within-population diversity seems to be largely due to the fact that differently oriented clines, from the southeast into the northwest and from south to north, converge in that area. As a consequence, haplogroups that were very rare or absent elsewhere tended to reach substantial frequencies in these populations. For northern-eastern samples, other peaks at three, four, and five differences were evident, resulting from differences between haplogroups 2 and 16, haplogroups 2 and 3, and haplogroups 3 and 16, respectively.

Mismatch distributions appeared to be bi- or multimodal, regardless of whether single samples (fig. 2) or groups thereof (fig. 3) were analyzed. Accordingly, insufficient sample size does not seem to be a plausible explanation for that finding. Bimodal distributions were also observed in the few Asian and North African samples available, and among Greenlanders. We do not know of any suitable Y-chromosome data set which could allow comparison with other continents.

Mitochondrial and Y-chromosome variation also differed when summarized by means of Tajima's D and Fu's F_S (seventh and ninth columns of tables 1 and 2). For mitochondrial data, both statistics were negative and significant (with the exception only of Saami), and all F_S values were significant at the 0.001 level. Conversely, when estimated from Y-chromosome data, most values of F_S, and especially of D, were insignificant, and the latter were even positive in 12 cases. Such positive D values were not associated with any spatial pattern that we could recognize. On the contrary, the four negative and significant values occurred in linguistic (Basques) or geographic isolates (Sardinians, Scots, Cornish; note, however, the small sample size in Sardinia), also showing low gene diversity. Gene diversity seems to be patterned in space (sixth column of table 1), with comparatively high values in the south and in the east, as also observed for mitochondrial variation (Comas et al. 1997).

By and large, taken at face value, mismatch distributions, Tajima's D, and Fu's F_S would suggest that the European male population has had a different history than the female population and that only the latter has increased substantially in numbers. Before drawing any conclusions, however, it is better to ask whether those apparent differences between sexes may simply be some sort of statistical artifact. Only 11 Y-chromosome polymorphic sites were studied, versus more than 200 for mtDNA. Might that have biased the results?

Reanalysis of Reduced Mitochondrial Data Sets

Initially, we repeatedly estimated the mitochondrial mismatch distribution and related statistics in one randomly chosen sample, the Cornish sample, each time considering a decreasing number of sites, from 360 to 10. Figure 4 shows that the characteristic, unimodal pattern of the mismatch distribution is always maintained through repeated reductions of molecular information in the 69 HVRI sequences considered. As expected, the mean moved left and the variance decreased as fewer and fewer nucleotide positions were considered. The reduction in the diversity was not linear; it was slow in the first steps, and it accelerated later, probably reflecting the fact that the 5′ and 3′ extremes of the HVRI are less variable than is the intermediate segment.

As the number of sites considered decreased, D always remained negative and lost significance when 50 sites were left, whereas F_S was significant even when

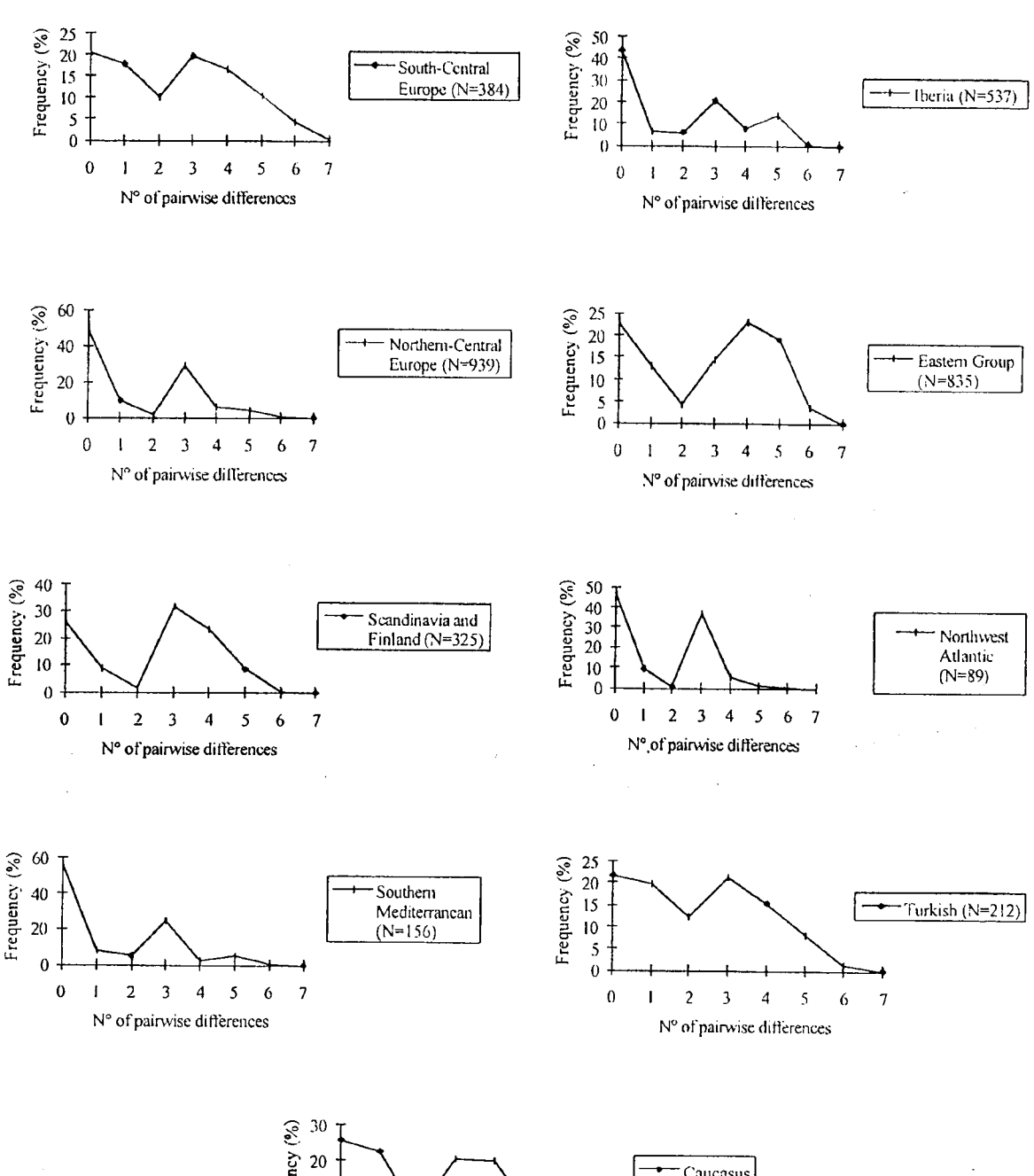


FIG. 3.—Mismatch distributions for Y-chromosome biallelic markers in the various population groups.

30 nucleotide sites were analyzed (the last 20 sites were monomorphic; table 3). This may indicate that although D is more robust than F_S for small sample sizes, F_S is more sensitive when few polymorphic positions are considered.

As a second test, we estimated the mismatch distributions from four different sets of 11 mtDNA polymorphic sites in the 21 populations for which both mtDNA and Y-chromosome data were available (table 4). The shape of the mismatch distribution was almost

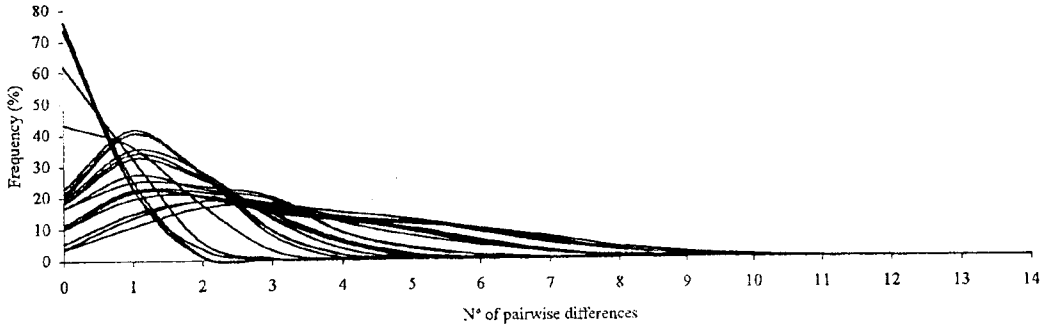


FIG. 4.—Mismatch distributions observed through analysis of subsets of data obtained by progressively removing 10 (constant or polymorphic) sites from the mitochondrial hypervariable region I of the Cornish population.

always unimodal, with only three vaguely bimodal shapes in a total of 81 mismatch distributions simulated (in the other three cases, all in data set D, no site was polymorphic among sequences, and therefore no statistic could be estimated). These results did not depend on the way polymorphic sites were selected, i.e., at random or on the basis of their levels of variation. Predictably, analyses of highly variable sites (data set C) yielded the highest means and variances.

Once again, Fu's F_S and Tajima's D values were negative with one exception (Saami), Also, confirming data obtained in the previous simulation, F_S negative values were statistically significant in the vast majority of the cases. In contrast, Tajima's D negative values did not often reach significance, although they tended to do so in data set D.

It is evident that even when the number of sites considered is the same as that available for the Y chromosome, mtDNA mismatch distributions are unimodal and therefore different from those calculated for the Y chromosome.

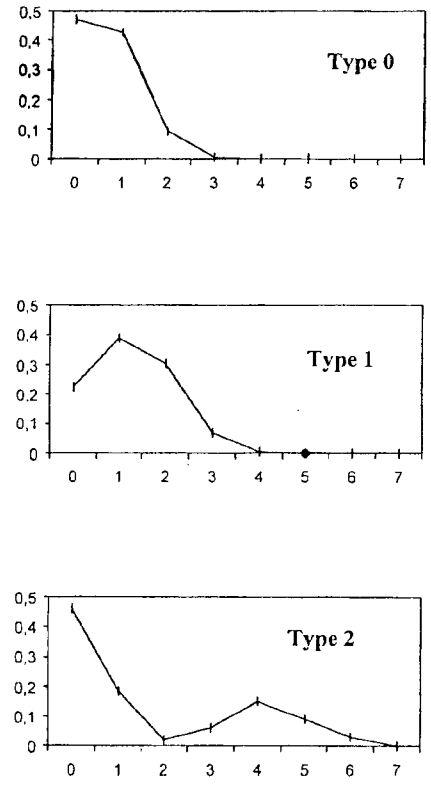


FIG. 5.—Scheme of the most common shapes of the mismatch distribution observed in the simulations.

Simulations

In simulated stationary populations, the average mismatch was higher than that for expanding populations (table 5) and close to the expected value, i.e., the parameter θ used to generate the simulated samples. This is what one expects under mutation-drift equilibrium (Rogers and Harpending 1992; Rogers et al. 1996). Also, the observed standard deviation (1.22) was close to the expectation (1.26) derived by Tajima (1983, eq. 30). In expanding populations, conversely, the average mismatch and its standard deviation were reduced, but both increased with the size of the population after the expansion and with the time since the expansion event.

Less than 12% of the mismatch distributions in the samples generated under stationarity showed a geometric form (type 0), and <20% had a single peak (type 1). Around 70% of the distributions showed multiple peaks (type 2). Conversely, for expanding populations, the number of bell-shaped mismatch distributions increased with the size of the population after the expansion and with the time to the expansion event and, correspondingly, the number of type 0 distributions decreased. Also, the proportion of distributions with multiple (generally two) peaks increased with the time since expansion and with the size of the population after expansion. The largest value was observed for populations that had expanded for 50,000 years, reaching an effective size N_0 of 200,000, where 21.8% of the mismatch distributions had two peaks (type 2).

In synthesis, the bimodal mismatch distributions observed in Y-chromosome European samples can be generated in simulations of both stationary and postexpansion populations. However, depending on the modes of the simulated expansion, bimodality is from 3–10

Table 3
Diversity Parameters Estimated in the Analysis of Mitochondrial Hypervariable Region I in the Cornish Population (N = 69), Considering a Decreasing Number of Sites Each Time

No. of Sites	Hp	Mean Mismatch	P(exp)	H	D	$P(D)$	F_S	$P(F_S)$
360	45	3.89 ± 1.98	0.540	0.965	−2.127	0.002	−25.973	0.000
310	44	3.62 ± 1.86	—	0.962	−2.186	0.000	−26.106	0.000
260	42	3.43 ± 1.77	0.640	0.945	−2.172	0.001	−26.209	0.000
210	38	2.55 ± 1.39	0.740	0.888	−2.292	0.000	−26.771	0.000
160	29	1.68 ± 1.00	0.920	0.819	−2.292	0.000	−27.730	0.000
110	23	1.33 ± 0.84	—	0.799	−2.127	0.000	−24.147	0.000
100	22	1.30 ± 0.83	—	0.786	−2.088	0.004	−22.365	0.000
90	20	1.22 ± 0.79	—	0.771	−1.950	0.006	−19.262	0.000
80	13	0.81 ± 0.59	0.700	0.567	−1.656	0.019	−10.700	0.000
70	8	0.50 ± 0.41	0.370	0.384	−1.878	0.005	−6.247	0.000
60	5	0.31 ± 0.32	0.550	0.266	−1.599	0.024	−3.135	0.021
50	5	0.28 ± 0.30	0.510	0.266	−1.410	0.058	−3.437	0.013
40	4	0.25 ± 0.29	0.480	0.240	−1.154	0.100	−2.239	0.042
30	4	0.25 ± 0.29	0.370	0.240	−1.154	0.105	−2.239	0.037
20/10.....	1	—	—	—	—	—	—	—

NOTE.—N = sample size; Hp = number of different haplogroups observed; mean mismatch = mean and standard deviation of mismatch distribution; P(exp) = probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time estimated from the data; H = gene diversity; D = Tajima's D; $P(D)$ = P value for D; F_S = Fu's F_S; $P(F_S)$ = P value for F_S. Some values are missing because the least-squares procedure to fit model mismatch distribution and observed distribution did not converge after 1,800 steps.

times as frequent in stationary populations as in expanding populations.

In the simulated stationary populations, both D and F_S showed a wide distribution centered on 0 (552 negative values out of 1,000 simulations for D, and 577 negative values for F_S). Only in 3.9% and 8.7% of the cases, respectively, were these values significant. Conversely, when expansions were simulated, F_S was always negative and was significant at the 5% level in >95% of the iterations. Tajima's D was also negative in nearly all cases of expansion and reached statistical significance in >85% of the simulations.

Discussion

The mismatch distributions inferred in this study from Y-chromosome biallelic markers were bimodal and did not resemble those inferred in the same populations from mtDNA data, which were unimodal. Statistical tests failed to reject a neutral equilibrium model for European Y-chromosome variation, whereas there was highly significant departure from equilibrium for mtDNA data (Merriwether et al. 1991; Excoffier and Schneider 1999) (table 5).

Models of population expansion do not predict, even transitorily, the presence of multiple peaks in the mismatch distribution (Rogers and Harpending 1992; Rogers and Jorde 1995). In Slatkin and Hudson's (1991) simulations, bimodal distributions with a peak at 0 were observed only for stationary populations. Conversely, expanding populations showed no instance of bimodality, and there were virtually no observations for mismatch = 0 (Slatkin and Hudson 1991, p. 560). Similar results were obtained by Harpending et al. (1998), who also showed that gene trees with a few well-differentiated alleles, much like those described in this study for the Y chromosome, are the rule in populations whose size has stayed constant or contracted.

Three lines of evidence suggest that the results of this study are not simply a statistical artifact:

1. By analyzing biallelic variation as we did in this study, one neglects other possible, but so far unde-

Table 4
Median Observed Values in the Analysis of Reduced mtDNA Data Sets Comprising 11 Sites in the 21 Population Samples of Table 2

Data Set[a]	Mean Mismatch	P(exp)	H	D	$P(D)$	No. significant D	F_S	$P(F_S)$	No. Significant F_S	No. of Unimodal Distributions
A	0.905	0.630	0.566	−1.117	0.132	1	−4.038	0.005	18	20/21
B	0.875	0.740	0.579	−0.914	0.185	1	−4.943	0.011	18	21/21
C	1.778	0.440	0.798	−0.411	0.315	0	−9.977	0.000	19	19/21
D	0.095	0.235	0.091	−1.418	0.039	11	−3.399	0.004	15	18/18[b]

NOTE.—P(exp) = probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time estimated from the data; H = gene diversity; D = Tajima's D; $P(D)$ = P value for D; F_S = Fu's F_S; $P(F_S)$ = P value for F_S.

[a] Sites considered: data set A—16051, 16069, 10693, 16104, 16124, 16126, 16129, 16145, 16163, 16172, and 16182; data set B—16187, 16189, 16192, 16222, 16223, 16224, 16245, 16249, 16256, 16261, and 16264; data set C—16069, 16126, 16183, 16189, 16192, 16223, 16270, 16278, 16294, 16298, and 16311; data set D—16051, 16104, 16142, 16167, 16176, 16215, 16248, 16257, 16265, 16290, and 16327.

[b] In three samples, no mismatch distribution could be calculated because all sequences were identical at the sites considered.

Table 5
Simulation Results

Simulated Population	Average Mismatch[a]	One Mode, Type 0[b]	One Mode, Type 1[b]	Two Modes Type 2[b]	D[c]	Signif.[d]	F_S[e]	Signif.[d]
Stationary.....	1.960 (1.218)	0.113	0.185	0.702	−0.041 (0.945) [−2.348, 2.775]	0.039	−0.159 (2.579) [−11.287, 10.955]	0.087
Expanding 1....	0.820 (0.375)	0.563	0.373	0.063	−1.826 (0.359) [−2.408, −0.019]	0.855	−10.193 (4.176) [−28.159, −0.568]	0.964
Expanding 2....	1.077 (0.457)	0.385	0.516	0.099	−1.959 (0.349) [−2.577, −0.244]	0.921	−14.208 (5.453) [−31.869, −1.039]	0.983
Expanding 3....	1.486 (0.651)	0.197	0.585	0.218	−1.987 (0.369) [−2.611, 0.005]	0.908	−17.245 (6.454) [−33.023, 1.320]	0.989
Expanding 4....	1.658 (0.563)	0.108	0.775	0.117	−2.015 (0.290) [−2.577, −0.746]	0.959	−19.018 (5.643) [−30.639, −4.013]	0.999

[a] Average mean of the mismatch distribution across 1,000 simulations. The standard deviation is shown in parentheses.
[b] Fraction of the 1,000 simulations showing shapes of the mismatch distribution.
[c] Average D (Tajima 1989*a*) over 1,000 simulations. The standard deviation is shown in parentheses, and the range of observed values is in brackets.
[d] Fraction of significant cases at the 5% level (out of 1,000 simulations).
[e] Average F_S (Fu 1997) over 1,000 simulations. The standard deviation is shown in parentheses, and the range of observed values is in brackets.

tected, polymorphism at other sites. However, that also applies to the mitochondrial RFLP studies which, as we have seen, show very different, unimodal mismatch distributions (Harpending 1994).

2. When we reanalyzed subsets of mitochondrial HVRI data, the shape of the distributions remained unimodal, and Tajima's D and Fu's F_S remained negative (although only the latter was always significant).
3. In our simulations, bimodal distributions appeared much more frequently in stationary than in expanding populations. We did not estimate a likelihood ratio because the numerical results depended on admittedly approximate expansion parameters and mutation rates. However, all factors considered, constant population sizes seemed roughly 3–10 times as likely as expansions.

The demographic scenarios we simulated were very simple. It is customary to model expansions as either instantaneous (see, e.g., Rogers and Jorde 1995) or exponential phenomena (this study; see also Excoffier and Schneider 1999), but populations may have grown in other ways. To mention just one, temporary contractions may have punctuated periods of general expansion, and that may have had an impact on levels and patterns of genetic diversity (Excoffier and Schneider 1999). However, in the absence of more sophisticated testable models, this study does not suggest that the Y-chromosome and the mitochondrial mismatch distributions differ from each other because of the limited resolution offered by the available Y-chromosome data.

Nonunimodal distributions (and insignificant Tajima's and Fu's statistics) are regarded as evidence that populations evolved under neutrality, without significantly increasing in size (Harpending et al. 1993). Departures from the expected shape can reflect adaptation (if one assumes constant population size), demographic changes (if one assumes neutrality), or both and can also occur when mutation rates vary across nucleotide sites (Aris-Brosou and Excoffier 1996). In principle, therefore, the different results obtained in the analysis of maternally and paternally transmitted genes in Europe may be due to differences in mutation mechanisms, differences in selective regimes, differences in past demographic history, or combinations thereof.

Might some sort of distorted mutational process have generated spurious multimodal mismatch distributions? Each of the 11 polymorphisms considered in this study probably results from a mutation that occurred only once (Rosser et al. 2000), and therefore these polymorphisms meet the assumptions of the model underlying the theory of mismatch distributions, the infinite-sites model (Rogers and Harpending 1992). In addition, simulations suggest anyway that the shape of the mismatch distribution tends to faithfully reflect the demographic history of a population, despite even substantial violations of the infinite-sites model (Rogers at al. 1996). Finally, recurrent mutation at some sites, reflecting mutation rate heterogeneity, may mimic the effects of population growth (Aris-Brosou and Excoffier 1996), but here the problem is the opposite, i.e., how to explain the apparent constancy of population size suggested by data. In short, we cannot rule out yet-to-be-discovered peculiarities of the Y-chromosome mutation process, but even if they existed, at present it is hard to imagine how such peculiarities could account for the results of this study.

Adaptation is the second factor. Tests based on the comparison of within-species and between-species nucleotide diversity have failed so far to reject the hypothesis of neutrality for Y-chromosome markers in comparisons between humans and mice (Nachman 1998), but not in comparisons between humans and Old World monkeys (Wyckoff, Wang, and Wu 2000). In addition, some loci of the Y chromosome are known to affect male fertility (Vogt 1997), and some detrimental mutations have been shown to occur more frequently on a particular Y-chromosome background (Jobling et al. 2000). Therefore, some role of selective pressures appears probable. However, Nachman's (1998) results and other simple calculations suggest that selection can ex-

plain, per se, only a small fraction of the human Y chromosome variation (Bertranpetit 2000).

The differences described here between mitochondrial and Y-chromosome data seem therefore to reflect, at least in part, the effects of past demographic phenomena. There are a few complications, though. Mismatch distributions from chromosomes subject to recombination contain little unambiguous evolutionary information. However, Tajima's D and Fu's F_S have been estimated within autosomal regions with no apparent recombination (Harding et al. 1997; Hey 1997; Zietkiewicz et al. 1998), and their values do not appear to depart from neutral, stationary expectations. Fay and Wu (1999) proposed that the smaller mitochondrial population size (one fourth that of autosomal genes) has caused a stronger impact of past population bottlenecks on mitochondrial variation. That interpretation seems at odds with the results of this study, because indices of Y-chromosome diversity resemble those estimated at the other nuclear loci, despite the fact that, in principle, Y-chromosome and mitochondrial effective population sizes should be the same.

It thus seems necessary to envisage either different demographic histories for males and females, with the former leaving a stronger mark on autosomal variation, or some combination of demographic changes and selective processes. Schematically, three hypotheses appear compatible with the available data:

1. The European female population increased in size; the male population did not.
2. Neither population increased in size, but there was disruptive selection for mtDNA.
3. Both populations increased in size, and there was purifying selection on the Y chromosome.

The data we analyzed do not allow, at present, discrimination among these hypotheses. However, it is worth noting that a small population size does not necessarily mean small numbers of individuals (of males in the present case). A high variance of reproductive success among individuals reduces the effective population size (Crow 1958). If the number of offspring has been generally more variable among males than among females, the effective population size inferred from Y-chromosome diversity is expected to be less (which is what this study suggests), and the genetic differences between populations tend to be greater (which has been demonstrated by, among others, Seielstad, Minch, and Cavalli-Sforza [1998] and Perez-Lezaun et al. [1999]). In other words, European men may have been approximately as numerous as European women, but a fraction of men may have left many descendants at each generation, and another fraction may have left just a few or none. The correlation in family size across generations, demonstrated in Canadian pedigrees (Austerlitz and Heyer 1998), would increase the evolutionary impact of this effect.

The first two hypotheses appear to contrast with the population expansions inferred from microsatellite (Pritchard et al. 1999) and sequence (Shen et al. 2000) Y-chromosome variation. Those studies, however, considered largely non-European samples and different Y-chromosome polymorphisms; it may be that the demographic history of Europe has been peculiar or that the biallelic polymorphisms we considered offer insight into a different period. Biallelic Y-chromosome markers, with their low mutation rates ($<10^{-8}$ per site per year; Hammer 1995; Jobling, Pandya, and Tyler-Smith 1997; Thomson et al. 2000), may only be able to reveal ancient population growth (see Takahata 1995). Conversely, fast-evolving markers in the mitochondrial genome (estimates of the mutation rate per site per year are 8.6×10^{-5} for the hypervariable region [Stoneking et al. 1992] and about 4.5×10^{-5} for RFLP [Rogers and Harpending 1992]) may contain information on more recent demographic changes.

At any rate, it is not impossible to reconcile the findings of Pritchard et al. (1999) and Shen et al. (2000) with those of the present study. If hypothesis 3 proved correct, the apparent constancy of the European male population size, as inferred from mismatch distributions, Tajima's tests, and Fu's tests, would be due to some form of purifying selection, ultimately concealing the effects of the demographic growth that previous studies of the Y-chromosome have recognized.

Acknowledgments

We are grateful to Giorgio Bertorelle, Laurent Excoffier, Antonio Amorim, and Chris Tyler-Smith, who discussed the results of this study with us and critically read the manuscript. Chris Tyler-Smith and Tatiana Zerjal also gave us access to unpublished data, and we thank them for that. This research was supported by funds from the Italian Ministry of the Universities (MURST COFIN 99) and from the University of Ferrara. L.P. was supported by a Ph.D. grant from Fundação para a Ciência e a Tecnologia (PRAXIS XXI/BD/13632/97), I.D. by a grant from the Swiss National Research Council (FNRS) for Perspective Investigators, Z.H.R. by a BBSRC Studentship, and M.A.J. by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant number 057559).

LITERATURE CITED

ALONSO, S., and J. A. ARMOUR. 2001. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc. Natl. Acad. Sci. USA 98:864–869.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL et al. (14 coauthors). 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

ARIS-BROSOU, S., and L. EXCOFFIER. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. 13:494–504.

AUSTERLITZ, F., and E. HEYER. 1998. Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. Proc. Natl. Acad. Sci. USA 95:15140–15144.

BARBUJANI, G., and G. BERTORELLE. 2001. Genetics and the population history of Europe. Proc. Natl. Acad. Sci. USA 98:22–25.

BERTRANPETIT, J. 2000. Genome, diversity, and origins: the Y chromosome as a storyteller. Proc. Natl. Acad. Sci. USA **97**:6927–6929.

CASALOTTI, R., L. SIMONI, M. BELLEDI, and G. BARBUJANI. 1999. Y-chromosome polymorphisms and the origins of the European gene pool. Proc. R. Soc. Lond. B Biol. Sci. **266**: 1959–1965.

COMAS, D., F. CALAFELL, E. MATEU, A. PEREZ-LEZAUN, E. BOSCH, and J. BERTRANPETIT. 1997. Mitochondrial DNA variation and the origin of the Europeans. Hum. Genet. **99**: 443–449.

CHIKHI, L., G. DESTRO-BISOL, V. PASCALI, V. BARAVELLI, M. DOBOSZ, and G. BARBUJANI. 1998. Clinal variation in the DNA of Europeans. Hum. Biol. **70**:643–657.

CROW, J. F. 1958. Some possibilities for measuring selection intensities in man. Hum. Biol. **30**:1–13.

DE KNIJFF, P. 2000. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am. J. Hum. Genet. **67**: 1055–1061.

DONNELLY, P. 1996. Interpreting genetic variability: the effects of shared evolutionary history. Pp. 25–50 *in* K. WEISS, ed. Variation in the human genome. Wiley, Chichester, England (CIBA Foundation Symposium).

EXCOFFIER, L. 1990. Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. **30**:125–139.

EXCOFFIER, L., and S. SCHNEIDER. 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA **96**:10597–10602.

FAY, J. C., and C. I. WU. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16**:1003–1005.

FU, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915–925.

HAMMER, M. F. 1995. A recent common ancestry for human Y chromosomes. Nature **378**:376–378.

HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60**:772–789.

HARPENDING, H. C. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66**:591–600.

HARPENDING, H. C., M. A. BATZER, M. A. GRUVEN, L. B. JORDE, A. R. ROGERS, and S. T. SHERRY. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95**:1961–1967.

HARPENDING, H. C., S. T. SHERRY, A. R. ROGERS, and M. STONEKING. 1993. The genetic structure of ancient human populations. Curr. Anthropol. **34**:483–496.

HEY, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14**:166–172.

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. FUTUYMA and J. ANTONOVICS, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford, England.

JOBLING, M. A., A. PANDYA, and C. TYLER-SMITH. 1997. The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med. **110**:118–124.

JOBLING, M. A., and C. TYLER-SMITH. 1995. Fathers and sons: the Y chromosome and human evolution. Trends Genet. **11**: 449–456.

———. 2000. New uses for new haplotypes: the human Y chromosome, disease, and selection. Trends Genet. **16**:356–362.

JOBLING, M. A., G. WILLIAMS, K. SCHIEBEL, A. PANDYA, K. MCELREAVEY, L. SALAS, G. A. RAPPOLD, N. A. AFFARA, and C. TYLER-SMITH. 2000. A selective difference between human Y-chromosomal DNA haplotypes. Curr. Biol. **8**: 1391–1394.

JORDE, L. B., W. S. WATKINS, M. J. BAMSHAD, M. E. DIXON, C. E. RICKER, M. T. SEIELSTAD, and M. A. BATZER. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am. J. Hum. Genet. **66**:979–988.

KARAFET, T. M., S. L. ZEGURA, O. POSUKH et al. (14 co-authors). 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am. J. Hum. Genet. **64**: 817–831.

MARJORAM, P., and P. DONNELLY. 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136**: 673–683.

MERRIWETHER, D. A., A. G. CLARK, S. W. BALLINGER, T. G. SCHURR, H. SOODYALL, T. JENKINS, S. T. SHERRY, and D. C. WALLACE. 1991. The structure of human mitochondrial DNA variation. J. Mol. Evol. **33**:543–555.

NACHMAN, M. W. 1998. Y-chromosome variation of mice and men. Mol. Biol. Evol. **15**:1744–1750.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

PEREZ-LEZAUN, A., F. CALAFELL, D. COMAS et al. (12 co-authors). 1999. Sex-specific migration patterns in central Asian populations, revealed by the analysis of Y-chromosome short tandem repeats and mtDNA. Am. J. Hum. Genet. **65**:208–219.

POLONI, E. S., G. PASSARINO, A. S. SANTACHIARA-BENERECETTI, O. SEMINO, A. LANGANEY, and L. EXCOFFIER. 1997. Human genetic affinities for Y chromosome p49a,f/Taq I haplotypes show strong correspondence with linguistics. Am. J. Hum. Genet. **61**:1015–1035.

PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN, and M. W. FELDMAN. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

QUINTANA-MURCI, L., O. SEMINO, E. MINCH, G. PASSARINO, A. BREGA, and A. S. SANTACHIARA-BENERECETTI. 1999. Further characteristics of proto-European Y chromosomes. Eur. J. Hum. Genet. **7**:603–608.

ROGERS, A. R. 1995. Genetic evidence for a Pleistocene population explosion. Evolution **49**:608–615.

ROGERS, A. R., A. E. FRALEY, M. J. BAMSHAD, W. S. WATKINS, and L. B. JORDE. 1996. Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. **13**:895–902.

ROGERS, A. R., and H. HARPENDING. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**:552–569.

ROGERS, A. R., and L. B. JORDE. 1995. Genetic evidence on modern human origins. Hum. Biol. **67**:1–36.

ROSSER, Z. H., T. ZERJAL, M. E. HURLES et al. (60 co-authors). 2000. Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. Am. J. Hum. Genet. **67**:1526–1543.

SAJANTILA, A., A. H. SALEM, P. SAVOLAINEN, K. BAUER, C. GIERIG, and S. PÄÄBO. 1996. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA **93**:12035–12039.

SCHNEIDER, S., D. ROESSLI, and L. EXCOFFIER. 2000. ARLEQUIN: a software for population genetics data analysis. Version 2.0. Department of Anthropology, University of Geneva, Switzerland.

SEIELSTAD, M., E. MINCH, and L. L. CAVALLI-SFORZA. 1998. Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20**:278–280.

SHEN, P., F. WANG, P. A. UNDERHILL et al. (13 co-authors). 2000. Population genetics implications from sequence variation in four Y chromosome genes. Proc. Natl. Acad. Sci. USA **97**:7354–7359.

SHERRY, S. T., A. R. ROGERS, H. C. HARPENDING, H. SOODYALL, T. JENKINS, and M. STONEKING. 1994. Pairwise differences of mtDNA reveal recent human population expansions. Hum. Biol. **66**:761–776.

SIMONI, L., F. CALAFELL, D. PETTENER, J. BERTRANPETIT, and G. BARBUJANI. 2000. Geographic patterns of mtDNA diversity in Europe. Am. J. Hum. Genet. **66**:262–278.

SLATKIN, M., and R. R. HUDSON. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. 3rd edition. Freeman, San Francisco.

STONEKING, M., S. SHERRY, and L. VIGILANT. 1992. Geographic origin of human mtDNA revisited. Syst. Biol. **41**:384–391.

TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

———. 1989*a*. Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. Genetics **123**: 585–595.

———. 1989*b*. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

TAKAHATA, N. 1995. A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. **26**:343–372.

THOMSON, R., J. K. PRITCHARD, P. SHEN, P. J. OEFNER, and M. W. FELDMAN. 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97**:7360–7365.

VOGT, P. H. 1997. Molecular basis of male (in)fertility. Int. J. Androl. **20**(Suppl. 3):2–10.

WISE, C. A., M. SRAML, D. C. RUBINSZTEIN, and S. EASTEAL. 1998. Comparative nuclear and mitochondrial genome diversity in humans and chimpanzees. Mol. Biol. Evol. **14**: 707–716.

WYCKOFF, G. J., W. WANG, and C. WU. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304–308.

ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK et al. (11 co-authors). 1998. Genetic structure of the ancestral population of modern humans. J. Mol. Evol. **47**:146–155.

JEFFREY LONG, reviewing editor

Accepted March 15, 2001

No Evidence of Population Growth in Y-chromosome Biallelic Variation

Isabelle Dupanloup[1], Luísa Pereira[2], Giorgio Bertorelle[1], Francesc Calafell[3], Maria João Prata[2], Antonio Amorim[2] and Guido Barbujani[1]

[1] Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, I-44100 Ferrara, Italy
[2] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias, s/n, 4200 Porto, and Faculdade de Ciências da Universidade do Porto, Praça Gomes Teixeira, 4050 Porto, Portugal
[3] Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Doctor Aiguader 80, 08003 Barcelona, Spain

Molecular genetic data contain information on the history of populations. Evidence of prehistoric demographic expansions have been detected in the mitochondrial diversity of most human populations, and in a Y-chromosome STR analysis, but not in the study of 11 Y-chromosome SNPs in Europeans. In this paper, we show that mismatch distributions and tests of mutation/drift equilibrium based on up to 166 Y-chromosome SNPs, in samples from all continents, also fail to support increase of the male effective population size. Computer simulations show that the low nuclear versus mitochondrial mutation rates cannot justify these results, but expansions may be more difficult to identify when only a few SNP sites are typed. However, the apparently constant effective size of the male population, observed over 46 world populations, is unlikely to simply reflect an ascertainment bias. The most plausible evolutionary scenarios that may account for these results, and for the recent expansion of male population size inferred from STR loci, include various combinations of demographic processes and selective forces.

Introduction

Patterns of DNA diversity in contemporary populations offer insight into the populations' past (von Haeseler et al., 1996; Cavalli-Sforza 1998; Bertranpetit 2000; De Knijff 2000). Processes such as migration, geographic dispersal, admixture, and changes in population size, leave recognizable marks at the DNA level. Episodes of rapid population growth and bottlenecks can be inferred from the distributions of pairwise sequence differences in a non-recombining DNA segment. Theory shows that these phenomena affects the shape of the gene genealogies (see e.g. Donnelly 1996). In particular, expansions result in star-like genealogies, and most mutations occurring on those genealogies do not tend to be shared among lineages. The resulting plots of differences between pairs of individuals, or mismatch distributions, are smooth and unimodal; their mode depends on the time passed since the expansion (Rogers and Harpending 1992). Most human mitochondrial mismatch distributions agree with expansion expectations, and the few exceptions have been explained as a result of demographic crises in hunting-gathering communities (Excoffier and Schneider 1999).

For the Y-chromosome the results are not equally straightforward. Two studies (Pritchard et al. 1999; Shen et al. 2000), concluded that Y-chromosome diversity is more compatible with exponentially-increasing than with constant population sizes, suggesting a recent (between 6,000 and 28,000 years ago) population growth. On the contrary, in a previous study, we found no evidence of growth in the mismatch distributions inferred from 11 Y-chromosome SNPs in Europe (Pereira et al. 2001). In agreement with what would be expected for a population of constant size, all distributions had multiple peaks, and statistics sensitive to demographic changes (Tajima's D and Fu's Fs) were insignificant. The limited number of polymorphic sites available might have reduced the sensitivity of the tests, although we observed that mitochondrial mismatch distributions maintain their unimodal shape even when a small subset of sites is considered (Pereira et al. 2001). We interpreted our results as reflecting either a combination of selective and demographic processes, or the fact that the effective population sizes of European males have really remained low until recently, while female population sizes increased sharply in prehistoric times. Three problems remained open, namely: (1) whether a greater number of SNPs could have led to differently shaped distributions; (2) whether the observed, non-expansion, pattern also occurs outside Europe; and (3) whether the existence of additional polymorphisms, neglected in our analysis but considered in studies based on complete DNA sequences (e.g. Shen et al. 2000), could have concealed an existing signal of expansion.

To address the first two questions, we analyzed two sets of data, comprising respectively 1007 Y chromosomes from Europe (Semino et al. 2000), and 1062 Y chromosomes from all continents (Underhill et al. 2000). Mismatch distributions were calculated, and test of mutation-drift equilibrium were carried out. As for the third question, we simulated the effects of factors such as different mutation rates and selection of polymorphic sites, in both stationary and expanding populations, and we compared the resulting mismatch distributions with those observed in the empirical analysis.

Materials and Methods

The data

We analyzed the datasets of Y-chromosome single-nucleotide polymorphisms published by Semino et al. (2000) (EU dataset) and Underhill et al. (2000) (WO dataset). They comprised respectively 1007 individuals from 25 European populations, typed at 22 polymorphic sites, and 1062 individuals from 21 populations of all continents, who either had been typed at 167 polymorphic sites, or whose genotype could be inferred with a high degree of confidence, assuming that all SNPs result from mutations that occurred only once in human evolution (Underhill et al. 2000). The only known violation of the assumption that no site mutated more than once is the M116 polymorphism, where three different alleles have been recorded. We chose to disregard that site, and therefore we considered as identical two groups of haplotypes that differ only by a substitution at M116, namely haplogroup 19 (which had been observed only once, in an African individual) and haplogroup 22. To evaluate the consequences of the pooling of different populations, in both cases we also jointly analyzed all chromosomes of the dataset. The European and Near Eastern samples of the WO dataset include part of the chromosomes of the Eu dataset. Therefore, the two datasets are not fully independent.

Mismatch distributions

Allele genealogies tend to have long deep branches in stationary populations, so that many mutations are shared by several individuals. Rapidly-expanding populations, conversely, will show long terminal branches in their gene trees (or *star-like* genealogies); the mutations occurring along those branches will tend to be unique to single individuals (Donnelly 1996). These different patterns of substitutions are reflected in the shape of the the distribution of pairwise differences between sequences, or mismatch distribution. Population subdivision (Marjoram and Donnelly 1994) and admixture (Bertorelle and Slatkin 1995) may act as confounding factors. As a rule, however, mismatch distributions are irregular in stationary or shrinking populations, whereas a smooth, unimodal shape is typical of expanding populations (Rogers and Harpending 1992; Rogers et al. 1996).

Mismatch distributions were estimated 48 times, namely for each of the 25 populations of the EU dataset, each of the 21 populations of the WO dataset, and for the two entire datasets. They were also estimated for a number of datasets generated by computer simulation, to represent a broad spectrum of evolutionary and demographic scenarios.

In all cases, mismatch distributions and gene diversity, i.e., the probability that two randomly sampled chromosomes differ from each other (Nei 1987), were estimated by ARLEQUIN 2.0 (Schneider et al. 2000), and the observed distribution of mismatches was fitted to expectations relative to an expanding population, by Monte-Carlo randomization. The null hypothesis was one of expansion, because there is no established expectation for the mismatch distribution in a stationary population (Harpending 1994). The age of the expansion, τ, was also estimated from the data

(Rogers and Jorde 1995) when the distribution was unimodal, and mutation-drift equilibrium (see below) could be rejected.

Tests of mutation-drift equilibrium

Departures from mutation-drift or mutation-selection equilibrium were tested, in each population and in the pooled samples, through Tajima's D and Fu's F_S. In Tajima's (1989a, b) test, a statistic D related with the parameter $\theta = 2N\mu$ (where N is the population size, and μ is the mutation rate) is independently estimated twice, from the number of polymorphic sites and from the average mismatch in the sample. Under equilibrium, the two D estimates should overlap. Differences between them may be caused by changes in the population size, or selection, or both. Fu's (1997) Fs statistic compares the observed number of alleles in a sample with the number of alleles expected in a population of constant size. Both D and Fs take negative values when the population expands, and positive values when it shrinks. Different selective regimes may affect the shape of the underlying gene tree, and hence mimic the effects of changes in population size.

The significance of D and Fs was tested by randomization. By the coalescent simulation program implemented in the ARLEQUIN package (Schneider et al. 2000), we repeatedly generated random samples from hypothetical stationary populations whose parameter θ was equal to the average number of pairwise differences observed in the sample of interest. In this way, empirical null distributions of the relevant statistics were generated by repeating the randomization procedure 1,000 times, and calculating in each case the values of D and F_S. It was straightforward to obtain empirical estimates of the probability of the observed D and F_S values from these distributions, under the hypothesis of neutrality and constant population size.

Monte-Carlo simulations

While complete sequences of hundreds or thousands base pairs are analysed in mitochondrial studies, and in Shen et al.'s (2000) Y-chromosome study, in most SNP analyses only sites known in advance to be variable are typed. In this way, private polymorphisms, and poorly polymorphic sites, of the Y chromosome are likely to be neglected, possibly affecting the inferred mismatch distributions. A second difference between mitochondrial and nuclear data lies in their different mutation rates. Biallelic Y-chromosome markers mutate slowly; estimated mutation rates per site and per year, μ, range between 2.5×10^{-8} (Hammer, 1995; Jobling et al. 1997) and 1.2×10^{-9} (Thomson et al. 2000). On the contrary, for the hypervariable region of the mitochondrial genome μ ranges from 4.6×10^{-7} (Soodyall et al. 1997; Jazin et al. 1998) to 3.2×10^{-6} (Sigurgardottir et al. 2000). It is conceivable that low rates of mutation may not allow Y-chromosome SNPs to reveal recent population growth.

To evaluate the effects of both ascertainment bias and different mutation rates on the possibility to detect an expansion, we generated samples from stationary and expanding populations by Monte Carlo simulation. The simulation algorithm was based on the coalescent process with superimposed mutations, as described by Hudson (1990). Each sample was obtained by first generating its genealogy. Mutations were then

randomly placed on the genealogy, assuming they occur according to a uniform and constant Poisson process. More details are in Pereira et al. (2001).

For four different mutation rates (from 5 x 10^{-9} to 1 x 10^{-7} per site and per year), we simulated 1,000 samples of chromosomes under the assumption of a large and constant population size (5,000 haploid individuals) from a single panmictic deme. Eighty chromosomes were sampled at random from each population, each characterized by 1,000 potentially variable SNP sites. For the same mutation rates we also simulated exponential population expansions using the same coalescent approach. The simulated expansion started 50,000 years (or 2,500 generations) ago, with a 100-fold increase of population size and a final effective size N_0= 100,000. Depending on mutation rates and population genealogies, variable numbers of these sites mutated, but the observed number of polymorphic sites never exceeded 199 in our simulations. Different rounds of analysis took into account either all sites, or only the 11 most variable ones, thus testing for the effect of the ascertainment bias.

In each of the simulated samples, Tajima's D and Fu's F_S statistics were estimated. The fit of the observed distribution of mismatches to a model of population expansion was tested by the bootstrap approach implemented in ARLEQUIN (Schneider et al. 2000). The statistic *SSD*, a sum of squared deviations from expansion expectations, was estimated and its empirical probability was computed by performing sets of 100 coalescent simulations of stepwise expansions. Finally, we defined three basic shapes of the mismatch distribution, namely unimodal with a maximum at 0 (Type 0), unimodal with a maximum > 0 (Type 1), and bi- or multi-modal (Type 2), and counted the number of occurrences of each type in each set of 1,000 simulations.

Results

Mismatch distributions

A ragged pattern, either bi- or tri-modal, is observed in the analysis of all populations in the EU dataset (Figure 1). Despite considering twice as many polymorphic sites as in the previous study of the same continent (Pereira et al. 2001), mismatch distributions with a peak at 0 differences are still common. Predictably, by doubling the number of sites considered, the average distance between modes increases. For instance, in the Iberian samples, mismatch distributions based on 11 sites (Pereira et al. 2001) displayed peaks at 0, 3 and 5 differences whereas, in this study, the peaks are located at 0, 4 and 7 differences.

However, by doubling the number of polymorphic sites analysed (Table 1; compare with Table 1 in Pereira et al. 2001), the number of different haplogroups did not increase much, from 2-8 per population using 11 SNPs to 3-13 using 22 SNPs. Accordingly, gene diversities were similar in the two studies, in agreement with Semino et al.'s (2000) observation that more than 95% of the chromosomes they typed could be assigned to clades of haplotypes defined by just 10 key mutations. Tajima's and Fu's statistics were insignificant, except for a negative *Fs* for Turks, who also showed a rather smooth distribution. However, this result was no longer significant after Bonferroni's correction for multiple tests (Sokal and Rohlf 1995).

Even when 166 biallelic markers were studied (WO dataset) most distributions were multimodal (Figure 2). The sub-Saharan samples were the ones that displayed the most irregular distributions, with peaks at 16 (Sudan), 17 (Ethiopia) or 18 differences (Khoisan), confirming extensive divergence of African Y chromosomes. The mismatches of the only hunting-gathering population (Khoisan) did not appear qualitatively different from those of the other, farming, populations, despite the significant differences between them at the mitochondrial level (Excoffier and Schneider 1999). In the European samples there were minor differences between the results of the analysis of 22 (EU dataset) and 166 (WO dataset) sites, both in terms of the shape of the mismatch distributions, and of the related statistics (Table 2). Sardinia shows a more irregular shape in the WO dataset, but that might reflect the small sample size, 22, in the study by Underhill et al. (2000), presumably a subset of the 77 individuals described by Semino et al. (2000).

Unimodal mismatch distributions were observed in three Central and Eastern Asian populations. Fu's *Fs* appeared negative and significant (as is the case for mitochondrial data) in these samples, and Tajima's *D* in one of them, but these significances did not stand Bonferroni's correction. A unimodal distribution was also observed in the American sample, but the peak is at 0 differences, reflecting the fact that 78% of the Y chromosomes belong to haplogroup 115, a haplogroup not found in other continents (Underhill et al. 2000). Taken at their face value, the negative Tajima's *D* and Fu's *Fs* (once again, both insignificant after Bonferroni's correction) would point to an Amerindian expansion. However, because the peak is at 0, the estimated effective population size is the same, before and after the expansion (Rogers and Harpending 1992) (see legend to Table 2). We do not know how well the sample considered represents the whole continent. However, based on the evidence available, it seems that Y chromosome diversity in this American sample reflects a severe bottleneck (Bonatto and Salzano 1997), with most Y-chromosome variation presumably restricted to STR sites (Ruiz-Linares et al. 1999).

To understand the effects of aggregation of individuals from distant populations, we ran two global analyses of the EU and WO datasets. For the European and Near-Eastern populations of the EU dataset, we observed a trimodal distribution, and insignificant, positive Tajima's *D* and Fu's *Fs* (Table 1 and Figure 3). For the more heterogeneous set of populations in the WO dataset, the mismatch distribution was still bimodal, but Tajima's *D* and Fu's *Fs* statistics were negative, and the former kept significance (P=0.044) even after Bonferroni's correction (Table 2 and Figure 3). If one picks up a few chromosomes from heterogeneous populations, many substitutions are likely to appear lineage-specific, although a broader sampling would show they are not. Our result suggests that this may leads to a mismatch distribution that looks similar to those resulting from expansions, and to values of *D* and *Fs* compatible with an expansion, even when expansions are not supported in any single population.

Effects of the mutation rates

In simulated stationary populations, when considering the whole set of sites, the average mismatch is close to the expected value, i.e. the parameter θ used to generate the simulated samples, and thus increased with the mutation rate in the different sets of

simulations (Table 3). As expected, in expanding populations both the average mismatch and its standard deviation are reduced.

For both stationary and expanding populations, multimodality is more frequent as the mutation rate increases, which does not support the view that the low Y-chromosome mutation rate increases the probability of observing multimodal mismatch distributions. Under stationarity, for example, for $\mu = 5 \times 10^{-9}$ per site and per year, more than 60% of the mismatch distributions have a single peak (types 0 and 1), but when μ is 10^{-7} all but one simulations yield distributions with multiple peaks (type 2). In expanding populations, multimodality is rare, especially at low mutation rates. When $\mu < 10^{-8}$, more than 90% of the mismatch distributions are bell-shaped or show a geometric form. Therefore, the low Y-chromosome mutation rate is not expected to affect much the shape of the mismatch distribution; our results show that, if anything, it may only enhance an existing signal of expansion.

In the simulated stationary populations, both D and F_S show wide distributions centered on 0, regardless of mutation rates. In all the simulated cases of stationarity, less than 7% of the D values are significant at the 95% level. On the contrary, after expansions D and F_S are always negative and in more than 78% of the cases significant. The absolute values of this statistics, and the number of times they reach significance, increase with the mutation rate, and their standard deviations decrease.

Effects of the selection of the most polymorphic sites

Several results change when only the11 most polymorphic RFLP sites are taken into account in the computation of the different statistics (Table 3). In stationary populations, the same fraction (0 to 3%) of 5% significant Tajima's tests is observed, both analyzing 11 sites and all of them. In expanding populations and when mutation rates are high, conversely, multimodal distributions become more frequent, D and F_S often take insignificant or even positive values, and the number of times they reach significance is reduced. That was to be expected, because higher mutation rates result in higher numbers of polymorphic sites, and therefore to a greater loss of information when all sites but 11 are neglected. Consider for instance the simulations with $\mu > 3 \times 10^{-8}$. More than 99% of D values are significant in expanding populations when considering the totality of sites, but when considering only the 11 most polymorphic sites, less than 12% of the D values reach significance.

Discussion

The results of this and of a previous study (Pereira et al. 2001), are not consistent with a rapid growth of the males' population size. Despite the number of polymorphic sites being now high in the WO database, nearly all mismatch distributions are multimodal, and there is no statistical support for departures from mutation-drift equilibrium. Patterns of Y-chromosome diversity incompatible with population growth are observed in all continents, and particularly in Africa.

Simulations based on a coalescent model show that the typing of sites that are already known to be polymorphic may reduce the probability to identify expansions.

However, that effect seems weak at the low mutation rates typical of Y-chromosome SNPs. In expanding populations, selection of 11 sites reduces the fraction of significant Tajima's tests, but only for mutation rates $\geq 5 \times 10^{-8}$. Current estimates of mutation rates for Y-chromosome SNPs are lower than 3×10^{-8} (Jobling et al. 1997; Thomson et al. 2000), and in this study we considered up to 166 sites. Also, mitochondrial mismatch distributions based on RFLPs, i.e. based on sites known to be polymorphic, are unimodal (Sherry et al. 1994). All these facts lead us to conclude that selection of sites may have affected only slightly, if at all, the mismatch distributions inferred from Y chromosome SNPs.

Another potential confounding factor is admixture (Bandelt and Forster 1997). Hybrid populations may fail to show signatures of past expansions, if gene flow between parental groups is low and the contact is recent (Marjoram and Donnelly 1994). There is empirical evidence of this phenomenon in a bimodal mitochondrial mismatch distribution described in North-West Africa, probably reflecting the presence of genes from the Subsaharan African and Mediterranean gene pools (Brakez et al 2001). Such bimodal distributions are likely to be evident only if the groups that hybridized were genetically differentiated. Because populations of the same continent tend to be more similar at the mitochondrial than at the Y-chromosome level (Seielstad et al. 1998), the potential impact of admixture might be greater on the mismatch distributions inferred from the latter. At this stage, however, the available data are too scanty for us to understand to what extent the differences between maternally- and paternally-transmitted genes are due to undetected admixture.

In addition, and in agreement with previous findings (Rogers et al. 1996), our simulations do not suggest that low mutation rates reduce the probability to detect an expansion once it occurred, although very recent phenomena are clearly unlikely to be recognized using slow-mutating markers, such as nuclear SNPs. In summary, it seems safe to conclude that Y-chromosome biallelic variation does not provide evidence of population growth.

There are doubtless more humans now than in the Pleistocene (see e.g. Zietkiewicz et al. 1998; Harpending et al. 1998), and so constant effective population sizes of males are counterintuitive. In fact, however, even studies concluding that the male population size did increase proposed much later expansions than studies of mitochondrial DNA (38,000 years ago or more: Sherry et al. 1994; Excoffier and Schneider 1999). Pritchard et al. (1999) analyzed variation at 8 microsatellite loci in 445 Y chromosomes, and found it better fitted a model of demographic growth, presumably in the last 6,000 years, than one of constant population sizes. Shen et al. (2000) reported evidence for a population expansion at 4 loci, sequenced in 53-72 Y-chromosomes. A Tajima's test and the unimodal mismatch distribution suggested an expansion, about 28,000 years ago. Therefore, if both females and males increased in numbers, they did not do that at the same time, which is not much easier to explain than absence of detectable growth in the Y chromosomes. Variation in other genome regions does not help to clarify the picture. Some autosomal loci seem to support population growth (Zhao et al. 2000; Alonso and Armour 2001), but other autosomal (Takahata et al. 1995; Harding et al. 1997) and X-linked (Harris and Hey 1999a) loci do not.

Apparent differences between the inferred demographic histories of males and females may reflect of selection, affecting either the mitochondrial genome (Excoffier

1990; Harris and Hey 1999b), the Y chromosome (Jobling and Tyler-Smith 2000), or both. A mitochondrial selective sweep may have led investigators to erroneously reject the hypothesis of constant female population size; stabilizing selection affecting the Y chromosome may have determined patterns compatible with constant population sizes, when those sizes, in fact, increased (Tajima 1989a; see also Pereira et al. 2001). The effects of selective pressures and demographic changes cannot be discriminated *a posteriori* from population-genetics data (see e.g. Takahata 1996), and hence the present study does not provide evidence relevant to this question. However, unless those selective pressures have really been strong (which would force us to reconsider most aspects of human evolution inferred from DNA evidence) the available data suggest at least that the human demographic past cannot be envisaged as a simple process of population growth, to which females and males contributed equally and in parallel.

At this stage, therefore, it seems that one should take seriously the idea that human male and female population sizes had a different dynamics, with several independent early female expansions, and low and constant male population sizes for much of our past. Widespread polygyny, and therefore sexual selection, are two necessary implications of this view. Note that in Shen et al.'s (2000) paper, male population growth was supported by a negative Tajima's test, and by the distribution of the number of mutants at independent sites fitting a model of population growth. For these calculations, however, 53-72 individuals from 46 different origins had been considered, which violates the assumptions of Tajima's (1989a, p. 593) test. In fact, our WO dataset also gave a significant Tajima's D in the global analysis (Table 2), although each population appeared stationary when individually analyzed. In Figure 4 we outline how a pattern of nucleotide substitutions resembling that caused by an expansion can be generated by sampling Y chromosomes from different non-expanding populations. As previously mentioned, in the starlike genealogies resulting from expansions, few mutations are shared among lineages (Figure 3A). However, if only few chromosomes are sampled in each stationary (Figure 3B) population, chances are that very few substitutions will be shared, leading one to reject constant population size (Figure 3C). These considerations do not apply to the study by Pritchard et al. (1999), which is based on a different, Bayesian, approach, and on larger population samples.

The composition of the sample considered may thus account, at least in part, for the differences between Shen at al.'s (2000) results and ours. Be that as it may, Pritchard et al. (2000) locate the expansion of male population size in a recent past, 6,000 years ago, and Shen et al.'s (2000) results are compatible with that date, although they suggest an older timing as more probable. In mutational terms, very little must have happened to the SNP sites we considered in that lapse of time. The Y-chromosome evidence available may thus be reconciled by saying that over much of human history the effective male populations were generally small and constant, and they grew to the current size only in the last few thousand years.

This paper was supported by grants of the Italian Ministry of Universities and Research (MIUR) and of the University of Ferrara. LP was supported by a PhD grant (PRAXIS XXI BD/13632/97) from Fundação para a Ciência e a Tecnologia and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

References

Alonso S, Armour JA (2001) A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc Natl Acad Sci USA 98:864-869

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

Bertorelle G, Slatkin M (1995) The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. Mol Biol Evol 12:887-892

Bertranpetit J (2000) Genome, diversity, and origins: The Y chromosome as a storyteller. Proc Natl Acad Sci USA 97:6927-6929.

Bonatto SL, Salzano FM (1997) A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA 94:1866-1871

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

Cavalli-Sforza LL (1998) The DNA revolution in population genetics. Trends Genet 14: 60-65.

De Knijff P (2000) Messages through bottlenecks: On the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet 67:1055-1061.

Donnelly P (1996) Interpreting genetic variability: The effects of shared evolutionary history. Pp. 25-50 *in* K. Weiss, ed. Variation in the human genome. Wiley, Chichester, UK (CIBA Foundation Symposium).

Excoffier L (1990) Evolution of human mitochondrial DNA: Evidence for departure from a pure neutral model of populations at equilibrium. J Mol Evol 30:125-139.

Excoffier L, Schneider S (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc Natl Acad Sci USA 96:10597-10602.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915-925

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376-378

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772-789.

Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66:591-600.

Harpending HC, Batzer MA, Gurven M, Jorde JB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Natl Acad Sci USA 95:1961-1967

Harris EE, Hey J (1999a) X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96:3320-3324

Harris EE, Hey J (1999b) Human demography in the Pleistocene: Do mitochondrial and nuclear genes tell the same story? Evol Anthropol 8:81-86.

Hudson RR (1990) Gene genealogies and the coalescent process. Pp.1-44. *in* D Futuyma, J Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

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

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

Jobling MA, Tyler-Smith C (2000) New uses for new haplotypes. The human Y chromosome, disease and selection. Trends Genet 16:356-362

Marjoram P, Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136:673-683

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York, NY, USA

Pereira L, Dupanloup I, Rosser Z, Jobling MA, Barbujani G (2001) Y-chromosome mismatch distributions in Europe. Mol Biol Evol 18:1259-1271

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Mol Biol Evol 16:1791-1798.

Rogers AR, Fraley AE, Bamshad MJ, Watkins WS, Jorde LB (1996) Mitochondrial mismatch analysis is insensitive to the mutational process. Mol Biol Evol 13:895-902.

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552-569

Rogers AR, Jorde LB (1995) Genetic evidence on modern human origins. Hum Biol 67:1-36.

Ruiz-Linares A, Ortiz-Barrientos D, Figueroa M, Mesa N, Munera JG, Bedoya G, Velez ID, et al. (1999) Microsatellites provide evidence for Y-chromosome diversity among the founders of the New World. Proc Natl Acad Sci USA 96:6312-6317

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

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

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, et al. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. Science 290:1155-1159.

Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, et al (2000) Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 97:7354-7359

Sherry ST, Rogers AR, Harpending HC, Soodyall H, Jenkins T, Stoneking M (1994) Pairwise differences of mtDNA reveal recent human population expansions. Hum Biol 66:761-776

Sigurgardottir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. Am J Hum Genet 66:1599-609

Sokal RR, Rohlf FJ (1995) Biometry. W. H. Freeman and Company, New York, NY, USA

Soodyall H, Jenkins T, Mukherjee A, du Toit E, Roberts DF, Stoneking M (1997) The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. Am J Phys Anthropol 104:157-166

Tajima F (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. Genetics 123:585-595.

Tajima F (1989b) The effect of change in population size on DNA polymorphism. Genetics 123:597-601

Takahata N (1996) Neutral theory of molecular evolution. Curr Opin Genet Devel 6:767-772

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Pop Biol 48:198-221

Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc Natl Acad Sci USA 97:7360-7365.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, et al. (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361.

von Haeseler A, Sajantila A, Pääbo S (1996) The genetical archaeology of the human genome. Nat Genet 14:135-140

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, et al. (2000) Worldwide DANN sequence variation in a 10-kilobase noncoding region of human chromosome 22. Proc Natl Acad Sci USA 97:11354-11358

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, et al. (1998) Genetic structure of the ancestral population of modern humans. J Mol Evol 47:146-155

Table 1

Measures of genetic diversity estimated from the EU dataset (Semino et al. 2000).

Population	*N*	*Hp*	Average mismatch	P(exp)	*H*	*D*	P (D)	*Fs*	P (*Fs*)	τ
Andalusia	29	7	2.49 ± 1.38	0.080	0.567	-0.363	0.420	0.182	0.581	-[a]
Basque-Spanish	45	5	1.00 ± 0.69	0.090	0.211	-1.471	0.055	-0.221	0.450	-[a]
Basque-French	22	3	1.18 ± 0.79	0.040	0.255	-0.882	0.218	1.730	0.821	-[a]
Catalan	24	5	1.69 ± 1.03	0.070	0.377	-1.228	0.107	0.502	0.657	-[a]
French	23	6	2.75 ± 1.51	0.030	0.700	-0.336	0.681	0.923	0.682	-[a]
Dutch	27	4	1.93 ± 1.13	0.050	0.470	-0.558	0.324	2.147	0.855	-[a]
German	16	4	2.70 ± 1.51	0.060	0.642	-0.017	0.529	2.309	0.887	-[a]
Czech+Slovakian	45	9	2.63 ± 1.43	0.120	0.784	-0.550	0.336	-0.330	0.464	-[a]
Centr. North Italian	50	6	2.19 ± 1.23	0.030	0.589	-0.313	0.442	1.476	0.789	-[a]
Calabrian	37	9	3.11 ± 1.65	0.053	0.829	-0.191	0.844	-0.104	0.517	-[a]
Sardinian	77	11	3.27 ± 1.70	0.040	0.801	-0.404	0.658	0.224	0.618	-[a]
Croatian	58	7	3.05 ± 1.61	0.020	0.707	0.804	0.817	2.224	0.837	-[a]
Albanian	51	8	3.57 ± 1.85	0.010	0.833	1.339	0.915	1.844	0.826	-[a]
Greek	76	11	3.75 ± 1.91	0.020	0.818	0.893	0.839	0.807	0.691	-[a]
Macedonian	20	6	3.72 ± 1.96	0.010	0.821	1.120	0.880	1.597	0.810	-[a]
Polish	55	4	2.37 ± 1.31	0.080	0.609	0.558	0.746	4.253	0.945	-[a]
Hungarian	45	7	2.67 ± 1.45	0.180	0.614	-0.085	0.539	1.196	0.742	-[a]
Ukrainian	50	10	3.08 ± 1.63	0.000	0.678	-0.243	0.447	-0.211	0.521	-[a]
Georgian	63	9	2.43 ± 1.34	0.050	0.778	0.732	0.803	-0.058	0.554	-[a]
Turkish	30	13	3.15 ± 1.68	0.310	0.828	-0.567	0.310	-3.969	**0.042**	4.519
Lebanese	31	9	3.36 ± 1.77	0.008	0.832	0.106	0.588	-0.211	0.503	-[a]
Syrian	20	8	3.04 ± 1.65	0.190	0.874	-0.067	0.517	-0.808	0.348	-[a]
Saami	24	4	2.26 ± 1.29	0.000	0.667	1.231	0.884	2.504	0.889	-[a]
Udmurt	43	8	3.12 ± 1.65	0.080	0.776	0.681	0.786	0.933	0.703	-[a]
Mari	46	6	1.88 ± 1.09	0.350	0.558	-0.313	0.722	0.807	0.729	-[a]
EU	1007	19	3.29 ± 1.70	0.050	0.847	0.544	0.748	0.304	0.644	-[a]

Note.- *N*: sample size; *Hp*: number of different haplogroups observed; P(exp): probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time τ estimated from the data; *H*: Gene diversity; *D*: Tajima's *D*; *Fs*: Fu's *Fs*; τ time since expansion; [a] not calculated due to absence of signal of expansion; Initial and final θ in the Turkish population, the only one showing evidence of expansion: 0.003 – 8.115

Table 2

Measures of genetic diversity estimated from the WO dataset (Underhill et al. 2000).

Population	*N*	*Hp*	Average mismatch	P(exp)	*H*	*D*	P (D)	*Fs*	P (*Fs*)	τ
Sudan	40	9	9.05 ± 4.25	0.000	0.776	1.283	0.912	5.673	0.952	-[a]
Ethiopia	88	15	6.62 ± 3.16	0.930	0.876	-0.053	0.551	1.791	0.775	-[a]
Mali	44	11	4.76 ± 2.37	0.030	0.814	-0.880	0.184	0.605	0.634	-[a]
Morocco	28	8	3.36 ± 1.78	0.470	0.722	-0.799	0.235	0.313	0.608	-[a]
C.Africa	36	9	5.21 ± 2.58	0.020	0.660	-0.343	0.417	1.869	0.771	-[a]
Khoisan	39	6	8.86 ± 4.17	0.000	0.802	1.692	0.961	9.876	0.993	-[a]
S.Africa	54	9	5.22 ± 2.56	0.210	0.697	-0.276	0.472	3.115	0.870	-[a]
Europe	60	12	3.05 ± 1.61	0.300	0.731	-1.201	0.102	-1.174	0.342	-[a]
Sardinia	22	7	5.75 ± 2.86	0.030	0.727	-0.959	0.164	2.513	0.877	-[a]
Basque	45	8	1.89 ± 1.10	0.010	0.638	-1.523	**0.041**	-0.779	0.358	0.882
Mid-east	24	11	5.73 ± 2.84	0.680	0.877	-0.886	0.200	-0.505	0.422	-[a]
C.Asia+Siberia	185	33	5.11 ± 2.49	0.940	0.936	-1.220	0.070	-8.009	**0.047**	5.536
Pakistan+India	88	24	5.82 ± 2.81	0.600	0.877	-1.203	0.088	-3.811	0.157	-[a]
Hunza	38	13	4.42 ± 2.23	0.280	0.875	-0.529	0.351	-1.397	0.319	-[a]
Japan	23	12	6.28 ± 3.09	0.530	0.893	0.202	0.635	-1.090	0.303	-[a]
China	20	12	3.41 ± 1.82	0.850	0.937	-1.233	0.105	-4.507	**0.010**	3.749
Taiwan	74	5	1.37 ± 0.86	0.090	0.554	-0.110	0.483	1.213	0.765	-[a]
Cambo+Laos	18	12	4.80 ± 2.46	0.280	0.895	-1.728	0.027	-3.496	**0.043**	6.101
NewGuinea	23	6	3.46 ± 1.83	0.240	0.826	0.937	0.857	1.689	0.809	-[a]
Australia	7	4	4.48 ± 2.51	0.030	0.810	1.161	0.883	1.486	0.760	-[a]
America	106	12	1.38 ± 0.86	0.360	0.384	-1.766	0.015	-4.082	**0.049**	3.027
WO	1062	115	7.33 ± 3.43	0.730	0.964	-1.895	**0.002**	-23.937	**0.007**	6.230

Note.- *N*: sample size; *Hp*: number of different haplogroups observed; P(exp): probability (estimated by bootstrap) of the observed mismatch distribution under the hypothesis of population expansion for a time τ estimated from the data; *H*: Gene diversity; *D*: Tajima's *D*; *Fs*: Fu's *Fs*; τ time since expansion; [a] not calculated, owing to absence of signals of expansion. Initial and final θ in the populations showing evidence of expansion are: C.Asia+Siberia: 0.567 – 16.595; China: 0.170 – 16.432; Cambodia + Laos: 0.034 – 13.963; for the Americas, the peak is at 0 differences, and so the estimated θ is the same, 0.560, both before and after the expansion.

Table 3

Simulation results

Simulated population	μ	N of sites	Average mismatch[a]	MM0[c]	MM1[c]	MM2[c]	SSD Signif[d]	*D*[e]	*Signif*[f]	F_S[g]	*Signif.*[f]
Stationary	5×10^{-9}	1000	1.005 (0.778)	0.429	0.185	0.386	0.356	-0.027 (0.949) [-1.939,3.282]	0.065	-0.042 (2.203) [-6.871,12.659]	0.120
	5×10^{-9}	11	1.000 (0.757)	0.430	0.184	0.385	0.372	-0.022 (0.958) [-1.939,3.913]	0.063	-0.019 (2.304) [-6.871,18.442]	0.119
	1×10^{-8}	1000	2.064 (1.243)	0.132	0.211	0.657	0.311	-0.022 (0.959) [-2.016,3.277]	0.044	-0.048 (2.728) [-9.404,14.086]	0.081
	1×10^{-8}	11	1.931 (1.035)	0.134	0.210	0.656	0.331	0.128 (1.090) [-1.933,4.027]	0.047	0.489 (3.311) [-9.404,18.692]	0.067
	5×10^{-8}	1000	9.715 (4.904)	0.000	0.027	0.972	0.269	-0.073 (0.892) [-2.175,3.010]	0.046	-0.526 (4.103) [-16.896,20.196]	0.088
	5×10^{-8}	11	4.183 (0.990)	0.107	0.022	0.871	0.577	2.382 (1.201) [-1.118,4.065]	0.000	8.184 (5.112) [-3.451,18.774]	0.000
	1×10^{-7}	1000	19.280 (9.536)	0.000	0.001	0.999	0.243	-0.043 (0.931) [-2.248,3.248]	0.042	-0.843 (4.747) [-23.947,23.828]	0.073
	1×10^{-7}	11	4.801 (0.644)	0.138	0.000	0.862	0.727	3.132 (0.782) [0.239,4.065]	0.000	9.908 (4.477) [-0.074,18.774]	0.000
Expanding	5×10^{-9}	1000	0.522 (0.312)	0.790	0.172	0.038	0.197	-1.740 (0.385) [-2.462,-0.114]	0.788	-7.741 (3.588) [-20.805,0.350]	0.926
	5×10^{-9}	11	0.507 (0.295)	0.801	0.161	0.038	0.197	-1.720 (0.386) [-2.362,-0.114]	0.794	-7.384 (3.219) [-18.252,0.350]	0.914
	1×10^{-8}	1000	1.045 (0.442)	0.436	0.499	0.065	0.164	-2.007 (0.317) [-2.595,-0.480]	0.939	-14.555 (5.449) [-1.362,0.000]	0.994
	1×10^{-8}	11	0.823 (0.368)	0.588	0.328	0.084	0.214	-1.686 (0.442) [-2.332,0.219]	0.768	-8.297 (3.421) [-17.423,0.229]	0.917
	5×10^{-8}	1000	5.187 (1.191)	0.000	0.825	0.175	0.049	-2.286 (0.195) [-2.673,-1.349]	0.999	-25.524 (0.371) [-26.808,-24.407]	1.000
	5×10^{-8}	11	2.491 (0.765)	0.021	0.388	0.591	0.189	0.328 (0.928) [-1.670,3.872]	0.005	-0.475 (2.855) [-11.493,14.692]	0.077
	1×10^{-7}	1000	10.402 (1.936)	0.000	0.634	0.366	0.028	-2.289 (0.176) [-2.674,-1.432]	1.000	-24.530 (0.190) [-25.209,-24.059]	1.000
	1×10^{-7}	11	3.523 (0.832)	0.014	0.133	0.853	0.248	1.581 (1.010) [-1.134,4.064]	0.000	2.344 (3.719) [-9.094,18.692]	0.007

[a] Mean of the mismatch distribution, and standard deviation, across 1,000 simulations.
[b] Mean raggedness of the mismatch distribution, and standard deviation, across 1,000 simulations.
[c] Fraction of the 1000 simulations showing the shapes of the mismatch distribution: MM0 (unimodal with a maximum at 0), MM1 (unimodal with a maximum > 0), and MM2 (bimodal/multimodal).
[d] Fraction of the cases where the P value of the *SSD* statistic is lower than 5%
[e] Average *D* . Standard deviation is within parentheses and the range of observed values is within brackets.
[f] Fraction of significant cases at the 5% level, out of 1,000 simulations.
[g] Average F_S (Fu 1997) over 1,000 simulations. Standard deviation is shown within parentheses and the range of observed values is within brackets.

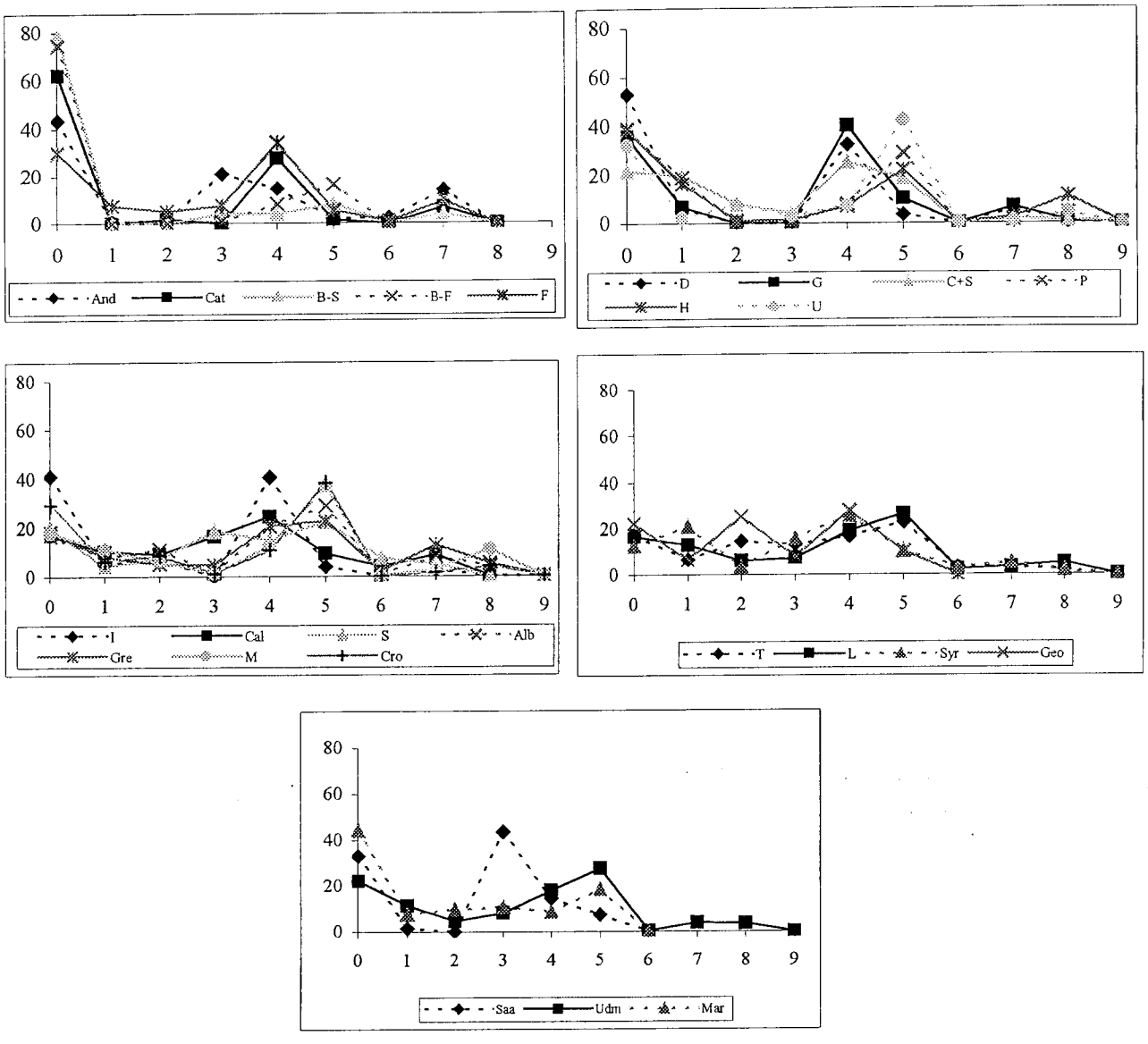


Figure 1 Mismatch distributions in the EU dataset (Semino et al. 2000). The X axis represents the no. of pairwise differences and the Y axis the frequency (%).

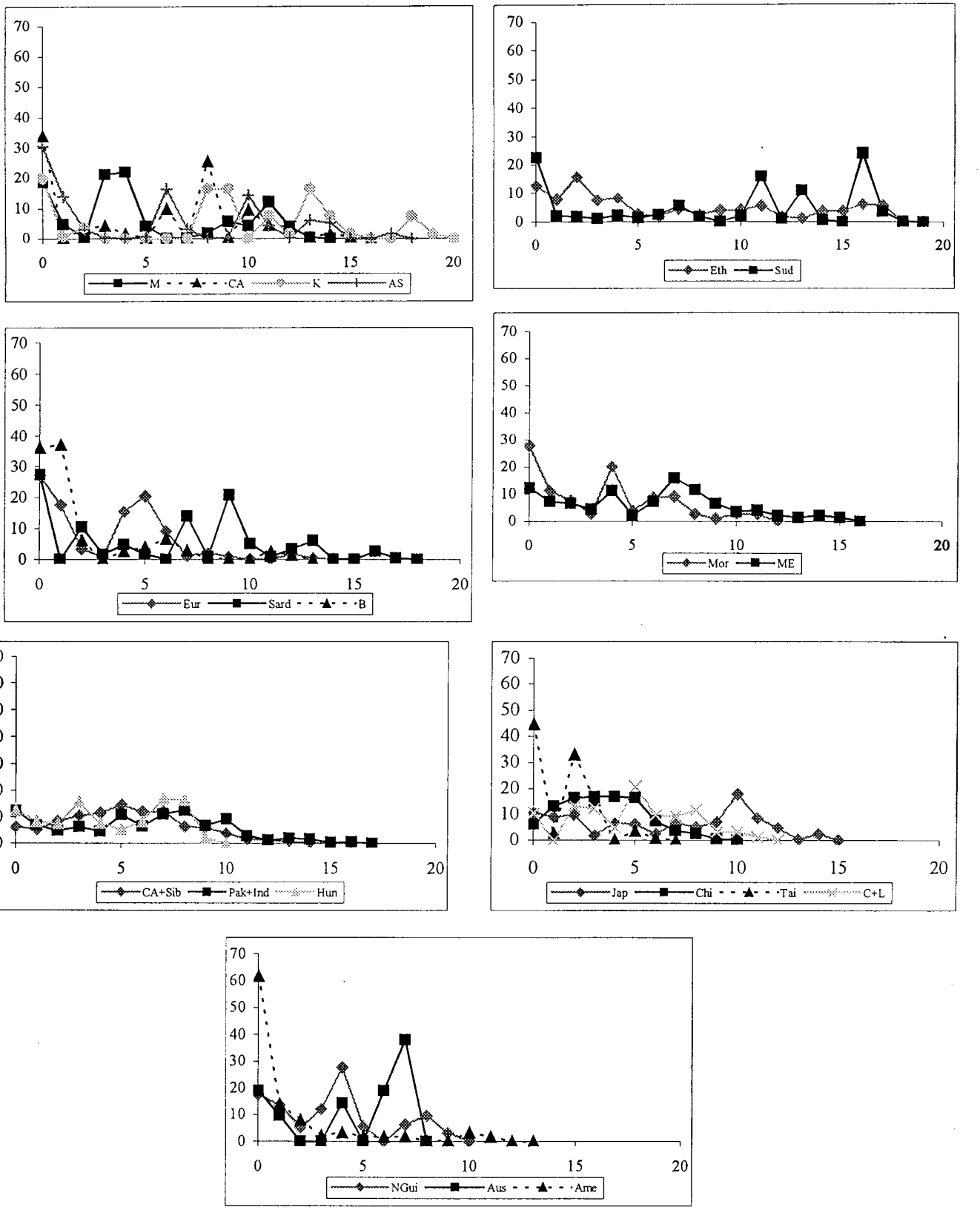


Figure 2 Mismatch distributions in the WO dataset (Underhill et al., 2000).

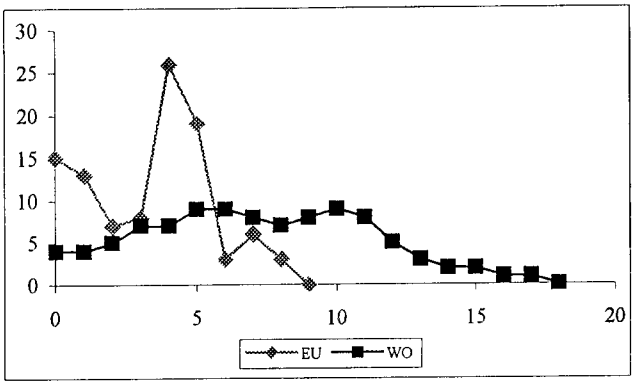
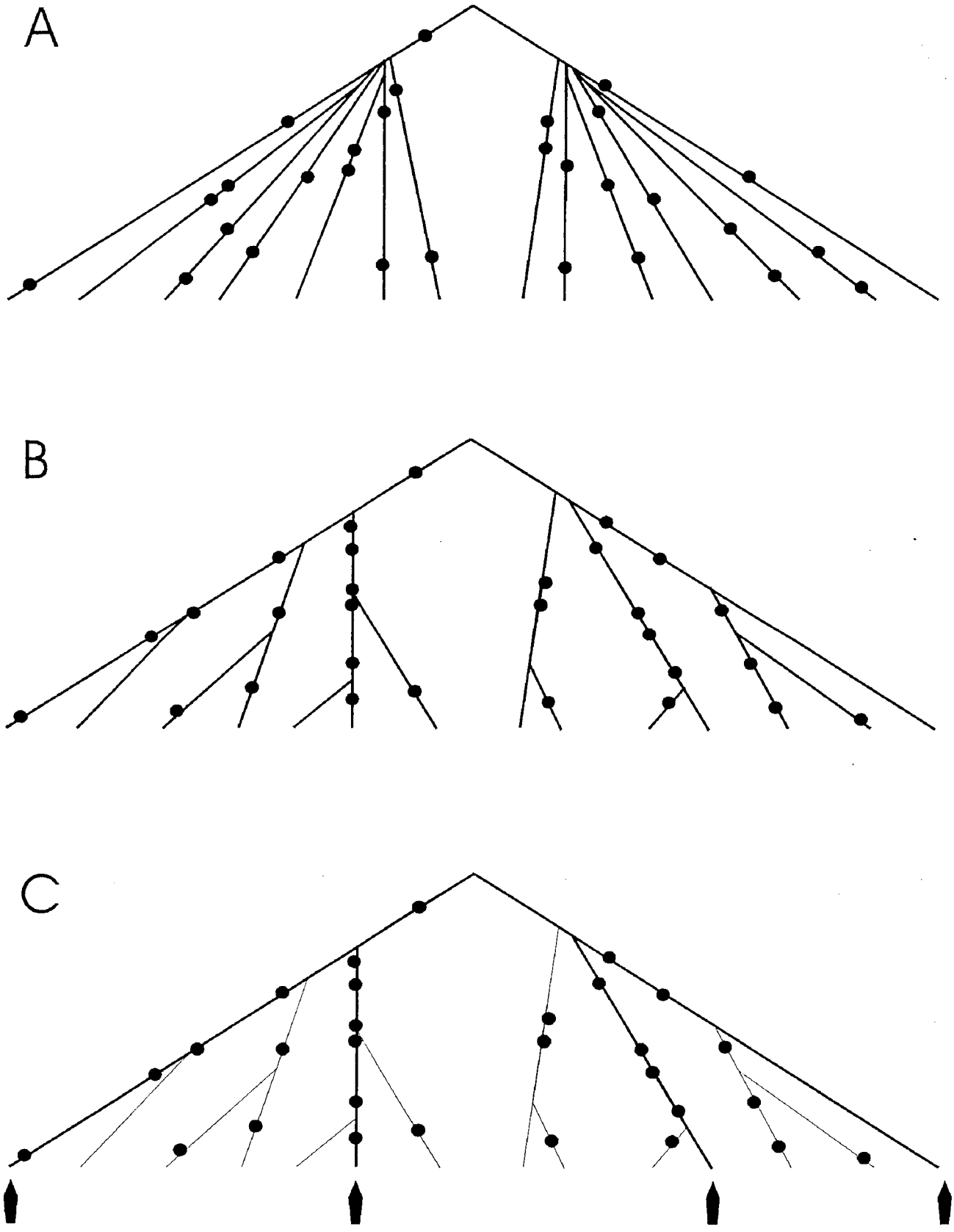


Figure 3 Mismatch distributions for the pooling of all populations in the EU and WO datasets.

Figure 4. Nucleotide substitutions (black dots) in: two expanding populations (a), two stationary populations (b), and two stationary populations from which a few individuals (marked by arrows) are sampled (c).



An evaluation of the proportion of identical Y-STR haplotypes due to recurrent mutation

L. Pereira[a,b], M.J. Prata[a,b] and A. Amorim[a,b]

[a]IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), R. Dr. Roberto Frias, s/n, 4200 Porto, Portugal
[b]Faculdade de Ciências da Universidade do Porto, Praça Gomes Teixeira 4050 Porto, Portugal

Corresponding author: Luísa Pereira, IPATIMUP, R. Dr. Roberto Frias s/n, 4200 Porto, Portugal. Phone: +351225570700 Fax: +351225570799 email: lpereira@ipatimup.pt

ABSTRACT

We present an approach to the evaluation of the probability of Y-STR haplotypes identity by state (IBS, in opposition to identity by descent, IBD) alternative to the empirical one for which it is necessary to have simultaneous information on STRs and SNPs, which is rare in the forensic field. It is based on the mismatch distribution analyses for the number of repeat unit differences between pairs of Y-STR haplotypes. The estimates of the IBS by both methods in a sample where STR and SNP information is available are compared and the relevance of population structure in determining widely different proportions of IBS vs. IBD.

KEYWORDS Recurrence, identity by state, identity by descent, mismatch distribution

1. INTRODUCTION

Search for haplotype matching is the main tool used in forensics to evaluate how significant is the observation of a particular haplotype in a certain sample. A neglected phenomenon that can be an important bias source to this kind of analysis is the possibility of recurrence, due to the high mutation rate of STRs, and hence, the rise of haplotypes identical by state, rather than by descent.

One approach that has been used to evaluate the proportion of these classes among identical haplotypes is to assess haplotype identity using SNPs, the same Y-STR haplotype in two different Y-SNP haplogroups pointing to identity by state and not by descent. De Knijff [1] estimated an IBS proportion of 0.2% of in a sample of 275 Dutch screened for 8 Y-STRs and 4-SNPs.

Unfortunately, combined information for both Y-STRs and Y-SNPs in the same sample is very rare, and it is therefore impossible to measure the proportion of IBS, at a large scale, using the above described approach.

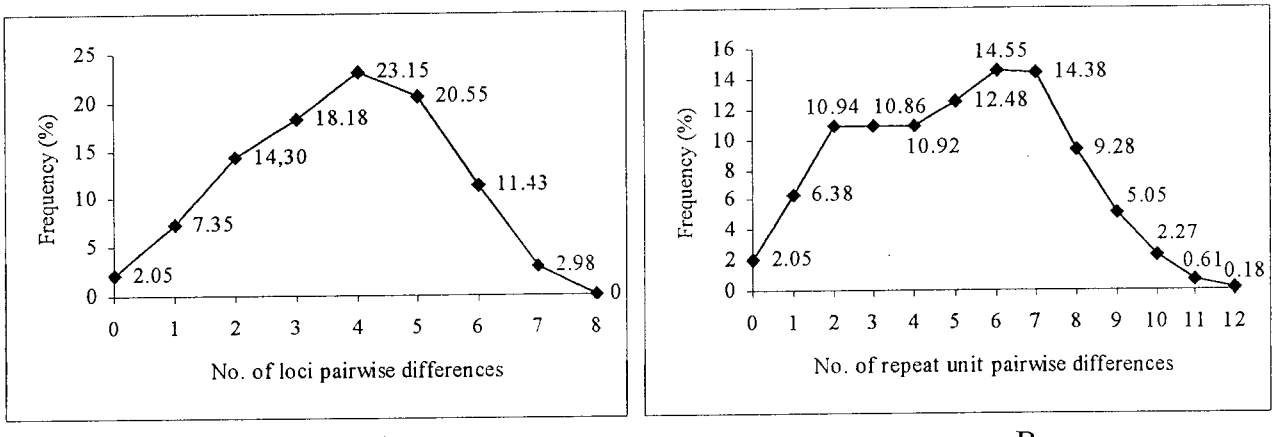
We suggest an approach based upon mismatch distributions/haplotype pairwise comparison that can at least indicate the proportion of Y-STR haplotypes in which mutation can originate IBS.

2. MATERIAL AND METHODS

A total of 123 different Y-STR haplotypes for 229 Dutch were collected from the Y-STR Haplotype Reference Database (http://ystr.charite.de). The samples of Holland (N=87), Friesland (N=44), Groningen (N=48) and Limburg (N=50) were considered together. The Y-STRs analysed were DYS19, DYS389I and II, DYS390, DYS391, DYS392 and DYS393. Mismatch distributions were calculated in ARLEQUIN 2.0 [2] software and the input files were: (1) the microsatellite input file, considering the number of repeat units observed in each locus, for obtaining the mismatch distribution for differences in number of locus; (2) a DNA input file, where each repeat unit is considered as an ambiguous position (N) and differences in the number of repeat units are considered as insertion (N) or deletion (-), for estimating the mismatch distribution for differences in repeat units between pairs of haplotypes.

3. RESULTS

The two mismatch distributions obtained for the 229 Dutch are displayed in figure 1. Graph A refers to the mismatch distribution for differences per locus and graph B differences in repeat units.



A B

Figure 1: Mismatch distributions for the number of loci (A) and number of repeat unit (B) pairwise differences for Y-STRs in a Dutch sample.

It is noteworthy that 86.8% (6.38%/7.35%) of the differences between haplotypes differing at one locus consist in a single repeat unit.

The probability of a certain Y-STR haplotype becoming identical by state to another one inside the class of those that differ by a single repeat can be roughly estimated as $nm(1-m)^{n-1}$, where n is the number of loci defining the haplotype (7 in our case) and m the average mutation rate (3.17×10^{-3} [3]). Applying it to the proportion of the class (6.38%) it results, for this sample, in a value of 0.14%, which is in agreement to the one inferred from 4 SNPs information on 8 STRs defined haplotypes [1].

We have not considered the rare cases of IBS resulting from more complex phenomena since these will add an almost insignificant contribution (for instance, two-step mutation rate is roughly 10 times lower [3]).

4. DISCUSSION

STRs have been markers of choice in forensics due to its high polymorphism. SNPs, although presenting high stability, are much less polymorphic, becoming less informative in the forensic field, while more advanced technologies allowing its easy and fast typing are not available. So, extensive forensic databases available for matching are now (and will be maintained for more years) exclusively based on STRs.

It seems therefore that the approach presented here can be useful in the forensic field, for the evaluation of the significance and evidential value of Y-STR haplotype matches.

Another important issue is the use of large scale databases. It has been claimed that, since there is no significant population substructuring for STRs in Europe (contrarily to SNPs [4]) matching against the whole database is legitimate. We demonstrate that this claim is at least debatable, since the opportunity for IBS varies a lot across Europe, as judged from the mismatch distributions (Table 1), in a way congruent with the well defined SNPs haplogroups' gradients [5]: higher SNP's diversity in south-central Europe associated with lower proportion of haplotype pairs prone to IBS; and decreasing SNP's diversity towards west and north associated with higher risk of recurrence.

Table 1: Proportions of pairs of haplotypes differing in 0 (h_0) and 1 (h_1) repeat unit and estimates of identity-by-state (IBS). Populations deposited in the Forensic database were grouped by country (data extracted in 11/04/2001; only Zeeland was not used).

Population	h_0 (%)	h_1 (%)	IBS (%)	Population	h_0 (%)	h_1 (%)	IBS (%)
Portugal (182)	1.85	6.70	0.146	The Netherlands (229)	2.05	6.38	0.139
Spain (365)	1.69	5.77	0.126	Poland (596)	1.51	5.83	0.127
Italy (553)	0.92	3.02	0.066	Estonia (133)	1.31	3.68	0.080
Austria (135)	0.98	3.45	0.075	Latvia (145)	0.93	4.53	0.099
Hungary (117)	0.43	2.18	0.047	Lithuania (151)	1.17	4.73	0.103
Switzerland (199)	1.11	4.26	0.093	Russia (85)	2.24	6.44	0.140
Germany (2096)	0.83	3.18	0.069	Norway (300)	1.41	3.75	0.082
Belgium (97)	1.55	5.61	0.122	Buenos Aires (100)	1.23	2.87	0.062

ACKNOWLEDGEMENTS

This work was partially supported by a research grant (PRAXIS XXI BD/13632/97) from Fundação para a Ciência e a Tecnologia and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

REFERENCES

[1] De Knifjj P. Y chromosome shared by descent or by state. In: Renfrew C, Boyle K, editors. Archaeogenetics: DNA and the population prehistory of Europe. Cambridge: McDonald Institute Monographs. Oxbow Books, 2000. p. 301-4.

[2] Schneider S, Roessli D, Excoffier L. Arlequin ver.2.0: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland 2000.

[3] Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Krüger C, Krawczak M, Nagy M, Dobosz T, Szibor R, De Knijff P, Stoneking M, Sajantila A. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Ann Hum Genet 2000;66:1580-8.

[4] Rosser ZH et al. Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. Am J Hum Genet 2000;67:1526-43.

[5] Pereira L, Dupanloup de Ceuninck I, Rosser ZH, Jobling MA, Barbujani G. Y-chromosome mismatch distributions in Europe. Mol Biol Evol 2001;18:1259-71.

RESULTS

II - POPULATION GENETICS' APPLICATIONS

A- PORTUGAL - THE COUNTRY CONTEXT

In this sub-section we include the papers that report the results on the characterisation of Portuguese Y-chromosome and mtDNA pools. These results are discussed in the framework of the major pre- and proto-historic events and also demographic phenomena that have occurred in the last 13 centuries.

In a country like Portugal, whose European political boundaries are essentially stable for more than 700 years, the comparative analyses of Y-chromosome and mtDNA population substructuring seemed particularly attractive.

An overview for both gene pools is presented in article 4. The permanent refinement of mtDNA haplogroups' classification led us to deeper analyses as reported in article 5:

ARTICLE 4

PEREIRA, L., PRATA M.J., JOBLING M.A., AMORIM, A. (2000) Analysis of the Y-chromosome and Mitochondrial DNA pools in Portugal. In Renfrew C., Boyle K. (eds) Archaeogenetics: DNA and the population prehistory of Europe. Chapter **20**:191-195. McDonald Institute Monographs. Oxbow Books, Cambridge.

ARTICLE 5

PEREIRA, L., PRATA, M.J., AMORIM, A. (2000) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. *Ann. Hum. Genet.* **64**:491 -506.

Chapter 20

Analysis of the Y-chromosome and Mitochondrial DNA Pools in Portugal

Luísa Pereira, Maria João Prata, Mark A. Jobling & António Amorim

Portugal is the most western country in Iberian Peninsula and, just focusing on historical times, it has been subject to more or less substantial inputs of foreign populations such as Romans, Barbarians and North Africans. The aim of this study was to characterize the Portuguese population relatively to Y-chromosome and the mtDNA variability. We have considered three main regions in the Portuguese population: North, Central and South. 10 Y-chromosome biallelic markers were studied as well as hypervariable regions I and II of the mtDNA control region. This allowed the construction of Y-chromosome and mtDNA haplogroups, both characterized by being highly geographically clustered, specially the first. Results on Y-chromosome biallelic markers revealed a low level of inter-regional diversity, although South Portugal was found to be significantly different from North Portugal. An interesting observation was a clear gradient of increasing frequency from North to South concerning haplogroup 21. As this haplogroup reaches the highest frequencies in Berber-speaking populations, the clinal variation registered might be interpreted as a sign of North African influence in the Portuguese genetic background. Levels of mtDNA diversity at HVRI in the Portuguese sample analyzed were higher than the values published for some neighbouring Iberian populations (Galicia, Basque country). This finding can be interpreted as a result of influx of new sequences from other populations. Sixteen different mtDNA haplogroups were found, including haplogroup U6 that reached the frequency of seven per cent in North Portugal. This haplogroup, which occurs in high frequency in Berber-speaking populations, has been reported to be absent in Europe except in Iberia.

Y-chromosome and mtDNA polymorphisms are particularly suitable for population genetics since they share the special features of haploidy, lack of recombination, uniparental inheritance and a four-fold reduction in effective population size relative to autosomes. Important inferences concerning population origins and movements are made possible by the simultaneous study of Y chromosome and mtDNA, since they give complementary information: the first about male and the second about female lineages. These inferences are relatively easy to derive because the Y and mtDNA haplogroups, especially the former, display a highly geographical clustering and, in some cases, haplotype frequency gradients can be followed through large or restricted geographic regions.

Male and female histories are not necessarily coincident and some results (Seielstad *et al.* 1998), based on Y chromosome and mtDNA SNPs pointed to very different patterns of dispersion of the sexes through the World.

In this work we have analyzed 10 Y-chromosome biallelic markers and HVRI and HVRII regions of mtDNA in Portugal. In the population screening we have considered three main regions in Portugal: North, Central and South.

Table 20.1. *Number of individuals screened in the three Portuguese samples.*

	Y chromosome	mtDNA
North Portugal	329	100
Central Portugal	118	82
South Portugal	49	59

Table 20.3. *Population pairwise F_{ST} P values between samples from North, Central and South Portugal. Values marked by an asterisk are significant at the 5 per cent level.*

	Central Portugal	North Portugal
North Portugal	0.12871 ± 0.0260	
South Portugal	0.16832 ± 0.0473	0.00000* ± 0.0000

Table 20.2. *Haplogroups identified with the ten Y-chromosome biallelic markers studied.*

	YAP	SRY-8299	92R7	SRY-1532	SRY-2627	Tat	SY81	M9	LLY 22g	12f2
2	0	0	0	1	0	0	0	0	0	0
16	0	0	0	1	0	1	0	1	1	0
21	1	1	0	1	0	0	0	0	0	0
8	1	1	0	1	0	0	1	0	0	0
3	0	0	1	0	0	0	0	1	0	0
1	0	0	1	1	0	0	0	1	0	0
22	0	0	1	1	1	0	0	1	0	0
9	0	0	0	1	0	0	0	0	0	1
4	1	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
12	0	0	0	1	0	0	0	1	1	0
26	0	0	0	1	0	0	0	1	0	0

The aims of the work were to characterize this Iberian country relative to Y-chromosome and mtDNA variability and to detect any possible genetic record of the several populations that crossed this southwestern region of Europe in prehistoric and historic times.

Materials and methods

Populations and sample sizes

The Portuguese populations analyzed were: North, Central and South, the country be divided up according to the major river basins: the Douro and Tejo. Sample sizes are presented in Table 20.1.

Y-chromosome biallelic markers

The ten biallelic markers analyzed are listed in Table 20.2. The table also displays the compound haplogroups defined by these markers which have already been observed (Jobling, unpublished results). Analytical conditions will be published by M.A. Jobling.

MtDNA

The two hypervariable regions, HVRI and HVRII, were analyzed. Primers for PCR amplification were: L15997 and H16401 for HVRI; L48 and H408 for HVRII. PCR amplification conditions were according to Wilson *et al.* (1995).

The amplified samples were purified with Microspin™ S-300 HR columns (Pharmacia Biotech). Sequencing reaction was performed using the Kit Big-Dye™ Terminator Cycle Sequencing Ready Reaction (Perkin-Elmer) with the primers above described, in the forward and reverse directions. Post cycle sequencing reaction sample purification was undertaken using a $MgCl_2$/ethanol-based protocol. Sequence run and analysis were performed in an ABI 377 sequencer.

Statistical analysis

Sequence diversity parameters and population pairwise differentiation test based upon F_{ST} were obtained with the software ARLEQUIN (Schneider *et al.* 1997).

Results and discussion

I - Y-chromosome biallelic markers

Analysis of the referred ten biallelic markers led to the detection of seven different haplogroups.

Figure 20.1 shows the haplogroup frequency distribution observed in the three Portuguese regions and Table 20.3 presents values of the popula-

tion pairwise differentiation test.

The major findings were:

- an increasing frequency gradient for haplogroup 21 from North to South Portugal;
- absence of haplogroup 8 (characteristic of Sub-Saharan populations) in the three regions;
- statistically significant difference between South and North Portugal.

II - MtDNA - HVRI and HVRII

Figure 20.2 represents the nucleotide pairwise difference distributions whereas some mtDNA diversity parameters are displayed in Table 20.4.

The different sequences were classified in haplogroups according to Richards *et al.* (1998) and the corresponding distributions are represented in Table 20.5.

For mtDNA, the main results were:

- the nucleotide pairwise difference distributions were very similar in all the Portuguese samples for HVRI and HVRII;
- concerning HVRI, absence of clear bell-shaped distributions of the number of nucleotide pairwise differences and high sequence diversity; both findings differentiate the Portuguese samples from other neighbouring Iberian populations (Côrte-Real *et al.* 1996; Salas *et al.* 1998), suggesting an older or mixed origin for Portugal;
- all major European clusters were detected in the Portuguese samples: about 80 per cent of the sequences belong to the so called Palaeolithic expanded clusters (Richards *et al.* 1998) (V, K, U, U3, U4, U5, W, X, T, I & H) and about 10 per cent to Neolithic expanded ones (J, J1, J2 & T1);
- about 10 per cent of the sequences may represent introductions from Africa;
- North African haplogroup U6 was only detected

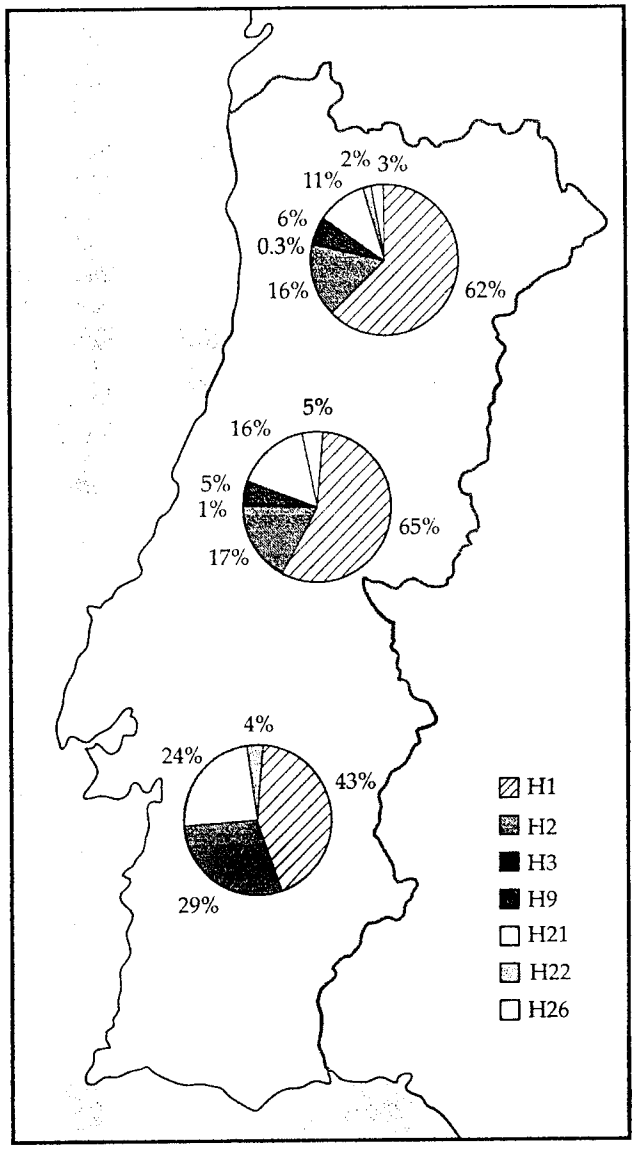


Figure 20.1. *Y-chromosome biallelic marker haplogroup distributions in Portugal.*

Table 20.4. *MtDNA diversity parameters in North, Central and South Portugal, considering HVRI and/or HVRII.*

		% of different haplotypes	Mean no. of nucleotide pairwise differences	Nucleotide diversity
HVRI	North Portugal	67.00	4.784243	0.013290
	Central Portugal	75.60	4.845529	0.013460
	South Portugal	69.49	4.509644	0.012527
HVRII	North Portugal	47.00	3.089091	0.011399
	Central Portugal	50.00	3.247516	0.011983
	South Portugal	61.02	3.552309	0.013060
HVRI	North Portugal	84.00	7.873333	0.012478
+	Central Portugal	92.68	8.093044	0.012826
HVRII	South Portugal	91.53	8.040912	0.012723

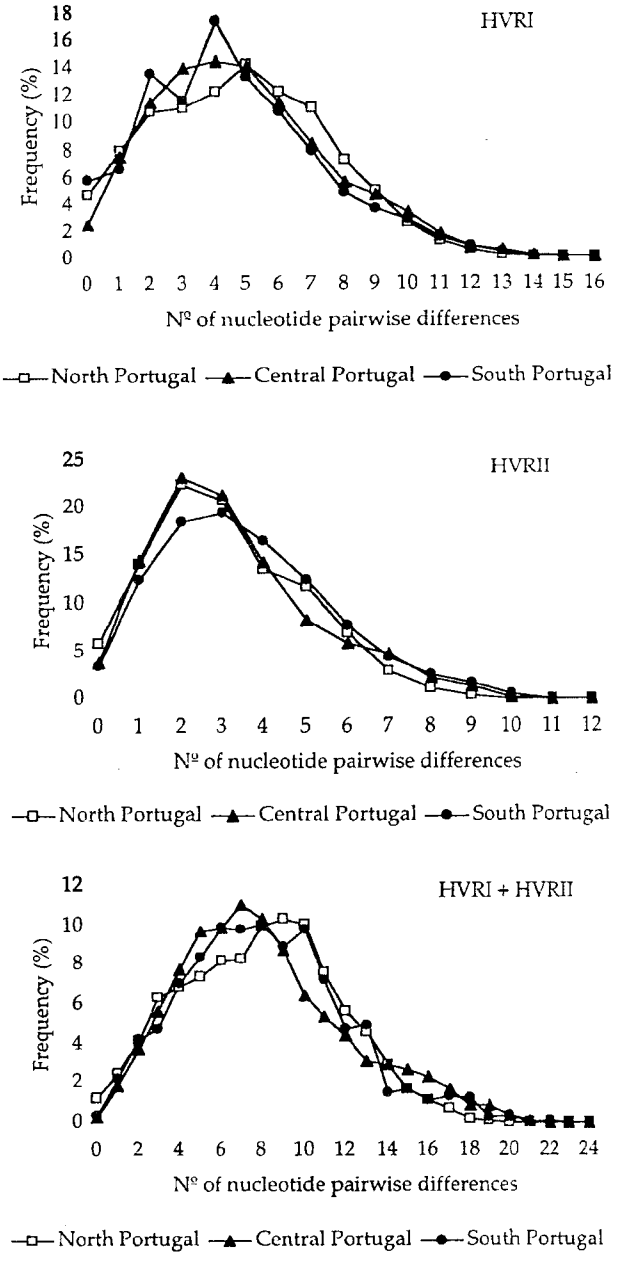


Figure 20.2. *Nucleotide pairwise difference distributions observed in North, Central and South Portugal, considering HVRI and/or HVRII.*

in North Portugal, where it reached the frequency of 7 per cent;

- other African haplogroups (Rando *et al.* 1998) were detected in all Portuguese regions analyzed, ranging from 5 per cent in North and South Portugal to 10 per cent in Central Portugal: L1b (West African), L2 (Pan African) and L3a (Sub-Saharan African).

Table 20.5. *MtDNA haplogroup distributions (values in %) in North (NP), Central (CP) and South (SP) Portugal.*

	NP	CP	SP		NP	CP	SP
H	41.00	36.59	38.98	**J1**	2.00	–	3.39
V	8.00	3.66	6.78	**J2**	2.00	1.22	–
K	3.00	7.32	6.78	**X**	–	3.66	5.08
U	6.00	9.76	10.17	**W**	2.00	1.22	–
U3	1.00	–	–	**T**	3.00	9.76	10.17
U4	2.00	3.66	–	**T1**	7.00	1.22	–
U5	7.00	6.10	6.78	**I**	1.00	1.22	1.70
U6	7.00	–	–	**L1b**	–	1.22	1.70
JT*	1.00	–	–	**L2**	3.00	2.44	1.70
J	2.00	4.88	5.08	**L3a***	2.00	6.10	1.70

Conclusions

In the Y-chromosome pool, the increasing frequency gradient of haplogroup 21 from north to south has also been described for other European regions (Hammer *et al.* 1998) and has been related to the Neolithic diffusion through the continent. However, since this haplogroup presents high frequencies in North African populations and it is known that multiple historic and pre-historic contacts between North Africans and Iberians have occurred, this cline could at least be enhanced (if not originated) by interchanges between these populations, besides those determined by Neolithic diffusion.

The importance of the Islamic influence in Western Iberia, which started at the beginning of the eighth century AD, was markedly heterogeneous (Saraiva 1993). While northern regions remained totally or practically untouched, in the central and especially the southern regions, Islamic administration lasted up to the thirteenth century. This pattern clearly mimics the clinal variation registered for haplogroup 21.

The absence of the characteristic Sub-Saharan haplogroup 8 suggests that interbreeding between African male slaves who entered Europe from the fifteenth to the last century, and Portuguese females must has been very restricted, with a minor impact in the paternally inherited Portuguese gene pool.

In the maternal gene pool, the distribution of the U6 haplogroup suggests a North African influence restricted, surprisingly, to North Portugal. This pattern is not consistent with the above-described chronology of Islamic administration in Portugal and may represent either another population movement not yet identified, or the effects of drift.

With respect to the detection of L haplogroups all over the country, it is possible that this reflects the recent presence of Black African slaves in Portugal. If this association is valid, then the African presence has led to a significant contribution to the present day mtDNA Portuguese gene pool.

The differential influence of the Black African genes in the Y chromosome and mtDNA seems to point to a much more frequent interbreeding between autochthonous males with Black African female slaves than the opposite, a conclusion which is in accordance with the socio-cultural relationships between Portuguese and African slaves.

Acknowledgements

This work was partially supported through a grant (PRAXIS BD/13632/97) and a project (PRAXIS/2/2.1/BIA/196/ 94) from Fundação para a Ciência e a Tecnologia. M.A.J. is a Wellcome Senior Research Fellow in Basic Biomedical Science (grant no. 057559/Z/99/Z).

References

Côrte-Real, H., V.A. Macaulay, M.B. Richards, G. Hariti, M.S. Issad, A. Cambon-Thomsen, S. Papiha, S. Bertranpetit & B.C. Sykes, 1996. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Annals of Human Genetics* 60, 331–50.

Hammer, M.F., T. Karafet, A. Rasanayagam, E.T. Wood, T.K. Altheide, T. Jenkins, R.C. Griffiths, A.R. Templeton & S.L. Zegura, 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molecular Biology and Evolution* 15(4), 427–41.

Rando, J.C., F. Pinto, A.M. González, M. Hernández, J.M. Larruga, V.M. Cabrera & H-J. Bandelt, 1998. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern and sub-Saharan populations. *Annals of Human Genetics* 62, 531–50.

Richards, M.B., V.A. Macaulay, H-J. Bandelt & B.C. Sykes, 1998. Phylogeography of mitochondrial DNA in western Europe. *Annals of Human Genetics* 62, 241–60.

Salas, A., D. Comas, M.V. Lareu, J. Bertranpetit & A. Carracedo, 1998. MtDNA analysis of the Galician population: a genetic edge of European variation. *European Journal of Human Genetics* 6, 365–75.

Saraiva, J.H., 1993. *História de Portugal*. 4th edition. Lisbon: Publicações Europa-América.

Schneider, S., J.M. Kueffer, D. Roessli & L. Excoffier, 1997. *Arlequin ver 1.1: a Software for Population Genetic Data Analysis*. Genetic and Biometry Laboratory, University of Geneva, Switzerland.

Seielstad, M.T., E. Minch & L. Cavalli-Sforza, 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20, 278–80.

Wilson, M.R., J.A. Di Zinnow, D. Polanskey, J. Replogle & B. Budowle, 1995. Validation of mitochondrial DNA sequencing for forensic casework analysis. *International Journal of Legal Medicine* 108, 68–74.

Ann. Hum. Genet. (2000), **64**, 491–506
Printed in Great Britain

Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation

L. PEREIRA[1,2], M. J. PRATA[1,2] AND A. AMORIM[1,2]

[1] *Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/no, 4200 Porto, PORTUGAL*

[2] *Faculdade de Ciências da Universidade do Porto, Pr. Gomes Teixeira, 4050 Porto, PORTUGAL*

(Received 11.1.00. Accepted 26.7.00)

SUMMARY

The analysis of the hypervariable regions I and II of mitochondrial DNA in Portugal showed that this Iberian population presents a higher level of diversity than some neighbouring populations. The classification of the different sequences into haplogroups revealed the presence of all the most important European haplogroups, including those that expanded through Europe in the Palaeolithic, and those whose expansion has occurred during the Neolithic. Additionally a rather distinct African influence was detected in this Portuguese survey, as signalled by the distributions of haplogroups U6 and L, present at higher frequencies than those usually reported in Iberian populations. The geographical distributions of both haplogroups were quite different, with U6 being restricted to North Portugal whereas L was widespread all over the country. This seems to point to different population movements as the main contributors for the two haplogroup introductions. We hypothesise that the recent Black African slave trade could have been the mediator of most of the L sequence inputs, while the population movement associated with the Muslim rule of Iberia has predominantly introduced U6 lineages.

INTRODUCTION

Since the description of the mitochondrial DNA sequence by Anderson *et al.* (1981), this peculiar genome, which is maternally inherited, non-recombining and fast-evolving, has been intensively investigated and applied to population studies. The initial screening based on restriction fragment length polymorphisms spread all over the molecule, was soon enlarged by direct sequencing of two hypervariable regions located in the control region (D-Loop): HVRI and HVRII.

Correspondence: Luísa Pereira, Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200 Porto PORTUGAL. Tel: +351 22 5570700; Fax: +351 22 5570799.
E-mail: lpereira@ipatimup.pt

Present day sequence variation of mtDNA is a valuable tool for making what Avise *et al.* (1987) referred to as phylogeographic inferences. MtDNA sequences can be used to construct networks or used in other methodological approaches which afford information about prehistoric population size and patterns of gene flow. The evolutionary history of haplogroups, their inferred origin and expansion through the world, provide the basis for reconstructing and dating major prehistoric and historic population movements.

Many published studies based on HVRI sequence diversity focus on the history of European populations (Richards *et al.* 1996; Côrte-Real *et al.* 1996; Richards *et al.* 1998). Richards *et al.* (1998), applying a phylogeographic approach to western Europe mtDNA diversity, concluded that the majority (85%) of

European sequences must have originated during the Upper Palaeolithic and suffered a considerable post-glacial expansion: about 15% of the sequences reflect a restricted Neolithic input, from the Near East toward the West of Europe, and only 1% of the sequences represent more recent influences of Asian and African mtDNA pools.

Focussing on the westernmost edge of Europe, the Iberian Peninsula, some studies (Côrte-Real *et al.* 1996; Salas *et al.* 1998) have pointed to a common origin of all Iberian populations in the Upper Palaeolithic. For most populations, diversity levels were found to be lower than the values reported for central European countries, a feature that was thought to support the expansion model of modern humans from the Middle East in the direction of Western Europe. The lowest Iberian diversity value was observed in Basques, reflecting the uniqueness of this Iberian population (Côrte-Real *et al.* 1996).

Another peculiarity of the Iberian mitochondrial pool is the presence of sequences belonging to the U6 group (Richards *et al.* 1998), signalling a North African influence that has not been detected elsewhere in other European populations.

In this work we have analysed HVRI and HVRII diversity in Portugal, the westernmost country of the Iberian Peninsula, with the aim of obtaining a better characterisation of European mtDNA variability. We have considered three main regions in Portugal: North, Central and South. This was done in parallel with a study of Y chromosome biallelic markers that has revealed statistical differences between the south compared to the north and central regions (Pereira *et al.* 2000).

Our main approach regarding the analysis of mtDNA diversity was the evaluation of patterns of mismatch distribution within the major haplogroups found in Portugal. We intended to assess whether inferences regarding the history of haplogroups were in agreement with those previously published based on network analysis, and, simultaneously, to deepen the evolutionary picture of mtDNA lineages in Europe.

MATERIAL AND METHODS

Population samples

Three population samples from Portugal were analysed: 100 unrelated individuals from the North, 82 from the Central region and 59 from the South, according to the country division by the major rivers Douro and Tagus. A total of 15 μl of blood was used to extract DNA by the resin Chelex-100 method (Lareu *et al.* 1994).

MtDNA amplification and sequencing

MtDNA was amplified using the primers L15997 (5′-CACCATTAGCACCC AAAGCT-3′) and H16401 (5′-TGATTTCACGGAGGATGGTG-3′) for HVRI, and L48 (5′-CTCACGGGAGCTCTCCATGC-3′) and H408 (5′-CTGTTAAAAGTGCATACCG CCA-3′) for HVRII. The temperature profile was 95 °C for 10 sec., 60 °C for 30 sec. and 72 °C for 30 sec., for 35 cycles of amplification.

The amplified samples were purified with Microspin™ S-300 HR columns (Pharmacia Biotech), according to the manufacturer's specifications. The sequencing reactions were carried out using the Kit Big-Dye™ Terminator Cycle Sequencing Ready Reaction (Perkin-Elmer), with one of the above described primers, in both forward and reverse directions.

A protocol based on $MgCl_2$/ethanol precipitation was used for post-sequence reaction purification of samples, which were then applied to a 6% PAGE gel and run in an automatic sequencer ABI 377.

Genetic analysis

The nucleotide positions considered for analysis were between bp 16024 and 16383 for HVRI and between 73 and 340 for HVRII (in the numbering system of Anderson *et al.* 1981).

Sequence classification into haplogroups was based on HVRI and position 00073 of HVRII, and the nomenclatures of Richards *et al.* (1998), Macaulay *et al.* (1999) and Rando *et al.* (1999) were followed for European, Sub-Saharan and North African clusters, respectively. Sequences are available in GenBank (accession

nos. AF277997 AF278237 and AF278238 AF278478).

Molecular diversity indexes and mismatch distributions were executed using the software ARLEQUIN 1.1 (Schneider *et al.* 1997).

RESULTS AND DISCUSSION

IIVRI and IIVRII diversity in Portugal

Some diversity parameters obtained in the three Portuguese regions studied for HVRI and/or II are presented in Table 1. HVRI presented a higher mean number of nucleotide differences than region II. However, when corrected for fragment sizes both regions showed a similar mean number of nucleotide pairwise differences: HVRII/HVRI mean ratios were 0.87, 0.90 and 1.05 in North, Central and South Portugal, respectively.

The proportion of polymorphic sites (Table 2) was higher in HVRI than in HVRII, averaging 17.68% and 11.20%, respectively. However, region II presented a slightly faster mutation rate than the former, a conclusion that was based on the comparison of estimates of the τ parameter of Rogers & Harpending (1992) in both regions. This parameter consists in $\tau = \mu l t$, where μ is mutation rate per nucleotide, l is sequence length and t is time in generations after a population expansion. Since t is equal in HVRI and HVRII in a certain population, and knowing the length of both regions, we can obtain μ_{II}/μ_I from the ratio τ_{II}/τ_I. The μ_{II}/μ_I ratios for Portugal were 1.044, 0.984 and 1.273, in North, Central and South respectively, with a mean value of 1.10, a value close to that reported by Salas *et al.* (2000) for other European populations (0.998 for British, 1.216 for Austrian and 0.943 for Tuscan), excepting Galicia, where the very high value found (1.845) is related to the low mean of nucleotide differences reported for HVRI in that population.

Combining both sets of observations it is clear that in HVRI mutations tend to occur more homogeneously along different nucleotide positions than in HVRII. Therefore, our data are in agreement with previous studies (Meyer *et al.* 1999; Torroni *et al.* 1996) that have described fast-mutating positions in the HVRII region: positions 146, 150, 152 and 195 are very prone to substitution events whereas at position 309 a length polymorphism is frequently found.

The differential mutational behaviour of HVRI and HVRII is also reflected in the pattern of mismatch distributions. The observed and expected numbers of pairwise differences were analysed in the three geographic regions, but as the distribution patterns were basically identical in the three regions, only those corresponding to the overall Portuguese sample will be presented (Figure 1). For HVRII, the observed distribution closely matches the expected distribution. However, the observed distributions for HVRI and HVRI+HVRII are characterised by the presence of slight shoulders and higher number of nucleotide differences compared to the expected values.

It is widely accepted that the mismatch distribution retains valuable information about demographic episodes undergone during the history of a population. Unimodal curves with modes at a small number of differences have been observed in western European populations (Côrte-Real *et al.* 1996; Salas *et al.* 1998) and were interpreted as signatures of relatively recent population expansions. By contrast, the tendency of African populations (Mateu *et al.* 1997) to display ragged and multimodal distributions has supported the idea of their being more ancient and stationary, or more diversified. It is worth mentioning that making population demographic inferences from the analysis of mismatch distributions must be tentative since several factors have to be considered: time of expansion, size before the expansion, gene flow between populations and sub-structuring of populations. Some simulation studies have shown that the statistical effects of some factors can be quite convergent (Marjoram & Donnelly, 1994).

In the presence of mismatch distributions showing some deviations from regular characteristic unimodal distributions, as we have found for HVRI and HVRI+HVRII in Portugal, we may be facing a population in which one or more

Table 1. *MtDNA diversity parameters in North (NP), Central (CP) and South (SP) Portugal, considering IIVRI and/or IIVRII*

	% of different haplotypes	Mean no. of nucleotide pairwise differences	Nucleotide diversity
HVRI			
NP	**67.0**	4.78	0.013
CP	**75.6**	4.87	0.014
SP	**69.5**	4.54	0.013
HVRII			
NP	**47.0**	3.09	0.012
CP	**50.0**	3.26	0.012
SP	**61.0**	3.55	0.013
HVRI+HVRII			
NP	84.0	7.87	0.012
CP	**92.7**	8.13	0.013
SP	**91.5**	8.09	0.013

Table 2. *IIVRI and IIVRII variability in North (NP), Central (CP) and South (SP) Portugal*

	HVRI (360 bp)	HVRII (268 bp)
Polymorphic sites (%)		
NP	71 (19.7)	28 (10.5)
CP	66 (18.3)	31 (11.6)
SP	54 (15.0)	31 (11.6)
No. of substitutions (%)		
NP	73 (100.0)	25 (89.3)
CP	69 (100.0)	30 (93.8)
SP	54 (100.0)	29 (90.6)
No. of transitions (%)		
NP	66 (90.4)	24 (96.0)
CP	62 (89.9)	28 (90.3)
SP	50 (92.6)	27 (93.1)
No. of transversions (%)		
NP	7 (9.6)	1 (4.0)
CP	7 (10.1)	2 (6.7)
SP	4 (7.4)	2 (6.9)
No. of indels (%)		
NP	0 (0.0)	3 (10.7)
CP	0 (0.0)	2 (6.3)
SP	0 (0.0)	3 (9.4)

of those factors has played a significant role. Even without discriminating which factors these might be, the mismatch distribution observed for HVRII suggests that this sequence stretch is more resistant to their effects. We cannot exclude the possibility that the regular unimodal distribution for HVRII could be the result of an (unknown) selective effect. However, if this was the case, this selection would also affect HVRI (since they are linked on the same molecule). Another explanation is related to the fact that, as already mentioned, mutational events in HVRII are more heterogeneously dispersed. Since some HVRII sites present very high mutation rates the nucleotide variation does not tend to be so strongly associated in blocks as happens in the HVRI region. Within HVRI, substitutions are more often found to be specific to certain haplogroups and, by turn, it is easier to classify the observed variation into haplogroups, which manifestly tend to show geographic clustering. Attending to all these features, we suggest that HVRI and HVRI +HVRII distributions are more sensitive to demographic factors and consequently their study seems to be more informative for making population demographic inferences.

Diversity comparison with other populations

For population comparison purposes, we have not considered the diversity values registered for HVRII, since population data for this region are more scarce.

Table 3 presents some diversity estimates for several populations. The Portuguese samples studied here display a mean number of nucleotide pairwise differences typical of European populations, which are, in general, lower than the values characteristic of Asian and African populations (Comas *et al.* 1998: Mateu *et al.* 1997). However, compared to a neighbouring Iberian population, namely Galicia (Salas *et al.* 1998), the populations analysed here present a higher level of diversity, even when assessed excluding the African L sequences (data not shown), which, as will be further discussed, are found at very low frequency in other Iberian populations.

It is difficult to explain the higher mtDNA diversity in Portugal compared to other Iberian neighbouring populations. The data might suggest an ancient settlement of this region of the Iberian Peninsula, a hypothesis that is not supported by historical or palaeonthological records. Alternatively, the use of North Portugal as a refuge zone during the last glacial maximum, as has been suggested for Andalusia (Côrte-Real

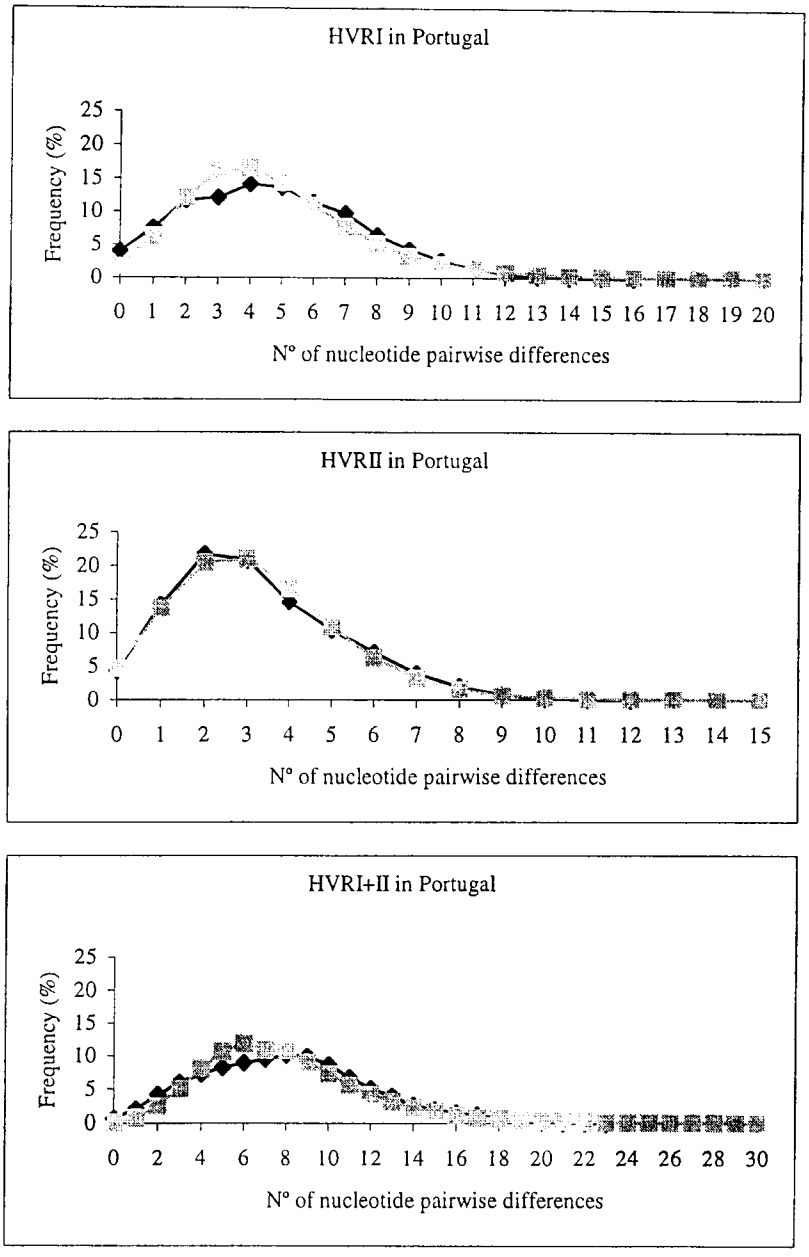


Fig. 1. Observed (black) and expected (grey) mismatch distributions for HVRI, HVRII and HVRI + HVRII in Portugal.

et al. 1996), maintaining the level of diversity while population size was reduced in other Iberian regions, is a possible explanation, but in contradiction with the known palaeoclimatic data (Mellars, 1998). A more likely explanation is the occurrence at different times of several significant influxes of different mtDNA lineages, a hypothesis that is in accordance with the multiple contacts with different people that have characterised the recent history (i.e. the last thousand years) of the region.

Haplogroup diversity analysis

The haplogroup frequencies observed in North, Central and South Portugal are summarised in Table 4. In HVRII, the CRS sequence was never found because two positions, not polymorphic in Portugal, were always different from the ref-

Table 3. *Sequence diversity observed in HVRI in several populations. N, sample size; K, number of different sequences found; A, number of variable nucleotides positions; B, mean nucleotide pairwise differences; C, percentage average pairwise difference per nucleotide; π, nucleotide diversity.*

	N	K	A	B	C	π	References
Basque	106	52	52	2.95	0.82	0.008	1,2
Galician	92	53	56	3.13	0.87	0.009	3
Portuguese	54	38	46	3.60	1.00	0.010	2
Catalonian	15	11	16	3.73	1.04	0.010	2
British	100	71	67	4.45	1.24	0.012	4
S. Portuguese	59	41	54	4.51	1.25	0.013	This study
N. Portuguese	100	67	71	4.78	1.33	0.013	This study
C. Portuguese	82	62	66	4.87	1.35	0.014	This study
Spanish	89	70	69	5.02	1.39	0.014	2,5
Tuscan	49	40	55	5.03	1.40	0.014	6
Turkish	96	79	82	5.45	1.51	0.015	7,8
Middle-Eastern	42	38	59	7.08	1.97	0.020	9
S. Tomean	50	32	53	7.56	2.10	0.021	10

[1] Bertranpetit *et al.* (1995); [2] Côrte-Real *et al.* (1996); [3] Salas *et al.* (1998); [4] Piercy *et al.* (1996); [5] Pinto *et al.* (1996); [6] Francalacci *et al.* (1996); [7] Calafell *et al.* (1996); [8] Comas *et al.* (1996); [9] Di Rienzo *et al.* (1991); [10] Mateu *et al.* (1997).

Table 4. *MtDNA haplogroup distributions (no. of individuals and % values in parenthesis) in North (NP), Central (CP) and South (SP) Portugal*

	NP	CP	SP
H	40+1? (41.00)	31 (37.81)	25+1? (50.0)
I	1 (1.00)	—	1 (1.70)
J*	2 (2.00)	4 (4.88)	3 (5.09)
J1	2 (2.00)	—	—
J1b	—	—	2 (3.39)
J2	2 (2.00)	1 (1.22)	—
K	3 (3.00)	6 (7.32)	4 (6.78)
L1b	—	1 (1.22)	1 (1.70)
L2	3 (3.00)	2 (2.44)	1 (1.70)
L3*	2 (2.00)	5 (6.10)	2 (3.39)
M1	—	1 (1.22)	—
T*	3+1? (4.00)	8 (9.76)	6 (10.17)
T1	6+1? (7.00)	1 (1.22)	—
U*	1+3? (4.00)	2+1? (3.66)	1? (1.70)
U2	—	2 (2.44)	—
U3	2 (2.00)	—	1 (1.70)
U4	2 (2.00)	2 (2.44)	—
U5	1 (1.00)	—	—
U5a	2 (2.00)	1 (1.22)	1 (1.70)
U5a1	2 (2.00)	1 (1.22)	1 (1.70)
U5a1a	1 (1.00)	2 (2.44)	1 (1.70)
U5a/b	1 (1.00)	—	—
U5b	1 (1.00)	1 (1.22)	1 (1.70)
U6	1 (1.00)	—	—
U6a	4 (4.00)	—	—
U6b	2 (2.00)	—	—
U7	—	—	1 (1.70)
V	8 (8.00)	3+3? (7.32)	4 (6.78)
W	2 (2.00)	1 (1.22)	–
X	–	3 (3.66)	1+1? (3.39)

Note: Haplogroups where the classification presented ambiguity are assigned with a question mark (?).

erence sequence: at position 00263 we found a G, and at position 00311 a length polymorphism with one more C. A recent revision of the CRS sequence (Andrews *et al.* 1999) confirmed the presence of an A in 00263 and 5C in 00311, both allelic states occurring at very low frequency.

Comparisons between North, Central and South Portugal did not reveal statistical differences between the three regions with respect to mtDNA variability (p values of F_{ST} pairwise genetic distances between populations were 0.327, 0.673 and 0.921 for North-Central, North-South and Central-South, respectively). Accordingly, the sequences of the three regions could be merged in a global Portuguese sample, for the analysis that will be presented next.

In order to evaluate if the overall diversity in Portugal was due to a high diversity within particular haplogroups or, alternatively, to a high diversity of haplogroups, after classifying the different sequences into haplogroups we have performed mismatch distribution analysis within each of the major haplogroups. We will only present the results based on HVRI diversity (Figure 2), as when the less informative HVRII region was used (data not shown) all clusters exhibited clear unimodal curves with low modes. In this section we will try to correlate the inferences taken from the analysis of haplogroups mismatch distributions with those made by Richards *et al.* (1998) based on network analysis.

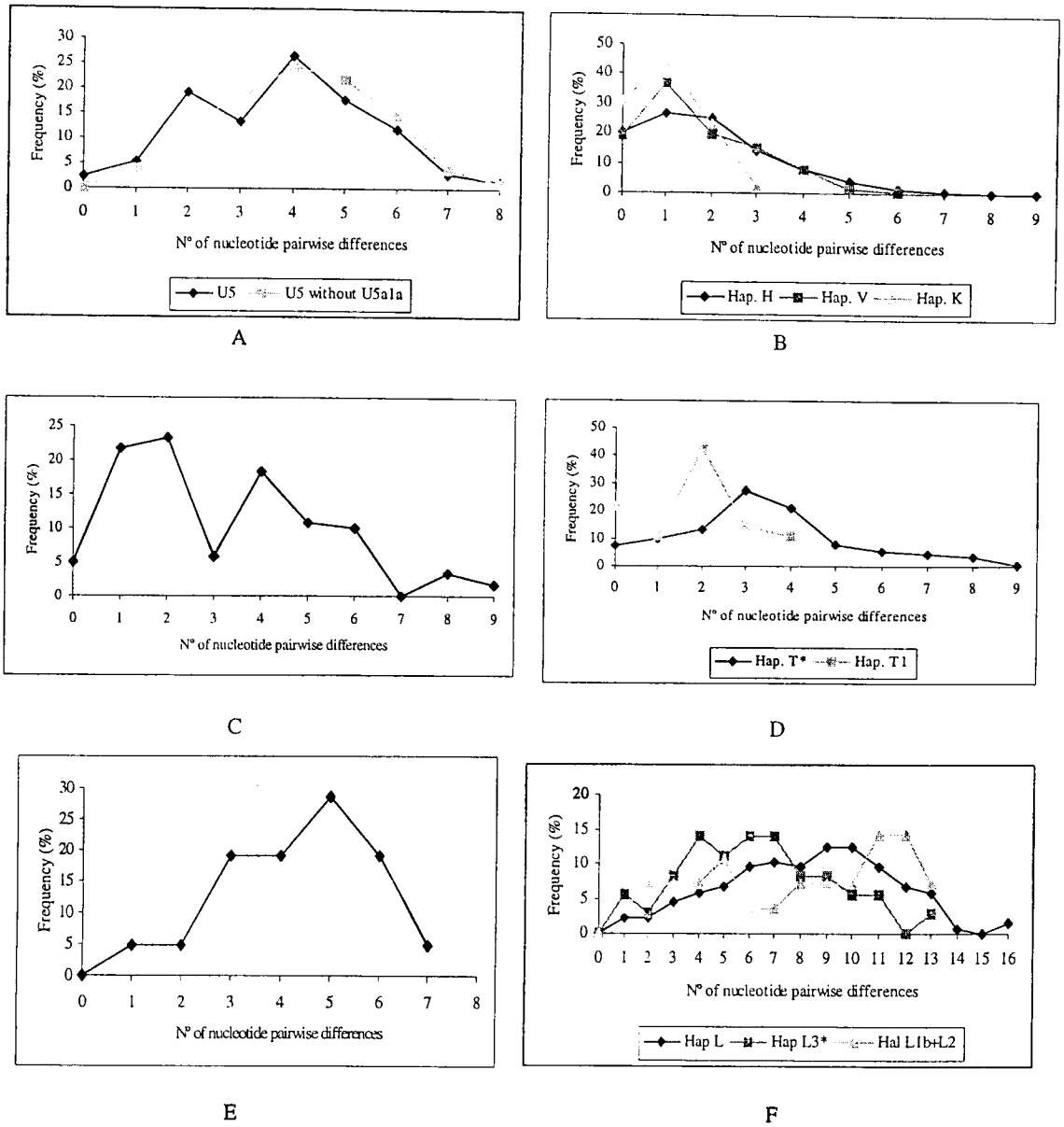


Fig. 2. Mismatch distributions for the haplogroups observed in Portugal (North, Central and South) considering only HVRI diversity. A- haplogroup U5 with and without U5a1a sequences. B- haplogroups H, V and K. C- haplogroup J. D- haplogroups T* and T1. E- haplogroup U6. F- haplogroups L1b + L2, L3* and all simultaneously.

Haplogroup U5 had the highest mode for the number of pairwise differences distribution, and showed a regular unimodal pattern (Figure 2A). Both features are in accordance with its being the oldest haplogroup in Europe which has registered a regional development. As expected, and depicted in Figure 2A, a slight bimodality appeared when the more recent sub-haplogroup U5a1a was considered in the analysis.

The three haplogroups H, V and K, all considered to be post-glacially expanded European haplogroups (Figure 2B), showed clear unimodal distributions but with low means, reflecting the shorter period since the accumulation of variation began.

Haplogroups J and T, said to have had a common origin in the Near East, presented very distinct mismatch distribution patterns. Haplogroup J, which seems to have been recently introduced into Europe during the Neolithic,

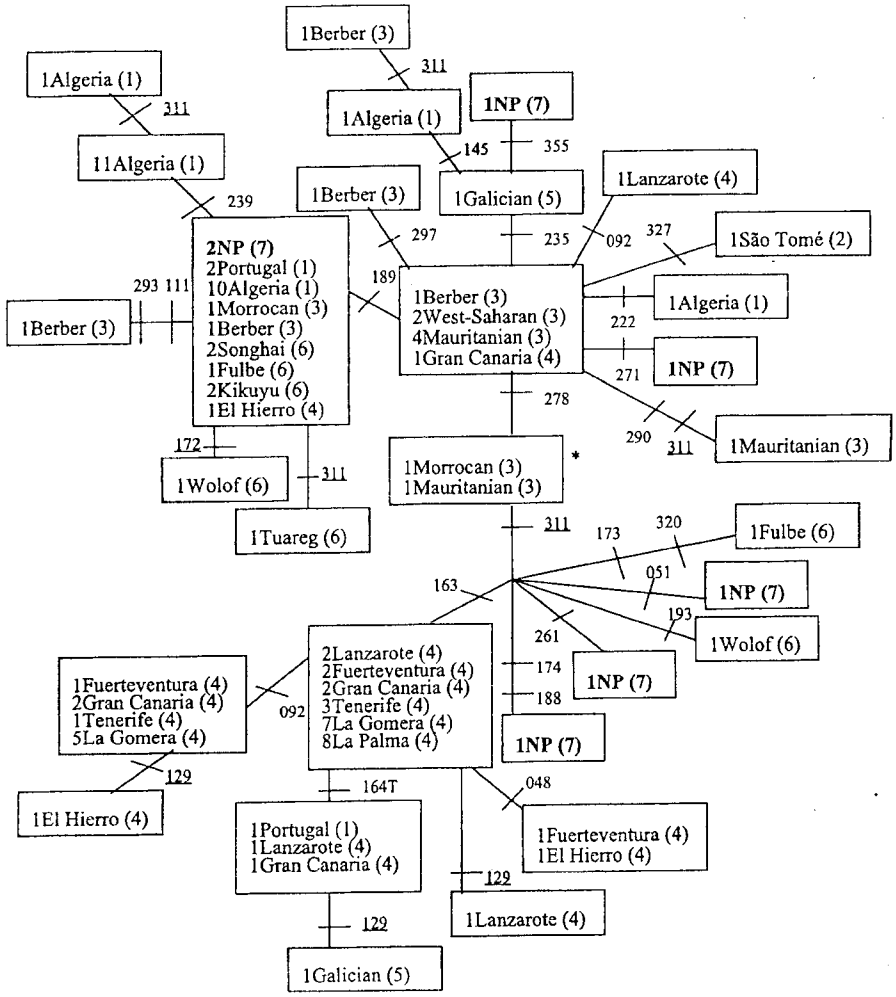


Fig. 3. A most parsimonious tree of sequences belonging to cluster U6. Root motif 172-219 indicated with an asterisk. Branches are labelled by the nucleotide positions in HRVI (minus 16000) to designate transitions; transversions are further specified and positions underlined represent parallel mutations. Numbers in brackets represent bibliographic reference (1) Côrte-Real *et al.* (1996), (2) Mateu *et al.* (1996), (3) Rando *et al.* (1998), (4) Rando *et al.* (1999), (5) Salas *et al.* (1998), Watson *et al.* (1996), (7) this work.

showed a very irregular curve with a ragged shape (Figure 2C). This pattern additionally suggests a high level of diversification of the founder sequences and also expresses the heterogeneity within the haplogroup: some well-defined J sub-haplogroups are distinctively spread through Europe, paralleling the clines observed for Y chromosomal, and some autosomal, markers.

In contrast, haplogroup T presented a unimodal and bell-shaped curve (Figure 2D), which can be explained by a more ancient introduction into Europe with subsequent accumulation, *in loco*, of homogenising mutations, starting from a diversity level that must also have been lower relatively to the initial package of J sequences. When the sub-haplogroup T1 was analysed separately, a slight bimodality and a low mean of pairwise differences were registered. Those features are quite compatible with its more recent introduction, although we cannot exclude some bias due to the low number of sequences considered.

In the Portuguese sample analysed two haplo-

groups, U6 and L, that have been reported as occurring sporadically in other European populations, were detected with comparatively high frequency. Both haplogroups were characterised by high levels of diversity and displayed very irregular mismatch distributions (Figure 2E and F). Moreover, haplogroup U6 was found to be restricted to the North region of the country, whereas the L sequences were spread all over the country.

These haplogroups have been reported to be characteristic of African populations, where their frequency is inversely correlated with the North-South axis: the frequency of U6 is high in North Africa and decreases in a southerly direction, being almost absent south of the equator: the L cluster has an opposite distribution (Rando *et al.* 1998, 1999; Watson *et al.* 1996; Mateu *et al.* 1996).

In Portugal, as well as generally in Iberia, many migration waves from both North and sub-Saharan African populations are well documented. The geographical proximity of North Africa and the Iberian Peninsula certainly afforded many opportunities for mutual population contacts. Among them, we stress the movement of Berbers and Arabs that took place during the very recent Muslim rule of Iberia (from the 8th century to the end of the 15th, in some regions). In addition, many sub-Saharan individuals entered the region during the slave trade period, from its very beginning (middle 15th century) until its total ban in the late 19th century.

As it would be interesting to find out the origin of the L and U6 sequences detected in Portugal, we have tried to compare the motifs of the sequences observed in Portugal with those described in the literature for several populations (Figures 3 and 4). However most of the matches found for the Portuguese sequences were with sequences widely distributed in Africa, and no clear pattern of geographic clustering was detected.

A striking aspect observed for the U6 haplogroup was that 5 out of 7 of the Portuguese sequences were unique to Portugal, not allowing, therefore, any accurate assignment of their geographical origin. The Canarian characteristic sub-haplogroup U6b1 (Rando *et al.* 1999), observed in other Iberian samples, was not detected in the present study.

Admitting that U6 sequences could have been at least partially introduced by Berber people during the Muslim rule of Iberia, it is strange to find them restricted to North Portugal. As a matter of fact, most historical sources document a deeper influence of Berber (as well as Arab) people in Central and particularly South Iberia (as judged from toponyms and general cultural affinities), compared to North Iberia where the Muslim presence is recorded to have been more ephemeral and consequently to have made less cultural and demographic impact. The data does not exclude the possibility that U6 introductions could have been additionally reinforced by later sub-Saharan inputs mediated by the African slave trade. Even if this mixed scenario is plausible, the presence of U6 sequences exclusively in North Portugal is a question that deserves further analysis. The hypothesis of an earlier introduction in the region does not seem to be favoured, neither by its presence in a restricted geographical area, nor by the high level of heterogeneity that characterises the set of sequences that were found among this haplogroup.

With respect to the L sequences, it is widely accepted that they have a sub-Saharan origin, excepting some L3* lineages that, as analysis of Figure 4 suggests, might indeed have a non-African origin. The presence of L sequences in North African regions does not allow us to exclude the possibility that population influxes from this region, namely the above referred Berber/Arab movement, have introduced a significant fraction of L sequences into Iberia. However, it seems more likely that most of the L lineages found nowadays in Portugal have been carried by African slaves, since the country was actively involved in the Transatlantic slave trade. Nine out of 17 L sequences found in this study showed matches with widespread African sequences, and with regard to the 8 remaining

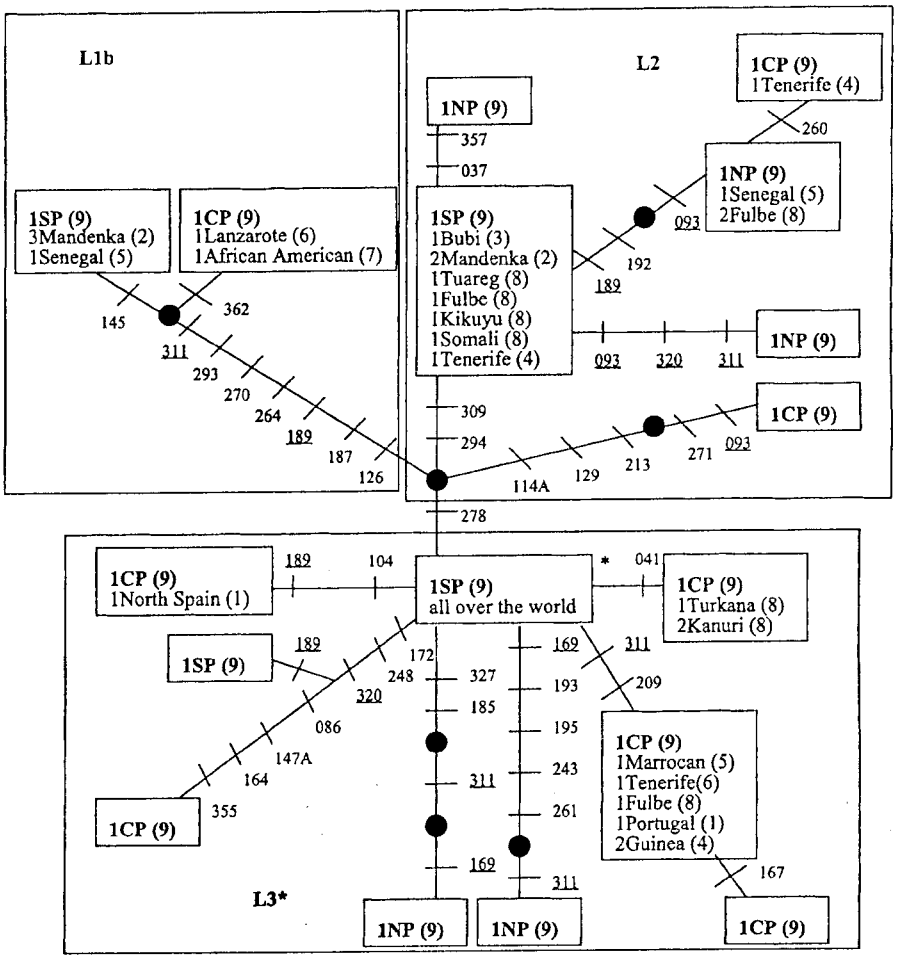


Fig. 4. A phylogeny of Portugese sequences belonging to African clades L1b and L2 and to the default cluster L3* (some members of which may have a non-African origin). The sequence with a transition from the CRS at np 16223 is indicated with an asterisk. Sequence matches in other populations are shown. Numbers in brackets represent bibliographic reference (1) Côrte-Real *et al.* (1996), (2) Graven *et al.* (1995), (3) Mateu *et al.* (1996), (4) Pinto *et al.* (1996), (5) Rando *et al.* (1998), (6) Rando *et al.* (1999), (7) Vigilant *et al.* (1991), (8) Watson *et al.* (1996), (9) this work. Solid circles represent sequences observed in the mtDNA database or branching nodes in the mtDNA phylogeny which aid in the identification of parallel mutations, which are shown with the position underlined.

sequences the absence of matches can be due to the present bias in the description of sub-Saharan mtDNA variability. Broad areas corresponding to Ivory Coast, Angola and Mozambique, which represented very important sources of African slaves, remain uncharacterised.

There were more African slaves in Portugal than in any other European country: in 1550, Lisbon boasted 10000 resident slaves in a population of 100000, and Portugal as a whole probably had over 40000 (Thomas, 1998). In the mid-sixteenth century the birth of slaves' children was stimulated in Portugal for internal traffic purposes. Inter-breeding between autochthonous individuals and African slaves certainly occurred and the predominant mating must have been between slave African females and autochthonous males, due to social pressures and also for legal reasons: offspring of slave females would be slaves, whereas offspring of slave males would not. Therefore, breeding between slave African males and white females, besides being socially repressed, would not bring any economic profit. If the pattern of genetic admixture was markedly sex influenced, the signature of this recent African influence would be expected to be very

different in the maternally inherited gene pool and in the paternally inherited one. In a recent study based on Y chromosome biallelic markers (Pereira *et al.* 2000) we have reported the absence of typical sub-Saharan haplogroups in the Y chromosome Portuguese pool. This finding, and the detection of L sequences at 7.1% in the mitochondrial pool, both seem to support the above-mentioned pattern of admixture with African slaves.

CONCLUSIONS

Studies of large population samples, designed to characterise the molecular diversity in restricted geographical contexts, can produce valuable insights concerning specific demographic features that would remain undetectable in broader scale surveys. In this work we have studied mtDNA variability in Portugal, considering North, Central and South regions as micro-screening sample units.

The level of mtDNA diversity found, although characteristic of European populations, is high when the westernmost location of the country in Europe, and the reported European tendency for reduction of diversity toward Western Iberia (Corte-Real *et al.* 1996; Salas *et al.* 1998), are considered. The observed HVRI and HVRI + HVRII mismatch distributions were unimodal but smoother than others previously found in neighbouring populations.

This finding, as well as the high level of haplogroup diversity, suggests the influence of specific demographic factors acting in the Portuguese population, and led us to hypothesise that an important modulator of the present Portuguese mtDNA variability could have been the influx of distinct mtDNA lineages at historically quite different times.

Sharing the features of mtDNA diversity generally registered in Europeans (all European haplogroups were detected), Portugal has in addition received significant North and sub-Saharan African influences. Frequencies of haplogroups specific to these regions were higher than those reported for other European populations: 7% of North African sequences were detected (restricted to North Portugal and representing almost 3% of the total sample), and sub-Saharan African sequences were found to be spread throughout the country, with frequencies between 5% and 9.8%. Although statistically significant differences were not detected between the three sub-samples considered, the geographic distribution pattern observed for U6 and L sequences strongly suggest that different population movements were responsible for their introduction into the country, although none of them had enough demographic impact to induce regional differentiation.

The introduction of L sequences in Portugal was tentatively imputed mainly to the modern slave trade that occurred between the 15th and 19th centuries. Both the great number of slaves that entered Portugal and their very diverse African geographic origin are consistent with the data set now reported. However, we cannot exclude some North-African contribution to present-day Portuguese L lineages.

While the population movement associated with the slave trade may be responsible by some U6 inputs, we suggest that U6 sequences were predominantly introduced into Portugal during the Berber/Arab invasion of the Peninsula. However, the observation that haplogroup U6 is restricted to North Portugal is puzzling, considering the more pronounced impact of the Muslin rule in south Iberia and the widespread presence of African slaves throughout the country, and deserves further investigation.

We are deeply in debt to Vincent Macaulay, whose critical discussion of several aspects of this paper was very important for its improvement. We also thank Martin Richards for his assistance in the classification of some sequences. This work was partially supported by a grant (PRAXIS BD/13632/97) financed by Fundação para a Ciência e a Tecnologia.

REFERENCES

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., *et al.* (1981). Sequence and organisation of the human mitochondrial genome. *Nature* **290**, 457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999). Reanalysis and revision of the Cambridge

reference sequence for human mitochondrial DNA. *Nature Genet.* **23**, 147.

Arnaiz-Villena, A., Martínez-Laso, J., Gómez-Casado, E., Díaz-Campos, N., Santos, P., Martinho, A., Breda-Coimbra, H. (1997). Relatedness among Basques, Portuguese, Spaniards and Algerians studied by HLA allelic frequencies and haplotypes. *Immunogenetics* **47**, 37–43.

Avise, J., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E. *et al.* (1987). Intraspecific phylogeography: the molecular bridge between population genetics and systematics. *Ann. Ver. Ecol. Syst.* **18**, 489–522.

Bertranpetit, J., Sala, I., Calafell, F., Underhill, P., Moral, P. & Comas, D. (1995). Human mitochondrial DNA variation and the origin of the Basques. *Ann. Hum. Genet.* **59**, 63–81.

Calafell, F., Underhill, P., Tolun, A., Aangelicheva, D. & Kaladjieva, L. (1996). From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann. Hum. Genet.* **60**, 35–49.

Comas, D., Calafell, F., Mateu, E., Pérez-Lezaun, A., Bosch, E., Martínez-Arias, R. *et al.* (1998). Trading genes along the Silk Road: mtDNA sequences and the origin of Central Asian populations. *Am. J. Hum. Genet.* **63**, 1824–1838.

Comas, D., Calafell, F., Mateu, E., Pérez-Lezaun, A. & Bertranpetit, J. (1996). Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European population. *Mol. Biol. Evol.* **13**, 1067–1077.

Côrte-Real, H., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A. *et al.* (1996). Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.* **60**, 331–350.

Di Rienzo, A. & Wilson, A. C. (1991). Branching pattern in the evolutionary tree for the human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **88**, 1597–1601.

Forster, P., Harding, R., Torroni, A. & Bandelt, H. J. (1996). Origin and evolution of native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59**, 935–945.

Francalacci, P., Bertranpetit, J., Calafell, F. & Underhill, P. A. (1996). Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. *Am. J. Phys. Anthrop.* **100**, 443–460.

Graven, L., Passarino, G., Semino, O., Boursot, P., Santachiara-Benerecetti, S., Langaney, A. & Excoffier, L. (1995). Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Sengalese Mandenka sample. Mol Biol Evol. **12**, 334–45.

Lareu, M. V., Phillips, C. P., Carracedo, A., Lincoln, A. J., Syndercombe-Court, D. & Thomson, J. A. (1994). Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats; UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci. Int.* **66**, 41–52.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V. *et al.* (1999). The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64**, 232–249.

Marjoram, P. & Donnelly, P. (1994). Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics.* **136**, 673–683.

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

Mellars, P. (1998). The Upper Paleolithic Revolution. In *Prehistoric Europe: an illustrated history* (ed. B. W. Cunliffe). Oxford: Oxford University Press.

Meyer, S., Weiss, G. & von Haeseler, A. (1999). Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* **152**, 1103–1110.

Pereira, L., Brión, M., Prata, M. J., Jobling, M. A., Carracedo, A. & Amorim, A. (2000). Gradient of Y chromosome haplogroup 21 across the Western Iberia. *Progress in Forensic Genetics* **8**, 281–283.

Piercy, R., Sullivan, K. M., Benson, N. & Gill, P. (1996). The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *Int. J. Leg. Med.* **106**, 85–90.

Pinto, F., González, A. M., Hernández, M., Larruga, J. M. & Cabrera, V. M. (1996). Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann. Hum. Genet.* **60**, 321–330.

Rando, J. C., Cabrera, V. M. Larruga, J. M., Hernández, M., González, A. M., Pinto, F., & Bandelt, H.-J. (1999). Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.* **63**, 413–428.

Rando, J. C., Pinto, F., González, A. M., Hernández, M., Larruga, J. M., Cabrera, V. M. & Bandelt, H.-J. (1998). Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. *Ann. Hum. Genet.* **62**, 531–550.

Richards, M. B., Macaulay, V. A., Bandelt, H.-J. & Sykes, B. C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Ann. Hum. Genet.* **62**, 241–260.

Richards, M. B., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A. *et al.* (1996). Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59**, 185–203.

Rogers, A. R. & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569.

Salas, A., Comas, D., Lareu, M. V., Bertranpetit, J. & Carracedo, A. (1998). mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur. J. Hum. Genet.* **6**, 365–375.

Salas, A., Lareu, M. V., Sánchez-Diz, P., Calafell, F. & Carracedo, A. (2000). mtDNA hypervariable region II (HVII) sequences in human evolution studies: impact of mutation rate heterogeneity. *Progress in Forensic Genetics* **8**, 329–331.

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

Thomas, H. (1998). *The slave trade – the history of the Atlantic slave trade 1440–1870*. London: Macmillan Publishers Ltd.

Torroni, A., Bandelt, H.-J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., *et al.* (1998). MtDNA analysis reveals a major late Palaeolithic population expansion from southwestern to northeastern European populations. *Genetics* **62**, 1137–1152.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R. *et al.* (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850.

Vigilant, L., Stoneking, M., Harpending, H. Hawkes, K., Wilson, A. C. (1991) African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–7.

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

APPENDIX

Haplotypes and their geographical distribution in North (N), Central (C) and South (S) Portugal.

HVRI	HVRII	N	C	S	Hap
051 162	73 263 311.1	—	1	—	H
051 257	152 263 303.1 311.1	1	—	—	H
075 183$^{A/C}$ 189 249 356	263 303.3 311.1	—	—	1	H
092 129 239	73 263 311.1	—	—	1	H?
093 126	199 263 303.1 311.1	1	1	—	H?
093 213 215 263$^{T/A}$	263 303.1 311.1	1	—	—	H
093 263	263 311.1	—	1	—	H
124	146 263 303.1 311.1	—	1	—	H
124 256	146 263 303.1 311.1	—	1	—	H
129	146 263 311.1	1	—	—	H
129	152 263 303.1 311.1	—	1	—	H
162 209	73 263 311.1	—	—	1	H
162 209 293	73 263 303.1 311.1	—	—	1	H
163	263 311.1	—	—	1	H
172	263 311.1	1	—	—	H
176	195 204 263 311.1	1	—	—	H
176 218	200 251 263 303.1 311.1	1	—	—	H
180 278	263 303.1 311.1	1	—	—	H
183$^{A/C}$ 189 356 362	263 311.1	1	1	—	H
183$^{A/C}$ 189	263 303.2 311	—	—	2	H
184	146 263 303.2 311.1	—	1	—	H
189	263 303.1 311.1	—	1	—	H
192 274 362	239 263 303.1 311.1	1	—	—	H
192 274 362	152 239 263 303.1 311.1	1	—	—	H
192	263 303.1 311.1	—	—	1	H
209	263 303.1 311.1	1	1	—	H
209 304	263 311.1	—	1	—	H
218 299	263 303.1 311.1	1	—	—	H
248	257 263 303.1 311.1	1	—	—	H
248	257 263 303.2 311.1	—	1	—	H
259	195 263 303.1 311.1	1	—	—	H
260	263 303.1 311.1	—	1	—	H
261	93 263 311.1	—	1	—	H
265$^{A/C}$	263 303.1 311.1	1	1	—	H
269 270	263 311.1	—	—	1	H
272 304	263 303.1 311.1	1	—	—	H
274 294$^{C/G}$	152 263 311.1	1	1	—	H
274	263 303.1 311.1	—	1	—	H
293 311	195 263 303.1 311.1	—1	—	H	
304	263 311.1	1	—	—	H
304 327	263 311.1	1	—	—	H
311	146 195 263 311.1	—	1	—	H
320$^{C/A}$	263 311.1	3	—	—	H
335	263 311.1	—	—	1	H
344	93 263 303.2 311	—	—	1	H
362	150 239 263 311.1	—	1	—	H
362	263 303.1 311.1	—	—	1	H
CRS	263 311.1	11	3	2	H
CRS	263 303.1 311.1	1	3	3	H

HVRI	HVRII	N	C	S	Hap
CRS	152 263 303.1 311.1	1	1	1	H
CRS	151 152 263 311.1	1	—	—	H
CRS	152 263 311.1	1	2	—	H
CRS	263 269$^{C/A}$ 303.2 311.1	1	—	—	H
CRS	185 263 303.1 311.1	1	—	—	H
CRS	195 257 263 303.2 311.1	1	—	—	H
CRS	263 303.2 311.1	—	1	2	H
CRS	146 263 303.1 311.1	—	1	—	H
CRS	150 263 311.1	—	1	1	H
CRS	151 262 263 303.2 311.1	—	—	1	H
CRS	151 263 303.2 311.1	—	—	1	H
CRS	150 263 303.1 311.1	—	—	1	H
CRS	195 257 263 303.1 311.1	—	—	1	H
CRS	263 303.2 311.1 338	—	—	1	H
129 223 278 311 (391)	73 199 204 250 263 303.2 311.1	1	—	—	I
129 172 223 311	73 199 203 204 250 263 311.1	—	—	1	I
063 069 126	73 228 263 295 311.1	—	1	—	J*
069 126 172	73 228 263 295 311.1	1	—	—	J*
069 126 286	73 185 228 263 295 303.1 311.1	1	—	—	J*
069 126	73 185 263 295 303.1 311.1	—	1	—	J*
069 126 311	73 185 263 295 303.1 311.1	—	1	—	J*
069 126 311	73 228 263 295 303.1 311.1	—	1	—	J*
069 126	73 185 207 228 263 295 311.1	—	—	1	J*
069 126	73 146 185 188 222 228 263 295 311.1	—	—	1	J*
069 126 319	73 185 228 263 295 303.1 311.1	—	—	1	J*
069 126 261	73 146 185 228 263 295 303.1 311.1	1	—	—	J1
069 126 261	73 146 185 228 263 295 311.1	1	—	—	J1
069 126 145 222 235 261 271	73 263 295 311.1	—	—	1	J1b
069 126 145 222 256 261 278	73 199 263 295 311.1	—	—	1	J1b
069 126 193 319 360	73 150 152 263 295 303.1 311.1	1	1	—	J2
069 126 193 319 360 362	73 150 152 263 295 303.1 311.1	1	1	—	J2
093 224 311	73 195 263 303.1 311.1	1	—	—	K
093 189 224 311	73 195 263 311.1	—	1	—	K
093 224 311	73 152 263 311.1	—	1	—	K
093 224 290 311	73 263 303.1 311.1	—	—	1	K
093 224 311	73 150 195 263 303.1 311.1	—	—	1	K
224 311	73 263 303.1 311.1	1	1	—	K
224 311	73 263 303.2 311.1	1	—	1	K
224 256 311	73 263 303.1 311.1	—	1	—	K
224 311	73 146 152 263 311.1	—	1	—	K
224 311	73 146 263 311.1	—	1	—	K
224 311	73 195 263 303.1 311.1	—	—	1	K
126 187 189 223 264 270 278 293 311 362	73 152 182 185$^{G/T}$ 195 247 263 303.1 311.1	—	1	—	L1b
126 145 187 189 223 264 270 278 293 311	73 152 182 185$^{G/T}$ 195 247 263 311.1 357	—	—	1	L1b
037 223 278 294 309 357 (390)	73 143 146 152 195 263 311.1	1	—	—	L2
093 189 192 223 278 294 309 (390)	73 143 146 152 195 263 303.1 311.1	1	—	—	L2
093 223 278 294 309 311 320 (390)	73 143 146 152 195 263 311.1	1	—	—	L2
093 114$^{C/A}$ 129 213 223 271 278	73 146 150 152 182 195 198 207 263 311.1	—	1	—	L2
093 189 192 223 260 278 294 309	73 143 146 152 195 207 263 303.1 311.1	—	1	—	L2
223 278 294 309	73 143 146 152 195 263 311.1	—	—	1	L2
041 223	73 150 263 311.1	—	1	—	L3*
086 147$^{C/A}$ 164 172 223 248 320 355	73 152 199 204 207 263 311.1	—	1	—	L3*
104 183$^{A/C}$ 189 223	73 263 311.1	—	1	—	L3*
167 209 223 311	73 189 200 263 311.1	—	1	—	L3*
169 185 223 311 327	73 150 185 189 200 263 311.1	1	—	—	L3*
169 193 195 223 243 261 311	73 150 200 235 249^{delA} 263 303.1 311.1	1	—	—	L3*
172 182$^{A/C}$ 183$^{A/C}$ 189 223 248 320	73 150 195 263 311.1	—	—	1	L3*
209 223 311	73 189 200 263 311.1	—	1	—	L3*

HVRI	HVRII	N	C	S	Hap
223	73 150 195 263 303.1 311.1	—	—	1	L3*
129 183$^{A/C}$ 189 223 249 311	73 195 263 303.1 311.1	—	1	—	M1
037 126 186 189 222	73 152 263 303.1 311.1	1	—	—	T*?
051 126 294 296 304	73 151 204 263 303.1 311.1	1	—	—	T*
093 126 271 294 296 304	73 151 263 303.1 311.1	—	1	—	T*
114 126 153 192 294	73 150 263 303.1 311.1	—	1	—	T*
126 153 294	73 150 263 311.1	1	—	—	T*
126 192 294 296 304	73 151 263 311.1	1	—	—	T*
126 256 294 296	73 152 263 303.1 311.1	—	1	—	T*
126 260 294 296 319	73 263 311.1	—	1	—	T*
126 292 294	73 263 303.1 311.1	—	1	1	T*
126 294 296 304	73 263 311.1	—	1	1	T*
126 294 296 304	73 195 263 311.1	—	1	—	T*
126 294 304	73 152 263 311.1	—	1	—	T*
126 153 189 294 296	73 150 263 303.1 311.1	—	—	1	T*
126 218 294 296 324	73 263 311.1	—	—	1	T*
126 294 296 304	73 151 263 311.1	—	—	1	T*
126 294 296 304	73 151 260 263 303.1 311.1	—	—	1	T*
037 126 163 186 189	73 152 263 303.1 311.1	1	—	—	T1?
126 163 186 187 189 294	73 152 195 263 303.1 311.1	1	—	—	T1
126 163 186 189 294	73 195 263 303.1 311.1	3	—	—	T1
126 163 186 189 249 294 311	73 263 303.1 311.1	2	1	—	T1
CRS	73 263 311.1	1	—	—	U*?
CRS	73 263 303.1 311.1	1	—	—	U*?
CRS	73 152 263 303.1 311.1	1	—	—	U*?
142$^{C/A}$ 311	73 263 311.1	—	1	—	U*
179	73 195 263 303.1 311.1	—	1	—	U*
184 264 291	73 263 303.1 311.1 316	1	—	—	U*
189	73 263 303.1 311.1	—	1	—	U*?
189	73 263 311.1	—	—	1	U*?
051 092 129$^{G/C}$ 174 183$^{A/C}$ 189 362	73 152 217 263 311.1	—	1	—	U2
051 129$^{G/C}$ 189 256 311	73 152 217 263 311.1 340	—	1	—	U2
343	73 150 263 311.1	1	—	—	U3
343 356 (390)	73 150 263 311.1	1	—	—	U3
343 356	73 150 195 263 303.1 311.1	—	—	—	U3
134 288 356	73 152 195 263 311.1	—	1	—	U4
179 356	73 150 195 263 311.1	1	—	—	U4
179 356	73 195 263 303.1 311.1	1	1	—	U4
224 270	73 150 199 263 279 311.1	1	—	—	U5
167 192 270 311 318 356	73 150 263 303.1 311.1	1	—	—	U5a
167 192 270 311 356	73 150 263 303.1 311.1	—	1	—	U5a
192 235 270 304	73 146 150 263 311.1	1	—	—	U5a
192 270	73 150 151 228 263 303.2 311.1	—	—	1	U5a
114$^{C/A}$ 192 270 294	73 263 303.1 311.1	1	—	—	U5a1
189 192 256 270 362	73 195 263 303.1 311.1	—	1	—	U5a1
192 256 270 291 (399)	73 263 303.1 311.1	1	—	—	U5a1
192 256 270	73 263 303.2 311.1	—	—	1	U5a1
189 256 270 362	73 185 204 263 303.1 311.1	1	—	—	U5a1a
256 270	73 263 311.1	—	2	—	U5a1a
256 270	73 146 263 284$^{A/C}$ 311.1	—	—	1	U5a1a
183$^{A/C}$ 187 189 192 270	73 150 195 200 263 311.1	1	—	—	U5a/b
093 111 189 270	73 150 263 311.1	—	—	1	U5b
114 189 192 270	73 140 150 263 311.1	—	1	—	U5b
189 270	73 150 263 311.1	1	—	—	U5b
051 172 219 311	73 263 311.1	1	—	—	U6
172 182$^{A/C}$ 183$^{A/C}$ 189 219 278	73 263 303.1 311.1	1	—	—	U6a
172 183$^{A/C}$ 189 219 278	73 263 303.1 311.1	1	—	—	U6a
172 219 235 278 355	73 146 263 303.1 311.1	1	—	—	U6a
172 219 271 278	73 152 263 303.1 311.1	1	—	—	U6a
172 174 188 219 311	73 263 311.1	1	—	—	U6b
172 219 261 311	73 263 303.1 311.1	1	—	—	U6b
309 318$^{A/T}$	73 152 263 303.2 311.1	—	—	1	U7
124 298 311 319	(72) 263 311.1	—	—	1	V
129 298	195 263 311.1	1	—	—	V
183$^{A/C}$ 189 259$^{C/G}$ 298	263 303.2 311.1	—	—	1	V

HVRI	HVRII	N	C	S	Hap
189 298	(72) 263 311.1	—	—	1	V
242	(72) 152 263 303.1 311.1	—	1	—	V?
254 298	(72) 263 303.1 311.1	—	1	—	V
264 298	195 263 303.1 311.1	0	—	—	V
264 298	(72) 195 227 263 311.1	—	—	1	V
298	(72) 131 263 303.2 311.1	1	—	—	V
298	(72) 263 303.1 311.1	2	1	—	V
298	(72) 263 311.1	1	—	—	V
298	(72) 195 263 303.1 311.1	1	1	—	V
298 311	(72) 263 311.1	1	—	—	V
298	73 263 311.1	—	1	—	V?
298 344	73 263 311.1	—	1	—	V?
192 223 292 325	73 189 194 195 204 207 263 303.1 311.1	—	1	—	W
223 292 311	73 189 195 204 207 263 311.1	1	—	—	W
223 292 362	73 189 194 195 204 207 263 303.1 311.1	1	—	—	W
048 189 223 255 278	73 146 153 195 225 263 303.1 311.1	—	—	1	X
172 183$^{A/C}$ 189 278	73 146 152 185 263 303.1 311.1	—	—	1	X?
183$^{A/C}$ 189 223 260 278	73 153 195 225 226 263 303.1 311.1	—	1	—	X
189 223 255 278	73 146 153 172$^{T/G}$ 195 225 226 263 303.1 311.1	—	1	—	X
189 223 255 278	73 146 153 195 225 226 263 311.1	—	1	—	X

Variant positions from the Cambridge Reference Sequence (CRS) of Anderson *et al.* (1981) are shown (minus 16000 in HVRI). Transversions are further specified by the appropriate base change. Haplogroups where the classification presents ambiguity are assigned with a question mark (?). In the case of positions 303 and 311, the presence of one, two or three Cs is referred by .1, .2 and .3, respectively, following the base position. In some cases, additional positions outside the referred regions are indicated inside brackets.

ERRATA

APPENDIX

HVRI	HVRII	N	C	S	Hap
051 162	73 263 311.1	-	1	-	H
051 257	152 263 303.1 311.1	1	-	-	H
075 183$^{A/C}$ 189 249 356	263 303.3 311.1	-	-	1	H
092 129 239	73 263 311.1	-	-	1	H?
093 126	199 263 303.1 311.1	1	-	-	H?
093 213 215 263$^{T/A}$	263 303.1 311.1	1	-	-	H
093 263	263 311.1	-	1	-	H
124	146 263 303.1 311.1	-	1	-	H
124 256	146 263 303.1 311.1	-	1	-	H
129	146 263 311.1	1	-	-	H
129	152 263 303.1 311.1	-	1	-	H
162 209	73 263 311.1	-	-	1	H
162 209 293	73 263 303.1 311.1	-	-	1	H
163	263 311.1	-	-	1	H
172	263 311.1	1	-	-	H
176	195 204 263 311.1	1	-	-	H
176 218	200 251 263 303.1 311.1	1	-	-	H
180 278	263 303.1 311.1	1	-	-	H
183$^{A/C}$ 189 356 362	263 311.1	1	1	-	H
183$^{A/C}$ 189	263 303.2 311	-	-	2	H
184	146 263 303.2 311.1	-	1	-	H
189	263 303.1 311.1	-	1	-	H
192 274 362	239 263 303.1 311.1	1	-	-	H
192 274 362	152 239 263 303.1 311.1	1	-	-	H
192	263 303.1 311.1	-	-	1	H
209	263 303.1 311.1	1	1	-	H
209 304	263 311.1	-	1	-	H
218 299	263 303.1 311.1	1	-	-	H
248	257 263 303.1 311.1	1	-	-	H
248	257 263 303.2 311.1	-	1	-	H
259	195 263 303.1 311.1	1	-	-	H
260	263 303.1 311.1	-	1	-	H
261	93 263 311.1	-	1	-	H
265$^{A/C}$	263 303.1 311.1	1	1	-	H
269 270	263 311.1	-	-	1	H
272 304	263 303.1 311.1	1	-	-	H
274 294$^{C/G}$	152 263 311.1	1	1	-	H
274	263 303.1 311.1	-	1	-	H
293 311	195 263 303.1 311.1	-	1	-	H
304	263 311.1	1	-	-	H
304 327	263 311.1	1	-	-	H
311	146 195 263 311.1	-	1	-	H
320$^{C/A}$	263 311.1	3	-	-	H
335	263 311.1	-	-	1	H

HVRI	HVRII	N	C	S	Hap
344	93 263 303.2 311	-	-	1	H
362	150 239 263 311.1	-	1	-	H
362	263 303.1 311.1	-	-	1	H
CRS	263 311.1	11	3	2	H
CRS	263 303.1 311.1	1	3	3	H
CRS	152 263 303.1 311.1	1	1	1	H
CRS	151 152 263 311.1	1	-	-	H
CRS	152 263 311.1	1	2	-	H
CRS	263 $269^{C/A}$ 303.2 311.1	1	-	-	H
CRS	185 263 303.1 311.1	1	-	-	H
CRS	195 257 263 303.2 311.1	1	-	-	H
CRS	263 303.2 311.1	-	1	2	H
CRS	146 263 303.1 311.1	-	1	-	H
CRS	150 263 311.1	-	1	1	H
CRS	151 262 263 303.2 311.1	-	-	1	H
CRS	151 263 303.2 311.1	-	-	1	H
CRS	150 263 303.1 311.1	-	-	1	H
CRS	195 257 263 303.1 311.1	-	-	1	H
CRS	263 303.2 311.1 338	-	-	1	H
129 223 278 311 (391)	73 199 204 250 263 303.2 311.1	1	-	-	I
129 172 223 311	73 199 203 204 250 263 311.1	-	-	1	I
063 069 126	73 228 263 295 311.1	-	1	-	J*
069 126 172	73 228 263 295 311.1	1	-	-	J*
069 126 286	73 185 228 263 295 303.1 311.1	1	-	-	J*
069 126	73 185 263 295 303.1 311.1	-	1	-	J*
069 126 311	73 185 263 295 303.1 311.1	-	1	-	J*
069 126 311	73 228 263 295 303.1 311.1	-	1	-	J*
069 126	73 185 207 228 263 295 311.1	-	-	1	J*
069 126	73 146 185 188 222 228 263 295 311.1	-	-	1	J*
069 126 319	73 185 228 263 295 303.1 311.1	-	-	1	J*
069 126 261	73 146 185 228 263 295 303.1 311.1	1	-	-	J1
069 126 261	73 146 185 228 263 295 311.1	1	-	-	J1
069 126 145 222 235 261 271	73 263 295 311.1	-	-	1	J1b
069 126 145 222 256 261 278	73 199 263 295 311.1	-	-	1	J1b
069 126 193 319 360	73 150 152 263 295 303.1 311.1	1	1	-	J2
069 126 193 319 360 362	73 150 152 263 295 303.1 311.1	1	-	-	J2
093 224 311	73 195 263 303.1 311.1	1	-	-	K
093 189 224 311	73 195 263 311.1	-	1	-	K
093 224 311	73 152 263 311.1	-	1	-	K
093 224 290 311	73 263 303.1 311.1	-	-	1	K
093 224 311	73 150 195 263 303.1 311.1	-	-	1	K
224 311	73 263 303.1 311.1	1	1	-	K
224 311	73 263 303.2 311.1	1	-	1	K
224 256 311	73 263 303.1 311.1	-	1	-	K
224 311	73 146 152 263 311.1	-	1	-	K
224 311	73 146 263 311.1	-	1	-	K
224 311	73 195 263 303.1 311.1	-	-	1	K
126 187 189 223 264 270 278 293 311 362	73 152 182 $185^{G/T}$ 195 247 263 303.1 311.1	-	1	-	L1b
126 145 187 189 223 264 270 278 293 311	73 152 182 $185^{G/T}$ 195 247 263 311.1 357	-	-	1	L1b
037 223 278 294 309 357 (390)	73 143 146 152 195 263 311.1	1	-	-	L2
093 189 192 223 278 294 309 (390)	73 143 146 152 195 263 303.1 311.1	1	-	-	L2
093 223 278 294 309 311 320 (390)	73 143 146 152 195 263 311.1	1	-	-	L2
093 $114^{C/A}$ 129 213 223 271 278	73 146 150 152 182 195 198 207 263 311.1	-	1	-	L2
093 189 192 223 260 278 294 309	73 143 146 152 195 207 263 303.1 311.1	-	1	-	L2
223 278 294 309	73 143 146 152 195 263 311.1	-	-	1	L2

HVRI	HVRII	N	C	S	Hap
041 223	73 150 263 311.1	-	1	-	L3*
086 147$^{C/A}$ 164 172 223 248 320 355	73 152 199 204 207 263 311.1	-	1	-	L3*
104 183$^{A/C}$ 189 223	73 263 311.1	-	1	-	L3*
167 209 223 311	73 189 200 263 311.1	-	1	-	L3*
169 185 223 311 327	73 150 185 189 200 263 311.1	1	-	-	L3*
169 193 195 223 243 261 311	73 150 200 235 249$^{del\ A}$ 263 303.1 311.1	1	-	-	L3*
172 182$^{A/C}$ 183$^{A/C}$ 189 223 248 320	73 150 195 263 311.1	-	-	1	L3*
209 223 311	73 189 200 263 311.1	-	1	-	L3*
223	73 150 195 263 303.1 311.1	-	-	1	L3*
129 183$^{A/C}$ 189 223 249 311	73 195 263 303.1 311.1	-	1	-	M1
037 126 186 189 222	73 152 263 303.1 311.1	1	-	-	T*?
051 126 294 296 304	73 151 204 263 303.1 311.1	1	-	-	T*
093 126 271 294 296 304	73 151 263 303.1 311.1	-	1	-	T*
114 126 153 192 294	73 150 263 303.1 311.1	-	1	-	T*
126 153 294	73 150 263 311.1	1	-	-	T*
126 192 294 296 304	73 151 263 311.1	1	-	-	T*
126 256 294 296	73 152 263 303.1 311.1	-	1	-	T*
126 260 294 296 319	73 263 311.1	-	1	-	T*
126 292 294	73 263 303.1 311.1	-	1	1	T*
126 294 296 304	73 263 311.1	-	1	1	T*
126 294 296 304	73 195 263 311.1	-	1	-	T*
126 294 304	73 152 263 311.1	-	1	-	T*
126 153 189 294 296	73 150 263 303.1 311.1	-	-	1	T*
126 218 294 296 324	73 263 311.1	-	-	1	T*
126 294 296 304	73 151 263 311.1	-	-	1	T*
126 294 296 304	73 151 260 263 303.1 311.1	-	-	1	T*
037 126 163 186 189	73 152 263 303.1 311.1	1	-	-	T1?
126 163 186 187 189 294	73 152 195 263 303.1 311.1	1	-	-	T1
126 163 186 189 294	73 195 263 303.1 311.1	3	-	-	T1
126 163 186 189 249 294 311	73 263 303.1 311.1	2	1	-	T1
CRS	73 263 311.1	1	-	-	U*?
CRS	73 263 303.1 311.1	1	-	-	U*?
CRS	73 152 263 303.1 311.1	1	-	-	U*?
142$^{C/A}$ 311	73 263 311.1	-	1	-	U*
179	73 195 263 303.1 311.1	-	1	-	U*
184 264 291	73 263 303.1 311.1 316	1	-	-	U*
189	73 263 303.1 311.1	-	1	-	U*?
189	73 263 311.1	-	-	1	U*?
051 092 129$^{G/C}$ 174 183$^{A/C}$ 189 362	73 152 217 263 311.1	-	1	-	U2
051 129$^{G/C}$ 189 256 311	73 152 217 263 311.1 340	-	1	-	U2
343	73 150 263 311.1	1	-	-	U3
343 356 (390)	73 150 263 311.1	1	-	-	U3
343 356	73 150 195 263 303.1 311.1	-	-	1	U3
134 288 356	73 152 195 263 311.1	-	1	-	U4
179 356	73 150 195 263 311.1	1	-	-	U4
179 356	73 195 263 303.1 311.1	1	1	-	U4
224 270	73 150 199 263 279 311.1	1	-	-	U5
167 192 270 311 318 356	73 150 263 303.1 311.1	1	-	-	U5a
167 192 270 311 356	73 150 263 303.1 311.1	-	1	-	U5a
192 235 270 304	73 146 150 263 311.1	1	-	-	U5a
192 270	73 150 151 228 263 303.2 311.1	-	-	1	U5a
114$^{C/A}$ 192 270 294	73 263 303.1 311.1	1	-	-	U5a1
189 192 256 270 362	73 195 263 303.1 311.1	-	1	-	U5a1
192 256 270 291 (399)	73 263 303.1 311.1	1	-	-	U5a1
192 256 270	73 263 303.2 311.1	-	-	1	U5a1

HVRI	HVRII	N	C	S	Hap
189 256 270 362	73 185 204 263 303.1 311.1	1	-	-	U5a1a
256 270	73 263 311.1	-	2	-	U5a1a
256 270	73 146 263 284$^{A/C}$ 311.1	-	-	1	U5a1a
183$^{A/C}$ 187 189 192 270	73 150 195 200 263 311.1	1	-	-	U5a/b
093 111 189 270	73 150 263 311.1	-	-	1	U5b
114 189 192 270	73 140 150 263 311.1	-	1	-	U5b
189 270	73 150 263 311.1	1	-	-	U5b
051 172 219 311	73 263 311.1	1	-	-	U6
172 182$^{A/C}$ 183$^{A/C}$ 189 219 278	73 263 303.1 311.1	1	-	-	U6a
172 183$^{A/C}$ 189 219 278	73 263 303.1 311.1	1	-	-	U6a
172 219 235 278 355	73 146 263 303.1 311.1	1	-	-	U6a
172 219 271 278	73 152 263 303.1 311.1	1	-	-	U6a
172 174 188 219 311	73 263 311.1	1	-	-	U6b
172 219 261 311	73 263 303.1 311.1	1	-	-	U6b
309 318$^{A/T}$	73 152 263 303.2 311.1	-	-	1	U7
124 298 311 319	(72) 263 311.1	-	-	1	V
129 298	195 263 311.1	1	-	-	V
183$^{A/C}$ 189 259$^{C/G}$ 298	263 303.2 311.1	-	-	1	V
189 298	(72) 263 311.1	-	-	1	V
242	(72) 152 263 303.1 311.1	-	1	-	V?
254 298	(72) 263 303.1 311.1	-	1	-	V
264 298	195 263 303.1 311.1	1	-	-	V
264 298	(72) 195 227 263 311.1	-	-	1	V
298	(72) 131 263 303.2 311.1	1	-	-	V
298	(72) 263 303.1 311.1	2	1	-	V
298	(72) 263 311.1	1	-	-	V
298	(72) 195 263 303.1 311.1	1	1	-	V
298 311	(72) 263 311.1	1	-	-	V
298	73 263 311.1	-	1	-	V?
298 344	73 263 311.1	-	1	-	V?
192 223 292 325	73 189 194 195 204 207 263 303.1 311.1	-	1	-	W
223 292 311	73 189 195 204 207 263 311.1	1	-	-	W
223 292 362	73 189 194 195 204 207 263 303.1 311.1	1	-	-	W
048 189 223 255 278	73 146 153 195 225 263 303.1 311.1	-	-	1	X
172 183$^{A/C}$ 189 278	73 146 152 185 263 303.1 311.1	-	-	1	X?
183$^{A/C}$ 189 223 260 278	73 153 195 225 226 263 303.1 311.1	-	1	-	X
189 223 255 278	73 146 153 172$^{T/G}$ 195 225 226 263 303.1 311.1	-	1	-	X
189 223 255 278	73 146 153 195 225 226 263 311.1	-	1	-	X

RESULTS

II - POPULATION GENETICS' APPLICATIONS

B- PORTUGAL - THE IBERIAN CONTEXT

The analysis of Portuguese genetic data in an Iberian context is justified at various levels: (1) the (relative) geographical isolation from the rest of the continent; (2) most of the pre-, proto-, and historic features are not correlated with modern political borders and, last but not least (3) the climatic aspect, since Iberia is thought to have been an important glacial refuge.

All these considerations are concordant in the sense they shape a potential axis of population structuring perpendicular to the east-west axis predominant in the rest of Europe. Islamic rule also followed and strengthened this pattern and special attention has been paid to it when analysing both Y-chromosome and mtDNA.

The two following papers intend to address these questions, in particular regarding genetic gradients and North African connections.

ARTICLE 6

PEREIRA, L., PRATA, M.J., BRIÓN, M., JOBLING, M.A., CARRACEDO, A., AMORIM, A. (2000) Clinal variation of the YAP^+ Y chromosome frequencies in Western Iberia. *Hum. Biol.* **72**: 937-944.

ARTICLE 7

PEREIRA, L., MACAULAY, V., PRATA, M.J., AMORIM, A. (2002) Phylogeny of the mtDNA haplogroup U6. Analysis of the sequences observed in North Africa and Iberia. In: Sensabaugh G.F., Lincoln, P.J., Olaisen, B. (eds) *Progress in Forensic Genetics* 9 (in press). Elsevier Science, Amsterdam.

Clinal Variation of YAP$^+$ Y-Chromosome Frequencies in Western Iberia

LUÍSA PEREIRA,[1,2] MARIA JOÃO PRATA,[1,2] MARIA BRIÓN,[3] MARK A. JOBLING,[4] ANGEL CARRACEDO,[3] AND ANTÓNIO AMORIM[1,2]

Abstract The potential of Y-chromosome biallelic marker haplotypes to infer population affiliations and structures was exploited to analyze four populations from the southwestern edge of Europe, namely north, central, and south Portugal and Galicia. Three markers subdividing the YAP$^+$ lineage were analyzed: the YAP *Alu* element insertion itself and the SRY_{8299} and sY81 base substitutions; these respectively define three haplotypes known as 4, 21, and 8. Only haplotype 21 was detected presenting an increasing north-to-south frequency gradient, from 9.6% (Galicia) to 24.5% (South Portugal). This clinal distribution most likely reflects the genetic input associated with the Neolithic spread of agriculture, but we cannot exclude other movements as potential contributors to the distribution. In this context, it is interesting to note the consistency between the clinal variation and the population movement associated with Islamic rule in Iberia. The absence of haplotype 8, a marker of sub-Saharan populations, suggests that, despite the massive introductions of African slaves in historical times, there was little admixture between the African males and Western Iberian populations.

The increasingly widespread interest in the use of Y-chromosome polymorphisms as a powerful information source for reconstructing the history of human populations, inferring major and local male migration movements and patterns, or, in the forensic field, for the establishment of disputed paternal lineages, has been mainly determined by (*a*) the striking features of the Y-chromosome inheritance patterns relative to other nuclear genomic regions and (*b*) the existence of several classes of Y-chromosome polymorphisms characterized by differential mutation rates, which allows the study of evolutionary or historical events over different time scales.

[1] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n°, 4200 Porto, Portugal.
[2] Faculdade de Ciências da Universidade do Porto, Porto, Pr. Gomes Teixeira, 4050 Porto, Portugal.
[3] Instituto de Medicina Legal, Genetica Forense, Santiago de Compostela, Galicia, Spain.
[4] Department of Genetics, University of Leicester, Leicester, United Kingdom.

Human Biology, December 2000, v. 72, no. 6, pp. 937–944.

KEY WORDS: SNP, HAPLOGROUP, FREQUENCY GRADIENT, NEOLITHIC EXPANSION

Y-chromosome biallelic markers are very slowly evolving polymorphisms and consequently less prone to recurrent mutational events when compared to other markers. The distribution patterns of haplotypes defined by biallelic markers tend to be geographically highly clustered (Jobling and Tyler-Smith 1995; Santos and Tyler-Smith 1996; Hammer et al. 1997; Underhill et al. 1997). For some specific markers, marked gene frequency gradients, particularly throughout Europe, have been reported (Jobling and Tyler-Smith 1995; Santoş and Tyler-Smith 1996; Hammer et al. 1997; Underhill et al. 1997; Semino et al. 1996).

In this work, we have used compound haplotypes defined by three Y-biallelic markers (YAP, SRY_{8299}, and sY81) to analyze the levels of intra- and interregional diversity in the western part of the Iberian Peninsula, in particular north, central, and south Portugal, and the northwesternmost Spanish region of Galicia.

Previous studies based on autosomal short tandem repeat (STR) loci (Amorim et al. 1996; Espinheira et al. 1996; Geada et al. 1996; Gusmão et al. 1995, 1997; Pereira et al. 1999a; Santos et al. 1996; Souto et al. 1996, 1998) have not revealed significant genetic heterogeneity among these populations, suggesting that factors such as substructuring, demography, and differential gene flow through migration have not played a major role in shaping genetic diversity within this region. The results obtained here point to a different picture, revealing an increasing north-to-south frequency gradient for haplotype 21, which follows a pattern in accordance with previously reported data from other European regions (Hammer et al. 1997).

This finding illustrates how the analysis of Y-chromosome markers can provide alternative insights into the structure and/or history of human populations, compared to their autosomal counterparts.

Materials and Methods

Unrelated males from four regions were analyzed: north Portugal, $n = 330$; central Portugal, $n = 118$; south Portugal, $n = 49$; and Galicia, $n = 104$.

An *Alu* insertion, YAP (Hammer and Horai 1995), and two single nucleotide polymorphisms (SNPs), SRY_{8299} (formerly SRY_{4064}; Whitfield et al. 1995) and sY81 (Seielstad et al. 1994), were typed.

DNA extraction was performed by using the resin Chelex 100, according to Lareu et al. (1994). Primers, polymerase chain reaction amplification conditions, and, when necessary, restriction enzyme treatment were as specified in the above-mentioned references. Amplified samples for YAP and restricted samples for SNPs were run on polyacrylamide gels (T9%; C5%) and visualized by silver staining (Budowle et al. 1991).

The three biallelic markers YAP, SRY_{8299}, and sY81 allow discrimination between a background haplotype (0 0 0) and haplotype 4 (1 0 0), 21 (1 1 0), and 8 (1 1 1); see Figure 1. Binary nomenclature indicates ancestral

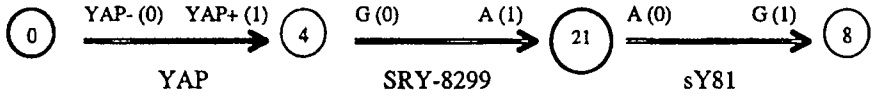


Figure 1. Phylogenetic relationships of haplotypes defined by the Y-chromosome biallelic markers YAP, SRY_{8299}, and sY81. The haplotypes observed in this study are in bold circles.

(0) and derived (1) forms of the polymorphism. Haplotype nomenclature is in accordance with Jobling et al. (1997) for haplotypes 4 and 8, and with M.A. Jobling and C. Tyler-Smith (unpublished observations) for haplotype 21. The background haplotype (0 0 0), which corresponds to a sum of different haplotypes discriminated by other SNPs not studied here, will be referred as haplotype 0.

Analysis of molecular variance (AMOVA) and the population pairwise differentiation test based upon F_{ST} were performed using the software Arlequin (Schneider et al. 1997).

Results and Discussion

In the four population samples analyzed, only two out of the four previously defined haplotypes (Figure 1) were observed, haplotypes 4 and 8 being absent; the frequency distributions of haplotypes 0 and 21 are graphically represented in Figure 2.

As can be seen, the populations studied exhibit a marked increasing north-to-south frequency gradient for haplotype 21. The frequency values for this haplotype were: 9.6% in Galicia, 10.6% in north Portugal, 16.1% in central Portugal, and 24.5% in south Portugal, which corresponds to an overall frequency increase of about 15%.

Haplotype 21 is defined by the presence of the *Alu* insertion at the YAP locus plus the derived state of SRY_{8299}. For this haplotype, decreasing south-to-north frequency gradients have been observed throughout Europe (Hammer et al. 1997).

Hammer et al. (1997) extensively surveyed the geographical variation of the YAP locus, performing compound haplotypic analysis with other Y-polymorphic sites, besides sY81 (designated *DYS271*). This allowed the definition of different YAP^+ haplotypes, haplotype 4 being markedly prevalent among European populations. Since this haplotype exhibited the highest frequency in North African populations and a decreasing frequency gradient towards northwest Europe, the authors pointed out that the pattern could be compatible with multiple human migrations out of Africa. Later, Hammer et al. (1998), on the basis of nested cladistic analysis, argued that the demic diffusion associated with the Neolithic spread of agriculture was the most likely explanation.

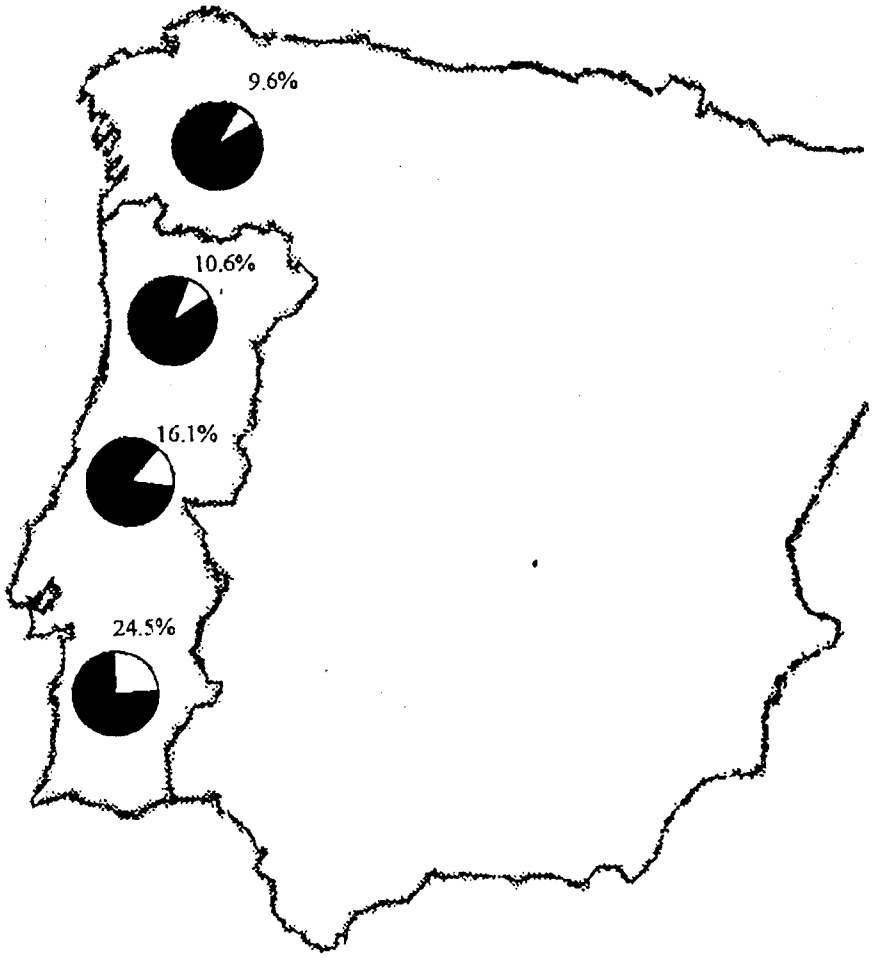


Figure 2. Frequency (%) distribution of haplotype 21 (white) in Galicia (n = 104), north (n = 330), central (n = 118), and south (n = 49) Portugal, from top to bottom, respectively.

We did not characterize YAP^+ chromosomes in a way that allowed us to distinguish haplotypes 3 and 4 of Hammer et al (1998). However, since haplotype 3 is almost absent from European populations, we can admit that the YAP^+ chromosomes detected must likely belong to haplotype 4 of Hammer et al. (1998). Thus, the pattern seen here for haplotype 21 may also reflect the spread of Neolithic farmers, which corresponds to a population movement with enough demographic and genetic impact to produce the clinal variation we now observe. However, we can not exclude more ancient or recent episodes of North African genetic influence in the Iberian gene pool, since other

waves of Mediterranean, including North African, migrations (namely Berbers) into Iberia have been historically documented (Mattoso 1994).

The importance of the Islamic influence in western Iberia, which started in the beginning of the 8th century A.D., was markedly heterogeneous. While northern regions remained almost untouched, in the central and especially the southern regions Islamic administration lasted up to the 13th century. This pattern clearly resembles the clinal variation registered for haplotype 21, and the question deserves further attention.

Another aspect of our data was the complete absence of haplotype 8 from any of the Iberian populations analyzed. Published data show this haplotype to be almost African-specific (Seielstad et al. 1994; Santos and Tyler-Smith 1996). This observation was recently confirmed by the distribution of the equivalent haplotype 5 of Hammer et al. (1997, 1998), which was found to be almost confined to sub-Saharan African populations, as the most frequent YAP^+ haplotype. Outside of Africa, it was detected only in west Asia, and there at much lower frequency values (Hammer et al. 1997). So, the presence of this haplotype elsewhere in the world will most likely represent very recent localized introductions (Bravi et al. 1997; Karafet et al. 1999).

For the five centuries following the Portuguese and Spanish discoveries, the slave trade has imported many thousands of Africans, mainly from central and western sub-Saharan Africa, into Iberia. In south and central Portugal, during the 15th and 16th centuries, individuals of African origin represented as much as 10% of the total population (Mattoso 1994). Despite this, no signs of genetic admixture with Africans are evident in the paternal inherited gene pool of the present-day Portuguese populations. On the contrary, specific sub-Saharan mtDNA haplotypes were detected, although at low frequency, in north, central, and south Portugal (Pereira et al. 1999b). Taken together, these findings suggest that interbreeding between autochthonous males and females of African ancestry was more frequent then the reciprocal mating, and the latter must have been demographically insignificant.

The most ancestral haplotype in the YAP^+ lineage, haplotype 4, was also completely absent from our sample. This is consistent with the findings of Hammer et al. (1997, 1998), which show the equivalent haplotype 3G to be confined to Asia—in particular Japan and Tibet. This distribution of the ancestral haplotype, and the predominance of the derived haplotypes in Africa, has led to the idea of a 'back to Africa' migration of an Asian population carrying YAP^+ chromosomes (Altheide and Hammer 1997; Hammer et al. 1998).

Population pairwise differentiation tests based on F_{ST}s were carried out, revealing that south Portugal was statistically different from north Portugal ($p = 0.00990 \pm 0.0100$) and Galicia ($p = 0.00990 \pm 0.0100$). These results contradict previous studies based upon autosomal short tandem repeats. It is known that the Y chromosome presents more differentiation with geograph-

ical distance than do autosomal and mtDNA loci (Seielstad et al. 1998), and the results obtained here might reflect this property.

Finally, the data were submitted to analysis of molecular variance (AMOVA), considering north, central, and south Portugal and Galicia all together. The major component of diversity was observed within populations (98.24%), while among populations the percentage of variation was found to be very low (1.76%).

Conclusions

In this study we have confirmed that haplotype analysis based on Y-chromosome biallelic markers can provide additional clues for inferring population movements and structures. The most striking result was the marked clinal distribution observed for haplotype 21. In south Portugal its frequency was more then twice that of Galicia and north Portugal, leading to statistically significant differences between south Portugal and these populations.

One interpretation of the distribution of haplotype 21 throughout Europe is as a genetic trace of the Neolithic diffusion process (Hammer et al. 1998). This population movement may also be responsible for the variation now reported for the southwestern edge of Europe. However, the observed pronounced gradient in so restricted a geographical area suggests that either the movement of farmers could have intensified a genetic profile already established in Iberia, or that later population movements have accentuated the genetic mark of the spread of agriculture. A more global picture of Y-chromosome variation throughout Europe, including more SNPs or other markers, is likely to bring useful insights into this question.

Finally, the absence of signs of recent sub-Saharan gene introductions in Iberia suggests that interbreeding between Iberian populations and African slaves, entering Europe from the 15th century until the last century, must have been very restricted, with a minor impact on the paternally inherited Iberian gene pool.

Acknowledgments This work was partially supported by Fundação para a Ciência e a Tecnologia through grant PRAXIS BD/13632/97 and project PRAXIS/2/2.1/BIA/196/94. M.A.J. is a Wellcome Senior Research Fellow in Basic Biomedical Science and received funding through grant 057559/Z/99/Z.

Received 18 January 2000; revision received 23 March 2000.

Literature Cited

Altheide, T.K., and M.F. Hammer. 1997. Evidence for a possible Asian origin of YAP$^+$ Y chromosomes. *Am. J. Hum. Genet.* 61(2):462–466.

Amorim, A., L. Gusmão, and M.J. Prata. 1996. Population and formal genetics of the STRs TPO, TH01 and VWFA31/A in North Portugal. *Adv. Forensic Haemogenet.* 6:486–488.

Budowle, B., R. Chakraborty, A.M. Giusti et al. 1991. Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* 48:137–144.

Bravi, C.M., M. Sans, G. Bailliet et al. 1997. Characterization of mitochondrial DNA and Y-chromosome haplotypes in a Uruguayan population of African ancestry. *Hum Biol.* 69(5):641–652.

Espinheira, R., H. Geada, T. Ribeiro et al. 1996. STR analysis—HUMTH01 and HUMFES/FPS for forensic application. *Adv. Forensic Haemogenet.* 6:528.

Geada, H., R. Espinheira, T. Ribeiro et al. 1996. Population genetics of D1S80, HUMVWFA31/A and HUMF13A1 from Portugal and Goa (India). *Adv. Forensic Haemogenet.* 6:465–467.

Gusmão, L., M.J. Prata, and A. Amorim. 1995. The STR system hTPO: Population and segregation data. *Int. J. Legal Med.* 108:167–169.

Gusmão, L., M.J. Prata, A. Amorim et al. 1997. Characterization of four short tandem repeat (STR) loci—TH01, VWA31/A, CD4 and TP53—in North Portugal. *Hum Biol.* 69:31–40.

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

Hammer, M.F., A.B. Spurdle, T. Karafet et al. 1997. The geographic distribution of human Y chromosome variation. *Genetics.* 145:787–805.

Hammer, M.F., T. Karafet, A. Rasanayagam et al. 1998. Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15(4):427–441.

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

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

Karafet, T.M., S.L. Zegura, O. Posukh et al. 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* 64(3):817–831.

Lareu, M.V., C.P. Phillips, A. Carracedo et al. 1994. Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats; UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci. Int.* 66:41–52.

Mattoso, J. 1994. *História de Portugal.* Lisbon: Círculo de Leitores.

Pereira, L., L. Gusmão, M.J. Prata et al. 1999a. Detection of additional structural variation at the FES/FPS system and population data from S. Tomé e Príncipe and North Portugal. *Int. J. Legal Med.* 112(3):204–206.

Pereira, L., M.J. Prata, and A. Amorim. 1999b. Analysis of mitochondrial DNA hypervariable regions I and II in a North Portuguese population. *Progress in Forensic Genetics* 8 (in press).

Santos, F.R., and C. Tyler-Smith. 1996. Reading the human Y chromosome: the emerging DNA markers and human genetic history. *Braz. J. Genet.* 19:665–670.

Santos, S.M.M., B. Budowle, J.B. Smerick et al. 1996. Portuguese population data on the six short tandem repeat loci—CSF1PO, TPOX, TH01, D3S1358, VWA and FGA. *Forensic Sci. Int.* 83:229–235.

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

Seielstad, M.T., J.M. Hebert, A.A. Lin et al. 1994. Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum. Mol. Genet.* 3(12):2159–2161.

Seielstad, M.T., E. Minch, and L.L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genet.* 20:278–280.

Semino, O., G. Passarino, A. Brega et al. 1996. A view of Neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am. J. Hum. Genet.* 59:964–968.

Souto, L., D.N. Vieira, F. Corte-Real et al. 1996. Allele frequencies in 4 STR's in a population of Portugal (central area). *Adv. Forensic Haemogenet.* 6:652–654.

Souto, L., A. Amorim, and M.C. Vide. 1998. Population and segregation data on the multiplex system (TH01, VWA, FES, F13A1) from central Portugal. *Progress in Forensic Genetics* 7:363–365.

Underhill, P.A., L. Jin, A.A. Lin et al. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7(10):996–1005.

Whitfield, L.S., E. Sulston, and P.N. Goodfellow. 1995. Sequence variation of the human Y chromosome. *Nature* 378:379–380.

Phylogeny of the mtDNA haplogroup U6. Analysis of the sequences observed in North Africa and Iberia

L. Pereira[a,b], V. Macaulay[c], M.J. Prata[a,b] and A. Amorim[a,b]

[a]IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), R. Dr. Roberto Frias, s/n, 4200 Porto, Portugal
[b]Faculdade de Ciências da Universidade do Porto, Praça Gomes Teixeira 4050 Porto, Portugal
[c]Department of Statistics, University of Oxford,1 South Parks Road, Oxford, OX1 3TG, United Kingdom

Corresponding author: Luísa Pereira, IPATIMUP, R. Dr. Roberto Frias s/n, 4200 Porto, Portugal. Phone: +351225570700 Fax: +351225570799 email: lpereira@ipatimup.pt

ABSTRACT

Comparison was performed between 41 U6 sequences observed in North Africans, 14 in Iberia and 42 in the Canary Islands. Only 2 of the 14 Iberian sequences belong to the Canarian specific U6b1 sub-haplogroup, and thus represent introduction from these islands. The remaining 12 sequences are very diverse (only 4 share the same haplotype), which do not support a single founder for the U6 lineages in Iberia.

KEYWORDS

MtDNA, haplogroup U6, Iberia, North Africa, Canary Islands

1. INTRODUCTION

The haplogroup U6, defined by transitions at np 16172 and 16219 in the hypervariable region I (HVRI) of the mitochondrial DNA (mtDNA) is characteristic of North African populations, reaching the highest frequencies (~20%) in Berbers [1]. The cluster U6 has been subdivided into two subgroups: U6a defined additionally by the transition at np 16278 and U6b having a transition at np 16311 [2].

Outside North Africa, the highest frequencies (10-16%) are observed in the Canary Islands [2], which are located in the Atlantic Ocean close to the northwest African coast, and which were a centre of the Iberian overseas expansion from the fifteenth century. There is indeed a typical Canarian U6 sub-haplogroup, U6b1, defined by an additional substitution at np 16163, which encompasses 93% of the U6 sequences observed in the Canary Islands, pointing to a unique introduction of the North African lineages into these islands around 2 800±900 years ago, with the less frequent sequences likely to have been introduced subsequently [2].

In Iberia, 14 U6 sequences have also been reported, a much higher frequency than in the rest of Europe (only 1 sequence in Sweden and another in Sicily, in a European-wide database). Out of 14 U6 sequences, 11 were from the north of Iberia: 7 in North Portugal (7.0%) [3], 2 in Galicia (2.2%) [4], and 2 in northeastern Spain (1.7%) [5], the remaining 3 being described in a general sample from Portugal (5.5%) [6].

In order to clarify whether the Iberian U6 sequences could have been introduced in one major prehistoric event, as it seems to have been the case in the Canary Islands, or if they are the outcome of more recent introductions, e.g., during the Islamic period, we took a close look at the different HVRI sequences.

2. MATERIAL AND METHODS

U6 sequences were collected from several publications summing a total of: 41 in North Africa [1,6,8,9], 42 in the Canary Islands [2], 14 in Iberia [3-6], and 9 in sub-Saharan Africa [7,8].

3. RESULTS

Of the Iberian sequences, 2 belong to the Canarian specific sub-haplogroup U6b1, and these were likely to have been introduced from these islands (a short-lived Guanche slave trade is recorded). The remaining 12 are very diverse, corresponding to 9 different haplotypes, 5 of which belong to U6a and 4 to U6b. Only 4 sequences share the same widespread U6a substitution pattern 16172-16189-16219-16278.

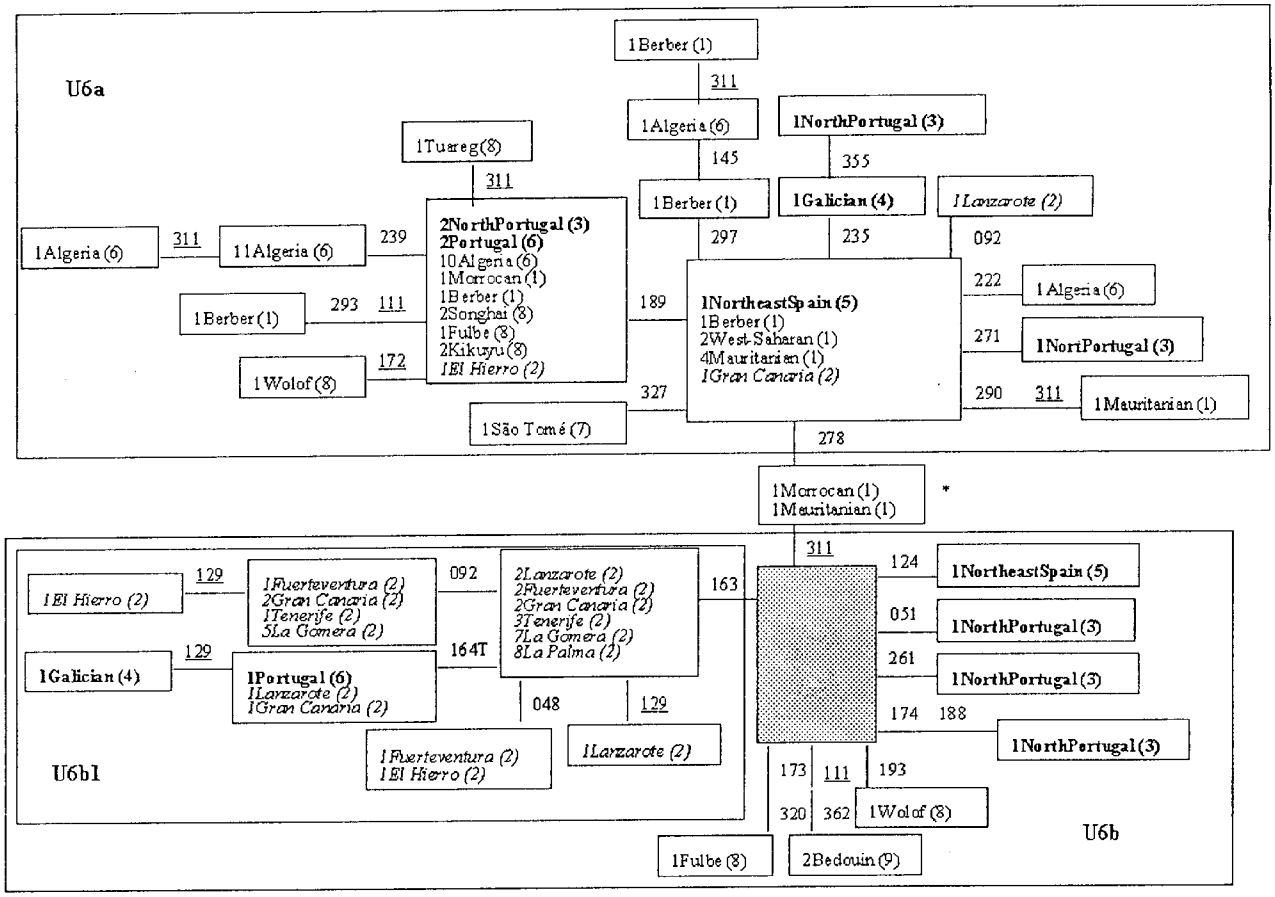


Figure 1- A most parsimonious tree of sequences belonging to cluster U6. Root motif 172-219 indicated with an asterisk. Branches are labelled by the nucleotide positions in HVRI (minus 16000) to designate transitions; transversions are further specified and positions underlined represent parallel mutations. The grey rectangle corresponds to an empty node. Iberian samples are in italic and Canarian in bold. Numbers preceding the population refer to the number of sequences observed with that motif and numbers inside brackets represent the bibliographic reference.

4. DISCUSSION

Taken together, these results do not support the hypothesis of a single founder for this haplogroup in Iberia. Whether it was introduced in a single event from a source

population with a diverse set of U6 lineages or whether it was introduced over a long period from several sources is not clear. Since it is currently not possible to infer the U6 founder types in Iberia, it is impossible to date its introduction there. An enlarged North African database would aid the solution of this puzzle.

ACKNOWLEDGEMENTS
This work was partially supported by a research grant (PRAXIS XXI BD/13632/97) from Fundação para a Ciência e a Tecnologia and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III. V.M. is a Wellcome Trust Research Career Development Fellow.

REFERENCES

[1] Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt H-J. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with Europeans, Near-Eastern, and sub-Saharan populations. Ann Hum Genet 1998;62:531-50.

[2] Rando JC, Cabrera VM, Larruga JM, Hernández M, González AM, Pinto F, Bandelt H-J. Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. Ann Hum Genet 1999;63:413-28.

[3] Pereira L, Prata MJ, Amorim A. Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann Hum Genet 2000;64:491-506.

[4] Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo A. MtDNA analysis of the Galician population: a genetic edge of European variation. Eur J Hum Genet 1998;6:365-75.

[5] Crespillo M, Luque JA, Paredes M, Fernández R, Ramírez E, Valverde JL. Mitochondrial DNA sequences for 118 individuals from northeastern Spain. Int J Legal Med 2000;114:130-2.

[6] Côrte-Real H, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, Bertranpetit J, Sykes BC. Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 1996;60:331-50.

[7] Mateu E, Comas D, Calafell F, Pérez-Lezaun A, Abade A, Bertranpetit J. A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea. Ann Hum Genet 1997;61:507-18.

[8] Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Pääbo S. MtDNA sequence diversity in Africa. Am J Hum Genet 1996;59:437-44.

[9] Di Rienzo A, Wilson AC. Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 1991;88:1597-601.

RESULTS

II - POPULATION GENETICS' APPLICATIONS

C- PORTUGAL - THE EUROPEAN CONTEXT

The analysis of Portugal in a broad European context is justified at length in the following paper, where the results on the Y-BMs are reported and analysed

ARTICLE 8

ROSSER, Z.H., ZERJAL, T., HURLES, M.E., ADOJAAN, M.A., ALAVANTIC, D., AMORIM, A., AMOS, W., ARMENTEROS, M., ARROYO, E., BARBUJANI, G., BECKMAN, L., BERTRANPETIT, J., BOSCH, E., BRADLEY, D.G., BREDE, G., COOPER, G.C., CÔRTE-REAL, H.B.S.M., DE KNIJFF, P., DECORTE, R., DUBROVA, Y.E., EVGRAFOV, O., GILISSEN, A., GLISIC, S., GÖLGE, M., HILL, E.W., JEZIOROWSKA, A., KALAYDJIEVA, L., KAYSER, M., KRAVCHENCO, S.A., LAVINHA, J., LIVSHITS, L.A., MARIA, S., MCELREAVEY, K., MEITINGER,T.A., BELA MELEGH, B., MITCHELL, R.J., NICHOLSON, J., NØRBY, S., NOVELLETTO, A., PANDYA, A., PARIK, J., PATSALIS, P.C., PEREIRA, L., PETERLIN, B., PIELBERG, G., PRATA, M.J., PREVIDERÉ, C., RAJCZY, K., ROEWER, L., ROOTSI, S., RUBINSZTEIN, D.C., SAILLARD, J., SANTOS, F.R., SHLUMUKOVA, M., STEFANESCU, G., SYKES, B.C., TOLUN, A., VILLEMS, R., TYLER-SMITH, C., JOBLING, M.A. (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. *Am. J. Hum. Genet.* **67**:1526-1543.

Am. J. Hum. Genet. 67:1526–1543, 2000

Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language

Zoë H. Rosser,[1] Tatiana Zerjal,[2] Matthew E. Hurles,[1,*] Maarja Adojaan,[5] Dragan Alavantic,[6] António Amorim,[7] William Amos,[8] Manuel Armenteros,[9] Eduardo Arroyo,[10] Guido Barbujani,[11] Gunhild Beckman,[12] Lars Beckman,[12] Jaume Bertranpetit,[13] Elena Bosch,[13,†] Daniel G. Bradley,[14] Gaute Brede,[15] Gillian Cooper,[8] Helena B. S. M. Côrte-Real,[16] Peter de Knijff,[17] Ronny Decorte,[18] Yuri E. Dubrova,[1] Oleg Evgrafov,[19] Anja Gilissen,[18] Sanja Glisic,[6] Mukaddes Gölge,[20] Emmeline W. Hill,[14] Anna Jeziorowska,[21] Luba Kalaydjieva,[22] Manfred Kayser,[23,‡] Toomas Kivisild,[3] Sergey A. Kravchenko,[24] Astrida Krumina,[25] Vaidutis Kučinskas,[26] João Lavinha,[16] Ludmila A. Livshits,[24] Patrizia Malaspina,[27] Syrrou Maria,[28] Ken McElreavey,[29] Thomas A. Meitinger,[30] Aavo-Valdur Mikelsaar,[4] R. John Mitchell,[31] Khedoudja Nafa,[32] Jayne Nicholson,[3] Søren Nørby,[33] Arpita Pandya,[2] Jüri Parik,[5] Philippos C. Patsalis,[28] Luísa Pereira,[7] Borut Peterlin,[34] Gerli Pielberg,[5] Maria João Prata,[7] Carlo Previderé,[35] Lutz Roewer,[23] Siiri Rootsi,[5] D. C. Rubinsztein,[36] Juliette Saillard,[33] Fabrício R. Santos,[2,§] Gheorghe Stefanescu,[37] Bryan C. Sykes,[32] Aslihan Tolun,[38] Richard Villems,[5] Chris Tyler-Smith,[2] and Mark A. Jobling[1]

Clinal patterns of autosomal genetic diversity within Europe have been interpreted in previous studies in terms of a Neolithic demic diffusion model for the spread of agriculture; in contrast, studies using mtDNA have traced many founding lineages to the Paleolithic and have not shown strongly clinal variation. We have used 11 human Y-chromosomal biallelic polymorphisms, defining 10 haplogroups, to analyze a sample of 3,616 Y chromosomes belonging to 47 European and circum-European populations. Patterns of geographic differentiation are highly nonrandom, and, when they are assessed using spatial autocorrelation analysis, they show significant clines for five of six haplogroups analyzed. Clines for two haplogroups, representing 45% of the chromosomes, are continentwide and consistent with the demic diffusion hypothesis. Clines for three other haplogroups each have different foci and are more regionally restricted and are likely to reflect distinct population movements, including one from north of the Black Sea. Principal-components analysis suggests that populations are related primarily on the basis of geography, rather than on the basis of linguistic affinity. This is confirmed in Mantel tests, which show a strong and highly significant partial correlation between genetics and geography but a low, nonsignificant partial correlation between genetics and language. Genetic-barrier analysis also indicates the primacy of geography in the shaping of patterns of variation. These patterns retain a strong signal of expansion from the Near East but also suggest that the demographic history of Europe has been complex and influenced by other major population movements, as well as by linguistic and geographic heterogeneities and the effects of drift.

Introduction

The earliest accepted date for the occupation of Europe by anatomically modern humans is ~40,000 years before the present (YBP) (Boyd and Silk 1997). Population size during the Paleolithic was probably stable and small, limited by the resources available from a hunting-gathering economy (Landers 1992). The development of agriculture (the Neolithic transition) was important, because the abundance of food supplies allowed populations to expand (Hassan 1973).

The origins of agriculture have become the focus of

Received July 10, 2000; accepted for publication September 25, 2000; electronically published November 9, 2000.

Address for correspondence and reprints: Dr. Mark A. Jobling, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom. Email: maj4@leicester.ac.uk

0002-9297/2000/6706-0018$02.00

[1]Department of Genetics, University of Leicester, Leicester; [2]CRC Chromosome Molecular Biology Group, Department of Biochemistry, and [3]Institute of Molecular Medicine, University of Oxford, Oxford; [4]Institute of General and Molecular Pathology, University of Tartu and [5]Estonian Biocentre, Tartu, Estonia; [6]Laboratory for Radiobiology and Molecular Genetics, Institute of Nuclear Sciences "Vinca," Belgrade; [7]5 IPATIMUP and Faculdade de Ciências, Universidade do Porto, Porto, Portugal; [8]Department of Zoology, University of Cambridge, Cambridge; [9]Centro de Investigacion y Criminalistica, Laboratorio de ADN, Policia Judicial, Guardia Civil, and [10]Laboratorio de Biología Forense, Departamento de Toxicología y Legislación Sanitaria, Univ-

attempts to interpret the genetic landscape of modern Europe. The fact that agriculture arose in the Near East ~10,000 YBP (evinced by the dating of archaeological sites) is not disputed; the argument has arisen over the mechanism of its subsequent dispersal. In the demic diffusion model (Ammerman and Cavalli-Sforza 1984), the spread is thought to be due to a movement of people and would therefore have substantially changed the genetic composition of European populations; the contrasting, cultural diffusion model (Dennell 1983; Zvelebil and Zvelebil 1988) holds that the ideas and technologies were transferred without substantial population movement and thus suggests that current patterns of genetic diversity should have their roots in the Paleolithic.

These opposing hypotheses are undoubtedly overly simplistic but have been widely adopted as models in genetic studies (Sokal et al. 1991; Cavalli-Sforza et al. 1993; Barbujani et al. 1994; Piazza et al. 1995; Semino et al. 1996; Chikhi et al. 1998*a*, 1998*b*; Richards and Sykes 1998; Simoni et al. 2000*a*), since they predict patterns of diversity that should be easily recognizable—in particular, demic diffusion is expected to result in clines with foci in the Near East. Principal components (PC) analysis of classical gene-frequency data reveals clines within Europe, and the first principal component, which indeed has a Near Eastern focus, has been taken to support the demic diffusion hypothesis (Menozzi et al. 1978; Cavalli-Sforza et al. 1993). A similar pattern has been observed in spatial autocorrelation analysis of DNA-based polymorphisms, including microsatellites, which have identified geographic patterns compatible with a substantial directional demographic expansion affecting much of the continent (Chikhi et al. 1998*a*). However, although these patterns in the genetic data are impressive and suggest major east-west population movements, their time depths are not known, and associating them with particular demographic events is usually speculative. They could be just as well due to the original peopling of Europe during the Upper Paleolithic as to the Neolithic transition. In this regard, some support for the latter does come from the finding of significant partial correlations between classical marker frequencies and the relative dates for the origin of agriculture in different locations (Sokal et al. 1991).

By contrast, analysis of diversity in European mtDNA reveals a relatively homogeneous landscape (Comas et al. 1997), with clines detectable only in the south (Simoni et al. 2000*a*). However, this is a contentious area, and conclusions may depend on the depth of analysis—for example, which sublineages are studied. An east-west gradient of pairwise differences has been discerned and claimed to be compatible with expansion from the Middle East (Comas et al. 1997). However, attempts to identify and date founding lineages (Richards et al. 1996) have suggested that Paleolithic lineages may persist in Europe to a degree that is inconsistent with the demic diffusion hypothesis, although an ancient origin of certain alleles or haplogroups (HGs) is certainly compatible with a later spread of those alleles within a geographic region (Langaney et al. 1992; Templeton 1993).

Language can provide additional evidence about past demography (Renfrew 1989), although direct information about past languages on the basis of writing is limited to the past 5,000 years, and inferences before that time are controversial (Renfrew 2000). Europe is remarkable for its linguistic homogeneity, languages of the Indo-European (IE) family being spoken by most populations from India to Ireland (Renfrew 1989). In one persuasive view, demic diffusion from the Near East provides a common explanation for the spread of both

ersidad Complutense, Madrid; [11]Dipartimento di Biologia, Universita di Ferrara, Ferrara, Italy; [12]Umeå University, Department of Medical Genetics, Umeå, Sweden; [13]Unitat de Biologia Evolutiva, Facultat de Ciecies de la Salut I de la Vida, Universitat Pompeu Fabra, Barcelona; [14]Department of Genetics, Trinity College, Dublin; [15]University of Oslo, Centre for Biotechnology, Oslo; [16]Instituto Nacional de Saúde Dr. Ricardo Jorge, Lisbon; [17]Forensic Laboratory for DNA Research, MGC-Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands; [18]Laboratory for Forensic Genetics and Molecular Archaeology, Center for Human Genetics, K.U. Leuven, Leuven, Belgium; [19]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, UFA Science Centre, Department of Biochemistry and Cytochemistry, Moscow; [20]Department of Physiology, University of Kiel, Kiel; [21]Department of Medical Genetics, Institute of Endocrinology, Medical University of Lódz, Lódz, Poland; [22]Department of Human Biology, Edith Cowan University, Joondalup Campus, and Western Australian Institute for Medical Research, Royal Perth Hospital, Perth; [23]Genetic Research Laboratory, Institute of Legal Medicine, Medical Faculty (Charité), Humboldt-University Berlin, Berlin; [24]Institute of Molecular Biology and Genetics, National Academy of Science of Ukraine, Kiev; [25]Department of Medical Biology and Genetics, Medical Academy of Latvia, Riga; [26]Center of Human Genetics, University of Vilnius, Vilnius, Lithuania; [27]Department of Biology, University of Rome "Tor Vergata," Rome; [28]The Cyprus Institute of Neurology and Genetics, Nicosia; [29]Unité d'Immunogénétique Humaine, Institut Pasteur, Paris; [30]Department of Medical Genetics, Kinderpoliklinik, Munich; [31]La Trobe University, School of Genetics and Human Variation, Bundoora, Australia [32]Department of Human Genetics, Memorial Sloan-Kettering Cancer Center, New York; [33]Laboratory of Biological Anthropology, Institute of Forensic Medicine, University of Copenhagen, Copenhagen; [34]Division of Medical Genetics, Department of Obstetrics and Gynaecology, Ljubljana, Slovenia; [35]Dipartimento di Medicina Legale e Sanita Pubblica, Pavia, Italy; [36]I.C. Biologice, Iasi, Romania; and [37]Bogazici University, Department of Molecular Biology and Genetics, Istanbul

* Present affiliation: McDonald Institute for Archaeological Research, University of Cambridge, Cambridge.

† Present affiliation: Department of Genetics, University of Leicester, Leicester, United Kingdom.

‡ Present affiliation: Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig.

§ Present affiliation: Departamento de Biologia Geral, Instituto Ciências Biológicas/Universidade Federal de Minas Gerais, Minas Gerais, Brazil.

agriculture and IE languages (Renfrew 1987). Other ideas have been put forward, however; one, which has been adopted by some geneticists because of its apparent compatibility with the pattern seen in the third principal component of variation of classical gene frequencies (Cavalli-Sforza et al. 1994), is that the IE language was spread by the movement, from north of the Caspian Sea, of the Kurgan people, pastoral nomads who domesticated the horse (Gimbutas 1970). An alternative view has it that the spread of IE language preceded the origins of agriculture and was due to the reexpansion of hunter-gatherers after the end of the Last Glacial Maximum (Adams and Otte 2000).

Despite the hegemony of IE languages, there is diversity within them, and some members of other language families also exist; one example, Basque, clearly represents a survival from an earlier era. Various methods for the detection of genetic barriers in autosomal gene frequencies within Europe (Barbujani 1991) show that most of these barriers correlate with linguistic boundaries, and it may be that language and geographic proximity are equally good predictors of genetic affinity (Barbujani 1997). However, some examples of non-IE languages reflect not persistence but recent acquisition through "elite dominance": for example, the Hungarians acquired their Uralic language from the invading Magyars only ~1,100 YBP (Cavalli-Sforza et al. 1994), and the Altaic language of the Turks was acquired as a result of the Turkic invasions during the 11th–15th centuries (Renfrew 1989). This process of language acquisition by elite dominance is not expected to be accompanied by a high degree of genetic admixture, and, if this is so, populations such as the Hungarians and Turks are unlikely to be separated from surrounding populations by genetic barriers.

Use of the Y chromosome to investigate human population histories (Jobling and Tyler-Smith 1995) is increasing as convenient polymorphic markers become available. However, the effective population size of this chromosome is one-quarter that of any autosome, and this means that it is particularly influenced by drift. Effective population size may be further reduced through the variance in the number of sons that a father has and perhaps by selective sweeps (Jobling and Tyler-Smith 2000). Conclusions about populations on the basis of this single locus must therefore be made with caution. One useful property of the Y chromosome is its high degree of geographic differentiation, compared with other parts of the genome, which has been explained by drift and a greater effective migration of women than of men, through the phenomenon of patrilocality (Seielstad et al. 1998), in which women are more likely to move from their birthplace after marriage than are men. The Y chromosome may therefore be a sensitive system for detecting the population movements that have shaped European genetic diversity; there again, it may be so susceptible to drift that ancient patterns have been obscured.

Published data on European Y-chromosome diversity are not extensive; markers have been of limited informativeness, and the distribution of population samples has often been unsatisfactory. By use of two "classical" Y-chromosome markers—the complex and highly polymorphic 49f/*Taq*I system (Ngo et al. 1986; Lucotte and Loirat 1999) and the biallelic marker 12f2 (Casanova et al. 1985)—patterns of diversity have been demonstrated that have been claimed to be clinal and to support the demic diffusion model (Semino et al. 1996). Subsequent analysis using Y-chromosome–specific microsatellites (Quintana-Murci et al. 1999) and a combination of microsatellites and two biallelic markers (Malaspina et al. 1998) showed similar east-west gradients. 49f has been exploited more fully to analyze the correlation between Y-chromosome diversity, mtDNA diversity, and language in a global sample, and it has been suggested that the Y chromosome shows the stronger correlation with language (Poloni et al. 1997).

Recent progress in the development of Y-chromosome polymorphic markers that can be assayed by use of PCR now allows us to explore these issues in greater detail. In this study, we use 11 such markers to assay the diversity of Y-chromosomal lineages in a large sample of men from 47 populations distributed over most of Europe.

Subjects and Methods

Subjects

Y chromosomes from 3,616 men from 47 populations (table 1) were included in this study; the majority were classified by birthplace of the paternal grandfather. DNA samples were from collections of the authors, and informed consent was obtained. A total of 311 samples from the Baltic region are from the study by T. Zerjal, L. Beckman, G. Beckman, A.-V. Mikelsaar, A. Krumina, V. Kučinskas, M. E. Hurles, and C. Tyler Smith (unpublished data). The 257 Irish Y chromosomes included 221 chromosomes studied elsewhere (Hill et al. 2000), which were typed here with three additional markers. The 129 North African samples were those studied elsewhere by Bosch et al. (1999); chromosomes with the M9 G allele and 92R7 C allele were additionally typed with LLY22g (see below). The 172 East Anglian samples were studied elsewhere by Cooper et al. (1996).

Biallelic Markers

A total of 11 biallelic markers were used in this study (fig. 1). These were chosen on the basis of previous work by us and by others (Santos and Tyler-Smith 1996; Sem-

Table 1

HG Frequency Data in 47 Populations

Population (no.)	Location	Language Family (Subfamily)	No. (%) of Individuals with HG												
			1	2	3	4	7	8	9	12	16	21	22	26
Icelandic (28)	64°1′N, 21°6′W	IE (Germanic)	13 (46)	9 (32)	6 (21)	0	0	0	0	0	0	0	0	0
Saami (48)	68°N, 22°E	Uralic (Finno-Ugric)	3 (6)	15 (31)	10 (21)	0	0	0	0	0	20 (42)	0	0	0
Northern Swedish (48)	63°7′N, 20°3′E	IE (Germanic)	11 (23)	22 (48)	9 (19)	0	0	0	1 (2)	0	4 (8)	1 (2)	0	0
Gotlander (64)	57°5′N, 18°5′E	IE (Germanic)	11 (17)	38 (59)	10 (16)	0	0	0	0	0	4 (6)	0	0	1 (2)
Norwegian (52)	59°9′N, 10°8′E	IE (Germanic)	15 (29)	17 (33)	16 (31)	0	0	0	1 (2)	0	2 (4)	1 (2)	0	0
Danish (56)	55°7′N, 12°6′E	IE (Germanic)	28 (50)	18 (32)	4 (7)	0	0	0	4 (7)	0	1 (2)	1 (2)	0	0
Finnish (57)	60°1′N, 25°E	Uralic (Finno-Ugric)	1 (2)	13 (23)	6 (10)	0	0	0	0	1 (2)	35 (61)	1 (2)	0	0
Estonian (207)	59°4′N, 24°7′E	Uralic (Finno-Ugric)	18 (9)	30 (14)	56 (27)	0	0	0	2 (1)	8 (4)	76 (37)	6 (3)	0	11 (5)
Latvian (34)	56°9′N, 24°1′E	IE (Balto-Slavic)	5 (15)	4 (12)	14 (41)	0	0	0	0	0	11 (32)	0	0	0
Lithuanian (38)	54°7′N, 25°3′E	IE (Balto-Slavic)	2 (5)	5 (13)	13 (34)	0	0	0	0	0	18 (47)	0	0	0
Russian (122)	55°8′N, 37°7′E	IE (Balto-Slavic)	8 (7)	21 (17)	57 (47)	0	0	0	5 (4)	5 (4)	17 (14)	8 (7)	0	1 (1)
Belarusian (41)	53°9′N, 27°5′E	IE (Balto-Slavic)	4 (10)	14 (34)	16 (39)	0	0	0	1 (2)	0	1 (2)	4 (10)	0	1 (2)
Ukrainian (27)	50°4′N, 30°5′E	IE (Balto-Slavic)	1 (4)	13 (48)	8 (30)	0	0	0	0	0	4 (11)	1 (4)	0	0
Mari (48)	56°5′N, 48°E	Uralic (Finno-Ugric)	5 (10)	2 (4)	14 (29)	0	0	0	3 (6)	8 (17)	16 (33)	0	0	0
Chuvash (17)	55°5′N, 47°E	Altaic (Turkic)	2 (12)	4 (24)	3 (18)	0	0	0	1 (6)	0	3 (18)	1 (6)	0	3 (18)
Georgian (64)	41°5′N, 44°5′E	Caucasian (Southern Caucasian)	12 (19)	31 (48)	4 (6)	0	0	0	15 (23)	0	0	1 (2)	0	1 (2)
Ossetian (47)	43°1′N, 44°5′E	IE (Indo-Iranian)	20 (43)	5 (11)	1 (2)	0	0	0	16 (34)	0	0	3 (6)	0	2 (4)
Armenian (89)	40°2′N, 44°5′E	IE (Armenian)	22 (25)	28 (31)	5 (6)	0	0	0	26 (29)	0	3 (3)	3 (3)	0	2 (2)
Turkish (167)	41°N, 29°E	Altaic (Turkic)	34 (20)	41 (25)	8 (5)	0	0	0	55 (33)	2 (1)	2 (1)	17 (10)	0	8 (5)
Cypriot (45)	35°3′N, 33°4′E	IE (Greek)	4 (9)	10 (22)	1 (2)	0	0	0	15 (33)	1 (2)	0	12 (27)	0	2 (4)
Greek (36)	38°N, 23°7′E	IE (Greek)	4 (11)	8 (22)	3 (8)	0	0	0	10 (28)	0	0	10 (28)	0	1 (3)
Bulgarian (24)	42°7′N, 23°3′E	IE (Balto-Slavic)	4 (17)	10 (42)	3 (12)	0	0	0	3 (12)	0	0	4 (17)	0	0
Czech (53)	50°2′N, 14°5′E	IE (Balto-Slavic)	10 (19)	10 (19)	20 (38)	0	0	0	6 (11)	3 (6)	0	4 (8)	0	0
Slovakian (70)	48°1′N, 17°1′E	IE (Balto-Slavic)	12 (17)	12 (17)	33 (47)	0	0	0	2 (3)	1 (1)	2 (3)	7 (10)	0	1 (1)
Romanian (45)	44°4′N, 26°1′E	IE (Italic)	8 (18)	12 (27)	9 (20)	0	0	0	11 (24)	0	0	3 (7)	1 (2)	1 (2)
Yugoslavian (100)	44°8′N, 20°5′E	IE (Balto-Slavic)	11 (11)	49 (49)	16 (16)	0	0	0	8 (8)	2 (2)	0	13 (13)	0	1 (1)
Slovenian (70)	46°1′N, 14°5′E	IE (Balto-Slavic)	15 (21)	19 (27)	26 (37)	0	0	0	4 (6)	0	0	5 (7)	1 (1)	0
Hungarian (36)	47°5′N, 19°1′E	Uralic (Finno-Ugric)	11 (30)	10 (28)	8 (22)	0	0	0	1 (3)	0	0	6 (17)	0	0
Polish (112)	51°7′N, 19°5′E	IE (Balto-Slavic)	20 (18)	19 (17)	61 (54)	0	0	0	4 (4)	1 (1)	5 (4)	2 (2)	0	0
Italian (99)	41°9′N, 12°5′E	IE (Italic)	44 (44)	14 (14)	2 (2)	0	0	0	20 (20)	0	0	13 (13)	0	6 (6)
Sardinian (10)	39°2′N, 9°1′E	IE (Italic)	3 (30)	4 (40)	0	0	0	1 (10)	0	0	0	2 (20)	0	0
Bavarian (80)	48°1′N, 11°6′E	IE (Germanic)	38 (48)	18 (23)	12 (15)	0	0	0	4 (5)	0	0	6 (8)	2 (3)	0
German (30)	52°5′N, 13°4′E	IE (Germanic)	12 (40)	6 (20)	9 (30)	0	0	0	1 (3)	0	1 (3)	0	0	1 (3)
Dutch (84)	52°3′N, 4°9′E	IE (Germanic)	36 (43)	27 (32)	11 (13)	0	0	0	6 (7)	0	0	3 (8)	1 (1)	0
French (40)	48°9′N, 2°3′E	IE (Italic)	20 (50)	10 (25)	2 (5)	0	0	1 (3)	2 (5)	0	0	3 (8)	2 (5)	0
Belgian (92)	50°8′N, 4°3′E	IE (Germanic)	58 (63)	21 (23)	4 (4)	0	0	0	5 (5)	0	0	2 (2)	1 (1)	1 (1)
Western Scottish (120)	57°2′N, 6°2′W	IE (Celtic)	87 (72)	23 (19)	8 (7)	0	0	0	0	0	0	2 (2)	0	0
Scottish (43)	56°N, 3°2′W	IE (Celtic)	34 (79)	5 (12)	3 (7)	0	0	0	0	0	0	0	0	1 (2)
Cornish (51)	50°3′N, 4°4′W	IE (Celtic)	42 (82)	9 (18)	0	0	0	0	0	0	0	0	0	0
East Anglian (172)	52°6′N, 1°3′E	IE (Germanic)	97 (56)	52 (30)	15 (9)	0	0	0	1 (1)	0	0	5 (3)	1 (1)	1 (1)
Irish (257)	53°3′N, 6°3′W	IE (Celtic)	207 (81)	39 (15)	2 (1)	0	0	0	2 (1)	0	1 (.5)	6 (2)	0	0
Basque (26)	43°3′N, 2°9′W	Basque (Basque)	19 (73)	2 (8)	0	0	0	0	0	0	0	0	5 (19)	0
Spanish (126)	40°4′N, 3°7′W	IE (Italic)	86 (68)	17 (13)	3 (2)	0	0	0	4 (3)	0	0	12 (10)	3 (2)	1 (1)
Southern Portuguese (57)	38°7′N, 9°1′W	IE (Italic)	32 (56)	8 (14)	1 (2)	0	0	0	5 (9)	0	0	10 (17)	0	1 (2)
Northern Portuguese (328)	41°2′N, 8°6′W	IE (Italic)	203 (62)	54 (16)	0	0	0	0	21 (6)	0	0	35 (11)	6 (2)	9 (3)
Algerian (27)	36°5′N, 3°E	Afro-Asiatic (Semitic)	0	1 (4)	0	0	0	1 (4)	11 (41)	0	0	14 (52)	0	0
Northern African (129)	35°5′N, 5°7′W	Afro-Asiatic (Berber and Semitic)	5 (4)	4 (3)	0	0	0	6 (5)	15 (12)	0	0	99 (77)	0	0
Total (3,616)			1,337 (37)	803 (22)	512 (14)	0	0	9 (0.3)	291 (8)	32 (1)	226 (6)	326 (9)	23 (0.7)	57 (2)

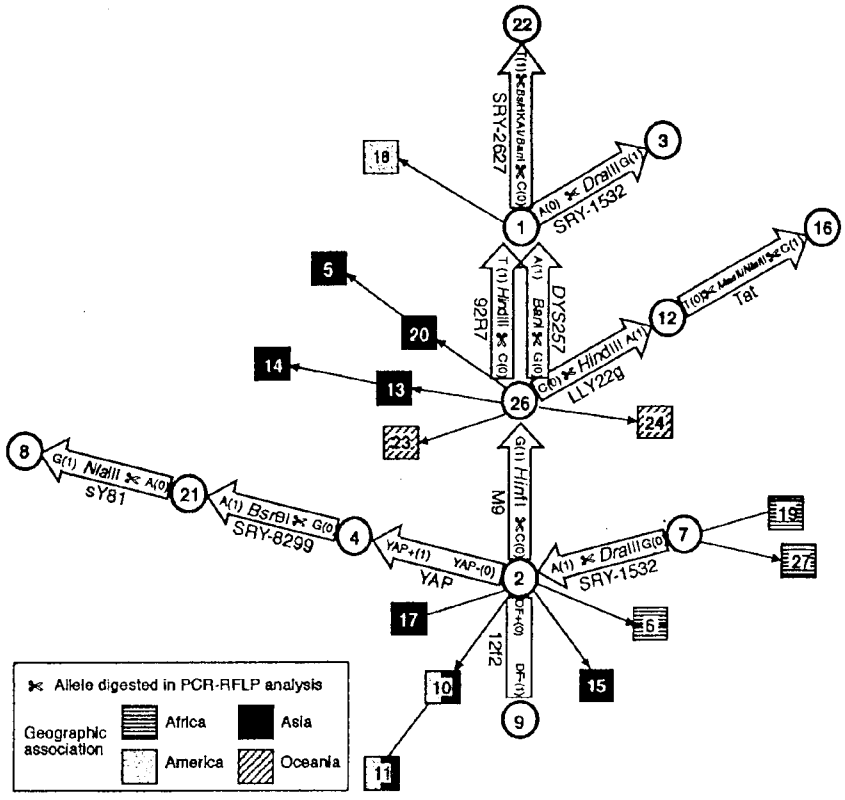


Figure 1 Maximum-parsimony network of Y-chromosomal biallelic HGs. Circles and squares represent compound haplotypes, or HGs; numbers within them are their arbitrarily assigned names; and arrows or lines between them represent the defining biallelic mutations. The order of occurrence of the 92R7 and *DYS257* mutations is not known, because the intermediate HG has not been found; arrows for these polymorphisms are shown adjacent to each other. Where ancestral state is known, arrows point to the derived state. HGs analyzed in this study are indicated by circles; arrows or boxes between them give the nature of the mutation (0, ancestral; 1, derived), and, where appropriate, the restriction enzyme used and the allele cleaved in PCR-RFLP analysis. For HGs not analyzed (*squares*), information on geographic association is provided by shading. The correspondence of some of these HGs with the haplotype nomenclature of Karafet et al. (1999) and Hammer et al. (2000), whose work is referred to in the text, is as follows: HGs 1 + 22, haplotype 1C; HG 3, haplotype 1D; HG 4, haplotype 3G; HG 7, haplotypes 1A + 2; HG 8, haplotype 5; HGs 12 + 26, haplotype 1U; HG 16, haplotype 1I; HG 21, haplotypes 3A + 4; and HG 9, haplotype "Med."

ino et al. 1996; Underhill et al. 1997; Zerjal et al. 1997; Hammer et al. 1998; Hurles et al. 1999), indicating that the HGs that they define are likely to be found within European populations. There are several nomenclature systems currently in use for Y-chromosomal lineages, and, since we refer to the data of Karafet et al. (1999) and Hammer et al. (2000) in the text, we give some correspondences in the legend to figure 1. HG 7 is specific to sub-Saharan African populations (Karafet et al. 1999) but is typed here by default, since it is defined by the ancestral state of the recurrent *SRY*-1532 polymorphism (fig. 1). Maximum-parsimony analysis of haplotypes defined by these markers generates a unique tree (figs. 1 and 2) in which *DYS257* (Hammer et al. 1998) and 92R7 (Mathias et al. 1994) are phylogenetically equivalent (Jobling et al. 1998; Z. H. Rosser, M. E. Hurles, and M. A. Jobling, unpublished data). For this part of the phylogeny, 92R7 was typed routinely, and *DYS257* was typed when necessary to confirm results. Nine of the markers have been described elsewhere: YAP (Hammer 1994) was typed according to the method of Hammer and Horai (1995), *SRY*-1532 (Whitfield et al. 1995) according to Kwok et al. (1996), *SRY*-2627 according to Veitia et al. (1997), 92R7 (Mathias et al. 1994) according to Hurles et al. (1999), *DYS257* according to Hammer et al. (1998), M9 (Underhill et al. 1997) according to Hurles et al. (1998), sY81 according to Seielstad et al. (1994), Tat according to Zerjal et al. (1997), and *SRY*-8299 (Whitfield et al. 1995) according to Santos et al. (1999).

12f2 (Casanova et al. 1985) was typed using a newly developed PCR assay. This polymorphism was originally suggested to be an ~2-kb insertion/deletion, but our analysis suggests that its molecular basis is more com-

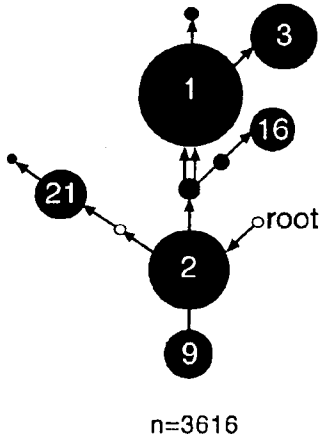


Figure 2 HG profile of the entire sample set. HG diversity within the complete sample set of 3,616 Y chromosomes, summarized on a simplified version of the network shown in figure 1. The area of each black circle is proportional to the frequency of the HG. Small unblackened circles indicate unobserved HGs (4 and 7). The position of the HG closest to the root (HG 7) is indicated.

plex than this. The PCR assay generates a 500-bp product from chromosomes carrying the *Taq*I/10-kb allele, but this product is absent from *Taq*I/8-kb–allele chromosomes (HG 9). An 820-bp amplicon from the *SRY* region, present in all chromosomes, is amplified as a control. Analysis of the 12f2 region gives no information about ancestral state, but we assume that presence of the 500-bp amplicon is ancestral. Primer sequences for the 12f2 amplicon are 12f2D (5′-CTG ACT GAT CAA AAT GCT TAC AGA TC-3′) and 12f2F (5′-TCT TCT AGA ATT TCT TCA CAG AAT TG-3′), and those for the *SRY* control amplicon are 3′*SRY*15 (5′-CTT GAT TTT CTG CTA GAA CAA G-3′) and 3′*SRY*16 (5′-TGT CGT TAC ATA AAT GGG CAC-3′). PCR conditions were 33–35 cycles of 94°C for 30 s, 59°C for 30 s, and 72°C for 45 s. An alternative assay, generating shorter amplicons, was used with degraded DNAs. The primers 12f2D (see above) and 12f2G (5′-GGA TCC CTT CCT TAC ACC TTA TAC-3′) produce an 88-bp product from *Taq*I/10-kb–allele chromosomes (and no product from *Taq*I/8-kb–allele chromosomes), which is coamplified with the Tat 112-bp amplicon (Zerjal et al. 1997) as a control, under the following conditions: 33–36 cycles of 94°C for 30 s, 59°C for 30 s, and 72°C for 30 s. All chromosomes known, from previous hybridization analysis, to carry *Taq*I/8-kb alleles lacked the 12f2 test amplicons in both of these assays. However, some YAP+ chromosomes belonging to HG 4 also lack the 12f2 amplicons, suggesting that the polymorphism may be recurrent (Blanco et al. 2000).

The LLY22g *Hin*dIII polymorphism was typed by a PCR-RFLP assay that will be described elsewhere (E. Righetti and C. Tyler-Smith, unpublished data). The deep-rooting markers *SRY*-1532, M9, YAP, and 92R7 were typed on all samples. For many samples, all other markers were also typed. However, in some cases, remaining markers were typed hierarchically—for instance, *SRY*-8299 and sY81 were, in some cases, typed only on chromosomes classified as YAP+.

Experimental Procedures

Haplotyping was carried out in Leicester; Oxford (both laboratories); Barcelona; Belgrade; Dublin; Leuven, Belgium; Lisbon; Porto, Portugal; Rome; and Tartu, Estonia. Procedures were based on those described by Hurles et al. (1998). To verify typing methodologies, a set of 12 quality-control DNAs was satisfactorily typed blindly by all participating laboratories.

Statistical Analysis

Spatial autocorrelation analysis was done by AIDA (Bertorelle and Barbujani 1995), for the entire data set, and SAAP (Sokal and Oden 1978), for individual HGs. PC analysis of covariances was carried out according to the method of Harpending and Jenkins (1973).

Mantel (1967) tests, done by ARLEQUIN version 2.0 (Schneider et al. 2000), were used to determine whether language or geography has the stronger impact on genetic differentiation. Genetic distances (as a pairwise F_{ST} matrix) were computed within ARLEQUIN, and geographic distances were calculated from latitude and longitude by use of great-circle distances, in a program written in Interactive Data Language 5.1 (Research Systems Inc.) by M. E. Hurles. Within IE languages, linguistic distances were adapted from Dyen et al. (1992), who used the lexicostatistical method of Swadesh (1952) on comparisons of 200-word lists: percentage similarities were first converted to dissimilarities, and these numbers then assigned as nonpercentage distances between languages (ranging from 9 [Czech to Slovak] to 88 [Armenian to Irish]). All IE languages within the data set were represented, with the exception of Scottish, which was assigned a distance of 10 from Irish; we also tested the effects of other values, in the range 5–20. The Belgian sample was divided into its two linguistic groups—those speaking French (56 individuals) and those speaking Dutch (36). An arbitrary and conservative, larger value, 200, was then assigned as a distance between language families. As was done by Poloni et al. (1997), Mantel tests were also performed using different inter–language-family distances, of 400 and 1,000. Two of the non–IE language families, Altaic and Uralic, are represented by more than one language within our data set. On the basis of a consideration of the classification by Ruhlen (1991) and of the inter–IE-language distances of Dyen et al. (1992), plausible distances were assigned within these families, and the effect of altering these values over

Am. J. Hum. Genet. 67:1526–1543, 2000

a range was tested. Within Uralic, values were as follows: Finnish to Estonian, 25 (altered value range 10–30); Finnish-Estonian to Saami, 30 (20–40); Finnish-Estonian-Saami to Mari, 40 (30–70); and Hungarian to all other Uralic languages, 80 (40–90). Values for Chuvash and Turkish (Altaic) were 40 (20–60).

To locate zones of abrupt genetic change, or genetic boundaries, and to assess their significance, we used the program ORINOCO, written in Interactive Data Language 5.1 (Research Systems) by M.E. Hurles (Hurles 1999), which adapts a method known as "wombling" (Barbujani et al. 1989), initially developed for the analysis of allele frequencies. First, an inverse-distance–squared weighted algorithm was used to interpolate the frequencies for each of the eight observed HGs at each grid point within a 100 × 100 array (with account taken of the curvature of the earth and with correspondence to a grid point every 0.36° latitude and 0.72° longitude). The derivatives of these eight interpolated surfaces were then calculated at every node of the grid, and the magnitudes of the derivatives were summed, thus giving a measurement of the slope of the combined surfaces—that is, the overall rate of Y-chromosomal genetic change in 10,000 rectangles covering Europe. The significance of these gradients was considered in two ways, both of which take into account isolation by distance within the landscape (Barbujani et al. 1989). First, a simple significance threshold was applied, with only the top 5% of values. Second, a Monte-Carlo algorithm was used to permute the HG data 1,000 times, and summed derivatives were calculated for each permutation. This algorithm maintains the observed sample sizes and positions and therefore controls for the conflated effects, in the generation of false positives, of sampling and heterogeneity in distances between sample sites. Grid points obtained with the original HG data were then retained only if the values of their summed derivatives were >95% of the values obtained from the permuted data. Grid points could then be plotted on a map, color coded to indicate the strength of the barrier, to show the positions of significant barriers, and were also displayed on Delaunay triangulation connections (Brassel and Reif 1979) between sample sites. The Algerian and northern-African samples were excluded from the barrier analysis, since their high degree of difference from all other samples (as shown in PC analysis) represents a strong genetic barrier that would bias the detection of barriers elsewhere.

Results

Y chromosomes from 3,487 males belonging to 47 populations (fig. 3*A*) were haplotyped using biallelic markers and were classified into HGs (table 1); data on 129 northern-African Y chromosomes (Bosch et al. 1999) were also included (see the Subjects and Methods section), giving a total of 3,616. The resulting frequency data for the entire sample are summarized in figure 2. Two HGs, 7 and 4, are absent, which is consistent with published information: HG 7 has been discussed above (see the Subjects and Methods section), and HG 4 is restricted to eastern and central Asia (Karafet et al. 1999).

No single population has a frequency distribution resembling that of the overall sample (fig. 2), emphasizing the strong geographic differentiation of Y-chromosomal variation in Europe. This is evident in the HG frequency data in figure 3: distributions of HGs are highly nonrandom, with, for example, a concentration of HG 1 chromosomes in the west, HG 9 chromosomes in the southeast, HG 16 chromosomes in the northeast, and HG 3 chromosomes in central and eastern Europe.

Clinal Distribution of Y-Chromosomal Lineages

To examine the geographic differentiation of these HGs more quantitatively, we used spatial autocorrelation analyses (Sokal and Oden 1978). These methods give a measure of the average level of genetic similarity, between populations within particular geographic distance classes, that can be represented as correlograms (fig. 4), and they allow clinal variation, reflecting population movement or natural selection, to be distinguished both from isolation by distance, reflecting short-range dispersals and drift, and from nonsignificance. We first used AIDA (Bertorelle and Barbujani 1995), which takes into account molecular distances between HGs and provides autocorrelation indices (Moran's I) for the entire data set, including the rare HGs. The pattern (fig. 4*A*) is strongly clinal, recognized as a change from positive to negative autocorrelation indices with increasing distance class. The SAAP analysis (fig. 4*B–G*), omitting low-frequency HGs (HGs 8, 12, 22, and 26), confirms this clinal pattern and reveals information about individual lineages. The distributions of all of the HGs examined, with the exception of HG 2, are strongly clinal (fig. 4), confirming the visual impression given by figure 3. In two cases (HGs 3 and 16), values become positive or zero in the longest-distance class (a "depression"), indicating the regional—rather than continentwide—influence of these clines.

HG 2 is the most ancestral lineage that we find within Europe, and it lies at a starlike node within the tree; chromosomes within this HG are essentially undefined and are likely to consist of a set of discrete sublineages that themselves probably have greater geographic coherence. Consistent with this, HG 2 chromosomes are widely distributed across the whole landscape and constitute the only high-frequency lineage that does not show clinal variation (figs. 3*B* and 4*C*). Because of this

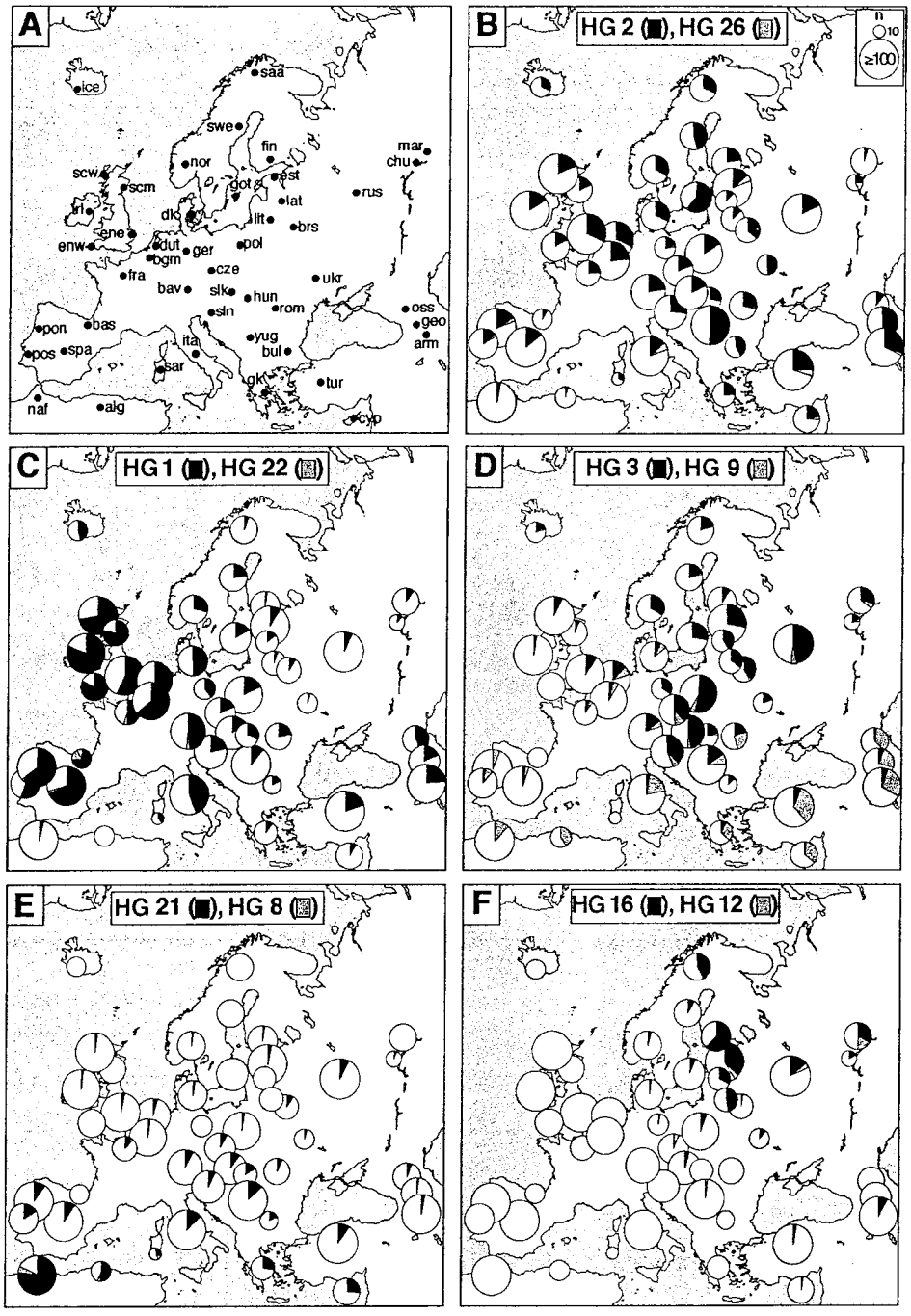


Figure 3 Distribution of populations sampled and geographic distribution of Y-chromosomal HG diversity. *A*, Abbreviated population names. alg = Algerian; arm = Armenian; bas = Basque; bav = Bavarian; bgm = Belgian; brs = Belarusian; bul = Bulgarian; chu = Chuvash; cyp = Cypriot; cze = Czech; dk = Danish; dut = Dutch; ene = East Anglian; enw = Cornish; est = Estonian; fin = Finnish; fra = French; geo = Georgian; ger = German; gk = Greek; got = Gotlander; hun = Hungarian; ice = Icelandic; irl = Irish; ita = Italian; lat = Latvian; lit = Lithuanian; mar = Mari; naf = northern African; nor = Norwegian, oss = Ossetian; pol = Polish; pon = northern Portuguese; pos = southern Portuguese; rom = Romanian; rus = Russian; saa = Saami; sar = Sardinian; scm = Scottish; scw = western Scottish; slk = Slovakian; sln = Slovenian; spa = Spanish; swe = northern Swedish; tur = Turkish; ukr = Ukrainian; yug = Yugoslavian. For a list of linguistic affiliations, see table 1. *B–F*, HG diversity within each of 47 populations, summarized on a map of Europe. The area of each pie chart is proportional to the sample size, up to a number of ≥100; sizes are indicated schematically within *B*. The area of each black or gray sector is proportional to the frequency of the corresponding HG.

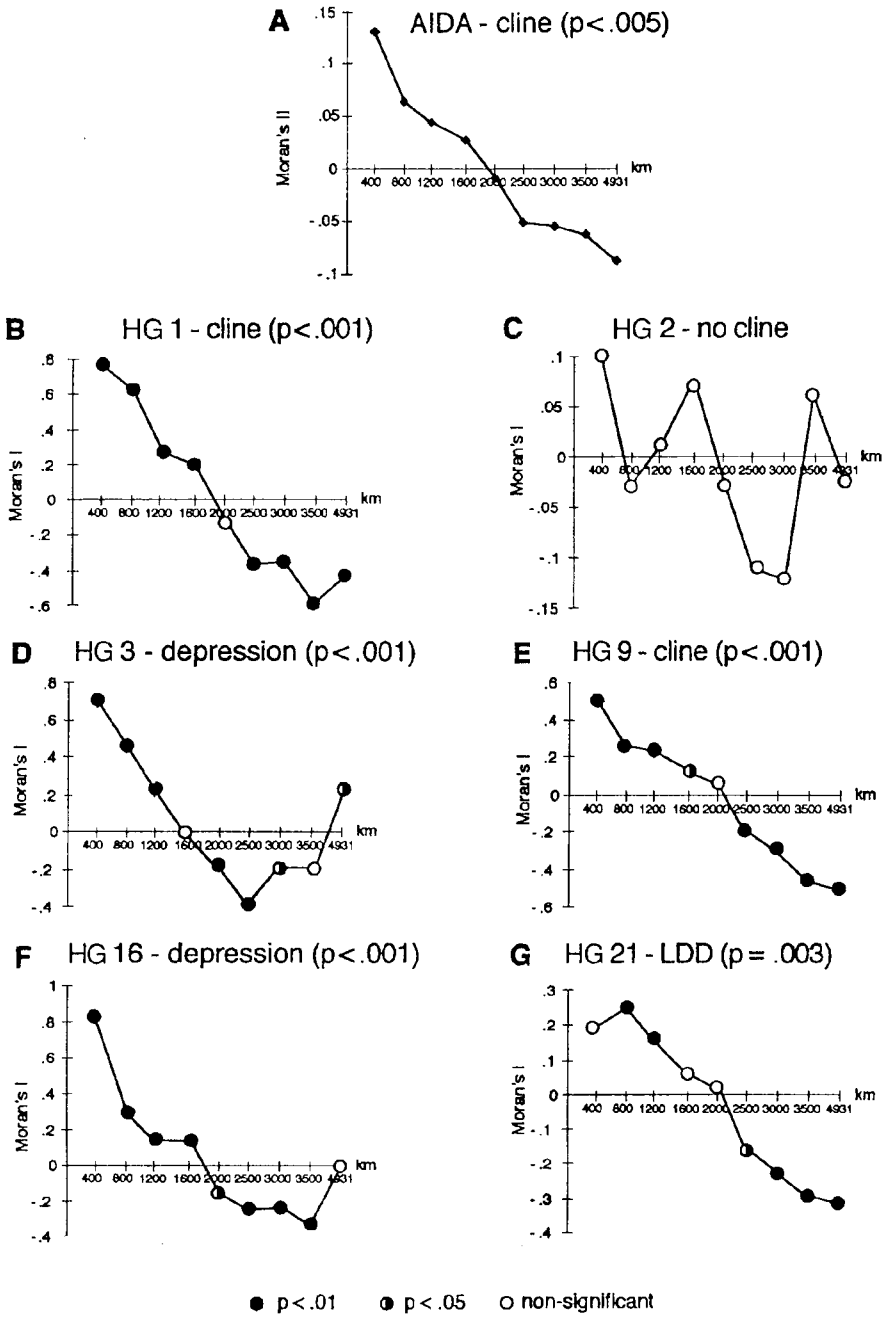


Figure 4 Spatial autocorrelation analyses. *A*, Correlogram, calculated using AIDA, for the entire data set. Overall significance is given. *B–G*, Correlograms, calculated using SAAP, for the six most frequent HGs. The significance of each point is indicated by its symbol, and the overall significance of each correlogram is also given. LDD = long-distance differentiation. In all correlograms, the *X*-axes show distance classes (km).

uninformativeness, HG 2 will not be further considered here. HG 26 occurs at low frequency (fig. 3*B*); like HG 2, it lies at a deep internal node within the tree and probably contains unidentified coherent sublineages.

We find two other HGs at low frequency—HGs 8 and 22. HG 8 is common in sub-Saharan Africa (Karafet et al. 1999) and is present in our northern-African samples at ~5% (fig. 3*E*). Only two European examples exist,

in Sardinia and France, which may represent recent admixture.

HG 22 chromosomes (fig. 3C) reach appreciable frequencies only in the French (5%) and Basques (19%). This HG has been analyzed in detail in a study elsewhere (Hurles et al. 1999), which suggested that it has a recent Iberian origin and that non-Iberian examples represent migrants. The distribution here is consistent with this analysis.

A Major Cline Consistent with the Demic Diffusion Model

HGs 1 and 9 show complementary clines on the continental scale, from the southeast of Europe to the northwest (figs. 3C and *D* and 4*B* and *E*): indeed, when the Irish sample is further subdivided on the basis of geographic information contained within surnames (Hill et al. 2000), HG 1 reaches near-fixation (98.5%) in the west of Ireland. HG 9 reaches its highest frequencies (~33%) in the Caucasus and in Anatolia (fig. 3*D*), where it is thought that agriculture originated (Cavalli-Sforza et al. 1994). The strong clinal pattern of these two HGs, which together account for almost half (45%) of the chromosomes in our study, resembles the first principal component of genetic variation of classical loci and is consistent with the demic diffusion hypothesis. However, distributions of the remaining HGs are very different from these and cannot be interpreted as a simple reflection of population movement from the Near East.

A Northeast/Southwest Cline Signaling an Expansion from North of the Black Sea

The distribution of HG 3 chromosomes is also strongly clinal (fig. 4*D*), but with a very different axis (fig. 3*D*) and more on a regional scale, and is likely to reflect population-historical events distinct from those responsible for the distributions of HGs 1 and 9. It reaches its highest frequencies in central-eastern Europe, comprising approximately half of the chromosomes in the Russian, Polish, and Slovakian samples; frequencies in the southeast and southwest are low. This distribution resembles the third principal component of variation of classical gene frequencies, which has been interpreted by some geneticists (Cavalli-Sforza et al. 1994) as marking the movement, from north of the Caspian Sea, of the Kurgan people, dated to ~7,000 YBP.

A North-South Cline: A Northern-African Influence?

Within Europe, HG 21 chromosomes are concentrated in the south (fig. 3*E*). Their frequency in the two northern-African samples is very high (52% and 77%), and their frequencies in the Greek and Cypriot samples are also high (~27%), which might reflect a barrier to gene flow between Africa and Europe, as is also shown by the analysis of autosomal protein markers (Simoni et al. 1999) and microsatellites (Bosch et al. 2000). In other southern-European populations, such as those in Spain, Portugal, Sardinia, Italy, Turkey, and Yugoslavia, frequencies are in the range of 10%–20%. The decline in frequencies to the north is rather uniform. This regional cline (fig. 4*G*) has similiarities to that detected in the second principal component of classical gene frequencies (Cavalli-Sforza et al. 1994), which has been interpreted on a climatic basis.

A Lineage Concentrated in the Northeast: A Contribution of Uralic Speakers?

HG 16 is at high frequency in the north, east of the Baltic Sea (fig. 3*F*), a distribution consistent with that noticed previously in a global survey (Zerjal et al. 1997). Its pattern is again clinal but regional (fig. 4*F*). HG 12, ancestral to HG 16, is at low frequency in the sample overall. However, its distribution overlaps that of HG 16, with no examples in the western half of the continent, and is concentrated more in the south (fig. 3*F*). It is most frequent (17%) in the Mari, who may be the population of origin of the Tat mutation, which defines HG 16 (T. Zerjal and C. Tyler-Smith, unpublished data).

With the exception of the Hungarians, who acquired their Uralic language through elite dominance by the Magyars during recent times (Cavalli-Sforza et al. 1994), all Uralic-speaking populations tested (Finnish, Estonians, Saami, and Mari) show a high frequency of HG 16. However, two nearby populations, the Lithuanians and Latvians, also show HG 16 at high frequency but speak languages of the IE family—for this lineage at least, the association appears to be geographic rather than linguistic. In the following section, we use methods that summarize variation among all lineages, to examine this issue in more detail.

Geography and Language as Causes of Genetic Differentiation

Population comparisons through PC analysis.—PC analysis is a method that allows the graphic display, in a few dimensions, of the maximum amount of variance within a multivariate data set, with minimum loss of information. Figure 5 shows the results of a PC analysis of the Y-chromosome HG data, in which populations are labeled according to linguistic affiliation. PC1–PC3 summarize 71.4% of the variance.

The major division is between the two populations from northern Africa and the others. This is unsurprising, given their high frequencies of HGs 21 and 9 and their near absence of HG 1, and indicates that the Mediterranean, even at its narrowest point, has represented a barrier to gene flow, as has been suggested previously by autosomal DNA analysis. The Mediterranean pop-

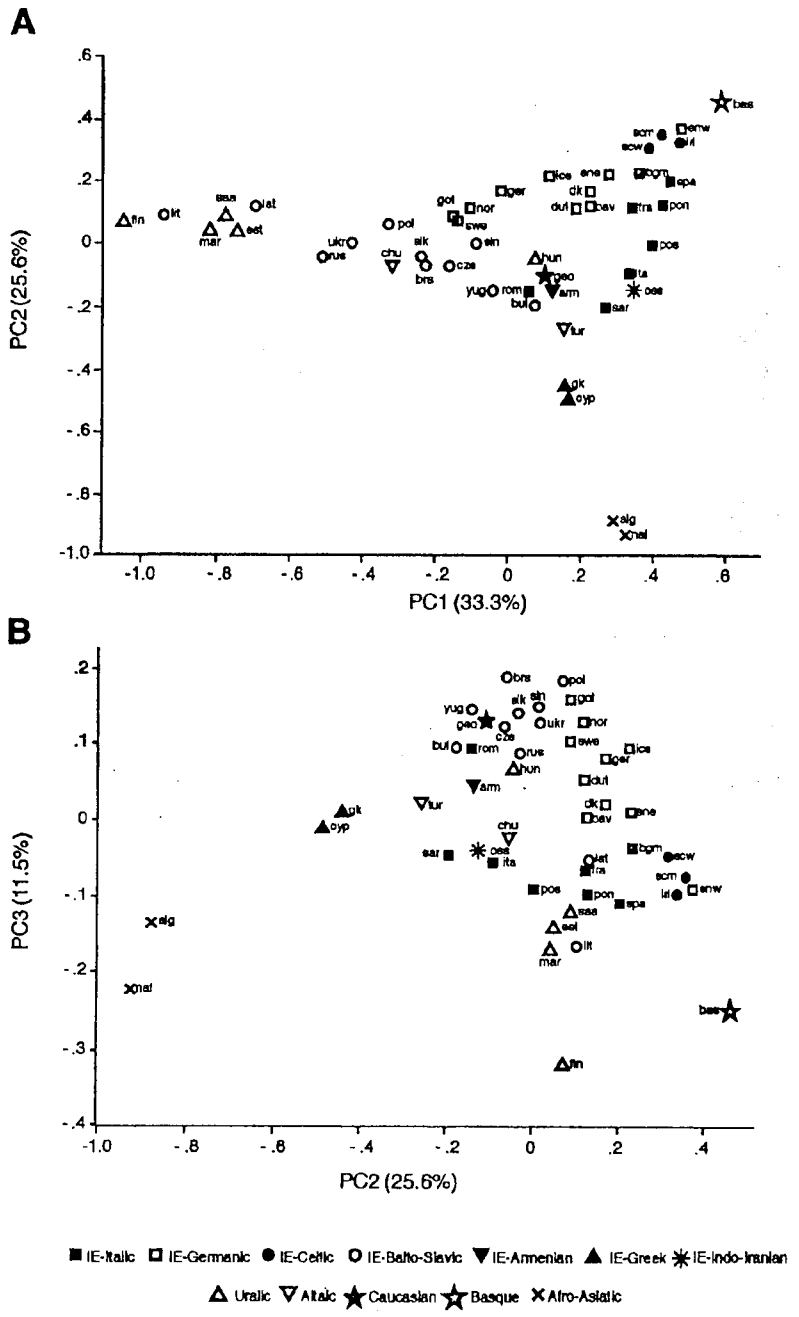


Figure 5 PC analysis of Y-chromosomal HG diversity. *A*, PC2 plotted against PC1. *B*, PC3 plotted against PC2. The percentage of variance explained by each component is given on the axes. Linguistic affiliation for each population is indicated symbolically; the Belgian sample is part Dutch-/part French-speaking and has a hybrid symbol. Abbreviations are as in figure 3.

ulations of Greece and Cyprus occupy an intermediate position between the northern Africans and the rest.

Basques speak a non-IE language unrelated to any other language (Ruhlen 1991) and thus represent the most striking example of a linguistic isolate in Europe. This isolation seems to be reflected in the PC analysis, in which they are separated from other populations (fig. 5*A*); however, this may be due to high frequency of a young lineage (HG 22; Hurles et al. 1999), rare elsewhere, rather than to persistence of ancient ones. Their

closest neighbors in the PC analysis are not the geographically close populations of Iberia but those of the Atlantic fringe, most of which speak Celtic-IE languages. In this context, the Cornish sample ("enw" in Figs. 3 and 5) is grouped not with the eastern English sample (ene) but with the Scottish and Irish—a reflection of geography or of the original Celtic language of this region (Ruhlen 1991) or both.

Among Uralic-speaking populations, this analysis confirms the impression given by figure 3*F*: with the exception of the Hungarians, who lie close to IE language speakers, these populations are grouped together with the Finns separated from the rest in PC3 (fig. 5*B*). Also within this group are the Lithuanians and Latvians, supporting the idea that this is primarily a geographic association.

The overall impression from figure 5 is that geographic proximity may be a better predictor of Y-chromosomal genetic affinity than is language: as well as the examples discussed above, the Italic-IE language–speaking Romanians are distant from other Italic language speakers, and the Turks lie between the geographically neighoring but linguistically distant Armenians and Greeks.

Correlating Geography, Language, and Genetics through Mantel Testing

Mantel (1967) tests provide an objective way of assessing the relative importance of different factors in the shaping of genetic diversity. In this method, correlation coefficients between pairs of factors (from genetics, geography, and language) can be calculated, together with significance values; partial correlation coefficients are then calculated between genetics and geography and between genetics and language, with the third factor kept constant to control for the strength of the correlation between geography and language. The populations from northern Africa are linguistically remote and geographically peripheral, and the PC analysis has shown their genetic differentiation. We therefore excluded them from the Mantel analysis, to examine effects within Europe itself. Genetics and geography (table 2) are strongly and significantly correlated ($P < .001$), and the correlation between genetics and language is less strong but still significant ($P = .014$). The partial correlation of genetics and geography, with language kept constant, is again strong and significant ($P < .001$); in contrast, the partial correlation of genetics and language is low and nonsignificant ($P = .095$). We examined the effect of changing the values that we had assigned to distances within Uralic and within Altaic and between Irish and Scottish (see the Subjects and Methods section), and this had a negligible influence on our results. Increasing the distance assigned between language families had the effect of reducing still further the partial correlation between genetics and language, as well as its significance. This analysis confirms the primacy of geography, rather than language, in the shaping of Y-chromosomal genetic diversity within Europe.

Table 2

Correlation and Partial Correlation Coefficients between Genetic, Geographic, and Linguistic Distance

Distance Considered	Correlation Coefficient	P[a]
Genetics and geography	.387	<.001
Genetics and language	.198	<.01
Genetics and geography, language held constant	.349	<.001
Genetics and language, geography held constant	.088	NS

[a] NS = not significant.

Location of Y-Chromosomal Genetic Barriers within Europe

Although the analysis above indicates a lack of large-scale correlation between language and genetics, it does not address local genetic differentiation, which may reflect local effects of language. Genetic-barrier analysis, which locates the zones of sharpest genetic change within a landscape, provides a way to do this.

Figure 6 shows the results of a genetic-barrier analysis of the Y-chromosome HG data for 45 populations, for the top 5% of barriers and a 95% significance filter (see the Subjects and Methods section). Within western Europe, minor barriers separate the Basques from some neighboring populations, the western from the eastern English, and the Dutch from the Belgians. In the east, there are two major barriers, one between the Uralic-speaking Mari and Altaic-speaking Chuvash and one between the Georgians and the Ossetians, who speak languages belonging to different families and who are also separated by the Caucasus Mountains. Most of the major barriers lie in the middle of the European landscape, running from Italy in the south to the Baltic Sea in the north, including one barrier around the island population of Gotland.

To what extent are linguistic differences contributing to Y-chromosomal barriers within Europe? Since 37 different languages are spoken among our 45 sample sites, we expect most genetic barriers to fall between populations speaking different languages. However, if language differences do constitute barriers to gene flow, then we might expect that the degree of linguistic difference between a pair of populations should correlate with the chance of a genetic barrier occurring—that is, the greatest proportion of genetic barriers should fall between populations speaking languages from different families, a lesser proportion between those speaking languages from different subfamilies, and the least between those speaking languages within a subfamily. There are

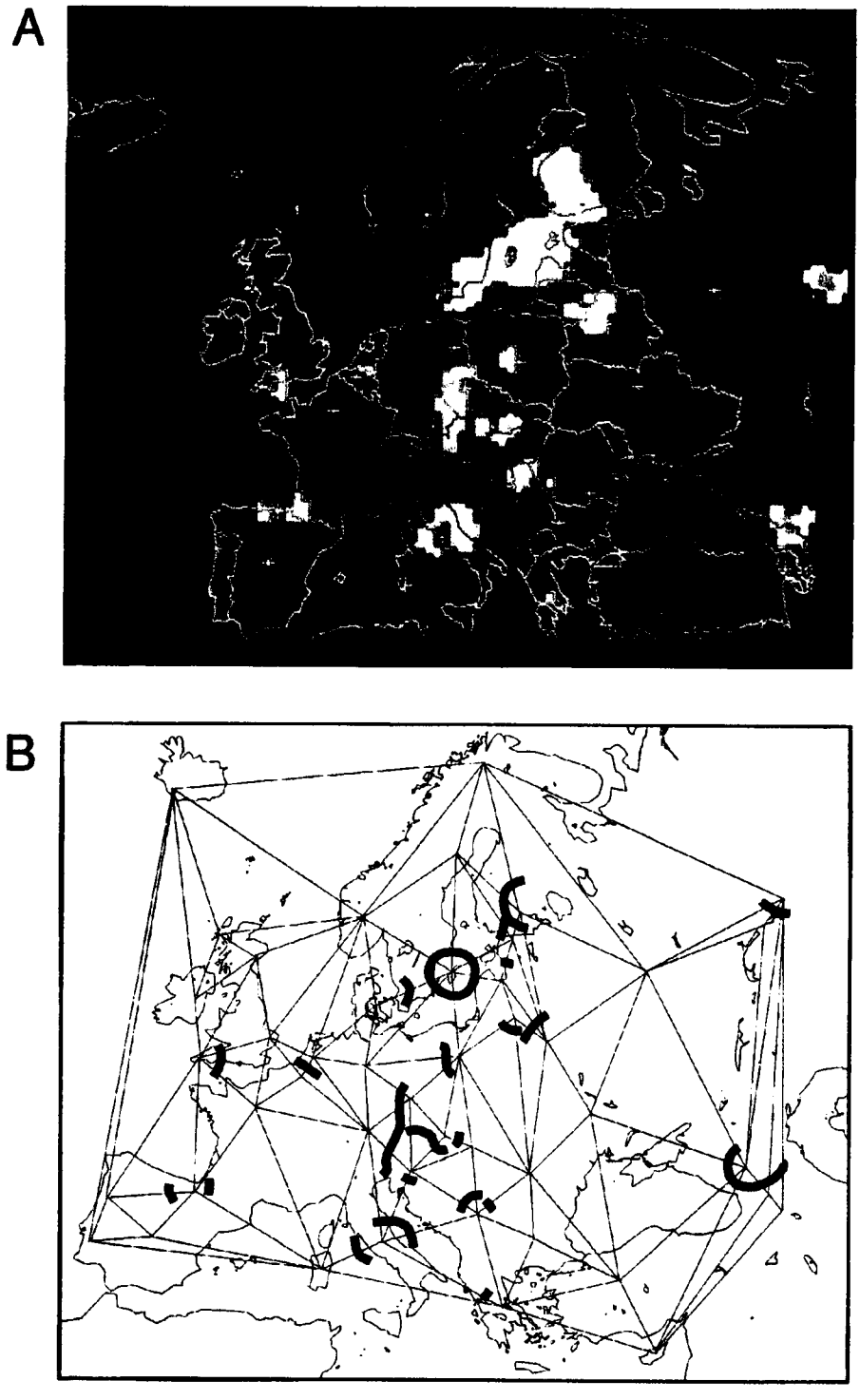


Figure 6 Significant Y-chromosomal genetic barriers within Europe. *A*, Output from the ORINOCO program. Positions of genetic barriers showing 95% significance after permutation (see the Subjects and Methods section) are indicated by blue through red areas on the black background, with sample sites indicated by stars. A three-dimensional animation of the actual output from the program can be viewed at the Molecular Genetics Laboratory of the McDonald Institute for Archaeological Research Web site. *B*, Schematic version of the output shown in *A*, with the positions of barriers indicated as thick lines on Delaunay connections (*thin lines*) between sample sites.

122 Delaunay connections in figure 6*B*, 48 of which are crossed by a genetic barrier. We count the proportion of connections that are crossed by a genetic barrier in each of the three classes, between language families, between subfamilies, and within subfamilies; these values are 46.2% (18/39), 40.5% (15/37), and 32.6% (15/46), respectively. Although the ranking of these three values is that expected under the hypothesis, differences between them are not significant ($P > .1$, three-way χ^2 test). This suggests that language may not be the primary force contributing to genetic barriers here. However, this analysis does not take into account the fact that two non-IE languages, Hungarian and Turkish, have been acquired recently: the PC analysis and the relative absence of Y-chromosomal genetic barriers around these populations supports the idea that elite dominance was not accompanied by extensive genetic admixture. If we remove these two populations and repeat the above analysis, differences between the proportions increase (to 50.0% [13/26], 43.2% [19/44], and 31.9% [15/47], respectively) but remain not significant ($P > .1$).

Discussion

We have described the most detailed survey to date of human Y-chromosomal diversity within Europe. Samples were distributed over most of the continent, including its western and eastern fringes; inclusion of these regions, omitted from some other studies, has allowed both the detection of influences from the east and clines extending to the extreme west, for example. However, some regions remain poorly sampled, and, if the possible effects of local differentiation are to be studied, more-extensive sampling is needed. At the eastern edge of Europe lie the steppes, which stretch uninterrupted to China. Analogous studies of Asian Y chromosomes are under way and will place the European data within a broader context (W. Bao, S. Zhu, M. E. Hurles, T. Zerjal, M. A. Jobling, J. Xu, Q. Shu, R. Du, H. Yang, and C. Tyler-Smith, unpublished data).

We used 11 biallelic markers in this study, but there is still a need for more. For instance, HG 2, constituting 22% of the total sample and as much as 49% in the sample from Yugoslavia, is poorly defined and therefore constitutes a potential source of error in our analyses, since equal weight is given both to this and to well-defined HGs. The pace of new marker discovery is increasing (Underhill et al. 1997; Shen et al. 2000), and soon the resources will be available to adequately define all major European lineages.

Consistent with global surveys (Underhill et al. 1997; Karafet et al. 1999), this continental study confirms the high degree of geographic differentiation of Y-chromosomal lineages. This differentiation makes the Y chromosome a sensitive indicator of either admixture, as demonstrated in studies of Polynesia (Hurles et al. 1998), South America (Bianchi et al. 1997), and Uruguay (Bravi et al. 1997), for example, or an absence of admixture, as has been shown in Jewish populations in Europe and northern Africa (Hammer et al. 2000). Knowledge about admixture is of particular importance in the choice of populations for studies that use linkage-disequilibrium analysis (McKeigue 1997) in both simple and complex disorders.

Clines of Y-Chromosomal HGs

The effects of drift on human Y-chromosome diversity are likely to be great. It is striking, therefore, to observe clear clinal variation in five of the six major lineages within Europe—this suggests that drift has not erased the patterns of variation established by past population movement. Natural selection on Y chromosomes (Jobling and Tyler-Smith 2000) provides an alternative explanation for such clines; possible effects of geographically variable factors (such as temperature) on fertility within specific lineages have yet to be investigated, but, in the absence of evidence to the contrary, we assume that the variation that we are assaying is selectively neutral and can therefore be interpreted in terms of population history.

The contrast between the clinal variation of Y-chromosomal lineages and the lack of clines in mtDNA data (Simoni et al. 2000*a*) is marked, although the latter is still a matter of debate (Simoni et al. 2000*b*; Torroni et al. 2000). It seems consistent with studies of global genetic diversity (Seielstad et al. 1998), which have ascribed such differences to patrilocality. However, direct evidence about mating practices in European prehistory is lacking—indeed, populations in some regions, such as northern Iberia, may have practiced matrilocality (Collins 1986).

Clines for HGs 1 and 9, encompassing 45% of the chromosomes—and doing so on a continental scale—show a pattern similar to that seen both in the first principal component of classical gene-frequency data and in the autocorrelation analysis of six Y-chromosomal microsatellites (Casalotti et al. 1999). A simplistic interpretation is that HG 9 chromosomes were carried in a major demographic expansion of agricultural migrants from the Near East and that HG 1 chromosomes were a preexisting predominant European lineage. Estimates of the ages of these lineages, from coalescent analysis, are not inconsistent with this scenario: the mutation defining HG 1 has been dated at ~23,000 YBP (Karafet et al. 1999), and that defining HG 9 has been dated at 14,800 ± 9,700 YBP (Hammer et al. 2000).

Demic diffusion—and, indeed, any major directional gene-flow process—is generally expected to generate clines for only a fraction of the alleles at one locus (Sokal

et al. 1989, 1997). Although two HGs show clines compatible with expansion from the Near East, three further lineages show different clinal patterns, indicating distinct population movements: southward and westward from north of the Black Sea (HG 3), from eastern Europe or northern Asia westward to the Baltic Sea (HG 16), and from south to north (HG 21). These clines are more regionally localized than those for HGs 1 and 9, pointing to phenomena affecting only part of the continent. It is tempting to assign known or surmised population-historical movements to these genetic gradients, but this should be done with caution.

The distribution of HG 3 chromosomes resembles the third principal component of variation of classical gene frequencies. There are several possible interpretations of this pattern. One explanation (Cavalli-Sforza et al. 1994) is that it marks the Kurgan expansion from north of the Caspian Sea, dated to ~7,000 YBP. However, alternative explanations—such as the spread of pastoralism, or east-to-west movements of people such as the Scythians, Mongols, and Huns—seem equally likely (Renfrew 2000). Globally, HG 3 chromosomes are absent from Africa and the Americas, but their distribution is wide within Asia as well as in Europe (Zerjal et al. 1999), consistent with their association with a recent and major expansion within Eurasia. Microsatellite diversity analysis (Zerjal et al. 1999) used the mutation-rate estimates of Heyer et al. (1997) to date the most recent common ancestor of a set of European and Asian HG 3 chromosomes to 3,800 YBP (95% confidence interval [CI] 1,600–13,000 YBP); the use of more-recent mutation-rate estimates (Kayser et al. 2000) would yield a date of 2,550 YBP (95% CI 1,650–4,260 YBP). Coalescent analysis has dated the *SRY*-1532 mutation defining HG 3 to ~7,500 YBP (Karafet et al. 1999). If these dates are to be relied on, they seem to suggest that the expansion of HG 3 chromosomes was due to population movements later than those of the Kurgan people.

Currently, dates cannot be attached to the clines, and the modern distributions of lineages are the outcome of many millennia of population movement. Assigning plausible dates to demographic movements is important, and here the Y chromosome can potentially contribute. Finer-scale definitions of monophyletic lineages within Europe, by use of new markers, and the analysis of these, by use of microsatellites, offers the possibility that timescales for the major demographic events can be inferred.

Language, Geography, and Y-Chromosomal Diversity

The Mantel tests demonstrate that patterns of Y-chromosomal genetic variation do not correlate as well with language as with geography. However, it should be borne in mind that geography and language together explain only 16.8% of the genetic variance (data not shown); therefore, other forces, such as founder effects and genetic drift, have also been important in determining the current patterns of spatial variation. Our findings seem at odds with those of Poloni et al. (1997), who showed that most of the population differentiation of Y-chromosome haplotypes was due to language. However, there are important differences between the two studies. The samples of Poloni et al. (1997) were global, rather than from a single continent, and showed a correspondingly greater linguistic and genetic diversity. The populations that we have studied are located within a single continent, and most speak languages belonging to one language family, IE; indeed, much of the genetic patterning that we now see may have its roots in the spread of that language family (Renfrew 1987). The effect of increasing genetic, geographic, and linguistic diversity in the input to the Mantel tests can be seen by including the northern-African samples (data not shown), which are both geographically and linguistically distant from most other populations. This increases the partial correlations between genetics and geography and between genetics and language and also increases the significance of the latter to $P = .024$, which, however, is still lower than the significance of the genetics-geography partial correlation ($P < .001$).

The results of genetic-barrier analysis (fig. 6) need to be interpreted with caution when, as in this case, sample distribution is uneven; the method is likely to be sensitive to the introduction of new populations, especially between existing sample sites that are far apart. However, the analysis has suggested that there is little correlation between genetic barriers and levels of linguistic separation, even when elite dominance is taken into account by removing the Hungarians and Turks from the analysis. Although cultural factors other than language (such as politics and religion) might also be associated with genetic barriers, we have examined language because it has the greatest time depth. However, this is still likely to be less than the age of geographic barriers, the relative importance of which cannot easily be analyzed. Twenty-five of 48 Delaunay connections crossed by genetic barriers also coincide with geographic barriers (under a conservative definition that considers only large stretches of water and the two major mountain ranges, the Alps and the Caucasus), which seems to emphasize the greater importance of geographic factors in subdividing populations, resulting in large differences in Y-chromosomal HG frequencies.

In synthesis, it seems that many kinds of barriers are probably recent, on an evolutionary timescale (see Renfrew 1987); after they have been established, fluctuations of allele frequencies have become partly or largely independent in the populations separated by those barriers. Therefore, it is perhaps not surprising to find little

correlation between the degree of language differentiation at a language boundary and the amount of genetic change observed across that boundary. As has been shown in the analysis of protein polymorphisms (Sokal et al. 1990), linguistic differences tend to cause some degree of population subdivision, regardless of whether such differences are between language families, between languages of the same family, or even between dialects of the same language.

Although we have dichotomized the forces of geography and language, in reality they work together; spatially coincident weak geographic and linguistic barriers may together form strong barriers to gene flow. Some of the strongest genetic barriers observed, in central Europe, coincide with neither strong linguistic nor strong geographic barriers. Linguistic and geographic heterogeneities and the effects of drift, on a background retaining a strong signal of expansion from the Near East and of other migrations, have combined to shape the genetic landscape of Europe.

Acknowledgments

We thank the DNA donors for making this study possible, and we thank Laurent Excoffier for assistance. Z.H.R. was supported by a BBSRC Studentship, T.Z. by a Wellcome Trust Bioarchaeology Studentship, M.E.H. by an MRC Studentship, F.R.S. by the Leverhulme Trust, and L.P. by Ph.D. grant PRAXIS XXI/BD/13632/97 from Fundação para a Ciência e a Tecnologia. D.C.R. is a Glaxo Wellcome Research Fellow. C.T.-S. is supported by the CRC, and M.A.J. is a Wellcome Trust Senior Fellow in Basic Biomedical Science, supported by grant 057559. Iberian sample collection was partially funded by multidisciplinary project grant PR182/96 6745 from Complutense University.

Electronic-Database Information

The URL for data in this article is as follows:

Molecular Genetics Laboratory of the McDonald Institute for Archaeological Research, http://www-mcdonald.arch.cam.ac.uk/Genetics/home.html

References

Adams J, Otte M (1999) Did Indo-European languages spread before farming? Curr Anthropol 40:73–77

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

Barbujani G (1991) What do languages tell us about human microevolution? Trends Ecol Evol 6:151–156

——— (1997) DNA variation and language affinities. Am J Hum Genet 61:1011–1014

Barbujani G, Oden NL, Sokal RR (1989) Detecting regions of abrupt change in maps of biological variables. Syst Zool 38: 376–389

Barbujani G, Pilastro A, de Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. paleolithic colonisation. Am J Phys Anthropol 95:137–154

Bertorelle G, Barbujani G (1995) Analysis of DNA diversity by spatial autocorrelation. Genetics 140:811–819

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

Blanco P, Shlumukova M, Sargent CA, Jobling MA, Affara N, Hurles ME (2000) Divergent outcomes of intra-chromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J Med Genet 37:752–758

Bosch E, Calafell F, Pérez-Lezaun A, Clarimón J, Comas D, Mateu E, Martínez-Arias R, Morera B, Brakez Z, Akhayat O, Sefiani A, Hariti G, Cambon-Thomsen A, Bertranpetit J (2000) Genetic structure of north-west Africa revealed by STR analysis. Eur J Hum Genet 8:360–366

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

Boyd R, Silk JB (1997) How humans evolved. WW Norton, New York

Brassel KE, Reif D (1979) A procedure to generate Thiessen polygons. Geogr Anal 11:289–303

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

Casalotti R, Simoni L, Belledi M, Barbujani G (1999) Y-chromosome polymorphisms and the origins of the European gene pool. Proc R Soc Lond B Biol Sci 266:1959–1965

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

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

——— (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G (1998*a*) Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. Proc Natl Acad Sci USA 95:9053–9058

Chikhi L, Destro-Bisol G, Pascali V, Baravelli V, Dobosz M, Barbujani G (1998*b*) Clinal variation in the nuclear DNA of Europeans. Hum Biol 70:643–657

Collins R (1986) The Basques. Blackwell, Oxford

Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Bertranpetit J (1997) Mitochondrial DNA variation and the origin of the Europeans. Hum Genet 99:443–449

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

Dennell R (1983) European economic prehistory: a new approach. Academic Press, London

Dyen I, Kruskal JB, Black P (1992) An Indoeuropean classification: a lexicostatistical experiment. Trans Am Philos Soc 82:1–132

Gimbutas M (1970) Proto-Indo-European culture: the Kurgan culture during the fifth, fourth and third millennia B.C. In: Cardona G, Hoenigswald HM, Senn A (eds) Indo-European and Indo-Europeans. University of Pennsylvania Press, Philadelphia, pp 155–195

Hammer MF (1994) A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol 11:749–761

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

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

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonné-Tamir B (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA 97:6769–6774

Harpending H, Jenkins T (1973) Genetic distance among Southern African populations. In: Crawford MH, Workman PL (eds) Methods and theories of anthropological genetics. University of New Mexico Press, Albuquerque, pp 177–199

Hassan FA (1973) On mechanisms of population growth during the neolithic. Curr Anthropol 14:535–542

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

Hill EW, Jobling MA, Bradley DG (2000) Y chromosomes and Irish origins. Nature 404:351–352

Hurles ME (1999) Mutation and variability of the human Y chromosome genetics. University of Leicester, Leicester

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

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Pérez-Lezaun A, Bosch E, Shlumukova M, Cambon-Thomsen A, McElreavey K, López de Munain A, Röhl A, Wilson IJ, Singh L, Pandya A, Santos FR, Tyler-Smith C, Jobling MA (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. Am J Hum Genet 65: 1437–1448

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

——— (2000) New uses for new haplotypes: the human Y chromosome, disease, and selection. Trends Genet 16:356–362

Jobling MA, Williams G, Schiebel K, Pandya A, McElreavey K, Salas L, Rappold GA, Affara NA, Tyler-Smith C (1998) A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol 8:1391-1394

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, deKnijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am J Hum Genet 64:817–831

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Krüger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580–1588

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

Landers J (1992) Reconstructing ancient populations. In: Jones S, Martin R, Pilbeam D (eds) The Cambridge encyclopedia of human evolution. Cambridge University Press, Cambridge, pp 402–405

Langaney A, Roessli D, van Blyenburgh NH, Dard P (1992) Do most human populations descend from phylogenetic trees? Hum Evol 7:47–61

Lucotte G, Loirat F (1999) Y-chromosome DNA haplotype 15 in Europe. Hum Biol 71:431–437

Malaspina P, Cruciani F, Ciminelli BM, Terrenato L, Santolamazza P, Alonso A, Banyko J, Brdicka R, Garcia O, Gaudiano C, Guanti G, Kidd KK, Lavinha J, Avila M, Mandich P, Moral P, Qamar R, Mehdi SQ, Ragusa A, Sefanescu G, Caraghin M, Tyler-Smith C, Scozzari R, Novelletto A (1998) Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. Am J Hum Genet 63:847–860

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

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

McKeigue PM (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 60:188–196

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

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J (1986) A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet 38:407–418

Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza LL (1995) Genetics and the origin of European languages. Proc Natl Acad Sci USA 92:5836–5840

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup L, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035

Quintana-Murci L, Semino O, Minch E, Passarino G, Brega A, Santachiara-Benerecetti AS (1999) Further characteristics of proto-European Y chromosomes. Eur J Hum Genet 7:603–608

Renfrew C (1987) Archaeology and language: the puzzle of Indo-European origins. Jonathan Cape, London

——— (1989) The origins of Indo-European languages. Sci Am 261:106–114

——— (2000) At the edge of knowability: towards a prehistory of languages. Camb Archaeol J 10:7–34

Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt H-J, Sykes B (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59: 185–203

Richards M, Sykes B (1998) Evidence for Paleolithic and Neolithic gene flow in Europe. Am J Hum Genet 62:491–492

Ruhlen M (1991) A guide to the world's languages. Edward Arnold, London

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

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19:665–670

Schneider S, Roessli D, Excoffier L (2000) ARLEQUIN ver 2.0: a software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

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

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

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

Shen P, Wang F, Underhill PA, Franco C, Yang W-H, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (2000) Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA 97:7354–7359

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

——— (2000*b*) Reconstruction of prehistory on the basis of genetic data. Am J Hum Genet 66:1177–1179

Simoni L, Gueresi P, Pettener D, Barbujani G (1999) Patterns of gene flow inferred from genetic distances in the Mediterranean region. Hum Biol 71:399–415

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. Biol J Linn Soc 10:199–249

Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim J, Thomson BA, Vaudor A, Harding RM, Barbujani G (1990) Genetics and language in European populations. Am Nat 135: 157–175

Sokal RR, Oden NL, Thomson BA (1997) A simulation study of microevolutionary inferences by spatial autocorrelation analysis. Biol J Linn Soc 60:73–93

Sokal RR, Oden NL, Wilson C (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143–145

Swadesh M (1952) Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. Proc Am Philos Soc 96:452–463

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. Am Anthropol 95:51–72

Torroni A, Richards M, Macaulay V, Forster P, Villems R, Nørby S, Savontaus M-L, Huoponen K, Scozzari R, Bandelt H-J (2000) mtDNA haplogroups and frequency patterns in Europe. Am J Hum Genet 66:1173–1177

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

Veitia R, Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, Tosi M, Meo T, Chibani J, Fellous M, McElreavey K (1997) Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46,XY female phenotype. Hum Genet 99:648–652

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

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60: 1174–1183

Zerjal T, Pandya A, Santos FR, Adhikari R, Tarazona E, Kayser M, Evgrafov O, Singh L, Thangaraj K, Destro-Bisol G, Thomas MG, Qamar R, Mehdi Q, Rosser ZH, Hurles ME, Jobling MA, Tyler-Smith C (1999) The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In: Papiha SS, Deka R, Chakraborty R (eds) Genomic diversity: applications in human population genetics. Plenum Press, New York, pp 91–102

Zvelebil M, Zvelebil KV (1988) Agricultural transition and Indo-European dispersal. Antiquity 62:574–583

RESULTS

II - POPULATION GENETICS' APPLICATIONS

D- PORTUGAL - AS CONTRIBUTOR FOR MOZAMBIQUE

The papers presented in this section report the genetic study of the Mozambican population, a former Portuguese colony. Mozambique is the only Portuguese former colony in the east African coast. This location puts it in the route for the Bantu migrations inside Africa towards the southernmost tip of the continent. Moreover, Mozambique was by the end of the slave trade an important outpost for the commerce towards, mainly, America.

These features have prompted us to work out the historic traces of Bantu expansion and slave trade routes through Mozambican mtDNA analysis, which is reported in the following paper:

ARTICLE 9

PEREIRA, L., MACAULAY, V., TORRONI, A., SCOZZARI, R., PRATA, M.J., AMORIM, A. (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann. Hum. Genet.* **65** (in press).

The second paper is a preliminary analysis, dealing with the low European impact registered in present Y-chromosome Mozambican pool and the consequences of the Bantu expansion on the reduction of its diversity.

ARTICLE 10

PEREIRA, L., GUSMÃO, L., ALVES, C., AMORIM, A., PRATA, M.J. Y-chromosome pool in the southeastern African population of Mozambique: the small European influence and the Bantu diversity reduction. (in preparation).

Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade

LUÍSA PEREIRA[1,2], VINCENT MACAULAY[3], ANTONIO TORRONI[4,5], ROSARIA SCOZZARI[5], MARIA JOÃO PRATA[1,2] and ANTÓNIO AMORIM[1,2]

[1] *Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP),*
R. Dr. Roberto Frias s/n, 4200 Porto, PORTUGAL
[2] *Faculdade de Ciências da Universidade do Porto,*
Pr. Gomes Teixeira, 4050 Porto, PORTUGAL
[3] *Department of Statistics, University of Oxford,*
1 South Parks Road, Oxford, OX1 3TG, UNITED KINGDOM
[4] *Dipartimento di Genetica e Microbiologia, Università di Pavia,*
Via Ferrata 1, 27100 Pavia, ITALY
[5] *Dipartimento di Genetica e Biologia Molecolare*
Università "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, ITALY

Running head: The mtDNA of Mozambique
Key words: mtDNA, HVRI, HVRII, haplogroups, sequence sharing

Correspondence:
Luísa Pereira
Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP),
R. Dr. Roberto Frias s/n, 4200 Porto, PORTUGAL
Phone: +351 22 5570700 Fax: +351 22 5570799 email: lpereira@ipatimup.pt

SUMMARY

A sample of the mitochondrial DNA (mtDNA) from the southeastern African population of Mozambique has been shown to have affinities with populations both to its north and south. From the north came sequences that may have been involved in the Bantu expansion (from western, through eastern, to southern Africa), such as members of haplogroups L3b, L3e1a and a subset of L1a. The dating of the major component of Mozambican mtDNAs, the subset L2a of haplogroup L2, displayed an age range compatible with the Bantu expansion. The southern influence was traced by the presence of sequence types from haplogroup L1d, a probable relict of Khoisan-speaking populations that inhabited the region prior to their displacement by the Bantu-speaking incomers. Within historical times, the forced displacement of Mozambicans as part of the slave trade, mainly documented as being to the Americas, generated a differential input of eastern African sequences into the mtDNA pools of the Americas and of Europe, as testified to by the greater number of sequences matches between Mozambique and the Americas compared to those between Mozambique and Europe.

INTRODUCTION

In the last decade, data on African mtDNA have been accumulated with the aim of unravelling something of the demographic phenomena that have contributed to the settlement of the continent. This task is still at an early stage and is made potentially more difficult in Africa than in the rest of the world because the characteristic demographic phenomena of recurrent migration, population expansion and contraction including bottlenecks, population sub-structure generated by limits on gene flow, and more recent admixture effects have occurred over a longer time depth. In addition, there is a rather poor archaeological context in which to place the genetics.

Several studies of restriction-fragment length polymorphisms (RFLPs) (Cann *et al.* 1987; Chen *et al.* 1995), of control region sequences (Vigilant *et al.* 1991; Soodyall 1993; Krings *et al.* 1999) or a combination of both (Graven *et al.* 1995; Watson *et al.* 1997; Chen *et al.* 2000) have involved African populations. However, the sampling is still surprisingly patchy, and quite poor in the southeast. Besides the intrinsic interest of seeing how this region fits into the emerging picture of African mtDNA diversity, it has potential implications on a wider scale too. South-east Africa was an important source of slaves, from 1643 onwards, when individuals from Mozambique and Madagascar constituted a major portion of the slaves shipped by the Portuguese to the former European colonies in America, e.g. Brazil and The Caribbean, in such a way that 'by the eighteenth century this commerce, directed to the Americas, was more important on that coast than anywhere else' (Thomas 1998). Records point to ~1000000 slaves originating from Mozambique/Madagascar in a total of ~13000000 leaving African ports (Thomas 1998). By that time, slave importation was already reduced in Europe, and the majority of European countries forbade entry of black slaves by the middle of the eighteenth century. So the eastern sub-Saharan contribution to the European African sequences, sporadically detected in some countries of this continent (Côrte-Real *et al.* 1996; Pereira *et al.* 2000) is expected to have been reduced, compared to its influence in America.

Here, we present hypervariable region I (HVRI) and II (HVRII) data for Mozambique, a south-east African population, which was a Portuguese colony between 1752 and 1975. This country contains several ethnic groups, nearly all Bantu-speaking.

Originally a linguistic classification, referring to a widespread group of languages within the South-Central Niger-Congo family, Bantu now refers to a complex of physical anthropological and genetic characters correlated with the linguistic distribution, which are explained by a large-scale Holocene range expansion. This expansion occurred in several waves and directions and was responsible for the dispersal of farming to southern and central Africa. The linguistic evidence points to a Bantu origin in the vicinity of the Cross River valley near the present-day border between Nigeria and Cameroon (Newman 1995). Around 5000 years ago, the Bantu expansion began in two directions: south-western, arriving at the equatorial rain forest by 3500 years ago and eastern, entering the fringes of the interlacustrine region in what is now Uganda, 3000 years ago, forming the eastern Bantu core area. From this new core, two new expansions moved towards South Africa: one group along the Ruvuma River toward the coast, reaching present-day Natal by the end of the third century A.D., and the other along the shores of Lake Malawi, through what is now eastern Zimbabwe, reaching the northern Transvaal around A.D. 500. At the mitochondrial level, some Bantu expansion markers have been proposed: Soodyall *et al.* (1996) and Watson *et al.* (1997) pointed to a 9-bp deleted subset of haplogroup L1a and to the haplogroup L3b, while Bandelt *et al.* (in press) proposed that haplogroup L3e1a must have been prominent in the southern Bantu expansion. However, particularly in the case of L1a, a detailed dissection of the phylogeography, with a well-resolved phylogeny, has still to be performed: the signal could easily be of an earlier Holocene event.

Our aims here are to describe the phylogeography of Mozambique sequences in the context of African variation, and to search for possible sequence matches outside Africa that might shed light on the slave trade.

MATERIAL AND METHODS

Subjects

A total of 109 unrelated individuals born in Mozambique were analysed and DNA was extracted from blood spots by the resin Chelex-100 method (Lareu *et al.* 1994). Individuals belonged to different ethnic groups (Changana, 35; Ronga, 21; Chope, 12; Bitonga, 8; and Matsua, 8; and 25 to various other groups), but all were Bantu speakers (http://www.sil.org/ethnologue/countries/Moza.html).

HVRI and II amplification and sequencing

Mitochondrial DNA was amplified using the primers L15997 (5'-CACCATTA GCACCCAAAGCT-3') and H16401 (5'-TGATTTCACGGAGGATGGTG-3') for HVRI and L48 (5'-CTCACGGGAGCTCTCCATGC-3') and H408 (5'-CTGTTAAAAGTGCATACCGCCA-3') for HVRII. The temperature profile was 95 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s, for 35 cycles of amplification. The amplified samples were purified with Microspin™ S-300 HR columns (Pharmacia Biotech), according to the manufacturer's specifications. The sequence reactions were carried out using the kit Big-Dye™ Terminator Cycle Sequencing Ready Reaction (Perkin-Elmer), with one of the above primers, in both forward and reverse directions. A protocol based on $MgCl_2$/ethanol precipitation was used for post-sequence reaction purification of samples, which were then applied to a 6% PAGE and run in an automatic sequencer ABI 377.

RFLP analyses of haplogroup L3 sequences

In order to check the assignment of a number of sequence types to haplogroup L3 and its subclusters, we checked the following RFLPs in putative members of L3: 2349*Mbo*I (present in L3e), 3592*Hpa*I (absent in L3 in general), 8616*Mbo*I (absent in L3d) and 10084*Taq*I (present in L3b), in the numbering system of the Cambridge Reference Sequence (CRS) of Anderson *et al.* (1981). PCR amplifications were performed using primers and conditions described by Torroni *et al.* (1992). Digestions were carried out according to the manufacturer's specifications and the resulting

fragments were run in 9% polyacrylamide gels and visualised by silver staining (Budowle *et al.* 1991).

Genetic analysis and population comparison

The nucleotide positions considered for the analysis of the sample were 16024–16383 for HVRI and 73–340 for HVRII. Length variation (often scored as transversions in HVRI) was not considered in the analysis (Bendall & Sykes 1995).

Sequence classification into haplogroups was according to Watson *et al.* (1997), Rando *et al.* (1998), Macaulay *et al.* (1999), Alves-Silva *et al.* (2000), Richards & Macaulay (2000) and Bandelt *et al.* (in press). Figure 1 displays a phylogeny of the African haplogroups used to classify the haplotypes. Given the relatively high mutation rate of mtDNA, especially in the control region, and the large time depth of some of the African haplogroups, recurrent mutations at motif positions occur quite often. It is necessary to be sensitive to this possibility when classifying haplotypes.

The Mozambique data were compared with data from several populations from various parts of Africa. The code, place of origin, sample size and bibliographic references of these population samples are displayed in Table 1. In some samples for which there were few individuals, we combined neighbouring populations or the same population from different studies. For this analysis, we only considered HVRI from nucleotide position 16090-16365.

Molecular diversity indexes, mismatch distributions, analysis of variance (AMOVA) and tests of the standard neutral model (via Tajima's D and Fu's F_s statistics) were calculated in ARLEQUIN 2.0 (Schneider *et al.* 2000). Principal component (PC) analyses were performed using POPSTR (H. Harpending, pers. comm.) on the haplogroup composition of the various African populations. For these analyses we used the frequencies of the following haplogroups: pre-HV, N1a together with N1b (Richards *et al.* 2000), I, J, K, T, U6, the rest of U, X (Macaulay *et al.* 1999), M1 (Quintana-Murci *et al.* 1999), L1a, L1b, L1c, L1d, L1e, L1f, L1* (consisting of non-L2/L3 types not classified as L1a-f), L2a, L2b, L2c together with L2* (since these cannot be distinguished without HVRII information), L3*, L3b, L3d, L3e1, L3e2, L3e3, L3e4 (Figure 1) and 'other'. A population neighbour-joining tree was obtained using PHYLIP (Felsenstein 1993) from pairwise F_{ST} values estimated in ARLEQUIN 2.0. These F_{ST}

values incorporate information both on haplotype frequencies and the genetic distances between haplotypes, calculated as the number of nucleotide positions different between pairs of sequences. Reduced median networks were constructed by hand and checked in NETWORK 2.0d (Bandelt *et al.* 1995). The dates of the most recent common ancestor of specific subclusters in the phylogeny were estimated using ρ, the average number of transitions from the ancestral sequence type to all sequences in the cluster, in conjunction with a mutation rate estimate of 20 180 years per transition in the sequence stretch 16090-16365 (Forster *et al.* 1996). Standard errors were calculated as in Saillard *et al.* (2000).

RESULTS AND DISCUSSION

Sequencing and RFLP results

The HVRI and HVRII sequences obtained and the results of the selective RFLP typing are shown in the Appendix.

HVRI and HVRII diversity in Mozambique

Diversity measures for both hypervariable regions are displayed in Table 2. As has been described previously, HVRII is rather less diverse than HVRI (Pereira *et al.* 2000). In both segments, the diversity in Mozambique is higher than in a typical European population, e.g., the mean pairwise difference is 1.5-2 times greater.

Mismatch distributions (Figure 2) in Mozambique were very ragged for both HVRI and HVRII, which is consistent with results for other sub-Saharan African populations (Bandelt & Forster 1997). An interesting point was the considerable number of identical sequences (especially for HVRII), which could point to a sample bias effect of some ethnic groups consisting of closely related individuals. However this does not seem to be the case, since most of the identical sequences belong to different ethnic groups.

We further investigated if there was any substructuring of mtDNA between individuals from the Mozambique ethnic groups for which we had a substantial number of samples. The application of AMOVA showed that there was no evidence of significant variation between ethnic groups ($p = 0.60$).

Comparison with other African populations

In order to set the diversity observed in Mozambique within a continental context, we compared this population with a database of dispersed African populations. Since data for HVRII are scarcer we considered only HVRI diversity.

We used several methods, some based on haplogroup frequencies, such as PC analysis, and others based on sequence diversity, such as AMOVA.

Molecular diversity

Overall diversity for HVRI in Africa (Table 3) is highest for western and eastern populations, followed closely by northern ones. The two Pygmy groups (here collectively referred to as Central African) and southern Africans displayed considerably lower diversity.

Departure from the standard (null) model of populations evolving at constant size in mutation-drift equilibrium with no selection was tested by employing Tajima's D and Fu's F_s statistics. For Tajima's D there were significant negative values only in northern populations and in Sudan. All the other populations showed non-significant values, and some southern (Khoisan-speaking) populations and the Pygmies even displayed positive values (although also non-significant). For Fu's more powerful F_s statistic, there were more populations with significant negative values, not only restricted to the north, but also present in the west and the east, while non-significant positive values occurred as for D. If we neglect the hypothesis of selection, the significant negative D and F_s values provide an indication of population expansion in most populations except in the Pygmies and Khoisan. This observation does not necessarily imply absence of expansion in the Pygmies and Khoisan, but possibly that the signal was lost by subsequent contractions (Bandelt & Forster 1997).

AMOVA analysis

We grouped the different populations into large geographic zones: northern, western, eastern, central and southern, as displayed in Table 1, and investigated how the proportion of variance was distributed between groups and between populations in the same group, by AMOVA. When Mozambique was assigned to the eastern group, the proportion of variance between populations in the same group took its minimum value (4.9%, compared to 5.5% when included in the central group, 6.5% when included in the southern group and 5.2% when included in the western group, all values significantly greater than zero at the 5% level). In this case, the proportion of variance between groups was 11.8%. This suggests that our Mozambique sample may have closest affinities to the populations to its north.

NJ tree

A population NJ tree constructed from pairwise F_{ST}s (not shown) revealed the geographic clustering of the different populations, with a Khoisan/Pygmy cluster, a northern African cluster and a western African cluster. The eastern populations fall between the Khoisan/Pygmy cluster and the rest. Mozambique clusters with western populations, although the branch that links it to them is long, in contrast to the short branches connecting western populations. As in the AMOVA analysis, the Mozambique population appears less like those populations to the south than to the north. A tree-like model of population evolution is unlikely to capture much of the reality of population history, so we proceeded with an exploratory PC analysis.

Principal component analysis

A preliminary PC analysis of all the population plus a sample of Herero (Vigilant *et al.* 1991) (for which there was enough information for haplogroup classification, but too many uncharacterised sites for its inclusion in the previous analyses) showed some consistent geographical clustering, although the western and eastern populations were intermingled (Fig. 3).

The first two principal components amount to only 38% of the variation, leaving the rest uncharacterised. However, the first principal component (PC1), responsible for 22% of the variance, splits northern, western and eastern populations from southern ones and Mozambique. The main haplogroup responsible for this PC is L1d, which is typical of Khoisan populations (Bandelt and Forster, 1997); its presence in non-Khoisan populations may represent recent admixture. L1d constitutes 7% of the Mozambique sample and was absent in all the other non-southern populations analysed here (except one individual in the Turkana). The second principal component (PC2), responsible for 16% of the variance, distinguishes the northern African populations. The main haplogroups responsible for this PC are L1c and pre-HV. The pattern in L1c probably reflects the extremely high frequency in the Biaka (probably due to drift) compared to its near absence in the north, while that in the predominantly western Eurasian haplogroup pre-HV is accounted for by its virtual absence south of the Sahara.

In order to remove these large signals, which are not especially informative with regard to Mozambique, we performed a refined PC analysis by excluding the peripheral

populations in Figure 3. Whereas in the previous analysis western and eastern populations were mixed up, the new PC1 (Fig. 4), responsible for 30% of the variance, splits the eastern populations and Mozambique from the western populations. Several haplogroups have a similar contribution: L3*, L1a, L1b, L2*+L2c and L1e. L3*, L1a and L1e are typically eastern haplogroups, and L1b and L2*+L2c are western haplogroups. The new PC2, responsible for 18% of the variance, splits the eastern and western populations in a north-south axis, and the main haplogroup responsible for this is pre-HV, as above.

Analyses and dating of sequence types

All the Mozambican sequences belong to sub-Saharan haplogroups (see Appendix). Past European (especially Portuguese) contact was not detected at the mtDNA level: there is a complete absence of European sequences (Richards *et al.* 2000). In addition, no east (Horai *et al.* 1996) or south (Kivisild *et al.* 1999) Asian mtDNAs were detected. No Near Eastern sequences (Richards *et al.* 2000), detectable in some northeastern African populations, were observed in Mozambique, and the north African haplogroup U6 (Côrte-Real *et al.* 1996; Macaulay *et al.* 1999) was also absent.

Certain haplogroups were present at high frequency in the Mozambique sample (table 4) and for these we performed a phylogenetic network analysis and examined their distributions across Africa, in an attempt to determine when they arrived in Mozambique.

Haplogroups L1a and L1d

Haplogroup L1a appears in our sample in two clusters of haplotypes. These clusters are inferred to correlate with the presence/absence of one instance of the intergenic COII/tRNA^{Lys} 9bp deletion (Soodyall *et al.* 1996). The non-deleted L1a types (usually with 16168T and 185A: Ingman et al., 2000) represent 10 mtDNAs in our sample (9%), while the six remaining L1a mtDNAs (6%) are likely 9bp-deleted.

Haplogroup L1d is by far the most common in Khoisan-speaking populations and, apart from one individual from the Turkana with an outlying L1d type (Watson *et al.* 1997), has so far not been observed north of Namibia. It is an early branch in the phylogeny, although its precise location is still subject to some uncertainty (compare Chen *et al.* 2000 with Watson *et al.* 1997). It is present in eight individuals (7%) of the

Mozambique sample, who display seven different HVRI/HVRII haplotypes. These types could well represent a relict of the populations that inhabited this area before the Bantu and earlier migrations, although recent gene flow with Khoisan-speaking populations (Bandelt & Forster 1997) cannot be excluded.

Haplogroup L2a

Figure 5 shows a reduced median network of haplogroup L2 in Mozambique. The majority (95%) of Mozambican L2 belong to the subcluster L2a defined in HVRI by a transition at 16294 in addition to the L2 motif. This subcluster is widely distributed in Africa and of considerable age (39000-51400 years: Chen et al., 2000). Its diversity is highest in eastern and western Africa (Table 5) and it is rare in Khoisan-speaking populations. Hence it probably has its origin at latitudes immediately south of the Sahara. In Mozambique it has a reduced diversity relative to the eastern and western populations. Ten out of the fourteen Mozambique L2a HVRI haplotypes had not been observed before, and belong to two or three subclusters. The largest of these Mozambique-specific clusters in L2a (based on 16189-16290-16294-16309 on top of the L2 motif), the root sequence of which is shared with a Zimbabwean (Horai & Hayasaka 1990), has an age of 13500±3000 years, suggesting a movement southwards in the late glacial or early Holocene. This subcluster contains another frequent haplotype (bearing 16192 in addition), which, although not observed elsewhere, is another potential founder type. If we make this assumption, the combined age falls to 6700±2100 years, a Holocene signal which could tentatively be attributed to the southern Bantu expansion.

In South African Y chromosomes from Bantu speakers, a substantial reduction in diversity is observed (Thomas *et al.* 2000): one (YAP+) haplotype, based on six microsatellite loci, together with its one-step neighbours, comprises almost half the Bantu Y chromosomes. This group of chromosomes is consistent with an expansion from a single type 3000–5000 years ago. The diversity reduction has no parallel in mtDNA, perhaps suggesting that local maternal lineages were assimilated during the expansion. Indeed, it is far from clear that the signature that we are detecting in L2a is not that of an earlier expansion, perhaps following climate change in the late-glacial, a

pattern which is becoming evident in other parts of the world (Forster *et al.* 1996; Torroni *et al.* 1998; Richards *et al.* 2000), albeit in regions where the changes in the climate were rather different.

Haplogroup L3

Figure 6 shows a reduced median network of haplogroup L3 in Mozambique. All well-characterised African L3 clusters are present in Mozambique, as well as one less well-characterised group, based on 16209-16223-16292-16311 (cf. Alves-Silva *et al.* 2000), which has a distribution south of the Sahara. All samples belonged to the African-specific branches of L3. L3b comprises 4% of the sample, is widespread in western Africans and has been implicated in the Bantu expansion (Watson *et al.* 1997). The single L3d haplotype presents a match with a Fulbe sequence. The full diversity of L3e is present in our sample (Bandelt *et al.* in press). We comment on two informative subclusters. L3e1a is most common in the south in both Bantu and Khoisan-speaking populations, although it has been suggested (Bandelt *et al.* in press) that its origin is further north and that it may have been carried south by the Bantu. L3e4, on the other hand, present in a single individual, was absent until now in the south and has an Atlantic western African distribution. Its presence in Mozambique presents a puzzle, since it suggests recent gene flow from the Atlantic west to southeastern Africa. A similar pattern occurs for haplogroup L1b, also present in one individual in our sample, which is concentrated in western Africa and very rare in the south (one Khwe, Chen *et al.* 2000, which differs at two positions in HVRI from the Mozambique individual).

Sequence matches

In order to investigate whether the contribution of Mozambique sequences to the mtDNA sequence pools of America was higher than in Europe, we searched for matches in a worldwide database. The African and European samples were substantial and widespread, but the Americas were represented by two populations, one from Brazil (Alves-Silva *et al.* 2000) and one from Santo Domingo, in the Caribbean (A. Torroni, unpublished data).

There was a considerable number of matches between Mozambique and American sequences from African haplogroups (Table 6), representing a total of 15

shared sequences in a total of 109 different haplotypes from African haplogroups in the American pool. Two out of these 15 matches correspond to sequences that, in Africa, have only been observed in Mozambique until now; five were not detected in western Africa (one was detected in Galicia, which probably also represents a slave introduction, three in Khoisan-speaking populations, and one in Sudan); and eight were also detected in western Africa. All the Mozambique-American matches for L3e1 were not shared elsewhere except in Khoisan-speaking populations. With respect to those American sequences with no match with Mozambique, there were 25 matches out of 94 different sequences, ten restricted to western African populations.

For the European L sequence pool (Table 7), in a total of 48 different haplotypes four matches were detected with Mozambique, but three of those sequences were also detected in western African populations. Besides the matches with Mozambique, a further nine were detected of which two were western African specific, one northern African, two eastern and four were widespread.

The comparison of the contribution to the American and to the European pools of African sequences via the Atlantic slave trade could be biased by the fact that these sequences had more ancient origins in Europe. Sub-Saharan and north African slaves are known to have been introduced during the Roman Period and also under the Muslim rule in Iberia, and the possibility of earlier, Neolithic contacts should not be discounted. Nevertheless, it is suggestive that the majority of sequences with a sub-Saharan origin within Europe are in Iberia and the Canary Islands (the first colonies of the Iberian kingdoms), which were most extensively involved in the slave trade.

There remained a large number of sequences from African haplogroups sampled in the Americas and Europe for which no match can be found in the current African database. This may be due in part to the fact that the main regions from where slaves were taken, such as Angola and the Slave Coast (Thomas 1998), remain uncharacterised.

CONCLUSIONS

The phylogeographic analysis of the Mozambique sequences in the study has revealed distinct components from the north and the south. An influence from Khoisan-speaking populations was detected as judged by the considerable proportion of distinct L1d sequences, a possible relict of the populations that inhabited this region before the arrival of the Bantu speakers. The Bantu expansion, although originating in a single western core, proceeded in two directions, western and eastern, both towards the south, although only the second reached the very south of Africa. Whether the Bantu mtDNA pool was or was not different in the west and the east remains to be clarified. Comparison with, for instance, Angolans, where some Khoisan-speaking groups are still present, would be essential in order to evaluate Bantu and Khoisan influences in both African coasts. Nowadays in Mozambique, there are no Khoisan-speaking ethnic groups. Although there has been a linguistic replacement, it is unlikely to have been a complete population replacement, as evidenced by the L1d types.

As possible remnants of the Bantu expansion through east towards south Africa, we detected all the haplogroups that have been implicated in this expansion, that is L3b, L3e1a and a subset of L1a sequences. A tentative dating of some L2a sequences, the most frequent haplogroup in the Mozambique sample, by postulating two founder types, as suggested by their low diversity and star-like phylogenies, displayed an age range overlapping the Bantu expansion, although an earlier arrival of these types cannot be excluded. Recent gene flow from Atlantic Africa seems the most probable explanation for the detection of one L1b and one L3e4 sequence in Mozambique.

With respect to the eastern African slave input to America and to Europe, the higher proportion of matches between sequences from Mozambique and the Americas compared to that between Mozambique and Europe, is in accordance with the historical documentation (Thomas 1998) of a differential slave trade, with eastern African slaves more likely to be taken to the Americas. This is particularly striking since other documented factors would have tended to weaken this signal. Firstly, the female/male proportion of the slaves taken to Europe was much higher than to America, and secondly, slave reproduction (particularly from female slave and white owner) was stimulated in Europe (especially after the ban on the importation of slaves after the middle of the eighteenth century) but was been repressed in America (Thomas 1998).

Acknowledgements: We thank Martin Richards and an anonymous referee for suggesting improvements to the manuscript. This work was supported by the following grants: a PhD grant (PRAXIS BD/13632/97) financed by Fundação para a Ciência e a Tecnologia to LP; a Wellcome Trust Career Development Fellowship to VM; by the "Istituto Pasteur Fondazione Cenci Bolognetti", Università di Roma "La Sapienza" (to R.S.), Grandi Progetti Ateneo, Università di Roma "La Sapienza" (to R.S.), Consiglio Nazionale delle Ricerche (99.02620.CT04) (to A.T.), Telethon-Italy E.0890 (to A.T.), Fondo d'Ateneo 2001 dell'Università di Pavia (to A.T.), the Italian Ministry of the University, Progetti Ricerca Interesse Nazionale 1999 and 2001 (to R.S and A.T.). The Mozambique samples were kindly provided by Dr. Albertino Damasceno and Dr. Benilde Soares of the Eduardo Mondlane University (Maputo).

REFERENCES

Alves-Silva, J, Santos, M. D. S., Guimarães, P. E. M., Ferreira, A. C. S., Bandelt, H.-J., Pena, S. D. J. & Prado, V. F. (2000). The ancestry of Brazilian mtDNA lineages. *Am. J. Hum. Genet.* **67**, 444-461.

Anderson, S., Bankier, A.T., Barrell, B.G., De Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I., Nierlich, D., Roc, B., Sanger, F., Schreier, P., Smith, A., Staden, A. & Young, I. (1981). Sequence and organisation of the human mitochondrial genome. *Nature* **290**, 457-465.

Bandelt, H.-J., Forster, P., Sykes, B. C. & Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743-753.

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

Bandelt, H.-J., Macaulay, V. & Richards, M. (2000). Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol. Phylogenet. Evol.* **16**, 8-28.

Bandelt, H.-J., Alves-Silva, J., Guimarães, P. E. M., Santos, M. S., Brehm, A., Pereira, L., Coppa, A., Larruga, J. M., Rengo, C., Scozzari, R., Torroni, A., Prata, M. J., Amorim, A., Prado, V. F. & Pena, S. D. J. Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann. Hum. Genet.* (in press).

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

Budowle, B., Chakraborty, R., Giusti, A. M., Eisenberg, A. J. & Allen, R. C. (1991). Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* **48**, 137-144.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* **325**, 31-36.

Chen, Y.-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S. & Wallace, D. C. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am. J. Hum. Genet.* **57**, 133-149.

Chen, Y.-S., Olckers, A., Schurr, T. G., Kogelnik, A. M., Huoponen, K. & Wallace, D. C. (2000). MtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am. J. Hum. Genet.* **66**, 1362-1383.

Côrte-Real, H., Macaulay, V. , Richards, M. B., Hariti, G., Issad, M. S., Cambon-Thomsen, A., Papiha, A., Bertranpetit, J. & Sykes, B. (1996). Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.* **60**, 331-350.

Di Rienzo, A. & Wilson, A.C. (1991). Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **88**, 1597-1601.

Dimo-Simonin, N., Grange, F., Taroni, F., Brandt-Casadevall, C. & Mangin, P. (2000). Forensic evaluation of mtDNA in a population from south west Switzerland. *Int. J. Legal Med.* **113**, 89-97.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package). Distributed by the author, Department of Genetics, University of Washington.

Forster, P., Harding, R., Torroni, A. & Bandelt, H.-J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. *Am. J. Hum. Genet.* **59**, 935-945.

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

Horai, S. & Hayasaka, K. (1990). Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *Am. J. Hum. Genet.* **46**, 828-842.

Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y., Fucharoen, G., Harihara, S., Park, K. S., Omoto, K. & Pan, I. H. (1996). MtDNA polymorphism in East Asian populations, with special reference to the peopling of Japan. *Am. J. Hum. Genet.* **59**, 579-590.

Ingman, M., Kaessmann, H., Pääbo, S. & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713.

Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos S, Parik, J., Watkins, W. S., Dixon, M. E., Papiha, S. S., Mastana, S. S., Mir, M. R., Ferak, V. & Villems, R. (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Current Biology* **9**, 1331-1334.

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

Lareu, M. V., Phillips, C. P., Carracedo, A., Lincoln, A. J., Syndercombe-court, D. & Thomson, J. A. (1994). Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats; UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci. Int.* **66**, 41-52.

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

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

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

Pereira, L., Prata, M. J. & Amorim, A. (2000). MtDNA diversity in Portugal: not a genetic edge of European variation. *Ann. Hum. Genet.* **64**, 491-506.

Rando, J. C., Pinto, F., González, A. M., Hernández, M., Larruga, J. M., Cabrera, V. M. & Bandelt, H.-J. (1998). Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations. *Ann. Hum. Genet.* **62**, 531-550.

Rando, J. C., Cabrera, V. M., Larruga, J. M., Hernández, M., González, A. M., Pinto, F. & Bandelt, H.-J. (1999). Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.* **63**, 413-428.

Richards, M., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H.-J. & Sykes, B. (1996). Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* **59**, 185-203.

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

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Nørby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozzari, R., Torroni, A. & Bandelt, H.-J. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251-1276.

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. (1999). Genetic evidence for an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat. Genet.* **23**, 437-441.

Saillard, J., Forster, P., Lynnerup, N., Bandelt, H.-J. & Nørby, S. (2000). MtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* **67**, 718-726.

Salas, A., Comas, D., Lareu, M. V., Bertranpetit, J. & Carracedo, A. (1998). mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur. J. Hum. Genet.* **6**, 365-375.

Schneider, S., Roessli, D. & Excoffier, L. (2000). Arlequin ver.2.0: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Soodyall, H. (1993). Mitochondrial DNA polymorphisms in Southern African populations. PhD thesis, University of the Witwatersrand, Johannesburg.

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

Thomas, H. (1998). *The slave trade – the history of the Atlantic slave trade 1440-1870.* London: Macmillan Publishers Ltd.

Thomas, M. G., Parfitt, T., Weiss, D. A., Skorecki, K., Wilson, J. F., le Roux, M., Bradman, N., & Goldstein, D. B. (2000). Y chromosome traveling south: the Cohen modal haplotype and the origins of the Lemba— the "black Jews of southern Africa". *Am. J. Hum. Genet.* **66**, 674-686.

Torroni, A., Schurr, T. G., Yang, C.-C., Szathmary, E. J. E., Williams, R. C., Schanfield, M. S., Troup, G. A., Knowler, W. C., Lawrence, D. N. & Weiss, K. M. (1992). Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics* **130**, 153-162.

Torroni, A., Bandelt, H.-J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savantaus, M.-L., Bonné-Tamir, B. & Scozzari, R. (1998). mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am. J. Hum. Genet.* **62**, 1137-1152.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503-1507.

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

Table 1: Code, place of origin, sample size and bibliographic references for the populations studied here.

Code	Place (ethnic group)	Sample size	Reference
Northern African			
SAH	Western Sahara	25	Rando et al. (1998)
MA	Mauritania	30	Rando et al. (1998)
MO	Morocco	32	Rando et al. (1998)
BM	Morocco (Berber)	60	Rando et al. (1998)
EGY	Egypt	68	Krings et al. (1999)
MZB	Algeria (Mozabite)	86	Côrte-Real et al. (1996); Macaulay et al. (1999)
Western African			
HA+KA	Niger (Hausa and Kanuri)	20+14	Watson et al. (1997)
FUL	Nigeria (Fulbe)	60	Watson et al. (1997)
SON+TU	Nigeria (Songhai and Tuareg)	10+23	Watson et al. (1997)
YOR	Nigeria (Yoruba)	21+14	Watson et al. (1997); Vigilant et al. (1991)
SEN	Senegal	50	Rando et al. (1998)
SER	Senegal (Serer)	23	Rando et al. (1998)
WO	Senegal (Wolof)	48	Rando et al. (1998)
MAN	Senegal (Mandenka)	119	Graven et al. (1995)
Central African			
MBU	Zaire (Mbuti)	20	Vigilant et al. (1991)
BIA	Central African Republic (Biaka)	17	Vigilant et al. (1991)
Eastern African			
TK	Kenya (Turkana)	36	Watson et al. (1997)
SO	Somalia	27	Watson et al. (1997)
KIK	Kenya (Kikuyu)	25	Watson et al. (1997)
NUB	Nubia	80	Krings et al. (1999)
SUD	Southern Sudan	76	Krings et al. (1999)
South-eastern African			
MOZ	Mozambique	109	This work
Southern African			
KNG1	Botswana (!Kung)	25	Vigilant et al. (1991)
KNG2	South Africa (!Kung)	43	Chen et al. (2000)
KWE	South Africa (Khwe)	31	Chen et al. (2000)
HER	South Africa (Herero)	26	Vigilant et al. (1991)

Table 2: Diversity measures in Mozambique within HVRI and HVRII.

	Haplotypes [1]	Segregating sites [2]	Gene diversity [3]	Mean pairwise difference
HVRI	50 (45.9)	60 (16.7)	0.962 ± 0.008	7.86
HVRII	35 (32.1)	29 (10.8)	0.846 ± 0.032	5.18
HVRI+HVRII	64 (58.7)	89 (14.1)	0.973 ± 0.007	13.04

[1] Number of distinct haplotypes in sample (percentage of sample size)

[2] Number of sites variable in sample (percentage of all sites)

[3] Average heterozygosity ± standard error

Table 3: HVRI (from np 16090 to 16365) diversity and neutrality measures in African populations.

Population	Sample size	Haplotypes [1]	Segregating sites [2]	Gene diversity [3]	Mean pairwise differences	D [4]	F_s [5]
Northern African							
SAH	25	20 (80.0)	29 (10.5)	0.973 ± 0.022	5.11	−1.25	−12.4 ***
MA	30	22 (73.3)	28 (10.1)	0.970 ± 0.018	5.83	−0.63	−11.5 ***
MO	32	29 (90.6)	44 (15.9)	0.988 ± 0.014	5.84	−1.70 *	−25.0 ***
BM	60	38 (63.3)	47 (17.0)	0.963 ± 0.015	4.44	−1.88 **	−25.7 ***
EGY	68	59 (86.8)	66 (23.9)	0.993 ± 0.005	6.82	−1.70 **	−25.1 ***
MZB	85[6]	29 (34.1)	35 (12.7)	0.942 ± 0.010	4.73	−1.02	−11.1 ***
Western African							
HA+KA	34	31 (91.2)	41 (14.9)	0.995 ± 0.009	6.19	−1.38	−25.2 ***
FUL	60	38 (63.3)	43 (15.6)	0.972 ± 0.010	6.82	−0.87	−23.2 ***
SON+TU	33	29 (87.9)	41 (14.9)	0.992 ± 0.009	7.26	−1.02	−21.3 ***
YOR	34[6]	32 (94.1)	44 (15.9)	0.996 ± 0.008	7.31	−1.16	−25.0 ***
SEN	50	42 (84.0)	41 (14.9)	0.989 ± 0.008	6.24	−1.08	−25.2 ***
SER	23	21 (91.3)	40 (14.5)	0.992 ± 0.015	8.09	−0.98	−12.0 ***
WO	48	39 (81.3)	42 (15.2)	0.991 ± 0.006	7.50	−0.71	−25.0 ***
MAN	110[6]	46 (41.8)	47 (17.0)	0.963 ± 0.008	6.23	−0.94	−24.5 ***
Central African							
MBU	13[6]	5 (38.5)	19 (6.9)	0.756 ± 0.097	7.13	0.70	3.8
BIA	17	8 (47.1)	20 (7.2)	0.890 ± 0.043	7.81	1.27	1.7
Eastern African							
TK	36	32 (88.9)	54 (19.6)	0.991 ± 0.010	9.66	−0.94	−20.8 ***
SO	27	24 (88.9)	41 (14.9)	0.992 ± 0.013	6.90	−1.32	−16.3 ***
KIK	25	23 (92.0)	45 (16.3)	0.993 ± 0.013	7.96	−1.27	−14.6 ***
NUB	80	50 (62.5)	64 (23.2)	0.974 ± 0.008	7.88	−1.29	−24.8 ***
SUD	76	63 (82.9)	73 (26.4)	0.993 ± 0.004	8.33	−1.47 *	−24.8 ***
MOZ	109	49 (45.0)	57 (20.7)	0.960 ± 0.008	7.78	−0.89	−23.6 ***
Southern African							
KNG1	24[6]	9 (37.5)	16 (5.8)	0.830 ± 0.053	2.97	−1.10	−1.3
KNG2	43	12 (27.9)	31 (11.2)	0.812 ± 0.045	7.30	0.07	1.8
KWE	31	10 (32.3)	34 (12.3)	0.884 ± 0.029	8.75	0.10	3.0

[1] Number of distinct haplotypes in sample (percentage of sample size)

[2] Number of sites variable in sample (percentage of all sites)

[3] Average heterozygosity ± standard error

[4] Tajima's D statistic (P-value: * = $0.01 < P \leq 0.05$; ** = $0.001 < P \leq 0.01$; *** = $P \leq 0.001$)

[5] Fu's F_s statistic (P-value: * = $0.01 < P \leq 0.05$; ** = $0.001 < P \leq 0.01$; *** = $P \leq 0.001$)

[6] Some sequences were not considered for this analysis since there were many positions not scored.

Table 4: Number of sampled individuals in Mozambique by haplogroup (and frequency with standard error).

Haplogroup	Frequency (%)
L1a	16 (14.7 ± 3.4)
L1b	1 (0.9 ± 0.9)
L1c	5 (4.6 ± 2.0)
L1d	8 (7.3 ± 2.5)
L1e	2 (1.8 ± 1.3)
L2a	47 (43.1 ± 4.7)
L2b	2 (1.8 ± 1.3)
L2c	1 (0.9 ± 0.9)
L3*	2 (1.8 ± 1.3)
L3b	4 (3.7 ± 1.8)
L3d	2 (1.8 ± 1.3)
L3e1*	9 (8.3 ± 2.6)
L3e1a	4 (3.7 ± 1.8)
L3e2a	1 (0.9 ± 0.9)
L3e2b	2 (1.8 ± 1.3)
L3e3	2 (1.8 ± 1.3)
L3e4	1 (0.9 ± 0.9)

Table 5: Haplogroup L2a diversity in Africa.

	Heterozygosity	Frequency (%)	Mean pairwise difference
Eastern African	0.953 ± 0.023	20.5	3.82
Western African	0.952 ± 0.013	19.0	3.14
Northern African	0.910 ± 0.068	4.3	2.36
Mozambique	0.832 ± 0.031	43.1	2.24
Central African	0.733 ± 0.155	29.7	2.27
Southern African	0.000 ± 0.000	3.0	0.00

Table 6: Number of matches between Mozambique (MOZ), Brazil (BRZ) and Santo Domingo (SD) sequences, and also matches for those sequences inside Africa. Information based only on HVRI between 16051 and 16362.

MOZ	BRZ [1]	SD [2]	Other	HVRI sequence	Haplo-group
1	1	—	—	093 129 148 168 172 187 188^{G} 189 223 230 278 293 311 320	L1a
8	2	—	2Equatorial Guinea[3] 1MO 1KIK	129 148 168 172 187 188^{G} 189 223 230 278 293 311 320	L1a
6	1	4	1Iraq[4] 1TK	148 172 187 188^{G} 189 223 230 311 320	L1a
1	—	1	3MAN 1SER 1SEN 1Portugal[5]	126 145 187 189 223 264 270 278 293 311	L1b
1	2	2	1SUD	129 163 187 189 209 223 278 293 294 311 360	L1c
2	1	3	2Canary[6] 1SER 1SEN 2WO 1SAH 1Syria[4] 1Equatorial Guinea[3] 1Portugal[5] 1SUD 1TU 1SO 1KIK 1FUL 1YOR	223 278 294 309	L2a
1	1	1	2KNG2	114^{A} 129 213 223 278 354	L2b
1	—	4	1Galicia[7]	093 124 223 278 362	L3b
2	—	1	1SUD 2NUB 1HA	124 223 278 311 362	L3b
2	3	—	1!Kung[8]	176 223 327	L3e1*
1	1	1	—	223 327	L3e1*
3	1	—	1Dama[8]	185 223 327	L3e1a
2	3	5	2KNG2 3KWE 1WO 1Syria[4] 1HA 1FUL 2MZB 1Israel[4]	172 189 223 320	L3e2b
2	—	1	1Israel[4] 1Equatorial Guinea[3]	223 265^{T}	L3e3
1	—	1	1SUD 1WO 4MAN	051 223 264	L3e4

Note: Besides the populations referred in Table 1, the survey included other populations with the following bibliographic references: [1] Alves-Silva et al. (2000); [2] A. Torroni (unpublished data); [3] Mateu et al. (1997); [4] Richards et al. (2000); [5] Pereira et al. (2000); [6] Rando et al. (1999); [7] Salas et al. (1998); [8] Soodyall (1993).

Table 7: Matches for sequences observed in Europe from African haplogroups to sequences in Africa, including Mozambique (MOZ). Information based only on HVRI between 16051 and 16362. African haplotypes observed in Europe but not observed in Africa: L1a (3); L1b (4); L2a (7); L3 (10); L3b (3); L3d (1); L3e (4).*

Europe	MOZ	Africa	HVRI sequence	Haplo-group
1Sardinia[9]	—	1TK 2MAN 9NUB 2SUD	129 148 168 172 187 188[G] 189 223 230 311 320	L1a
1Canary[6]	—	5FUL 1WO 1Equatorial Guinea[3]	093 126 187 189 223 264 270 278 293 311	L1b
1Portugal[5]	1	1SEN 1SER 3MAN	126 145 187 189 223 264 270 278 293 311	L1b
2Canary[6]	—	1EGY 1MO 1MA 2WO 1SER 1KWE 2Equatorial Guinea[3]	126 187 189 223 264 270 278 311	L1b
1Portugal[5]	—	1SEN 1FUL	093 189 192 223 278 294 309	L2a
1Italy[4]	10	1SO 1SEN	223 278 286 294 309	L2a
1Portugal[5] 2Canary[6]	2	1SAH 2WO 1SER 1SEN 1YOR 1FUL 1Equatorial Guinea[3] 1SUD 1TU 1SO 1KIK	223 278 294 309	L2a
1Portugal[5] 1Portugal (born in Angola)[10] 1Canary[6]	—	1FUL 1SUD 1MO	209 223 311	L3*
1Portugal[5] 1Portugal[10]	—	1NUB 2SUD 1MA	223	L3*
1Basque[4]	—	1KIK	223 311	L3*
1Switzerland[11]	—	1SUD	176 188 209 223 234 311 355	L3*
1Galicia[7]	1	—	093 124 223 278 362	L3b
1Spain[12]	—	1MO	124 223 234 278 362	L3b

Note: [1] to [8] as in Table 6; [9] Di Rienzo et al. (1991); [10] Côrte-Real et al. (1996); [11] Dimo-Simonin et al. (2000); [12] Richards et al. (1996).

FIGURE LEGENDS

Fig 1. A schematic phylogeny of African haplogroups used in our classification of sequences from Mozambique. We have drawn on information from Watson *et al.* (1997), Chen *et al.* (2000) (although our naming scheme is different from theirs), Alves-Silva *et al.* (2000), Richards and Macaulay (2000), Ingman *et al.* (2000), Bandelt *et al.* (in press) and also on some unpublished information. Triangles represent well-characterized clades in the mtDNA phylogeny. Circles and ellipses correspond to possibly paraphyletic groupings of less well-characterized haplotypes. The node marked "L3*" has the motif 16223T, 073G, 263G in the control region with respect to the CRS. Mutations (shown on the branches) are transitions unless a nucleotide is specified. Underlining of a position indicates that it mutates more than once in the figure. Four diagnostic RFLPs are also shown since they were checked in a number of samples; the direction of site loss and gain is indicated with respect to the node L3*. We keep the existing L1 nomenclature despite the fact that L1 is not a clade: it harbours the root of the tree, which occurs on a branch separating L1a and certain lineages present in Khoisan-speaking populations from the rest of the tree (Ingman *et al.* 2000). There are still uncertainties in the phylogeny, e.g., whether the 16124 mutation occurs independently in L3b and L3d: however these ambiguities did not affect the classification of the haplotypes reported here.

Fig. 2. Mismatch distributions for HVRI and HVRII in Mozambique.

Fig. 3. The first two principal components of haplogroup frequency profiles for all African populations.

Fig. 4. The first two principal components of haplogroup frequency profiles for western and eastern African populations, including the Mbuti.

Fig. 5. The reduced median network of the 50 L2 sequences in the Mozambique sample. The circles are combined HVRI/HVRII haplotypes, the areas of which are proportional to the frequency in the sample. The smallest circles are singletons, the largest has frequency 12. Mutations (shown on the branches) are transitions unless a base change is explicitly indicated. Underlining indicates resolved recurrent mutations, unresolved events being shown by reticulation. The solid node (not observed in the

sample) has the motif 16223-16278-16390-073-146-152-195-263. Branches are shown compressed in L2b and L2c for convenience. Two putative founder sequences are indicated by the symbols † and ‡.

Fig. 6. The reduced median network of the 27 L3 sequences in the Mozambique sample. The circles are combined HVRI/HVRII haplotypes, the areas of which are proportional to the frequency in the sample. The smallest circles are singletons, the largest has frequency two. Also included is information on the RFLP markers assayed (the arrows indicate the direction of a site gain). Control-region mutations (shown on the branches) are transitions unless a base change is explicitly indicated. Underlining indicates resolved recurrent mutations, unresolved events being shown by reticulation. The node marked with an asterisk has the motif 16223-073-150-263. The evolution at hypervariable sites such as 150, 152 and 195 (Bandelt *et al.* 2000) is probably not accurately reconstructed. For example, there is coding-region information to suggest that L3e3 is more closely related to L3e4 than to L3e2 (Bandelt *et al.* in press).

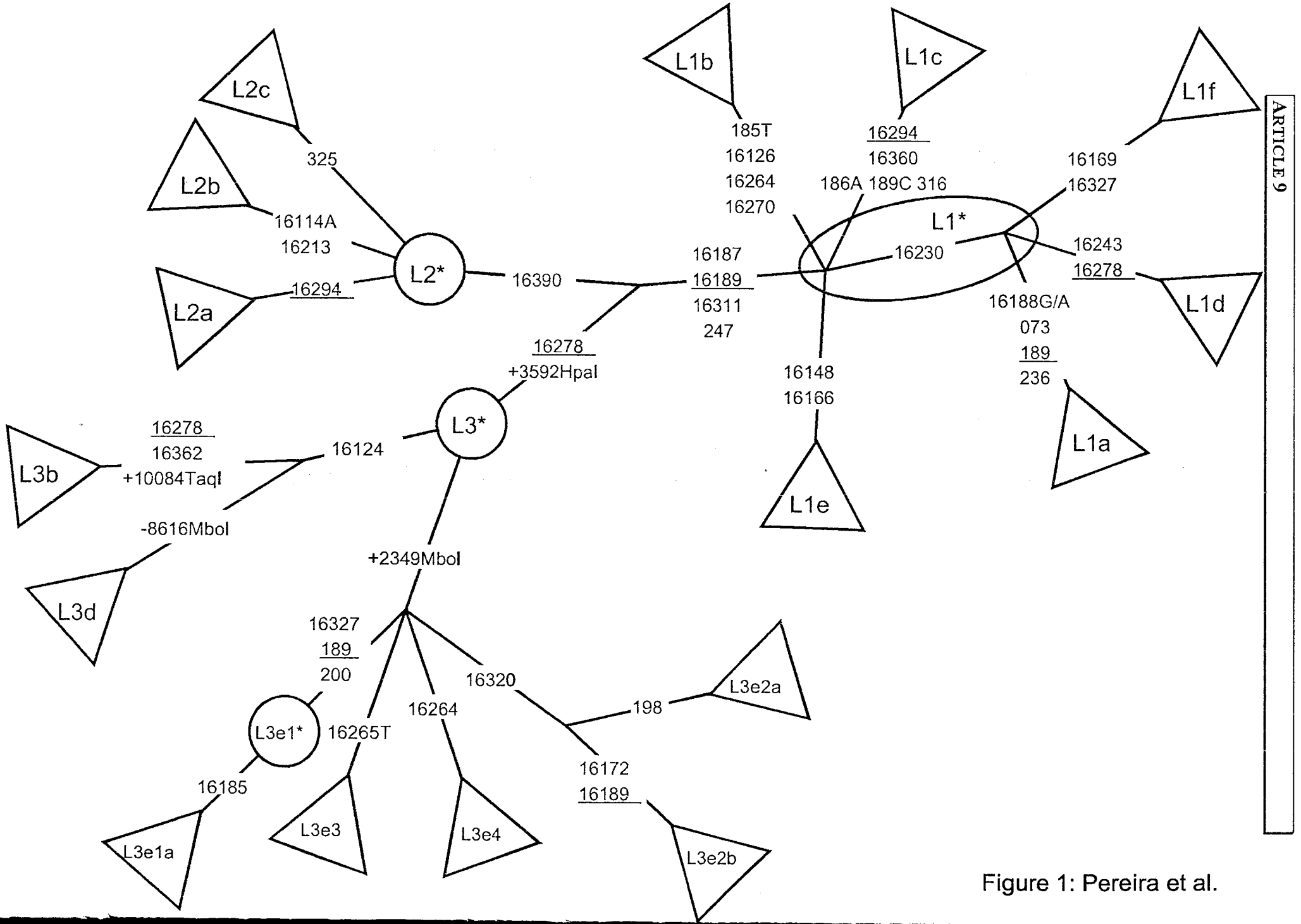
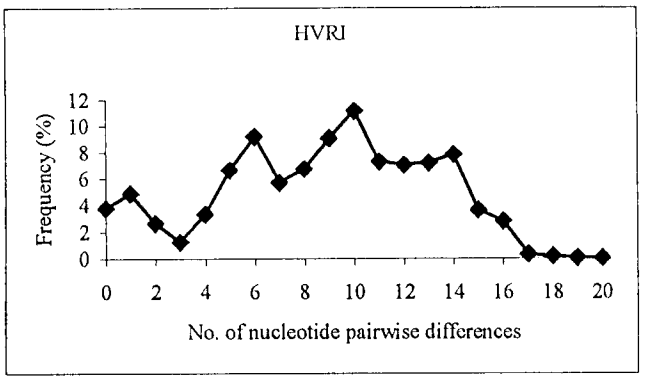
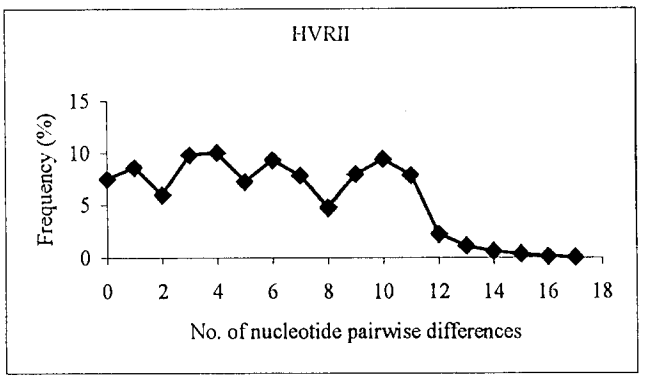


Figure 1: Pereira et al.





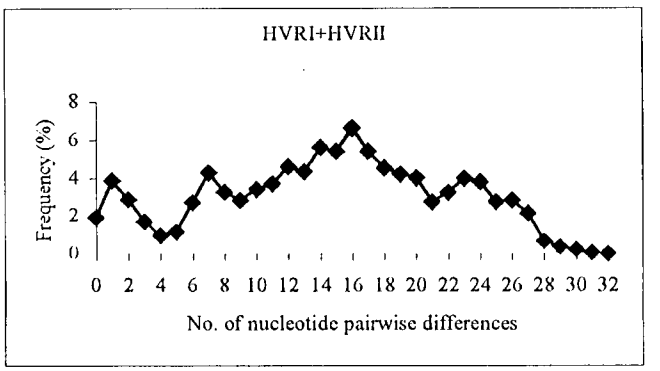


Figure 2: Pereira et al.

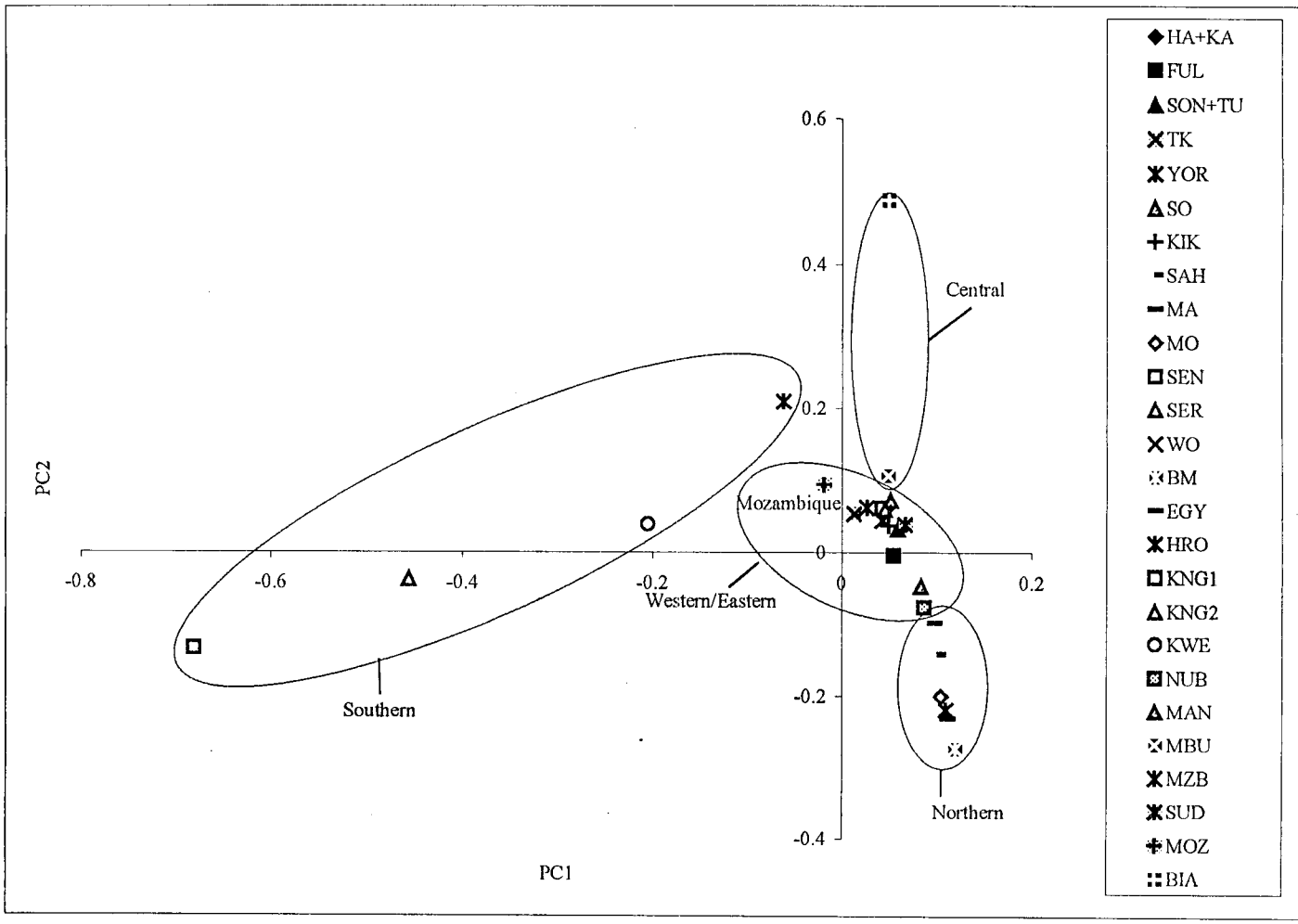


Figure 3: Pereira et al.

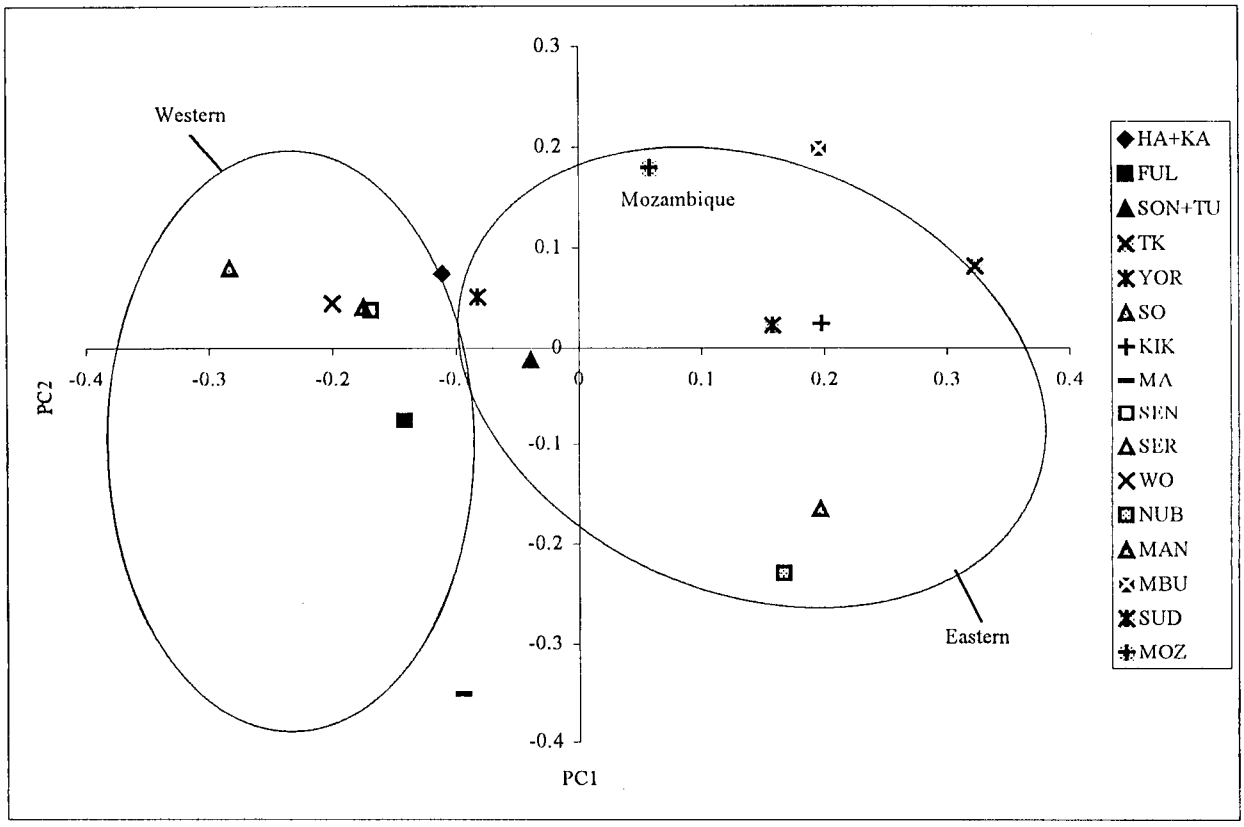


Figure 4: Pereira et al.

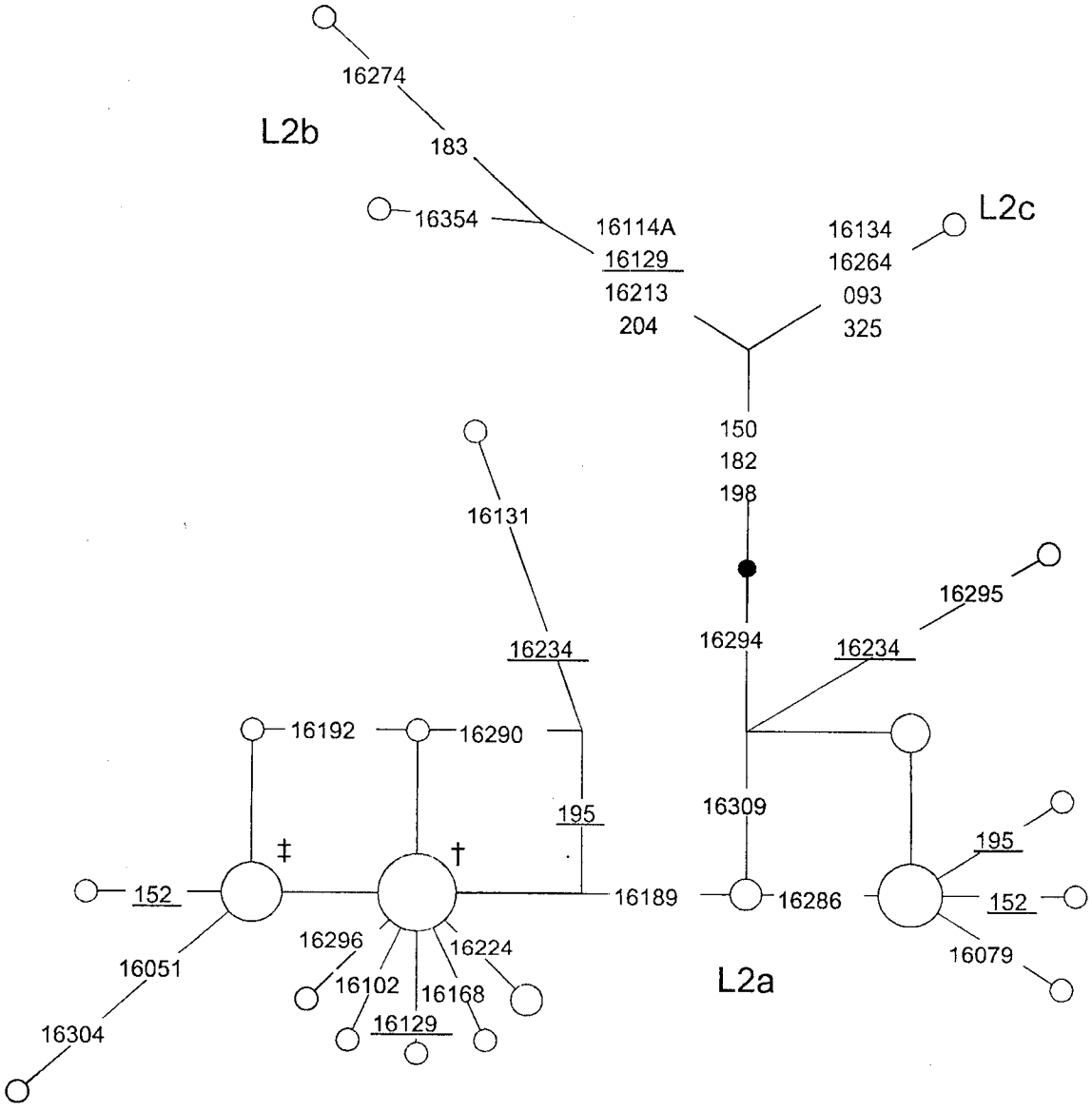


Figure 5: Pereira et al.

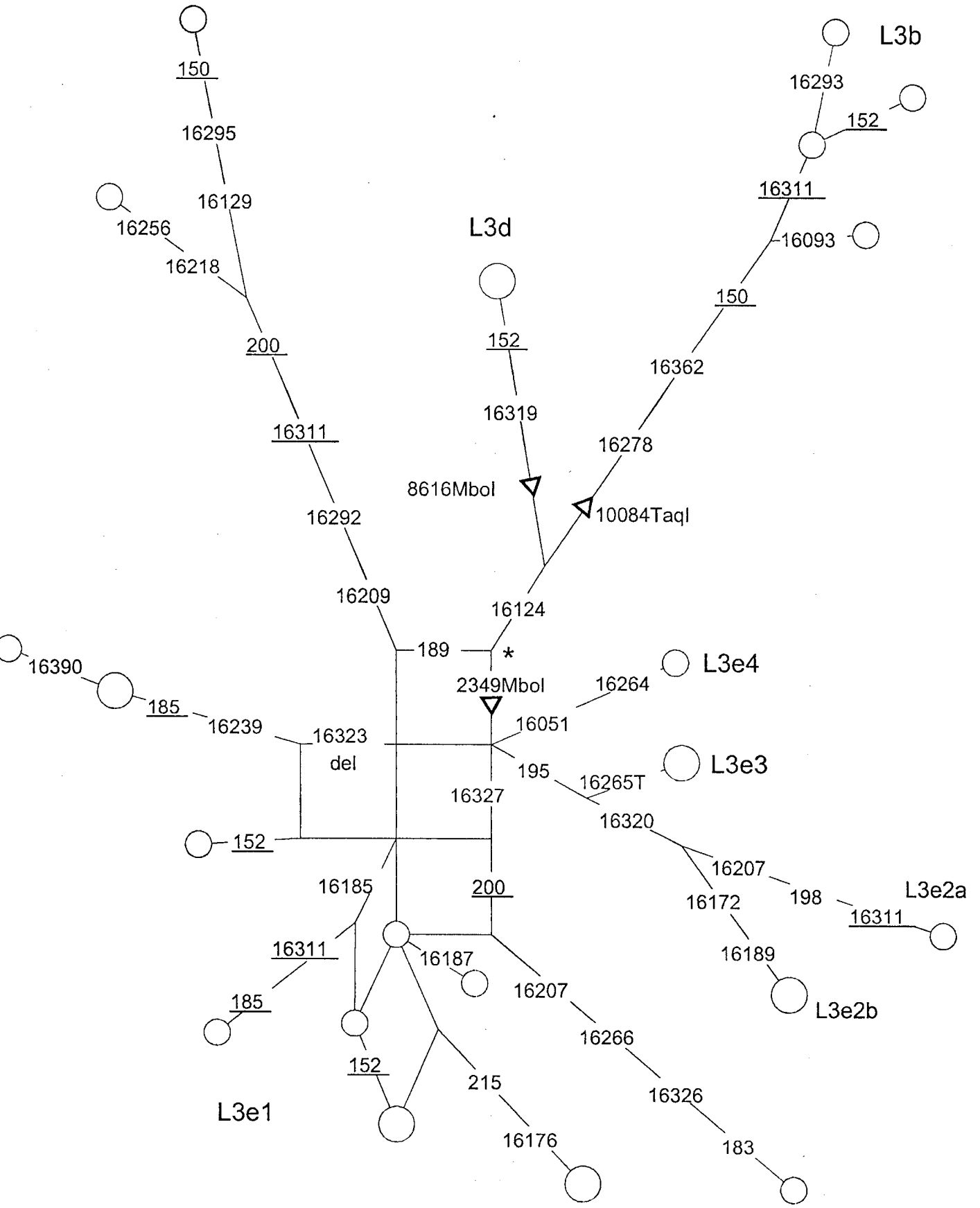


Figure 6: Pereira et al.

Appendix 1. *Control region sequences in Mozambique*

Variant positions from the CRS are shown between 16017-16390 in HVRI (minus 16000) and 73-340 in HVRII. The substitution indicated in italics was observed to be heteroplasmic. Substitutions are transitions unless the base change or a deletion is explicitly indicated. Insertions of one and two cytosines are shown by appending '.1' and '.2', respectively. Also shown are the results of the partial RFLP typing. "+" indicates the presence of the restriction site, a "–" the absence, "n/a" = not available. The haplogroup assignment is discussed in the text.

Freq.	HVRI sequence	HVRII sequence	RFLP				Haplo-group
			3592*Hpa*I	10084*Taq*I	8616*Mbo*I	2349*Mbo*I	
1	93 129 148 168 172 187 188$^{C/G}$ 189 223 230 278 293 311 320	93 95$^{A/C}$ 185 189 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
5	129 148 168 172 187 188$^{C/G}$ 189 223 230 278 293 311 320	93 95$^{A/C}$ 185 189 236 247 263 303.1 311.1	n/a	n/a	n/a	n/a	L1a
1	129 148 168 172 187 188$^{C/A}$ 189 223 230 278 293 311 320	93 95$^{A/C}$ 185 189 236 247 263 303.1 311.1	n/a	n/a	n/a	n/a	L1a
2	129 148 168 172 187 188$^{C/G}$ 189 223 230 278 293 311 320	93 95$^{A/C}$ 185 189 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
1	129 148 168 172 187 188$^{C/G}$ 189 223 230 278 293 311 320	93 95$^{A/C}$ 185 189 236 247 263 303.2 311.1	n/a	n/a	n/a	n/a	L1a
1	148 172 187 188$^{C/G}$ 189 223 230 311 320	93 150 152 189 204 207 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
1	148 172 187 188$^{C/G}$ 189 223 230 311 320	93 152 189 199 204 207 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
3	148 172 187 188$^{C/G}$ 189 223 230 311 320	93 152 189 204 207 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
1	148 172 187 188$^{C/G}$ 189 223 230 311 320	93 152 189 236 247 263 311.1	n/a	n/a	n/a	n/a	L1a
1	126 145 187 189 223 264 270 278 293 311	73 152 182 185$^{G/T}$ 195 247 263 311.1 357	n/a	n/a	n/a	n/a	L1b
1	17 129 163 187 189 209 223 278 293 294 311 360	73 151 152 182 186$^{C/A}$ 189$^{A/C}$ 247 263 311.1 316	n/a	n/a	n/a	n/a	L1c
1	17 129 187 189 223 278 293 294 311 360	73 151 152 182 186 $^{C/A}$ 189 $^{A/C}$ 247 263 303.1 311.1 316	n/a	n/a	n/a	n/a	L1c
1	71 129 145 187 189 213 223 234 265$^{A/C}$ 278 286$^{C/G}$ 294 311 360	73 151 152 182 186$^{C/A}$ 189$^{A/C}$ 195 198 247 263 297 311.1 316	n/a	n/a	n/a	n/a	L1c
1	71 145 187 189 213 223 234 265$^{A/C}$ 278 286$^{C/G}$ 294 311 360	73 93 151 152 182 186$^{C/A}$ 189$^{A/C}$ 195 198 247 263 297 303.1 311.1 316	n/a	n/a	n/a	n/a	L1c
1	129 183$^{A/C}$ 189 215 223 278 294 311 360	73 151 152 182 186$^{C/A}$ 189$^{A/C}$ 247 263 303.1 311.1 316	n/a	n/a	n/a	n/a	L1c
1	129 145 187 189 212 223 230 243 311 390	73 146 152 195 198 247 311.1	n/a	n/a	n/a	n/a	L1d
1	129 187 189 212 223 230 243 291 311	73 146 152 188 195 247 295 311.1	n/a	n/a	n/a	n/a	L1d
1	129 187 189 212 223 230 243 291 311	73 146 152 188 195 228 247 311.1	n/a	n/a	n/a	n/a	L1d

2	129 187 189 223 230 243 311	73 146 152 195 247^{delG} 294$^{T/A}$ 311.1	n/a	n/a	n/a	n/a	L1d
1	129 187 189 223 230 243 311 390	73 146 152 195 198 247 311.1	n/a	n/a	n/a	n/a	L1d
1	129 187 189 223 239 243 294 311	73 146 152 195 247 311.1	n/a	n/a	n/a	n/a	L1d
1	187 189 223 230 234 243 294$^{C/G}$ 311	73 146 152 195 247 311.1	n/a	n/a	n/a	n/a	L1d
1	129 148 166 183^{delA} 187 189 192 223 278 311 355 362	73 152 182 247 263 311.1	n/a	n/a	n/a	n/a	L1e
1	129 148 166 183^{delA} 187 189 192 223 278 311 355 362 390	73 152 182 195 247 263 311.1	n/a	n/a	n/a	n/a	L1e
1	223 234 278 294 295 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
1	51 182$^{A/C}$ 183$^{A/C}$ 189 192 223 278 290 294 304 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	79 223 278 286 294 309 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
1	102 182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	129 182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	131 189 223 234 278 294 309 390	73 146 152 263 303.2 311.1	n/a	n/a	n/a	n/a	L2a
1	168 182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
3	182$^{A/C}$ 183$^{A/C}$ 189 192 223 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	182$^{A/C}$ 183$^{A/C}$ 189 192 223 278 290 294 309 390	73 146 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
2	182$^{A/C}$ 183$^{A/C}$ 189 223 224 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
11	182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 309 390	73 146 152 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
4	182$^{A/C}$ 183$^{A/C}$ 189 192 223 278 290 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	182$^{A/C}$ 183$^{A/C}$ 189 192 223 278 290 294 309 390	73 146 152 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
1	182$^{A/C}$ 183$^{A/C}$ 189 223 278 290 294 296 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
5	223 278 286 294 309 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
2	223 278 286 294 309 390	73 146 152 195 263 303.2 311.1	n/a	n/a	n/a	n/a	L2a
1	223 278 286 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	223 278 286 294 309 390	73 146 152 263 303.2 311.1	n/a	n/a	n/a	n/a	L2a
1	223 278 286 294 309 390	73 146 195 263 311.1	n/a	n/a	n/a	n/a	L2a
3	223 278 286 294 390	73 146 152 195 263 303.1 311.1	n/a	n/a	n/a	n/a	L2a
1	223 278 294 309 390	73 146 152 195 263 303.2 311.1	n/a	n/a	n/a	n/a	L2a
1	223 278 294 309 390	73 146 152 195 263 311.1	n/a	n/a	n/a	n/a	L2a
1	114$^{C/A}$ 129 213 223 274 278 390	73 146 150 152 182 183 195 198 204 263 303.1 311.1	n/a	n/a	n/a	n/a	L2b
1	114$^{C/A}$ 129 213 223 278 354 390	73 146 150 152 182 195 198 204 263 311.1	n/a	n/a	n/a	n/a	L2b
1	134 223 264 278 390	73 93 146 150 152 182 195 198 263 311.1 325	n/a	n/a	n/a	n/a	L2c
1	129 209 223 292 295 311	73 189 *200* 263 303.1 311.1	–	–	+	–	L3*
1	209 218 223 256 292 311	73 150 189 200 263 311.1	–	–	+	–	L3*
1	93 124 223 278 362	73 263 311.1	–	+	+	–	L3b
1	124 223 278 293 311 362	73 263 303.1 311.1	–	+	+	–	L3b

1	124 223 278 311 362	73 152 263 311.1	–	+	+	–	L3b
1	124 223 278 311 362	73 263 311.1	–	+	+	–	L3b
2	124 223 319	73 150 152 263 303.1 311.1	–	–	–	–	L3d
2	223 239 323^{delT}	73 150 185 189 263 303.1 311.1	–	–	+	+	L3e1*?
1	223 239 323^{delT} 390	73 150 185 189 263 303.1 311.1	–	–	+	+	L3e1*?
1	223 323^{delT} 327	73 150 152 189 263 303.1 311.1	–	–	+	+	L3e1*
2	176 223 327	73 150 152 189 200 215 263 311.1	–	–	+	+	L3e1*
1	187 223 327	73 150 189 200 263 311.1	–	–	+	+	L3e1*
1	207 223 266 326 327	73 150 183 200 263 303.1 311.1	–	–	+	+	L3e1*
1	223 327	73 150 189 200 263 311.1	–	–	+	+	L3e1*
1	185 223 311 327	73 150 185 189 263 311.1	–	–	+	+	L3e1a
2	185 223 327	73 150 152 189 200 263 311.1	–	–	+	+	L3e1a
1	185 223 327	73 150 189 200 263 303.1 311.1	–	–	+	+	L3e1a
1	207 223 311 320	73 150 195 198 263 311.1	–	–	+	+	L3e2a
2	172 183$^{A/C}$ 189 223 320	73 150 195 263 311.1	–	–	+	+	L3e2b
2	223 265$^{A/T}$	73 150 195 263 311.1	–	–	+	+	L3e3
1	51 223 264	73 150 263 311.1	–	–	+	+	L3e4

Total: 109

Y-chromosome pool in the southeastern African population of Mozambique: the small European influence and the Bantu diversity reduction

PEREIRA, L.[1,2], GUSMÃO, L.[1], ALVES, C.[1], AMORIM, A.[1,2], PRATA, M.J.[1,2]

[1] *Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200 Porto, PORTUGAL*

[2] *Faculdade de Ciências da Universidade do Porto, Pr. Gomes Teixeira, 4050 Porto, PORTUGAL*

Running head: Bantu and European Y-chromosome lineages in Mozambique

Key words: Y-BM, Y-STR, haplogroups, haplotypes, matching

Correspondence:
Luísa Pereira
Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP),
R. Dr. Roberto Frias s/n, 4200 Porto, PORTUGAL
Phone: +351 22 5570700
Fax: +351 22 5570799
email: lpereira@ipatimup.pt

Summary

Analysis of the Mozambican Y-chromosome pool led to the conclusion that the genetic impact of Portuguese colonialism was low. In fact, undisputable European lineages amount to just 1.5%, and even adding all the haplogroups of uncertain origin (including some most likely sub-Saharan), the upper limit is 26.5%. Haplotype matching performed for putative Bantu lineages in several sub-Saharan populations allowed to confirm the increasing founder effect towards south, with stronger reduction of diversity along the western coast. A mixed frequency distribution for the Bantu haplotypes in South Africa, relatively to the western and eastern packages, seems to be evidence for the intermingling between both Bantu waves in that region.

Introduction

Y-chromosome markers, in contrast to mtDNA and autosomal loci, are the only ones for which sub-Saharan populations do not show the highest diversity at a worldwide scale. This has been observed not only for Y-biallelic markers (Y-BMs) (Underhill et al., 2001), but also for Y-STRs (Pritchard et al., 1999). The failure to detect intermediates for deeply divergent Y-SNPs lineages is consistent with a model where several episodes of population contractions have eliminated diversity accumulated during periods of expansion (Underhill et al., 2001). One of these episodes occurred during the last three thousand years, through the expansion and migration of Bantu populations towards sub-Sahel regions. Underhill et al. (2001) have suggested as markers of that expansion the haplotypes defined by M2/PN1/M180 polymorphisms, which are the analogues of haplogroup 8 defined by YAP^+/sY81G polymorphisms, in the nomenclature of Rosser et al. (2000). The first authors presented evidence of strong founder effects in that sub-clade (~40% of the members share the M191 mutation), which was independently supported by results from Y-STR haplotypes in a South African Bantu population (Thomas et al., 2000), where the proportion of YAP^+/sY81G lineages was 80.5%, from which more than a half shared the same 6 Y-STR based haplotype or its one-step neighbours.

Mozambique is a southeast African country located along one of the two eastern routes by which Bantus reached the southern tip of the continent (along the Ruvuma River toward the coast, reaching present-day Natal by the end of the third century A.D.), as a result of the re-expansion of the eastern Bantu core area, in present Uganda, around 3,000 years ago (Newman, 1995). Linguistically, Khoisan replacement was total, since all the 33 languages belong to the Bantu group of languages within the South-Central Niger-Congo family (http://www.sil.org/ethnologue/countries/Moza.html). However, mtDNA shows that some admixture must have occurred since some Khoisan lineages are still detected in Mozambique (Pereira et al., 2001a). In fact, 7.3% of the mtDNA sequences belong to the L1d haplogroup, considered to be a Khoisan marker (Watson et al., 1997). Anyway, a considerable proportion (18%) of the Mozambican mtDNA lineages belong to those tagged as markers of Bantu expansion (Soodyall et al., 1996; Watson et al., 1997; Bandelt et al., in press). Nevertheless, the most frequent (43.1%) Mozambican haplogroup is L2a, which is not currently considered as a Bantu marker.

However, its estimated age of 6,700±2,100 (Pereira et al., 2001a), is coincident with Bantu expansion, and shows a pattern of southwards diversity reduction, in agreement to what is observed for Y-chromosome, at a deeper scale.

It has been harder to establish the genetic impact of the western Bantu wave of expansion, since no mtDNA data have been published yet for this region. It is worth to mention that some Khoisan-speaking groups are still present in Angola. The two waves of expansion are thought to have got in contact around 2,000 years ago, in the region of nowadays Zimbabwe (Newman, 1995).

Mozambique was a Portuguese colony between 1752 and 1975 (although trading posts were established as early as 1493). Portuguese colonialism was characterised (Russell-Wood, 1998) by: (1) the presence of extraordinary low numbers of Portuguese (except in Brazil); (2) which were predominantly males; (3) and inter-breed with local females. From all Portuguese colonies in Asia, America and Africa, the smallest demographic impact was in this last continent (Russell-Wood, 1998). At the mtDNA level, the European influence in Mozambique was proved to be nil (Pereira et al., 2001), but the male counterpart was not yet studied.

The aims of this study were: (1) to evaluate the proportion of European haplogroups detected at the Mozambican Y gene pool; and (2) to investigate, by haplotype matching analysis, if the detected Y-STR diversity patterns in the context of the available data from sub-Saharan populations were consistent with the reduction associated with the Bantu expansion.

MATERIAL AND METHODS

Subjects

We studied a sample of unrelated individuals from Mozambique, belonging to different ethnic groups (Changana, 25; Ronga, 15; Chope, 8; Matsua, 6; Tonga, 5; and the remaining to various other groups), but all Bantu speakers (http://www.sil.org/ethnologue/countries/Moza.html). The DNA was extracted from blood spots by the resin Chelex-100 method (Lareu et al., 1994). A total of 68 individuals was analysed for Y-BMs, from which 66 were further characterised for Y-STRs.

Y-chromosome markers

The Y-BMs analysed in this work were the 10 reported by Rosser et al. (2000), being 8 single nucleotide polymorphisms (SRY-8299, 92R7, SRY-1532, SRY-2627, Tat, sY81, M9 and LLY22g) and 2 insertion/deletion polymorphisms (YAP and 12f2). The screening conditions and nomenclature used were according to Rosser et al. (2000). Following those authors, we will use the notation "haplogroup" for describing a particular combination of Y-BMs. In order to make the diversity data comparable with those reported in other populations (Rosser et al., 2000; Pereira et al., 2001b), we considered also the marker DYS257, as in Figure 1 of Rosser et al. (2000).

Seven short tandem repeats (STRs) were analysed: DYS19, DYS388I, DYS388II, DYS390, DYS391, DYS392 and DYS393. The order just described will be followed from now on when referring to a certain "haplotype" (notation used to refer a particular combination of Y-STRs). PCR amplification was done in two multiplex reactions: a pentaplex system as described by González-Neira et al. (2000) for the markers DYS19, DYS388I, DYS388II, DYS390 and DYS393; and a duplex reaction using 0.12μM and 0.2μM of each pair of primers described in Kayser et al. (1997) for DYS391 and DYS392, respectively. Genotyping was performed using the ABI Prism 310 sequencer, using the automatic fragment sizing software provided by the manufacturer (AB Applied Biosystems).

Statistical analyses

The diversity measures were obtained by using the Arlequin 2.0 (Schneider et al., 2000). Y-STR haplotypes were either coded (1) in the standard way, considering the

number of repeat units observed in each locus, or (2) as a DNA strand, where each repeat unit is considered as an ambiguous position (N) and differences in the number of repeat units are considered as insertion (N) or deletion (-). In the first case, we measure how many loci are different between pairs of haplotypes; while in the second, we measure the number of differences in repeat units.

The reduced median networks were calculated using the program Network3.0 (www.fluxus-engineering.com), by considering all the haplotypes and setting the reduction threshold to the default value of 2 (Forster et al., 2000). Only the Y-STR information or the joined Y-STR and Y-BM information was used.

Databases and haplotype matching

Besides the present Mozambican sample (66 individuals), we have collected other sub-Saharan population datasets, for Y-STRs, available in the literature: 47 individuals from Cape Verde, 33 from Guinea Bissau, 50 from Angola and another sample of 37 individuals from Mozambique (Côrte-Real et al., 2000); 34 plus 104 from São Tomé and Príncipe (Côrte-Real et al., 2000 and Trovoada et al., 2000, respectively); 34 central African Pigmies (Kayser et al., 2001); and 77 South African Bantus from the Pretoria region (Thomas et al., 2000).

The haplotype matching was restricted to the common loci to these databases, namely DYS19, DYS390, DYS391 DYS392 and DYS393.

Results and Discussion

Y-BM diversity in Mozambique - the small European influence

The combined haplogroup and haplotype diversity observed in Mozambique is displayed in Table 1.

With respect to the Y-BMs, gene diversity was estimated to be 0.434±0.066, much lower than the average observed for Europeans (Pereira et al., 2001b). In fact, the vast majority of the samples belonged to African haplogroups, 73.5% to the typical sub-Saharan H8 and 11.8% to the characteristic North African H21. H8 is absent from the majority of European populations (Rosser et al., 2000), including the Portuguese population (Pereira et al., 2000a), but H21 is present in considerable frequencies in southern Europe, where a north-south increasing gradient was detected. Therefore, a Portuguese introduction of H21 lineages in Mozambique cannot be disregarded, especially having in mind that H21 is present in very low amount (2.6%; Thomas et al., 2000) in South African Bantus, where the European influence was northern (England and the Netherlands), characterised by much lower H21 frequencies than Portugal (Rosser et al., 2000).

A minor frequency of the sample (1.5%) belonged to H1 and the remaining (13.2%) to H2. While for the first it is safe to assume a European origin (and possibly western, where this haplogroup is especially frequent; Rosser et al., 2000), this assignment is more difficult for H2, since the markers analysed in this work do not allow to distinguish between the characteristic Euro-Asian H2 and the African haplogroups that belong to groups I and II, using the nomenclature of Underhill et al. (2000). These are said to contain very diverse and variable lineages, present in Khoisan and Bantu speakers from South Africa, Pygmies from central Africa but also observed in Sudan, Ethiopia and Mali (Underhill et al., 2001).

Figure 1 displays the reduced network for the Y-STR haplotypes observed in Mozambique. There is a high correlation between the haplotype relatedness and haplogroup classification, evidencing, for these lineages, that Y-STRs are highly structured by Y-BM diversity. Further confirmation of this structuring results from the essentially identical pattern displayed by the reduced network obtained with the joined Y-STR/Y-BM information (Figure 2). Its most striking feature is the high level of

reticulation between haplotypes that belong to H8, with many one-step neighbours. The intermediate position displayed by H2 lineages, instead of an extreme position as expected from the most parsimonious trees for Y-BMs (Rosser et al., 2000), seems to favour the hypothesis that they belong to African lineages instead of resulting from European flow.

Figure 3 displays the mismatch distribution for the Mozambican Y-BMs. A bimodal pattern is evident and is associated with non-significant values for Tajima's *D* (-0.499; p=0.356) and Fu's *Fs* (1.121; p=0.750) neutrality tests, similarly to what was observed in Europe (Pereira et al., 2001b) and at a worldwide scale (Dupanloup et al., in preparation). Taken together, there are no signs of exponential growth in the Y-BM Mozambican pool, in contrast to what was shown by the mtDNA counterpart (Pereira et al., 2001a).

Y-STRs - the Bantu diversity reduction

Thomas et al. (2000) have described strong founder effects for Y-STRs in South African Bantus (from the Pretoria area), pointing to a possible founder haplotype, dated to 3,000 years ago, an age compatible with the Bantu expansion.

We tried to assess if these founder effects were detectable in other African populations touched by the Bantu expansion, collecting Y-STR data already reported. The majority of the populations available for comparison were Portuguese former colonies, including Cape Verde, São Tomé and Principe, Guinea Bissau and Angola. A Central African Pigmy sample was also available, representing the ancient African population, but Bantu admixture was possible since they have established camps near the Bantu villages (Newman, 1995). Both Cape Verde and São Tomé and Príncipe archipelagos were inhabited in consequence of the Atlantic slave trade; from the remaining, Guinea Bissau is the only one not located in the sub-Sahel portion of Africa. Thus, in the routes of Bantu expansion, the available samples were: Angola for the west coast; Mozambique for the east; and Pretoria Bantus for the south.

Table 2 displays the haplotypes for which at least a match between 2 populations was observed. The Bantu founder haplotype (15-21-10-11-13), defined by Thomas et al. (2000), is the most abundant in Angola (28.00%) and in São Tomé and Príncipe (10.14%), while in Mozambique it shows the same frequency (10.68%) as its one-step

neighbour, 15-21-10-11-14. These observations are therefore consistent with the hypothesis of Thomas et al. (2000).

Adding up the frequencies for all the one-step neighbour haplotypes around the presumed founder, the new values are: 31.15% for São Tomé and Príncipe; 48.00% for Angola; 28.16% in Mozambique; and 49.37% in Pretoria Bantus. However, the internal composition displayed by this common set of haplotypes was different in the western and the eastern populations. In the west, haplotype 15-21-10-11-13 was predominant, and haplotypes 15-21-11-11-13 and 15-21-10-11-14 were present in decreasing frequencies, while the remaining 16-21-10-11-13 and 15-22-10-11-13 were the less (and equally) frequent. In the east, haplotypes 15-21-10-11-13 and 15-21-10-11-14 were equally predominant, then 16-21-10-11-13, and at last 15-21-11-11-13. The Pretoria Bantus displayed an intermediate frequency distribution for the three most frequent haplotypes (pooling the frequencies from Angola, Mozambique and Pretoria).

In summary, it seems that the Bantu expansion could be the cause for the increasing reduction of diversity towards south (as shown by the increase of the Bantu core haplotype frequencies), which is especially strong in the western coast (when comparing Angola and Mozambique, located at almost the same latitude). In the south, both waves intermingled (Newman, 1995), and so in consonance, Pretoria Bantus display a mixed frequency distribution for the core haplotypes, showing therefore higher diversity than the other two samples (Table 3).

Assuming that those haplotypes are Bantu, and that the Bantu package was different from the Pigmy one, the admixture in Central African Pigmies was of 12.91%.

However populations that were not in the way of the Bantu expansion displayed also some of these haplotypes (24.29% in Guinea Bissau and 14.93% in Cape Verde). The high value observed in Guinea Bissau raises doubts if these haplotypes are indeed good Bantu markers, since, at least linguistically, no instance of Bantu language is nowadays present there. Various explanations are possible namely: (1) the Sahel populations shared some ancestral haplotypes with the Bantu core, but they have underwent also a diversity reduction, affecting especially haplotype 15-21-10-11-14; (2) there were migrations of Bantu lineages to the Sahel, either before or during colonial period (which is surely the case for Cape Verde islands).

When considering the overall haplotype diversity defined by the 5 Y-STRs (Table 4), several measures seem to show the southwards decrease in diversity, stronger in the western coast than in the eastern, but with the south displaying an intermediate level of diversity. There is no correlation between average haplotype pairwise differences, when they are measured in the number of differences (a) per locus or (b) per repeat units. For instance, comparing Angola, Mozambique and Pretoria (Table 4), we see namely that the reduction in diversity is more pronounced in Angola when measured as in (a), but is strongest in Pretoria using the alternative (b).

Pigmies show a low level of gene diversity, what can be explained by population contractions said to have occurred and can be detected at the mtDNA level (Excoffier and Schneider, 1999). Nevertheless, this low level of diversity is not accompanied by the averages of pairwise differences both at locus and (especially) at repeat units.

The archipelagos of Cape Verde and São Tomé and Príncipe display the highest diversities, which can be due to the fact that their colonisation was recent and made from individuals original from several locations of the western African coast (and in particular for Cape Verde, European admixture).

Conclusions

The haploid markers of the Y-chromosome allow the tracing of migrations occurred throughout time, and the study of their male component, which is know to have been predominant, at least in recent history, as for the European "Discovery" period.

The characterisation of 10 Y-BMs in Mozambique, allowed the observation that the European male contribution to present population is minimal, since the main western European H1 is present only in 1.5%. Additional European influence could have been contributed through a portion of the 11.8% of H21 and 13.2% of H2 detected in Mozambique, although a North African and a Khoisan/Pigmy influence cannot be disregarded, respectively. This makes the estimate of the proportion of European lineages in Mozambique male pool to vary from 1.5% to a very theoretical upper limit of 26.5%. No matter how low, even the lower limit proves that Mozambican male history is quite different from the female one, since for mtDNA, no European sequences were observed (Pereira et al., 2001a). A similar bias in the mating pattern between Europeans and sub-Saharans was also observed in Portugal (Pereira et al., 2000a,b); and, while in Europe it is responsible for the absence of sub-Saharan lineages in the male pool, in the former colony lead to the absence of European lineages in the female pool. The same bias was also observed in American ex-colonies, such as Colombia (Carvajal-Carmona et al., 2000; Mesa et al., 2000), USA (Parra et al., 1998) and Brazil (Alves-Silva et al., 2000; Carvalho-Silva et al., 2000).

Nevertheless, the main demographic episodes modelling sub-Sahel diversity were those resulting from the Bantu expansion waves towards south. The enlarged comparison of Y-STR haplotypes in that region showed a reduction of diversity towards south, which seems stronger along the western African coast. The western and eastern waves of Bantu migration seem to have shared a common founder set, but differing in haplotype frequencies, just slightly more diverse in the east. The south region of Pretoria seems the result of a mix between both western and eastern waves of Bantu advance.

The proposed Bantu haplotypes are also detected in Central African Pigmies (12.91%), which can indicate gene flow between both, but it seems that the Bantu expansion is not sufficient to account for the diversity reduction observed in this

population, as the haplotype frequency distribution suggests. Population contractions are therefore a more plausible cause for the Pigmy low diversity. The presence of those Bantu haplotypes in populations not located in the way of Bantu advance, and associated with high levels of diversity, can be explained by the existence of an ancient common Y-pool for sub-Saharan populations or by Bantu inputs not connected with the classical Bantu expansion.

These results give evidence to another important fact: that an increase in the number of individuals does not necessarily conduct to an increase of diversity (especially for the fast-evolving STRs) and to signs of expansion in neutral tests (for the Y-BMs). The absence of signs for expansion in the Y-BMs in opposition to its presence in mtDNA has been discussed at the European scale (Pereira et al., 2001b), but demographic phenomena are known to have been different in this African region. Indeed, the Bantu migration corresponded in fact to a demographic expansion, which was stronger in the east coast (at least in part attributable to more productive food and agriculture technologies from Asia; Fage, 1997), coinciding with its slightly higher Y-chromosome diversity.

ACKNOWLEDGEMENTS

LP was supported by a PhD grant (PRAXIS XXI BD/13632/97) from Fundação para a Ciência e a Tecnologia, and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III. Dr. Albertino Damasceno and Dr. Benilde Soares of the Eduardo Mondlane University (Maputo) kindly provided the Mozambican samples.

References:

Alves-Silva J, Santos MDS, Guimarães PEM, Ferreira ACS, Bandelt H-J, Pena SDJ, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67:444-461

Bandelt H-J, Alves-Silva J, Guimarães PEM, Santos MS, Brehm A, Pereira L, Coppa A, Larruga JM, Rengo C, Scozzari R, Torroni A, Prata MJ, Amorim A, Prado VF, Pena SDJ (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Ann Hum Genet 65 (in press)

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

Carvalho-Silva D, Santos FR, Rocha J, Pena SDJ (2000) The phylogeography of Brazilian Y-chromosome lineages. Am J Hum Genet 68:281-286

Côrte-Real F, Carvalho M, Andrade L, Anjos MJ, Pestoni C, Lareu MV, Carracedo A, Vieira DN, Vide MC (2000) Chromosome Y STRs analysis and evolutionary aspects for Portuguese spoken countries. In: Sensabaugh GF, Lincoln PJ, Olaisen B (eds) Progress in Forensic Genetics 8:272-274. Elsevier Science, Amsterdam

Excoffier L, Schneider S (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc Natl Acad Sci USA 96:10597-10602

Fage JD (1997) História da África. Edições 70. Lisboa.

Forster P, Röhl A, Lünnermann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B (2000) A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet 67:182-196

González-Neira A, Gusmão L, Barral S, Lareu MV, Carracedo A (2000) Multiplexing Y chromosome STRs: analysis of artifactual bands and PCR strategies. In: Sensabaugh GF, Lincoln PJ, Olaisen B (eds) Progress in Forensic Genetics 8:436-438. Elsevier Science, Amsterdam

Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, Roewer L (1997) Applications of microsatellite-based Y chromosome haplotyping. Electrophoresis 18:1602-1607

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68:990-1018

Lareu MV, Phillips CP, Carracedo A, Lincoln AJ, Syndercombe-court D, Thomson JA (1994) Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats; UK and Galician Caucasian population surveys and usefulness in paternity investigations. Forensic Sci Int. 66:41-52

Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortíz-Barrientos D, García LF, Velez ID, Bravo ML, Múnera JG, Bedoya G, Bortolini M-C, Ruiz-Linares A (2000) Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-columbian patterns of gene flow in South America. Am J Hum Genet 67:1277-1286.

Newman JL (1995) The peopling of Africa: a geographic interpretation. Yale University Press

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

Pereira L, Prata MJ, Brión M, Jobling MA, Carracedo A, Amorim A (2000a) Clinal variation of the YAP^+ Y chromosome frequencies in Western Iberia. Hum Biol 72:937-944

Pereira L, Prata MJ, Amorim A (2000b) Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation. Ann Hum Genet 64:491-506

Pereira L, Dupanloup I, Rosser ZH, Jobling MA, Barbujani G (2001a) Y-chromosome mismatch distributions in Europe. Mol Biol Evol 18:1259-1271

Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A (2001b) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. Ann Hum Genet 65 (in press)

Pritchard JK, Seielstad MT, Pérez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791-1798

Rosser ZH, Zerjal T, Hurles ME, Adojaan MA, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper GC, Côrte-Real HBSM, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Gölge M, Hill EW, Jeziorowska A, Kalaydjieva L,

Kayser M, Kravchenco SA, Lavinha J, Livshits LA, Maria S, McElreavey K, Meitinger TA, Melegh B, Mitchell RJ, Nicholson J, Nørby S, Novelletto A, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previderé C, Rajczy K, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Shlumukova M, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. Am J Hum Genet 67:1526-1543

Russell-Wood AJR (1998) The Portuguese empire, 1415-1808. A world on the move. The Johns Hopkins University Press. Baltimore

Schneider S, Roessli D, Excoffier L (2000) Arlequin ver.2.0: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

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

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

Trovoada MJ, Alves C, Gusmão L, Abade A, Amorim A, Prata MJ (2001) Evidence for population sub-structuring in São Tomé e Príncipe as inferred from Y-chromosome STR analysis. Ann Hum Genetics 65:271-283

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

Underhill PA, Passarino G, Lin AA, Shen P, Lahr MM, Foley RA, Oefner PJ, Cavalli-Sforza LL (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet 65:43-62

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

Table 1: Haplogroup and haplotypes observed in the Mozambican sample.

Sample	N	Haplogroup	Haplotype	Sample	N	Haplogroup	Haplotype
M1	1	H1	ND	M23	2	H8	16 13 17 21 10 11 14
M2	1	H2	15 13 16 23 9 11 12	M24	4	H8	16 13 17 21 10 11 15
M3	1	H2	15 13 17 22 10 11 12	M25	2	H8	16 13 17 21 10 11 16
M4	1	H2	15 13 18 26 10 11 13	M26	1	H8	16 13 17 21 11 11 15
M5	4	H2	15 14 18 24 10 11 13	M27	1	H8	16 13 17 22 10 11 15
M6	1	H2	15 14 19 24 10 11 13	M28	1	H8	16 13 17 22 10 11 16
M7	1	H2	16 13 18 24 11 11 12	M29	2	H8	16 13 18 21 10 11 13
M8	1	H8	14 13 17 21 11 11 14	M30	1	H8	16 13 19 21 8 11 14
M9	1	H8	15 12 17 20 10 11 13	M31	1	H8	16 13 19 21 10 11 13
M10	1	H8	15 12 17 21 10 11 13	M32	1	H8	16 14 16 21 10 11 14
M11	2	H8	15 12 17 21 10 11 14	M33	1	H8	16 14 17 21 10 11 14
M12	1	H8	15 13 17 20 10 11 14	M34	1	H8	17 13 17 21 10 11 13
M13	3	H8	15 13 17 21 10 11 13	M35	5	H8	17 13 17 21 10 11 14
M14	3	H8	15 13 17 21 10 11 14	M36	2	H8	17 13 18 21 10 11 13
M15	1	H8	15 13 17 21 11 11 14	M37	1	H8	17 13 18 21 10 11 14
M16	1	H8	15 13 17 21 11 11 15	M38	1	H8	ND
M17	2	H8	15 13 18 21 10 11 13	M39	1	H21	13 10 17 23 11 11 14
M18	1	H8	15 13 18 21 10 11 14	M40	1	H21	13 13 17 24 10 11 14
M19	1	H8	15 13 18 21 10 11 15	M41	1	H21	14 12 16 24 10 11 13
M20	2	H8	15 14 17 21 10 11 14	M42	2	H21	14 12 16 24 11 11 13
M21	1	H8	16 13 16 21 10 11 13	M43	2	H21	14 12 16 25 10 11 13
M22	2	H8	16 13 17 21 10 11 13	M44	1	H21	14 12 16 25 11 11 13

Note: ND - not determined.

Table 2: Y-STR haplotype (DYS19, DYS390, DYS391, DYS392 and DYS393) matches between sub-Saharan populations. In dark grey the hypothetical Bantu founder haplotype and in light grey its one-step neighbour haplotypes.

Haplotype	Cape Verde (47[1])	Guinea-Bissau (33[1])	São Tomé and Príncipe (34[1]+104[2])	Angola (50[1])	Mozambique (37[1]+66[3])	Central African Pigmies (34[4])	South African Bantu (77[5])
13 24 9 11 13	1 (2.13)		1 (0.72)				
13 24 10 11 13	3 (6.38)		3 (2.17)				
13 24 11 11 13	1 (2.13)		1 (0.72)				
14 21 10 11 14				1 (2.00)			1 (1.30)
14 23 9 11 13	1 (2.13)		1 (0.72)				
14 23 10 11 12	1 (2.13)		3 (2.17)				
14 24 10 11 13			2 (1.45)		1 (0.97)		
14 24 11 13 13	6 (12.80)		5 (3.62)				
14 25 10 11 13				2 (4.00)	2 (1.94)		1 (1.30)
14 25 11 11 13				1 (2.00)	1 (0.97)	2 (6.45)	
15 21 10 11 13		2 (6.06)	14 (10.14)	14 (28.00)	11 (10.68)	1 (3.23)	17 (22.10)
15 21 10 11 14	6 (12.80)	5 (15.20)	8 (5.80)	4 (8.00)	11 (10.68)		7 (9.09)
15 21 10 11 15		1 (3.03)	5 (3.62)		1 (0.97)		2 (2.60)
15 21 11 11 13			16 (11.59)	5 (10.00)	1 (0.97)	2 (6.45)	7 (9.09)
15 21 11 11 14		1 (3.03)		1 (2.00)	2 (1.94)		2 (2.60)
15 21 12 11 13			1 (0.72)	1 (2.00)			
15 22 10 11 13	1 (2.13)		2 (1.45)			1 (3.23)	1 (1.30)
15 22 10 11 14	1 (2.13)	1 (3.03)	2 (1.45)	2 (4.00)			
15 22 11 11 13	3 (6.38)	2 (6.06)					1 (1.30)
15 23 10 11 13	1 (2.13)		1 (0.72)				
15 24 10 11 13				1 (2.00)	6 (5.83)		5 (6.49)
15 24 10 13 13			1 (0.72)		1 (0.97)		
15 24 11 11 13					1 (0.97)		2 (2.60)
15 24 11 13 13	1 (2.13)		1 (0.72)				
15 25 10 11 13			1 (0.72)			3 (9.68)	
15 25 11 11 13	1 (2.13)				1 (0.97)		
16 21 10 11 13		1 (3.03)	3 (2.17)	1 (2.00)	6 (5.83)		6 (7.79)
16 21 10 11 14	2 (4.26)	3 (9.09)	2 (1.45)	2 (4.00)	5 (4.85)		4 (5.19)
16 21 10 11 15	1 (2.13)	1 (3.03)	3 (2.17)	4 (8.00)	4 (3.88)	5 (16.10)	3 (3.90)
16 21 11 11 13			2 (1.45)	2 (4.00)			1 (1.30)
16 21 11 11 14			2 (1.45)	1 (2.00)			
16 22 10 11 13	1 (2.13)	2 (6.06)	1 (0.72)				
16 21 10 12 15		1 (3.03)					2 (2.60)
16 22 10 11 16					1 (0.97)		1 (1.30)
16 22 11 11 13	1 (2.13)	1 (3.03)	1 (0.72)				
16 23 10 11 13	1 (2.13)		1 (0.72)			1 (3.23)	
16 24 10 11 13			1 (0.72)	1 (2.00)			2 (2.60)
16 25 10 11 13				1 (2.00)	1 (0.97)		
17 21 10 11 13	2 (4.26)			1 (2.00)	3 (2.91)		2 (2.60)
17 21 10 11 14	1 (2.13)	1 (3.03)	4 (2.90)		11 (10.68)	2 (6.45)	2 (2.60)
17 21 10 11 15			2 (1.45)	1 (2.00)			1 (1.30)

References: [1]Côrte-Real et al. (2000); [2]Trovoada et al. (2001); [3]this work; [4]Kayser et al. (2001); [5]Thomas et al. (2000).

Table 3: Diversity measures for the Bantu core Y-STR haplotypes (DYS19, DYS390, DYS391, DYS392 and DYS393) in the Sub-Saharan populations.

	N	No. of different haplotypes	Gene diversity	Average pairwise differences
Cape Verde	7	2	0.286 ± 0.196	0.571 ± 0.521
Guinea Bissau	8	3	0.607 ± 0.164	0.786 ± 0.633
Central African Pigmies	4	3	0.833 ± 0.222	1.167 ± 0.928
São Tomé and Príncipe	43	5	0.731 ± 0.035	1.012 ± 0.693
Angola	24	4	0.612 ± 0.087	0.717 ± 0.556
Mozambique	29	4	0.692 ± 0.004	0.897 ± 0.643
South African Bantu	38	5	0.726 ± 0.048	0.943 ± 0.661

Table 4: Diversity measures for all the Y-STR haplotypes (DYS19, DYS390, DYS391, DYS392 and DYS393) in the Sub-Saharan populations.

	N	No. of different haplotypes (%)	Gene diversity	Average pairwise differences in loci	Average pairwise differences in repeat units
Cape Verde	47	30 (63.83)	0.964 ± 0.015	3.122 ± 1.649	5.064 ± 2.502
Guinea Bissau	33	19 (57.58)	0.941 ± 0.024	2.402 ± 1.339	3.114 ± 1.658
Central African Pigmies	31	15 (48.39)	0.929 ± 0.026	2.900 ± 1.564	4.710 ± 2.369
São Tomé and Príncipe	138	70 (50.72)	0.969 ± 0.007	2.858 ± 1.514	4.761 ± 2.342
Angola	50	22 (44.00)	0.904 ± 0.031	1.980 ± 1.138	3.180 ± 1.672
Mozambique	103	43 (41.75)	0.956 ± 0.008	2.367 ± 1.300	4.044 ± 2.035
South African Bantu	77	28 (36.36)	0.926 ± 0.018	2.021 ± 1.150	2.974 ± 1.572

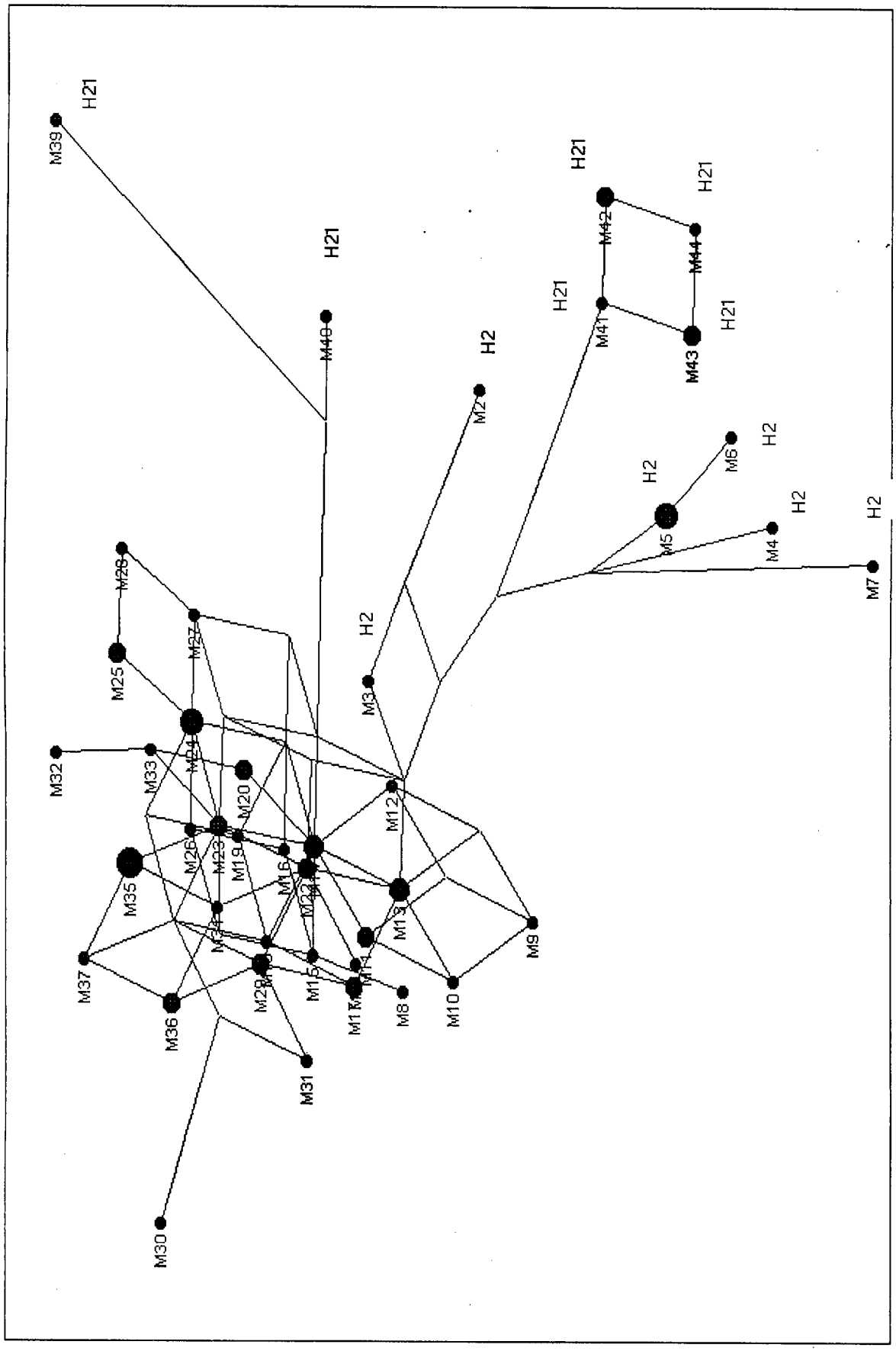


Figure 1- Reduced median network for the Mozambican Y-STR haplotypes. Lineages that belong to other haplogroups than H8 are indicated, but Y-BM information was not used for the network construction.

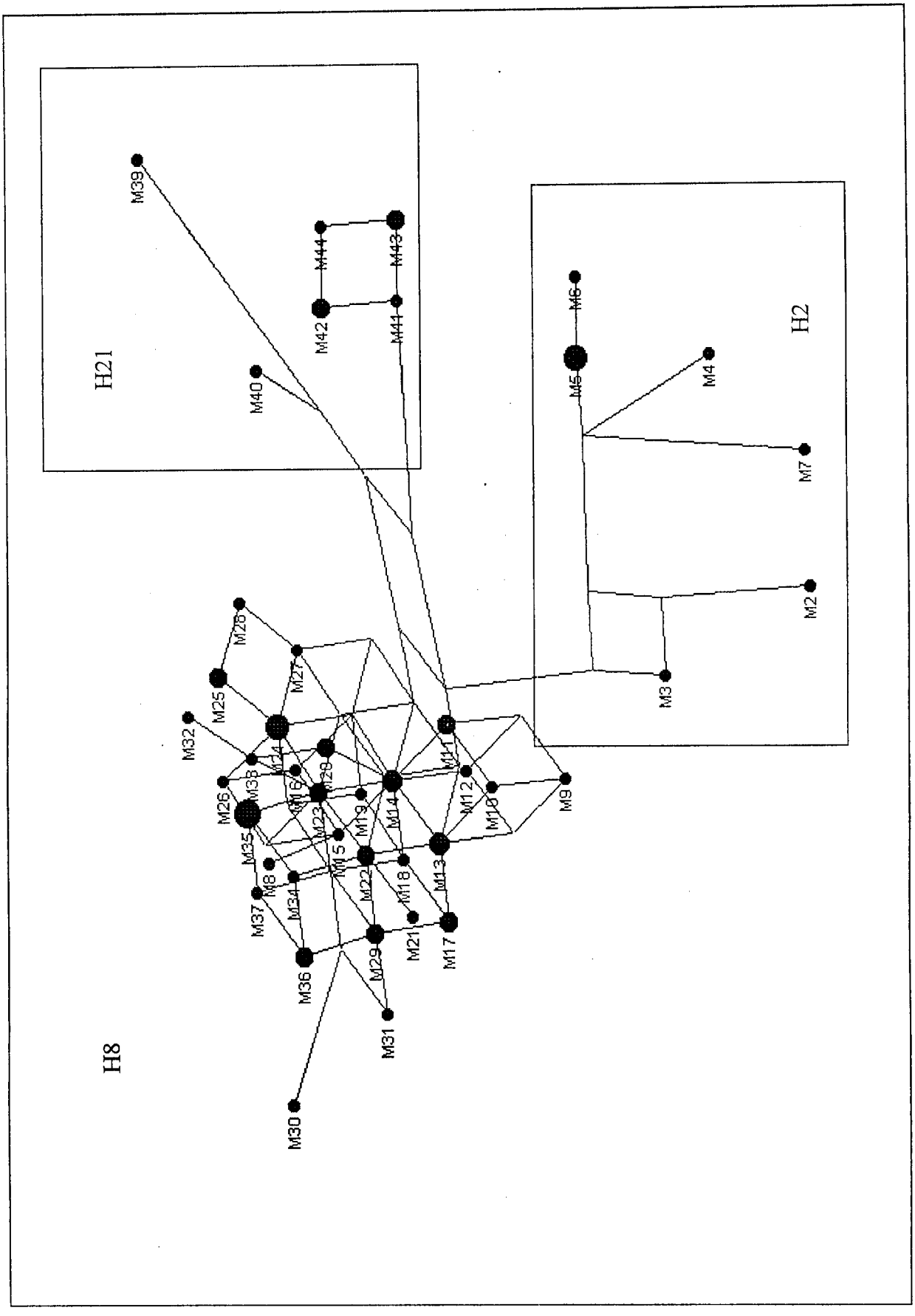


Figure 2- Reduced median network for the Mozambican Y-STR haplotypes, adding the Y-BM information for the network construction.

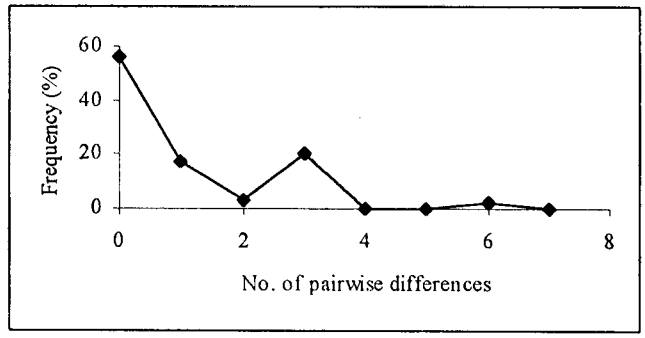


Figure 3: Mismatch distribution for the Mozambican Y-BMs.

Conclusions

I - PROPERTIES OF Y-CHROMOSOME AND MTDNA MARKERS

The Y-BM mismatch distributions

Y-BM mismatch distributions are, in general, multimodal at a worldwide scale and statistical tests failed to reject a neutral equilibrium model. This pattern does not resemble those detected for mtDNA data, which were unimodal, showing highly significant departure from neutrality (Excoffier and Schneider, 1999). As described for other demographic parameters such as migration rates (Belledi et al., 2000), Y-BMs display a very different pattern when compared with mtDNA and Y-STRs.

Considering that the underlying causes responsible for these differences could lie outside the purely demographic ones, some possible sources of bias were tested.

One of these could be the different number of polymorphic sites used within each type of markers. This possible bias was tested in two ways: (1) by decreasing the number of mtDNA polymorphic sites analysed; and (2) by comparing the Y-chromosome data obtained by Rosser et al. (2000), Semino et al. (2000) and Underhill et al. (2000) for 11, 22 and 166 Y-BMs, respectively. Results showed that (1) the characteristic unimodal pattern and significant negative departure from neutrality observed for mtDNA were maintained even after reduction of polymorphic sites; and (2) bi- and tri-modal distributions were observed, no matter what number (11, 22 or 166) of Y-BMs were analysed.

With respect to the effect of the demographic test applied, simulations under stationary and expansion scenarios for Y-BMs were assayed and showed that bimodal distributions appeared much more frequently, roughly 3 to 10 times, in stationary than in expanding populations. Nevertheless, these numerical results obviously depend upon admittedly approximate values for the expansion parameters and mutation rates.

So, the effects of mutation rate and/or selection of sites were also assayed. Results showed that in a stationary population, no differences were found when considering just 11 sites or all of them. However, in an expanding population (and for relatively high mutation rates), by selecting a few sites we may miss the signal of

expansion. In any case, the mutation rates for Y-BMs are such that the number of significant tests is not affected by selection of sites. Therefore, mutation rate does not seem to be the explanation for the differences displayed by Y-BMs comparatively to Y-STRs and mtDNA.

The contradiction with the reported expansion signals for four Y-genes by Shen et al. (2000) can be resolved as resulting from a bias on the selection of a heterogeneous worldwide population.

When a similar exercise was performed with Semino et al. (2000) and Underhill et al. (2000) Y-BMs datasets, grouping all the populations, we showed that for the first one, a trimodal distribution and non-significant positive Tajima's D and Fu's Fs values were obtained, while for the second one (a much more heterogeneous group) the distribution was still bimodal, but with Tajima's D and Fu's Fs statistics already significantly negative.

So, the apparently opposite Y-chromosome and mtDNA mismatch scenarios do not seem to be caused by an artefact, and do reflect, at least in part, past demographic phenomena. Various explanations remain possible, which are not mutually exclusive:

1. The European female population increased in size, the male population did not.
2. Neither population increased in size, but there was diversifying selection for mtDNA.
3. Both populations increased, and there was purifying selection on the Y chromosome.

The data under analysis do not provide evidence favouring any of these hypotheses, but it is worth to mention that a small population size does not necessarily mean a small number of individuals (of males, in this case) since a high variance of reproductive success among individuals reduces the effective population size (Crow, 1958). If the number of offspring is potentially more variable among males than among females, the effective population size inferred from Y-chromosome diversity is expected to be smaller (as suggested in this study), and the corresponding estimates of genetic differences between populations to be greater (as demonstrated, e.g., by Seielstad et al., 1998).

The mismatch distribution as an evaluation method for the proportion of Y-STR haplotypes that can become identical-by-state

For Y-STRs, mismatch distribution has shown to be a useful tool for the evaluation of the molecular distance (in loci or in repeat units) between all pairs of haplotypes in a certain population. This distribution is therefore important in the forensic field, in the sense that allows the evaluation of the proportion of pair of haplotypes that are prone to become identical-by-state, by single-step mutation, in the next generation. It presents some advantages when compared to the approach based in the concomitant information from Y-BM/Y-STR, the only one that has been so far applied to the evaluation of recurrence (de Knijff, 2000). Indeed, (1) it allows the comparison of results in a larger geographical range, since most available samples are typed for only Y-STRs; (2) being one-generational, avoids the complex phenomenon of re-recurrence, being more indicated for estimations intended to be used in the forensic field (where we are interested in establishing links within one generation or between one or two generations apart); (3) it does not imply the very restrictive assumption that recurrence only acts in haplotypes between-haplogroups, evaluating also recurrence within-haplogroups.

The comparison between both strategies for the population of The Netherlands produced values of the same order of magnitude. Having in mind the referred differences between the mismatch and the Y-BM/Y-STR approaches, one can conclude that the within-haplogroup recurrence must be considerable higher than between-haplogroups. This is not surprising since haplotypes within-haplogroups are more similar than between-haplogroups (the degree of similarity being inversely related to the age of the haplogroup), and hence more prone to recurrence than the second ones.

The geographic distribution, at an European scale, of these IBS probabilities showed that, contrarily to the described absence of population stratification with respect to Y-STR diversity, it followed clines that matched those observed for Y-BMs. South-central European populations displayed the lowest IBS probabilities, which increased towards west and north.

In conclusion, matching evaluation for forensic purposes, at least at an European scale, needs a lot more of refinement and, at the moment, a lot more of prudence.

II – POPULATION GENETICS' APPLICATIONS

Portugal – the country context

In general, the Portuguese genetic landscape is what would be expected for a western European population, that is, briefly, the characteristic European haplogroups are present and the most frequent ones (Y-BM haplogroup 1 and mtDNA haplogroup H) reach here high proportions. Additionally, the lower frequencies of the Near Eastern haplogroups is consistent with a weaker Neolithic influence: in the Y-chromosome, the frequency of haplogroup 9 was estimated as 5.2%; and, in the mtDNA, considering as Neolithic lineages the ones belonging to haplogroups J, U3 and T1 (Richards et al., 2000), their combined frequencies were 11.0%.

At the mtDNA level, Portuguese genetic diversity was found to be higher than expected when considering the neighbouring populations. The difference is mainly due to the presence of African haplogroups in higher frequencies as compared with the rest of Europe.

The African input shows, however, a distinct pattern for the Portuguese Y-chromosome and mtDNA pools, as it seems to be the rule for almost all populations.

At least for sub-Saharan lineages, this is probably explainable by a sex bias in the mating between Portuguese and Africans. In this way, no specific sub-Saharan lineages were observed in the male pool, in opposition to its presence in the mtDNA pool. Phylogenetic analysis of the sub-Saharan mtDNA sequences observed in Portugal showed that all were very differentiated, which makes more probable a scenario of a recent introduction, but not necessarily as a single event, of a multi-origin package of sequences. The sub-Saharan slaves' intake, between the mid of fifteenth and eighteenth centuries, fits this hypothesis. Although no significant sex bias in the slaves taken to Portugal is registered, a very negative social pressure against Portuguese female/African male mating and resulting offspring, explains this asymmetric African contribution to the present Portuguese Y-chromosome and mtDNA pools.

With respect to the North African influence in Portugal, differences between the female and the male pools were also registered. An increasing gradient for the Y-North African characteristic haplogroup 21, from North to South Portugal, was found, while in mtDNA, the typical Berber haplogroup U6 was only detected in North Portugal (7%, the highest value registered outside North Africa and Canarias, and moreover the sequences were phylogenetically very differentiated). So, the observed Y-chromosome gradient is in accordance with the stronger Islamic influence and closeness with North Africa. MtDNA data are, however, in opposition to these facts. Again a bias in the mating pattern could be invoked: Islamic males inter-bred with Portuguese, in higher numbers and for longer in the south, while the mating between Islamic females and Portuguese was more restricted and limited, for unknown reasons (perhaps a slave-like introgression of Berber females in the North). Other explanations cannot be disregard, such as differential inputs (and/or at distinct times) for both genetic pools.

Concerning population substructuring, it was observed for male lineages (South Portugal being significant different from North), but not for mtDNA (the restriction of U6 sequences to North Portugal was quantitatively insufficient).

Portugal – the Iberian context

The north-to-south increasing gradient for haplogroup 21, in the Y-chromosome, was maintained when studying the entire western Iberian coast, structured in four regions: Galicia (the northwestern most Spanish province), North Portugal, Central Portugal and South Portugal. This observation reinforces the interpretation of the presence of a cline for this haplogroup, since it is more significant its maintenance between four than only three geographic points. Frequencies observed for this haplogroup were of 9.6%, 10.6%, 16.1% and 24.5%, from north to south. South Portugal was found to be statistically different from the two northernmost regions.

This study represents one of the most detailed micro-regional surveys of a cline performed so far, and showed that Y-BMs give information enough to resolve population structure even at such a small geographic scale.

With respect to the Iberian distribution of mtDNA U6 lineages, the available results are in accordance with the described Portuguese picture, being detected, so far, in the north: 7% in North Portugal, 2.2% in Galicia and 1.8% in northeastern Spain. Samples from Central and South Portugal and Spain have not displayed yet any U6 sequences.

Phylogenetic analysis of the Iberian U6 sequences has not allowed the establishment of a putative founder. Therefore, their presence can be explained either by (1) a single event from a source population with a very diverse set of U6 lineages, or (2) several introductions over a long period from various sources.

Portugal – the European context

The entire European screening of 11 Y-BMs, defining 10 observed haplogroups, showed that diversity is geographically structured as clines for five of them.

At a continental range, haplogroups 1 and 9, amounting jointly to 45%, displayed opposite west-east gradients, which was interpreted as representing, respectively, the previous (Palaeolithic) and the newcomer (Neolithic) lineages.

A north-south increasing gradient, encompassing the southern half of the continent, was observed for haplogroup 21. This could represent a North-African influence, since its frequency was highest in the 2 North African samples analysed. This cline shows similarities with the second principal component of classical gene frequencies (Cavalli-Sforza et al., 1994), which was then interpreted on a climatic basis, linking it to the post Last Glacial Maximum expansion.

Two further clines, restricted to smaller areas were observed for haplogroups 3 and 16, reflecting probably population movements in the eastern Europe.

The application of the Mantel test to assess the relative importance of geography and language, in the shaping of genetic differentiation at the Y-BMs, showed a high correlation with geography, but not with language. Language seemed to act at a more restricted geographical level, as in the case of the two strongest eastern European barriers between the (1) Uralic-speaking Mari and Altaic-speaking, and (2) Georgians and Ossetians, who speak languages belonging to different families.

Portugal – as contributor for Mozambique

In order to study the Portuguese genetic impact in Mozambique, the assessment of the effects of other migrations, namely the Bantu expansion and slave trade, had to be made. In consequence, the corresponding results are also presented in this section.

At the cultural level, as judged by linguistics, the Bantu impact was overwhelming in Mozambique, since all dialects spoken there belong to the Bantu linguistic supra family.

This impact showed a different pattern in its male and female components, a much more pronounced reduction of genetic diversity being observed for the first. In fact, 73.5% of the Y-chromosomes belonged to haplogroup 8, and even for the fast-evolving Y-STRs, it was possible to define candidate founder haplotypes for the Bantu expansion. The high diversity for mtDNA was testified by the finding of all typical sub-Saharan lineages. The relative reduction in diversity for haplogroup L2a (43.1% of the Mozambican sample) could not be directly linked to the Bantu expansion, since an earlier (late-glacial) expansion could not be excluded.

As it was possible for the Y-chromosome to compare both western and eastern Bantu waves of expansion, we could infer that the reduction of male genetic diversity was stronger in the west than in the east. Moreover, the analysis of the Y-STR haplotype distributions in Angola, Mozambique and Pretoria, allowed us to postulate that a common Bantu package was present in both waves of expansion, but their relative frequencies were different, being intermingled in South Africa.

In the Mozambican mtDNA pool, a frequency of 7% of L1d was observed, which can be a relict of the ancient Khoisan people that inhabited the region before Bantu arrival (although recent gene flow cannot be excluded). In the Y-chromosome, it was observed that a central African population of Pigmies displayed 12.5% of the haplotypes implicated in the Bantu expansion, which can indicate gene flow between both or the sharing of some ancient haplotypes.

Concerning the European contribution, from the fifteenth century onwards, it is known that the migrants from that continent were in very small numbers and mainly males.

This seemed to be reflected in the Mozambique present gene pool. For the mtDNA, no European lineages were detected (neither North African or Asian), while for the Y-chromosome, H1, which can be assigned to a European (probably western) ancestry, reached a frequency of 1.5%. For some Y-lineages it was not possible to assign their origin: H21 (11.8%) that although typically North African is very frequent in Mediterranean Europe; and H2 (13.2%), which, defined by default, could be either European or ancient African. So, at the moment, the European genetic influence in the male Mozambican pool can be estimated to lie between 1.5-26.5%, although the true value is expectedly closer to the first one.

These facts are in accordance with the mating bias already detected in Portugal (favouring the crossing between European males and African females), as well as in most European ex-colonies (as discussed in the Introduction).

Finally, we address the question of Mozambican contribution to the slave trade to Europe and Americas.

When comparing Mozambican, European and American mtDNA pools, a higher proportion of matches were found between Mozambique and Americas than between Mozambique and Europe, in agreement with the recorded history of the slave trade to these continents. In this context, L3e1* haplogroup seemed to be a good marker to follow the dispersion of southeastern African sequences throughout the world.

In agreement with the mating bias referred above, the same kind of comparative analysis for the male counterpart is at the moment impossible to perform, since European gene pools do not show H8 lineages, and they are present at low frequencies in American Y-BM databases.

References

ALVES-SILVA, J, SANTOS, M.D.S., GUIMARÃES, P.E.M., FERREIRA, A.C.S., BANDELT, H.-J., PENA, S.D.J., PRADO, V.F. (2000) The ancestry of Brazilian mtDNA lineages. *Am. J. Hum. Genet.* 67:444-461.

AMMERMAN, A.J., CAVALLI-SFORZA, L.L. (1984) *The Neolithic transition and the genetics of population in Europe*. Princeton (NJ): Princeton University Press.

AMORIM, A., GUSMÃO, L., PRATA, M.J. (1996) Population and formal genetics of the STRs TPO, TH01 and VWFA31/A in North Portugal. *Adv. Forensic Haemogenet.* 6:486-488.

AMORIM, A. (1999) Archaeogenetics. *Journal Iberian Archaeology.* 1:15-25.

ANDERSON, S., BANKIER, A.T., BARRELL, B.G., DE BRUIJN, M.H., COULSON, A.R., DROUIN, J., EPERON, I.C., NIERLICH, D.P., ROE, B.A., SANGER, F., SCHREIER, P.H., SMITH, A.J., STADEN, R., YOUNG, I.G. (1981) Sequence and organisation of the human mitochondrial genome. *Nature* 290:457-465.

ANDREWS, R.M., KUBACKA, I., CHINNERY, P.F., LIGHTOWLERS, R.N., TURNBULL, D.M., HOWELL, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* 23:147.

ARNAIZ-VILLENA, A., MARTÍNEZ-LASO, J., GÓMEZ-CASADO, E., DÍAZ-CAMPOS, N., SANTOS, P., MARTINHO, A., BREDA-COIMBRA, H. (1997) Relatedness among Basques, Portuguese, Spaniards and Algerians studied by HLA allelic frequencies and haplotypes. *Immunogenetics.* 47:37-43.

ARNAIZ-VILLENA, A., MARTÍNEZ-LASO, J., ALONSO-GARCÍA, J. (1999) Iberia: population genetics, anthropology, and linguistics. *Hum. Biol.* 71:725-743.

ARIS-BROSOU, S., EXCOFFIER, L. (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.* 13:494-504.

AWADALLA, P., EYRE-WALKER, A., SMITH, J.M. (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science.* 286:2524-2525.

AYUB, Q., MOHYUDDIN, A., QAMAR, R., MAZHAR, K., ZERJAL, T., MEHDI, S.Q., TYLER-SMITH, C. (2000) Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res.* 2, e8.

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

BANDELT, H.-J., ALVES-SILVA, J., GUIMARÃES, P.E.M., SANTOS, M.S., BREHM, A., PEREIRA, L., et al. (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann. Hum. Genet.* 65 (in press).

BARBUJANI, G., PILASTRO, A., DE DOMENICO, S., RENFREW, C. (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. Paleolithic colonisation. *Am. J. Phys. Anthropol.* 95:137-154.

BELLEDI, M., SIMONI, L., CASALOTTI, R., DESTRO-BISOL, G. (2000) Male and female differential patterns of genetic variation in human populations. In RENFREW, C., BOYLE, K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe.* Cambridge. McDonald Institute for Archaeological Research, pp. 295-300.

BERTORELLE, G., SLATKIN, M. (1995) The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters. *Mol. Biol. Evol.* 12:887-892.

BOCQUET-APPEL, J.-P., DEMARS, P.Y. (2000) Neanderthal contraction and modern human colonization of Europe. *Antiquity.* 74:544-552.

BOSCH, E., CALAFELL, F., SANTOS, F.R., PÉREZ-LEZAUN, A., COMAS, D., BENCHEMSI, N., TYLER-SMITH, C., BERTRANPETIT, J. (1999) Variation in short tandem repeats is

deeply structured by genetic background on the human Y chromosome. *Am. J. Hum. Genet.* 65:1623-1638.

BOSCH, E., CALAFELL, F., COMAS, D., OEFNER, P.J., UNDERHILL, P.A., BERTRANPETIT, J. (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between Northwestern Africa and the Iberian Peninsula. *Am. J. Hum. Genet.* 68:1019-1029.

BOWCOCK, A.M., RUIZ-LINARES, A., TOMFOHRDE, J., MINCH, E., KIDD, J.R., CAVALLI-SFORZA, L.L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature.* 368:455-457.

BROWN, W.M., GEORGE, M. JR., WILSON, A.C. (1979) Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 76:1967-1971.

BUDOWLE, B., CHAKRABORTY, R., GIUSTI, A.M., EISENBERG, A.J., ALLEN, R.C. (1991) Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* 48:137-144.

CANN, R., STONEKING, M., WILSON, A. (1987) Mitochondrial DNA and human evolution. *Nature.* 325:31-36.

CARVAJAL-CARMONA, L.G., SOTO, I.D., PINEDA, N., ORTÍZ-BARRIENTOS, D., DUQUE, C., OSPINA-DUQUE, J., MCCARTHY, M., MONTOYA, P., ALVAREZ, V.M., BEDOYA, G., RUIZ-LINARES, A. (2000) Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am. J. Hum. Genet.* 67:1287-1295.

CARVALHO-SILVA, D., SANTOS, F.R., ROCHA, J., PENA, S.D.J. (2000) The phylogeography of Brazilian Y-chromosome lineages. *Am. J. Hum. Genet.* 68:281-286.

CASALOTTI, R., SIMONI, L., BELLEDI, M., BARBUJANI, G. (1999) Y-chromosome polymorphisms and the origins of the European gene pool. *Proc. R. Soc. Lond. B Biol. Sci.* 266:1959–1965.

CAVALLI-SFORZA, L.L., MENOZZI, P., PIAZZA, A. (1993) Demic expansions and human evolution. *Science.* 259:639-646.

CAVALLI-SFORZA, L.L., MENOZZI, P., PIAZZA, A. (1994) *The history and geography of human genes.* Princeton University Press. Princeton. New Jersey.

CHAKRABORTY, R., KIMMEL, M., STIVERS, D.N., DAVISON, L.J., DEKA, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. USA.* 94:1041-1046.

CHEN, Y.-S., TORRONI, A., EXCOFFIER, L., SANTACHIARA-BENERECETTI, A. S., WALLACE, D.C. (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am. J. Hum. Genet.* 57:133-149.

CHEN, Y.-S., OLCKERS, A., SCHURR, T. G., KOGELNIK, A. M., HUOPONEN, K., WALLACE, D. C. (2000) mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am. J. Hum. Genet.* 66:1362-1383.

CHIKHI, L., DESTRO-BISOL, G., BERTORELLE, G., PASCALI, V., BARBUJANI, G. (1998a) Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. *Proc. Natl. Acad. Sci. USA.* 95:9053-9058.

CHIKHI, L., DESTRO-BISOL, G., PASCALI, V.,BARAVELLI, V., DOBOSZ, M., BARBUJANI, G. (1998b) Clinal variation in the DNA of Europeans. *Hum. Biol.* 70:643-657.

COMAS, D., CALAFELL, F., MATEU, E., PÉREZ-LEZAUN, A., BOSCH, E., BERTRANPETIT, J. (1997) Mitochondrial DNA variation and the origin of the Europeans. *Hum. Genet.* 99:443-449.

COMAS, D., MATEU, E., CALAFELL, F., PÉREZ-LEZAUN, A., BOSCH, E., MARTÍNEZ-ARIAS, R., BERTRANPETIT, J. (1998) HLA class I and class II DNA typing and the origin of Basques. *Tissue Antigens*. 51:30-40.

CÔRTE-REAL, H., MACAULAY, V.A., RICHARDS, M.B., HARITI, G., ISSAD, M.S., CAMBON-THOMSEN, A. *et al.* (1996). Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann. Hum. Genet.* 60:331-350.

CROW, J. F. (1958) Some possibilities for measuring selection intensities in man. *Hum. Biol.* 30:1-13.

DENNELL, R. (1983) *European economic prehistory: a new approach*. London: Academic Press.

DONNELLY, P. (1996) Interpreting genetic variability: the effects of shared evolutionary history. In WEISS, K. (ed) *Variation in the human genome.* Wiley, Chichester, England (CIBA Foundation Symposium), pp. 25-50.

DUARTE, C., MAURÍCIO, J., PETTITT, P., SOUTO, P., TRINKAUS, E., VAN DER PLICHT, ZILHÃO, J. (1999) The early Upper Palaeolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and the modern human emergence in Iberia. *Proc. Natl. Acad. Sci. USA* 96:7604-7609.

ELSON, J.L., SAMUELS, D.C., TURNBULL, D.M., CHINNERY, P.F. (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.* 68:802-806.

ESPINHEIRA, R., GEADA, H., RIBEIRO, T., REYS, L. (1996) STR analysis-HUMTH01 and HUMFES/FPS for forensic application. *Adv. Forensic Haemogenet.* 6:528.

EXCOFFIER, L. (1990) Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *J. Mol. Evol.* 30:125-139.

EXCOFFIER, L., SCHNEIDER, S. (1999) Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proc. Natl. Acad. Sci. USA* 96:10597-10602.

FARINHA, A.D. (1999). *Os Portugueses em Marrocos.* Instituto Camões. Colecção Lazúli.

FISHER, E.M., BEER-ROMERO, P., BROWN, L.G., RIDLEY, A., MCNEIL, J.A., LAWRENCE, J.B., WILLARD, H.F., BIEBER, F.R., PAGE, D.C. (1990) Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell.* 63:1205-1218.

FOOTE, S., VOLLRATH, D., HILTON, A., PAGE, D.C. (1992) The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science.* 258:60-66.

GILES, R.E., BLANC, H., CANN, H.M., WALLACE, D.C. (1980) Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA.* 77:6715-6719.

GRAVEN, L., PASSARINO, G., SEMINO, O., BOURSOT, P., SANTACHIARA-BENERECETTI, S., LANGANEY, A., EXCOFFIER, L. (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol. Biol. Evol.* 12:334-345.

GUSMÃO, L., PRATA, M.J., AMORIM, A., SILVA, F., BESSA, I. (1997) Characterisation of four short tandem repeat (STR) loci - TH01, VWA31/A, CD4 and TP53 - in North Portugal. *Hum Biol.* 69:31-40.

HAMMER, M.F. (1995) A recent common ancestry for human Y chromosome. *Nature* 378:376-378.

HAMMER, M.F., KARAFET, T., RASANAYAGAM, A., WOOD, E.T., ALTHEIDE, T.K., JENKINS, T., GRIFFITHS, R.C., TEMPLETON, A.R., ZEGURA, S.L. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15:427-441.

HARDING, R.M., FULLERTON, S.M., GRIFFITHS, R.C., BOND, J., COX, M.J., SCHNEIDER, J.A., MOULIN, D.S., CLEGG, J.B. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* 60:772-789.

HARPENDING, H.C., SHERRY, S. T., ROGERS, A.R., STONEKING, M. (1993) The genetic structure of ancient human populations. *Curr. Anthropol.* 34:483-496.

HARPENDING, H.C. (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Hum. Biol.* 66:591-600.

HARPENDING, H.C., BATZER, M.A., GRUVEN, M.A., JORDE, L.B., ROGERS, A.R., SHERRY, S.T. (1998) Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* 95:1961-1967.

HARRIS, E.E., HEY, J. (1999) X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. USA.* 96:3320-3324.

HASSAN, F.A. (1973) On mechanisms of population growth during the Neolithic. *Curr. Anthropol.* 14:535-542.

HOWELL N., KUBACKA I., MACKEY D.A. (1996) How rapidly does the human mitochondrial genome evolve? *Am. J. Hum. Genet.* 59:501-509.

HOWELL, N. (1997) MtDNA recombination: what do in vitro data mean? *Am. J. Hum. Genet.* 61:18-22.

HURLES M.E., IRVEN C., NICHOLSON J., TAYLOR P.G., SANTOS F.R., LOUGHLIN J., JOBLING M.A., SYKES B. (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. *Am. J. Hum. Genet.* 63:1793-1806.

HURLES M.E., VEITIA R., ARROYO E., ARMENTEROS M., BERTRANPETIT J., PÉREZ-LEZAUN A., BOSCH E., SHLUMUKOVA M., CAMBON-THOMSEN A., MCELREAVEY K.,

LÓPEZ DE MUNAIN A., RÖHL A., WILSON I.J., SINGH L., PANDYA A., SANTOS F.R., TYLER-SMITH C., JOBLING M.A. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am. J. Hum. Genet.* 65:1437-1448.

INGMAN, M., KAESSMANN, H., PÄÄBO, S., GYLLENSTEN, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713.

IZAGIRRE, N., DE LA RÚA, C. (1999) An mtDNA analysis in ancient Basque populations: implications for haplogroup V as a marker for a major paleolithic expansion from southwestern Europe. *Am. J. Hum. Genet.* 65:199-207.

JAZIN E, SOODYALL H, JALONEN P, LINDHOLM E, STONEKING M, GYLLENSTEN U (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat Genet* 18:109-110.

JOBLING, M.A. (1994) A survey of long-range DNA polymorphisms on the human Y chromosome. *Hum. Mol. Genet.* 3:107-114.

JOBLING, M.A., PANDYA, A., TYLER-SMITH, C. (1997) The Y chromosome in forensic analysis and paternity testing. *Int. J. Legal Med.* 110:118-124.

JOBLING, M.A., BOUZEKRI, N., TAYLOR, P.G. (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). *Hum. Mol. Genet.* 7:643-653.

JOBLING, M.A., TYLER-SMITH, C. (2000) New uses for new haplotypes: the human Y chromosome, disease, and selection. *Trends Genet.* 16:356-362.

JORDE, L.B., ROGERS, A.R., BAMSHAD, M., WATKINS, W.S., KRAKOWIAK, P., SUNG, S., KERE, J., HARPENDING, H.C. (1997) Microsatellite diversity and the demographic history of modern humans. *Proc. Natl. Acad. Sci. USA.* 94:3100-3103.

JORDE, L.B., WATKINS, W. S., BAMSHAD, M.J., DIXON, M.E., RICKER, C.E., SEIELSTAD, M.T., BATZER, M. A. (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* 66:979-988.

KAYSER, M., DE KNIJFF, P., DIELTJES, P., KRAWCZAK, M., NAGY, M., ZERJAL, T., PANDYA, A., TYLER-SMITH, C., ROEWER, L. (1997) Applications of microsatellite-based Y chromosome haplotyping. *Electrophoresis.* 18:1602-1607.

KAYSER, M., ROEWER, L., HEDMAN, M., HENKE, L., HENKE, J., BRAUER, S., KRÜGER, C., KRAWCZAK, M., NAGY, M., DOBOSZ, T., SZIBOR, R., DE KNIJFF, P., STONEKING, M., SAJANTILA, A. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* 66:1580-1588.

KIVISILD, T., VILLEMS, R. (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science.* 288:1931.

DE KNIJFF, P. (2000) Y chromosome shared by descent or by state. In RENFREW C., BOYLE K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe.* Cambridge. McDonald Institute for Archaeological Research, pp. 301-304.

KRINGS, M., STONE, A., SCHMITZ, R.W., KRAINITZKI, H., STONEKING, M., PÄÄBO, S. (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* 90:19-30.

KRINGS, M., SALEM, A.H., BAUER, K., GEISERT, H., MALEK, A., CHAIX, L. *et al.* (1999) mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration? *Am. J. Hum. Genet.* 64:1166-1176.

KRINGS, M., CAPELLI, C., TSCHENTSCHER, F., GEISERT, H., MEYER, S., VON HAESELER, A., GROSSSCHMIDT, K., POSSNERT, G., PAUNOVIC, M., PÄÄBO, S. (2000) A view of Neandertal genetic diversity. *Nat Genet.* 26:144-146.

LAHN, B.T., PAGE, D.C. (1997) Functional coherence of the human Y chromosome. *Science* 278:675-679.

LAHR, M.M., FOLEY, R.A., PINHASI, R. (2000) Expected regional patterns of Mesolithic-Neolithic human population admixture in Europe based on archaeological evidence. In RENFREW, C., BOYLE, K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe.* Cambridge. McDonald Institute for Archaeological Research, pp. 81-88.

LANDERS, J. (1992) Reconstructing ancient populations. In JONES, S., MARTIN, R., PILBEAM, D. (eds) *The Cambridge encyclopedia of human evolution.* Cambridge University Press. Cambridge, pp. 402-405.

LAREU, M.V., PHILLIPS, C.P., CARRACEDO, A., LINCOLN, A.J., SYNDERCOMBE-COURT, D, THOMSON, J.A. (1994) Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats; UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci. Int.* 66:41-52.

MALASPINA, P., PERSICHETTI, F., NOVELLETTO, A., IODICE, C., TERRENATO, L., WOLFE, J., FERRARO, M., PRANTERA, G. (1990) The human Y chromosome shows low level of DNA polymorphism. *Ann. Hum. Genet.* 54:297-305.

MARJORAM, P., DONNELLY, P. (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics.* 136:673-683.

MEDINA, J. (1998) *História de Portugal. Vol. II O mundo Luso-Romano.* Ediclube. Lisboa.

MENOZZI, P., PIAZZA, A., CAVALLI-SFORZA, L. (1978) Synthetic maps of human gene frequencies in Europeans. *Science.* 201:786-792.

MESA, N.R., MONDRAGÓN, M.C., SOTO, I.D., PARRA, M.V., DUQUE, C., ORTÍZ-BARRIENTOS, D., GARCÍA, L.F., VELEZ, I.D., BRAVO, M.L., MÚNERA, J.G., BEDOYA,

G., BORTOLINI, M.-C., RUIZ-LINARES, A. (2000) Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-columbian patterns of gene flow in South America. *Am. J. Hum. Genet.* 67:1277-1286.

MEYER, S., WEISS, G, VON HAESELER, A. (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics.* 152:1103-1110.

MOUNTAIN, J.L., CAVALLI-SFORZA, L.L. (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc. Natl. Acad. Sci. USA.* 91:6515-6519.

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

OVCHINNIKOV, I.V., GÖTHERSTRÖM, A., ROMANOVA, G.P., KHARITONOV, V.M., LIDÉN, K., GOODWIN, W. (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* 340:490-493.

PARSONS TJ, MUNIEC DS, SULLIVAN K, WOODYATT N, ALLISTON-GREINER R, WILSON MR, BERRY DL, HOLLAND KA, WEEDN VW, GILL P, HOLLAND MM (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* 15:363-368.

PARRA, E.J., MARCINI, A., AKEY, J., MARTINSON, J., BATZER, M.A., COOPER, R., FORRESTER, T., ALLISON, D.B., DEKA, R., FERRELL, R.E., SHRIVER, M.D. (1998) Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* 63:1839-1851.

PINHASI, R., FOLEY, R.A., LAHR, M.M. (2000) Spatial and temporal patterns in the Mesolithic-Neolithic archaeological record of Europe. In RENFREW C., BOYLE K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe.* Cambridge. McDonald Institute for Archaeological Research, pp. 45-56.

PRITCHARD, J.K., SEIELSTAD, M.T., PÉREZ-LEZAUN, A., FELDMAN, M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16:1791-1798.

RANDO, J.C., CABRERA, V.M., LARRUGA, J.M., HERNÁNDEZ, M., GONZÁLEZ, A.M., PINTO, F., BANDELT, H.-J. (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann. Hum. Genet.* 63:413-428.

RENFREW, C. (1987) *Archaeology & languages. The puzzle of Indo-European origins.* London: Pimlico.

RENFREW, C. (2000) Archaeogenetics: towards a population prehistory of Europe. In RENFREW C., BOYLE K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe.* Cambridge. McDonald Institute for Archaeological Research, pp. 3-11.

RENFREW, C., BAHN, P. (2000) *Archaeology: theories, methods and pratice.* London: Thames and Hudson. 3rd ed.

RICHARDS, M.B., CÔRTE-REAL, H., FORSTER, P., MACAULAY, V., WILKINSON-HERBOTS, H., DEMAINE, A. *et al.* (1996). Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. *Am. J. Hum. Genet.* 59:185-203.

RICHARDS, M., MACAULAY, V., HICKEY, E., VEGA, E., SYKES, B., GUIDA, V., REI GO, C., SELLITTO, D., CRUCIANI, F., KIVIVILD, T., VILLEMS, R., THOMAS, M., RYCH OV, S., RYCHKOV, O., RYCHKOV, Y., GÖLGE, M., DIMITROV, D., HILL, E., BRADLEY D., ROMANO, V., CALÌ, F., VONA, G., DEMAINE, A., PAPIHA, S., TRIANNTAPHYLLIDIS C., STEFANESCU, G., HATINA, J., BELLEDI, M., DI RIENZO, A., NOVELLETTO, A., OPPENHEIM, A., NØRBY, S., AL-ZAHERI, N., SANTACHIARA-BENERECETTI, S., SCOZZARI, R., TORRONI, A., BANDELT, H.-J. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* 67:1251-1276.

ROGERS, A.R., HARPENDING, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9:552-569.

ROGERS, A.R., JORDE, L.B. (1995) Genetic evidence on modern human origins. *Hum. Biol.* 67:1-36.

ROGERS, A.R., FRALEY, A.E., BAMSHAD, M.J., WATKINS, W.S., JORDE, L.B. (1996) Mitochondrial mismatch analysis is insensitive to the mutational process. Mol. Biol. Evol. 13:895-902.

ROSSER, Z.H., ZERJAL, T., HURLES, M.E., ADOJAAN, M.A., ALAVANTIC, D., AMORIM, A., AMOS, W., ARMENTEROS, M., ARROYO, E., BARBUJANI, G., BECKMAN, L., BERTRANPETIT, J., BOSCH, E., BRADLEY, D.G., BREDE, G., COOPER, G.C., CÔRTE-REAL, H.B.S.M., DE KNIJFF, P., DECORTE, R., DUBROVA, Y.E., EVGRAFOV, O., GILISSEN, A., GLISIC, S., GÖLGE, M., HILL, E.W., JEZIOROWSKA, A., KALAYDJIEVA, L., KAYSER, M., KRAVCHENCO, S.A., LAVINHA, J., LIVSHITS, L.A., MARIA, S., MCELREAVEY, K., MEITINGER,T.A., BELA MELEGH, B., MITCHELL, R.J., NICHOLSON, J., NØRBY, S., NOVELLETTO, A., PANDYA, A., PARIK, J., PATSALIS, P.C., PEREIRA, L., PETERLIN, B., PIELBERG, G., PRATA, M.J., PREVIDERÉ, C., RAJCZY, K., ROEWER, L., ROOTSI, S., RUBINSZTEIN, D.C., SAILLARD, J., SANTOS, F.R., SHLUMUKOVA, M., STEFANESCU, G., SYKES, B.C., TOLUN, A., VILLEMS, R., TYLER-SMITH, C., JOBLING, M.A. (2000) Y-chromosomal diversity within Europe is clinal and influenced primarily by geography rather than language. *Am. J. Hum. Genet.* 67:1526-1543.

RUSSELL-WOOD, A.J.R. (1998) The Portuguese empire, 1415-1808. A world on the move. The Johns Hopkins University Press. Baltimore.

SALAS, A., COMAS, D., LAREU, M.V., BERTRANPETIT, J., CARRACEDO, A. (1998). mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur. J. Hum. Genet.* 6:365-375.

SANTOS, F.R., TYLER-SMITH, C. (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. *Braz. J. Genet.* 19:665-670.

SARAIVA, J.H. (1993) *História de Portugal.* Publicações Europa-América. 4ª Ed. Lisboa.

SEIELSTAD, M.T., MINCH, E., CAVALLI-SFORZA, L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genet.* 20:278-280.

SEIELSTAD, M., BEKELE, E., IBRAHIM, M., TOURÉ, A., TRAORÉ, M. (1999) A view of modern human origins from Y chromosome microsatellite variation. *Genome Res.* 9:558-567.

SEMINO, O., PASSARINO, G., BREGA, A., FELLOUS, M., SANTACHIARA-BENERECETTI, A.S. (1996) A view of Neolithic Demic Diffusion in Europe through two Y chromosome-specific markers. *Am. J. Hum. Genet.* 59:964-968.

SEMINO, O., PASSARINO, G., OEFNER, P.J., LIN, A.A., ARBUZOVA, S., BECKMAN, L.E., DE BENEDICTIS, G., FRANCALACCI, P., KOUVATSI, A., LIMBORSKA, S., MARCIKIAE, M., MIKA, A., MIKA, B., PRIMORAC, D., SANTACHIARA-BENERECETTI, A.S., CAVALLI-SFORZA, L.L., UNDERHILL, P.A. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science.* 290:1155-1159.

SERRÃO, J., MARQUES, A.H.O. (1993) *Nova História de Portugal. Vol II Portugal das Invasões Germânicas à "Reconquista"*. Editorial Presença. Lisboa.

SERRÃO, J., MARQUES, A.H.O. (1996) *Nova História de Portugal. Vol III Portugal em definição de fronteiras (1096-1325). Do Condado Portucalense à crise do século XIV.* Editorial Presença. Lisboa.

SHEN, P., WANG, F., UNDERHILL, P.A., FRANCO, C., YANG, W.H., ROXAS, A., SUNG, R., LIN, A.A., HYMAN, R.W., VOLLRATH, D., DAVIS, R.W., CAVALLI-SFORZA, L.L., OEFNER, P.J. (2000) Population genetics implications from sequence variation in four Y chromosome genes. *Proc. Natl. Acad. Sci. USA.* 97:7354-7359.

SHERRY, S.T., ROGERS, A.R., HARPENDING, H.C., SOODYALL, H., JENKINS, T., STONEKING, M. (1994) Pairwise differences of mtDNA reveal recent human population expansions. *Hum. Biol.* 66:761-776.

SIGURÐARDÓTTIR S., HELGASON A., GULCHER J.R., STEFANSSON K., DONNELLY P. (2000) The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.* 66:1599-1609.

SIMONI, L., GUERESI, P., PETTENER, D., BARBUJANI, G. (1999) Patterns of gene flow inferred from genetic distances in the Mediterranean region *Hum. Biol.* 71:399-415.

SIMONI, L., CALAFELL, F., PETTENER, D., BERTRANPETIT, J., BARBUJANI, G. (2000) Geographic patterns of mtDNA diversity in Europe. *Am. J. Hum. Genet.* 66:262-278.

SOODYALL, H. (1993) Mitochondrial DNA polymorphisms in Southern African populations. PhD thesis, University of the Witwatersrand, Johannesburg.

SOODYALL, H., VIGILANT, L., HILL, A. V., STONEKING, M. & JENKINS, T. (1996). MtDNA control-region sequence variation suggests multiple origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am. J. Hum. Genet.* 58:595-608.

SPURDLE, A., JENKINS, T. (1992) The search for Y chromosome polymorphism is extended to negroids. *Hum. Mol. Genet.* 1:169-170.

SOUTO, L., AMORIM, A., VIDE, M.C. (1998) Population and segregation data on the multiplex system (TH01, VWA, FES, F13A1) from central Portugal. *Progress in Forensic Genetics.* 7:363-365.

STRAUS, L.G., BICHO, N., WINEGARDNER, A.C. (2000) The Upper Palaeolithic settlement of Iberia: first-generation maps. *Antiquity.* 74:553-566.

TATTERSALL, I., SCHWARTZ, J. (1999) Hominids and hybrids: the place of Neanderthals in human evolution. *Proc. Natl. Acad. Sci. USA.* 96:7117-7119.

THYAGARAJAN, B., PADUA, R.A., CAMPBELL, C. (1996) Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.* 271:27536-27543.

THOMAS, H. (1998). *The slave trade – the history of the Atlantic slave trade 1440-1870.* London: Macmillan Publishers Ltd.

THOMAS, M.G., PARFITT, T., WEISS, D.A., SKORECKI, K., WILSON, J.F., LE ROUX, M., BRADMAN, N., GOLDSTEIN, D.B. (2000) Y chromosome traveling south: the Cohen modal haplotype and the origins of the Lemba - the "black Jews of southern Africa". *Am. J. Hum. Genet.* 66:674-686.

TISHKOFF, S.A., DIETZSCH, E., SPEED, W., PAKSTIS, A.J., KIDD, J.R., CHEUNG, K., BONNE-TAMIR, B., SANTACHIARA-BENERECETTI, A.S., MORAL, P., KRINGS, M. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science.* 271:1380-1387.

TORRONI, A., SCHURR, T. G., YANG, C.-C., SZATHMARY, E. J. E., WILLIAMS, R. C. *et al.* (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics* 130:153-162.

TORRONI, A., BANDELT, H.-J., D'URBANO, L., LAHERMO, P., MORAL, P., SELLITTO, D., RENGO, C., FORSTER, P., SAVONTAUS, M.-L., BONNÉ-TAMIR, B., SCOZZARI, R. (1998) MtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern European populations. *Am. J. Hum. Genet.* 62:1137-1152.

TORRONI, A., BANDELT, H.-J., VINCENT, M., RICHARDS, M., CRUCIANI, F., RENGO, C., MARTINEZ-CABRERA, V., VILLEMS, R., KIVISILD, T., METSPALU, E., PARIK, J., TOLK, H.-V., TAMBETS, K., FORSTER, P., KRAGER, B., FRANCALACCI, P., RUDAN, P., JANICIJEVIC, B., RICKARDS, O., SAVONTAUS, M.-L., HUOPONEN, K., LAITINEN, V., KOIVUMÄKI, S., SYKES, B., HICKEY, E., NOVELLETTO, A., MORAL, P., SELLITTO, D., COPPA, A., AL-ZAHERI, N., SANTACHIARA-BENERECETTI, A.S., SEMINO, O., SCOZZARI, R. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am. J. Hum. Genet.* 69:844-852.

UNDERHILL, P.A., JIN, L., LIN, A.A., MEHDI, S.Q., JENKINS, T., VOLLRATH, D., DAVIS, R.W., CAVALLI-SFORZA, L.L., OEFNER, P.J. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7:996-1005.

UNDERHILL, P.A., SHEN, P., LIN, A.A., JIN, L., PASSARINO, G., YANG, W.H., KAUFFMAN, E., BONNÉ-TAMIR, B., BERTRANPETIT, J., FRANCALACCI, P., IBRAHIM, M., JENKINS, T., KIDD, J.R., MEHDI, S.Q., SEIELSTAD, M.T., WELLS, R.S., PIAZZA, A., DAVIS, R.W., FELDMAN, M.W., CAVALLI-SFORZA, L.L., OEFNER, P.J. (2000) Y chromosome sequence variation and the history of human populations. *Nature Genet.* 26: 358-361.

VIGILANT, L., PENNINGTON, R., HARPENDING, H., KOCHER, T.D. (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA.* 86:9350-9354.

VIGILANT, L., STONEKING, M., HARPENDING, H., HAWKES, K., WILSON, A. C. (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507.

WAKELEY, J. (1994) Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11:436-442.

WARD, R.H., FRAZIER, B.L., DEW-JAGER, K., PÄÄBO, S. (1991) Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA.* 88:8720-8724.

WATSON, E., FORSTER, P., RICHARDS, M. & BANDELT, H.-J. (1997) Mitochondrial footprints of human expansions in Africa. *Am. J. Hum. Genet.* 61:691-704.

WEBER, J.L., WONG, C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2:1123-1128.

WISE, C.A., SRAMIL, M., RUBINSZTEIN, D.C., EASTEAL, S. (1998) Comparative nuclear and mitochondrial genome diversity in humans and chimpanzees. *Mol. Biol. Evol.* 14:707-716.

WYCKOFF, G.J., WANG, W., WU, C. (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304-308.

WOLPOFF, M.H., WU, X., THORNE, A.G. (1984) Modern *Homo sapiens* origins: a general theory of hominid evolution involving the fossil evidence from East Asia. In: *The origins of Modern Humans: a World survey of the fossil evidence*. Liss. New York.

ZERJAL, T., DASHNYAM, B., PANDYA, A., KAYSER, M., ROEWER, L., SANTOS, F.R., SCHIEFENHOVEL, W., FRETWELL, N., JOBLING, M.A., HARIHARA, S., SHIMIZU, K., SEMJIDMAA, D., SAJANTILA, A., SALO, P., CRAWFORD, M.H., GINTER, E.K., EVGRAFOV, O.V., TYLER-SMITH, C. (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am. J. Hum. Genet.* 60:1174-1183.

ZILHÃO, J. (1997) Maritime pioneer colonisation in the Early Neolithic of the west Mediterranean: testing the model against the evidence. *Porocilo* 24:19-42.

ZHIVOTOVSKY, L.A., FELDMAN, M.W., GRISHECHKIN, S.A. (1997) Biased mutations and microsatellite variation. *Mol. Biol. Evol.* 14:926-933.

ZVELEBIL, M., ZVELEBIL, K.V. (1988) Agricultural transition and Indo-European dispersal. *Antiquity*. 62:574-583.

ZVELEBIL, M. (2000) The social context of the agricultural transition in Europe. In RENFREW C., BOYLE K. (eds) *Archaeogenetics: DNA and the population prehistory of Europe*. Cambridge. McDonald Institute for Archaeological Research, pp. 57-79.