Susana Alves Seixas

Análise dos padrões de diversidade genética em três regiões funcionalmente relevantes do genoma humano

Apolipoproteína E (APOE)

α1-antitripsina (PI)

Grupo Sanguíneo Duffy (FY)

PORTO

2002

Susana Alves Seixas

Análise dos padrões de diversidade genética em três regiões funcionalmente relevantes do genoma humano
Apolipoproteína E (APOE)
α1-antitripsina (PI)
Grupo Sanguíneo Duffy (FY)

PORTO

2002

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

Agradecimentos

Ao chegar ao fim do presente trabalho gostaria de expressar a minha gratidão a todos aqueles que de algum modo contribuíram para a sua concretização.

Ao Doutor Jorge Rocha queria prestar os meus sinceros agradecimentos por todo o seu empenho na difícil tarefa de orientador e pela confiança e liberdade dadas na gestão e execução dos trabalhos. Estes anos de convivência nem sempre foram fáceis mas foi através de inúmeras discussões, umas mais acesas que outras, que tive a oportunidade de crescer quer ao nível científico quer pessoal.

Ao Professor Doutor Sobrinho Simões agradeço o interesse devotado ao meu trabalho e o acolhimento no IPATIMUP que proporcionou todas as condições materiais para a sua realização.

Ao Professor Doutor António Amorim agradeço igualmente o apreço que desde sempre revelou pelo meu trabalho e a integração no grupo de Genética Populacional do IPATIMUP que me permitiu aumentar os conhecimentos nesta fascinante área de investigação.

A todos os co-autores das publicações apresentadas no âmbito desta dissertação agradeço a preciosa colaboração.

Aos clínicos agradeço a cedência de amostras de material biológico e a pronta disponibilidade.

A todos aqueles que comigo partilharam a bancada do laboratório (Jorge Reis Sá, Isabel Alonso, Luísa Azevedo, Luísa Seco, Gil Tomás, Paulo Gaspar e Teresa Pacheco) quero agradecer a cumplicidade e apoio, fundamentais nos momentos mais críticos e o privilégio de conjuntamente ampliarmos a nossa formação científica.

Aos restantes colegas que pertencem ou pertenceram ao Grupo de Genética Populacional (Fátima Santos, Leonor Gusmão, Luís Filipe, Cíntia, Sandra Alves,

Luísa Pereira, Sandra Beleza, Sandra Martins, Alexandra Lopes, Solange e Carla) agradeço a boa dose de discussões espirituosas, o companheirismo e amizade.

A todos os membros do IPATIMUP um muito especial obrigado pela vossa cooperação e bom ambiente de trabalho.

Aos elementos do Centro de Estudos de Ciência Animal do ICETA agradeço o vosso caloroso acolhimento e troca de experiências durante os meus breves estágios laboratoriais.

Ao Guillaume Queney agradeço a fabulosa revisão do resumo em francês.

Ao Instituto de Cooperação Científica e Tecnológica Internacional agradeço a concessão de um subsídio que permitiu a deslocação a São Tomé e Príncipe para a realização de um trabalho de campo de recolha de amostras de material biológico. Ao Ministério da Saúde de São Tomé e Príncipe agradeço o suporte prestado na execução do trabalho, que em muito contribuiu para o seu sucesso.

À Fundação para Ciência e a Tecnologia agradeço a concessão da bolsa de Doutoramento sem a qual não poderia ter realizado os trabalhos conducentes a esta dissertação.

Aos meus amigos e ao Miguel agradeço a vossa presença, que tantas vezes me serviu de apoio e o incansável incentivo.

Finalmente agradeço aos meus familiares toda a compreensão e a possibilidade de escolher o meu próprio caminho.

Índice

Resumo

Os estudos de variabilidade genética ao nível populacional podem-se centrar na história evolutiva das populações ou na história de certos genes em particular. A análise da história das populações está na base de progressos muito importantes na interpretação das relações interpopulacionais e dos principais padrões de migração, mas tende a não aprofundar o estudo das propriedades específicas de cada marcador genético. A caracterização da variação genética *locus* a *locus* é inadequada para a descrição da história populacional, mas permite uma análise mais detalhada dos aspectos que condicionaram a actual diversidade das regiões codificantes do genoma humano, incluindo o estudo das variedades de processos mutacionais, a pesquisa de diferenças funcionais entre alelos e o seu impacto no risco de doença, a avaliação do papel da selecção, bem como a identificação do local de origem de certas mutações e a análise das forças que influenciaram a sua subsequente dispersão.

Esta tese constitui uma contribuição para a compreensão dos padrões de variação genética em três genes polimórficos: apolipoproteína E (APOE), α1-antitripsina (PI) e grupo sanguíneo Duffy (FY). A variação alélica do *locus* da APOE determina uma importante fracção da variabilidade nos níveis de colesterol total e está associada ao risco de doença de Alzheimer, proporcionando um bom exemplo da influência de um polimorfismo genético comum na susceptibilidade a doenças degenerativas e não infecciosas. O gene PI apresenta um elevado grau de polimorfismo com diversos alelos normais e muitas variantes patogénicas que causam a deficiência da proteína, constituindo um bom modelo para o estudo da história natural de um polimorfismo. O grupo sanguíneo FY está associado à resistência à malária por *Plasmodium vivax* e é particularmente adequado ao estudo do impacto das doenças infecciosas na modulação da variação genética.

Com o objectivo de caracterizar a diversidade genética da apolipoproteína E, estudaram-se as distribuições conjuntas dos fenótipos da APOE e de um polimorfismo de restrição no *locus* vizinho da apolipoproteína C1 (APOC1), em amostras da população portuguesa, do País Basco e de São Tomé e Príncipe (Golfo da Guiné), uma

ex-colónia portuguesa originalmente povoada com escravos importados da África continental. As frequências dos alelos da APOE (*2, *3 e *4) nas amostras ibéricas e são-tomenses incluem-se na gama de valores geralmente observados nas populações europeias e africanas, respectivamente. Em São Tomé, observaram-se diferenças significativas entre as frequências da APOE nas localidades situadas a norte e a sul da ilha, como seria de esperar de um padrão de povoamento caracterizado pela restrição do fluxo génico entre as duas regiões. De acordo com a tendência global da distribuição da variação genética humana, a amostra europeia de Portugal apresentou desequilíbrios gaméticos mais intensos do que a amostra africana de São Tomé onde, apesar da pequena distância de 5 kb que separa os dois *loci*, o nível de associação entre os alelos de APOC1 e APOE*4 não é significativo. Esta diferença pode contribuir para explicar a heterogeneidade dos resultados dos estudos de associação entre APOE e várias patologias realizados em diferentes populações. Adicionalmente, a análise haplotípica demostrou que, quer em Portugal, quer em São Tomé, a intensidade do desequilíbrio gamético é mais elevada para APOE*2 e mais reduzida para APOE*4, o que indica que a geração dos alelos da APOE se deu pela ordem: 4→3→2. As estimativas da idades dos alelos APOE*2, APOE*3 e APOE*4, calculadas com base na intensidade do desequilíbrio gamético, foram de 109 200 (5 730 - 235 800), 575 100 (353 400 - 945 000) e mais de 1 300 000 anos, respectivamente. Estas estimativas indicam que os três alelos da APOE estavam presentes numa população ancestral comum, anterior à hipotética migração para fora de África postulada pelo modelo unirregional de origem do Homem Moderno. Por outro lado, os alelos APOE*2 e APOE*3 parecem ser suficientemente antigos para terem alcançado as suas actuais frequências em condições de neutralidade selectiva.

A fim de investigar a história natural do polimorfismo da α1-antitripsina (PI), analisaram-se os níveis de diversidade haplotípica associados aos diferentes alelos de PI, através do estudo de três microssatélites, localizados na proximidade dos *loci* da globulina transportadora de corticoesteróides (CBG), da α1-antitripsina [PI(TG)n] e do inibidor da proteína C [PCI(TG)n], em quatro populações com diferentes histórias evolutivas: Portugal, País Basco, São Tomé e ameríndios de língua Quechua dos Andes centrais peruanos. A análise da diversidade dos microssatélites revelou que há

diferenças substanciais nos padrões de variação interpopulacionais e interalélicos. De acordo com as respectivas histórias populacionais, as amostras basca e ameríndia apresentam níveis de variação mais reduzidos. Nos ameríndios, a redução de variação no *locus* PI(TG)n é compatível com um estrangulamento de efectivo suficientemente acentuado para eliminar completamente a diversidade, seguido de uma expansão populacional ocorrida há cerca de 9 300 anos. Na amostra africana, apesar de se terem registado os valores mais elevados de diversidade no *locus* PCI(TG)n, observou-se uma redução da heterozigotia associada à linhagem mais antiga M1Ala213 no marcador PI(TG)n, que pode ter resultado de 1) selecção num *locus* flanqueante ou 2) enviezamentos no processo mutacional do microssatélite, conducentes a um equilíbrio caracterizado por uma distribuição estável. Ao contrário do microssatélite mais distante PCI(TG)n, a variação alélica em PI(TG)n reflecte distintas fases de recuperação mutacional da diversidade, associadas a níveis consideráveis de diferenciação entre as variantes proteicas da α1-antitripsina. As estimativas da idade dos alelos da α1-antitripsina, baseadas na variação acumulada nos microssatélites, sugerem que o alelo de deficiência PI*Z terá surgido há 3 600 - 7 400 anos e se terá dispersado no período Neolítico ou pós-Neolítico. Por outro lado, a observação de que os níveis de diversidade associados aos alelos Z na amostra portuguesa são semelhantes, ou até superiores, aos encontrados no norte da Europa indica que a mutação Z pode não ter tido origem na Escandinávia, como foi proposto com base apenas na distribuição das frequências génicas. A idade da mutação S foi estimada em 8 300 - 16 300 anos e sugere que as frequências elevadas deste alelo na Península Ibérica não resultaram de deriva genética recente e que PI*S pode ter tido origem nesta região, tendo sido dispersado, posteriormente, através da expansão populacional associada à recolonização da Europa no período pós-glaciar.

Para caracterizar o espectro mutacional da α1-antitripsina foram sequenciadas 36 variantes raras independentes pertencentes a 15 classes alélicas, detectadas por focagem isoeléctrica em amostras do País Basco, Portugal e São Tomé: I (Arg39Cys; n=7); Slisboa (Ser47Arg; n=2); Mmalton (Phe52del; n=2); T (Arg101His ou Glu264Val; n=3); M4 (Arg101His ou Asp376Glu; n=8); V (Gly148Arg; n=1); Plowell (Asp256Val; n=1); Ieuskadi (Arg281del, n=1); Pdonauwoerth (Asp341Asn; n=2);

Zaugsburg (Glu342Lys; n=1); Q0ourém (Leu353framStop376; n=1); São Tomé (Pro362His; n=1); Mheerlen-2 (Pro369Leu; n=3); Mwürzburg (Pro369Ser; n=2); Q0porto (IVS1C+1G-A; n=1). Cinco destas mutações foram descritas pela primeira vez no decurso do presente trabalho: Ser47Arg, Arg281del, Pro362His, Pro369Ser e IVS1C+1G-A. As mutações Ser47Arg, Arg281del e Pro362His estão associadas a concentrações de PI normais. A substituição de Pro369Ser está associada a uma deficiência da proteína de 85% causada pela acumulação intrahepática da α1-antitripsina. A mutação IVS1C+1G-A afecta o processamento normal do RNA produzindo um alelo nulo. As restantes mutações já tinham sido anteriormente descritas noutras populações. A revisão das mutações no gene da α1-antitripsina anteriormente publicadas demonstrou que, de acordo com resultados obtidos para outros *loci*, os diferentes eventos mutacionais não estão distribuídos ao acaso. As indels estão concentradas na proximidade de motivos TG(A/G)(A/G)(G/T)(A/C), ou no seio de repetições de mononucleotídeos ou de sequências com motivos reiterados. As substituições nucleotídicas estão sobrerepresentadas nos dinucleotídos CpG. A análise das relações genealógicas entre diferentes variantes de PI permitiu identificar nove codões homoplásicos que podem ser especialmente mutáveis: 51/52, 101, 115, 148, 256, 342, 353, 362 e 369. A comparação da sequência da PI com as de outras proteínas homólogas revelou que as mutações patogénicas estão concentradas nos aminoácidos mais conservados.

A história evolutiva do grupo sanguíneo FY foi analisada através do estudo da distribuição da diversidade do microssatélite D1S2635, de elevada mutabilidade, nas linhagens mais estáveis definidas pelos alelos FY*A e FY*B de Portugal e FY*O de São Tomé. A análise conjunta de polimorfismos de sequência encontrados na região flanqueante do microssatélite e de duas posições polimórficas descritas previamente permitiu definir diferentes haplótipos na região 5' do gene FY. De acordo com o actual padrão de distribuição dos alelos FY, a maioria dos haplótipos derivados encontra-se repartida pelos alelos FY*O ou não FY*O. No entanto, verificou-se que duas linhagens derivadas são partilhadas por todos os alelos FY, o que mostra que os alelos FY*O e não FY*O estiveram presentes numa mesma população ancestral antes da fixação de FY*O em África, tendo trocado sequências por recombinação ou conversão

génica. A análise das distribuições das repetições do microssatélite D1S2635 nos haplótipos associados aos alelos FY sugere que as duas principais linhagens de FY*O podem ter surgido por recorrência da transição T46-C que caracteriza este alelo. Com base na acumulação de diversidade no microssatélite D1S2635, calculou-se um limite superior para a data de fixação do alelo FY*O em África de 14 700 (5 100 - 31 800) anos. Esta estimativa é consideravelmente mais recente do que um cálculo anterior de 33 000 anos e aproxima a data de fixação de FY*O das alterações ecológicas e demográficas associadas à transição Paleolítico/Neolítico e à concomitante dispersão da malária como uma pressão selectiva generalizada em África. Tendo em conta estes resultados, é possível que o aumento das frequências de FY*O iniciado no final do Paleolítico na África Ocidental tenha sido amplificado pela introdução da agricultura tropical e facilitado pelos movimentos populacionais associados à expansão Bantu.

Summary

The study of genetic variation at the population level may focus either on the evolutionary history of populations or on the history of particular genes. The analysis of population history has led to substantial advances in the interpretation of population relationships and major migration patterns but tends to overlook the specific properties of each genetic marker. The characterisation of single *locus* variation cannot lead to an adequate description of population history, but it allows a more detailed analysis of the features that have shaped the present diversity in the coding regions of the human genome. This includes the study of the variety of mutation processes, the assessment of functional differences between alleles and its impact on disease risk, the evaluation of the role of selection, as well as the identification of the place of origin of specific mutations and the analysis of the forces that have influenced their subsequent spread.

This thesis is a contribution to the understanding of genetic variation in three polymorphic genes: apolipoprotein E (APOE), α1-antitrypsin (PI) and the Duffy blood group (FY). The allelic variation at the APOE *locus* determines an important fraction of the variability in total cholesterol levels and is associated with the risk of Alzheimer disease, thus providing a good example of the influence of common genetic polymorphisms in the susceptibility to non-infectious and degenerative disease. The PI gene his highly polymorphic and has several normal alleles and several pathogenic variants that cause protein deficiency, thus providing a model for studying the natural history of a polymorphism. The FY blood group is associated with the resistance to *Plasmodium vivax* malaria and is particularly adequate to study the impact of infectious disease in the shaping of genetic variation.

To characterise the apolipoprotein E genetic variation, the joint distributions of phenotypes from APOE and from a closely linked restriction site polymorphism at the apolipoprotein C1 *locus* (APOC1) were studied in population samples from Portugal, the Basque Country and São Tomé island (Gulf of Guinea), a former Portuguese colony that was originally populated by imported slaves from the African mainland. The frequencies of the APOE alleles (*2, *3 and *4) in the Iberian samples and São

Tomé fitted the ranges of variation generally observed in European and African populations, respectively. Within São Tomé, there were significant differences between the APOE allele frequencies from localities in the northern and southern regions of the island, reflecting a settlement pattern characterized by the restriction of gene flow between the two regions. In accordance with global trends in the distribution of human genetic variation, the European sample from Portugal presented more intense linkage disequilibrium between APOE and APOC1 than the African sample from S. Tomé where, despite the short 5 kb distance that separates the two *loci*, the level of association between the APOC1 alleles and APOE*4 was non-significant. This difference may provide a basis for understanding the heterogeneity of disease association studies involving the APOE *locus* in distinct populations. Haplotype analysis has also shown that, both in Portugal and São Tomé, the strength of linkage disequilibrium was highest for APOE*2 and lowest for APOE*4, indicating that the origin of the APOE alleles followed a 4→3→2 pathway. Age estimates of the APOE*2, APOE*3 and APOE*4 alleles based on the decay of linkage disequilibrium were calculated at 109 200 (5 730 - 235 800), 575 100 (353 400 - 945 000) and more than 1 300 000 years, respectively. These estimates indicate that all three APOE alleles were present in a common ancestral population prior to the putative out of Africa migration postulated by the uniregional model of modern human origins. Moreover, both APOE*2 and APOE*3 alleles seem to be old enough to have reached their current average frequencies under selective neutrality.

In order to investigate the natural history of the α1-antitrypsin (PI) polymorphism the levels of haplotype diversity associated with different PI alleles were assessed through the analysis of three microsatellites located within or close to corticosteroid-binding globulin (CBG), α1-antitrypsin [PI(TG)n] and protein C inhibitor [PCI(TG)n] *loci* in four populations with different historic backgrounds: Portugal, Basque Country, São Tomé and Quechua speaking Amerindians from the Peruvian Central Andes. This survey of microsatellite diversity has revealed substantial differences in the patterns of allelic variation both between populations and between protein variants in the same population. In accordance with population history, the Basque and the Amerindian samples presented overall reduced levels of microsatellite variation.

Among the Amerindians, this reduction at the PI(TG)n *locus* is compatible with a strong bottleneck effect leading to a complete reset to zero of variation followed by a population expansion occurring 9 300 years ago. The African sample, while presenting the highest PCI(TG)n diversity, showed a lineage specific reduction in PI(TG)n heterozygosity within the oldest M1Ala213 variant that could have been caused by 1) selection at a closely linked *locus* or 2) biases in the microsatellite mutation process leading to a stable equilibrium distribution. Unlike the more distant PCI(TG)n repeat, allelic variation at PI(TG)n reflected distinct phases of mutational recovery of microsatellite diversity around different founder alleles and showed a considerable differentiation between α1-antitrypsin protein variants. Age estimates of α1-antitrypsin variants based on microsatellite variation suggest that the Z deficiency allele has appeared 3 600-7 400 years ago and was spread in Neolithic or post-Neolithic times. Moreover, the observation of levels of microsatellite diversity in Z types from Portugal that are similar to, or even higher than those found in northern Europe indicates that Z mutation might have not originated in Scandinavia as proposed on the basis of gene frequencies alone. The S mutation has an older 8 300 - 16 300 years age estimation, indicating that its high frequencies in Iberia didn't result from a recent bottleneck and that PI*S could have been originated in this region and dispersed by population expansions associated with postglacial recolonization in Europe.

To characterise the mutation spectrum of α1-antitrypsin we have sequenced 36 unrelated rare variants, belonging to 15 allelic classes, which were detected by protein isoelectric focusing in samples from the Basque Country, Portugal and São Tomé: I (Arg39Cys; n=7); Slisboa (Ser47Arg; n=2); Mmalton (Phe52del; n=2); T (Arg101His or Glu264Val; n=3); M4 (Arg101His or Asp376Glu; n=8); V (Gly148Arg; n=1); Plowell (Asp256Val; n=1); Ieuskadi (Arg281del, n=1); Pdonauwoerth (Asp341Asn; n=2); Zaugsburg (Glu342Lys; n=1); Q0ourém (Leu353framStop376; n=1); São Tomé (Pro362His; n=1); Mheerlen-2 (Pro369Leu; n=3); Mwürzburg (Pro369Ser; n=2); Q0porto (IVS1C+1G-A; n=1). Five mutations were reported for the first time during the course of this work: Ser47Arg, Arg281del, Pro362His, Pro369Ser and IVS1C+1G-A. The mutations Ser47Arg, Arg281del and Pro362His are associated with normal PI

concentrations. The Pro369Ser substitution is associated with a 85% protein deficiency and causes intrahepatic α1-antitrypsin accumulation. The IVS1C+1G-A affects normal RNA splicing and produces a null allele. All the remaining mutations had been previously described in other populations. The review of the data on published mutations has shown that, in accordance with the results from other *loci*, mutation events are not randomly distributed in the PI gene. Indels are either concentrated in the vicinity of TG(A/G)(A/G)(G/T)(A/C) motifs or within mononucleotide runs and reiterated sequence stretches. Nucleotide substitutions are disproportionately represented among CpG dinucleotides. The analysis of the genealogical relations between different PI variants has led to the identification of 9 homoplasic codons that can be especially mutable: 51/52, 101, 115, 148, 256, 342, 353, 362 and 369. Sequence comparisons with PI protein homologues has shown that pathogenic mutations are disproportionately concentrated in the most conserved residues.

The evolutionary history of the FY blood group was approached through the study of the distribution of the faster-mutating D1S2635 microsatellite polymorphism within the more stable lineages defined by FY*A and FY*B alleles from Portugal and FY*O alleles from São Tomé. The combined analysis of microsatellite flanking sequence variation and previously described polymorphic positions led to the definition of different haplotypes in the 5' region upstream the FY gene. In accordance with the present patterns of distribution of FY alleles, most derived haplotypes were sorted along with either FY*O or non-FY*O alleles. However, two derived lineages were found to be shared by all FY alleles, indicating that FY*O and non-FY*O lineages were present in the same ancestral population and exchanged sequences through gene conversion or recombination, prior to FY*O fixation in Africa. The analysis of the distributions of D1S2635 repeat sizes within the haplotypes carried by FY alleles suggests that two major lineages linked to FY*O could have arisen by recurrence of the T-46C transition that characterises this allele. Based on the accumulation of D1S2635 microsatellite diversity the upper limit of the date of fixation of FY*O in Africa was estimated at 14 700 (5 100 - 31 800) years. This estimate is considerably more recent than a previous 33 000 years calculation and

places the fixation date of FY*O closer to the ecological and demographic changes associated with the Palaeolithic/Neolithic transition and to the concomitant spreading of malaria as a generalised selective pressure in Africa. On the basis of these results, it is possible that a selective increase of FY*O frequencies, starting in late Palaeolithic in West Africa, was further amplified by the introduction of tropical agriculture and followed by the dispersion of the allele during the Bantu expansion.

Résumé

L'étude de la variation génétique au niveau populationnel peut être centré, soit dans l'histoire évolutive des populations, soit dans l'histoire de certains gènes en particulier. L'analyse de l'histoire populationnelle a conduit à des progrès considérables dans l'interprétation des relations entre populations et de leurs principaux patrons de migration, mais elle a tendance à ne pas caractériser les propriétés spécifiques de chaque marqueur génétique avec une profondeur suffisante. La caractérisation de la variation à chaque *locus* ne peut pas conduire à une description adéquate de l'histoire populationnelle, mais elle permet une analyse plus détaillée des aspects qui ont déterminés l'actuelle diversité des régions codantes du génome humain. Ce type d'analyse comprend l'étude de la variété des processus de mutation, l'évaluation des différences fonctionnelles entre les allèles et son impact dans le risque de maladie, la caractérisation du rôle de la sélection, aussi bien que l'identification du lieu d'origine de certaines mutations et l'analyse des forces qui ont influencé leur dispersion.

Cette thèse est une contribution à la compréhension de la variation génétique en trois *loci* polymorphes: apolipoprotéine E (APOE), α1-antitrypsine (PI) et le groupe sanguin Duffy (FY). La variation allélique au *locus* APOE est un bon exemple de l'influence de la variation génétique commune à la susceptibilité aux maladies dégénératives et aux maladies non-infectieuses, puisqu'elle détermine une fraction importante de la variabilité des niveaux de cholestérol total et est associée au risque de la maladie d'Alzheimer. Le gène PI fourni un bon modèle pour étudier l'histoire naturelle d'un polymorphisme, puisqu'il est hautement polymorphe et a plusieurs allèles normaux et plusieurs variantes pathologiques qui produisent un déficit de protéine. Le groupe FY est associé à la résistance au paludisme provoqué par *Plasmodium vivax* et il est particulièrement adéquat pour étudier l'impact des maladies infectieuses dans la variation génétique.

Pour caractériser la variation génétique de l'apolipoprotéine E, les distributions des phénotypes de APOE et d'un site de restriction polymorphe du *locus* de

l'apolipoprotéine C1 (APOC1), étroitement lié à APOE, ont été étudiés dans les populations du Portugal, du Pays Basque et de l'île de São Tomé (Golfe de Guinée), une ancienne colonie portugaise qui a été peuplée par des esclaves importés des côtes voisines d'Afrique. Les fréquences des allèles de APOE (*2, 3 et 4) dans les populations Ibériques et de São Tomé sont comprises dans les intervalles de variation observés en Europe et en Afrique, respectivement. A São Tomé, les fréquences des allèles de APOE dans les villages du nord et du sud de l'île sont significativement différentes et reflètent un patron de peuplement caractérisé par la restriction des échanges génétiques entre les deux régions. En accord avec la tendance globale de distribution de la variation génétique humaine, l'échantillon européen du Portugal a montré un plus fort déséquilibre de liaison entre APOE et APOC1 que l'échantillon Africain de São Tomé où, malgré la petite distance de 5 kb qui sépare les deux *loci*, le niveau d'association entre les allèles de APOC1 et APOE*4 n'est pas significatif. Cette différence peut expliquer l'hétèrogenité des études d'association entre le *locus* APOE et différentes maladies dans plusieurs populations. L'analyse haplotypique a aussi montré que, dans les échantillons du Portugal et de São Tomé, le déséquilibre de liaison est plus fort pour APOE*2 et plus faible pour APOE*4, ce qui indique que l'origine des allèles de APOE a suivi l'ordre 4→3→2. Les âges des allèles APOE*2, APOE*3 et APOE*4, calculé à partir des intensités du déséquilibre de liaison, ont été estimés à 109 200 (5 730 - 235 800), 575 100 (353 400-945 000) et plus de 1 300 000 années, respectivement. Ces valeurs indiquent que les trois allèles dehors étaient présents dans une population ancestrale commune avant l'hypothétique migration en d'Afrique proposée par le modèle unirrégional de l'origine de l'Homme Moderne. En plus, les âges calculés suggèrent que les allèles APOE*2 et APOE*3 sont suffisamment anciens pour atteindre leurs fréquences moyennes actuelles en condition de neutralité sélective.

Afin de caractériser l'histoire naturelle du polymorphisme de l'α1-antitrypsine, les niveaux de diversité haplotypique associés aux différents allèles de PI ont été évalués par l'analyse de trois microsatellites localisés près des *loci* de la globuline de transport des corticostéroïdes (CBG), de l' α1-antitrypsine [PI(TG)n] et de l'inhibiteur de la protéine C [PCI(TG)n] pour quatre populations avec différentes histoires:

Portugal, Pays Basque, São Tomé et amérindiens parleurs du Quechua des Andes centrales du Pérou. L'étude de la diversité des microsatellites a révélé d'importantes différences dans les patrons de variation, soit entre populations, soit entre variants protéiques de la même population. En accord avec l'histoire des populations, les échantillons basque et amérindien ont présenté un niveau global de diversité plus réduit. Chez les amérindiens, au *locus* PI(TG)n, cette réduction est compatible avec un fort *bottleneck* génétique conduisant à la perte complète de variation, suivie d'une expansion populationnelle il y a 9 300 années. L'échantillon africain a présenté des niveaux plus hauts de diversité au *locus* PCI(TG)n, mais a subi une réduction d'hétérozygotie pour l'allèle plus ancien M1Ala213 qui peut être causée par 1) la sélection dans un *locus* voisin ou par 2) les particularités du processus de mutation qui pourraient conduire à une distribution d'équilibre stable. Contrairement au *locus* PCI-(TG)n plus éloigné, la variation allélique du microsatellite PI-(TG)n est associée à différentes phases de récupération mutationnelle de la diversité et est caractérisée par une différenciation considérable entre les allèles de l'α1-antitrypsine. Les estimations de l'âge des allèles de l'α1-antitrypsine, basées sur la variation des microsatellites, suggèrent que l' allèle de déficit Z est né il y a 3 600 - 7 400 années et que sa dispersion date de la période Néolithique ou post-Néolithique. Le fait que les niveaux de diversité des allèles Z du Portugal et du nord de l'Europe sont équivalents indique que la mutation Z n'est pas originaire de Scandinavie comme on l'a proposé en tenant compte seulement de la distribution des fréquences de l'allèle. L'âge de la mutation S est estimé à 8 300-16 300 années ce qui indique que les plus hautes fréquences de PI*S observées en Ibérie ne sont pas dues à un *bottleneck* récent et que l'allèle peut être apparu dans cette région et avoir été dispersé par les expansions populationnelles associées à la recolonisation de l'Europe après le dernier climax de glaciation.

Afin de caractériser le spectre des mutations de l' α1-antitrypsine, on a séquencé 36 produits géniques rares et indépendants, appartenant à 15 classes alléliques qui ont été détectées par focalisation isoélectrique des protéines dans des échantillons du Pays Basque, du Portugal et de São Tomé : I (Arg39Cys; n=7); Slisboa (Ser47Arg; n=2); Mmalton (Phe52del; n=2); T (Arg101His ou Glu264Val; n=3); M4 (Arg101His ou Asp376GluVal; n=8); V (Gly148Arg; n=1); Plowell (Asp256Val; n=1); Ieuskadi

(Arg281del, n=1); Pdonauwoerth (Asp341Asn; n=2); Zaugsburg (Glu342Lys; n=1); Q0ourém (Leu353framStop376; n=1); Sãotomé (Pro362His; n=1); Mheerlen-2 (Pro369Leu; n=3); Mwürzburg (Pro369Ser; n=2); Q0porto (IVS1C+1G-A; n=1). Cinq mutations ont été rapportées pour la première fois pendant la réalisation du travail: Ser47Arg, Arg281del, Pro362His, Pro369Ser et IVS1C+1G-A. Les mutations Ser47Arg, Arg281del et Pro362His sont associées à des concentrations normales de PI. La substitution Pro369Ser est associée à un déficit de protéine de 85% et cause l'accumulation intra-hépatique de l'α1-antitrypsine. La mutation IVS1C+1G-A perturbe la maturation du RNA et produit un allèle nul. Toutes les autres mutations ont été décrites antérieurement. La révision des mutations publiées a montré que, en accord avec les résultats d'autres *loci*, les mutations ne sont pas distribuées au hasard dans le gène de PI. Les indels se trouvent près des motifs TG(A/G)(A/G)(G/T)(A/C) ou dans les répétitions de mononucléotides et dans d'autres séquences répétitives. Les substitutions de nucléotides sont surreprésentées dans les dinucléotides CpG. L'analyse des relations généalogiques entre les différents variants de PI a permis l'identification de 9 codons homoplasiques qui peuvent être spécialement mutables : 51/52, 101, 115, 148, 256, 342, 353, 362 et 369. Les comparaisons avec les séquences homologues ont montrés que les mutations pathogeniques sont beaucoup plus fréquentes dans les aminoacides plus conservés.

L'histoire évolutive du groupe sanguin FY a été étudiée à travers l'analyse de la distribution du polymorphisme du microsatellite D1S2635, qui a un taux de mutation élevé, parmi les lignées plus stables définies par les allèles FY*A et FY*B du Portugal et par l'allèle FY*O de São Tomé. L'analyse combinée de la variation de séquence dans les régions flanquantes du microsatellite et au niveau des positions polymorphes déjà connues a permis la définition de différents haplotypes dans la région 5' voisine du gène FY. En accord avec les patrons actuels de distribution des allèles FY, la plupart des lignées dérivées sont associées, soit avec l'allèle FY*O, soit avec les allèles non FY*O. Cependant, deux haplotypes dérivés sont partagés par tous les allèles FY, ce qui indique que les lignées FY*O et non FY*O étaient présentes dans la même population ancestrale et qu'elles ont échangées des séquences par conversion génique ou recombinaison, avant la fixation de FY*O en Afrique. L'analyse des

distributions du nombre de répétitions du *locus* D1S2635 parmi les haplotypes associés aux différents allèles FY suggère que les deux principaux haplotypes de FY*O ont pu apparaître par récurrence de la transition T-46C, caractéristique de l'allèle FY*O. L'accumulation de diversité au microsatellite D1S2635 a pu être utilisée pour établir une limite supérieure de l'âge de fixation de FY*O en Afrique qui a été estimée à 14 700 (5 100 - 31 800) années. Cet âge est plus récent que la valeur de 33 000 années proposée antérieurement et place la fixation de FY*O plus proche des altérations écologiques et démographiques associées à la transition du Paléolithique/Néolithique et de la généralisation de la pression sélective du paludisme. Dans le cadre de ces résultats, il est possible que l'augmentation sélective des fréquences de FY*O, a commencé à la fin du Paléolithique dans l'Afrique Occidentale, et a été amplifiée par l'introduction de l'agriculture tropicale et suivie par une plus grande dispersion pendant l'expansion Bantou.

INTRODUÇÃO GERAL

Os recentes progressos tecnológicos registados na área da biologia molecular e o desenvolvimento do Projecto do Genoma Humano conduziram a um aumento sem precedentes da precisão com que a variação genética da nossa espécie pode ser descrita e interpretada.

Hoje em dia, é possível utilizar, com relativa facilidade e um elevado grau de automatização, vários tipos de polimorfismos de DNA altamente informativos, localizados em diferentes regiões do genoma, com diferentes categorias de variação de sequência e distintas taxas de mutação. Estes polimorfismos têm sido fundamentais para a análise de um grande número de questões em áreas de investigação muito diversas, que vão desde as ciências biomédicas, à genética evolutiva e à antropogenética ou à genética forense (Przeworski *et al.*, 2000; Risch, 2000; Roses, 2000; Jorde *et al.*, 2001).

Apesar desta amplitude, podem-se distinguir, no que respeita aos estudos centrados na genética das populações, pelo menos dois grandes grupos de abordagens da diversidade humana (Livingstone, 1991; Bertranpetit, 2000). Na primeira dessas abordagens, procura-se analisar as implicações evolutivas dos actuais padrões de variabilidade usando os diferentes tipos de marcadores genéticos para reconstruir a história das populações humanas. Numa fase inicial, a maior parte destes estudos baseou-se apenas na distribuição das frequências dos alelos de *loci* independentes, sem que houvesse informação disponível sobre as relações genealógicas entre as diferentes formas alélicas (Cavalli-Sforza *et al.*, 1994; Templeton, 1996; Avise, 2000). Mais recentemente, tem-se privilegiado a investigação da distribuição geográfica de haplótipos (ou linhagens) definidos por locais polimórficos adjacentes, cujas relações evolutivas podem ser reconstituídas, de tal modo que é possível incorporar princípios da análise filogenética interespecífica na interpretação da microevolução das populações de uma mesma espécie (Avise, 2000). Este tipo de análise, correntemente designado por filogeografia (Avise, 2000), começou por se alicerçar nos resultados do estudo de locais polimórficos do DNA mitocondrial (mtDNA) e, posteriormente, do cromossoma Y, devido à ausência de recombinação e à relativa simplicidade da inferência haplotípica associada ao modo transmissão uniparental destas regiões do genoma (Jorde *et al.*, 1998). No entanto, o interesse crescente na ressequenciação de segmentos do DNA autossómico tem permitido alargar o âmbito dos estudos

filogeográficos a um número cada vez maior de *loci* situados em diferentes cromossomas (Przeworski *et al.*, 2000; Jorde *et al.*, 2001).

Em conjunto, os resultados obtidos com diferentes tipos de polimorfismos de DNA permitiram esclarecer muitas questões importantes sobre a origem e evolução das populações humanas e, apesar de ainda persistirem diferenças na interpretação das causas dos padrões observados (Relethford, 2001), é inegável que as relações históricas entre várias populações já se encontram relativamente bem estabelecidas.

Um dos princípios fundamentais dos estudos sobre a evolução das populações é o de que as inferências se devem alicerçar na análise combinada do maior número possível de sistemas genéticos, geralmente localizados em regiões não codificantes, de modo a evitar os enviezamentos associados ao comportamento individual de cada *locus* e assegurar que os efeitos da selecção não interferem com o pressuposto de neutralidade selectiva com base no qual as principais conclusões são estabelecidas (Livingstone, 1991; Cavalli-Sforza *et al.*, 1994). Neste contexto, os resultados obtidos podem ser considerados tendências médias em que as propriedades específicas de cada *locus* não têm de ser conhecidas com detalhe.

Pelo contrário, no segundo tipo de abordagem da diversidade humana privilegiam-se as análises *locus* a *locus* e, em vez de se reconstruir a história das populações, procura-se fazer a caracterização pormenorizada de cada gene com o objectivo de elucidar a história natural da variação observada. Por motivos óbvios, a maior parte destes estudos é realizada em regiões funcionalmente relevantes do genoma, em particular nos *loci* cuja variação alélica está associada a doenças hereditárias ou à susceptibilidade a patologias de determinismo multifactorial. Nestes casos, é dada especial ênfase às diferenças funcionais associadas aos vários alelos e à influência da diversidade genética no desenvolvimento de doenças e na heterogeneidade da resposta a riscos ambientais (Bertranpetit e Calafell, 1996; Hill e Motulsky, 1999; Tishkoff e Williams, 2002). As análises da distribuição geográfica da variação alélica são muitas vezes acompanhadas da tentativa de determinar a origem e a idade das diferentes mutações e de identificar os factores que condicionam a sua frequência e dispersão, incluindo a pesquisa de correlações com variáveis ambientais e a ponderação de possíveis efeitos selectivos (Lewontin, 1991; Bertranpetit e Calafell, 1996; Weiss, 1996; Barbujani, 2000). Adicionalmente, em particular nos *loci*

envolvidos em doenças de determinismo simples, há a preocupação de identificar o maior número possível de mutações, a fim de esclarecer os principais mecanismos de mutagénese e caracterizar as relações entre a estrutura e função de cada produto génico (Cooper *et al.*, 1995).

Como é evidente, as duas perspectivas da diversidade genética são complementares. O conhecimento da história das populações é uma condição necessária à interpretação dos padrões de heterogeneidade geográfica observados em cada *locus* individualmente considerado. Sem esse conhecimento, não é possível avaliar se as distribuições dos alelos dos genes estudados são condicionadas por factores histórico-demográficos ou se, pelo contrário, estão dissociadas da tendência média observada e resultam de particularidades dos mecanismos mutacionais ou da acção da selecção (Barbujani, 2000). Por outro lado, a estrutura genética das populações pode influenciar o grau da associação entre variação alélica e patologia. Actualmente, há mesmo um debate importante sobre a utilidade relativa de populações com diferentes histórias evolutivas para a identificação de *loci* associados a doenças de determinismo genético simples ou complexo (Jorde, 2000; Jorde *et al.*, 2001; Tishkoff e Williams, 2002).

Por sua vez, os estudos *locus* a *locus*, apesar de terem um carácter monográfico, podem contribuir para abordar questões de âmbito geral relacionadas com a estrutura e evolução das populações humanas. A análise das propriedades dos espectros mutacionais, por exemplo, permite definir modelos de mutação e estimar parâmetros que são usados para calcular distâncias genéticas e calibrar relógios moleculares (Nei, 1987; Krawczak e Cooper, 1996). No estudo de *loci* com interesse clínico, a elevada intensidade de amostragem num grande número de populações permite documentar a diversidade genética com um detalhe poucas vezes alcançado, o que facilita a análise da dispersão de alelos raros, relativamente recentes, que nem sempre são incluídos em comparações globais da diversidade humana (Bertorelle e Rannala, 1998). Por último, o estudo detalhado de cada *locus* é fundamental para determinar o grau de concordância dos resultados obtidos com vários marcadores genéticos e identificar as causas de possíveis discrepâncias entre reconstruções da história das populações baseadas em diferentes tipos de polimorfismos (Barbujani, 2000; Przeworski *et al.*, 2000; Jorde *et al.*, 2001).

O trabalho desenvolvido nesta tese insere-se no âmbito das análises *locus* a *locus* e pretende contribuir para o melhor entendimento dos actuais padrões de diversidade genética de três *loci* polimórficos: apolipoproteína E (APOE), α1-antripsina (PI) e grupo sanguíneo Duffy (FY). A tese está dividida em três partes que correspondem a cada um dos polimorfismos estudados e incluem resultados de artigos já publicados ou aceites para publicação.

Para além da sua importância específica, estes polimorfismos permitem ilustrar a utilidade do estudo aprofundado de regiões funcionalmente relevantes do genoma para caracterizar o contexto evolutivo da variabilidade genética das populações humanas e elucidar os factores que condicionam a frequência e distribuição geográfica das mutações responsáveis por essa variabilidade. A variação alélica da APOE, determina uma fracção importante dos níveis de colesterol e está associada ao risco de doença de Alzheimer, pelo que este *locus* é um bom exemplo da influência que um polimorfismo pode ter na predisposição a doenças crónicas de determinismo complexo. O gene da PI, com vários alelos normais e alelos patogénicos que causam a deficiência da proteína, é um excelente modelo para estudar a história natural de um polimorfismo e entender os mecanismos envolvidos na origem e dispersão das mutações que provocam doenças hereditárias. O grupo sanguíneo FY está associado à resistência à malária provocada por *Plasmodium vivax* e permite estudar a importância dos agentes infecciosos como factores selectivos que influenciam a distribuição da diversidade genética.

Referências Bibliográficas

Avise, J. C. (2000). *Phylogeography. The history and Formation of Species.* Harvard University Press, Cambridge, Massachusetts.

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

Bertorelle, G., Rannala, B. (1998). Using rare mutations to estimate population divergence times: a maximum likelihood approach. *Proc Natl Acad Sci U S A* **95**:15452-15457.

Bertranpetit, J. (2000). Genome, diversity, and origins: the Y chromosome as a storyteller. *Proc Natl Acad Sci U S A* **97**:6927-6929.

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

Cavalli-Sforza, L. L., Menozzi, P., Piazza, A. (1994). *The history and geography of human genes.* Princeton University Press, Princeton, New Jersey.

Cooper, D. N., Krawczak, M, Antonarakis, S. E. (1995). The Nature and Mechanisms of Human Gene Mutation. In: Sriver, C. R., Beaudet, A. L., Sly, W., Valle, D. (eds). *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill, New York, pp. 259-291.

Hill, A. V. S, Motulsky, A. G. (1999). Genetic variation and human disease: the role of natural selection. In: Stearns, S. C. (ed). *Evolution in Health and Disease.* Oxford University Press, Oxford, pp. 50-61.

Jorde, L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**:1435-1444.

Jorde, L. B., Bamshad, M., Rogers, A. R. (1998). Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays* **20**:126-136.

Jorde, L. B., Watkins, W. S., Bamshad, M. J. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* **10**:2199-2207.

Krawczak, M., Cooper, D. N. (1996). Single base-pair substitutions in pathology and evolution: two sides to the same coin. *Hum Mutat* **8**:23-31.

Lewontin, R. C. (1991). Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? *Genetics* **128**:657-662.

Livingstone, F. B. (1991). Phylogenies and the forces of evolution. *Am J Hum Biol* **3**:83-89.

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

Przeworski, M., Hudson, R. R., Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends Genet* **16**:296-302.

Relethford, J. H. (2001). *Genetics and the Search for Modern Human Origins.* Wiley-Liss, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**:847-856.

Roses, A. D. (2000). Pharmacogenetics and the practice of medicine. *Nature* **405**:857-865.

Templeton, A. R. (1996). Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. *Ciba Found Symp* **197**::259-277.

Tishkoff, S. A., Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet* **3**:611-621.

Weiss, K. M. (1996). Is there a paradigm shift in genetics? Lessons from the study of human diseases. *Mol Phylogenet Evol* **5**:259-265.

PARTE I

Apolipoproteína E

1. Introdução

A apolipoproteína E (APOE) é uma glicoproteína polimórfica que se associa aos quilomicra e às lipoproteínas de tipo VLDL (*very low density lipoproteins*) e HDL (*high density lipoproteins*) e promove a remoção destas partículas do plasma através da ligação a receptores celulares de superfície (Kamboh, 1995; Mahley e Rall, Jr., 2000). A proteína é também o principal veículo de transporte de lípidos no fluído cerebroespinal e intervém ainda nos processos de regeneração do tecido nervoso, na regulação imunológica e na modulação do crescimento e diferenciação celulares (Siest *et al.*, 1995; Mahley e Rall, Jr., 2000).

A APOE constitui, juntamente com outros nove tipos de apolipoproteínas, uma família de moléculas que apresentam semelhanças estruturais assinaláveis e cujos genes terão divergido a partir de um ancestral comum com mais de 680 milhões de anos através de processos de duplicação e translocação (Li *et al.*, 1988). Do ponto de vista funcional, a participação no metabolismo dos lípidos constitui o principal denominador comum das apolipoproteínas, embora se registem diferenças importantes no tipo de partícula lipoproteica a que cada molécula se liga e no modo como cada uma interfere no processo. A unidade funcional destas proteínas é ilustrada pela atenção crescente que vem sendo dedicada à influência da sua variação genética no risco de desenvolvimento de doenças cardiovasculares (Kardia *et al.*, 1999; Mahley e Rall, Jr., 2000).

Embora a maioria dos genes das apolipoproteínas se situe em diferentes regiões do genoma, o *locus* da APOE constitui, juntamente com os *loci* das apolipoproteínas CI (APOCI), CII (APOCII) e CIV (APOCIV) e um pseudogene de APOCI (APOCI'), um agrupamento génico sinténico que ocupa uma região de ~ 44 kb no cromossoma 19q13.2 (Figura. I.1). De entre as apolipoproteínas que se encontram ligadas à APOE, a apolipoproteína CIV (APOCIV) é a que foi descrita mais recentemente e a sua estrutura foi deduzida a partir da identificação do gene correspondente, por não se terem encontrado vestígios de proteína circulante (Jong *et al.*, 1999). É assim improvável que esta proteína tenha um papel relevante no metabolismo dos lípidos (Jong *et al.*, 1999). A apolipoproteína C-II liga-se aos quilomicra e às partículas lipoproteicas VLDL e actua como cofactor da lipase das lipoproteínas (LPL) que hidrolisa os triglicerídeos e liberta os ácidos gordos que são absorvidos pelos tecidos (Bowman, 1992). A apolipoproteína CI liga-se preferencialmente às partículas VLDL

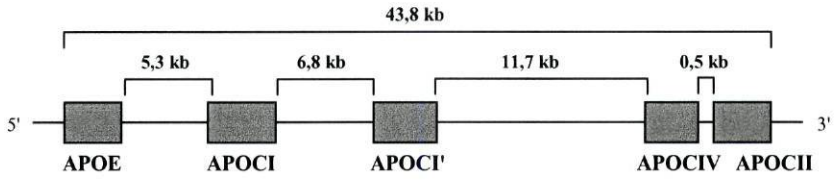


Figura I.1- Agrupamento génico das apolipoproteínas na região cromossómica 19q13.2. As distâncias intergénicas estão de acordo com a descrição de Jong *et al.* (1999), e com a sequência disponível do cromossoma 19 (Genbank Accession nº NT_011109).

e HDL e é um cofactor da acil-transferase de lecitina/colesterol (LCAT) que promove a esterificação do colesterol durante o transporte reverso dos tecidos periféricos para o fígado (Smit *et al.*, 1988; Jong *et al.*, 1999).

O *locus* da APOE apresenta, na maioria das populações humanas, três produtos génicos comuns que diferem na sua composição em arginina (Arg) e cisteína (Cys) nas posições 112 e 158 das respectivas cadeias polipeptídicas (Rall, Jr. *et al.*, 1982b): APOE*2 (Cys112-Cys158), APOE*3 (Cys112-Arg158) e APOE*4 (Arg112-Arg158). Estes alelos têm diferentes capacidades de ligação aos receptores celulares e influenciam de forma significativa a variação interindividual dos níveis de colesterol no plasma. Em média, calcula-se que a variação alélica da APOE determine cerca de 5 a 10% da variação total nos níveis de colesterol em populações de origem europeia (Sing e Davignon, 1985; Boerwinkle *et al.*, 1987; Davignon *et al.*, 1988).

O produto génico APOE*2 tem uma capacidade de ligação correspondente a apenas 2% do normal e a sua ocorrência em homozigotia é uma condição necessária, mas não suficiente, para o desenvolvimento da hiperlipidémia de tipo III, que se caracteriza pela elevação dos níveis de colesterol devido à persistência de partículas VLDL em circulação (Mahley e Rall, Jr., 2000). Como apenas cerca de 10% destes homozigóticos acabam por manifestar a doença, é necessária a presença de outros factores, genéticos e/ou ambientais, para que haja progressão da patologia (Mahley e Rall, Jr., 1995). Na ausência destas contribuições adicionais, o alelo APOE*2 está associado a um decréscimo efectivo da concentração de colesterol total e é o alelo APOE*4 que está ligado a um aumento significativo dos níveis deste factor de risco de

doenças cardiovasculares (Davignon *et al.*, 1988; Kardia *et al.*, 1999; Mahley e Rall, Jr., 2000).

Para além do envolvimento no metabolismo dos lípidos e nas patologias que lhe estão associadas, o polimorfismo da APOE é um dos principais factores genéticos de susceptibilidade às formas comuns de manifestação tardia da doença de Alzheimer (Petersen *et al.*, 1996; Selkoe, 2001). O alelo APOE*4 está associado a um claro aumento do risco de doença e à antecipação dos seus sintomas, e este efeito é mais pronunciado nos homozigóticos que nos heterozigóticos (Corder *et al.*, 1993; Strittmatter *et al.*, 1993). O alelo APOE*2, pelo contrário, parece conferir protecção contra o desenvolvimento da patologia (Corder *et al.*, 1994).

A associação entre a variação alélica na APOE e a doença de Alzheimer, ou as patologias do metabolismo lipídico, tem sido considerada um caso exemplar de predisposição genética às doenças crónicas de determinismo complexo, comuns na espécie humana (Kardia *et al.*, 1999; Martin *et al.*, 2000; Roses, 2000). Por esta razão, a APOE é uma das apolipoproteínas mais bem estudadas e o seu polimorfismo está documentado num grande número de populações. Embora haja uma significativa variação geográfica nas frequências génicas, verifica-se que APOE*3 é o alelo mais comum em todas as populações analisadas e APOE*2 o alelo mais raro (Kamboh, 1995). Em muitos estudos realizados em diferentes populações foi possível replicar a tendência para um aumento dos riscos de patologia associados a APOE*4 e, em determinadas regiões, observaram-se correlações positivas entre as diferenças interpopulacionais na prevalência de algumas patologias e a variação na frequência do alelo. Por exemplo, na Europa observa-se uma associação entre o aumento gradual da frequência de APOE*4 com a latitude e a elevação da prevalência das doenças cardiovasculares nos países nórdicos (Lucotte *et al.*, 1997; Kardia *et al.*, 1999). No entanto, apesar destas correlações, também se verificou que há uma grande variabilidade interindividual e interpopulacional nos riscos correspondentes a cada genótipo e que a intensidade das associações pode variar consideravelmente de população para população (Kardia *et al.*, 1999; Fullerton *et al.*, 2000; Nickerson *et al.*, 2000). Esta variabilidade mostra que a influência do polimorfismo da APOE depende de múltiplos factores ambientais e da composição genética das populações em outros *loci* (Kardia *et al.*, 1999; Fullerton *et al.*, 2000).

Em contraste com a abundância de estudos centrados apenas no *locus* da APOE, a investigação da variação haplotípica associada aos seus alelos mais comuns tem sido reduzida e limitada a um número relativamente pequeno de populações (Fullerton *et al.*, 2000). Este tipo de análise pode, no entanto, revelar-se fundamental para a interpretação da relação entre a variação genética da APOE e diferentes patologias, uma vez que é possível que as associações observadas sejam modificadas pelos níveis de desequilíbrio gamético com mutações de outros *loci* (Templeton, 1995). É o caso, por exemplo, das mutações recentemente descritas em regiões reguladoras da expressão do *locus* da APOE, que podem influenciar significativamente os níveis de síntese dos seus produtos génicos (Mui *et al.*, 1996; Artiga *et al.*, 1998; Bullido *et al.*, 1998; Lambert *et al.*, 1998).

Outra das questões importantes relacionadas com o polimorfismo da APOE, e com a epidemiologia das doenças a ele associadas, consiste na interpretação do significado evolutivo das distribuições de frequências dos seus principais alelos e na identificação dos factores que determinaram a sua origem e dispersão geográfica. Para tal, é fundamental estabelecer as relações filogénicas entre esses alelos e conhecer a respectiva antiguidade, dado que a idade das mutações que lhes deram origem é um parâmetro crucial a ter em conta na ponderação dos processos que podem ter influenciado os actuais padrões de diversidade genética (Clark, 1999; Fullerton *et al.*, 2000; Mahley e Rall, Jr., 2000).

Os alelos APOE*3 e APOE*4 têm sido alternativamente considerados as formas ancestrais do polimorfismo da APOE com base em diferentes critérios. A frequência mais elevada de APOE*3 e a possibilidade de ele originar as variantes APOE*2 e APOE*4 através de um único passo mutacional nos codões 112 e 158, respectivamente, constituem os principais argumentos favoráveis à hipótese de que este alelo seria o mais antigo (Figura I.2 A) (Rall, Jr. *et al.*, 1982a; Mahley e Rall, Jr., 2000). Pelo contrário, a presença de resíduos de Arg nas posições 112 e 158 das proteínas homólogas de Primatas não humanos sugere que, apesar de ocorrer em frequências mais baixas, APOE*4 pode representar a forma alélica primitiva (Figura I.2 B) (Hixson *et al.*, 1988; Hanlon e Rubinsztein, 1995). No entanto, a possibilidade de esta situação se poder dever a uma repartição de linhagens alélicas de um

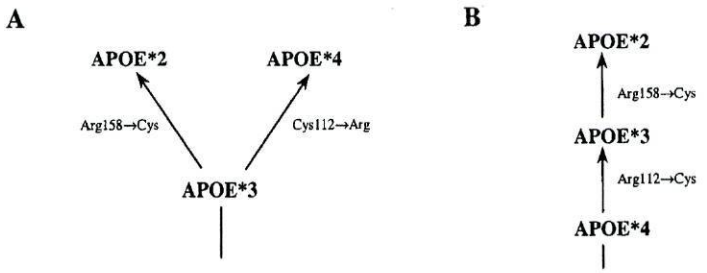


Figura I.2- Hipóteses alternativas sobre a filogenia dos alelos da APOE. **A-** Assumindo APOE*3 como o alelo ancestral. **B-** Assumindo APOE*4 como a forma alélica mais primitiva.

polimorfismo ancestral, anterior à divergência das espécies comparadas, ou à ocorrência de mutações paralelas (Clark, 1997; Hacia *et al.*, 1999), implica que as relações filogenéticas entre os alelos de APOE devam ser confirmadas com outro tipo de evidência.

A identificação da filogenia dos alelos da APOE está intimamente ligada à avaliação do impacto relativo da selecção natural e dos factores histórico-demográficos na sua actual distribuição geográfica. Em princípio, considera-se que se APOE*3 for o alelo mais antigo, a distribuição global das frequências génicas pode ter resultado inteiramente de factores demográficos, uma vez que em condições de neutralidade o alelo mais comum é, normalmente, o ancestral (Watterson e Guess, 1977; Clark, 1997; Hacia *et al.*, 1999). Inversamente, a hipótese de que APOE*4 é o alelo mais antigo tem sido associada à necessidade de invocar mecanismos selectivos que justifiquem o aumento da frequência de APOE*3 e a diminuição da frequência de APOE*4 (Fullerton *et al.*, 2000; Mahley e Rall, Jr., 2000). A razoabilidade da hipótese selectiva depende, contudo, da idade absoluta das mutações características de cada alelo e do intervalo de tempo que decorreu entre a sua origem. Por exemplo, se APOE*4 for suficientemente antigo, é natural que a sua frequência esteja a decrescer a partir de um estado anterior em que o alelo estava fixado sem que seja necessário invocar a acção da selecção (Kimura, 1983).

No trabalho que agora se apresenta, analisa-se a evolução do polimorfismo da APOE através da caracterização da sua diversidade haplotípica em diferentes populações e da datação dos seus alelos com base nos níveis de desequilíbrio gamético

que lhes estão associados. Os resultados iniciais desta abordagem foram publicados no artigo 1, que é apresentado após uma discussão mais alargada que integra novos dados obtidos posteriormente.

2. Resultados e Discussão

2.1 Variação alélica no *locus* APOE

A variação alélica no *locus* APOE foi estudada em indivíduos não aparentados provenientes das populações de São Tomé e Príncipe, Portugal e País Basco. Na população de São Tomé e Príncipe, foram analisados dois conjuntos independentes de amostras. O primeiro conjunto engloba cerca de 60% de indivíduos residentes na Cidade de São Tomé (capital do país) e 40% de residentes em localidades situadas no norte da ilha de São Tomé (Figura I.3). Este conjunto, designado por "São Tomé-Norte", foi estudado previamente no artigo 1. O segundo conjunto é constituído por indivíduos provenientes de diferentes comunidades situadas na costa leste da ilha de São Tomé, incluindo a cidade capital (Figura. I.3). Em Portugal, a principal amostra é constituída por indivíduos nascidos na área urbana do distrito do Porto, tendo-se incluído, adicionalmente, os resultados obtidos na localidade rural de Ribeira de Pena (distrito de Vila Real), que foi analisada no âmbito de uma colaboração num estudo longitudinal de avaliação do perfil lipídico daquela

Figura I.3- Locais de amostragem na ilha de São Tomé. À área a sombreado corresponde o conjunto designado por São Tomé-Norte. Os pontos a vermelho indicam as diferentes comunidades da costa leste da ilha.

população. A amostra do País Basco é constituída por indivíduos residentes nas províncias da Biscaia e Guipuzcoa com oito ancestrais de apelido Basco e quatro avós nascidos em território Basco.

As distribuições de frequências no distrito de Porto e no País Basco (Tabela I.1) são semelhantes às encontradas noutras amostras do sul da Europa e enquadram-se no gradiente Norte-Sul de decréscimo das frequências de APOE*4 e aumento das frequências de APOE*3 previamente observado neste continente (Figura I.4) (Lucotte *et al.*, 1997). No entanto, verificou-se que na amostra do País Basco não há conformidade com o formalismo de Hardy-Weinberg devido a um défice de heterozigóticos, o que sugere a possível ocorrência de subestruturação no seio desta população ou a inadequação do esquema amostral que foi seguido. Na amostra rural de Ribeira de Pena observou-se uma distribuição significativamente diferente das duas anteriores, com um aumento das frequências de APOE*2 e APOE*4. Esta diferença indica que, mesmo no seio de sociedades industrializadas, é ainda possível detectar vestígios de microdiferenciação em núcleos populacionais mais isolados e periféricos.

Tabela I.1- Distribuição das frequências dos alelos de APOE nas duas amostras Portuguesas e no País Basco.

APOE	Portugal (N^a =149)	Ribeira de Pena (N=55)	P. Basco (N=92)
*2	0,044±0,012	0,091±0,027	0,038±0,014
*3	0,882 ±0,019	0,791±0,039	0,918±0,020
*4	0,074±0,015	0,118±0,031	0,043±0,015
HW[b]	p=0,424±0,001	p=0,796±0,002	p=0,002±0,000

[a] Número de indivíduos.
[b] Probabilidades correspondentes ao teste exacto do equilíbrio de Hardy-Weinberg disponibilizado no programa ARLEQUIN (Schneider *et al.*, 1997).

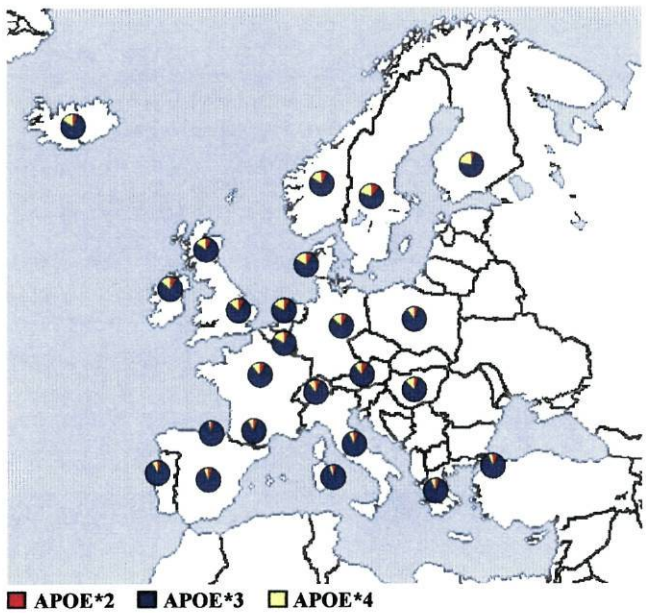


Figura I.4 - Distribuição das frequências dos alelos de APOE no continente Europeu de acordo com o presente trabalho e Lucotte *et al.* (1997).

Na ilha de São Tomé (Tabela I.2) observou-se uma notável variação nas frequências de APOE calculadas em diferentes localidades, com 40% (6/15) das comparações entre amostras a revelarem valores significativos de diferenciação (Tabela I.3). As frequências observadas na região Norte ("São Tomé-Norte" e Trindade) enquadram-se no intervalo de variação definido pela maioria das populações do continente africano onde, apesar das diferenças interétnicas, a frequência média de APOE*4 é superior à registada na Europa (Zekraoui *et al.*, 1997) (Figura I.5). Nas amostras do sul da ilha (São João dos Angolares e Ribeira Afonso), a frequência média de APOE*4 (0,443) é superior em cerca de 70% à encontrada no Norte (0,255) e representa um dos valores mais elevados alguma vez registados nas populações até agora estudadas (Sandholzer *et al.*, 1995, Zekraoui *et al.*, 1997). Em África, apenas a amostra de Pigmeus estudada por Zekraoui *et al.* (1997), e a população de Khoi-San analisada por Sandholzer *et al*, (1995) exibem valores semelhantes (0,41 e 0,37, respectivamente). As amostras da Cidade de São Tomé e de Santana apresentam frequências intermédias, a que também correspondem valores intermédios de probabilidade de diferenciação (Tabelas I.2 e I.3).

Tabela I.2- Distribuição das frequências dos alelos de APOE nas diferentes amostras recolhidas na ilha de São Tomé.

APOE	S. Tomé-Norte (N^a =165)	Trindade (N=63)	Cidade de S. Tomé (N=60)	Santana (N=56)	Ribeira Afonso (N=56)	S. João .Angolares (N=86)
*2	0,100±0,017	0,071±0,023	0,158±0,033	0,152±0,034	0,107±0,029	0,145±0,027
*3	0,652 ±0,026	0,667±0,042	0,533±0,046	0,518±0,047	0,420±0,047	0,442±0,038
*4	0,248±0,024	0,262±0,039	0,308±0,042	0,330±0,044	0,473±0,047	0,413±0,036
HW[b]	p=0,449±0,001	p=0,942±0,001	p=0,639±0,001	p=0,135±0,001	p=0,467±0,002	p=0,485±0,001

[a] Número de indivíduos.
[b] Probabilidades correspondentes ao teste exacto do equilíbrio de Hardy-Weinberg disponibilizado no programa ARLEQUIN (Schneider *et al.*, 1997).

Tabela I.3- Distribuição dos valores de probabilidade do teste exacto de diferenciação interpopulacional baseado nas frequências dos alelos de APOE, calculadas em diferentes localidades da ilha de São Tomé[a].

	S. Tomé Norte	Trindade	Cid de S. Tomé	Santana	Ribeira Afonso
Trindade	0,887±0,001				
Cid. S. Tomé	**0,044±0,010**[b]	0,207±0,012			
Santana	**0,035±0,007**	0,186±0,012	0,800±0,010		
R. Afonso	**0,000**	**0,002±0,001**	0,133±0,019	0,094±0,010	
S. J. Angolares	**0,000**	**0,001±0,001**	0,273±0,033	0,109±0,010	0,930±0,009

[a] Teste exacto de diferenciação populacional disponibilizado no programa ARLEQUIN (Schneider *et al.*, 1997).
[b] Valores significativos em destaque.

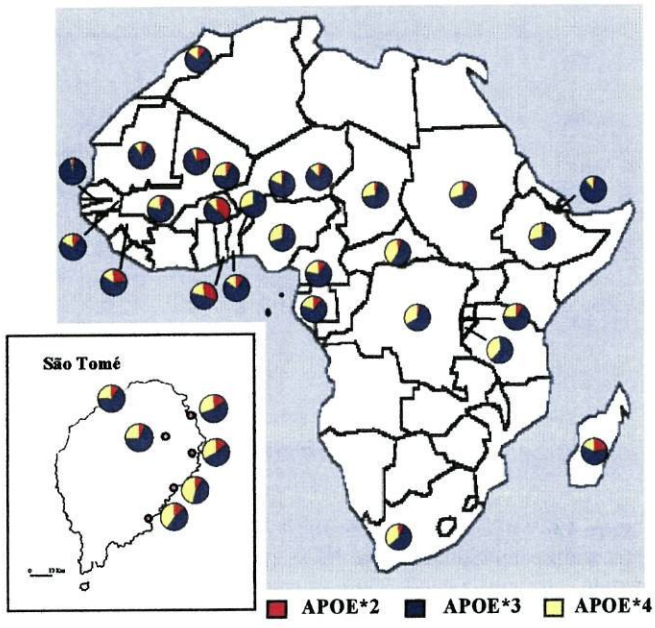


Figura I.5- Distribuição das frequências dos alelos de APOE no continente Africano de acordo com o presente trabalho e Zekraoui *et al.* (1997).

A sobreposição entre as diferenças de frequência de APOE e a posição geográfica relativa das amostras de São Tomé (Figura I.3) é ilustrada no ordenamento produzido na Figura I.6, em que é evidente a presença de dois grupos extremos, constituídos pelas amostras do Norte e do Sul, e de um grupo intermédio que inclui a amostra de Santana e a população mais cosmopolita da Cidade de São Tomé.

Esta microdiferenciação pronunciada ao longo de um percurso de apenas 50 Km da costa oriental da Ilha de São Tomé poderá ser explicada pelas características do povoamento de São Tomé durante a sua colonização com escravos trazidos das costas africanas adjacentes a partir do início do século XVI. Durante este processo, e até ao século XIX, as regiões do sudeste da ilha não foram afectadas pela instalação de plantações de cana do açúcar e foram povoadas por comunidades de africanos livres, mais tarde designados por *angolares* (Tenreiro, 1961). Segundo a tradição oral, as comunidades angolares, onde se fala um tipo de crioulo mais influenciado por dialectos Bantu do que no resto da ilha (Hagemeijer, 1999), teriam sido fundadas por sobreviventes do naufrágio de um navio negreiro proveniente de Angola, em meados

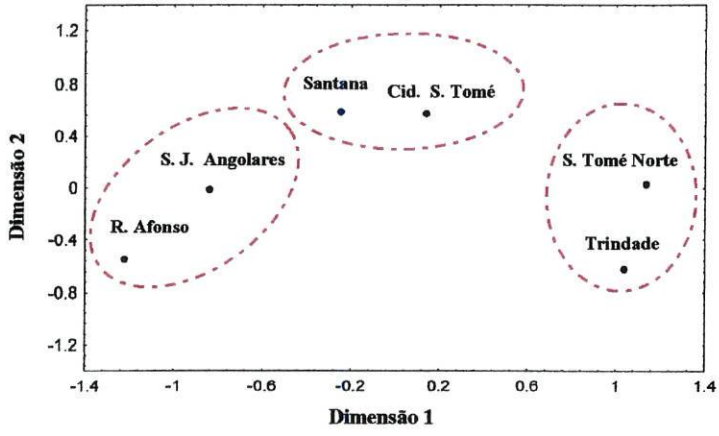


Figura I.6- Ordenação bidimensional das seis amostras da ilha de São Tomé após análise multidimensional MDS (*multidimensional scaling*), utilizando a medida de distância genética F_{ST} calculada com base nas frequências de APOE. As distâncias genéticas foram calculadas com o programa ARLEQUIN (Schneider *et al*, 1997) e a análise multidimensional efectuada com o programa STATISTICA (http://www.statsoftinc.com).

do século XVI (Tenreiro, 1961; Caldeira, 1999). De acordo com uma versão alternativa, preferida pelos historiadores, os angolares seriam descendentes de escravos fugitivos, que teriam abandonado as plantações e se teriam instalado em aldeias autónomas, semelhantes aos *quilombos* ou *mocambos* mais tarde encontrados no Brasil (Caldeira, 1999). Em qualquer dos casos estas populações ter-se-iam mantido numa situação periférica, afastadas do contacto com as regiões mais densamente povoadas do norte e nordeste da ilha de São Tomé. A diferenciação genética da população *angolar* foi recentemente notada através da análise de microssatélites do cromossoma Y em indivíduos que se auto-identificaram com este grupo etno-linguístico numa amostra não estratificada geograficamente (Trovoada *et al.*, 2001). Os resultados agora obtidos com o polimorfismo da APOE, apesar de limitados a um único sistema genético, sugerem que esta diferenciação terá promovido uma marcada divisão entre o Norte e o Sul da ilha, posteriormente atenuada pelo fluxo génico que terá originado o gradiente Norte-Sul agora observado. A análise de um maior número de marcadores de diferentes tipos, em amostras seleccionadas segundo critérios geográficos, permitirá avaliar melhor a extensão da diferenciação observada e

verificar se ela resulta de uma diversificação *in situ* ou da ocupação de diferentes áreas da ilha por escravos provenientes de distintas regiões do continente africano.

2.2 Diversidade haplotípica no *locus* APOE

2.2.1 *Diferenças interpopulacionais*

A diversidade haplotípica foi avaliada através da análise da distribuição conjunta dos fenótipos da APOE e de um polimorfismo de restrição que resulta da inserção/delecção do tetranucleotídeo CGTT na região promotora do gene da APOCI, localizada a 317 bp do local de início da transcrição (Figura. I.1) (Smit *et al.*, 1988; Xu *et al.*, 1999). Os pormenores dos métodos de fenotipagem destes polimorfismos encontram-se no artigo 1. Nas amostras de Trindade, Cidade de São Tomé e Santana, não foi possível realizar a análise de APOCI. A amostra do País Basco foi excluída da análise conjunta devido à ausência de verificação do equilíbrio de Hardy-Weinberg nos *loci* APOE e APOC1.

A comparação das medidas de desequilíbrio gamético nas duas amostras portuguesas (Tabela I.4) mostra que a diferenciação observada na distribuição das frequências génicas de APOE em Ribeira de Pena é acompanhada de um aumento da intensidade do desequilíbrio gamético que conduz a uma associação completa entre os alelos APOE*4/APOCI*2, APOE*3/APOCI*1 e APOE*2/APOCI*2. Este resultado sugere que na origem da referida diferenciação poderão ter estado efeitos de deriva genética que homogeneizaram a variação haplotípica pré-existente.

Uma análise semelhante nas amostras de São Tomé não mostra uma tendência tão marcada para a correlação entre a diferenciação das comunidades do Norte e do Sul da ilha e a variação nos níveis de associação entre alelos dos *loci* APOE e APOCI. Em nenhuma amostra há um aumento sistemático de todos os valores de δ^* (ou D'). Quando as comunidades de São João dos Angolares e Ribeira Afonso são comparadas com a de "São Tomé-Norte", que se encontra no pólo oposto de diferenciação das frequências de APOE, apenas em São João dos Angolares se registam maiores valores de desequilíbrio gamético, com associações mais acentuadas entre os alelos APOE*3/APOCI*1 e APOE*2/APOCI*2 (Tabela I.4). Por outro lado, a amostra de

Tabela I.4- Frequências haplotípicas e intensidade do desequilíbrio gamético entre os *loci* APOE e APOC1 nas populações portuguesas e são-tomenses.

		Frequências		Medidas de desequilíbrio gamético		
		APOC1				
Populações	APOE	*1	*2	p[a]	D' [b]	δ* [c]
Portugal	*2	0,000	0,044±0,012	0,000	-1,00	1
	*3	0,865±0,022	0,017±0,007	0,000	0,865	0,841
	*4	0,014±0,006	0,060±0,013	0,000	-0,785	0,785
Ribeira de	*2	0,000	0,091±0,028	0,000	-1	1
Pena	*3	0,791 ±0,039	0,000	0,000	1	1
	*4	0,000	0,118±0,030	0,000	-1	1
São Tomé	*2	0,012±0,006	0,088±0,016	0,000	-0,825	0,825
Norte	*3	0,521±0,033	0,131±0,022	0,000	0,362	0,362
	*4	0,150±0,022	0,098±0,021	0,186±0,001	-0,117	0,117
Ribeira	*2	0,010±0,010	0,097±0,029	0,000	-0,869	0,869
Afonso	*3	0,343±0,044	0,077±0,029	0,116±0,001	0,359	0,359
	*4	0,361±0,047	0,112±0,036	0,392±0,002	0,172	0
São João dos	*2	0,000	0,145±0,026	0,000	-1	1
Angolares	*3	0,413±0,037	0,029±0,016	0,000	0,795	0,795
	*4	0,268 ±0,034	0,145±0,028	0,770±0,002	-0,068	0,046

[a] Probabilidades correspondentes ao teste de desequilíbrio gamético disponibilizado no programa ARLEQUIN (Schneider *et al.*, 1997).
[b] D' coeficiente de desequilíbrio gamético segundo Lewontin (1964) obtido considerando a associação dos alelos de APOE ao alelo marcador APOCI*1.
[c] Fracção de cromossomas que mantêm o arranjo haplotípico ancestral (Bengtsson e Thomson, 1981; Devlin e Risch, 1995).

Ribeira Afonso, é única em que a associação entre os alelos de APOC1 e APOE*3 se revela não significativa.

A comparação dos padrões de desequilíbrio gamético observados em Portugal e São Tomé mostra que os valores de D' e δ* são consistentemente mais baixos nas amostras africanas, onde os níveis de associação entre APOE*4 e os alelos de APOCI são mesmo não-significativos, apesar da curta distância de 5,3 kb que separa os dois *loci* (Tabela I.4). Também no que se refere às medidas sintéticas da diversidade genética se observam diferenças entre os dois grupos de amostras, com a população de São Tomé a registar os valores mais elevados de heterozigotia, quer em cada *locus* considerado individualmente, quer no conjunto dos dois *loci* estudados (Tabela I.5).

Tabela I.5- Valores de heterozigotia esperada calculados com base nas frequências génicas e haplotípicas dos *loci* APOE e APOCI nas diferentes amostras estudadas.

	Heterozigotia esperada		
Populações	APOE	APOC1	APOE-APOC1
Portugal	0,215±0,031	0,213±0,029	0,247±0,032
Ribeira de Pena	0,355±0,053	0,334±0,045	0,355±0,056
São Tomé-Norte	0,505±0,024	0,433±0,019	0,673±0,022
Trindade	0,485±0,039	na	na
Cidade de S, Tomé	0,601±0,027	na	na
Santana	0,605±0,026	na	na
Ribeira Afonso	0,594±0,021	0,411±0,036	0,731±0,023
S, João Angolares	0,617±0,016	0,438±0,026	0,719±0,019

na - não analisado.

A diferença entre a diversidade haplotípica das amostras portuguesas e de São Tomé sugere a existência de padrões de desequilíbrio gamético muito distintos em amostras de origem europeia e africana nos *loci* da APOE e da APOCI. O mesmo tipo de diferença, inicialmente observado no artigo 1, foi confirmado em pelo menos um estudo independente, em que se detectaram níveis menos elevados de desequilíbrio em afro-americanos (Xu *et al.*, 1999). Estes resultados mostram que a tendência registada noutros marcadores para a ocorrência de maior diversidade haplotípica em populações de origem africana (Tishkoff *et al.*, 1996; Reich *et al.*, 2001, Relethford, 2001) também é captada pelo sistema formado pelos *loci* APOE/APOCI.

Esta observação tem consequências adicionais importantes ao nível da interpretação dos estudos de associação entre a variação alélica da APOE e diferentes patologias. Como se referiu anteriormente, se estas associações não resultarem de um efeito directo dos alelos da APOE e puderem ser modificadas pela ligação a outras mutações num mesmo haplótipo, é provável que a variação interpopulacional na sua intensidade se possa dever à heterogeneidade dos níveis de desequilíbrio gamético. Por exemplo, a incapacidade para detectar, em amostras de origem africana, a forte associação entre o alelo APOE*4 e a doença de Alzheimer (Tang *et al.*, 1996; Sahota *et al.*, 1997) pode resultar da quebra de desequilíbrio gamético naquelas populações. O mesmo se aplica à variabilidade das associações entre o polimorfismo da APOE e os níveis lipídicos encontradas em diferentes populações (Xu *et al.*, 1999).

2.2.2 História evolutiva do polimorfismo da APOE

A comparação da variação haplotípica nas diferentes amostras indica que em todos os casos em que se registam valores significativos de desequilíbrio gamético, os alelos APOE*4 e APOE*2 encontram-se associados a APOC1*2 (valores de D' negativos), enquanto o alelo APOE*3 está sempre associado a APOC1*1 (Tabela I.4). Este padrão está de acordo com resultados de outros estudos efectuados em amostras de indivíduos Afro-Americanos (Xu *et al.*, 1999), europeus (Klasen *et al.*, 1987; Chartier-Harlin *et al.*, 1994; Poduslo *et al.*, 1995) e japoneses (Kamino *et al.*, 1996) e sugere que APOE*4-APOCI*2, APOE*3-APOCI*1 e APOE*2-APOCI*2 representam os haplótipos ancestrais originados pelas mutações que criaram a variação alélica de APOE.

Por outro lado, a análise dos níveis de desequilíbrio gamético associados aos diferentes alelos de APOE mostra que, com excepção da amostra portuguesa de Ribeira de Pena, os alelos APOE*4 e APOE*2 apresentam, em todas as populações, os níveis mais baixos e mais altos de desequilíbrio gamético, respectivamente (Tabela I.4). Tendo em conta que a ruptura das associações ancestrais entre alelos dos *loci* APOE e APOCI, promovida pela recombinação, depende do tempo decorrido desde a sua origem, estes resultados indicam que os alelos de APOE foram gerados pela ordem 4→3→2, através de substituições sequenciais Arg→Cys nos codões 112 e 158 (Figura I.2 B).

A quantificação dos níveis de desequilíbrio gamético também pode ser usada para tentar realizar a datação absoluta dos alelos da APOE. Se os alelos APOE*2, APOE*3 e APOE*4 forem tratados como populações sujeitas a fluxos migratórios bidireccionais que em cada geração promovem a troca de alelos da APOCI com a população geral devido à recombinação, a frequência de qualquer alelo de APOCI no seio de cada alelo da APOE pode-se expressar por (Hartl e Clark, 1989):

$$p_g = (1-\theta)^g p_0 + \theta^g p \approx e^{-\theta g} p_0 + \left(1 - e^{-\theta g}\right) p \qquad \text{I.1}$$

em que p_0 é a frequência de um dado alelo de APOCI nos cromossomas portadores do alelo de APOE que se pretende datar, calculada na geração inicial em que este surgiu por mutação; p_g corresponde a essa frequência na geração g; p é a frequência do mesmo alelo de APOCI na população geral; θ é a fracção de recombinação entre APOE e APOCI; g é número de gerações decorrido desde a origem do alelo de APOE que se pretende datar.

Assumindo que na geração inicial (g=0) há desequilíbrio gamético total entre APOE e APOCI (p_0=1) e que as frequências dos alelos de APOCI (p) não sofrem alteração ao longo do tempo na população geral, a resolução da expressão I.1 em ordem a g dá:

$$g \approx \left\{-\ln\left[(p_g - p) \div (1-p)\right]\right\} \div \theta \qquad \text{I.2}$$

Esta expressão é equivalente à do método dos momentos normalmente utilizado na avaliação da idade de alelos patogénicos raros com base nos valores de desequilíbrio gamético, assumindo a ausência de selecção ou deriva genética (Risch *et al.*, 1995; Jorde, 2000; Slatkin e Rannala, 2000). A quantidade $(p_g\text{-}p)\div(1\text{-}p)$, usualmente designada por δ^*, corresponde à fracção de cromossomas que mantêm o arranjo haplotípico ancestral e já foi usada para avaliar a intensidade do desequilíbrio gamético na tabela I.4. O intervalo de confiança mínimo de 95% para as estimativas de g, pode ser obtido através da avaliação do modo como a incerteza no cálculo de p_g se reflecte nas idades calculadas para os alelos de APOE. Esta avaliação foi feita através do cálculo de δ^* com valores extremos de p_g obtidos através de $p_g \pm 2 \times \sqrt{\{[p_g (1- p_g)]/n_g\}}$ (Goldstein *et al.*, 1999), em que n_g é o número de cromossomas com o alelo de APOE que se pretende datar.

Na Figura I.7 exemplifica-se graficamente a evolução esperada das frequências do alelo marcador APOCI*1 no seio dos alelos da APOE de acordo com o modelo de datação utilizado. Na Tabela I.6 apresentam-se os resultados da datação absoluta dos alelos de APOE obtida com a expressão I.2.

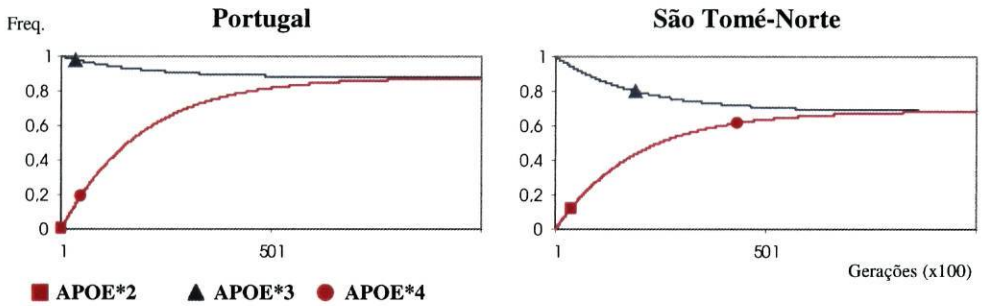


Figura I.7- Exemplos da evolução esperada da frequência do alelo marcador APOCI*1 no seio dos alelos de APOE de acordo com o modelo utilizado na datação dos alelos. A curva azul representa a evolução de APOCI*1 no seio dos alelos APOE*3 e a curva vermelha a evolução de APOCI*2 no seio dos alelos APOE*2 e APOE*4. Os símbolos indicam a posição correspondente à idade dos alelos de APOE nas respectivas curvas teóricas.

Tabela I.6- Estimativas da idade dos alelos da APOE, em número de gerações, obtidas com a expressão I.2 assumindo uma fracção de recombinação $\theta = 0{,}000053$ (1cM=1000kb) entre o *locus* da APOE e o *locus* marcador APOCI.

APOE	Portugal	S.Tomé Norte	R. Afonso	S J Angolares	R. Afonso + Angolares
*2	0	3640 (191-7860)	2660 (0-8620)	0	857 (0-2550)
*3	3490 (366-6720)	19170 (11780-31500)	≥43450	4340 (528-9110)	8270 (3560-14580)
*4	4600 (483-9810)	≥43450	≥43450	≥43450	≥43450

Os intervalo de confiança a 95% são dados entre parênteses.

A ausência de desequilíbrio gamético entre APOE*4 e o *locus* APOCI nas amostras de São Tomé apenas permitiu realizar uma estimativa mínima do número de gerações decorrido desde a origem deste alelo. Esta estimativa foi feita através do cálculo do tempo necessário para desfazer 90% do desequilíbrio inicial, obtendo-se um valor de ~ 43 450 gerações para a idade mínima de APOE*4 (Tabela I.6), correspondente a 1 300 500 anos, se cada geração corresponder a 30 anos (Tremblay e Vezina, 2000). Este valor é da mesma ordem de grandeza das datações da coalescência da diversidade alélica realizadas noutros *loci* autossómicos (Labuda *et al.*, 2000; Templeton, 2002) e é muito próximo, por exemplo, de uma estimativa de ~40 000 gerações obtida para a idade da linhagem mais antiga do gene da β-globina (Harding *et al.*, 1997).

As idades mais elevadas de APOE*3 também foram obtidas nas amostras de São Tomé, embora as estimativas reflictam a maior diversidade dos valores de desequilíbrio gamético das diferentes localidades. Na amostra de "São Tomé-Norte" o valor encontrado de 19 170 (11 780-31 500) gerações corresponde a uma idade de 575 100 anos (Tabela I.6). Nas restantes amostras obtiveram-se valores mais extremos que vão de ~ 43 450 gerações, correspondentes à ausência de desequilíbrio gamético na amostra de Ribeira Afonso, a 4 340 (528-9 110) gerações calculados com os dados de São João dos Angolares (Tabela I.6). Estas duas amostras têm contudo um efectivo reduzido, e a sua combinação, após confirmação da inexistência de diferenças significativas nas respectivas distribuições haplotípicas, conduz a uma estimativa de 8 270 (3 560-14 580) gerações (Tabela I.6).

A datação mais antiga de APOE*2, também obtida em São Tomé, sugere que este alelo se terá originado há cerca de 3 640 (191-7 860) gerações ou 109 200 anos (Tabela I.6).

Tal como seria de esperar, na amostra do distrito do Porto, que é a amostra portuguesa com maior diversidade haplotípica, as datações dos alelos da APOE correspondem sempre a valores mais recentes do que as estimativas equivalentes obtidas em São Tomé (Tabela I.6). Por outro lado os intervalos de variação das estimativas de idade de APOE*3 e APOE*4 encontram-se sobrepostos, indicando que ao aumento do desequilíbrio gamético nesta amostra correspondeu um esbatimento das diferenças na intensidade das associações haplotípicas dos alelos de APOE.

Apesar de as datações absolutas aqui tentadas assentarem em pressupostos demasiado simplificadores, as idades mais antigas calculadas para os três alelos comuns de APOE, obtidas em São Tomé, são inteiramente compatíveis com actual distribuição geográfica desses alelos e com a sua ocorrência na grande maioria das populações humanas. A idade mínima de ~ 43 450 gerações (1 300 500 anos) calculada para APOE*4 é da mesma ordem de grandeza dos tempos de coalescência obtidos noutros sistemas autossómicos, e corresponde a aproximadamente o quádruplo das idades dos ancestrais comuns das linhagens dos cromossomas Y e mitocondrial, como seria esperar num alelo neutro (Templeton, 2002). Por outro lado, a idade máxima de 109 200 anos obtida para o alelo mais recente APOE*2 indica que ele poderá ter sido originado pouco tempo antes da expansão das populações humanas para fora de África há cerca de 100 000 anos, segundo o modelo unirregional de origem do Homem Moderno (Relethford, 2001).

A diferença observada nas idades dos alelos de APOE nas amostras portuguesas e são-tomenses também é compatível com as previsões do modelo unirregional e pode ser atribuída ao estrangulamento de efectivo populacional que hipoteticamente terá estado associado à dispersão do Homem Moderno de África para as regiões euro-asiáticas. Assumindo que a obtenção de estimativas de idade sistematicamente mais recentes na população portuguesa resulta desse efeito de fundador, e que o processo conduziu ao restabelecimento de todas as associações haplotípicas ancestrais, a data mais antiga associada a APOE*4 em Portugal indica que esse efeito poderá ter ocorrido há ~ 4 600 gerações, ou seja há 138 000 anos (Tabela I.6). Esta datação coincide notavelmente com estimativas obtidas noutros sistemas com outras metodologias (Tishkoff *et al.*, 1996; Relethford, 2001). Futuramente, será necessário alargar a proveniência geográfica das amostras, a fim de garantir que o padrão agora observado não resulta de peculiaridades da história demográfica das populações estudadas. Por outro lado, deve ter-se em conta que a presença de níveis menos elevados de desequilíbrio gamético nas populações africanas pode não implicar uma relação de ancestralidade com outras populações, mas apenas uma menor intensidade de deriva genética resultante de um maior efectivo populacional (Relethford, 2001).

Recentemente, Fullerton *et al.* (2000), ressequenciaram o gene da APOE em 96 indivíduos de quatro populações e identificaram 23 polimorfismos de sequência

distribuídos por 31 haplótipos. Aplicando a estes polimorfismos intragénicos um método de máxima verosimilhança baseado na teoria da coalescência, estes autores obtiveram idades dos alelos APOE*2 e APOE*3 que não são significativamente diferentes das aqui obtidas (Tabela I.7), embora o acordo verificado se deva, em parte, à grande amplitude dos intervalos de confiança associados a ambos os tipos de estimativa. Relativamente a APOE*4, os métodos de máxima verosimilhança conduziram a uma datação bastante mais recente, correspondente a um máximo de 28 950 gerações ou 868 500 anos (Tabela I.7).

É difícil comentar esta discrepância entre cálculos efectuados com metodologias e pressupostos muito diferentes. Em princípio, ao maior número de polimorfismos usado por Fullerton *et al.* (2000), deverá corresponder uma estimativa com maior precisão. No entanto, o facto de nessa estimativa não terem sido usadas populações Africanas, mas apenas uma amostra de negros americanos, provavelmente miscigenada, poderá ter conduzido a uma subavaliação da diversidade haplotípica e do tempo necessário para acumulá-la. Em qualquer caso, os dois grupos de estimativas mostram que os alelos da APOE foram gerados pela ordem 4→3→2, que entre a sua origem decorreram consideráveis intervalos de tempo e que todos se originaram antes da divergência das populações humanas postulada pelo modelo unirregional.

Tabela I.7- Comparação das estimativas médias da idade dos alelos da APOE, em número de gerações, obtidas no presente trabalho com as publicadas por Fullerton *et al.* (2000).

	Idade dos alelos	
	Este Trabalho	Fullerton *et al.* (2000)
APOE		
*2	3640	4500
	(191-7860)	
*3	19170	11000
	(11780-31500)	(6100-22000)
*4	≥43450	15550
		(8800-28950)

Os intervalo de confiança a 95% são dados entre parênteses.

Como já se referiu, o reconhecimento de que APOE*4 é o alelo mais antigo está normalmente associado à hipótese de que só com a acção favorável da selecção se poderiam ter atingido as elevadas frequências do alelo APOE*3 que se observam na maioria das populações humanas. No entanto, o facto de as patologias associadas a APOE*4 até agora descritas se manifestarem depois da idade reprodutiva, dificulta a aceitação de uma hipótese selectiva consensual.

Fullerton *et al.* (2000), propuseram que o aumento da frequência de APOE*3 pode ter resultado de uma vantagem selectiva com base na observação de níveis de diversidade intraespecífica mais baixos do que seria de esperar a partir da elevada divergência interespecífica registada na região vizinha da mutação que originou o alelo.

Uma forma alternativa de testar a acção da selecção, consiste em comparar as datações obtidas com a frequência de uma mutação e com a diversidade haplotípica acumulada desde a sua origem. O princípio desta abordagem é o de que a selecção pode levar a que mutações muito recentes atinjam frequências elevadas sem terem tido tempo para acumular diversidade haplotípica (Slatkin e Rannala, 2000; Rannala e Bertorelle, 2001).

No caso da APOE, pode utilizar-se a teoria de Kimura e Ohta (1973), para calcular o número de gerações necessário para que a substituição Arg112Cys, que separa APOE*4 dos restantes alelos, atinja a sua actual frequência assumindo neutralidade selectiva. Usando a expressão $-4Ne[p(\ln p)/(1-p)]$, em que p é a frequência da mutação que se pretende datar e Ne o tamanho efectivo da população e considerando o um valor de p=0,75, correspondente à frequência de Cys na posição 112 (APOE*3+APOE*2) na amostra com a data mais antiga deduzida pela variação diversidade haplotípica ("São Tomé- Norte"; Tabelas I.6 e I.7), e assumindo que Ne=10 000 (Harpending *et al.*, 1998; Zietkiewicz *et al.*, 1998), obtém-se uma datação de ~ 34 500 gerações, com um intervalo de 10 800 a 39 000 gerações, calculado segundo Slatkin e Rannala (2000). Embora a estimativa anterior de 19 170 (11 780-31 500) gerações, baseada na diversidade haplotípica de APOE*3, sugira que a mutação Arg112Cys é mais recente do que seria necessário para, em condições de neutralidade selectiva, alcançar a sua presente frequência, verifica-se uma sobreposição dos intervalos de variação dos dois tipos de datação. Esta sobreposição sugere que poderá

ser prematuro abandonar a hipótese de neutralidade selectiva do polimorfismo da APOE, pelo menos com base na evidência reunida pela datação dos alelos. No entanto, a grande amplitude dos intervalos de confiança observados, bem como a introdução de simplificações excessivas nos cálculos efectuados neste exercício, obrigam a um estudo mais aprofundado, envolvendo mais amostras provenientes de diferentes regiões e a simulação de vários cenários demográficos alternativos.

Artigo 1

Seixas, S., Trovoada, M. J., Rocha, J. (1999). Haplotype analysis of the apolipoprotein E and apolipoprotein C1 loci in Portugal and São Tomé e Príncipe (Gulf of Guinea): linkage disequilibrium evidence that APOE*4 is the ancestral APOE allele. *Hum Biol* **71**:1001-1008.

*Haplotype Analysis of the Apolipoprotein E and Apolipoprotein C1 Loci in Portugal and São Tomé e Príncipe (Gulf of Guinea): Linkage Disequilibrium Evidence That APOE*4 Is the Ancestral APOE Allele*

SUSANA SEIXAS,[1,2] MARIA JESUS TROVOADA,[3] AND JORGE ROCHA[1,2]

Abstract The joint distributions of phenotypes from the apolipoprotein E gene (*APOE*) and from a closely linked restriction site polymorphism at the apolipoprotein C1 locus (*APOC1*) were studied in population samples from Portugal and São Tomé e Príncipe (Gulf of Guinea), a former Portuguese colony that was originally populated by slaves imported from the African mainland. The frequencies of the *APOE* alleles (**2*, **3*, and **4*) in Portugal and São Tomé fitted the ranges of variation generally observed in European and African populations, respectively. Haplotype analysis showed that in both populations the strength of linkage disequilibrium was highest for the *APOE*2* allele and lowest for the *APOE*4* allele, suggesting that the origin of the *APOE* alleles followed a $4 \rightarrow 3 \rightarrow 2$ pathway and thus providing independent confirmation of the results from sequence homology studies with nonhuman primates. In accordance with global trends in the distribution of human genetic variation, the European sample from Portugal presented more intense linkage disequilibrium between *APOE* and *APOC1* than the African sample from São Tomé where, despite the short 4-kb distance that separates the 2 loci, the level of association between the *APOC1* alleles and *APOE*4* was nonsignificant.

Apolipoprotein E (apoE) is a polymorphic glycoprotein that is found in several major classes of lipoproteins and functions as a ligand for the receptor-mediated uptake of lipoprotein particles from plasma [reviewed by Mahley (1988)]. The apolipoprotein E gene (*APOE*) is located on chromosome 19q13.2 in a gene cluster of related apolipoprotein genes that spans approximately 48 kb and includes the apolipoprotein C1 (*APOC1*) and apolipopro-

[1] Instituto de Patologia e Imunologia Molecular, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200 Porto, Portugal.
[2] Faculdade de Ciências, Universidade do Porto, Porto, Portugal.
[3] Departamento de Antropologia, Universidade de Coimbra, Coimbra, Portugal.

Human Biology, December 1999, v. 71, no. 6, pp. 1001–1008.

KEY WORDS: HAPLOTYPE ANALYSIS, *APOE*, *APOC1*, PORTUGAL, SÃO TOMÉ

tein C2 (*APOC2*) loci, which code for activating cofactors of lecithin-cholesterol acyltransferase and lipoprotein lipase, respectively (Smit et al. 1988).

The *APOE* locus encodes 3 common isoforms that differ in their cysteine and arginine composition at positions 112 and 158 (Rall et al. 1982): E2 (Cys 112, 158), E3 (Cys 112, Arg 158) and E4 (Arg 112, 158). These isoforms have different affinities for the apoB/apoE cellular receptor and influence the variability of plasma cholesterol levels (Mahley 1988). In addition, the *APOE*4* allele has been found to be associated with an increased risk of Alzheimer's disease [reviewed by Kamboh (1995)].

Because of its clinical significance, the *APOE* polymorphism has been studied in several different human populations where, despite the wide variation observed in allele frequency distributions, *APOE*3* was found to be the most common *APOE* allele (Hallman et al. 1991; Gerdes et al. 1992; Kamboh 1995). Interspecies sequence comparisons have shown that *APOE* molecules from nonhuman primates possess arginine residues in the positions corresponding to human codons 112 and 158, indicating that *APOE*4* is more likely to be the ancestral human *APOE* allele than the common *APOE*3* allele (Hixson et al. 1988; Hanlon and Rubinsztein 1995). This has been further supported by the observation that the highest *APOE*4* frequencies occur among African populations, such as the Khoi San and the Pygmies, who are thought to retain a more archaic genetic make-up than other human groups (Sandholzer et al. 1995; Zekraoui et al. 1997).

In this study we present additional population genetic evidence that *APOE*4* is the human ancestral allele by evaluating the association between the *APOE* alleles and a closely linked restriction site polymorphism at the *APOC1* locus in 2 different populations: Portugal and São Tomé, the major island of the São Tomé e Príncipe archipelago located in the Gulf of Guinea, which was a former Portuguese colony that started to be peopled at the end of the 15th century by slaves imported from the African mainland.

Materials and Methods

Blood samples collected by venipuncture into tubes containing EDTA were obtained from unrelated individuals born and living in northern Portugal (n = 149) and in different places covering São Tomé island (n = 165). DNA was prepared from the whole blood using standard phenol-chloroform extraction protocols.

APOE types were determined using the polymerase chain reaction (PCR) followed by restriction enzyme digestion with *Cfo*I, as described by Wenham et al. (1991). Genetic variation at the *APOC1* locus was assessed by typing an *Hpa*I restriction site polymorphism located upstream from the 5′ end of the *APOC1* gene (Klasen et al. 1987; Smit et al. 1988) after PCR amplification according to the conditions described by Nillesen et al. (1990).

Digested DNA fragments were separated by horizontal electrophoresis in 12% polyacrylamide gels using the buffer system described by Luis and Caeiro (1995). DNA was visualized by silver staining according to the method of Budowle et al. (1991).

Allele frequencies at the individual loci were calculated by gene counting.

Haplotype frequency estimates, expected heterozygosity calculations, and statistical analyses for testing Hardy-Weinberg equilibrium and linkage disequilibrium were done using the ARLEQUIN software package (Schneider et al. 1997). In ARLEQUIN divergence from Hardy-Weinberg expectations is assessed by a modified version of the exact test described by Guo and Thompson (1992). Expected heterozygosities are estimated according to the method of Nei (1987). Maximum-likelihood estimates of haplotype frequencies are computed using an expectation-maximization algorithm (Dempster et al. 1977), and linkage disequilibrium is tested using a likelihood ratio test whose distribution is obtained by permutation (Slatkin and Excoffier 1996).

The relative amounts of association between the *APOE* and *APOC1* alleles were quantified by the standardized linkage disequilibrium coefficient D' (Lewontin 1964).

To test for linkage disequilibrium in the different *APOE* alleles, we split the overall 6 × 3 matrices with the joint *APOE-APOC1* phenotype distributions into 3 different 3 × 3 separate distributions in which *APOE* was treated as a biallelic locus with each of its alleles segregating with the sum of the remaining 2 alleles.

Results and Discussion

Table 1 presents the joint phenotype distributions and allele frequencies of *APOE* and *APOC1* in Portugal and São Tomé. No significant departures from Hardy-Weinberg equilibrium were found for either locus in the 2 populations.

The *APOE* allele frequency distribution in Portugal is similar to distributions observed in other populations from southern Europe and supports the previously reported gradual decrease in *APOE*4* frequencies and increase in *APOE*3* frequencies along a north to south cline in European populations [reviewed by Lucotte et al. (1997)].

In São Tomé the frequencies of the *APOE* alleles fall within the range of variation so far reported for other African populations from the continental mainland where, despite the interethnic variation, the average *APOE*4* frequency is higher than the frequency from Europe (Zekraoui et al. 1997).

Table 2 presents the haplotype frequency estimates, the probability values from the linkage disequilibrium tests, and the D' values for the nonrandom association between the 3 *APOE* alleles and the *APOC1*1* allele in Portugal

Table 1. Joint Phenotype Distributions and Frequencies of *APOE* and *APOC1* Alleles in the Populations of Portugal and São Tomé

Joint APOE-APOC1 Phenotype Distribution				*APOE Allele Frequency Distribution*			*APOC1 Allele Frequency Distribution*		
APOE Phenotype	*APOC1 Phenotype*	*Portugal* (n = 149)[a]	*São Tomé* (n = 165)[b]	*APOE Allele*	*Portugal*[c]	*São Tomé*[c]	*APOC1 Allele*[d]	*Portugal*[c]	*São Tomé*[c]
2									
	1	0	0	*2	0.044 ± 0.012	0.100 ± 0.017	*1	0.879 ± 0.019	0.685 ± 0.026
	1,2	0	1	*3	0.882 ± 0.019	0.652 ± 0.026	*2	0.121	0.315
	2	0	1	*4	0.074 ± 0.015	0.248 ± 0.024			
2,3									
	1	0	1						
	1,2	13	18						
	2	0	3						
2,4									
	1	0	1						
	1,2	0	4						
	2	0	2						
3									
	1	112	42						
	1,2	3	20						
	2	1	4						
3,4									
	1	3	35						
	1,2	15	21						
	2	0	5						
4									
	1	0	1						
	1,2	1	2						
	2	1	4						

a. Exact test for Hardy-Weinberg equilibrium: *APOE*, p = 0.431 (SD, ±0.002); *APOC1*, p = 1.000.

b. Exact test for Hardy-Weinberg equilibrium: *APOE*, p = 0.444 (SD, ±0.002); *APOC1*, p = 0.384 (SD, ±0.002).

c. Plus or minus standard deviation.

d. *APOC1*1* = absence of the *HpaI* restriction site.

Table 2. Haplotype Frequency Estimates, Probability Values from Linkage Disequilibrium Tests, and Linkage Disequilibrium Values (D') for the *APOE* and *APOC1* Loci in the Populations of Portugal and São Tomé

	Portugal				*São Tomé*			
APOE	*APOC1*1*[a]	*APOC1*2*[a]	*p*	D'[b]	*APOC1*1*[a]	*APOC1*2*[a]	*p*[a]	D'[b]
**2*	0.000	0.044 ± 0.012	0.000	−1.00	0.012 ± 0.006	0.088 ± 0.016	0.000	−0.825
**3*	0.865 ± 0.022	0.017 ± 0.007	0.000	0.865	0.521 ± 0.033	0.131 ± 0.022	0.000	0.362
**4*	0.014 ± 0.006	0.060 ± 0.013	0.000	−0.785	0.150 ± 0.022	0.098 ± 0.021	0.186 ± 0.001	−0.117

a. Plus or minus standard deviation.

b. Association between the *APOC1*1* allele and the corresponding *APOE* allele.

Table 3. Single Locus Expected Heterozygosities and Haplotype Diversities in Portugal and São Tomé

	Expected Heterozygosity		
Population	*APOE*[a]	*APOC1*[a]	*Haplotype*[b]
Portugal	0.214 ± 0.031	0.213 ± 0.029	0.246 ± 0.032
São Tomé	0.505 ± 0.024	0.433 ± 0.019	0.674 ± 0.022

a. Calculated using the allele frequencies for any given locus.
b. Calculated using the haplotype frequencies for *APOE* and *APOC1*.

and São Tomé. In both populations the *APOE*4* and *APOE*2* alleles are associated with the *APOC1*2* allele (negative *D'* values), whereas *APOE*3* is associated with *APOC1*1* (positive *D'* values), in accordance with *APOE-APOC1* haplotype information derived from case-control studies of European patients with type III hyperlipidemia (Klasen et al. 1987) or Alzheimer's patients of European and Japanese ancestry (Chartier-Harlin et al. 1994; Poduslo et al. 1995; Kamino et al. 1996). Such associations in different populations suggest that *APOE*4-APOC1*2*, *APOE*3-APOC1*1*, and *APOE*2-APOC1*2* represent the ancestral haplotypes arising from the mutations that created the different *APOE* alleles.

In addition, clear differences in the amount of linkage disequilibrium were found between the 3 *APOE* alleles in both Portugal and São Tomé, with *APOE*4* and *APOE*2* presenting the lowest and the highest absolute *D'* values, respectively. Because the initial amount of linkage disequilibrium between *APOE* and the marker *APOC1* locus is expected to decrease over time as a result of recombination, these results indicate that the origin of *APOE* gene products followed a 4 → 3 → 2 pathway with successive single Arg → Cys substitution events at codons 112 and 158. This is independent confirmation of previous suggestions that *APOE*4* is the ancestral *APOE* allele on the basis of sequence homology with nonhuman primates (Hixson et al. 1988; Hanlon and Rubinsztein 1995).

Comparison of *D'* values in the 2 study populations shows that linkage disequilibrium for all *APOE* alleles is consistently stronger in Portugal than in São Tomé, where the allelic association between *APOE*4* and *APOC1* alleles is not even statistically significant ($p = 0.186$), despite the short 4-kb distance that separates the 2 loci (Smit et al. 1988). Together with the higher single locus heterozygosity and haplotype diversity observed in São Tomé (Table 3), this weaker linkage disequilibrium supports the global patterns of genetic diversity showing that African populations were less subjected than non-African human groups to genetic drift events interrupting the breakdown of linkage disequilibrium (Tishkoff et al. 1996). On a more local scale this observation also suggests that the peopling of São Tomé island may

not have been accompanied by a significant reduction in genetic diversity and therefore agrees with recent research by Mateu et al. (1997), who found that mitochondrial DNA genetic variation in São Tomeans remains comparable to that of mainland African populations.

The extension of the *APOE-APOC1* haplotype analysis to a more significant number of human populations will lead to a more reliable interpretation of the evolution and dispersion of the *APOE* alleles.

Acknowledgments Susana Seixas is supported by Praxis XXI through grant BD/13885/97.

Received 21 January 1999.

Literature Cited

Budowle, B., R. Chakraborty, A.M. Giusti et al. 1991. Analysis of the VNTR locus *D1S80* by PCR followed by high-resolution PAGE. *Am. J. Hum. Genet.* 48:137–148.

Chartier-Harlin, M.-C., M. Parfitt, S. Legrain et al. 1994. Apolipoprotein E, ε4 allele as a major risk factor for sporadic early- and late-onset forms of Alzheimer's disease: Analysis of the 19q13.2 chromosomal region. *Hum. Molec. Genet.* 3:569–574.

Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc.* 39:1–38.

Gerdes, L.U., I.C. Klausen, I. Sihm et al. 1992. Apolipoprotein E polymorphism in a Danish population compared to findings in 45 other study populations around the world. *Genet. Epidemiol.* 9:155–167.

Guo, S.W, and E.A. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics* 48:361–372.

Hallman, D.M., E. Boerwinkle, N. Saha et al. 1991. The apolipoprotein E polymorphism: A comparison of allele frequencies and effects in nine populations. *Am. J. Hum. Genet.* 49:338–349.

Hanlon, C.S., and D.C. Rubinsztein. 1995. Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. *Atherosclerosis* 112:85–90.

Hixson, J.E., L.A. Cox, and S. Borenstein. 1988. The baboon apolipoprotein E gene: Structure, expression, and linkage with the gene for apolipoprotein C-I. *Genomics* 2:315–323.

Kamboh, M.I. 1995. Apolipoprotein E polymorphism and susceptibility to Alzheimer's disease. *Hum. Biol.* 67:195–215.

Kamino, K., A. Yoshiiwa, Y. Nishiwaki et al. 1996. Genetic association study between senile dementia of Alzheimer's type and *APOE/C1/C2* gene cluster. *Gerontology* 42(suppl. 1):12–19.

Klasen, E.C., P.J. Talmud, L. Havekes et al. 1987. A common restriction fragment length polymorphism of the human apolipoprotein E gene and its relationship to type III hyperlipidemia. *Hum. Genet.* 75:244–247.

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49–67.

Lucotte, G., F. Loirat, and S. Hazout. 1997. Pattern of gradient of apolipoprotein E allele *4 frequencies in Western Europe. *Hum. Biol.* 69:253–262.

Luis, J.R., and B. Caeiro. 1995. Application of two STRs (*VWA* and *TPO*) to human population profiling: Survey in Galicia. *Hum. Biol.* 67:789–795.

Mahley, R.W. 1988. Apolipoprotein E: Cholesterol transport protein with expanding role in cell biology. *Science* 240:622–633.

Mateu, E., D. Comas, F. Calafell et al. 1997. A tale of two islands: Population history and mitochondrial DNA sequence variation of Bioko and S. Tomé, Gulf of Guinea. *Ann. Hum. Genet.* 61:507–518.

Nei, M. 1987. *Molecular Evolutionary Genetics.* New York: Columbia University Press.

Nillesen, W.M., H.J. Smeets, and B.A. van Oost. 1990. Human ApoCI *HpaI* restriction site polymorphism revealed by the polymerase chain reaction. *Nucleic Acids Res.* 18:3428.

Poduslo, S.E., M. Neal, and J. Schwankhaus. 1995. A closely linked gene to apolipoprotein E may serve as an additional risk factor for Alzheimer's disease. *Neurosci. Lett.* 201:81–83.

Rall, S.C., K.H. Weisgraber, and R.W. Mahley. 1982. Human apolipoprotein E: The complete amino acid sequence. *J. Biol. Chem.* 257:4171–4178.

Sandholzer, C., R. Delport, H. Vermaak et al. 1995. High frequency of the apo ε4 allele in Khoi San from South Africa. *Hum. Genet.* 95:46–48.

Schneider, S., J.-M. Kueffer, D. Roessli et al. 1997. *Arlequin ver. 1.1: A Software for Population Genetic Data Analysis.* Geneva, Switzerland: Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva.

Slatkin, M., and L. Excoffier. 1996. Testing for linkage disequilibrium in genotypic data using the EM algorithm. *Heredity* 76:377–383.

Smit, M., E. van der Kooij-Meijs, R.R. Frants et al. 1988. Apolipoprotein gene cluster on chromosome 19: Definite localization of the *APOC2* gene and the polymorphic *HpaI* site associated with type III hyperlipoproteinemia. *Hum. Genet.* 78:90–93.

Tishkoff, S.A., E. Dietzsch, W.C. Speed et al. 1996. Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. *Science* 271:1380–1387.

Wenham, P.R., W.H. Price, and G. Blundell. 1991. Apolipoprotein E genotyping by one-stage PCR. *Lancet* 337:1158–1159.

Zekraoui, L., J.P. Lagarde, A. Raisonnier et al. 1997. High frequency of the apolipoprotein *E*4* allele in African Pygmies and most of the African populations in sub-Saharan Africa. *Hum. Biol.* 69:575–581.

3. Referências Bibliográficas

Artiga, M. J., Bullido, M. J., Sastre, I., Recuero, M., Garcia, M. A., Aldudo, J., Vazquez, J., Valdivieso, F. (1998). Allelic polymorphisms in the transcriptional regulatory region of apolipoprotein E gene. *FEBS Lett* **421**:105-108.

Bengtsson, B. O. Thomson, G. (1981). Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* **18**:356-363.

Boerwinkle, E., Visvikis, S., Welsh, D., Steinmetz, J., Hanash, S. M., Sing, C. F. (1987). The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of the apolipoprotein E polymorphism in determining levels, variability, and covariability of cholesterol, betalipoprotein, and triglycerides in a sample of unrelated individuals. *Am J Med Genet* **27**:567-582.

Bowman, B. H. (1992). *Hepatic plasma proteins: mechanisms of function and regulation.* Academic Press Inc.

Bullido, M. J., Artiga, M. J., Recuero, M., Sastre, I., Garcia, M. A., Aldudo, J., Lendon, C., Han, S. W., Morris, J. C., Frank, A., Vazquez, J., Goate, A., Valdivieso, F. (1998). A polymorphism in the regulatory region of APOE associated with risk for Alzheimer's dementia. *Nat Genet* **18**:69-71.

Caldeira, A. M. (1999). *Mulheres, sexualidade e casamento em São Tomé e Príncipe (séculos XV-XVII).* Edições Cosmo, Lisboa.

Chartier-Harlin, M. C., Parfitt, M., Legrain, S., Perez-Tur, J., Brousseau, T., Evans, A., Berr, C., Vidal, O., Roques, P., Gourlet, V. (1994). Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. *Hum Mol Genet* **3**:569-574.

Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proc Natl Acad Sci U S A* **94**:7730-7734.

Clark, A. G. (1999). Chips for chimps. *Nat Genet* **22**:119-120.

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Jr., Rimmler, J. B., Locke, P. A., Conneally, P. M., Schmader, K. E. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet* **7**:180-184.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**:921-923.

Davignon, J., Gregg, R. E., Sing, C. F. (1988). Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis* **8**:1-21.

Devlin, B., Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**:311-322.

Fullerton, S. M., Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Stengard, J. H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., Sing, C. F. (2000). Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* **67**:881-900.

Goldstein, D. B., Reich, D. E., Bradman, N., Usher, S., Seligsohn, U., Peretz, H. (1999). Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. *Am J Hum Genet* **64**:1071-1075.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P., Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* **22**:164-167.

Hagemeijer, T. (1999). As ilhas de Babel: a crioulização no Golfo da Guiné. *Camões* **6**:74-88.

Hanlon, C. S., Rubinsztein, D. C. (1995). Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. *Atherosclerosis* **112** :85-90.

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., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., Sherry, S. T. (1998). Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* **95**:1961-1967.

Hartl, D. L., Clark, A. G. (1989). *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts.

Hixson, J. E., Cox, L. A., Borenstein, S. (1988). The baboon apolipoprotein E gene: structure, expression, and linkage with the gene for apolipoprotein C-1. *Genomics* **2**:315-323.

Jong, M. C., Hofker, M. H., and Havekes, L. M. (1999). Role of ApoCs in lipoprotein metabolism: functional differences between ApoC1, ApoC2, and ApoC3. *Arterioscler Thromb Vasc Biol* **19**:472-484.

Jorde, L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res* **10**:1435-1444.

Kamboh, M. I. (1995). Apolipoprotein E polymorphism and susceptibility to Alzheimer's disease. *Hum Biol* **67**:195-215.

Kamino, K., Yoshiiwa, A., Nishiwaki, Y., Nagano, K., Yamamoto, H., Kobayashi, T., Nonomura, Y., Yoneda, H., Sakai, T., Imagawa, M., Miki, T., Ogihara, T. (1996). Genetic association study between senile dementia of Alzheimer's type and APOE/C1/C2 gene cluster. *Gerontology* **42**:12-19.

Kardia, L. R., Stengärd, J., Templeton, A. (1999). An evolutionary perspective on the genetic architecture of susceptibility to cardiovascular disease. In: Stearns, S. C., (ed). *Evolution in Health and Disease.* Oxford University Press, Oxford, pp. 231-245.

Kimura, M. (1983). *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

Kimura, M., Ohta, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics* **75**:199-212.

Klasen, E. C., Talmud, P. J., Havekes, L., de Wit, E., Kooij-Meijs, E., Smit, M., Hansson, G., Humphries, S. E. (1987). A common restriction fragment length polymorphism of the human apolipoprotein E gene and its relationship to type III hyperlipidaemia. *Hum Genet* **75**:244-247.

Labuda, D., Zietkiewicz, E., Yotova, V. (2000). Archaic lineages in the history of modern humans. *Genetics* **156**:799-808.

Lambert, J. C., Berr, C., Pasquier, F., Delacourte, A., Frigard, B., Cottel, D., Perez-Tur, J., Mouroux, V., Mohr, M., Cecyre, D., Galasko, D., Lendon, C., Poirier, J., Hardy, J., Mann, D., Amouyel, P., Chartier-Harlin, M. C. (1998). Pronounced impact of Th1/E47cs mutation compared with -491 AT mutation on neural APOE gene expression and risk of developing Alzheimer's disease. *Hum Mol Genet* **7**:1511-1516.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**:49-67.

Li, W. H., Tanimura, M., Luo, C. C., Datta, S., Chan, L. (1988). The apolipoprotein multigene family: biosynthesis, structure, structure-function relationships, and evolution. *J Lipid Res* **29**:245-271.

Lucotte, G., Loirat, F., Hazout, S. (1997). Pattern of gradient of apolipoprotein E allele *4 frequencies in western Europe. *Hum Biol* **69**:253-262.

Mahley, R. W., Rall, S. C., Jr. (1995). Type III hyperlipoproteinemia (dysbetalipoproteinemia): the role of apolipoprotein e in normal and abnormal lipoprotein metabolism. In: Sriver, C. R., Beaudet, A. L., Sly, W., Valle, D. (eds). *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill, New York, pp. 1953-1980.

Mahley, R. W., Rall, S. C., Jr. (2000). Apolipoprotein E: far more than a lipid transport protein. *Annu Rev Genomics Hum Genet* **1**:507-537.

Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, J., Finch, K. L., Stevens, J. F., Livak, K. J., Slotterbeck, B. D., Slifer, S. H., Warren, L. L., Conneally, P. M., Schmechel, D. E., Purvis, I., Pericak-Vance, M. A., Roses, A. D., Vance, J. M. (2000). SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am J Hum Genet* **67**:383-394.

Mui, S., Briggs, M., Chung, H., Wallace, R. B., Gomez-Isla, T., Rebeck, G. W., Hyman, B. T. (1996). A newly identified polymorphism in the apolipoprotein E enhancer gene region is associated with Alzheimer's disease and strongly with the epsilon 4 allele. *Neurology* **47**:196-201.

Nickerson, D. A., Taylor, S. L., Fullerton, S. M., Weiss, K. M., Clark, A. G., Stengard, J. H., Salomaa, V., Boerwinkle, E., Sing, C. F. (2000). Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res* **10**:1532-1545.

Petersen, R. C., Waring, S. C., Smith, G. E., Tangalos, E. G., Thibodeau, S. N. (1996). Predictive value of APOE genotyping in incipient Alzheimer's disease. *Ann N Y Acad Sci* **802**:58-69.

Poduslo, S. E., Neal, M., Schwankhaus, J. (1995). A closely linked gene to apolipoprotein E may serve as an additional risk factor for Alzheimer's disease. *Neurosci Lett* **201**:81-83.

Rall, S. C., Jr., Weisgraber, K. H., Innerarity, T. L., Mahley, R. W. (1982a). Structural basis for receptor binding heterogeneity of apolipoprotein E from type III hyperlipoproteinemic subjects. *Proc Natl Acad Sci U S A* **79**:4696-4700.

Rall, S. C., Jr., Weisgraber, K. H., Mahley, R. W. (1982b). Human apolipoprotein E. The complete amino acid sequence. *J Biol Chem* **257**:4171-4178.

Rannala, B., Bertorelle, G. (2001). Using linked markers to infer the age of a mutation. *Hum Mutat* **18**:87-100.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**:199-204.

Relethford, J. H. (2001). *Genetics and the Search for Modern Human Origins*. Wiley-Liss, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* **9**:152-159.

Roses, A. D. (2000). Pharmacogenetics and the practice of medicine. *Nature* **405**:857-865.

Sahota, A., Yang, M., Gao, S., Hui, S. L., Baiyewu, O., Gureje, O., Oluwole, S., Ogunniyi, A., Hall, K. S., Hendrie, H. C. (1997). Apolipoprotein E-associated risk for Alzheimer's disease in the African-American population is genotype dependent. *Ann Neurol* **42**:659-661.

Sandholzer, C., Delport, R., Vermaak, H., Utermann, G. (1995). High frequency of the apo epsilon 4 allele in Khoi San from South Africa. *Hum Genet* **95**:46-48.

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

Selkoe, D. J. (2001). Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* **81**:741-766.

Siest, G., Pillot, T., Regis-Bailly, A., Leininger-Muller, B., Steinmetz, J., Galteau, M. M., Visvikis, S. (1995). Apolipoprotein E: an important gene and protein to follow in laboratory medicine. *Clin Chem* **41**:1068-1086.

Sing, C. F., Davignon, J. (1985). Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *Am J Hum Genet* **37**:268-285.

Slatkin, M., Rannala, B. (2000). Estimating allele age. *Annu Rev Genomics Hum Genet* **1**:225-249.

Smit, M., Kooij-Meijs, E., Frants, R. R., Havekes, L., Klasen, E. C. (1988). Apolipoprotein gene cluster on chromosome 19. Definite localization of the APOC2 gene and the polymorphic Hpa I site associated with type III hyperlipoproteinemia. *Hum Genet* **78**:90-93.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A* **90**:1977-1981.

Tang, M. X., Maestre, G., Tsai, W. Y., Liu, X. H., Feng, L., Chung, W. Y., Chun, M., Schofield, P., Stern, Y., Tycko, B., Mayeux, R. (1996). Effect of age, ethnicity, and head injury on the association between APOE genotypes and Alzheimer's disease. *Ann N Y Acad Sci* **802**:6-15.

Templeton, A. (2002). Out of Africa again and again. *Nature* **416**:45-51.

Templeton, A. R. (1995). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics* **140**:403-409.

Tenreiro, F. (1961). *A ilha de São Tomé.* Memórias da Junta de Investigação do Ultramar. Lisboa.

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.

Tremblay, M., Vezina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet* **66**:651-658.

Trovoada, M. J., Alves, C., Gusmão, L., Abade, A., Amorim, A., Prata, M. J. (2001). Evidence for population sub-structuring in São Tomé e Príncipe as inferred from Y-chromosome STR analysis. *Ann Hum Genet* **65**:271-283.

Watterson, G. A. and Guess, H. A. (1977). Is the most frequent allele the oldest? *Theor Popul Biol* **11**:141-160.

Xu, Y., Berglund, L., Ramakrishnan, R., Mayeux, R., Ngai, C., Holleran, S., Tycko, B., Leff, T., Shachter, N. S. (1999). A common Hpa I RFLP of apolipoprotein C-I increases gene transcription and exhibits an ethnically distinct pattern of linkage disequilibrium with the alleles of apolipoprotein E. *J Lipid Res* **40**:50-58.

Zekraoui, L., Lagarde, J. P., Raisonnier, A., Gerard, N., Aouizerate, A., Lucotte, G. (1997). High frequency of the apolipoprotein E *4 allele in African pygmies and most of the African populations in sub-Saharan Africa. *Hum Biol* **69**:575-581.

Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K. K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M., Labuda, D. (1998). Genetic structure of the ancestral population of modern humans. *J Mol Evol* **47**:146-155.

PARTE II

α1-antitripsina

1. Introdução

A α1-antitripsina (PI) é uma glicoproteína monomérica de síntese predominantemente hepática, que se localiza na fracção electroforética das globulinas α1 e constitui o principal inibidor plasmático das proteases que possuem serina no centro activo. Apesar do nome utilizado para a designar - resultante de os primeiros estudos sobre a sua capacidade inibidora terem sido realizados com tripsina – e apesar de poder inibir um amplo espectro de proteases, a principal função da α1-antitripsina é a protecção do trato respiratório inferior da acção proteolítica da elastase dos neutrófilos (Cox, 1995).

O *locus* da PI situa-se na região cromossómica 14q32.1, num agrupamento sinténico que ocupa aproximadamente 320 kb e que inclui um pseudogene da α1-antitripsina (PIL) e os genes parálogos da α1-antiquimotripsina (AACT), da globina transportadora de corticosteróides (CBG), do inibidor da proteína C (PCI) e da calistatina (PI4) (Figura II.1) (Byth *et al.*, 1994; Rollini e Fournier, 1997). Estas proteínas fazem parte de uma família mais vasta de moléculas homólogas, que terão sido originadas por processos de duplicação a partir de um único *locus* ancestral e que são designadas pelo acrónimo SERPIN (*serine protease inhibitors*) (Salzet *et al.*, 1999; Irving *et al.*, 2000). Com excepção de um pequeno conjunto de moléculas que, à semelhança da CBG, têm funções de transporte, a maioria das proteínas da família SERPIN participa no sistema defensivo que regula os fenómenos proteolíticos envolvidos em diferentes processos fisiológicos, incluindo a regeneração dos tecidos, a activação do complemento e o controlo da coagulação (Salzet *et al.*, 1999; Irving *et al.*, 2000).

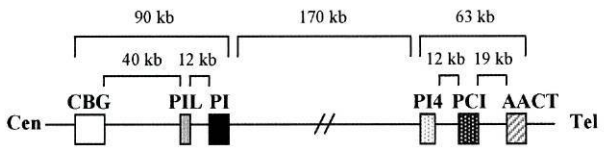


Figura II.1- Agrupamento génico das serpinas na região cromossómica 14q32.1 (distâncias intergénicas de acordo Rollini e Fournier, 1997).

A α1-antitripsina é a proteína mais bem caracterizada da família SERPIN e apresenta um amplo espectro de formas variantes que inclui mais de 75 produtos génicos diferentes que podem ser, na sua maior parte, discriminados por técnicas de focagem isoeléctrica. Os alelos PI*M1, PI*M2, PI*M3, PI*M4, PI*S e PI*Z são os mais comuns e atingem frequências polimórficas em várias populações humanas. A análise da base molecular destes alelos conduziu à diferenciação de dois subtipos comuns de M1 (M1Ala213 e M1Val213) - que apenas podem ser distinguidos por análise de DNA - e permitiu estabelecer uma sequência filogenética em que M1Ala213 é considerado o alelo mais antigo por apresentar maior homologia com as sequências ortólogas de Primatas não humanos (Figura II.2) (Brantly *et al.*, 1988; Nukiwa *et al.*, 1996a).

As mutações que caracterizam os diferentes alelos da PI podem causar variações significativas na concentração sérica da proteína. Os alelos M mais comuns estão associados a valores considerados normais. Os produtos génicos S e Z têm níveis circulantes que correspondem, respectivamente, a apenas 60% e 15% daqueles valores e originam constelações genotípicas em que se verifica a deficiência aprofundada da proteína, a que correspondem aumentos do risco de ocorrência de condições patológicas (Cox, 1995). A deficiência associada ao alelo Z é causada pela substituição Glu342Lys e resulta de uma alteração conformacional que induz a polimerização da proteína e leva à acumulação no retículo endoplasmático dos hepatócitos de cerca de 85% da α1-antitripsina produzida (Lomas *et al.*, 1992). Esta acumulação intra-hepática gera depósitos que podem ser observados histologicamente como grânulos PAS positivos diastase-resistentes (Cox, 1995; Massi, 1996). O produto génico S tem uma mutação Glu264Val que também aumenta a instabilidade

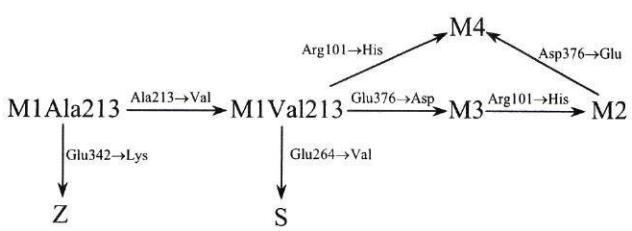


Figura II.2- Filogenia dos alelos da α1-antitripsina.

conformacional e a susceptibilidade à polimerização. Contudo, a retenção intracelular associada a esta mutação é muito mais moderada do que a da variante Z e não leva à formação de grânulos visíveis (Elliott *et al.*, 1996; Mahadeva *et al.*, 1999).

Além das substituições Glu342Lys e Glu264Val, já se identificaram pelo menos 30 mutações, muito mais raras, que conduzem à deficiência de α1-antitripsina (Crystal, 1990; Brantly, 1996). A maioria destas mutações ocorre nas regiões codificantes do gene da PI e provoca a redução dos níveis circulantes ou a ausência total de proteína no plasma (alelos nulos). A diminuição da concentração plasmática está sobretudo associada a substituições de aminoácidos que alteram a estabilidade conformacional da proteína e causam a sua retenção intra-hepática, ou o aumento da sua susceptibilidade a processos de degradação intra e extracelulares. Os alelos nulos resultam de um conjunto mais variado de processos mutacionais que incluem a substituição de aminoácidos, a delecção parcial ou total do gene, ou a formação de codões de terminação prematuros, com síntese de proteínas truncadas ou degradação de mRNA (Crystal, 1990; Lee e Brantly, 2000).

A principal manifestação clínica do défice de α1-antitripsina é o enfisema pulmonar precoce, que se desencadeia na terceira ou quarta década de vida como resultado do decréscimo dos níveis de protecção das fibras elásticas alveolares do ataque proteolítico da elastase leucocitária a que estão cronicamente sujeitas (Cox, 1995; WHO, 1997). A homozigotia para o alelo Z representa cerca de 95% das formas mais graves de deficiência de PI e é o principal genótipo associado a este tipo de enfisema. No entanto, todas as combinações alélicas que conduzam a níveis séricos da proteína inferiores a 11μM (cerca de 30% da concentração normal) comprometem a barreira antiproteolítica e aumentam o risco de patologia pulmonar, independentemente dos mecanismos moleculares envolvidos na deficiência (Crystal, 1990). O tabagismo contribui significativamente para gravidade da doença e antecipa em cerca de 10 a 15 anos o aparecimento dos seus primeiros sintomas (Cox, 1995; WHO, 1997). Nas situações de deficiência menos acentuadas, como na heterozigotia SZ, a que correspondem concentrações da proteína situadas no limiar de risco, e mesmo em condições de défice mais atenuado, a existência de hábitos tabágicos pode desencadear enfisema precoce em indivíduos que, de contrário, não estariam expostos a uma probabilidade acrescida de o contrair (Cox, 1995; WHO, 1997).

A segunda consequência clínica mais frequente do défice de α1-antitripsina é constituída por formas de doença hepática, geralmente associadas ao genótipo ZZ, que podem manifestar-se na infância ou em adultos de idade mais avançada (Crystal, 1990; Cox, 1995; Massi, 1996). Na infância, a patologia mais comum é a síndroma da hepatite neonatal, caracterizada pela ocorrência de hiperbilirrubinémia conjugada, hepatomegalia e elevação das transaminases.

No estudo longitudinal mais completo até hoje realizado (Sveger, 1976), verificou-se que 17% das crianças ZZ apresentaram sinais de disfunção hepática. Destas, 41% (~7% do total de indivíduos ZZ) tiveram formas mais graves da doença que evoluiu para cirrose fatal em cerca de 1/3 dos casos (~2,5% do total de indivíduos ZZ). Nos adultos, a doença manifesta-se sob a forma de hepatite crónica, cirrose, hipertensão portal ou carcinoma hepatocelular (Eriksson *et al.*, 1986; Cox, 1995; Massi, 1996). Os dados epidemiológicos sobre estas manifestações clínicas são mais escassos e difíceis de interpretar do que as formas pediátricas, mas indicam que o risco de doença é maior no sexo masculino e que a frequência da cirrose pode atingir valores entre 15 e 20% em indivíduos ZZ com mais de 50 anos (Cox e Smyth, 1983; Cox, 1995; Massi, 1996). A evidência disponível sugere que a patologia do fígado resulta, quer nas crianças quer nos adultos, dos efeitos tóxicos da acumulação da α1-antitripsina nos hepatócitos, e não da deficiência da proteína em circulação. A doença não se manifesta em indivíduos que possuem genótipos raros em que, apesar de níveis de PI mais baixos do que os causados pela homozigotia para o alelo Z, não há deposição intrahepática da proteína (Crystal, 1990). Inversamente, foram descritos casos de patologia hepática associados a outras variantes raras que partilham com o alelo Z a tendência para formar agregados visíveis no interior dos hepatócitos, apesar de haver alguma secreção da proteína para o plasma (Curiel *et al.*, 1989a; Seyama *et al.*, 1991). Por outro lado, verificou-se que em ratos transgénicos que expressam o alelo Z, mas têm níveis de PI normais, formam-se glóbulos da proteína no interior dos hepatócitos e há marcas histológicas e sintomatologia típicas de disfunção do fígado (Dycaico *et al.*, 1988; Carlson *et al.*, 1989). Para além da constituição genotípica, há factores adicionais, como a elevação da temperatura corporal ou o aumento da síntese da α1-antitripsina no decurso de inflamações sistémicas, que podem favorecer a polimerização da proteína, exacerbar a sua acumulação e contribuir para a

manifestação da doença (Lomas *et al.*, 1992; Dafforn *et al.*, 1999). A diminuição da capacidade de degradar a proteína acumulada também pode provocar o aparecimento da sintomatologia e justificar a sua distribuição etária. Nos casos de expressão pediátrica, essa diminuição poderá estar relacionada com a imaturidade do metabolismo hepático, enquanto que nos adultos a expressão clínica resultará da perda da capacidade de degradação com o avanço da idade (Carrell e Lomas, 2002). Há também evidência de que diferenças constitucionais no metabolismo de degradação intracelular podem contribuir para a variação interindividual na manifestação da doença, em qualquer das idades de risco (Wu *et al.*, 1994; Teckman *et al.*, 2001).

Devido à sua associação com as condições patológicas mencionadas e ao seu interesse como marcador genético, o polimorfismo da α1-antitripsina foi já estudado num número muito significativo de populações humanas. Os alelos M1, M2 e M3 têm ampla distribuição e foram descritos em quase todas as populações analisadas, embora haja variações consideráveis nas respectivas frequências (Figura II.3). O alelo M4 é menos comum e mais difícil de discriminar por focagem isoeléctrica, mas também foi detectado em amostras geograficamente afastadas. A ocorrência de PI*S e PI*Z limita-se, com excepção de casos de miscigenação, às populações de origem europeia (Figura II.3). As frequências mais elevadas de PI*Z são observadas na região da Escandinávia (Figura II.4 A), onde podem atingir valores médios próximos de 0,02 correspondentes a uma incidência do défice grave de α1-antitripsina de 1 em cada 2000 recém-nascidos (Sveger, 1976; Hutchison, 1998). Nas regiões do Sul da Europa, a frequência de PI*Z é mais baixa, mas mantém-se nos limites das proporções polimórficas (~0,01) (Figura II.4 A). Estas frequências fazem com que o défice de PI devido a homozigotia ZZ constitua uma das principais causas de transplante hepático pediátrico e colocam-no, juntamente com a fibrose quística, entre as doenças hereditárias mais comuns nas populações europeias. Apesar de também se encontrar confinado à Europa, o alelo S tem uma distribuição muito diferente, caracterizada por um vincado gradiente em que as frequências decrescem de sudoeste para nordeste (Figura II.4 B). As frequências mais altas de PI*S observam-se na Península Ibérica onde chegam a atingir valores de 0,15, cerca dez vezes mais elevados do que os observados, por exemplo, em algumas amostras do norte e no centro da Europa.

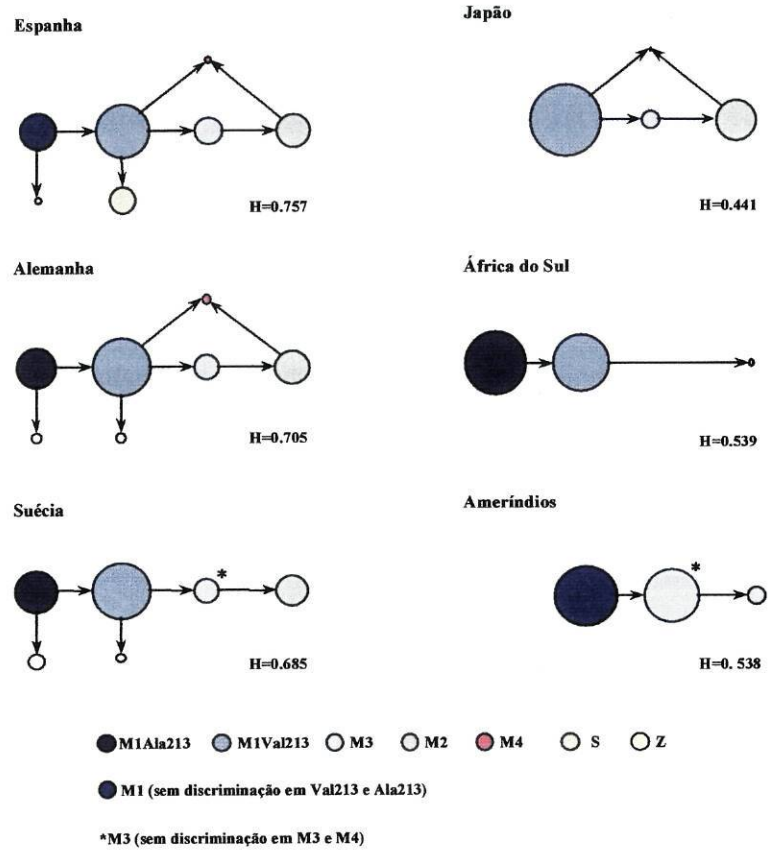


Figura II.3- Exemplos de frequências dos alelos de PI em diferentes populações humanas. A cada alelo corresponde um círculo cuja área é proporcional à sua frequência na população. As setas indicam as relações filogenéticas esquematizadas na figura II.2. H-heterozigotia.

O polimorfismo da α1-antitripsina constitui, sob muitos aspectos, um modelo que permite investigar algumas das questões centrais relacionadas com a interpretação das causas e consequências da variação genética em regiões funcionalmente relevantes do genoma humano. Do ponto de vista bioquímico, a molécula é considerada um arquétipo das proteínas da família SERPIN e muitas das suas mutações têm permitido compreender melhor as relações entre a estrutura e a função destas proteínas (Stein e Carrell, 1995; Irving *et al.*, 2000). O entendimento dessas relações também conduziu à identificação de mecanismos comuns de patogénese em doenças aparentemente muito díspares que não envolvem inibidores de proteases (Carrell e Lomas, 2002). Por exemplo, a alteração conformacional que ocorre no produto génico Z é hoje em dia considerada um protótipo dos processos moleculares que poderão estar na base de

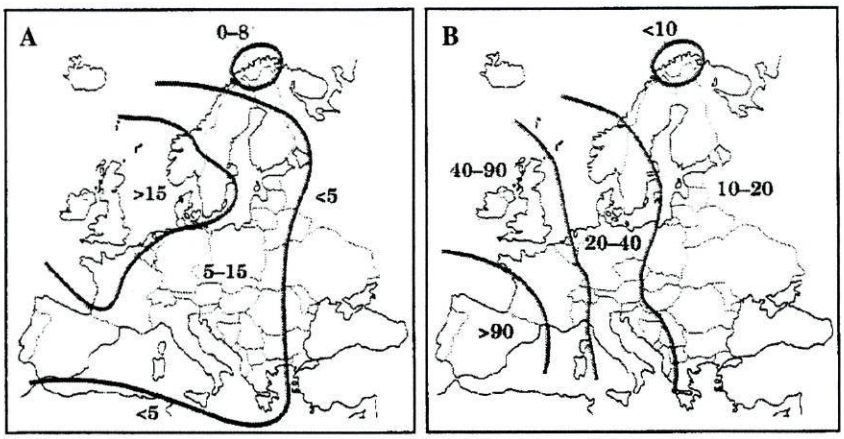


Figura II.4- Distribuição da frequência dos alelos PI*Z e PI*S no continente europeu. **A-** PI*Z **B-** PI*S. Os valores de frequência encontram-se expressos em numero de genes Z e S por 1000 (figura de Hutchinson 1998).

patologias como as encefalopatias espongiformes, a doença de Alzheimer ou a paramiloidose, em que se regista a acumulação anormal de proteínas muito diversificadas devido a uma tendência partilhada para a excessiva agregação (Dafforn *et al.*, 1999; Carrell e Lomas, 2002). Muitos dos conceitos e metodologias envolvidos na caracterização molecular da fisiopatologia do défice de α1-antitripsina e no seu diagnóstico são comuns as inúmeras outras doenças de determinismo genético simples. Por outro lado, a variedade das manifestações clínicas que lhe estão associadas e a importância de factores adicionais, hereditários ou ambientais, no desencadeamento ou gravidade dessas doenças, conferem à deficiência da α1-antitripsina características próprias de algumas patologias de determinismo complexo cujas bases genéticas têm sido recentemente estudadas. O elevado grau de variação genética da α1-antitripsina permite ainda investigar questões relacionadas com a sua diversidade populacional que são comuns a muitos outros *loci* polimórficos, tais como a comparação dos padrões de distribuição de alelos patogénicos ou não patogénicos e a análise do contexto evolutivo que influenciou a heterogeneidade geográfica observada nessas distribuições. Por fim, a localização do gene da α1-antitripsina na vizinhança de genes que lhe são homólogos, cria uma oportunidade para a realização de estudos sobre a diversidade haplotípica e a organização da variabilidade genética em grupos de *loci* filogeneticamente relacionados.

O estudo da variação genética da α1-antitripsina realizado neste trabalho foi dividido em duas partes. Na primeira parte, procurou-se elucidar a história natural do polimorfismo através da análise dos padrões de diversidade haplotípica associados aos seus alelos mais comuns em diferentes populações. Para tal, seleccionaram-se microssatélites localizados em diferentes regiões da família de genes SERPIN do cromossoma 14q32.1 e as respectivas distribuições de frequências alélicas foram ancoradas nas linhagens definidas pelas mutações dos produtos génicos da PI. Com esta abordagem, os microssatélites puderam ser usados como relógios moleculares capazes de registar a diversidade acumulada por cada um dos alelos da α1-antitripsina ao longo do seu percurso evolutivo. A informação assim obtida foi analisada no âmbito da história das populações estudadas e no contexto filogenético definido pelas mutações pontuais características de cada alelo. Reciprocamente, a repartição da variação de microssatélites por segmentos genómicos com diferentes idades permitiu analisar aspectos da dinâmica evolutiva deste tipo de marcadores. O essencial dos resultados deste estudo é apresentado no artigo 2. A integração desses resultados com novos dados obtidos depois da publicação deste trabalho e a discussão mais detalhada de abordagens alternativas ou complementares às interpretações que nele foram expostas são apresentadas, posteriormente, sob a forma de um comentário.

A segunda parte do estudo foi dedicada à caracterização do espectro mutacional da α1-antitripsina e resultou de um esforço de identificação das bases moleculares de alelos raros que foram, na sua maioria, detectados no decurso da fenotipagem da proteína como meio auxiliar de diagnóstico. Apesar da sua baixa frequência, estes alelos são numerosos e muito heterogéneos, pelo que a sua caracterização é fundamental para evitar os problemas de diagnóstico muitas vezes associados a estratégias de rastreio que apenas se centram na detecção das mutações típicas das variantes mais comuns. Por outro lado, a caracterização dos alelos raros, normais ou patogénicos, também é essencial para a localização das regiões funcionalmente mais importantes da proteína, para o conhecimento da variedade dos mecanismos que provocam a sua deficiência e para a avaliação das consequências clínicas que lhes estão associadas. As contribuições já publicadas no âmbito deste estudo encontram-se condensadas nos artigos 3 a 6. Estes artigos são apresentados depois de uma discussão

mais alargada em que se procurou integrar os resultados obtidos na análise das propriedades gerais do espectro mutacional da α1-antitripsina.

2. Resultados e Discussão

2.1 Diversidade genética da α1-antitripsina e história natural do polimorfismo

Artigo 2

Seixas, S., Garcia, O., Trovoada, M. J., Santos, M. T., Amorim, A., Rocha, J. (2001). Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the α1-antitrypsin polymorphism. *Hum Genet* **108**:20-30.

Hum Genet (2001) 108:20–30
DOI 10.1007/s004390000434

ORIGINAL INVESTIGATION

Susana Seixas · Oscar Garcia · M. Jesus Trovoada
M. Teresa Santos · António Amorim · Jorge Rocha

Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the α1-antitrypsin polymorphism

Received: 9 August 2000 / Accepted: 27 October 2000 / Published online: 9 December 2000

Abstract The levels of haplotype diversity associated with different α1-antitrypsin (PI) alleles were assessed by the analysis of three microsatellites located within or close to corticosteroid-binding globulin (CBG), α1-antitrypsin [PI-$(TG)_n$] and protein C inhibitor [PCI-$(TG)_n$] loci in three populations with different historic backgrounds: Portugal, the Basque Country and São Tomé e Príncipe (Gulf of Guinea). Unlike the more distant PCI-$(TG)_n$ repeat, allelic variation at PI-$(TG)_n$ reflected distinct phases of mutational recovery of microsatellite diversity around different founder alleles and showed a considerable differentiation between α1-antitrypsin protein variants. In accordance with population history, the Basque sample presented overall reduced levels of microsatellite variation. The African sample, although presenting the highest PCI-$(TG)_n$ diversity, showed a lineage-specific reduction in PI-$(TG)_n$ heterozygosity within the oldest M1Ala213 variant that could have been caused by (1) selection at a closely linked locus or (2) biases in the microsatellite mutation process leading to a stable equilibrium distribution. Age estimates of α1-antitrypsin variants based on microsatellite variation suggest that the Z deficiency allele appeared 107–135 generations ago and could have been spread in Neolithic times. The S mutation has an older 279- to 470-generation age, indicating that its high frequencies in Iberia did not result from a recent bottleneck and that PI*S could have originated in this region. M2 and M3 types had lower age estimates than would be expected from their wide geographical distributions, suggesting that their dispersion in Europe might have been preceded by important bottlenecks.

S. Seixas · A. Amorim · J. Rocha (✉)
Instituto de Patologia e Imunologia Molecular
da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n,
4200 Porto, Portugal
e-mail: jrocha@ipatimup.pt,
Tel.: +351 225570700, Fax: +351 225570799

S. Seixas · M. T. Santos · A. Amorim · J. Rocha
Faculdade de Ciências, Universidade do Porto, Porto, Portugal

O. Garcia
Area de Laboratorio de la Ertzaintza, Sección de Biología,
Bilbao, Spain

M. J. Trovoada
Departamento de Antropologia, Universidade de Coimbra,
Coimbra, Portugal

Introduction

Human α1-antitrypsin (PI) is a glycoprotein that is mainly synthesised in the liver and functions as the major inhibitor of neutrophil elastase in the lower respiratory tract. The PI gene is located on chromosome 14q32.1, within a cluster of related serine protease inhibitor (serpin) genes that spans approximately 320 kb and additionally includes a α1-antitrypsin pseudogene sequence (PIL) and the loci coding for α1-antichymotrypsin (ACT), corticosteroid-binding globulin (CBG), protein C inhibitor (PCI) and kallistatin (PI4; Sefton et al. 1990; Billingsley et al. 1993; Byth et al. 1994; Rollini and Fournier 1997; Fig. 1).

PI has a wide spectrum of protein variants that can be distinguished by isoelectric focusing including several rare variants and six common alleles that reach polymorphic frequencies in different human populations: M1, M2, M3, M4, S and Z. Sequence characterisation of these alleles has led to the differentiation of two subtypes of M1 (M1Ala213 and M1Val213) and has shown that common M variants differ in their amino acid composition at codons 101, 213 and 376 (reviewed in Brantly et al. 1988). Interspecies sequence comparisons have shown that PI molecules from non-human primates possess Ala at the position corresponding to human codon 213 suggesting that common M alleles originated according to the following pathway: M1Ala213 (Arg101-Ala213-Glu376) → M1Val213 (Arg101-Val213-Glu376) → M3 (Arg101-Val213-Asp376) → M2 (His101-Val213-Asp376). Within this framework, the M4 allele (His101-Val213-Glu376) could have been derived either from M1Val213 or M2 or from intragenic recombination between both, whereas the Z variant was derived from a Glu342Lys substitution on

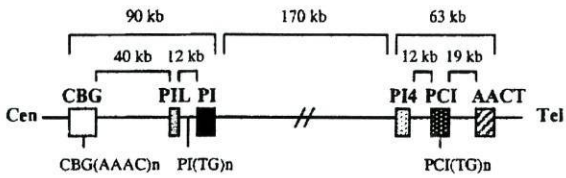


Fig. 1 Map of the serpin gene cluster showing the relative locations of the microsatellites used in the haplotype studies. Distances are as in Rollini and Fournier (1997). The position of PI-$(TG)_n$ is as in BAC R-349I1 (GenBank Accession no. AL132708)

an M1Ala213 background and PI*S resulted from a Glu264Val replacement on a M1Val213 ancestral allele (Brantly et al. 1988; Nukiwa et al. 1996).

Whereas the PI M alleles have normal circulating protein concentrations and are widely dispersed, the S and Z types have serum levels 60% and 15% of normal, respectively, and are virtually restricted to populations of European descent (for a review, see Cox 1995). PI deficiency caused by homozygosity for PI*Z is one of the most common genetic disorders among Europeans and is associated with an increased risk for early onset pulmonary emphysema in adults and liver disease in children (Cox 1995). Since the frequency of PI*Z is highest in Scandinavia (>0.015), it has been suggested that its present distribution resulted from a spreading process along a north to south gradient in European populations (Cox et al. 1985; Hutchison 1998). On the basis of the observation of the highest PI*S frequencies in Portugal and Spain (>0.090), the Iberian Peninsula has been considered the most likely place of origin for the S mutation (Hutchison 1998).

Recently, several microsatellite polymorphisms have been described for each of the serpin genes (Byth and Cox 1993a, 1993b; Byth et al. 1993, 1994). The use of these polymorphisms to assess the haplotype homogeneity of the Z deficiency allele in northern Europe has shown a considerable amount of allelic association throughout the cluster, indicating that Z gene products are derived from a single mutation that has been dated to approximately 66 generations ago (Byth et al. 1994). However, much less is known about the relative antiquity of the more dispersed, common protein variants.

Using a highly polymorphic dinucleotide repeat close to the PI gene, we have previously shown that the levels of microsatellite variation within both normal and disease-associated PI alleles in the Portuguese population are in good agreement with their sequence of origin and therefore represent a useful tool for studying the evolution and dispersion of the different PI variants (Rocha et al. 1997). In an attempt to provide a deeper insight into the natural history of the PI polymorphism, we have used microsatellites located within or close to the PI, CBG and PCI loci to perform an extended evaluation of haplotype diversity among PI alleles from three populations with different historic backgrounds: Portugal, the Basque Country and São Tomé e Príncipe (Gulf of Guinea).

Materials and methods

Sample composition

The Portuguese sample included 318 unrelated individuals from the northern part of the country and families with at least one member for whom PI typing was requested to confirm protein deficiency after low serum level determinations. These were a superset of the data presented in a previous study (Rocha et al. 1997), totalling 528 unrelated PI chromosomes (60 M1Ala213, 123 M1Val213, 32 M3, 97 M2, 6 M4, 103 S and 107 Z variants). The Basque sample included 97 unrelated autochthonous individuals from the provinces of Biscay and Guipuzcoa whose eight immediate forebears had Basque surnames and whose four grandparents were natives of the Basque Country. The sample from the São Tomé e Príncipe archipelago, a former Portuguese colony lying approximately 240 km off the coast of Gabon, in the Gulf of Guinea, included 189 unrelated individuals from various places on the major island of São Tomé.

Blood samples were collected by venipuncture into EDTA tubes. DNA was prepared by using standard phenol-chloroform extraction protocols after plasma separation for PI typing.

PI typing

The different PI protein variants were discriminated by means of polyacrylamide gel hybrid isoelectric focusing (Rocha et al. 1997). The Ala213/Val213 and Arg101/His101 variation to distinguish M1 subtypes and M3 (Arg 101) and M4 (His101) alleles, respectively, was analysed by polymerase chain reaction/restriction fragment length polymorphism (PCR-RFLP) methods according to Rocha et al. (1997) and Faber et al. (1990).

Microsatellite typing

Haplotype heterogeneity associated with the different PI alleles was assessed by the analysis of three microsatellites (Fig. 1): (1) a $(TG)_n$ repeat that was originally located upstream the PI gene, 3.6 kb from the start of exon IC (Byth and Cox 1993a) but that appears to lie 7.1 kb from its 3' end in a recently released genomic sequence of a bacterial artificial chromosome (BAC) from human chromosome 14 (BAC R-349I1; GenBank Accession no. AL132708); (2) a $(TG)_n$ repeat located in the first intron of the PCI gene (Byth et al. 1993; GenBank Accession no. M68516); (3) a $(AAAC)_n$ repeat from a bacteriophage clone containing the CBG gene (Byth et al. 1994).

PI and PCI microsatellites were amplified with fluorescently labelled primers and separated in a ABI 310 DNA sequencer. CBG repeat PCR products were separated by horizontal electrophoresis in 9% polyacrylamide gels and visualised by silver staining. All PCR amplifications were performed as described (Byth and Cox 1993a; Byth et al. 1993, 1994). Direct comparison of a common set of Centre d'Étude du Polymorphisme Humain (CEPH) reference individuals (10201, 10202, 88401, 88402) has shown that the apparent sizes of microsatellite alleles are 22 bp and 11 bp shorter than those initially reported for PI and PCI loci, respectively (Byth and Cox 1993a; Byth et al. 1993). For comparison purposes, we have chosen to keep the original allele nomenclatures but will refer to the allele lengths determined according to our sizings.

DNA sequencing

DNA sequencing of rare PI alleles was performed after PCR amplification of all coding exons (II–V) as previously described (Seixas et al. 2000). Sequencing of PI microsatellite alleles was carried out by cloning their corresponding PCR products into a pCR4 vector with the TOPO TA cloning kit for sequencing (Invitrogen). The products of ligation were used to transform One Shot TOP 10 chemically competent *Escherichia coli* cells and sequenc-

ing reactions were then performed by using the ABI Prism Big Dye DydeoxyTerminator Cycle sequencing kit. Sequencing products were analysed in an automatic DNA sequencer (ABI 377).

Data analysis

Allele frequencies at the individual loci were calculated by direct gene counting. In the Portuguese sample, haplotypes were determined either directly in homozygous individuals or by two-generation family studies. In the Basque Country and São Tomé samples, maximum-likelihood estimates of haplotype frequencies were computed by using an expectation-maximisation algorithm (Dempster et al. 1977). Following an approach similar to that reported in Bosch et al. (1999), the degree of structuring of microsatellite variation within PI alleles was assessed by an analysis of molecular variance (AMOVA), with either Fst or Rst as measures of dissimilarity. All these analyses were performed with the ARLEQUIN package (Schneider et al. 1997). Unbiased estimates of expected heterozygosity were calculated according to Nei (1987); 95% confidence limits of these estimates were established by 10,000 bootstrap simulations with GENETIX software (Belkhir et al. 1998), as described in Carvalho-Silva et al. (1999).

To estimate the age G of the most recent common ancestor of haplotypes from the different PI variants, we simulated the over-time decay in linkage disequilibrium in a population of infinite size and calculated the microsatellite allele frequency distribution in each generation by using the following relation:

$$p_{(g,i)}=p_{(g-1,i)}\,(1-\mu-\theta)+\theta q_i+(\mu/2)\,[p_{(g-1,i-1)}+p_{(g-1,\,i+1)}] \quad (1)$$

where $p_{(g,i)}$ is the frequency of a marker microsatellite allele with i repeats in generation g within a given PI variant, q_i is the frequency of that allele in the whole population, and θ and μ represent the recombination fraction and the microsatellite mutation rate, respectively. The number of generations leading to a match with the observed frequency of the ancestral microsatellite allele within each protein variant was chosen as the estimate of G.

Results

α1-Antitrypsin protein variation

The frequencies of various PI alleles in Portugal, the Basque Country and São Tomé are presented in Table 1. In accordance with previous results from other populations of West-African origin (DeCroo et al. 1991), the allele frequency distribution in São Tomé is characterised by the predominance of M1 types and by much lower frequencies of M2 and M3 than in most Eurasian populations. The frequency of M1Ala213, which is though to be the most ancient PI allele, represents 73% of all M1 alleles in São Tomé, a considerably higher proportion than that found in the Iberian samples and other European and Asiatic populations where the M1Val213 allele predominates (Cox and Billingsley 1986; Nukiwa et al. 1996). A higher frequency of M1Ala213 (0.55) over M1Val213 (0.45) has been also observed by Gaillard et al. (1994) in black South Africans.

The major differences between the Iberian samples are attributable to the reduction of the M1Ala213 frequency in the Basque Country and an increase in the frequency of M1Val213, which represents 84% of the PI M1 types, whereas the Portuguese sample has a 66% percentage, similar to the previously reported values in other populations of European origin (Cox and Billingsley 1986).

Table 1 PI allele frequency distributions in Portugal, the Basque Country and São Tomé (n number of individuals)

Alleles	Populations		
	Portugal n=318	Basque Country n=97	São Tomé n=189
M1Ala213	0.194	0.090	0.661
M1Val213	0.377	0.482	0.246
M2	0.200	0.216	0.019
M3	0.072	0.098	0.052
M4	0.010	0.021	–
S	0.124	0.072	0.008
Z	0.011	0.015	–
I	0.002	–	–
T	–	–	0.005
V	–	–	0.003
Pdonauwoerth	–	–	0.003
Pro362His	–	–	0.003
Arg281del	–	0.005	–

Apart from the common PI alleles, different rare variants have been also identified and subsequently characterised by DNA sequencing. Arg281del and Pro362His have only been found so far in the Basque Country and in São Tomé, respectively, and have been characterised elsewhere (Seixas et al. 1999a, 2000). PI*I, PI*T, PI*Pdonauwoerth and PI*V have also been reported in other populations of European descent. The occurrence of PI*T and PI* Pdonauwoerth alleles in São Tomé is probably a result of admixture, since they share the same microsatellite alleles at the closest PI and CBG loci with variants drawn from our Portuguese panel of rare alleles (results not shown). PI*V was never found in the Portuguese material and probably has an independent African origin, since it has been found to reach polymorphic frequencies in South African blacks and is absent in a sample from the white population from the same region (Halkas et al. 1998).

Microsatellite variability within common PI alleles

In Fig. 2, the most relevant data on the distribution of CBG microsatellite alleles in the three populations studied are summarised and compared with the information previously reported from a sample of northern European origin (Byth et al. 1994). The CBG tandem repeat behaves as a stable bi-allelic polymorphism in all populations (Fig. 2A). Whereas, in the African sample, these alleles are evenly distributed within the different PI types, the CBG-1 (90 bp) allele occurs mainly among M1Val213 in the Iberian populations where it defines a common M1Val213-CBG-1 subtype (Fig. 2B). As in northern European populations (Byth et al. 1994), 95% of the Portuguese Z types have been found to be associated with the rare CBG-1 allele, whereas variant haplotypes are associated with CBG-2 (86 bp) and have probably resulted from recombination between PI and CBG (Fig. 2B). This further confirms that PI*Z originated from a single mutational event, which

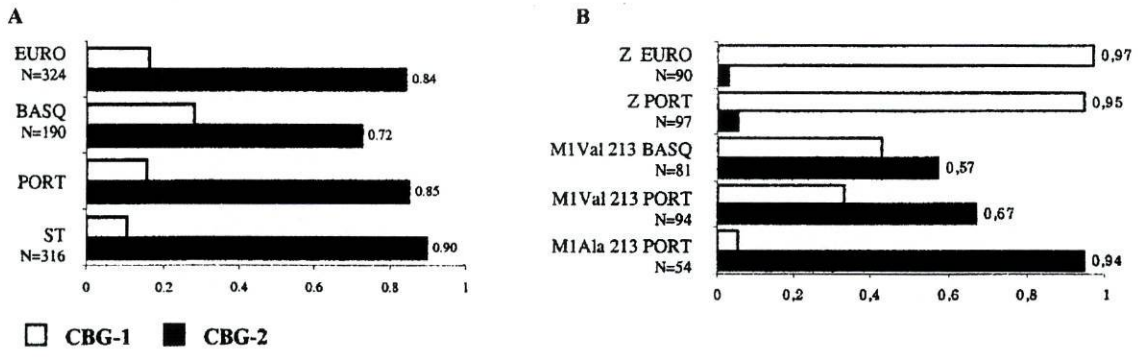


Fig. 2 **A** Comparison of CBG microsatellite allele frequencies in northern Europe (*EURO*), the Basque Country (*BASQ*), Portugal (*PORT*) and São Tomé (*ST*). Portuguese distributions were estimated by weighting each CBG allele frequency with PI frequencies from Table 1. **B** CBG microsatellite allele frequencies within PI alleles from various populations (*N* number of chromosomes)

most probably occurred in a very rare haplotype, since CBG-1 is infrequent within the putative M1Ala213 ancestral allele (Fig. 2B).

The overall allele frequency distributions for the PI-$(TG)_n$ and PCI-$(TG)_n$ microsatellites in the three populations are presented in Fig. 3. The Portuguese distributions are similar to those previously reported from unrelated European individuals taken from CEPH pedigrees (Byth and Cox 1993a; Byth et al. 1993). Compared with the Portuguese sample, the Basques show a consistent trend of reduced diversity at both loci, with higher homozygosities, lower repeat number variances and increased frequencies of the most common allele. Interestingly, PI-$(TG)_n$ in the African population has a smooth unimodal allele distribution with reduced diversity, although the PCI-$(TG)_n$ locus presents the highest heterozygosity.

When PI-$(TG)_n$ allele frequency distributions are split by PI gene product, there are considerable differences between the patterns of allelic diversity within the protein variants in each population (Fig. 4). In contrast, PCI-$(TG)_n$ allele distributions within the different protein variants are generally similar to the pattern of variation observed in each population as a whole (results not shown). This is reflected in the results from the analysis of molecular variance, which show that, independently of the measure used to assess dissimilarity, the fractions of microsatellite variation that can be attributed to differences among PI alleles are consistently higher for PI-$(TG)_n$ than for PCI-$(TG)_n$ (Table 2).

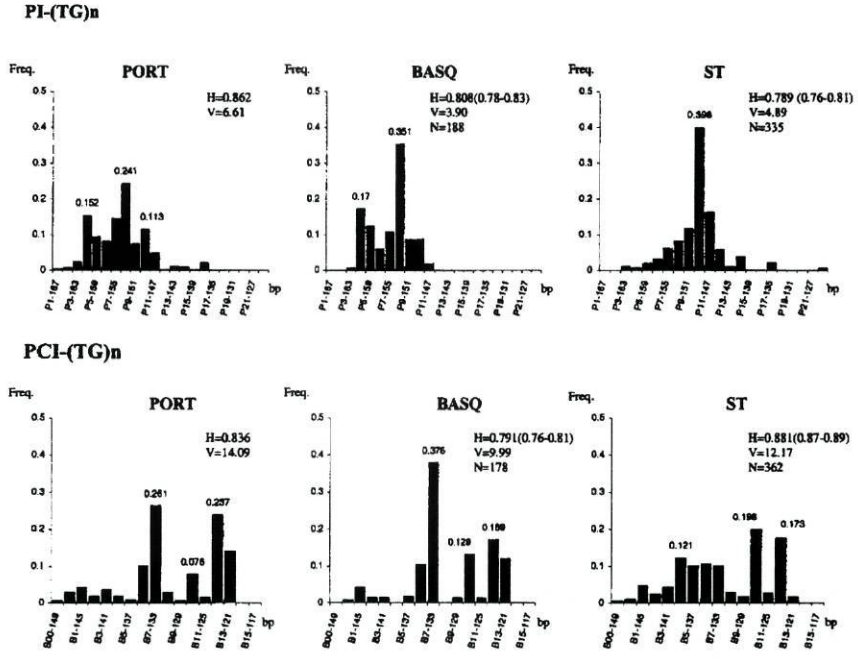


Fig. 3 Allele frequency distributions of PI-$(TG)_n$ and PCI-$(TG)_n$ microsatellites in Portugal (*PORT*), the Basque Country (*BASQ*) and São Tomé (*ST*). The Portuguese distributions were estimated as in Fig. 2 (*H* heterozygosity, *V* variance in repeat number, *N* number of chromosomes)

24

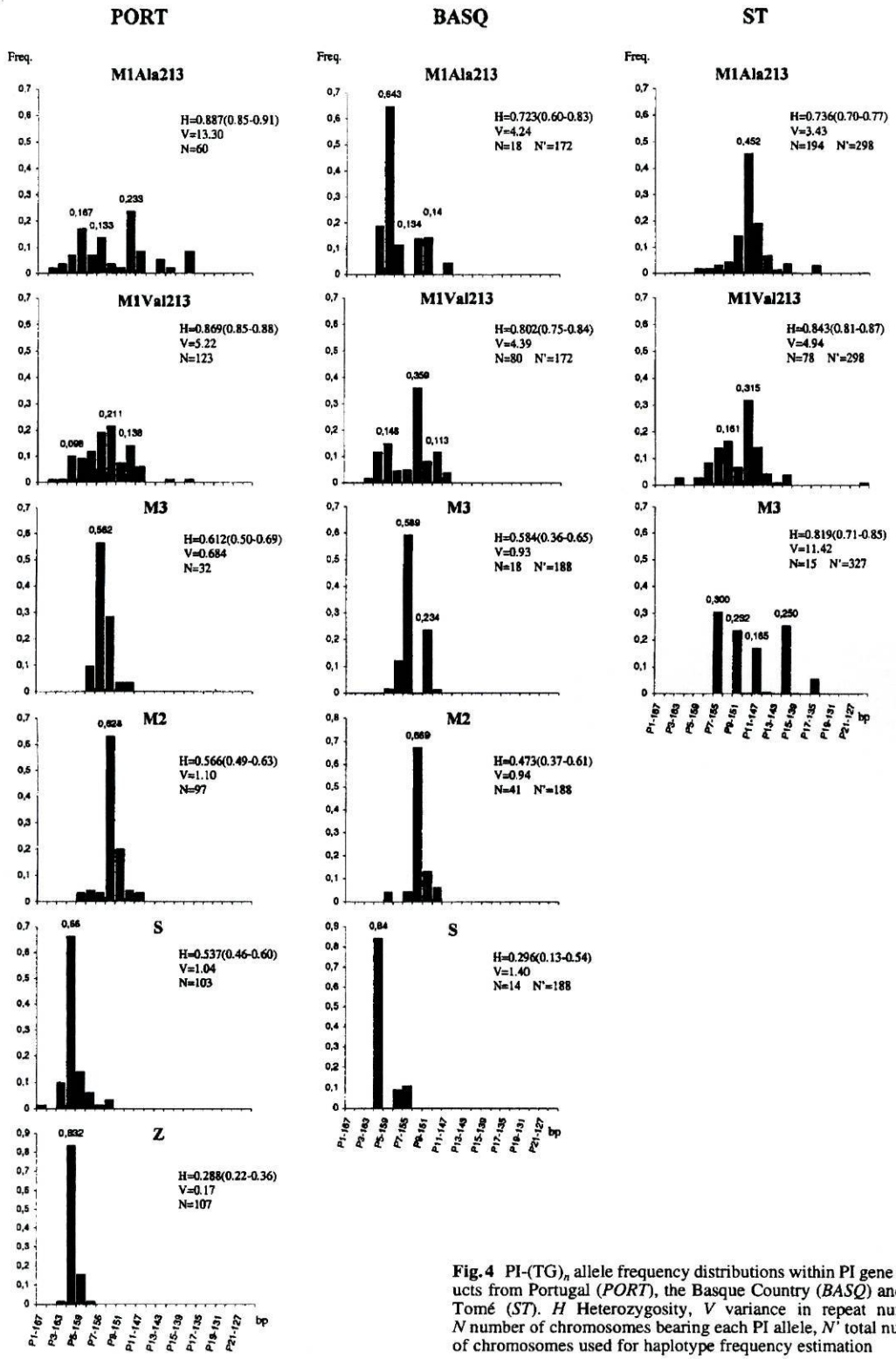


Fig. 4 PI-$(TG)_n$ allele frequency distributions within PI gene products from Portugal (*PORT*), the Basque Country (*BASQ*) and São Tomé (*ST*). *H* Heterozygosity, *V* variance in repeat number, *N* number of chromosomes bearing each PI allele, *N'* total number of chromosomes used for haplotype frequency estimation

Table 2 Fractions of the total genetic variation that can be attributed to differences in microsatellite allele distributions among PI gene products from the three sampled populations

Microsatellite	Percent fractions of genetic variation among PI alleles in the three samples[a]					
	Portugal		Basque Country		São Tomé	
	Fst[b]	Rst[c]	Fst	Rst	Fst	Rst
PI	28.86 (P<0.001)	53.47 (P<0.001)	23.81 (P<0.001)	30.77 (P<0.001)	5.36 (P<0.001)	11.95 (P<0.001)
PCI	1.81 (P<0.001)	3.54 (P=0.018)	8.95 (P<0.001)	11.39 (P=0.003)	1.45 (P=0.009)	−0.63[d] (P=0.564)

[a]As measured by AMOVA
[b]Fraction of the total variance in microsatellite allele frequencies attributable to differences among PI gene products
[c]Fraction of the total variance in microsatellite repeat number attributable to differences among PI gene products
[d]Negative variance components should be regarded as zero

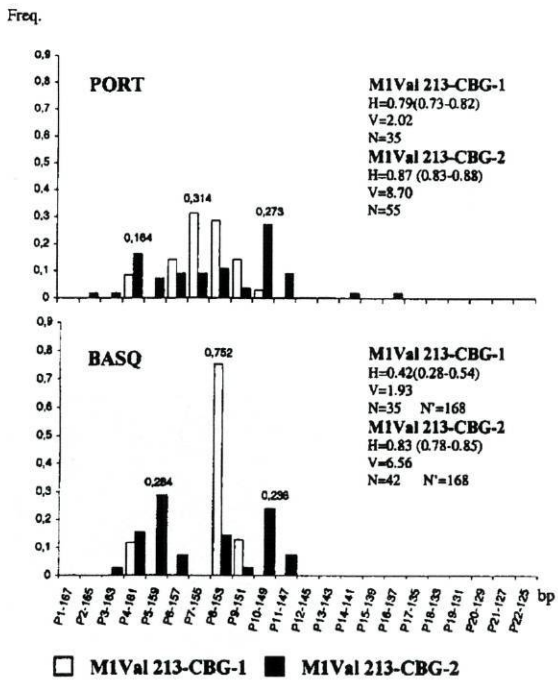


Fig. 5 PI-$(TG)_n$ allele frequency distributions within M1Val213 sublineages defined by CBG alleles in the Portuguese (*PORT*) and Basque (*BASQ*) samples (*H* heterozygosity, *V* variance in repeat number, *N* number of chromosomes bearing each CBG-PI haplotype, *N'* total number of chromosomes used for haplotype frequency estimation)

A comparison of PI-$(TG)_n$ allele frequencies among M1 variants from the three populations shows that P10 (149-bp) is the shared allele that remains common in most distributions and probably represents either the ancestral type originally associated with the M1Ala213 and M1Val213 lineages or the most stable of the PI-$(TG)_n$ length variants. M1Ala213 from the Portuguese sample has the highest heterozygosity and repeat number variance and presents a multimodal distribution that shares a peak at the P10 allele with the African sample (Fig. 4). In the Basque Country, M1Ala213 has a subset of the Portuguese microsatellite allele distribution and shows a derived pattern characterised by the loss of the ancestral P10 allele. In the African sample, M1Ala213 also has markedly reduced levels of variation but shows a regular pattern compatible with the stepwise accumulation of mutations around the hypothetical ancestral modal P10 allele (Fig. 4).

M1Val213 lineages present less divergent diversity levels between the three populations and have similar repeat number variances and overlapping ranges of heterozygosities (Fig. 4). In the Iberian samples, M1Val213 distributions can be further subdivided according to the associated allele at the CBG locus (Figs. 2B, 5). M1Val213-CBG-2 is likely to be the older sublineage and exhibits, in both populations, a more diverse bimodal distribution, which shares a frequent ancestral P10 allele with the African sample. M1Val213-CBG-1 is less diverse and presents more pronounced differences between the two samples, with the Basque Country showing a marked predominance of the P8 (153 bp) allele.

Unlike the M1 types, the distributions of microsatellite variation among M3, M2 and S variants in the Iberian samples generally have more regular patterns, typically centred around one modal allele, shared by both populations. The modal P4 (161 bp) microsatellite allele from Portuguese Z types is also strongly predominant in Z variants from northern Europe (Byth et al. 1994) and occurs among the only three Z-bearing phenotypes observed in the Basque sample, whereas the predominant P8 allele in Iberian M2 variants is also present in all seven M2-bearing phenotypes observed in the African sample. The M3 variant from São Tomé has a different PI-$(TG)_n$ distribution characterised by a highly irregular allele spectrum (Fig. 4).

DNA sequences of PI microsatellite alleles

In order to investigate whether the observed differences in the diversity levels of the PI microsatellite could be related to specific structural features affecting the mutation process, we sequenced several microsatellite alleles from

the Portuguese and São Tomean samples: three P22 (125 bp), one P16 (137 bp), five P10 (149 bp), three P9 (151 bp), one P8 (153 bp), seven P7 (155 bp), one P5 (159 bp) and six P4 (161 bp). Sequence information confirmed that the allele sizes in this microsatellite are 22 bp shorter than those originally reported by Byth and Cox (1993a). All microsatellite alleles presented an interrupted $(TG)_nAT(TG)_5$ structure and no differences were observed between the base composition of the flanking regions. The P7 (155 bp) allele was found to have a $(TG)_{23}AT(TG)_5$ structure identical to the sequence found in BAC R-349I1 from human chromosome 14 (GenBank Accession no. AL132708).

Age estimates of α1-antitrypsin variants

Table 3 shows the age estimates for PI variants presenting more regular microsatellite distributions in the Iberian populations (M3, M2, S and Z). Calculations for northern Europe were performed with data taken from Byth et al. (1994). PI-$(TG)_n$ haplotype diversification was assumed to be entirely caused by stepwise mutation and an average figure of 0.001 was used as the mutation rate (Weber and Wong 1993). Estimates for the Z allele based on CBG variation were made by assuming that recombination was the only factor leading to the breakdown of linkage disequilibrium, by using the general relationship 1000 kb= 1 cM. Calculations with the PCI-$(TG)_n$ marker took into account the effects of both mutation and recombination. In addition to the genetic distance calculated from the standard 1 cM to 1000 kb translation, a 2 cM recombination fraction between PI and PCI was also used. This was based on previous studies derived from family data, which suggested that the levels of recombination between the CBG-PIL-PI and the PI4-PCI-ACT subclusters could be higher than would be expected from their physical distance (Sefton et al. 1990; Byth et al. 1994; GDB Mapview http://www.gdb.org).

Although values derived from CBG tend to be lower than those calculated from PI-$(TG)_n$, age estimates for PI*Z based on these loci are globally concordant in the Portuguese and northern European samples, leading to an average estimate of approximately 4070 years, if a mean value of 30 years per generation is assumed (Tremblay and Vezina 2000). Results based on PCI-$(TG)_n$ show larger differences between the estimates from the two European samples. Byth et al. (1994) have found that 41% of northern European Z chromosomes share a unique extensive haplotype encompassing the whole serpin gene cluster and defined by alleles 1 (90 bp), P4 (161-bp) and B13 (121-bp) at the CBG, PI and PCI microsatellite loci, respectively. However, we have found the same predominant haplotype only in 21 of 79 (27%) Z chromosomes in which the three locus haplotypes could be defined (results not shown), leading to older age estimates in the Portuguese sample (Table 3). This discrepancy could have resulted from an early recombination or a multistep mutation leading to additional founder PCI-$(TG)_n$ haplotypes that could have persisted only in the Portuguese sample; neverthelss, no further predominant microsatellite allele could be found in this population (results not shown). When the data from PCI-$(TG)_n$, PI-$(TG)_n$ and CBG are combined, average estimates of 6070 and 3210 years are obtained, depending on whether the recombination rate between PCI and PI is calculated from physical distance or from family data, respectively. The use of the 2-cM distance leads to PCI-$(TG)_n$ age estimates closer to the values obtained from CBG (Table 3), thus adding indirect evidence supporting previous suggestions that the region between the two serpin subclusters recombines in excess of the genome average (Sefton et al. 1990).

Age estimates for PI*S based on PI-$(TG)_n$ are higher in Portugal (14,100 years) than in the Basque sample (5490 years), which showed substantially lower levels of microsatellite diversity, consistent with the general pattern observed in other PI variants (Fig. 4). Estimates based on

Table 3 Estimates of the number of generations to the most recent common ancestral of haplotypes from different PI variants

PI variants	Sample	Estimates based on different marker loci			
		CBG[a]	PI[b]	PCI[c]	PCI[d]
Z	Portugal	102 (20–187)[e]	194 (104–297)	448 (322–657)	66 (48–95)
	Northern Europe	61 (0–145)	185 (73–319)	224[f]	34[f]
S	Portugal	–	470 (312–672)	619 (451–973)	88 (65–139)
	Basque Country	–	183 (0–511)	–	–
M2	Portugal	–	536 (351–782)	–	–
	Basque Country	–	453 (216–807)	–	–
M3	Portugal	–	739 (360–1489)	–	–
	Basque Country	–	625 (210–1573)	–	–

[a]Assuming θ=0.0006, μ=0 and 60 kb between CBG and PI
[b]Assuming θ=0 and μ=0.001
[c]Assuming θ=0.002, μ=0.001 and 200 kb between PCI and PI
[d]Assuming θ=0.02, μ=0.001
[e]Support bounds calculated by using ±2× the standard deviation of $p_{(g,i)}$ in equation (1)
[f]No data available for support interval calculation

PCI-$(TG)_n$ were derived from the B10 (127 bp) allele, which was found to be shared by 22% of the Portuguese S alleles and had a 0.073 frequency in the whole population (data not shown). Combination of the values obtained from both PCI-$(TG)_n$ and PI-$(TG)_n$ loci leads to estimates of 16,335 and 8370 years for PI*S in the Portuguese population, according to the two different values used for the recombination rate between PI and PCI (Table 3). In the Basque sample, no B10 alleles were found among the S variant; B14 (119 bp) and B6 (135 bp) were the predominant PCI-$(TG)_n$ alleles. This discrepancy could have resulted from the loss of the ancestral microsatellite allele in the Basque population caused by drift.

Age estimates for PI*M2 were found to be very close to PI*S in the two Iberian populations and showed overlapping support intervals with the M3 dates. Both variants showed PCI-$(TG)_n$ distributions similar to the whole population confirming that the traces of ancestral types in this locus are lost much more rapidly than in PI-$(TG)_n$ as indicated by AMOVA (Table 2).

Discussion

Patterns of microsatellite variation within α1-antitrypsin alleles

We have characterised the patterns of haplotype variation within the lineages defined by α1-antitrypsin alleles in three populations with different histories. São Tomé e Príncipe was uninhabited before its discovery by Portuguese sailors in the early 1470s and began to be peopled by slaves brought from different regions of Africa, becoming an important stepping-stone in the massive relocation of African populations promoted by the Atlantic slave trade. This settlement pattern is reflected in the retention of the globally high levels of mitochondrial and nuclear DNA diversity that are generally observed in the continental mainland (Mateu et al. 1997; Albarrán et al. 1998; Prata et al. 1996; Seixas et al. 1999b). Portugal and the Basque Country are two geographically close populations that have been differently exposed to cultural and genetic exchange, with the Basques remaining substantially more isolated and maintaining a remarkable linguistic differentiation. This is reflected in the position of the two populations in the genetic tree of Europe where the Portuguese are clustered together with most other populations, whereas the Basques are outliers, although they have noticeable similarities with their neighbours (Cavalli-Sforza et al. 1994).

Our survey of serpin microsatellite diversity has revealed substantial differences in the patterns of allelic variation both between populations and between protein variants within each population. Differences in PI-$(TG)_n$ distributions among PI variants reflected distinct levels of variability restoration around different founder alleles and can be used for age estimation. In addition, since microsatellite distributions are anchored in point mutations with a different age, the analysis of microsatellite variation within PI alleles provides a set of natural replicates that allows the study of the interplay between mutation processes and population history along different stages of microsatellite evolution. In contrast, the more distant PCI-$(TG)_n$ locus presents a much lower level of diversity compartmentalisation indicating that recombination between the two serpin subclusters is sufficiently high to wipe out the diversity differences between PI gene products with a different age.

As expected, the Iberian populations presented the most similar allele distributions, although the Basques showed a consistent pattern of reduced diversity at both PI-$(TG)_n$ and PCI-$(TG)_n$ loci and had a subset of Portuguese genetic variation, in accordance with the pattern of restricted gene flow and increased drift that is believed to have produced their present genetic peculiarity (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994).

The African population, although presenting the highest PCI diversity, showed unexpectedly reduced levels of variation at the PI-$(TG)_n$ locus, especially within the oldest M1Ala213 allele. Unlike the Basque sample, this locus- and lineage-specific reduction is difficult to explain by local population history and might have resulted either from (1) selection at a closely linked locus or (2) special features of the microsatellite mutation process.

If positive selection were favouring an advantageous allele at any locus tightly linked to PI microsatellite, the allele at the microsatellite locus that was initially linked to the advantageous variant would become predominant, leading to reduced levels of genetic variability (Slatkin 1995; Schlötterer and Wiehe 1999). A possible target for selection could be the PI gene itself and, more specifically, the M1Ala213 allele. However, the type of effect that could account for the observed microsatellite distribution would imply an increase in the frequency of a previously rare variant. As M1Ala213 is likely to be the primitive human PI allele, any hypothetical selective advantage over other normal variants would not be linked to a specific reduction in microsatellite variation. CBG is the next closest active serpin gene and has been suggested to be implicated, through linkage disequilibrium, in inflammatory diseases that appear to be associated with the PI Z deficiency allele (Billingsley et al. 1993; Byth et al. 1994). Whereas this points to a possible link between CBG functional variants and PI alleles that could lead to lineage-specific selective effects within the CBG-PIL-PI subcluster, no systematic mutational screening in the CBG gene has been performed so far to establish this concept.

As an alternative to selection, the PI-$(TG)_n$ allele distribution of M1Ala213 in São Tomé could have resulted from increased mutational stability of the modal P10 (149 bp) allele. There is increasing evidence that microsatellite alleles from the same locus can have different mutation rates and that these can be influenced by factors such as the number of repeats or the regularity of the sequence structure of the repeat blocks (Jin et al. 1996; Goldstein and Pollock 1997; Brinkmann et al. 1998; Carvalho-Silva et al. 1999). We have found no difference be-

tween the structures of the PI-$(TG)_n$ alleles. However, pedigree data, population surveys and theoretical analyses of microsatellite variation suggest that, even in the absence of structural differences, allele size constraints caused by centrally directed mutation biases can lead to stationary allele distributions (Garza et al. 1995; Nauta and Weissing 1996; Amos 1999; Xu et al. 2000). PI microsatellite distribution in the M1Ala213 lineage from São Tomé could thus represent an example of such a stationary distribution resulting from mutation biases leading to an increased mutational stability at the critical modal P10 (149 bp) allele.

The finding of equilibrium distributions caused by size constraints would depend on the combination of large population effective sizes and increased age of the lineages with which microsatellite variation is associated. Generally, the effects of genetic drift are thought to be more important than mutation in the shaping of microsatellite distributions in different human populations (Pérez-Lezaun et al. 1997). In addition, the origin of new lineages defined by recent unique mutation events resets microsatellite variation to zero, so that younger variants would be further from mutational equilibrium than the older ones. Since African populations have been less subjected than non-African human groups to genetic drift events (Tishkoff et al. 1996), stable equilibrium distributions shaped by mutational patterns are expected to be more easily found in Africa among ancestral lineages that have remained common in these populations, as in the present case of M1Ala213 from São Tomé. Distributions in other M1 alleles may have been considerably disturbed by genetic drift and therefore show more irregular patterns, which are shifted towards large sized regions of the allele spectrum where mutation occurring at a higher rate could lead to a transitory increase in microsatellite variation.

Age estimates of α1-antitrypsin variants

The use of microsatellite variation to estimate the age of lineages defined by unique mutational events is typically hampered by the need to rely on assumptions that cannot be verified about the relevant parameters. However, attempts to establish relationships between time and levels of variation can still be useful for evaluating the relative antiquity of different variants or for analysing the factors that might produce inconsistencies between age estimates and the geography of each mutation.

Our age estimates for the S and Z alleles are consistent with their virtual confinement to populations of European descent. Moreover, the finding that both variants share a large sized P4 (161 bp) PI microsatellite allele common only in the non-African M1 alleles adds further support for a European origin of S and Z and allows relative age comparisons to be made without the distortions that could be caused by allele-specific mutation rates (Carvalho-Silva et al. 1999).

The global estimate of 4070–3210 years for the coalescent times of Z haplotypes obtained from the information provided by the CBG, PI-$(TG)_n$ and PCI-$(TG)_n$ loci lies between a previous estimate of 6840 years based on RFLP haplotypes (Cox et al. 1985) and one of 2000 years derived solely from CBG microsatellite variation (Byth et al. 1994). This suggests that the Z mutation could have been dispersed during Neolithic times, similar to the cystic fibrosis ΔF508 mutation, if its more recent age estimates are accepted (Serre et al. 1990; but see Morral et al. 1994). The observation of levels of microsatellite diversity in Z types from Portugal that are similar to, or even higher than, those found in northern Europe contrasts with the stepwise reduction in variation that would be expected if the mutation was spread from north to south. Thus, the Z mutation might not have originated in Scandinavia as proposed on the basis of gene frequencies alone (Hutchison 1998).

The estimate of 14,100–8370 years for the S allele in the Portuguese population indicates that the high frequencies of this variant in the Iberian Peninsula did not result from a recent bottleneck and that the occurrence of PI*S in this region could have predated the pronounced differentiation of the Basque population. Subsequent drift in the Basque Country could have reduced microsatellite diversity, leading to younger age estimates in this population. When the ages of S and Z are compared with their present gene frequencies in different populations (Hutchison 1998), it is interesting to see that, in spite of an age estimate that is more than two times younger than that of S, the Z variant is more homogeneously distributed, indicating that its spread has probably been aligned with major population movements within Europe. On the contrary, the older S variant still presents a well-defined SW-NE gradient consistent with an origin in Iberia followed by dispersion against the main directions of demic diffusion in Europe. A controversial proposal of a population expansion from the Iberian Peninsula occurring 10,000–15,000 years ago based on the distribution of mtDNA haplogroup V could provide a basis for such dispersion (Torroni et al. 1998; but see Izaguirre and de la Rúa 1999).

In contrast to the S and Z alleles, M2 and M3 variants in the two Iberian samples had much lower age estimates than would be expected from their wide distributions in most human populations. If the variation in mutation rates along the microsatellite allele spectrum is not strong enough to have produced such distortions, this indicates that previous bottlenecks might have affected their microsatellite variation. A comparison of M3 microsatellite distributions in African and Iberian populations (Fig. 4) suggests that the modal allele in Iberian samples could have become fixed after a drastic bottleneck upon the pre-existing African distribution. However, the possibility of an independent mutation from a non-African M1Val213 variant implying a multi-regional origin for the PI M3 allele cannot be excluded. Given the small number of M3 chromosomes analysed, a more extensive study involving other African, European and Asian populations, including the further evaluation of the molecular homogeneity of M3 and M2 in different samples, will be needed to provide a more reliable interpretation of the present results.

Acknowledgements The authors wish to thank the people and the Ministry of Health of the Democratic Republic of São Tomé e Príncipe. We are especially grateful to Dr. Nuno Ferrand and an anonymous reviewer for comments on the manuscript and helpful suggestions. Field work in São Tomé was supported by Instituto de Cooperação Científica e Tecnológica Internacional (ICCTI). Susana Seixas is supported by grant BD/13885/97 from Praxis XXI.

References

Albarrán C, García O, Alonso A, Deka R, Martín P, Trovoada MJ, Amorim A, Sancho M (1998) Patterns of haplotype variation at the D1S80 locus and a flank sequence polymorphism in African and non-African populations. Prog Forensic Genet 7: 401–403

Amos W (1999) A comparative approach to the study of microsatellite evolution. In: Goldstein DB, Schlötterer C (eds) Microsatellites. Evolution and applications. Oxford University Press, Oxford, pp 66–79

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F (1998) Genetix, logiciel sous Windows pour la génétique des populations. Laboratoire Génome et populations, CNRS UPR 9060, Université de Montpellier II, Montpellier, France

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

Billingsley GD, Walter MA, Hammond GL, Cox DW (1993) Physical mapping of four serpin genes: α1-antitrypsin, α1-antichymotrypsin, corticosteroid-binding globulin, and protein C inhibitor, within a 280-kb region on chromosome 14q32.1. Am J Hum Genet 52:343–353

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 human Y chromosome. Am J Hum Genet 65:1623–1638

Brantly M, Nukiwa T, Crystal RG (1988) Molecular basis of alpha-1-antitrypsin deficiency. Am J Med (Suppl 6A) 84:13–31

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in microsatellites: influence of structure and length of the tandem repeats. Am J Hum Genet 62:1408–1415

Byth BC, Cox DW (1993a) A $(CA)_n$ repeat polymorphism at the 5' end of the α1-antitrypsin (PI) gene. Hum Mol Genet 2:1752

Byth BC, Cox DW (1993b) Two consecutive dinucleotide repeats constitute an informative marker at the α1-antichymotrypsin gene. Hum Mol Genet 2:1085

Byth BC, Meijers JCM, Cox DW (1993) A $(CA)_n$ polymorphism in the protein C inhibitor (PCI) gene. Hum Mol Genet 2:1752

Byth B, Billingsley GD, Cox DW (1994) Physical and genetic mapping of the serpin gene cluster at 14q32.1: allelic association and a unique haplotype associated with α1-antitrypsin deficiency. Am J Hum Genet 55:126–133

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

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

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

Cox DW (1995) α1-Antitrypsin deficiency. In: Scriver CH, Beaudet AL, Sly WS, Valle D (eds) The metabolic and molecular bases of inherited disease, vol 3, 7th edn. McGraw-Hill, New York, pp 4125–4158

Cox DW, Billingsley (1986) Restriction enzyme *Mae*III for prenatal diagnosis of α1-antitrypsin deficiency. Lancet 2:741–742

Cox DW, Woo SLC, Mansfield T (1985) DNA restriction fragments associated with α1-antitrypsin indicate a single origin for deficiency allele PI Z. Nature 316:79–81

DeCroo S, Kamboh MI, Ferrel RE (1991) Population genetics of alpha-1-antitrypsin polymorphism in US whites, US blacks and African blacks. Hum Hered 41:215–221

Dempster A, Laird N, Rubin D (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J R Statistics Soc 39:1–38

Faber J-P, Weidinger S, Olek K (1990) Sequence data of the rare deficient alpha-1-antitrypsin variant PI Zaugsburg. Am J Hum Genet 46:1158–1162

Gaillard MC, Zwi S, Nogueira CM, Ludewick H, Feldman C, Frankel A, Tsilimigras C, Kilroe-Smith TA (1994) Ethnic differences in the occurrence of the M1(ala213) haplotype of alpha-1-antitrypsin in asthmatic and non-asthmatic black and white South Africans. Clin Genet 45:122–127

Garza JC, Slatkin M, Freimer NB (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol 12:594–603

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J Hered 88:335–342

Halkas AC, Gaillard MC, Thomson PD, Green SL, Ludewick H, Kala U (1998) Variants of α1-proteinase inhibitor in black and white South African patients with focal glomerulosclerosis and minimal change nephrotic syndrome. J Med Genet 35:6–9

Hutchison DCS (1998) α1-Antitrypsin deficiency in Europe: geographical distribution of Pi types S and Z. Respir Med 92: 367–377

Izaguirre N, Rua C de la (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

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E (1996) Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci USA 93:15285–15288

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

Morral N, Bertrandpetit J, Estivil X, Nunes V, Casals T, Giménez J, Reis A, Varon-Mateeva R, Macek Jr M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, Ferec C, Arce M de, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L (1994) The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169–175

Nauta MJ, Weissing FJ (1996) Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics 143:1021–1032

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

Nukiwa T, Ogushi F, Crystal RG (1996) Alpha-1-antitrypsin gene evolution. In: Crystal RG (ed) Alpha-1-antitrypsin deficiency. Biology, pathogenesis, clinical manifestations, therapy. Dekker, New York, pp 33–43

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

Prata MJ, Amorim A, Gusmão L, Trovoada MJ (1996) Population genetics of the STRs TPO, TH01 and VWFA31/A in São Tomé e Príncipe. Adv Forensic Haemogenet 7:604–606

Rocha J, Pinto D, Santos MT, Amorim A, Amil-Dias J, Cardoso-Rodrigues F, Aguiar A (1997) Analysis of the allelic diversity of a $(CA)_n$ repeat polymorphism among α1-antitrypsin gene products from northern Portugal. Hum Genet 99:194–198

Rollini P, Fournier REK (1997) A 370-kb cosmid contig of the serpin gene cluster on human chromosome 14q32.1: molecular linkage of the genes encoding α1-antichymotrypsin, protein C inhibitor, kallistatin, α1-antitrypsin, and corticosteroid-binding globulin. Genomics 46:409–415

Schlötterer C, Wiehe T (1999) Microsatellites, a neutral marker to infer selective sweeps. In: Goldstein DB, Schlötterer C (eds) Microsatellites. Evolution and applications. Oxford University Press, Oxford, pp 238–248

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, Department of Anthropology, University of Geneva, Switzerland

Sefton L, Kelsey G, Kearney P, Povey, Wolfe J (1990) A physical map of the human PI and ACT genes. Genomics 7:382–388

Seixas S, Trovoada MJ, Santos MT, Rocha J (1999a) A novel alpha-1-antitrypsin P362H variant found in a population sample from S. Tomé e Príncipe (Gulf of Guinea, West Africa). Hum Mutat 13:414

Seixas S, Trovoada MJ, Rocha J (1999b) Haplotype analysis of the apolipoprotein E (APOE) and apolipoprotein C1 (APOC1) loci in Portugal and S. Tomé e Príncipe (Gulf of Guinea): linkage disequilibrium evidence that apolipoprotein E*4 is the ancestral APOE allele. Hum Biol 71:1001–1008

Seixas S, Garcia O, Amorim A, Rocha J (2000) A novel alpha-1-antitrypsin R281del variant found in a population sample from the Basque Country. Hum Mutat 15:121–122

Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillandier A, Boué J, Boué A (1990) Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. Hum Genet 84:449–454

Slatkin M (1995) Hitchhiking and associative overdominance at a microsatelite locus. Mol Biol Evol 12:473–480

Tishkoff SA, Dietzsch E, Speed WC, Pakstis AJ, Cheung K, Kidd JR, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Watson E, Krings M, Pääbo S, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellito D, Rengo C, Forster P, Savontaus M-L, Bonné-Tamir B, Scozzari R (1998) mt-DNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62:1137–1152

Tremblay M, Vezina H (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am J Hum Genet 66:651–658

Weber JL, Wong C (1993) Mutation of short tandem repeats. Hum Mol Genet 2:1123–1128

Xu X, Peng M, Fang Z, Xu X (2000) The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24: 396–399

2.1.1 Comentário

2.1.1.1 Distribuição das frequências dos alelos da α1-antitripsina

Às distribuições de frequências dos alelos comuns da α1-antitripsina em Portugal, País Basco e São Tomé analisadas no artigo 2 foi possível acrescentar os resultados do estudo de uma amostra de 68 ameríndios de língua Quechua das províncias de Tayacaja e Arequipa dos Andes centrais peruanos (Luiselli *et al.*, 2000; Tarazona-Santos *et al.*, 2001). Esta amostra caracteriza-se pela quase total ausência do alelo ancestral M1Ala213 (0,016), pela baixa frequência de M2 (0,023) e por uma frequência de M3 (0,555) situada muito acima do intervalo de variação registado em populações de fora do continente americano (Figura II.5). A presença do alelo S com uma frequência muito baixa (0,008) é provavelmente devida a miscigenação com indivíduos de origem espanhola.

A relativa raridade de PI*M1Ala213 não tinha sido notada anteriormente em populações ameríndias devido à não discriminação de PI*M1Ala213 e PI*M1Val213. Dado que PI*M1Ala213 não foi detectado no Japão, na única análise de uma amostra asiática em que se procurou identificar este alelo (Figura II.3) (Nukiwa *et al.*, 1996b), é provável que a distribuição agora observada reflicta a proximidade genética entre as populações da Ásia e das Américas. A elevação da frequência de PI*M3 é semelhante à registada noutras amostras originárias de regiões muito dispersas da América do Sul (Figura II.3 e II.5) e constitui uma característica distintiva das populações do sub-continente, à semelhança da quase fixação do alelo O do grupo sanguíneo ABO, da maior frequência do alelo DI*A do grupo sanguíneo Diego ou da predominância de um único haplogrupo do cromossoma Y (Cavalli-Sforza *et al.*, 1994; Pena *et al.*, 1995; Santos *et al.*, 1996; Underhill *et al.*, 1996; Crawford, 1998). A ausência de dados sobre populações de diferentes locais da América do Norte e da Ásia em que todos os produtos génicos tenham sido discriminados impede, contudo, a identificação da área em que se deu a alteração das frequências dos alelos α1-antitripsina durante a colonização do continente americano.

A análise das distâncias genéticas entre as populações em que se discriminaram completamente os alelos comuns da α1-antitripsina revela uma hierarquia de semelhanças parecida com a que se obtém com a utilização de vários *loci* polimórficos simultaneamente (Figura II.6). Se a amostra Quechua for removida (Figura II.6 A),

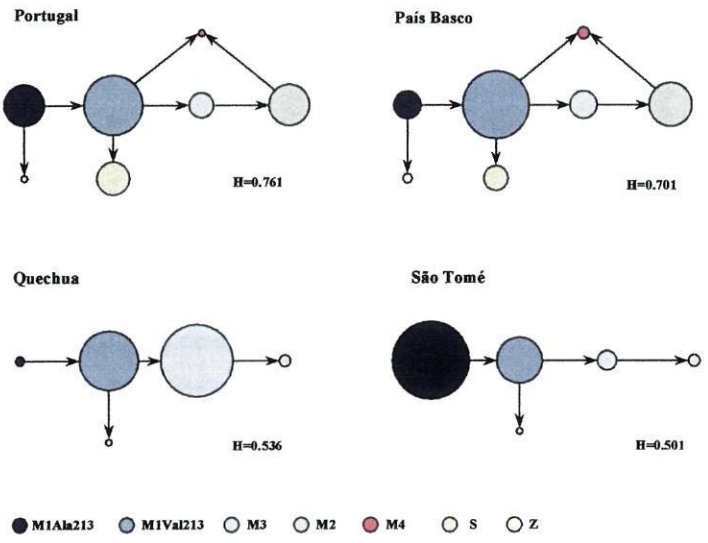


Figura II.5- Distribuição das frequências dos alelos de PI nas populações estudadas no presente trabalho. A área dos círculos é proporcional à frequência dos alelos. H-heterozigotia.

verifica-se que a primeira dicotomia separa as populações africanas das populações euroasiáticas que só posteriormente se dividem entre si, tal como acontece na maioria dos padrões de agrupamento das populações humanas (Figuras II.6 B e C). Na população Quechua, o aumento da frequência de M3 é o resultado provável da acção marcada da deriva genética que acompanhou a colonização das Américas (Cavalli-Sforza *et al.,* 1994; Pena *et al.,* 1995; Santos *et al.*, 1996; Mesa *et al.,* 2000) embora, neste caso, o grau de diferenciação seja superior à média e coloque aquela amostra numa posição excêntrica que não reproduz a proximidade entre os ameríndios e as populações asiáticas revelada pelo estudo de um maior número de marcadores (Figuras II.6).

À semelhança do que se observa com outros polimorfismos electroforéticos, ou com polimorfismos de tamanho de fragmentos de restrição (RFLPs), é também nas populações europeias que se registam os valores mais elevados de heterozigotia para a α1-antitripsina (Figura II.6 A), ao contrário do que acontece com os microssatélites e o DNA mitocondrial, que apresentam maior diversidade genética na África sub-shariana (Relethford, 2001). A discrepância entre as estimativas de diversidade obtidas

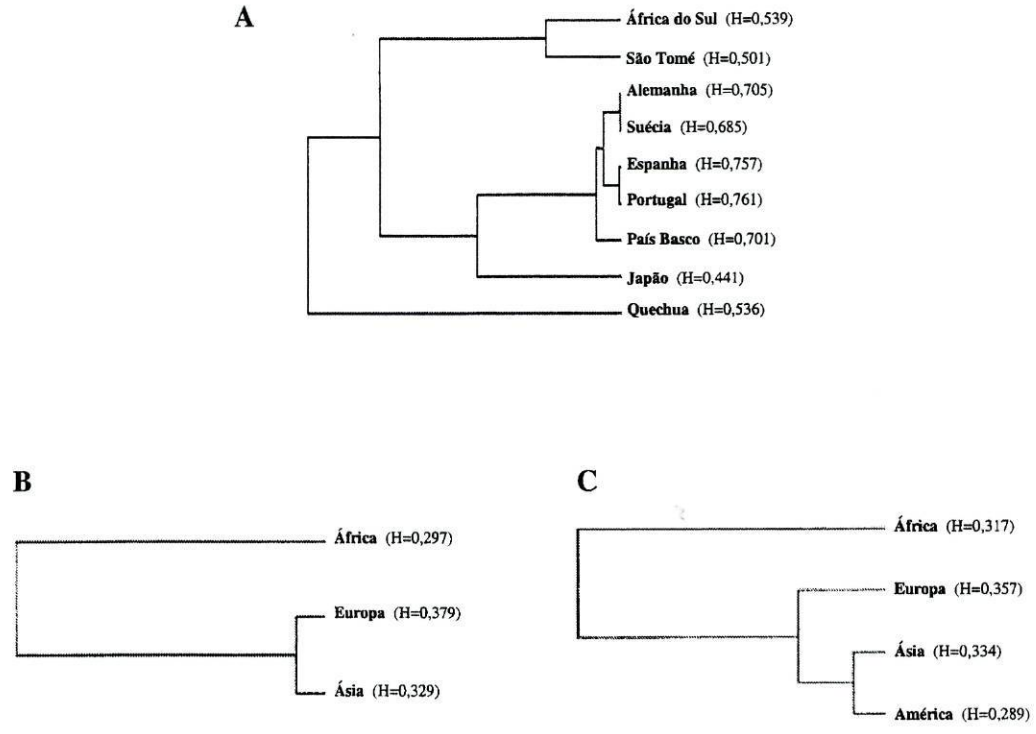


Figura II.6- Agrupamentos de populações humanas obtidos com o método de UPGMA (*Unweigthted Pair Group Method with Arithmetic Mean*) com base na distância genética F_{ST} utilizando diferentes tipos de *loci*. **A**- Agrupamento com base no *locus* PI. **B**- Agrupamento com base em 79 RFLPs (polimorfismos de tamanho de fragmentos de restrição) (Mountain e Cavalli-Sforza, 1994). **C**- Agrupamento com base em 42 polimorfismos de proteínas (Cavalli-Sforza *et al*, 1994). As distâncias genéticas foram calculadas com o programa ARLEQUIN (Schneider *et al*, 1997) e os agrupamentos efectuados com o programa STATISTICA (http://www.statsoftinc.com). H- heterozigotia.

com diferentes tipos de marcadores é geralmente atribuída ao enviezamento provocado pela amostragem preferencial de populações europeias na pesquisa de polimorfismos electroforéticos e RFLPs (Mountain e Cavalli-Sforza, 1994). Como a intensidade do enviezamento é inversamente proporcional à heterozigotia real, é mais provável que os *loci* com baixa taxa de mutação, nos quais se incluem aqueles polimorfismos, tenham uma diversidade genética sobreavaliada (Mountain e Cavalli-Sforza, 1994; Rogers e Jorde, 1996; Relethford, 1997). Se as relações filogenéticas entre alelos forem conhecidas, é contudo possível diminuir os efeitos desta distorção e conciliar os padrões de variação observados em diferentes tipos de sistemas genéticos. Por exemplo, a observação de maior variabilidade genética em África em estudos de DNA mitocondrial e microssatélites é normalmente considerada uma indicação de que

as populações africanas estiveram historicamente menos sujeitas à acção da deriva genética do que as populações euro-asiáticas (Rogers e Jorde, 1995; Relethford, 2001). A ser assim, é de esperar que o enviezamento amostral nos sistemas electroforéticos e RFLPs conduza a uma maior frequência de alelos derivados na Europa, enquanto que à menor diversidade registada em África deve estar associada a retenção de frequências elevadas dos alelos ancestrais. Por outro lado, a maior proximidade genética das populações asiáticas relativamente à fonte de enviezamento (Europa) e a colonização das Américas partir da Ásia Oriental implicam que, nessas populações, existam também frequências proporcionalmente mais elevadas dos alelos derivados. Estas previsões foram confirmadas em estudos de RFLPs em que a distinção de alelos primitivos e derivados foi feita através de comparações com Primatas não-humanos (Moutain e Cavalli-Sforza 1994; Watkins *et al.*, 2001). O mesmo se verificou numa análise alargada da distribuição de inserções *Alu* polimórficas (Watkins *et al.*, 2001), embora neste caso as populações africanas registassem as heterozigotias mais elevadas e o enviezamento amostral se restringisse à sobrevalorização da frequência do alelo derivado (presença da inserção).

O conhecimento das relações filogenéticas entre as mutações da α1-antitripsina permite verificar que a distribuição dos seus alelos derivados também tem um padrão concordante com os resultados obtidos em *loci* com propriedades semelhantes (Figuras II.3, II.5 e II.6): a baixa heterozigotia das amostras africanas resulta da preservação de frequências elevadas do alelo ancestral M1Ala213; a maior diversidade das amostras europeias está associada à elevação das frequências dos alelos derivados M1Val213, M2 e M3; à menor heterozigotia da amostra japonesa corresponde a perda de M1Ala213; a diferenciação da amostra Quechua resulta do forte aumento da frequência do alelo derivado M3. Tendo em conta as irregularidades que caracterizam a análise da diversidade de um único *locus*, é notável o modo como a distribuição das frequências alélicas da α1-antitripsina capta os principais sinais da história evolutiva das populações humanas reconstruída a partir da ponderação de diferentes *loci*.

2.1.1.2 Variação genética em microssatélites associados aos alelos da α1-antitripsina

2.1.1.2.1 *Comparações interpopulacionais*

Os padrões de diversidade haplotípica definidos pela combinação de microssatélites com as mutações pontuais da α1-antitripsina podem ser interpretados a nível interpopulacional ou a nível interalélico. As comparações entre populações permitem examinar o impacto dos principais acontecimentos demográficos da sua história evolutiva e aumentam a resolução dos estudos que apenas se baseiam em distribuições de frequências alélicas. As comparações interalélicas podem ser usadas para avaliar a antiguidade relativa das mutações pontuais ou, em casos mais favoráveis, para fazer a sua datação absoluta. Embora estas duas perspectivas não possam nem devam ser completamente separadas, é conveniente começar por comentar a congruência entre a diversidade haplotípica associada aos alelos da α1-antitripsina e a história evolutiva das populações estudadas no artigo 2, às quais se acrescentam agora resultados obtidos na amostra Quechua.

A comparação das distribuições dos microssatélites PI(TG)n e PCI(TG)n nos alelos M1Ala213, M1Val213 e M3 de Portugal e do País Basco ilustra bem os efeitos que a história populacional pode provocar na variação haplotípica observada (Figuras II.7 a II.9). Estes alelos são os que apresentam maiores diferenças de frequência entre as duas populações (Tabela 1 e Figura 4 do artigo 2) e registam, na amostra basca, uma diminuição da diversidade genética em ambos os microssatélites. Para além da redução de diversidade, as distribuições das frequências alélicas dos microssatélites no País Basco têm configurações mais irregulares, caracterizadas por um número elevado de modas e descontinuidades (Figuras II.7 a II.9), como acontece, tipicamente, nas populações sujeitas à acção da deriva genética (Reich e Goldstein, 1998). De um modo geral, estas observações estão de acordo com a história das duas populações ibéricas e apoiam a hipótese de que a singularidade basca terá resultado de uma redução do efectivo populacional durante um período de maior isolamento em relação às regiões vizinhas (Calafell e Bertranpetit, 1994; Cavalli-Sforza *et al.*, 1994). No entanto, deve salientar-se que muitos outros marcadores, incluindo polimorfismos

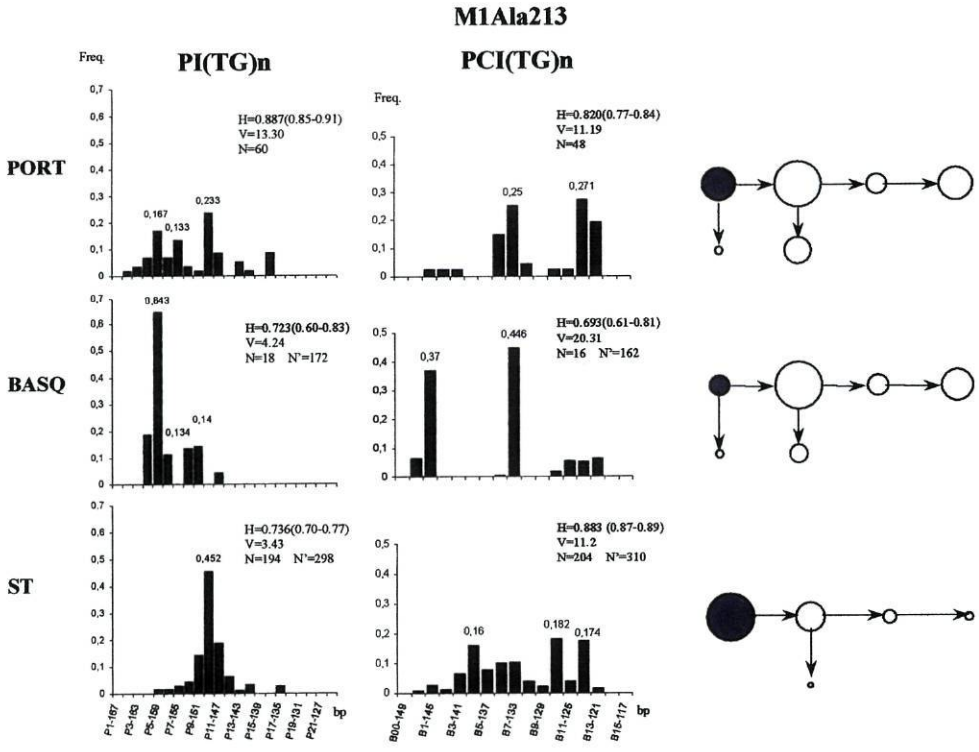


Figura II.7- Distribuição das frequências dos alelos dos *loci* PI(TG)n e PCI(TG)n no seio dos alelos M1Ala213 na população portuguesa (PORT), basca (BASQ) e são-tomense (ST). Na coluna do lado direito mostra-se a relação filogenética e as frequências dos alelos da α1-antitripsina em cada população. O alelo em análise está representado a cheio. H- heterozigotia , V- variância do número de repetições, N- cromossomas M1Ala213 e N'- total de cromossomas utilizados na estimativa das frequências haplotípicas.

do cromossoma Y e do DNA mitocondrial, não mostram uma diferenciação tão clara do País Basco (Bertranpetit *et al.*, 1995; Hurles *et al.*, 1999), o que sugere que o seu isolamento pode já ter sido contrariado, em grande parte, pelo fluxo génico causado pela migração. Harpending e Eller (2000), por exemplo, calculam que se tiver havido um nível de fluxo génico, a partir das populações adjacentes, equivalente a 1% por geração durante 250 gerações, é provável que apenas cerca de 8% dos genes bascos retenham os níveis de diferenciação ancestrais. Em face dos resultados obtidos, pode ser que o *locus* da α1-antitripsina seja um desses genes, juntamente com os grupos sanguíneos ABO e RH e com um número restrito de polimorfismos electroforéticos (Calafell e Bertranpetit, 1994; Cavalli-Sforza *et al.*, 1994) que contribuem para que a população basca se distinga das restantes populações europeias.

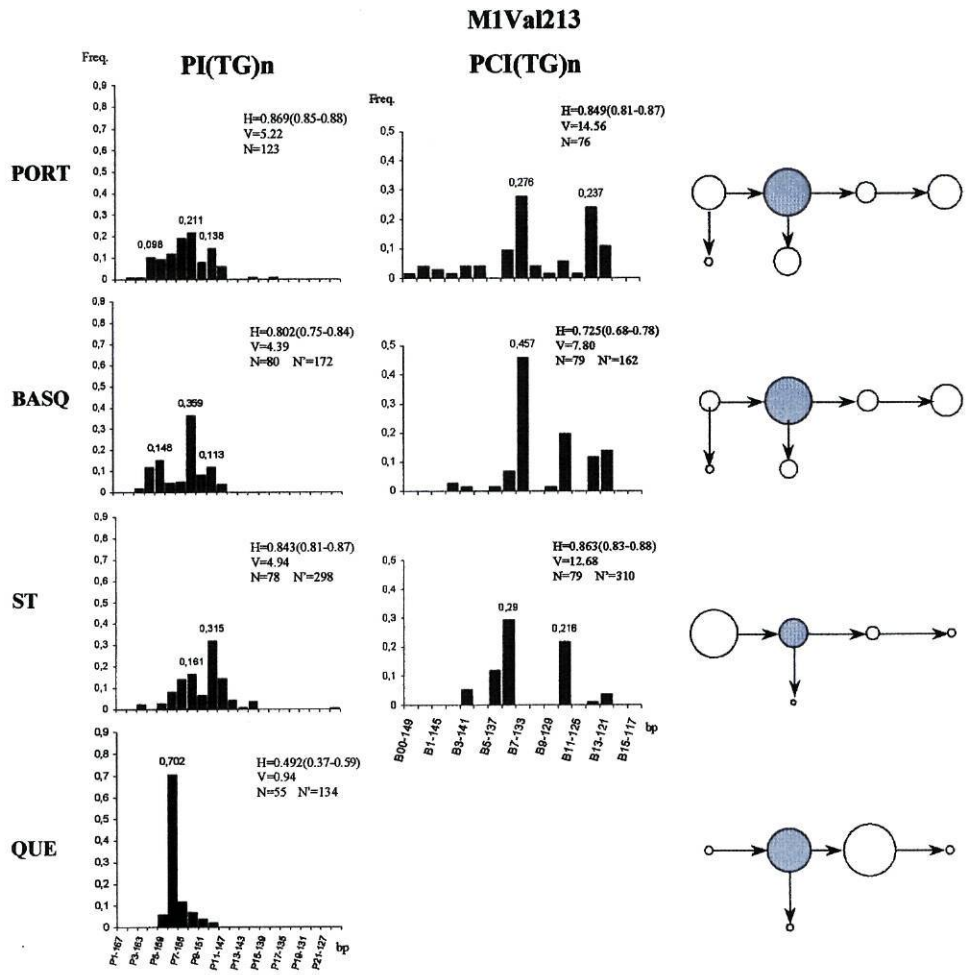


Figura II.8- Distribuição das frequências dos alelos dos *loci* PI(TG)n e PCI(TG)n no seio dos alelos M1Val213 na população portuguesa (PORT), basca (BASQ) são-tomense (ST) e quechua (QUE). Na coluna do lado direito mostra-se a relação filogenética e as frequências dos alelos da α1-antitripsina em cada população. O alelo em análise está representado a cheio. H- heterozigotia , V- variância do número de repetições, N- cromossomas M1Val213 e N'- total de cromossomas utilizados na estimativa das frequências haplotípicas.

Ainda assim, é interessante verificar que a diferenciação observada na α1-antitripsina só é visível se a variação haplotípica for tida em conta e se todos os alelos puderem ser discriminados. Na ausência de informação sobre os microssatélites e em caso de junção de M1Ala213 e M1Val213 num único electromorfo, a distribuição de frequências obtida no País Basco é praticamente indistinguível da observada noutras regiões da Península Ibérica.

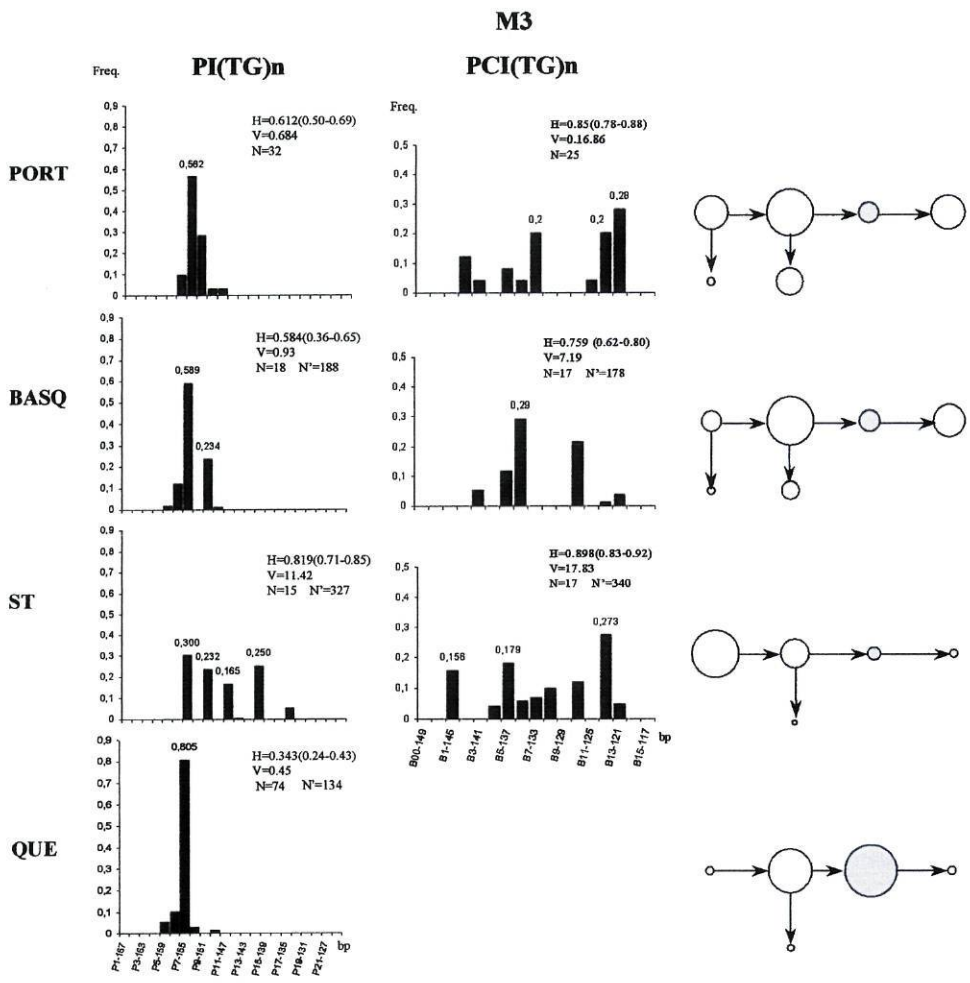


Figura II.9- Distribuição das frequências dos alelos dos *loci* PI(TG)n e PCI(TG)n no seio dos alelos M3 na população portuguesa (PORT), basca (BASQ) são-tomense (ST) e quechua (QUE). Na coluna do lado direito mostra-se a relação filogenética e as frequências dos alelos da α1-antitripsina em cada população. O alelo em análise está representada a cheio. H- heterozigotia , V- variância do número de repetições, N- cromossomas M3 e N'- total de cromossomas utilizados na estimativa das frequências haplotípicas.

A análise da amostra da população Quechua é também um bom exemplo da forma como a variação haplotípica pode acrescentar informações importantes para a interpretação dos padrões de heterogeneidade geográfica de um polimorfismo. As distribuições de frequências do *locus* PI(TG)n, o único microssatélite que pôde ser testado nesta amostra, revelam que houve uma redução significativa da diversidade genética associada a M1Val213 e M3 (Figuras II.8 e II.9). Isto indica claramente que a alteração nas frequências destes dois alelos foi causada por deriva genética resultante

de um estrangulamento do efectivo populacional. No entanto, contrariamente ao País Basco, as distribuições da população Quechua são mais regulares e unimodais e sugerem que esse estrangulamento terá tido maior duração e intensidade nas populações da América do Sul. Este resultado está de acordo com a observação de uma diminuição da diversidade genética em outros *loci* autossómicos (Cavalli-Sforza *et al.*, 1994; Mesa *et al.*, 2000), no DNA mitocondrial e (Horai *et al.*, 1993) e sobretudo, no cromossoma Y que, como já se referiu, tem um haplogrupo predominante em regiões do continente americano geograficamente muito afastadas (Pena *et al.*, 1995; Santos *et al.*, 1996; Underhill *et al.*, 1996). Admitindo que a colonização dos Andes pelos Quechua se fez a partir da expansão de uma população fundadora que perdeu completamente a variação alélica do microssatélite PI(TG)n nas linhagens M1Val213 e M3, é possível tentar datar essa colonização através da estimativa do tempo necessário à recuperação dos níveis de diversidade actualmente observados. Recorrendo ao formalismo exposto no Artigo 2 (equação 1), e assumindo que houve uma acumulação regular de mutações à taxa média de 0,001 por cada geração de 30 anos, o tempo necessário para as frequências dos alelos ancestrais P6 (157 bp) e P7 (155 bp) das linhagens M1Val213 e M3 (Figuras II.8 e II.9) baixarem para os valores actualmente observados é de 11 820 (6 090-19 470) e 6 930 (3 450-11 400) anos, respectivamente. Estas estimativas referem-se apenas ao início da expansão no efectivo populacional que permitiu acumular as mutações do microssatélite e podem ter pelo menos dois tipos opostos de enviezamento. Se não tiver havido perda total de diversidade, o que é sugerido pela data mais antiga ter sido obtida com a linhagem mais antiga (M1Val213), a idade pode estar sobreavaliada. Se, no entanto, a perda total de diversidade tiver de facto ocorrido, a expansão de tamanho pode ter acontecido muito depois da fundação da população e as datas estarão subestimadas. De qualquer modo, as duas datas têm intervalos de variação sobreponíveis e um valor médio de 9 375 (3 450-19 470) anos, semelhante a estimativas recentes de 9 000-11 000 anos para a idade do haplogrupo predominante do cromossoma Y em índios colombianos (Ruiz-Linares *et al.*, 1999). Por outro lado, a aplicação da metodologia agora utilizada aos dados de Tarazona-Santos *et al.* (2001), sobre a variação de microssatélites desse haplogrupo na população Quechua conduz a uma estimativa pontual de 8 710 anos, compatível com a obtida com as linhagens da α1-antitripsina.

Estes valores são também concordantes com a evidência paleo-ecológica, que indica que o povoamento maciço da região Andina só teria sido possível no fim da última era glaciar há cerca de 10 000 - 12 000 anos (Dillehay, 1999; Tarazona-Santos *et al.*, 2001). No futuro, seria interessante fazer uma pesquisa de polimorfismos nas sequências de DNA dos intrões e das zonas adjacentes do gene da α1-antitripsina para estabelecer linhagens que permitam utilizar esta região do genoma numa abordagem filogeográfica mais detalhada da colonização das Américas.

Ao contrário das amostras do País Basco e do Peru, os padrões de variação haplotípica na população de São Tomé só muito dificilmente são compatíveis com a sua história evolutiva. Tendo em conta que esta população retém, para os *loci* até agora analisados, os níveis elevados de diversidade genética característicos do continente africano (Prata *et al.*, 1996; Mateu *et al.*, 1997; Albarrán *et al.*, 1998), seria de esperar que a variabilidade dos microssatélites associados às linhagens da α1-antitripsina fosse mais elevada em São Tomé, à semelhança do que acontece noutras populações da África sub-sahariana para este tipo de *loci* (Jorde *et al.*, 1997; Calafell *et al.*, 1998; Jorde *et al.*, 2000). No entanto, apesar de se observarem valores de heterozigotia mais elevados no microssatélite PCI(TG)n, a variabilidade associada ao alelo M1Ala213 no *locus* PI(TG)n é significativamente mais baixa do que a registada na população portuguesa (Figura II.7).

A primeira questão que se coloca com este resultado é a de saber se a redução de diversidade que se observou é excepcional ou se, pelo contrário, é um acontecimento fortuito que se inclui na gama de variação esperada quando se analisam diferentes *loci* com a mesma história evolutiva. No artigo 2, o facto de, ao contrário da população basca, a redução de heterozigotia em São Tomé se restringir ao microssatélite PI(TG)n e à linhagem mais antiga M1Ala213, foi tomado como uma indicação informal de que este resultado não se obteve por acaso. No entanto, só a comparação entre os valores de diversidade observados em São Tomé e na população portuguesa para vários *loci* poderá sustentar formalmente essa conclusão. Na impossibilidade de realizar esta análise, procurou-se avaliar até que ponto os resultados obtidos se enquadravam nas diferenças de diversidade registadas numa recente comparação global de populações africanas e não-africanas com um conjunto de 94 microssatélites (Schlötterer, 2002). De acordo com a metodologia desenvolvida nesse estudo, a comparação da

diversidade genética de duas populações para um dado microssatélite pode ser feita através do logaritmo natural da razão entre as variâncias do respectivo número de repetições (ln RV). Quando calculado para vários microssatélites, o parâmetro ln RV tem uma distribuição normal que permite atribuir a cada *locus* uma probabilidade *P* do respectivo valor de ln RV se dever ao acaso. No conjunto dos *loci* comparados por Schlötterer (2002), os dois únicos microssatélites com reduções significativas de diversidade em África, D10S249 e D6S305, tinham valores de ln RV de –2,107 e –1,239 com *P* igual a 0,002 e 0,023, respectivamente. Uma vez que o valor de ln RV para o microssatélite PI(TG)n na linhagem M1Ala213 é igual a –1,355, é provável que a redução de diversidade agora observada não tenha sido originada por acaso.

Normalmente, considera-se que a selecção pode causar desvios significativos em relação ao padrão médio observado na generalidade dos *loci*, com base no conceito de que os fenómenos selectivos tendem a repercutir-se apenas em regiões delimitadas do genoma (Lewontin e Krakauer, 1973). Teoricamente, se um alelo for favorecido (selecção positiva) ou se, em alternativa, houver a remoção recorrente de mutações deletérias (selecção negativa), verificar-se-á uma perda de diversidade genética numa extensão inversamente proporcional à fracção de recombinação entre a região seleccionada e as regiões vizinhas (Schlötterer e Wiehe, 1999). Por esta razão, a redução na heterozigotia do microssatélite PI(TG)n na linhagem M1Ala213 pode ter sido provocada por selecção nos genes da CBG e da PI que lhe estão próximos. Como o processo de eliminação de mutações deletérias é equivalente a uma redução localizada do tamanho da população (Charlesworth *et al.*, 1995; Schlötterer e Wiehe, 1999), seria de esperar que, em caso de selecção negativa, a perda de diversidade no microssatélite fosse acompanhada de uma distribuição irregular de frequências alélicas, o que não se verifica (Figura II.7). Pelo contrário, a selecção positiva é equivalente a uma expansão populacional e as distribuições dos microssatélites associados às regiões seleccionadas tendem a apresentar uma moda predominante (Slatkin, 1995), como de facto acontece na linhagem M1Ala213.

A pressão causada pelas proteases de agentes infecciosos pode ser um factor importante de selecção positiva e tem sido apontada como uma das causas da acelaração da divergência evolutiva das proteínas da família SERPIN (Hill e Hastie, 1987; Goodwin *et al.*, 1996). Estas proteases estão envolvidas nos processos de

penetração no hospedeiro e na digestão da matriz extracelular de muitos tecidos afectados. Por exemplo, a actividade elastinolítica é um dos factores de virulência mais importantes nas infecções causadas por bactérias do género *Pseudomonas* e a elastase produzida pelas formas larvares de *Schistosoma mansoni* é essencial para destruir a pele durante a fase invasiva do hospedeiro (Werb *et al.*, 1982; Potempa *et al.*, 1994). É assim possível que, em condições de elevada carga parasitária, as mutações que melhorem a capacidade inibidora da PI, ou aumentem a sua concentração plasmática, possam ser selectivamente favorecidas. Por outro lado, a participação da CBG no controlo da resposta inflamatória, também pode expor este *locus* a pressões selectivas equivalentes, relacionadas com a optimização da distribuição de glucocorticóides. De qualquer forma, só uma análise sistemática das regiões codificantes e reguladoras destes genes, permitirá verificar se há mutações associadas à linhagem M1Ala213 que possam sustentar a hipótese de selecção.

Tal como foi discutido no artigo 2, as particularidades do processo mutacional dos microssatélites também podem conduzir a situações de redução de diversidade sem ser necessário invocar a acção da selecção. Há evidências crescentes de que a frequência e o sentido das mutações podem variar com o número de repetições dos alelos num mesmo *locus* (Jin *et al.*, 1996; Goldstein e Pollock, 1997; Brinkmann *et al.*, 1998; Carvalho-Silva *et al.*, 1999). Em geral, os alelos de menor tamanho tendem a expandir-se, enquanto os alelos mais longos podem ter taxas de mutação mais elevadas e uma maior tendência para a contracção (Garza *et al.*, 1995; Nauta e Weissing, 1996; Xu *et al.*, 2000; Buard *et al.*, 2002). Uma das consequências interessantes deste tipo de comportamento mutacional é a possibilidade de as taxas de mutação de um mesmo *locus* poderem variar de população para população, ou de linhagem para linhagem, devido às diferenças nas distribuições do tamanho dos alelos. Por outro lado, se o efectivo de uma população for suficientemente grande, é provável que o tamanho dos alelos convirja para distribuições estacionárias, regulares e unimodais, que se manterão inalteradas se não forem perturbadas pelos efeitos da deriva genética (Garza *et al.*, 1995; Nauta e Weissing, 1996; Xu *et al.*, 2000; Buard *et al.*, 2002). Nestas condições, é concebível que, na linhagem M1Ala213 de São Tomé, o *locus* PI(TG)n possa ter atingido um equilíbrio inteiramente determinado pela dinâmica mutacional, com valores de heterozigotia paradoxalmente menores do que os

observados em linhagens menos antigas ou em populações mais sujeitas a deriva genética.

A pesquisa de polimorfismos na sequência dos intrões do gene da PI poderá ser uma boa forma de discriminar as hipóteses de selecção e de equilíbrio mutacional. Se o padrão observado tiver sido provocado por selecção, a redução de diversidade do microssatélite deverá ser acompanhada de menor variação de sequência na linhagem M1Ala213. Se, pelo contrário, houver uma distribuição de equilíbrio causada pelas propriedades mutacionais específicas do microssatélite, a linhagem M1Ala213 deverá ser mais variável ao nível das sequências intrónicas e haverá uma discordância entre os padrões observados com os dois tipos de polimorfismo.

2.1.1.2.2 *Comparações interalélicas*

De cada vez que uma variante da α1-antitripsina é originada numa população, forma-se uma nova linhagem em que a diversidade dos microssatélites adjacentes é completamente eliminada. Em condições ideais de efectivo infinito, a recuperação da diversidade será proporcional ao tempo decorrido desde a origem da nova linhagem e ocorrerá a um ritmo que só depende da taxa de mutação dos microssatélites e da sua distância ao *locus* da PI. Assumindo que as taxas de mutação não são muito desiguais, a recombinação restaurará mais rapidamente a variabilidade nos microssatélites mais distantes e as distribuições de frequência nas diferentes linhagens tenderão a assemelhar-se. É o que acontece com o microssatélite PCI(TG)n, localizado a cerca de 200 kb da PI (Figura 1 do artigo 2), no qual apenas uma pequena fracção da diversidade genética se deve a diferenças entre as linhagens definidas pelas mutações pontuais da α1-antitripsina (Tabela 2 do artigo 2). Pelo contrário, no *locus* PI(TG)n, situado a apenas 7,1 kb (Figura 1 do artigo 2), ainda há diferenças consideráveis entre as distribuições observadas nos alelos da PI que correspondem, pelo menos em parte, a distintas fases de recuperação de diversidade (Tabela 2 e Figura 4 do artigo 2). Por exemplo, na população portuguesa, o decréscimo de heterozigotia de PI(TG)n segundo a ordem M1Ala213>M1Va213>M3>M2 (Figura 4 do artigo 2) está de acordo com a antiguidade relativa destes alelos, inferida a partir da sequência filogenética das respectivas mutações (Figura II.2). Por outro lado, os baixos níveis de diversidade

associados a S e Z, indicam que estes alelos são os mais recentes e que, por isso, têm uma distribuição geográfica mais confinada. No entanto, o alargamento das comparações interalélicas a outras populações mostra que, devido aos factores de diferenciação interpopulacional acima discutidos, podem ocorrer excepções importantes a este tipo de comportamento. Uma forma de avaliar as influências cruzadas que a história evolutiva das populações e a filogenia dos alelos têm nas distribuições dos microssatélites consiste em agrupar as diferentes linhagens de acordo com a distância genética calculada com base nessas distribuições. Se os factores de diferenciação interpopulacional tivessem uma importância menor, seria de esperar que os diferentes alelos da α1-antitripsina se agrupassem entre si independentemente da sua proveniência geográfica. Porém, a análise dos agrupamentos das linhagens de PI obtidos com base no microssatélites PI(TG)n mostra que nem sempre isso acontece (Figura II.10). Com excepção de M1Val213 de Portugal e do País Basco, as linhagens M1Ala213 e M1Val213 das diferentes populações não se agrupam entre si.

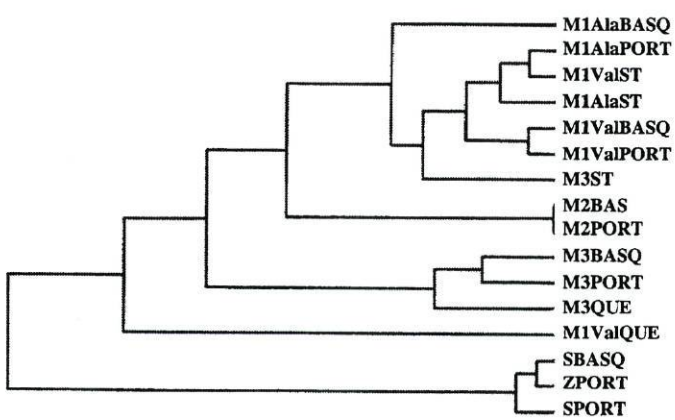


Figura II.10- Agrupamento dos alelos da α1-antitripsina de diferentes populações obtidos com o método de UPGMA (*Unweigthted Pair Group Method with Arithmetic Mean*) a partir da distância genética F_{ST} calculada com base nas frequências dos alelos do microssatélite PI(TG)n. As distâncias genéticas foram calculadas com o programa ARLEQUIN (Schneider *et al*, 1997) e os agrupamentos efectuados com o programa STATISTICA (http://www.statsoftinc.com). PORT- Portugal, BASQ- País Basco, ST- São Tomé e QUE- Quechua.

O alelo M1Val213 na população Quechua tem uma distribuição completamente distinta da das outras amostras estudadas. A distribuição do alelo M3 em São Tomé é muito diferente da que se observa nos Quechua e nas duas populações ibéricas. Estas discrepâncias mostram que só será possível fazer uma descrição completa das diferenças interalélicas se forem analisadas várias populações com histórias evolutivas diferentes. A combinação de mutações pontuais com microssatélites será tanto menos informativa sobre a filogenia dos alelos quanto mais for influenciada pela história das populações.

2.1.1.2.3 *Datação dos alelos S e Z da α1-antitripsina*

Como foi acima discutido, há vários factores que podem afectar a estimativa correcta da antiguidade absoluta ou relativa de mutações pontuais a partir da variação acumulada em microssatélites. Por um lado, fenómenos como a deriva genética associada à história das populações, a selecção, ou a convergência para distribuições de equilíbrio, fazem com que muitos dos pressupostos do cálculo das idades das mutações não sejam verificados (de Knijff, 2000). Por outro lado, os parâmetros básicos necessários para realizar estes cálculos - tais como a taxa de mutação, a fracção de recombinação ou intervalo de tempo entre gerações - não são conhecidos com precisão. Ainda assim, vale a pena tentar utilizar a informação disponível para realizar datações, uma vez que o conhecimento da idade de uma mutação contribui para identificar os factores que determinaram a sua dispersão geográfica e permite formular hipóteses que poderão ser depois analisadas à luz de outros tipos de evidência (Slatkin e Rannala, 2000; Rannala e Bertorelle, 2001).

Apesar da sua relativa simplicidade, o método de datação dos alelos da α1-antitripsina desenvolvido no artigo 2 tem a desvantagem de se basear na média das estimativas obtidas com cada um dos microssatélites isoladamente, não permitindo incorporar informação sobre a respectiva mutabilidade relativa. A fim de combinar a informação dos *loci* CBG, PI(TG)n e PCI(TG)n numa única estimativa de idade, foi possível aplicar um método alternativo a uma subamostra de 79 alelos Z e 59 alelos S da população portuguesa, em que se pôde determinar a distribuição dos haplótipos definidos simultaneamente pelos três *loci* (Tabelas II.1 A e B). Neste método, baseado

numa modificação da abordagem de Bertranpetit e Calafell (1996), começa-se por reconstruir a relação filogenética entre os haplótipos associados a cada um dos alelos que se pretende datar, utilizando critérios de máxima parcimónia (Figura II.11). Para tal, é necessário definir os haplótipos ancestrais associados às mutações que originaram os alelos S e Z e inferir o número mínimo de acontecimentos mutacionais ou recombinacionais que o ligam aos restantes haplótipos (Figura II.11). No caso presente, assumiu-se, tal como no artigo 2, que a variação haplotípica foi gerada por recombinação nos *loci* CBG e PCI(TG)n, e por mutação no *locus* PI(TG)n. Na linhagem Z, o haplótipo ancestral (HZ1) pôde ser facilmente inferido por ser o que apresenta a combinação dos alelos mais comuns dos três microssatélites (Tabela II.1 A). Na linhagem S existem três haplótipos equifrequentes (HS2, HS3 e HS4) que diferem no alelo do microssatélite PCI(TG)n que lhes está associado (Tabela II.1 B). No entanto, a análise da correlação entre as frequências alélicas de PCI(TG)n na linhagem S e na população geral, mostra claramente que é o alelo B10 (127bp) que mais se afasta da média, pelo que HS3 é muito provavelmente o haplótipo ancestral (Figura II.12).

Conhecida a genealogia dos haplótipos associados a cada linhagem, é possível estimar as idades de S e Z utilizando a expressão (Thomson *et al.*, 2000):

$$T = \sum_{i=1}^{n} xi \; /(n \mu) \qquad \text{II.1}$$

em que T é idade em gerações da linhagem a datar, xi o número de diferenças mutacionais e recombinacionais entre o i-ésimo haplótipo e o haplótipo ancestral, n o número total de haplótipos e μ uma taxa agregada de modificação haplotípica que é a soma das taxas de mutação e das taxas efectivas de recombinação entre os microssatélites e o *locus* da PI.

Nos 79 alelos Z, há 14 haplótipos derivados com um total de 67 alterações (mutações e recombinações): 2 em CBG, 16 em PI(TG)n e 49 em PCI(TG)n a que correspondem mutabilidades relativas de 1:8:24,5, respectivamente. Aceitando que a taxa de mutação em PI(TG)n é de 0,001 e tendo em conta as diferenças de mutabilidade entre os três *loci*, obtêm-se taxas efectivas de recombinação de 0,000125

Tabela II.1- Distribuição dos haplótipos definidos pelos *loci* PI(TG)n, PCI(TG)n e CBG em cromossomas Z (**A**) e S (**B**). A nomenclatura dos alelos é a mesma do artigo 2. Os haplótipos que partilham alelos dos microssatélites CBG e PI(TG)n estão destacados com a mesma cor. Os haplótipos ancestrais estão assinalados a negro. N- número de cromossomas.

A

	CBG	PI(TG)n	PCI(TG)n	N
HZ1	**CBG1**	**P4-161**	**B13-121**	**21**
HZ2	CBG1	P4-161	B12-123	9
HZ3	CBG1	P4-161	B10-127	8
HZ4	CBG1	P4-161	B7-133	11
HZ5	CBG1	P4-161	B6-135	6
HZ6	CBG1	P4-161	B3-141	4
HZ7	CBG1	P4-161	B2-143	1
HZ8	CBG1	P4-161	B1-145	1
HZ9	CBG1	P5-159	B13-121	8
HZ10	CBG1	P5-159	B12-123	1
HZ11	CBG1	P5-159	B10-127	1
HZ12	CBG1	P5-159	B7-133	5
HZ13	CBG1	P3-163	B12-123	1
HZ14	CBG2	P4-161	B13-121	1
HZ15	CBG2	P4-161	B7-133	1
			Total	**79**

B

	CBG	PI(TG)n	PCI(TG)n	N
HS1	CBG2	P4-161	B13-121	3
HS2	CBG2	P4-161	B12-123	11
HS3	**CBG2**	**P4-161**	**B10-127**	**11**
HS4	CBG2	P4-161	B7-133	11
HS5	CBG2	P4-161	B6-135	1
HS6	CBG2	P4-161	B3-141	1
HS7	CBG2	P4-161	B1-145	1
HS8	CBG2	P5-159	B13-121	2
HS9	CBG2	P5-159	B12-123	2
HS10	CBG2	P5-159	B10-127	1
HS11	CBG2	P5-159	B7-133	2
HS12	CBG2	P5-159	B6-135	1
HS13	CBG2	P3-163	B13-121	1
HS14	CBG2	P3-163	B12-123	3
HS15	CBG2	P3-163	B10-127	1
HS16	CBG2	P3-163	B7-133	2
HS17	CBG2	P3-163	B6-135	1
HS18	CBG2	P6-157	B10-127	1
HS19	CBG2	P6-157	B7-133	2
HS20	CBG1	P4-161	B6-135	1
			Total	**59**

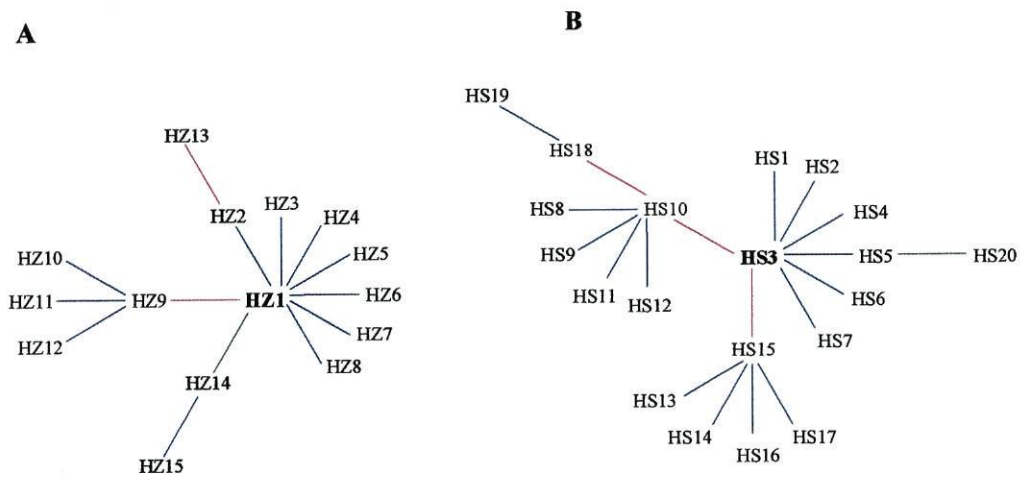


Figura II.11- Filogenia dos haplótipos da tabela II.1 construída segundo critérios de máxima parcimónia. **A-** Haplótipos de cromossomas Z. **B-** Haplótipos de cromossomas S. Os haplótipos ancestrais estão assinalados a negro. A cada passo mutacional ou recombinacional corresponde uma linha, vermelha, azul ou verde consoante a sua ocorrência no *locus* PI(TG)n, PCI(TG)n ou CBG, respectivamente.

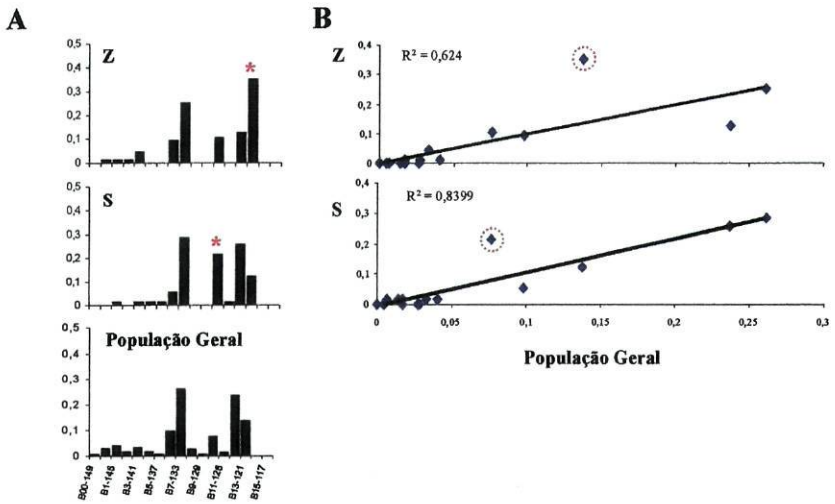


Figura II.12- A - Distribuição das frequências dos alelos do *locus* PCI(TG)n no seio dos alelos Z e S e na população portuguesa em geral. **B** - Gráficos de correlação entre as distribuições do microssatélite observadas nos alelos Z e S (ordenadas) e na população geral (abcissas). Os alelos de PCI(TG)n originalmente associados aos produtos génicos Z e S (B13 e B10) estão assinalados a vermelho.

e 0,0031 para CBG e PI(TG)n, respectivamente. Nestas condições, μ=0,0042 e a estimativa de T será igual a 202 gerações ou ~ 6000 anos.

Nos 59 alelos S, há 19 haplótipos derivados com um total de 68 alterações (mutações e recombinações): 1 em CBG, 22 em PI(TG)n e 45 em PCI(TG)n, a que correspondem mutabilidades relativas de 1:22:45, respectivamente. Usando de novo uma taxa de mutação em PI(TG)n igual a 0,001, as taxas efectivas de recombinação para CBG e PI(TG)n serão de 0,00005 e 0,002, com μ=0,003 e T=384 gerações ou ~11 500 anos.

As idades dos alelos Z e S estimadas com esta nova abordagem são inteiramente compatíveis com os valores obtidos no artigo 2 (Tabela II.2). As datas da mutação Z calculadas em Portugal com diferentes tipos de pressupostos e metodologias variam entre 3 630 e 7 440 anos sugerindo que, à semelhança da mutação ΔF508 responsável pela fibrose quística (Serre *et al.*, 1990), a dispersão deste alelo pode ter ocorrido durante a expansão populacional do Neolítico, ou no período de intensificação de movimentos migratórios e crescimento urbano que se lhe seguiu. No norte da Europa,

Tabela II.2- Estimativas das idades dos alelos Z e S obtidas a partir da análise dos *loci* CBG, PI(TG)n e PCI(TG)n, com base em duas metodologias distintas.

Alelos	Populações	Estimativas de idade em anos[a]	
		Análise independente[b]	Análise simultânea[c]
Z	Portugal	3630-7440	6060
	Europa Norte	2790-4710	---
S	Portugal	8370-16335	11490
	País Basco	5460	---

[a] 30 anos =1 geração.
[b] Média dos valores apresentados na tabela 3 do artigo 2.
[c] Estimativa efectuada com base na variação haplótipica definida pelos três microssatélites simultaneamente.

apesar da frequência de Z ser mais elevada (Figura II.4), obtêm-se datações mais recentes que variam entre 2 790 e 4 710 anos (Tabela II.2), possivelmente devido a uma maior exposição das populações dessas regiões aos efeitos da deriva genética. É assim pouco provável que a dispersão de Z na Europa se tenha dado a partir de migrações no sentido Norte-Sul depois de uma origem na Escandinávia, tal como foi proposto apenas com base nas suas frequências (Hutchison, 1998). Também neste aspecto o padrão do alelo Z é semelhante ao da mutação ΔF508, que é mais frequente nas populações norte-europeias, onde foram calculadas datações de 2 000 anos, do que no Sul da Europa, onde se observa mais diversidade haplotípica e a idade da mutação foi estimada em cerca de 6 900 anos (Serre *et al.*, 1990; Guo e Xiong, 1997). Serre *et al.* (1990), explicaram estas diferenças de idade com base na difusão da mutação ΔF508 através de duas ondas migratórias neolíticas distintas. No sul da Europa, a dispersão teria ocorrido ao longo da costa mediterrânica e teria sido menos afectada por estrangulamentos acentuados de efectivo populacional. A chegada ao norte da Europa teria resultado de uma sucessão de efeitos de fundador ao longo de uma rota continental alinhada segundo um eixo definido pelos rios Danúbio e Reno. A ideia de diferentes intensidades de deriva genética é também apoiada pelas estimativas de Bertranpetit e Calafell (1996), que sugeriram que, durante o período de dispersão das mutações da fibrose quística, o efectivo populacional das populações do sul da Europa era cerca de quatro vezes superior ao das populações do norte e centro do continente.

É provável que o padrão geográfico das idades e frequências do alelo Z também se possa dever a este tipo de diferença.

Quanto às datas da mutação S, o intervalo de variação entre 8 370 e 16 335 anos (Tabela II.2) indica que o alelo se pode ter originado na transição do final do Paleolítico para o Mesolítico, depois do clímax do último período glaciar ocorrido há cerca de 18 000 anos. Recentemente, Torroni *et al.* (2001), detalhando as conclusões de um trabalho anterior (Torroni *et al.*, 1998), identificaram uma linhagem de DNA mitocondrial (haplogrupo V) com cerca de 11 200 a 16 300 anos e um gradiente de frequências muito semelhante ao do alelo S, centrado na Península Ibérica e no sul de França e orientado de sudoeste para nordeste (Figura II.13). Este gradiente foi interpretado como o resultado do movimento de populações que se terão concentrado nos refúgios glaciares da Península Ibérica e posteriormente reexpandido para o resto do continente europeu após a última glaciação (Torroni *et al.*, 2001). A concordância com as datações e as distribuições do haplogrupo V, sugere que o alelo S da α1-antitripsina também pode ser um marcador dessa reexpansão.

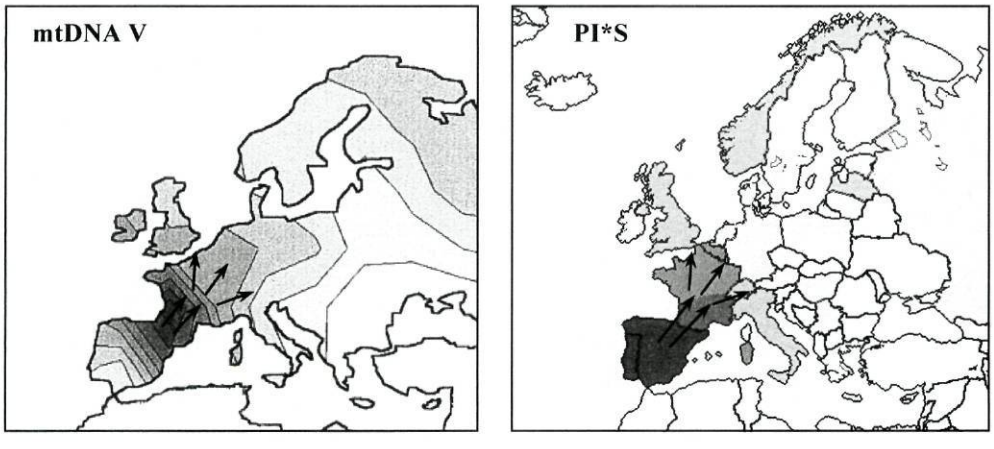


Figura II.13- Distribuição das frequências do haplogrupo V do DNA mitocondrial (adaptado de Torroni *et al.*, 2001) e do alelo PI*S no continente europeu. A redução da intensidade de cor corresponde ao gradiente decrescente das frequências. As setas indicam o sentido de um provável movimento migratório que terá favorecido a difusão destes marcadores genéticos a partir da Península Ibérica.

2.2 Espectro mutacional da α1-antitripsina

2.2.1 Mutações identificadas em alelos raros

A sequenciação de DNA de 36 variantes raras da α1-antitripsina, detectadas na sua maioria por focagem isoléctrica híbrida, permitiu identificar 14 mutações distribuídas por 15 alelos diferentes nas populações de São Tomé, País Basco e Portugal (Tabela II.3, Figura II.14). Foi ainda possível caracterizar a diversidade haplotípica associada a cada variante através da análise dos microssatélites estudados no artigo 2 (CBG, PI(TG)n e PCI(TG)n) e de uma posição polimórfica A/G localizada no nucleotídeo 95, no intrão IC (Long *et al.*, 1984; Cox *et al.*, 1985; Faber *et al.*, 1994). As mutações identificadas em São Tomé e no País Basco foram observadas nas amostras descritas no artigo 2. A amostra portuguesa inclui indivíduos clinicamente triados com suspeita de défice da proteína. Os pormenores dos métodos de focagem isoléctrica e de sequenciação encontram-se nos artigos 3 a 6.

Das 14 mutações identificadas, 5 foram pela primeira vez observadas no decurso deste trabalho: IVS1C+1G→A, Ser47Arg, Arg281del, Pro362His e Pro369Ser.

As mutações IVS1C+1G→A e Pro369Ser foram identificadas na amostra portuguesa e provocam deficiências acentuadas da proteína. A transição IVS1C+1G→A, caracterizada no artigo 6, causa a ausência total de α1-antitripsina no plasma devido à alteração do processamento do mRNA e está associada ao enfisema pulmonar. Esta mutação ainda não foi descrita noutras populações. A substituição Pro369Ser, analisada no artigo 5, ocorre numa posição altamente conservada (Figura II.15) (Huber e Carrell, 1989; Irving *et al.*, 2000) e, à semelhança do alelo Z, leva à retenção intrahepática da proteína, podendo estar associada tanto à doença hepática como ao enfisema pulmonar. Inicialmente depositada na base de dados HGMD (The Human Gene Mutation Database Cardiff http://archive.uwcm.ac.uk/uwcm/mg/hgmd) (Rocha *et al.*, 1999), a mutação Pro369Ser foi independentemente identificada na Alemanha por Poller *et al.* (1999), numa variante designada por Mwürzburg, também num alelo base M1Val213. Recentemente, Jardi *et al.* (2000), observaram a mutação em Espanha, num alelo base M1Ala213, e designaram a variante correspondente por Mvall d'hebron. A posição 369 é também modificada pela mutação Pro369Leu, anteriormente identificada noutras regiões da Europa (Hofker *et al.*, 1989; Kalsheker *et al.*, 1992) e agora observada na amostra portuguesa (Tabela II.3). No entanto, ao

Tabela II.3- Mutações raras da α1-antitripsina identificadas no presente trabalho e respectiva caracterização haplotípica[a].

Mutação	Variante	Alelo Base	Conc. plasmatica	Haplótipo	CBG[b]	PI(TG)n[b]	Intão IC[c]	PCI(TG)n[b]	N	Origem	Outras ocorrências
IVS1C+1G-A	Q0porto	M1Val213	0%		CBG1	P8-153	G	B7-133	1	Portugal	
Arg39Cys	I	M1Val213	60%	H1	CBG1	P11-147	A	B7-133	2	Portugal	Europa
				H2	CBG1	P11-147	A	B2-143	1	Portugal	
				H3	CBG1	P11-147	A	B6-135	1	Portugal	
				H4	CBG1	P11-147	A	B9-129	1	Portugal	
				H5	CBG1	P8-153	G	B8-127	1	Portugal	
				H6	CBG2	P10-149	A	B13-121	1	Portugal	
Ser47Arg	Slisboa	M2	100%	H1	CBG2	P8-153	A	B2-143	1	Portugal	
				H2	CBG2	P8-153	A	B9-129	1	Portugal	
Phe52del	Mmalton	M2	12%	H1	CBG2	P7-155	A	B7-133	1	Portugal	Europa/EUA
				H2	CBG2	P8-153	A	B12-123/B11-125	1	Portugal	
Arg101His	T	S	60%	H1	CBG2	P3-163	A	B6-135	1	Portugal	Europa
ou		ou		H2	CBG2	P3-163	A	B6-135	1	São Tomé	
Glu264Val		M4		H3	CBG2	P3-163	A	B12-123	1	São Tomé	
Arg101His	M4	M1Val213	100%	H1	CBG2	P10-149	A	B7-133	1	Portugal	Europa
ou		ou		H1	CBG2	P10-149	A	B7-133	3	País Basco	
Asp376Glu		M2		H2	CBG2	P10-149	A	B12-123	1	Portugal	
				H3	CBG2	P5-159	A	B7-133/B1-145	1	País Basco	
				H4	CBG2	P8-153	G	B6-135	1	Portugal	
				H5	CBG2	P8-153	G	B7-133	1	Portugal	
Gly148Arg	V	M1Val213	100%		CBG2	P10-149	A/G	B9-129/B7-133	1	São Tomé	África/Europa
Asp256Val	Plowell	M1Val213	30%		CBG2	P2-165	G	B6-135	1	Portugal	Europa/EUA
Arg281del	Ieuskadi	M1Val213	100%		CBG2	P8-153/P4-161	A/G	B7-133/B2-143	1	País Basco	
Asp341Asn	Pdonauwoerth	M1Val213	100%	H1	CBG2	P10-149	A	B8-127	1	Portugal	Europa
				H2	CBG2	P10-149	A	B7-133/B6-135	1	Portugal	
Glu342Lys	Zaugsburg	M2	15%		CBG2	P8-153	nd	nd	1	Portugal	Europa
Leu353fram	Q0ourém	M3	0%		CBG2	P7-155	A	B6-135	1	Portugal	
Pro362His	São Tomé	M1Val213	100%		nd	P9-151/P7-155	A/G	B4-139	1	São Tomé	
Pro369Leu	Mheerlen-2	M1Val213	2%		CBG1	P8-157	A	B3-137	3	Portugal	Europa
Pro369Ser	Mwürzburg	M1Val213	15%	H1	CBG1	P8-153	G	B7-133	1	Portugal	Europa
				H2	CBG1	P8-153	G	B3-137	1		

Intrão IC[c]	A	G
M1Ala213	0,77	0,23
M1Val213	0,27	0,73
M3	0,11	0,89
M2	0,84	0,16
S	0,05	0,95
Z	0,86	0,14

[a] As cores delimitam regiões partilhadas por diferentes haplótipos associados a cada mutação.
[b] A nomenclatura dos alelos dos microssatélites é a do artigo 2.
[c] Polimorfismo A/G do intrão IC localizado no nucleotídeo 95 da sequência do gene (Long *et al.*, 1984). As frequências deste polimorfismo nos diferentes alelos de PI apresentam-se na tabela enquadrada.
nd- não determinado N- número de haplótipos.

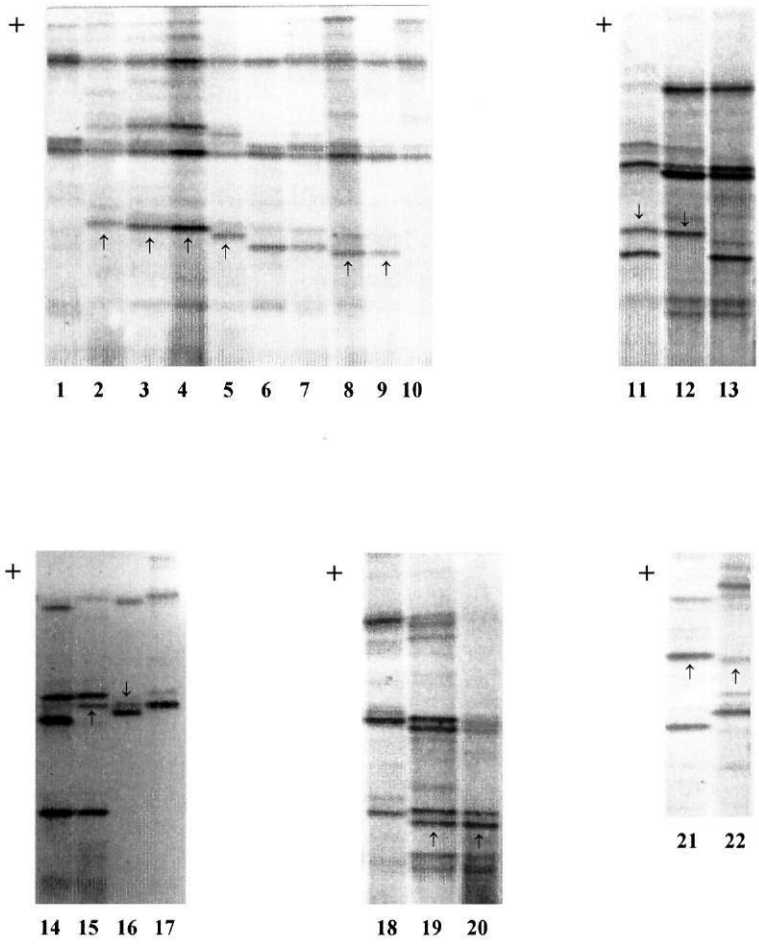


Figura II.14- Padrões electroforéticos dos variantes raros da α1-antitripsina. As amostras numeradas de 1 a 22 têm os seguintes fenótipos: **1**- M1; **2**-M1Sãotomé; **3** e **4**-M1Pdonauwoerth; **5**- M1V; **6** e **7**-M1S **8** e **9**-M1T; **10**-M1; **11**- SPlowell; **12**- M1Plowell; **13**- M1S; **14**-M2S; **15**-MwürzburgS; **16**-M3Mwürzburg; **17**-M1; **18**-M1T; **19**-M3Slisboa; **20**-MwürzburgSlisboa; **21**-M2Ieuskadi; **22**- M1I. As bandas correspondentes aos diferentes produtos génicos raros estão assinaladas com ↑ ou ↓.

contrário de Pro369Ser, esta substituição não está associada a qualquer banda identificável por focagem isoeléctrica e o único estudo de material de biópsia hepática efectuado num doente adulto com enfisema refere a ausência de acumulação intracelular de α1-antitripsina associada a Pro369Leu (Hofker *et al.*, 1989). Estes resultados mostram que apesar da perda do resíduo Pro369 provocar a deficiência

Figura II.15- Alinhamentos parciais da sequência polipeptídica da α1-antitripsina humana em que se destacam os resíduos de aminoácidos 47, 362 e 369 onde ocorreram as substituições que caracterizam as variantes Slisboa, São Tomé e Mwürzburg. **A-** Alinhamento com sequências ortólogas de sete espécies de mamíferos. **B-** Alinhamento com sequências parálogas de outras proteínas humanas. As sequências foram retiradas da base de dados NCBI (http://www.ncbi.nlm.nih.gov/) e têm as seguintes referências: ***H. sapiens*** (P01009), ***P. cynocephalus*** (P01010), ***C. aethiops*** (BAA20264), ***S. scrofa*** (CAA61259); ***O. aries*** (CAA33561), ***E. caballus*** (AAC83412), ***M. auratus*** (BAA08557), ***M. Musculus*** (P07758), **PI4-** calistatina (P29622); **PCI-** inibidor da proteína C (P05154); **TBG -** globulina transportadora da tiroxina (P05543); **CBG -** globulina transportadora dos corticoesteróides (P08185); **AACT -** antiquimiotripsina (P01011); **ZPI -** inibidor de proteases da proteína Z-dependente (Q9UK55) e **AT3-** antitrombina 3 (P01008). O alinhamento das sequências foi efectuado com o programa ClustalW (http://www.ebi.ac.uk/clustalw).

severa da proteína, o tipo de aminoácido que o substitui pode influenciar a intensidade e o mecanismo dessa deficiência.

A substituição Ser47Arg ocorre no local de N-glicosilação Asn-Ser-Thr formado pelos aminoácidos 46, 47 e 48, até agora, só foi observada na amostra portuguesa. Neste tipo de locais de glicosilação há uma sequência consensual Asn-X-Thr em que a segunda posição (X) pode ser ocupada por qualquer aminoácido, com excepção de prolina, sem que haja perturbação do processo de glicosilação (Creighton, 1992). O padrão de focagem isoeléctrica do produto génico correspondente a Ser47Arg está associado a concentrações plasmáticas normais da α1-antitripsina e é o que seria de esperar da substituição de um aminoácido neutro por um aminoácido positivo, confirmando a ausência de alterações no padrão de glicosilação da proteína (Figura II.14). Tendo em conta que o ponto isoeléctrico desse produto génico é semelhante ao

do alelo S e atendendo ao local de nascimento do probando em que foi pela primeira vez observado, o alelo correspondente foi designado por Slisboa. Um dos dois alelos Slisboa agora detectados foi observado em heterozigotia com Pro369Ser, numa criança com sintomas de colestase neonatal de evolução favorável e evidência de acumulação difusa de α1-antitripsina nos hepatócitos, sem formação de grânulos evidentes. Como a mutação Ser47Arg está associada a concentrações normais, é provável que esta acumulação se deva apenas à substituição Pro369Ser (artigo 5). O baixo grau de conservação da posição 47 em outras moléculas ortólogas e parálogas sugere que a substituição Ser47Arg não é patogénica (Figura II.15) (Huber e Carrell, 1989; Stein e Carrell, 1995; Irving *et al.*, 2000).

A delecção Arg281del, descrita no artigo 4, foi observada num único indivíduo da amostra basca e é aqui designada por Ieuskadi de acordo os critérios acima expostos. A presença de uma banda de focagem isoeléctrica com intensidade normal (Figura II.2.1) indica que a variante não provoca a acumulação da proteína nem uma maior susceptibilidade à degradação intra ou extra-celular. No entanto, dado que não se pode excluir que a perda do resíduo Arg na posição 281 não cause uma diminuição da capacidade inibidora de proteases, só a realização de ensaios funcionais permitirá avaliar as consequências desta delecção.

O mesmo problema coloca-se em relação à substituição Pro362His, detectada na amostra de São Tomé e descrita no artigo 3. Esta mutação também está associada a concentrações normais da proteína, mas pode originar uma molécula disfuncional, uma vez que o resíduo Pro362 se situa numa região rica em prolinas (posições 357, 361, 362) que definem o domínio catalítico da α1-antitripsina no qual o aminoácido Met358 é o centro activo (Stein e Carrell, 1995). Com base em argumentos semelhantes, Faber *et al.* (1994), propuseram que a substituição Pro→Thr do alelo Loffenbach, que também ocorre na posição 362, pode diminuir a capacidade inibidora da proteína apesar de estar associada a concentrações plasmáticas normais. A análise dos padrões de conservação também parece indicar que a posição 362 é funcionalmente relevante (Figura II.15). A conservação do resíduo Pro362 em sequências ortólogas, juntamente com Pro357 e Pro361, sugere que este aminoácido pode ser essencial para a actividade inibidora característica da α1-antitripsina (Figura II.15 A). A ausência de conservação de Pro362 em moléculas parálogas, à semelhança

do que acontece com Pro357 e Pro361, indica que a alteração do resíduo 362 pode estar associada ao processo de divergência funcional que originou outras proteínas da família SERPIN com diferente especificidade de substrato (Figura II.15 B).

Todas as restantes mutações agora encontradas foram já anteriormente descritas noutras populações (Tabela II.14.1). A mutação Leu353framStop376, que caracteriza o alelo nulo Q0ourém analisado no artigo 6, ocorre num alelo base diferente do originalmente descrito num doente de origem canadiana (Cox e Levison, 1988; Curiel *et al.*, 1989b) e resulta de um fenómeno de recorrência mutacional. As outras mutações foram observadas em alelos base iguais aos de variantes previamente identificados em populações de origem europeia (Tabela II.3) e não se pode excluir que o fluxo génico interpopulacional tenha tido um papel essencial na sua actual distribuição geográfica.

A caracterização de polimorfismos flanqueantes permitiu analisar a homogeneidade haplotípica das várias mutações e aprofundar a discussão de hipóteses alternativas sobre a sua origem e dispersão (Tabela II.3). A maior parte dos diferentes haplótipos associados a cada mutação partilha os alelos da posição polimórfica A/G e dos microssatélites CBG e PI(TG)n e apenas se distingue quanto ao microssatélite PCI(TG)n. Este padrão indica que as mutações raras são relativamente homogéneas e que a sua diversidade haplotípica se deve essencialmente a recombinação entre os subagrupamentos CBG-PIL-PI e PI4-PCI-AACT, que se encontram mais distanciados no cromossoma 14q32.1 (Figura II.1). No entanto, os alelos I e M4 têm uma heterogeneidade considerável. No caso do alelo I, resultante da substituição Arg39Cys, os haplótipos H5 e H6 diferem do grupo constituído pelos haplótipos H1 a H4 em três dos quatro *loci* analisados e sugerem a possibilidade de recorrência da mutação que provoca aquela alteração de aminoácidos. O alelo M4, que combina mutações específicas de mais do que um alelo comum e pode ter sido originado a partir de M1Val213 ou M2 (Figura II.16 A), está associado a dois grupos de haplótipos definidos pela posição polimórfica A/G : H1 a H3 e H4 e H5 (Tabela II.3). Os haplótipos H1 e H2 do alelo M4 têm um dos alelos do microssatélite PI(TG)n mais comuns entre o variante M1Val213, bem como o alelo A da posição polimórfica A/G, que tem uma frequência de 84% entre as variantes M2 (Tabela II.3). É assim provável que a linhagem M4 a que estes haplótipos estão associados tenha resultado da

recombinação entre M1Val213 e M2 (Figura II.16 A). Os alelos do microssatélite PI(TG)n e da posição polimórfica A/G nos haplótipos H5 e H6 são ambos comuns em M1Val213 e é possível, neste caso, que o alelo M4 possa ter resultado da recorrência da mutação causadora da substituição Arg101His (ver secção -2.2.2.3). À semelhança de M4, o alelo T, que foi observado em Portugal e São Tomé, também combina mutações específicas de mais do que um alelo comum (Tabela II.3). Faber *et al.* (1994), sugeriram que o alelo T resultou da recorrência da mutação característica da variante S numa sequência base M4. No entanto, a associação com o alelo P3-163bp do microssatélite PI(TG)n, que está ausente de M4 e ocorre, ainda que com baixa frequência, no alelo S (artigo 2), indica que este alelo teve um papel fundamental na origem da variante T. A presença na posição polimórfica A/G do nucleotídeo A, que é raro nos alelos S mas predomina nos alelos M2 e M4 (Tabela II.3), sugere que a recombinação entre S e M2 ou S e M4 foi o processo que provavelmente originou a variante T (Figura II.16 B). A homogeneidade haplotípica deste alelo indica, por outro lado, que a sua ocorrência em São Tomé e Portugal terá resultado de miscigenação.

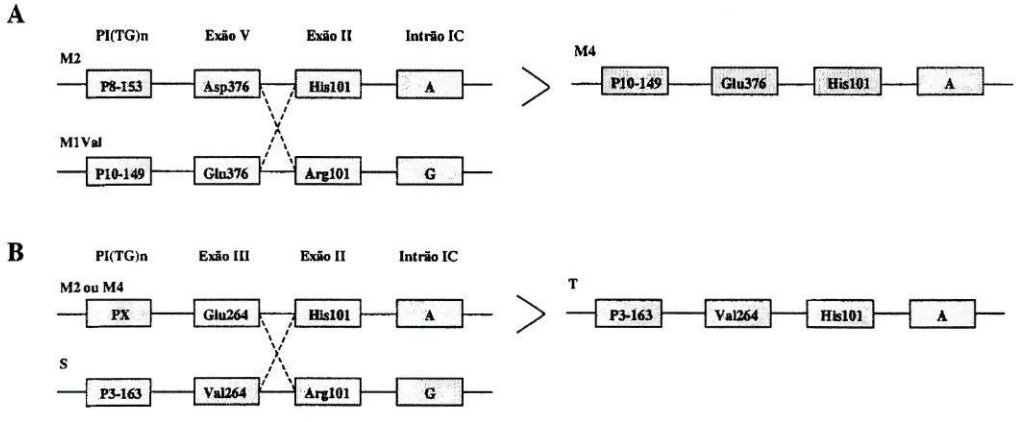


Figura II.16- Representação esquemática dos processos de recombinação que podem estar na origem dos produtos génicos M4 e T. O cruzamento define a região na qual terá ocorrido a recombinação e os diferentes blocos correspondem às posições polimórficas informativas.

2.2.2 Propriedades gerais do espectro mutacional

Com o objectivo de analisar as propriedades gerais do espectro mutacional da α1-antitripsina e identificar os principais factores que o influenciam, foi possível compilar um total de 54 mutações diferentes a partir da consulta da literatura e das bases de dados disponíveis (Tabela II.4). Destas 54 mutações, 24 não estão associadas a deficiência de α1-antitripsina e são aqui classificadas como não patogénicas, embora nalguns casos, como por exemplo nas alterações que envolvem o resíduo Pro362, a ausência de défice não garanta que não haja perturbações funcionais na proteína (ver acima). Das restantes 30 mutações, uma, Met358Arg, altera o centro do local reactivo e transforma a α1-antitripsina numa potente antitrombina que provoca desequilíbrios da hemostase e está associada a doença hemorrágica fatal (Owen *et al.*, 1983). As outras 29 causam deficiências mais ou menos acentuadas da proteína.

As substituições nucleotídicas e as inserções ou delecções (indels) representam 77% (23/30) e 23% (7/30) das mutações patogénicas, respectivamente. Setenta e oito por cento (18/23) das substituições nucleotídicas patogénicas causam alterações de aminoácidos, 13% (3/23) introduzem um codão de terminação prematuro e 9% (2/23) afectam o processamento normal do mRNA (Tabela II.4). Com excepção da delecção Arg281del agora observada na amostra do país Basco, todas as mutações não patogénicas resultam de substituições de aminoácidos (Tabela II.4).

A análise de bases de dados que reúnem os espectros mutacionais de diferentes *loci* tem vindo a demonstrar que os vários tipos de mutação não ocorrem ao acaso e tendem a concentrar-se num pequeno número de regiões hipermutáveis, definidas por segmentos específicos das sequências de DNA (Cooper e Krawczak 1990; Cooper et al., 1995; Krawczak et al., 1998). A fim de esclarecer a influência das regiões potencialmente hipermutáveis na diversidade alélica da α1-antitripsina, realizou-se um estudo da associação entre essas regiões e os principais tipos de mutações observadas.

Tabela II.4 - Mutações no gene da α1-antitripsina identificadas neste trabalho e disponíveis em bases de dados[a].

Mutação	Sequência de DNA[b]	Consequências	Nº
Ser-19Leu	TCG→TTG		1
Asp2Ala	GAT→GCT		2
Ala34Thr	GCC→ACC		3
Tyr38Stop	TAC→TAA	ausência de proteína	4
Arg39Cys	CGC→TGC	deficiência proteica (60%)	5
Leu41Pro	CTG→CCG	deficiência proteica (4%)	6
Ser45Phe	TCC→TTC		7
Ser47Arg	AGC→CGC		8
Phe52del	TTCttcTCC	deficiência proteica (12%) + acumulação intrahepática	9
Ser53Phe	TCC→TTC	deficiência proteica (7%) + acumulação intrahepática	10
Ala60Thr	GCC→ACC		11
Gly67Glu	GGG→GAG	deficiência proteica (nd)	12
Thr68Ile	ACC→ATC	ausência de proteína	13
Thr85Met	ACG→ATG	deficiência proteica (60%)	14
Pro88Thr	CCG→ACG		15
Ile92Asn	ATC→AAC	ausência de proteína	16
Arg101His	CGT→CAT		17
Gly115Ser	GGC→AGC	deficiência proteica (nd)	18
Leu118Trp	CTG→TTG		19
Gly148Trp	GGG→TGG		20
Gly148Arg	GGG→AGG		21
Gln156Glu	CAG→GAG		22
Tyr160framStop160	TAcGTG	ausência de proteína	23
Trp194Stop	TGG→TGA	ausência de proteína	24
Glu204Lys	GAG→AAG		25
Ala213Val	GCG→GTG		26
Lys217Stop	AAG→TAG	ausência de proteína	27
Met221Thr	ATG→ACG		28
Arg223Cys	CGT→TGT	deficiência proteica (15%)	29
Pro255Thr	CCT→ACT		30
Asp256Val	GAT→GTT	deficiência proteica (30%)	31
Asp256Asp	GAT→GAC		32
Glu264Val	GAA→GTA	deficiência proteica (60%)	33
Arg281del	CagaAGG ou CAGaagG		34

Tabela II.4 -(Continuação)

Mutação	Sequência de DNA[b]	Consequências	Nº
Leu318framStop334	CtcTCC	ausência de proteína	35
Gly320Glu	GGG→GAG	ausência de proteína	36
Ser330Phe	TCC→TTC		37
Ala336Thr	GCT→ACT	deficiência proteica (60%)	38
Asp341Asn	GAC→AAC		39
Glu342Lys	GAG→AAG	deficiência proteica (15%) + acumulação intrahepática	40
Leu353framStop376	TTT^T	ausência de proteína	41
Met358Arg	ATG→AGG	função alterada	42
Pro362Thr	CCC→ACC		43
Pro362His	CCC→CAC		44
Pro362framStop373	CccGA	ausência de proteína	45
Pro362framStop376	CCC^C	ausência de proteína	46
Glu363Lys	GAG→AAG		47
Pro369Leu	CCC→CTC	deficiência proteica (2%)	48
Pro369Ser	CCC→TCC	deficiência proteica (15%) + acumulação intrahepática	49
Glu376Asp	GAA→GAC		50
Pro391His	CCC→CAC	deficiência proteica (nd)	51
Zonas de junção exão/intrão			
IVS1C+1GA	GT→AT	ausência de proteína	52
IVS2+1GT	GT→TT	ausência de proteína	53
IVS2-1Gdel	AgGA	ausência de proteína	54

[a] SwissProt (http://www.expasy.ch/sprot/), Brantly (1996) e WHO (1997).
[b] As substituições estão sublinhadas, as delecções estão representadas com letras minúsculas e as inserções são precedidas por ^.
nd- não determinado.

2.2.2.1 *Inserções e delecções (indels)*

No total de mutações compiladas observaram-se 5 delecções e 2 inserções na região codificante do gene da α1-antitripsina. Das 5 delecções, 3 (Tyr160framStop160, Arg281del e Leu318framStop334) foram encontradas a uma distância igual ou inferior a 5bp de uma sequência TG(A/G)(A/G)(G/T)(A/C), que corresponde a locais de paragem da síntese de DNA mediada pela polimerase α e é um factor conhecido de promoção deste tipo de mutações (Krawczak e Cooper, 1991).

Tendo em conta que nas 1258bp da porção codificante do gene 256bp estão ocupados pela sequência TG(A/G)(A/G)(G/T)(A/C) e pelos 10 bp flanqueantes, a percentagem de delecções que lhe estão associadas é cerca de 3 vezes superior à que seria de esperar por acaso [(1258x3)/(5x256)]. A delecção Phe52del ocorre numa região ATC-TTC-TTC-TCC definida pelos codões 50 a 53, em que há uma repetição do motivo TCT: A**TCT/TCT/TCT**CC. Este tipo de repetições, que favorece os desalinhamentos das sequências durante a replicação do DNA, é considerado o principal factor de mutagénese em microssatélites e também se encontra associado a delecções e inserções em porções codificantes do genoma (Cooper *et al.*, 1995). No caso da mutação Phe52del, a sequência ATC-TTC-TCC, que resulta na perda do codão 51 ou 52, pode ter sido provocada pela remoção de um trinucleotídeo TCT durante um desalinhamento.

As 3 indels restantes (Leu353framStop376, Pro362framStop373 e Pro362framStop376) são provocadas pela remoção ou inserção de um único nucleotídeo. A inserção Pro362framStop376 e a delecção Pro362framStop373 ocorrem numa região repetitiva (C)7 que abrange os codões 360 a 362. A inserção Leu353framStop376 ocorre numa sequência (T)5 nos codões 352 e 353. Os motivos com pelo menos 5 mononucleotídeos também são factores reconhecidos de promoção de inserções e delecções através de desalinhamentos, especialmente quando compostos por pirimidinas (Kunkel, 1985). No caso da PI as sequências em que o mesmo nucleotídeo ocorre 5 ou mais vezes seguidas correspondem a um total de 37 bp, o que conduz à estimativa de que a percentagem de indels de um único nucleotídeo associado a este tipo de sequências é cerca de 25,5 vezes superior à que seria de esperar por acaso (1258x3/37x4).

2.2.2.2 *Substituições nucleotídicas*

O dinucleotídeo CpG é desde há muito reconhecido como o principal motivo hipermutável associado às substituições de nucleotídeos e tem uma taxa de modificação que é cerca de 10 vezes mais elevada que a de qualquer outro dinucleotídeo (Cooper e Krawczack, 1990; Cooper *et al.*, 1995). Esta hipermutabilidade é essencialmente provocada pela metilação da citosina, que nos

genomas eucarióticos ocorre quase exclusivamente no motivo CpG, e que por desaminação origina timina (Cooper *et al.*, 1995; Krawczack *et al.*, 1998). Duas das consequências mais visíveis desta situação são a frequência inusual de substituições dos tipos CG→TG e CG→CA e a marcada subrepresentação de CpG naqueles genomas (Cooper e Krawczak, 1990; Cooper *et al.*, 1995).

Nas regiões codificantes do gene da α1-antitripsina há um total de 27 pares CpG quando seria de esperar 83 com base nas frequências de C e G. Esta diferença é altamente significativa (χ^2= 40,45; 1gl; p<0,001) e corresponde a um nível de subrepresentação de 32,5%, semelhante ao observado nas porções codificantes de outros *loci* onde, em média, a frequência observada de CpG é cerca de 37% do valor esperado (χ^2= 0,71; 1gl; p>0,25) (Cooper e Krawczak, 1989; Cooper e Krawczak, 1990).

A percentagem de substituições que ocorrem em dinucleotídeos CpG é de 24% (5/21) nas mutações patogénicas e de 39% (9/23) nas substituições não patogénicas (Tabela II.5). Estes valores não são significativamente diferentes (χ^2= 1,22; 1 gl; p>0,25) e levam a uma estimativa conjunta de ~ 32% (14/44) que também não se afasta da percentagem média de 25% encontrada em bases de dados que compilam o espectro das mutações patogénicas de vários *loci* autossómicos (χ^2= 1,09 ; 1 gl; p>0,25) (Krawczak *et al.*, 1998). A fracção das substituições em dinucleotídeos CpG que pertencem aos tipos CG→ TG ou CG→ CA, compatíveis com a desaminação da citosina promovida pela metilação, é de 93% (13/14) e assemelha-se à média de 90% calculada com outros *loci* (Cooper e Krawczak, 1989). Tendo em conta que os 27 pares CpG ocupam 54bp na região codificante da α1-antitripsina, a frequência de substituições nestes motivos é 7,4 vezes superior ao esperado [(1258x14)/(44x54)], o que representa uma mutabilidade ~ 10 vezes maior do que a de qualquer outro dinucleotídeo [(1204x14)/(30x54)] e está de acordo com as médias obtidas para outras regiões codificantes do genoma (Cooper e Krawczak, 1990; Cooper *et al.*, 1995).

O excesso de transições em relação a transversões é outra das consequências da hipermutabilidade dos pares CpG (Cooper e Krawczak, 1990; Cooper *et al.*, 1995; Krawczak *et al.*, 1998). Se as mutações se dessem ao acaso seria de esperar que 33% das substituções nucleotídicas fossem transições. No entanto, a percentagem média observada em bases de dados de mutações patogénicas é de 62,5% (Cooper e

Krawczak, 1990; Krawczak *et al.*, 1998). A análise dos padrões de substituição nucleotídica (Tabela II.5) mostra que a fracção total de transições nos exões da α1-antitripsina é de 63,6% (28/44) e não difere da média geral (χ^2= 0,02; 1 gl; p>0,75). Também não há diferenças significativas entre a percentagem de transições nas mutações patogénicas [66,7% (14/21)] e não patogénicas [60,9% (14/23)] (χ^2= 0,14; 1 gl; p>0,50). As mutações do tipo C→T ou G→A são claramente predominantes no seio das transições. Representam 89,3% (25/28) do total e 92,9% (13/14) e 85,7% (12/14) das transições patogénicas e não patogénicas, respectivamente (Tabela II.5). Este predomínio é em grande parte explicável pela ocorrência das substituições CG→TG e CG→CA nos dinucleotídeos hipermutáveis CpG, que correspondem a 52% (13/25) do total de transições do tipo C→T ou G→A (Tabela II.5).

Tabela II.5- Tipos de substituições nucleotídicas observadas na região codificante do gene da α1-antitripsina.

Substituições	Patogénicas	Não Patogénicas	Total
Transições			
C→T	7 (33,3%)	5 (21,7%)	12 (27,3%)
T→C	1 (4,8%)	2 (8,7%)	3 (6,8%)
A→G	0	0	0
G→A	6 (28,6%)	7 (30,4%)	13 (29,5%)
Total	14 (66,7%)	14 (60,9 %)	28 (63,6%)
Transversões			
T→G	1 (4,8%)	0	1 (2,3%)
G→T	0	1 (4,3%)	1 (2,3%)
T→A	1 (4,8%)	0	1 (2,3%)
A→T	3 (14,3%)	0	3 (6,8%)
C→G	0	1 (4,3%)	1 (2,3%)
G→C	0	0	0
C→A	2 (9,5%)	4 (17,4%)	6 (13,6%)
A→C	0	3 (13,0%)	3 (6,8%)
Total	7 (33,3%)	9 (39,1%)	16 (36,4%)
Total de substituições	21 (100%)	23 (100%)	44 (100%)
mCpG[a]			
CG→TG	3 (14,3%)	2 (8,7%)	5 (11,4%)
CG→CA	2 (9,5%)	6 (26,1%)	8 (18,2%)
Total	5 (23,8%)	8 (34,8%)	13 (29,5%)
CpG total	5 (23,8%)	9 (39,1%)	14 (31,8%)

[a] Mutações em dinucleotídeos CpG compatíveis com a desaminação da citosina metilada.

A utilização diferencial de codões sinónimos também pode ser influenciada pela hipermutabilidade do motivo CpG através da diminuição, por acção da selecção, do uso de codões que contêm este dinucleotídeo, como forma de evitar a substituição de aminoácidos essenciais ao funcionamento das proteínas (Cooper e Krawczak, 1990). Na α1-antitripsina verifica-se que, para os aminoáciodos treonina e alanina, há uma subrepresentação dos codões que contêm CpG (Tabela II.6), à semelhança da média dos genomas humanos e de outros vertebrados (Cooper e Krawczak, 1990). Nos aminoácidos serina, prolina e arginina, os níveis de subrepresentação não são significativos (Tabela II.6). Curiosamente, observaram-se mutações em 7 dos 8 codões com o dinucleotídeo CpG, três das quais associadas a défices da proteína: Arg39Cys, Arg223Cys e Thr85Met (Tabela II.6).

Tabela II.6- Distribuição dos codões sinónimos que podem ser influenciados pela hipermutabilidade do motivo CpG no gene da α1-antitripsina.

Aminoácido	Codão			
Serina N=25	TCT	5	(2%)	χ^2=2,93 1 gl, p>0.05
	TCC	9	(36%)	
	TCA	1	(4%)	
	TCG[a]	1	(4%)	
	AGT	0	(0%)	
	AGC	9	(36%)	
Prolina N=19	CCT	4	(21%)	χ^2=2,13 1 gl, p>0.10
	CCC	9	(47%)	
	CCA	4	(21%)	
	CCG	2	(11%)	
Treonina N=29	ACT	6	(21%)	χ^2=7,18 1 gl, p<0.01
	ACC	17	(58%)	
	ACA	5	(17%)	
	ACG[b]	1	(3%)	
Alanina N=27	GCT	11	(41%)	χ^2=6,53 1 gl, p<0.05
	GCC	11	(41%)	
	GCA	4	(15%)	
	GCG[b]	1	(4%)	
Arginina N=7	CGT	2	(29%)	χ^2=1,28 1 gl, p>0.25
	CGC	1	(14%)	
	CGA	0	(0%)	
	CGG	0	(0%)	
	AGA	3	(43%)	
	AGG	1	(14%)	

[a]Os dinucletídeos CG estão sublinhados.
[b]Codões subrepresentados.

Resumindo, os resultados da análise agora efectuada indicam que a distribuição dos principais tipos de mutações da α1-antitripsina (indels e substituições nucleotídicas) não é aleatória e é fortemente condicionada pelas sequências hipermutáveis, à semelhança do que é observado noutras regiões codificantes do genoma. Apesar dos espectros mutacionais destas regiões serem obtidos a partir do estudo de mutações patogénicas, a ausência de homogeneidade na distribuição da variação genética não parece resultar da selecção preferencial de mutações que causem consequências clínicas evidentes, uma vez que também é observada em comparações interespecíficas (Krawczak e Cooper, 1996; Miller e Kumar, 2001). É assim provável que a análise dos *loci* envolvidos em doenças hereditárias permita destacar algumas das principais propriedades mutacionais do genoma humano (Krawczak e Cooper, 1996; Miller e Kumar, 2001) A semelhança entre os padrões de substituições patogénicas e não patogénicas na α1-antitripsina apoia esta conclusão.

2.2.2.3 *Locais hipermutáveis no gene da α1-antitripsina*

No estudo acima efectuado, analisou-se a associação entre as mutações da α1-antripsina e locais hipermutáveis previamente reconhecidos, tendo-se contabilizado apenas um exemplo de cada mutação, a fim de evitar distorções causadas pela dificuldade em distinguir as mutações independentes das mutações idênticas por descendência. Contudo, esta abordagem tem a desvantagem de não identificar os episódios de recorrência mutacional e não permite reconhecer pontos hipermutáveis que ocorram em zonas diferentes das que foram anteriormente descritas. Com o objectivo de minorar estas lacunas, utilizou-se a evidência disponível para verificar se, entre os locais variáveis até agora descritos, há pontos que se destacam pela sua maior mutabilidade.

Uma das propriedades notáveis do espectro mutacional da α1-antripsina é a existência de vários casos de homoplasia, em que a mesma mutação ocorre em diferentes alelos-base, de modo que diversas variantes podem resultar da combinação de um repertório limitado de mutações (Hildesheim *et al.*, 1993). Na revisão agora efectuada foi possível identificar nove posições homoplásicas correspondentes aos codões 51/52, 101, 115, 148, 256, 342, 353, 362 e 369 (Figura II.17).

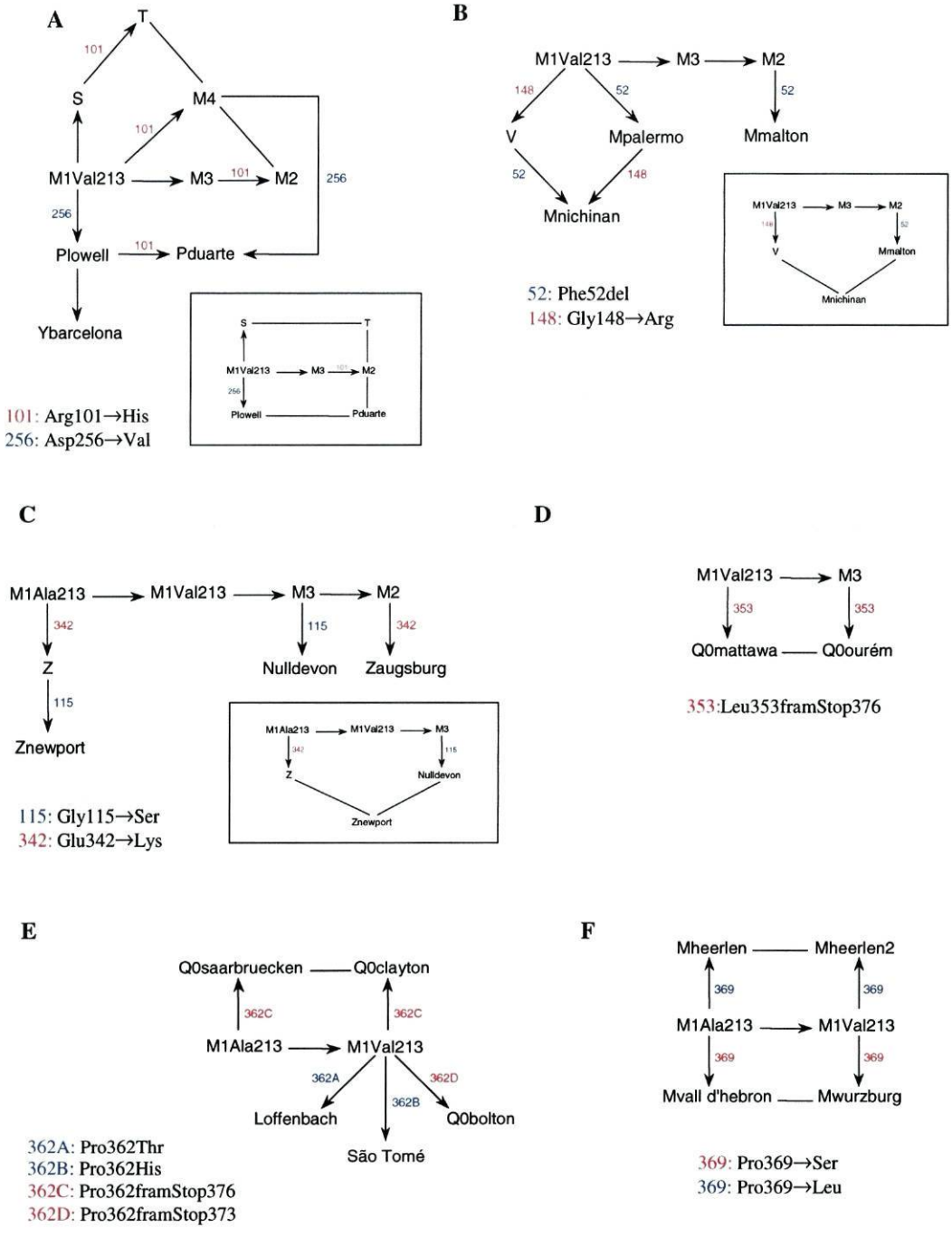


Figura II.17 - Filogenia dos alelos da α1-antitripsina associados a mutações homoplásicas. Os passos mutacionais entre alelos estão assinalados com setas e os processos recombinatórios estão assinalados com linhas. As posições homoplásicas estão representadas a cor. As opções filogenéticas que envolvem possíveis recombinações entre alelos que diferem em mais do que uma posição, estão representadas nos esquemas enquadrados.

Estas situações de homoplasia podem, em princípio, ser originadas, quer por recorrência mutacional, quer por recombinação intragénica, sendo difícil fazer uma distinção clara entre estas duas causas. No entanto, é possível estabelecer alguns critérios adicionais que, a verificarem-se, favorecem a hipótese de hipermutabilidade. Entre esses critérios, destacam-se os seguintes: 1) combinação no mesmo alelo de mutações que envolvam duas ou mais posições potencialmente hipermutáveis; 2) fixação do aminoácido derivado de uma mutação potencialmente recorrente numa sequência ortóloga ou paráloga; 3) ocorrência de duas ou mais mutações diferentes num mesmo codão ou nucleotídeo (Cooper *et al.*, 1995); 4) observação de mutações em codões vizinhos de posições potencialmente hipermutáveis; 5) observação de uma mutação potencialmente recorrente em populações com baixo fluxo genético (Krawczak *et al.*, 1995); 6) localização de mutações em regiões com motivos hipermutáveis identificados *a priori*.

Por exemplo, a análise da variação haplotípica de mutações raras mostrou que a recombinação esteve provavelmente na origem do alelo T e de parte dos alelos M4, ambos com a mutação Arg(C<u>G</u>T)→His(C<u>A</u>T) da posição homoplásica 101 (Figura II.16 e II.17 A). No entanto, a evidência adicional sugere que algumas das situações de homoplasia observadas para esta posição também podem ter sido causadas por mutação recorrente: a mutação Arg101(C<u>G</u>T)→His(C<u>A</u>T) ocorre num dinucleotídeo CpG; o aminoácido correspondente à posição 101 na sequência paráloga da calistatina e na sequência ortóloga da α1-antripsina de *Cercophitecus aethiops* é a histidina; no alelo Pduarte a mutação Arg101(C<u>G</u>T)→His(C<u>A</u>T) aparece combinada com a mutação Asp(G<u>A</u>T)→Val(G<u>T</u>T) da posição homoplásica 256 (Figura II.17 A). Por sua vez, nesta posição, há uma mutação adicional Asp256(GA<u>T</u>)→Asp(GA<u>C</u>) e no codão 255 vizinho há uma substituição Asp(<u>C</u>CT)→Thr(<u>A</u>CT) (Tabela II.4). A mutação Asp256(G<u>A</u>T)→Val(G<u>T</u>T) também está associada à mutação rara Pro391(CCC)→ His(C<u>A</u>C) na variante Ybarcelona (Figura II.17 A). Embora a mutação Asp256(G<u>A</u>T)→Val(G<u>T</u>T) não esteja associada a um dinucleotídeo CpG, existe nos codões 252 a 253 uma repetição do motivo CTT, que é um local consensual de clivagem da topoisomerase I e, segundo Cooper *et al.* (1995), ocorre com uma frequência superior à esperada nas 10bp flanqueantes de substituições nucleotídicas.

A delecção de um resíduo de fenilalanina nas posições 51 ou 52 e a mutação Gly148(GGG)→Arg(AGG) estão associadas no alelo Mnichinan (Figura II.17 B). A delecção ocorre numa região repetitiva (ver acima - secção 2.2.2.1) e foi observada em europeus e japoneses. A substituição ocorre num dinucleotídeo CpG e foi observada em japoneses, europeus e africanos. Por outro lado, o aminoácido arginina encontra-se na posição 148 das sequências parálogas do inibidor de proteases da proteína Z-dependente, do inibidor da proteína C e do pseudogene da α1-antitripsina, bem como nas sequências homólogas da α1-antitripsina de *Sus scrofa*, *Ovis aries* e *Equus caballus*. No primeiro nucleotídeo do codão 148, observa-se ainda outra mutação associada ao motivo CpG: Gly148(GGG)→Trp(TGG) (Tabela II.4).

A mutação Gly115(GGC)→Ser(AGC) e a substituição característica do alelo Z Glu342(GAG)→Lys(AAG) estão associadas no alelo Znewport e ambas ocorrem em dinucleotídeos CpG (Figura II.17 C). No codão vizinho da posição 342 há uma substituição Asp341(GAC)→Asn(AAC), também associada a um dinucleotídeo CpG (Tabela II.4). Apesar da sua potencial hipermutabilidade, a esmagadora maioria dos alelos Z deriva de uma única linhagem (ver artigo 2) e a sua distribuição na Europa deve-se ao fluxo génico entre populações.

A delecção homoplásica que origina os alelos Q0mattawa e Q0ourém ocorre numa região repetitiva (T)5 que engloba os codões 353 e 355 e foi encontrada em regiões relativamente isoladas de Portugal e do Canadá (Figura II.17 D e artigo 6).

A posição 362 está associada a 2 indels e duas substituições de aminoácidos diferentes (Figura II.17 E). As indels Pro362framStop376 e Pro362framStop373, que originam os alelos Q0saarbruecken, Q0clayton e Q0bolton, ocorrem na região repetitiva (C)7 que abrange os codões 360 a 362. As substituições Pro362(CCC)→Thr(ACC) e Pro362(CCC)→His(CAC) não estão associadas a qualquer motivo CpG ou outra sequência hipermutável definida *a priori*, nem são facilmente explicáveis pelos desalinhamentos produzidos pela sequência repetitiva (Cooper *et al.*, 1995). No entanto, a sua ocorrência simultânea no mesmo codão e a presença adicional da mutação Glu(GAG)→Lys(AAG) no codão vizinho 363 indicam que esta região tem uma taxa de substituição nucleotídica elevada. Esta conclusão é ainda reforçada pela observação do aminoácido treonina nas posição 362 das sequências parálogas da calistatina e da globulina transportadora de tiroxina.

No codão 369 ocorrem duas substituições de nucleotídeos diferentes: Pro369(CCC)→Leu(CTC) e Pro369(CCC)→Ser(TCC), ambas homoplásicas (Figura II.2.4 F). O aminoácido serina também é observado na posição 369 da sequência paráloga da globulina transportadora de corticosteróides. À semelhança do que acontece no codão 362, nenhuma destas mutações está associada ao dinucleótido CpG, embora nos codões 371 a 373 exista uma repetição do motivo CTT, tal como acontece na vizinhança do codão 256.

Em conjunto, estes resultados mostram que todas as regiões homoplásicas analisadas satisfazem simultaneamente pelo menos três critérios adicionais de hipermutabilidade, e sugerem que nos locais variáveis do gene da α1-antitripsina há pontos que têm uma taxa de mutação mais elevada (Tabela II.7). O facto de apenas duas das nove posições hipermutáveis não estarem associadas à deficiência da proteína, deve-se, provavelmente, ao enviezamento provocado pela selecção de mutações que chamem mais a atenção devido às suas consequências clínicas.

Tabela II.7 - Propriedades das nove posições hipermutáveis identificadas no gene da α1-antitripsina.

	Critérios de hipermutabilidade*						
Codões	1	2	3	4	5	6	7
51/52	+	+			+	+	+
101	+	+	+				+
115	+	+				+	+
148	+	+	+	+		+	+
256	+	+		+	+	+	+
342	+	+			+		+
353	+					+	+
362	+			+	+	+	+
369	+		+	+			+

* 1) homoplasia;
2) combinação no mesmo alelo de mutações que envolvam duas ou mais posições potencialmente hipermutáveis;
3) fixação do aminoácido derivado de uma mutação potencialmente recorrente numa sequência ortóloga ou paráloga;
4) ocorrência de duas ou mais mutações diferentes num mesmo codão ou nucleotídeo;
5) observação de mutações em codões;
6) observação de uma mutação potencialmente recorrente em populações com baixo fluxo genético;
7) localização de mutações em regiões com motivos hipermutáveis identificados *a priori*.

Por outro lado, esta situação indica que os locais hipermutáveis não foram selectivamente removidos das regiões funcionalmente relevantes e que, apesar das desvantagens associadas, há um potencial de diversificação funcional que se pode manifestar após a criação de genes ortólogos por duplicação.

2.2.2.4 *Implicações funcionais*

A análise da distribuição de substituições de aminoácidos na α1-antitripsina tem sido fundamental para a identificação das regiões mais relevantes para o funcionamento da proteína. Nos estudos exaustivos que realizaram, Huber e Carrell (1989), Stein e Carrell (1995) e Irving *et al.* (2000), demonstraram que as substituições patogénicas permitem confirmar a importância de zonas críticas previamente definidas pela análise da estrutura tridimensional da molécula e pelo grau de conservação observado em comparações conjuntas de sequências parálogas e ortólogas de várias proteínas da família SERPIN. Com o objectivo de complementar a informação fornecida por esses estudos, procurou-se fazer uma comparação das distribuições das mutações patogénicas e não patogénicas e relacioná-las com os padrões de conservação observados separadamente em sequências ortólogas e parálogas.

A análise das sequências ortólogas mostra que 95% (18/19) das substituições patogénicas e apenas 47,8% das substituições não patogénicas (11/23) localizam-se em resíduos que são conservados numa fracção igual ou superior a 75% das espécies comparadas (Figura II.18). Estas distribuições são significativamente diferentes (χ^2=10,88; 1gl; p<0.001) e indicam que há uma clara associação entre as consequências de uma mutação e o grau de conservação evolutiva dos resíduos em que ocorre, à semelhança do que se observa em muitas outras proteínas (Miller e Kumar, 2001).

Ao contrário das sequências ortólogas em que a percentagem total de aminoácidos conservados em pelo menos 75% das sequências analisadas é de 72%, nas sequências parálogas apenas 28,5 % dos codões têm um nível equivalente de conservação, devido ao relaxamento da pressão da selecção negativa após duplicação génica. Ainda assim, as distribuições das mutações patogénicas e não patogénicas são significativamente

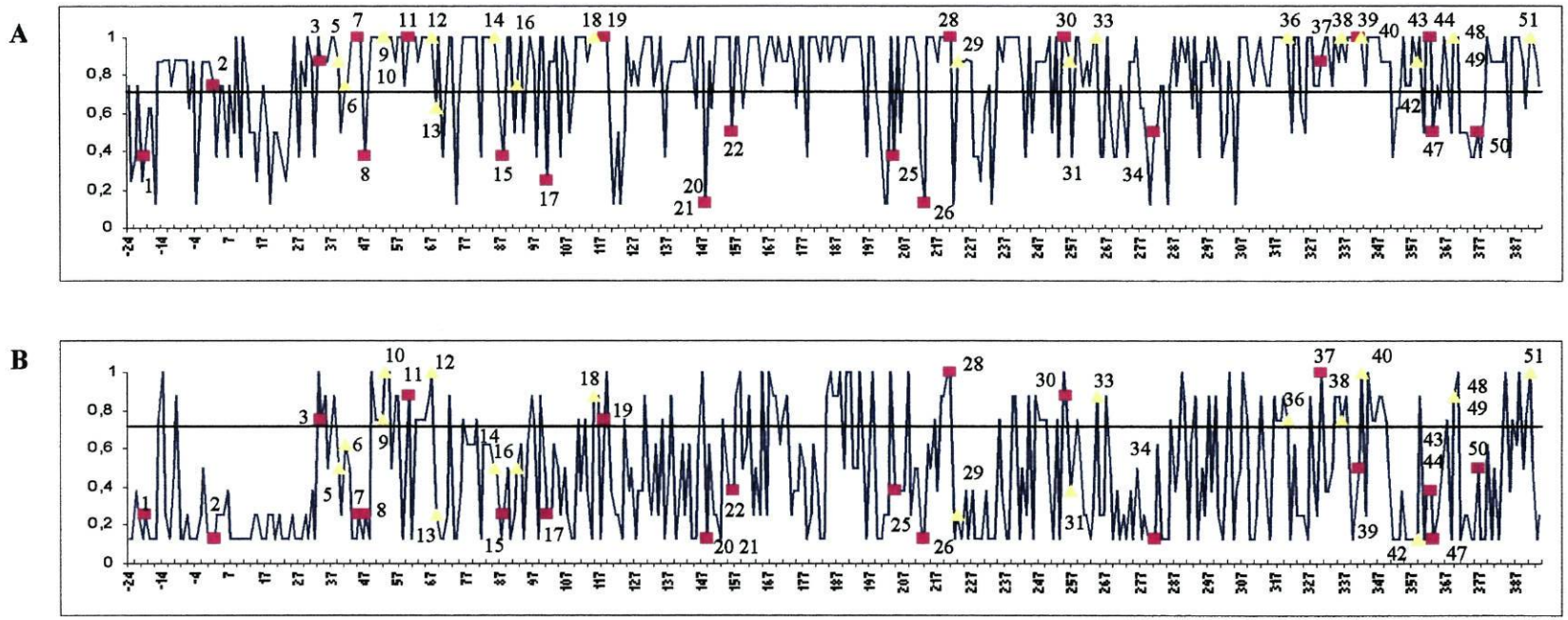


Figura II.18 - Espectros de conservação da sequência polipeptidica da α1-antitripsina humana obtidos através da comparação das sequências ortólogas (**A**) e parálogas (**B**) referidas na figura II.15. Nas abcissas apresentam-se os números correspondentes às posições de cada aminoácido. Nas ordenadas mostram-se as percentagens de sequências homólogas que contêm um dado aminoácido observado na α1-antitripsina humana. As mutações descritas na tabela II.4 estão representadas pelos símbolos ▲ e ■ correspondentes às alterações patogénicas e não patogénicas, respectivamente. Os números atribuídos às mutações correspondem aos da tabela II.4. A linha horizontal define o patamar de conservação de 75%.

diferentes (χ^2= 4,37; 1 gl; p<0,05), registando-se uma sobrerepresentação das mutações que conduzem ao défice da α1-antitripsina nas regiões mais conservadas das proteínas da família SERPIN. Estas regiões são provavelmente fundamentais para a estrutura da proteína desta família e é neles que ocorrem todas as substituições patogénicas em que a deficiência da mutação é provocada pela acumulação intracelular (Figura II.18; Tabela II.4).

Estes resultados indicam que, apesar de as mutações que provocam o défice da α1-antitripsina não estarem aparentemente associadas a fortes mortalidades pré-reprodutivas, a selecção negativa condicionou fortemente a divergência evolutiva desta proteína.

Artigo 3

Seixas, S., Trovoada, M. J., Santos, M. T., Rocha, J. (1999). A novel alpha-1-antitrypsin P362H variant found in a population sample from São Tomé e Príncipe (Gulf of Guinea, West Africa). *Hum Mutat* **13**:414.

HUMAN MUTATION Mutation and Polymorphism Report #50(1999) Online

Mutation and Polymorphism Report

Title : A novel alpha-1-antitrypsin P362H variant found in a population sample from São Tomé e Príncipe (Gulf of Guinea, West Africa).

Authors: Susana Seixas (1,2), Mª Jesus Trovoada (3), Mª Teresa Santos(2), Jorge Rocha(1,2)

Affiliations: 1 Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Portugal. 2 Faculdade de Ciências da Universidade do Porto, Portugal. 3 Departamento de Antropologia da Universidade de Coimbra, Portugal.

Species: Homo sapiens

Gene/Locus

Name: Alpha-1-antitrypsin

Symbol: PI

HUGO-approved when available

Genbank accession number: K02212

OMIM accession number: 107400

Locus specific database:
web address when available

Chromosomal location: 14q32.1
e.g., 12q24.1

Inheritance: codominant

Mutation / polymorphism name: transversion
To follow nomenclature guide.

Nucleotide change–Systematic name: g-1050 C->A
Sequential no. in genomic or cDNAsequence. *e.g., c1227c->T*

Amino acid change–Trivial name: P362H
Codon number and change. *e.g., R108W*

Mutation / polymorphism type: Missense
Missense, deletion, splice, etc.

Polymorphism frequency:
e.g., 40/60 C/T

Detection method: Sequencing of PCR amplified DNA from all coding exons after detection of variant protein by high resolution hybrid isoelectric focusing on a 4.3-4.8 pH gradient.
DGGE, etc.

Detection conditions

sequence of primers: PCR primers (Graham et al. 1990): A1 and A2 (exon 2); A3 and A4 (exon 3); A5 and A6 (exon 4); A7 and A8 (exon 5). Sequencing primers: S1 from Graham et al. 1990 and 5'-GGAGATTCCGGAGGCTCAGAT-3' (exon2), S3, S5 and S7 from Graham et al. 1990 for exons 3, 4 and 5, respectively.

PCR conditions: Subjected to the conditions described in Graham et al. 1990

electrophoresis: Sequencing products obtained using the ABI Prism Big Dye Dideoxy Terminator Cicle Sequencing Kit were separated and analysed in a 6% acrylamide gel in an automatic ABI377 DNA sequencer.

Diagnosis method developed:
ASO, etc.

Evidence for existence and effect (mutation) or lack of effect (polymorphism):

	Yes	No	Don't know
1. Mutation found on repeat PCR sample	☒	☐	☐
2. Mutation segregates or appears with trait	☐	☒	☐
3. Mutation affects conserved residue	☐	☒	☐
4. Expression analysis supports hypothesis	☐	☐	☒
5. Mutation not found in 50 normal subjects	☒	☐	☐
6. Polymorphism	☒	☐	☐

Ancillary data

1. **Haplotype association :**	PI STRP-P7,P9/ PCI-B4,B4 the sample is heterozygous for the novel mutation
2. **Population association :**	unexplored
3. **Geographic association :**	unexplored
4. **Clinical phenotype of proband :**	normal subject
5. **Homologous allele** (if recessive trait):	
6. **Frequency:**	1/368 p=0.003
7. **PIC:**	
8. **Ethnic background:**	184 unrelated individuals from S. Tomé (São Tomé e Príncipe archipelago in the Gulf of Guinea; West Africa)
9. **Other:**	
10. **Present in HGMD listing:** (http://www.cf.ac.uk/uwcm/mg/hgmd0.html)	Yes ☐ No ☒

Comments

Alpha-1-antytripsin (PI), the major inhibitor of neutrophil elastase in the lower respiratory tract, is a highly polymorphic glycoprotein synthesized in the liver with several variants that can be detected by protein isoelectric focusing. While most variants are associated with normal circulating concentrations, several rare alleles have been described in which serum PI values are reduced or even non-detectable due to defective secretion, intracellular degradation or lack of synthesis (reviewed in Cox, 1995). Early- onset pulmonary emphysema and neonatal cholestasis are the two most common clinical manifestations of PI deficiency and are mainly associated with PI*Z (E342K), the most common PI deficient allele. Knowledge of the molecular basis of rare variants, both normal and pathogenic, can be useful for the identification of the residues most important for protein function. The rare novel variant we now describe was found by means of hybrid isoelectric focusing in heterozygous state with the common normal M1Ala 213 allele in a healthy individual from S. Tomé island (São Tomé e Príncipe archipelago, Gulf of Guinea; West Africa). The variant protein has an isoelectric point only slightly more acidic than that of the normal rare PI*Pdonauwoerth (D341N; Faber et al. 1995) and its P362H substitution has occurred on the common normal M3 base allele. The relative band intensity of the variant P362H protein upon densitometric analysis is compatible with normal plasma PI level. To our knowledge this represents the fourth mutation involving codon 362 which, besides the substitution now described, include the normal PI*Loffenbach variant (P362T; Faber et al. 1995) and the null alleles PI*Q0saarbruecken (P362CCCC insertion; Faber et al. 1995) and PIQ0bolton (P362CC deletion; Frazier et al.1989).

Acknowledgments

Susana Seixas is supported by grant BD/13885/97 from Praxis XXI.

References

Cox DW (1995) Alpha-1-antitrypsin deficiency: in Scriver CR, Beaudet AL, Sly Ws, Valle D (eds): The metabolic and molecular bases of inherited disease. New York, McGraw-Hill, vol3, pp4125-4158.

Graham A, Kalsheker NA, Bamford FJ, Newton CR, Markham AF (1990) Molecular characterisation of two alpha-1-antitrypsin deficiency variants: proteinase inhibitor (Pi) Null Newport (Gly115-Ser) and (Pi) Z Wrexham (Ser-19-Leu). Hum Genet 85: 537-540.

Faber J-P, Poller W, Weidinger S, Kirchgesser M, Schaab R, Bidlingmeier F, Olek K (1994) Identification and DNA sequence analysis of 15 new alpha-1-antitrypsin variants, including two PI*Q0 alleles and one deficient PI*M allele. Am J Hum Genet 55: 1113-1121.

Fraizier GC, Siewertsen M, Harrold TR, Cox DW (1989) Deletion frameshift mutation in the alpha-1-antitrypsin null allele, PI*Q0bolton. Hum Genet 83:377-382.

HUMAN MUTATION Mutation and Polymorphism Report #50(1999) Online

Keywords
Alpha-1-antitrypsin, rare P362H normal substitution, West Africa.

Corresponding Author Information (address, phone, fax, e-mail)
Jorge Rocha
Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)
Rua Dr. Roberto Frias s/n, 4200, Porto, Portugal
Tel.: +351+2+5570700; Fax: +351+2+5570799; e-mail: jrocha@ipatimup.pt

Artigo 4

Seixas, S., Garcia, O., Amorim, A., Rocha, J. (2000). A novel alpha-1-antitrypsin R281del variant found in a population sample from the Basque country. *Hum Mutat* **15**:121-122.

HUMAN MUTATION Mutation and Polymorphism Report #90 (1999) Online

Mutation and Polymorphism Report

Authors: Susana Seixas[1,2], Oscar Garcia[3], António Amorim[1,2], and Jorge Rocha[1,2]

Affiliations: [1]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Portugal; [2]Faculdade de Ciências da Universidade do Porto, Portugal; [3]Area de Laboratorio Ertzaintza, Bilbao, Spain

Corresponding Author Address and E-mail: Jorge Rocha

Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)

Rua Dr. Roberto Frias s/n 4200, Porto, Portugal

E-mail: jrocha@ipatimup.pt

Title : A novel alpha-1-antitrypsin R281del variant found in a population sample from the Basque country

Keywords: Alpha-1-antitrypsin, R281del mutation, Basque Country.

Species: Homo sapiens

Change is: Polymorphism

Gene/Locus

Name:	Alpha-1-antitrypsin
Symbol:	PI
Genbank accession number:	K02212
OMIM accession number:	107400
Locus specific database:	
Chromosomal location:	14q32.1
Inheritance:	Codominant

Mutation / polymorphism name

Nucleotide change–Systematic name:	g.7727-7729delAGA or g.7729-7731delAAG
Amino acid change–Trivial name:	R281del
Mutation / polymorphism type:	Deletion
Polymorphism frequency:	
Detection method:	Sequencing of PCR amplified DNA from all coding exons after detection of variant protein by high resolution hybrid isoelectric focusing on a 4.3-4.8 pH gradient. Heteroduplexes are formed after separation of exon 3 PCR products in 6% non-denaturing polyacrylamide gels.
Detection conditions:	**Sequence of primers:**

PCR primers as in (Graham et al. 1990): A1 and A2 (exon 2); A3 and A4 (exon 3); A5 and A6 (exon 4); A7 and A8 (exon 5).

Sequencing primers: S1 from Graham et al. 1990 and 5'-GGAGATTCCGGAGGCTCAGAT-3' (exon2), S3, S5 and S7 from Graham et al. 1990 for exons 3, 4 and 5, respectively.

PCR conditions:

As described in Graham et al. 1990

Sequencing:

Sequencing products obtained using the ABI Prism Big Dye Dideoxy Terminator Cycle Sequencing Kit were separated and analyzed in a 6% acrylamide gel in an automatic ABI377 DNA sequencer.

Diagnosis method developed:

Evidence for existence and effect of mutation:

	Yes	No	Don't know
1. **Base change found on repeat PCR sample**	X		
2. **Base change segregates or appears with trait**		X	
3. **Base change affects conserved residue**		X	
4. **Expression analysis supports hypothesis for causation**			X
5. **Normals tested (50 required)**	X		

Ancillary data

1. **Haplotype association :**	PI STRP-P4, P8 (153 bp, 161 bp)/ CBG-2, 2 (86bp, 86 bp). The sample is heterozygous for the novel mutation.
2. **Ethnic background/Population association :**	98 unrelated individuals from the Basque Country autochthonous population.
3. **Geographic association :**	Unexplored
4. **Frequency** (of mutation) **in population:**	1/196 p=0.005
5. **Clinical phenotype of proband :**	Normal subject.
6. **Homologous allele** (if recessive trait):	
7. **PIC:** (if microsatellite)	
8. **Other:**	
9. **Present in HGMD listing:** (http://www.cf.ac.uk/uwcm/mg/hgmd0.html)	Yes: No: X

Comments

Alpha-1-antytripsin (PI), the major inhibitor of neutrophil elastase in the lower respiratory tract, is a highly polymorphic glycoprotein synthesized in the liver. In addition to a majority of normal alleles, several rare variants have been described in which serum PI values are reduced or even non-detectable. Early- onset pulmonary emphysema and neonatal cholestasis are the most common manifestations of protein deficiency caused by these alleles (reviewed in Cox, 1995). The novel variant we now describe was found by means of hybrid isoelectric focusing in heterozygous state with the common normal M2 allele in a healthy individual from the Basque Country autochthonous population. The R281del occurs on the common normal M1Val 213 base allele and could have resulted from the in- frame deletion of AGA or AAG in the last five nucleotides of exon 3. Consistently with the loss of a positively charged amino acid, the variant gene product has an isoelectric point only slightly more acidic than that of the rare mildly deficient PI*I variant (R39C; Graham et al., 1989). The relative band intensity of the R281del variant protein upon densitometric evaluation is compatible with normal plasma PI levels. To our knowledge, this represents the second documented mutation involving the loss of just one amino acid in alpha-1-antitrypsin, together with the deletion of Phe in codon 51 or 52 shared by Mmalton, Mnichinan and Mpalermo severe deficiency alleles. However, the finding of normal gene product levels upon isoelectric focusing indicates that, contrary to these alleles, the deletion now observed is neither associated with defective protein secretion nor to intracellular degradation due to decreased stability of the molecule.

Acknowledgments

Susana Seixas is supported by grant BD/13885/97 from Praxis XXI.

References

Cox DW (1995) Alpha-1-antitrypsin deficiency: in Scriver CR, Beaudet AL, Sly Ws, Valle D (eds): The metabolic and molecular bases of inherited disease. New York, McGraw-Hill, vol3, pp4125-4158.

Graham A, Kalsheker NA, Bamforth FJ, Newton CR, Markham AF (1990) Molecular characterisation of two alpha-1-antitrypsin deficiency variants: proteinase inhibitor (Pi) Null Newport (Gly115-Ser) and (Pi) Z Wrexham (Ser-19-Leu). Hum Genet 85: 537-540.

Graham A, Kalsheker NA, Newton CR, Bamforth FJ, Powell SJ, Markham AF (1989) Molecular characterisation of three alpha-1-antitrypsin deficiency variants: proteinase inhibitor (Pi) nullcardiff (Asp256→Val); Pi Mmalton (Phe51→ del) and Pi I (Arg39→Cys). Hum Genet 84: 55-58.

Artigo 5

Seixas, S., Lopes, A. I., Rocha, J., Silva, L., Salgueiro, C., Salazar-de-Sousa, J., Batista, A. (2001). Association between the defective Pro369Ser mutation and in vivo intrahepatic α1-antitrypsin accumulation. *J Med Genet* **38**:472-474.

Association between the defective Pro369Ser mutation and in vivo intrahepatic α1-antitrypsin accumulation

Susana Seixas, Ana Isabel Lopes, Jorge Rocha, Lídia Silva, Carlos Salgueiro, Jaime Salazar-de-Sousa, Amélia Batista

EDITOR—α1-antitrypsin (PI), the major inhibitor of neutrophil elastase in the lower respiratory tract, is a highly polymorphic glycoprotein synthesised in the liver that has several rare gene products in which serum protein levels are reduced or even undetectable.[1] Early onset pulmonary emphysema, resulting from unopposed elastase activity, and neonatal cholestasis probably resulting from the retention of the defective protein in the liver,[2] are the two most common clinical manifestations of PI deficiency and are mainly associated with PI*Z, the most common deficient allele. In addition, other rare alleles occasionally associated with liver injury have been shown to share with PI*Z an increased tendency for intracellular accumulation. Recently, a complete intracellular transport block has been reported for a newly identified[3 4] defective Pro369Ser allele (Mwürzburg) by in vitro expression studies in human cell cultures. Adenovirus mediated transfer of the mutant gene into the mouse reproduced the consequences of this block and no traceable amounts of the variant protein could be detected in the plasma after in vivo recombinant expression.[3] However, no detectable intrahepatocytic PI inclusions were found in the mice expressing the Mwürzburg mutant[3] and no liver biopsy material has yet been presented from patients with this defective allele.

Case report

We report a carrier of the Mwürzburg allele with evidence for in vivo intrahepatic accumulation of PI. The patient is a white Portuguese boy who presented at the age of 1.5 months with cholestasis associated with a recent cytomegalovirus (CMV) infection. A percutaneous liver biopsy performed at the age of 2.5 months showed significant portal fibrosis with porto-portal bridging, giant cell transformation, moderate cholestasis, and an intense portal-acinar inflammatory infiltrate. Periodic acid-Schiff staining after diastase treatment (PAS-D) additionally showed the presence of positive, diastase resistant, intracellular inclusions. Immunoperoxidase staining specific for PI was positive (fig 1). Serum PI concentration, determined by automated nephelometry (Behring), was found to be 92% of normal on admission and dropped to 45% of normal at the age of 24 months, after CMV serology (IgM) and antigens became negative and following progressive decline of transaminase levels to their normal upper limit. The PI

J Med Genet 2001;**38**:472–474

Instituto de Patologia e Imunologia Molecular, Universidade do Porto (IPATIMUP), R Dr Roberto Frias s/n, 4200 Porto, Portugal
S Seixas
J Rocha

Faculdade de Ciências, Universidade do Porto, Portugal
S Seixas
J Rocha

Unidade de Gastrenterologia Pediátrica, Hospital de Santa Maria, Lisboa, Portugal
I A Lopes
J Salazar de Sousa

Serviço de Pediatria, Hospital do Barreiro, Barreiro, Portugal
L Silva
C Salgueiro

Serviço de Anatomia Patológica, Hospital de Santa Maria, Lisboa, Portugal
A Batista

Correspondence to:
Dr Rocha,
jrocha@ipatimup.pt

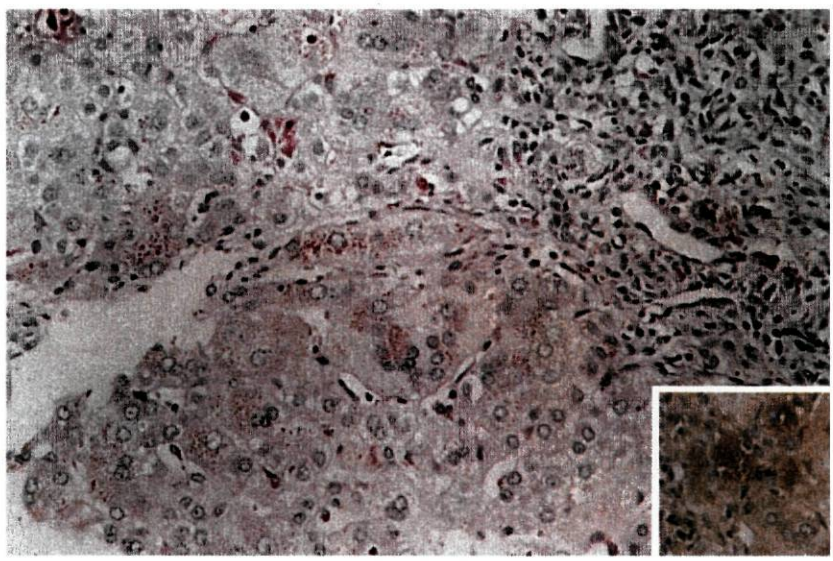


Figure 1 Tissue section from the patient's percutaneous liver biopsy. PAS positive diastase resistant inclusions were found in the cytoplasm of several hepatocytes (PAS-D). PI is identified by immunoperoxidase in inset (immunoperoxidase staining).

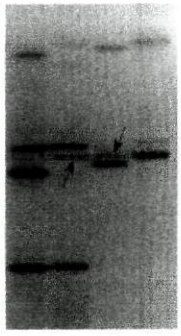


Figure 2 PI patterns after polyacrylamide gel hybrid isoelectric focusing (pH range of 4.3 to 4.8) and Coomassie blue staining.[3] From left (1) M2/S = patient's mother, (2) Mwürzburg/S = patient, (3) Mwürzburg/M3 = patient's father, (4) M1 = control. The major band corresponding to the Mwürzburg gene product is indicated with an arrow.

concentration in the patient's father was 49% of normal.

Isoelectric focusing analysis of PI types showed that the patient was heterozygous for the S allele and for a deficient variant gene product, inherited from his father, which has an isoelectric point identical to M1 and a band with decreased intensity (fig 2). These patterns were confirmed by print immunofixation (not shown). DNA sequencing of all PI coding exons (II-V), performed as previously described,[6 7] has shown that the patient and his father shared a C to T transition leading to a 369Pro (CCC)→Ser (TCC) substitution in the common M1Val213 allele, as in the variant Mwürzburg.[3 4] The presence of the Mwürzburg allele was also confirmed by partial PCR amplification of exon V with a mismatched primer that generates an 118 bp fragment: 5'-CCCGAGGTCAAGTTCAA CAgA-3' (bases 10049-10069, mismatched base in lower case); 5'- GAGGAGCGAGAGG CAGTTATT-3' (bases 10166-10146). Thirty five cycles of PCR were performed for one minute at 94°C, one minute at 58°C, and one minute at 72°C. The mismatched primer artificially introduces a *Nde*II restriction site in the mutated Pro369Ser allele during PCR amplification (fig 3, lanes 1 and 2). In addition, the primer also generates a further *Hin*fI restriction site in the presence of the Pro369Leu mutation (fig 3, lanes 3 and 4), which characterises the severely deficient variant Mheerlen.[8]

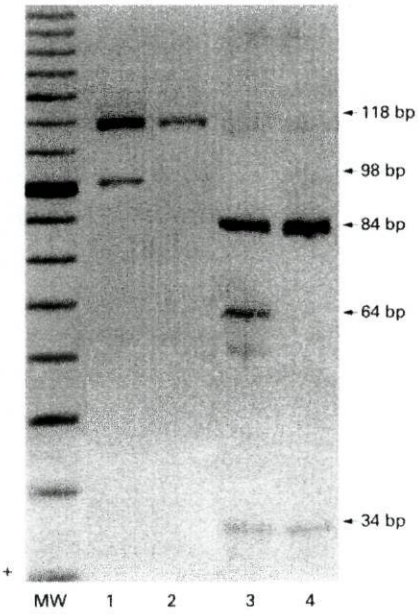


Figure 3 Detection of the Pro369Ser and Pro369Leu mutations after PCR amplification with a mismatched primer. (Lanes 1 and 2) NdeII digested fragments from the Mwürzburg/S patient (1) and from a normal M1 control (2). The Pro369Ser mutated sequence generates one 98 bp fragment and one 20 bp fragment (not visible). (Lanes 3 and 4) HinfI digested fragments from a M2/Mheerlen subject (3) and from a normal M1 control (4). The normal sequence produces one 84 bp fragment and one 34 bp fragment. The Pro369Leu mutated sequence generates an additional restriction site and 64 bp, 34 bp, and 20 bp (not visible) fragments are produced. MW= molecular weight marker. DNA fragments were visualised by silver staining after non-denaturing electrophoresis on 9% polyacrylamide gels.

Discussion

To our knowledge the present case represents the first reported association between the defective α1-antitrypsin Pro369Ser mutation and in vivo intrahepatic protein accumulation. Although the S variant has been found to show increased intracellular retention and the ability to form heteropolymers with Z,[9-11] this increase is only marginal and no evident inclusions of S α1-antitrypsin have been found in most pathology samples observed so far. Therefore, the observation of PI liver inclusions in the Mwürzburg/S patient is most likely to be predominantly caused by the Mwürzburg variant and provides further in vivo evidence that the severe deficiency resulting from the Pro369Ser mutation is caused by protein accumulation, as in the case of the Z allele.

Since the patient's PI type would be expected to be similar to SZ, which is not associated with increased risk of liver disease in infancy, his liver injury is probably related more to the CMV infection than to the Mwürzburg variant. However, the similar behaviour of Mwürzburg and Z both in vitro and in vivo indicates that it may lead to liver disease in Mwürzburg homozygotes or in Mwürzburg/Z heterozygous combinations.

Contrary to previous observations,[3] the present detection of a faint PI band with the same isoelectric point of M1, both in the patient and his father, indicates that Mwürzburg can still be secreted in limited amounts into the plasma. However, since the variant will remain unidentified in combination with the common normal M1 allele, the PCR introduction of a *Nde*II restriction site is a simple alternative tool to detect the Pro369Ser mutation, especially in cases where no unusual isoelectric focusing patterns are associated with decreased PI serum levels or with intrahepatic protein accumulation.

We thank Ms Lúcia Ramires and Ms Piedade Mendonça for technical assistance. This work was supported in part by Conselho de Prevenção do Tabagismo. Susana Seixas is supported by grant BD/13885/97 from Praxis XXI.

1 Cox DW. α1-antitrypsin deficiency. In: Scriver CH, Beaudet AL, Sly WS, Valle D, eds. *The metabolic and molecular bases of inherited disease*. Vol 3. New York: McGraw-Hill, 1995:4125-58.
2 Permultter DH. The cellular basis for liver injury in α1-antitrypsin deficiency. *Hepatology* 1991;**13**:172-85.
3 Poller W, Merklein F, Schneider-Rasp S, Haack A, Fechner H, Wang H, Anagostopoulos I, Weidinger S. Molecular characterisation of the defective α1-antitrypsin alleles PI Mwürzburg (Pro369Ser), Mheerlen (Pro369Leu), and Q0lisbon (Thr68Ile). *Eur J Hum Genet* 1999;7:321-31.
4 Rocha J, Seixas S, Lopes AI, Silva L, Salgueiro C, Salazar-de-Sousa J, Batista A. Human gene mutation report. *Hum Genet* 1999;**104**:114.
5 Rocha J, Pinto D, Santos MT, Amorim A, Amil-Dias J, Cardoso-Rodrigues F, Aguiar A. Analysis of the allelic diversity of a (CA)n repeat polymorphism among α1-antitrypsin gene products from northern Portugal. *Hum Genet* 1997;**99**:194-8.
6 Graham A, Kalsheker NA, Bamforth FJ, Newton CR, Markham AF. Molecular characterization of two alpha-1-antitrypsin deficiency variants: proteinase inhibitor (Pi) Nullnewport (Gly115→Ser) and (Pi) Zwrexham (Ser-19→Leu). *Hum Genet* 1990;**85**:537-40.
7 Seixas S, Garcia O, Amorim A, Rocha J. A novel alpha-1-antitrypsin R281del variant found in a population sample from the Basque Country. *Hum Mutat* 2000;**15**: 121-2.

8 Hofker MH, Nukiwa T, van Paassen, Nelen M, Kramps JA, Klasen EC, Frants RR, Crystal RG. A Pro→Leu substitution in codon 369 of the alpha-1-antitrypsin deficiency variant PI Mheerlen. *Hum Genet* 1989;**81**:264-8.
9 Teckman JH, Permutter DH. The endoplasmatic reticulum degradation pathway for mutant secretory proteins α1-antitrypsin Z and S is distinct from that for an unassembled membrane protein. *J Biol Chem* 1996;**271**:13215-20.
10 Elliot PR, Stein PE, Bilton D, Carrell RW, Lomas DA. Structural explanation for the deficiency of S α1-antitrypsin. *Nat Struct Biol* 1996;**3**:910-11.
11 Mahadeva R, Chang WSW, Dafforn TR, Oakley DJ, Foreman RC, Calvin J, Wight DGD, Lomas DA. Heteropolymerization of S, I an Z α1-antitrypsin and liver cirrhosis. *J Clin Invest* 1999;**103**:999-1006.

Artigo 6

Seixas, S., Mendonça, C., Costa, F., Rocha, J. (2002a). α1-Antitrypsin null alleles: evidence for the recurrence of the L353fsX376 mutation and a novel G-A transition in position +1 of intron IC affecting normal mRNA splicing. *Clin Genet* **62:**175-180.

α1-Antitrypsin null alleles: evidence for the recurrence of the L353fsX376 mutation and a novel G→A transition in position +1 of intron IC affecting normal mRNA splicing

Seixas S, Mendonça C, Costa F, Rocha J. α1-Antitrypsin null alleles: evidence for the recurrence of the L353fsX376 mutation and a novel G→A transition in position +1 of intron IC affecting normal mRNA splicing.
Clin Genet 2002:62:00-00.

α1-Antitrypsin (PI) deficiency is a common autosomal recessive disorder associated with emphysema and liver disease, which may result from a wide spectrum of mutations causing a reduction of serum levels (deficient alleles) or a total lack of circulating protein (null alleles). We report two different alleles associated with the absence of isoelectric focusing banding patterns in Portuguese patients with emphysema. The first allele, $Q0_{ourém}$, results from the recurrence of the defining mutation of the $Q0_{mattawa}$ variant (L353fsX376) on a M3 normal background. The second allele, $Q0_{porto}$, has a novel G→A mutation at position +1 of the intron IC (IVS1C+1G→A), which restricts mononuclear phagocyte transcripts to mRNA species resulting from the direct splice of exon IA to exon II. The absence of this normal splice alternative in the liver, where PI is primarily synthesized, provides a basis for the pathogenic effects of this mutation.

S Seixas[a,b], C Mendonça[c], F Costa[d] and J Rocha[a,b]

[a]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal; [b]Faculdade de Ciências, Universidade do Porto, Portugal; [c]Serviço de Pneumologia, Hospital dos Covões, Coimbra, Portugal; [d]Serviço de Pneumologia, Hospital de São João de Deus, VN Famalicão, Portugal

Key words: α1-antitrypsin null alleles, recurrent mutation, novel RNA splicing mutation

Corresponding author: Jorge Rocha, Instituto de Patologia e Imunologia Molecular, Universidade do Porto (IPATIMUP),R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
Tel.: +351+22-5570700;
Fax: +351+22-5570799
e-mail:jrocha@ipatimup.pt

Received 9 January 2002, revised and accepted for publication 11 April 2002

Human α1-antitrypsin (PI) is a highly polymorphic glycoprotein that functions as the major inhibitor of neutrophil elastase in the lower respiratory tract. The protein is coded for by a 12.2 kb gene, located on chromosome 14q32.1, which is organized into three non-coding exons (IA, IB, IC), four coding exons (II-V) and six introns (1-3). The liver is the major site of PI production, but lower levels of expression are also found in diverse extrahepatic cell types, including neutrophils, blood monocytes and tissue macrophages (3-5). While in hepatocytes transcription starts within exon IC, mononuclear phagocytes use macrophage-specific start sites located in exons IA and IB and generate a variety of alternatively spliced mRNA species (2,6).

PI deficiency is one of the most common genetic disorders among populations of European descent (7) and may result from a wide spectrum of mutations, which lead either to a reduction of serum protein levels (deficient alleles) or to a total lack of circulating PI (null alleles) (3, 8, 9). Early-onset pulmonary emphysema, due to unopposed elastase activity, and diverse liver diseases in children and adults resulting from the retention of the defective protein in hepatocytes are the most common clinical manifestations of PI deficiency and are mainly associated with PI*Z, the predominant deficient allele (3, 8, 9). While deficient alleles result predominantly from missense mutations, null alleles can be caused by a variety of different mechanisms, including gene deletion or the formation of premature stop codons leading to the synthesis of structurally unstable truncated proteins or to mRNA degradation (3, 8-10). No known promoter mutations have been described so far and only two null variants were found to be associated with abnormal mRNA splicing ($Q0_{bonny\ blue}$ and $Q0_{west}$) (8, 10, 11).

In this report, we extend the sequence data on PI deficiency by characterizing two different null alleles associated with emphysema in two Portuguese families. The first allele was found to have the same insertional mutation of the $Q0_{mattawa}$ variant (12) on a M3 instead of a M1Val213 normal background. The second allele results from a novel mutation affecting normal mRNA splicing caused by a G→A transition at position +1 of intron IC.

Materials and methods

Identification of the α1-antitrypsin null alleles

The first null allele, named $Q0_{ourém}$ was identified in two siblings (index case 1 and index case 2) born in Ourém, a small village in central Portugal. Index case 1, a 46-year-old male, has had dyspnea for 5 years and a 12 pack-year cigarette smoking history. Index case 2, a 44 year-old non-smoker female, had had dyspnea for 10 years, asthma and an obstetric history of four spontaneous abortions after an initial successful gestation. Both patients had panlobular emphysema, documented by chest computed tomography (CT) scan, absence of liver disease and no detectable serum PI levels, as determined by automated nephelometry (Behring). Serum isoelectric focusing analysis of PI types followed by print immunofixation, performed as previously described (13), failed to reveal any PI bands in both patients, indicating homozygosity for $Q0_{ourém}$. Extension of PI typing to other family members was compatible with the segregation of this null allele and showed that both index cases had inherited $Q0_{ourém}$ from a $M3Q0_{ourém}$ father and a $M1Q0_{ourém}$ mother. Although no consanguinity was acknowledged by the parents of the index cases, they were both born in Ourém and are likely to be related.

The second null allele, named $Q0_{porto}$, was identified in a 52 year-old male patient (index case 3), born in the urban area of Porto district (northern Portugal), with panlobular emphysema documented by chest CT scan since the age of 44 and no evidence of liver disease. He had a 42 pack-year cigarette smoking history until the age of 45, when he stopped smoking. Serum PI concentration was found to be 10% of normal. PI isoelectric focusing has revealed a Z phenotype, but confirmatory DNA analysis, using a PCR-RFLP assay (14), has shown that the patient was heterozygous for the Z specific mutation, suggesting he was carrying an allele with no detectable isoelectric focusing banding patterns. Two offspring from the patient were both typed as M2Z.

RNA and DNA extraction and cDNA synthesis

RNA and DNA were isolated from whole peripheral blood. RNA extraction, was done by using the Tripure Isolation Reagent (Boehringer Mannheim) according to the instructions provided by the manufacturer. DNA was prepared with a salting-out extraction protocol (15). First strand cDNA synthesis was carried out by reverse transcription-polymerase chain reaction (RT-PCR) using the Superscript II RT-PCR system (Life Technologies, Gibco, BRL) and the manufacturers recommended conditions.

DNA sequencing

Direct sequencing of genomic DNA was performed after PCR amplification of all coding exons (II-V) and exon-intron junctions as previously described (16, 17). Additional sequencing of exon IC, responsible for hepatocyte transcription, was done by using the following primers anchored in exons IB and intron IC: 5'-ATCAGGCATTTTGG-GGTGACT-3'; 5'-ACTGGGAAACAGGACACA-AC-3'.

To further elucidate the basis for the PI null allele in index case 3, cDNA was used as a substrate for producing several PCR amplified fragments that were subsequently cloned and sequenced (Fig. 1). The various primers used in these amplifications are as follows: IA, 5'-TCCTGTGCCTGCCAGA-AGAG-3'; IB, 5'-ATCAGGCATTTTGGGGTG-ACT-3'; IC, 5'-CTCTGGATCCACTGCTTAAAT-3'; ICintR, 5'-CGCAGTGAAAGGCATACTTA-3'; IIR, 5'-CTGATCATGGTGGGAT-GTAT-3'; II, 5'-GGAGATTCCGGAGGCTCAGAT-3'; IIIR, 5'-GATGATATCGTGGGTGAGAACATTT-3'. The reaction conditions of these experiments were: 94°C (1 min); 52-60°C (1 min); 72°C (2 min) for 30

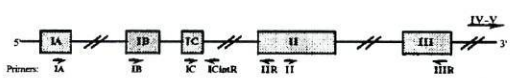


Fig. 1. Schematic representation of the 5'-portion of the PI gene. The exons are represented as boxes (IA-III) and the introns as horizontal lines. Approximate positions of the different primers used to amplify RT-PCR products from index case 3 are indicated below the scheme by arrows.

cycles. Amplifications using primer pairs IA/IIIR (Fig. 1) were performed with the Expand Long Template PCR System (Roche).

PCR products were cloned into a pCR4 vector using the TOPO TA Cloning Kit for sequencing (Invitrogen) and the products of ligation were used to transform One Shot TOP 10 chemically competent *E. coli* cells.

All sequencing products were obtained by using the ABI Prism Big Dye Dydeoxy Terminator Cycle Sequencing Kit and were analyzed in an automatic ABI 377 DNA sequencer.

Results

$Q0_{ourém}$

DNA sequencing of exons IC-V and all exon-intron junctions in index cases 1 and 2 has shown that $Q0_{ourém}$ resulted from the insertion of a single T into a stretch of 5 thymidines in codons 352-353 (exon V), occurring in the background of a normal M3 allele (Val213-Asp376). This insertion results in a frameshift that produces an altered reading frame starting in codon 353 and generates a premature stop codon (TGA) at position 376. Direct sequencing of exon V in the remaining family members has confirmed the segregation of the mutant allele.

$Q0_{porto}$

DNA sequencing in index case 3 has shown that $Q0_{porto}$ differed from the normal M1Val213 allele by a G→A transition at position +1 of the splice donor site of intron IC (IVS1C+1G→A). This position is part of the invariant GT dinucleotide found in splice donor sites at the 5' end of introns.

In an attempt to characterize the effects of the G→A mutation on the PI primary transcripts, the mRNA extracted from whole peripheral blood of the Z $Q0_{porto}$ index case 3 was reverse transcribed and the resulting cDNA used as substrate for PCR reactions with various combinations of seven different primers (Fig. 1).

Using primer pairs II/IIIR, IC/IIIR, IB/IIIR and IA/IIIR, four specific RT-PCR products were obtained, all including codon 213 in exon III. Since the Z Glu342Lys defining mutation in exon V has occurred in a M1Ala213 background (3), while $Q0_{porto}$ was found to be derived from a M1Val213 allele, codon 213 could be used to discriminate between the mRNA species from the two variants without the need of further extension of PCR reactions to exon V. To analyze the Ala213/Val213variation, all four types of PCR products were tested for the presence of a BstEII

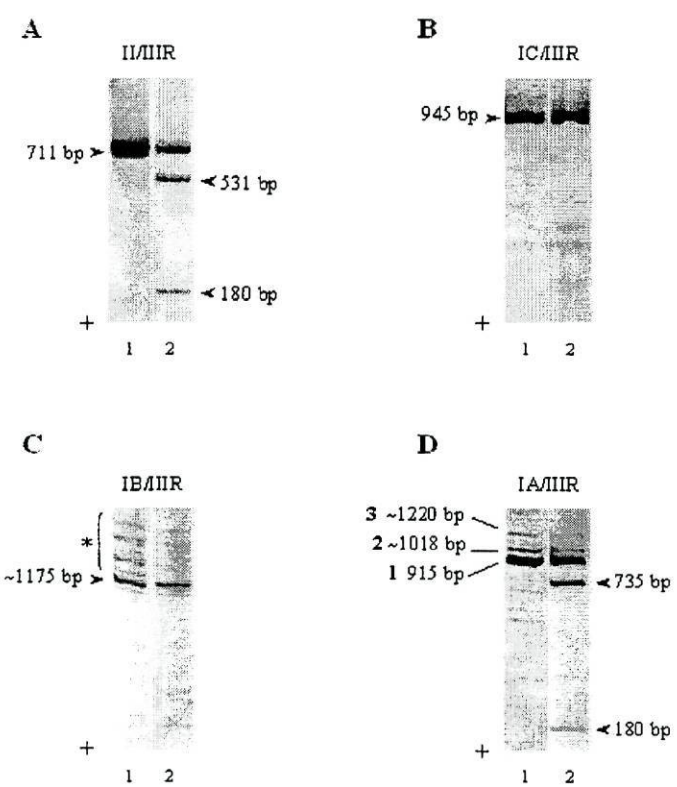


Fig. 2. Detection of the Ala213/Val213 variation in RT -PCR products from index case 3 (ZQ0porto) amplified with primer pairs II/IIIR (A), IC/IIIR (B), IB/IIIR (C) and IA/IIIR (D). Non-digested (1) and BstEII digested (2) fragments are shown for each primer combination. Val213 is associated with Q0porto and generates a 180 bp band and an additional fragment whose size depends on the length of the amplification product (A and D). Bands marked with a * in (C) are likely to be heteroduplexes resulting from sequence heterogeneity due to alternative splice donor/acceptor site utilization at the exon IB and or IC boundaries. Bands 1, 2 and 3 in (D) correspond to mRNA species with exons IA-II-III, IA-IC-II-III and IA-IB-IC-II-III, respectively. DNA fragments were visualized by silver staining after non-denaturing electrophoresis in 9% polyacrilamide gels.

Table 1. Summary results from the analysis of RT-PCR products in index case 3

Primer pair[a]	BstEII type[b]	Exon composition[c]
II/IIIR	+/-	ND[d]
IC/IIIR	-	ND[d]
IB/IIIR	-	IB-IC-II-III(Ala213)[e]
IA/IIIR	+/-	IA-IB-IC-II-III(Ala213)[e] IA -IC-II-III(Ala213)[e] IA -II-III(Ala213)[e] IA -II-III(Val213)[e]

[a] Approximate positions of primer pairs are given in Fig. 1
[b] Determined by BstEII digestion of the corresponding amplification products; + = Val 213; - = Ala213
[c] As determined after cloning and sequencing of the amplification products
[d] ND: not determined
[e] The presence of Ala or Val at codon 213 in exon III is additionally indicated

restriction site in exon III, which is specifically created by the Val213 codon (Fig. 2). This evaluation has shown that $Q0_{porto}$ was associated with mRNA production but, contrary to the Z allele, lacked mRNA species containing exons IB and IC (Fig. 2; Table 1).

To further elucidate the exon composition of the different mRNA species, the amplification products from primer pairs IB/IIIR and IA/IIIR were cloned and sequenced. Sequence analysis of 24 patient clones confirmed that PCR products obtained with primer pair IB/IIIR were exclusively associated with the Z allele and had the expected exon structure of mononuclear phagocyte transcripts containing exon IB, which are invariably associated with exon IC (2, 6, 18) (Table 1). Moreover, sequence information has also revealed an alternative splice donor/acceptor site utilization at the exon IB and or IC boundaries, precisely as recently described in RT-PCR analyses of RNA products transcribed from the human macrophage-specific promoter (18) (results not shown).

When 23 clones containing the amplification products from primer pair IA/IIIR were sequenced, three different mRNAs were found to be associated with Ala213, and therefore to the Z allele: IA-IB-IC-II-III (2 clones), IA-IC-II-III (4 clones) and IA-II-III (7 clones) (Table 1). Equivalent clones from the patient's M2Z sons yielded the same mRNA populations (results not shown). The isoforms IA-IC-II-III and IA-IB-IC-II-III have been previously recognized as alternatively spliced species from mononuclear phagocytes (2, 6). The IA-II-III results from the direct splice of exon IA to exon II and was only recently reported as a splice variant of human PI after RT-PCR analysis of macrophage-promoter transcripts (18). Interestingly, this was the only RNA species that was found to be associated with Val213 in the remaining 10 patient clones (Table 1). This finding indicates that the IVS1C+1G→A mutation has impaired the splice donor site of intron IC and restricted the alternatively spliced transcripts from $Q0_{porto}$ to the sole product that does not use this site (Fig. 3). Since PI primary transcripts in hepatocytes lack the splice donor site in intron IA due to transcription initiation in exon IC, no normally spliced $Q0_{porto}$ RNA products are expected to be produced in the liver under basal conditions.

In order to search for mRNA species resulting from a read-through of intron IC that could have failed to be amplified due to the long size of this intron, we have performed additional cDNA amplifications from index case 3 using primer pairs IA/ ICintR and IB/ICintR. In both cases no further amplification products were detected, while control genomic DNA samples yielded expected 2072bp and 452bp fragments from primers IA/ICintR and IB/ICintR, respectively. This lack of detectable mRNA sequences including the initial portions of intron IC, suggests that aberrantly spliced transcripts are likely to be degraded.

Discussion

We have characterized the molecular basis for two different PI variants, which illustrate the variety of

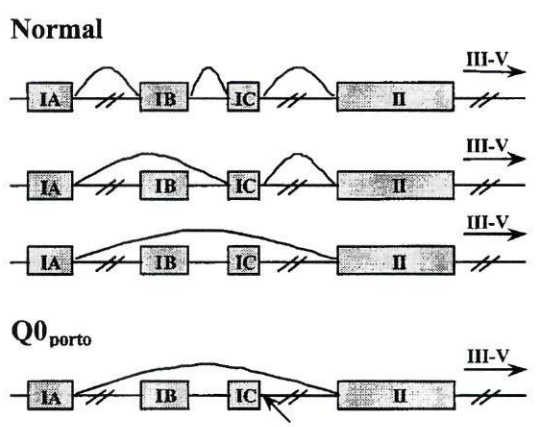


Fig. 3. Schematic representation of normal and $Q0_{porto}$ patterns of alternative splicing as deduced from the exon composition of mRNA species produced by macrophage-specific transcription start sites and detected by RT-PCR analysis. In the presence of the intron IC +1 G→A mutation, indicated by an arrow, production of normally spliced mRNA depends on the use of the intron IA splice donor site that is absent in liver transcripts.

mechanisms that may cause the absence of detectable circulating protein.

The insertional mutation observed in $Q0_{ourém}$, which occurs in a M3 base allele, has been previously described in a different M1Val213 (Val213-Glu376) normal allelic background in the $Q0_{mattawa}$ variant, which was found in two heterozygous patients from Canada that had also inherited the $Q0_{bellingham}$ allele (12, 19). While the possibility of a recombination event between M3 and $Q0_{mattawa}$ cannot be definitely excluded, the geographical dispersion of $Q0_{ourém}$ and $Q0_{mattawa}$, together with an expected increase in mutational likelihood at codons 352-353 due to slippage-mispairing promotion, suggests that it is more probable that both variants have resulted from independent mutations (20, 21). In this respect, the 352-353 region may be analogous to the mutational hot spot located in the run of seven cytosines encompassing codons 360-362, where a C deletion and two likely independent C insertions have been reported in variants $Q0_{bolton}$, $Q0_{saarbrucken}$ and $Q0_{clayton}$, respectively (8, 22-24).

The variant $Q0_{porto}$ is, to our knowledge, the third reported PI deficiency allele resulting from mutations in RNA splice sites, together with $Q0_{west}$ and $Q0_{bonny\ blue}$, which were found to have base substitutions in splice donor and acceptor sites of intron II, respectively (8, 10, 11). The present finding of a novel IVS1C+1G→A transition extends the mutation spectrum to the 5' untranslated region of the PI gene and highlights the potential pathogenic effects that may result from restriction of transcription start sites and splice alternatives in liver-specific expression. Although it has been shown that, during modulated expression by the acute phase mediator interleukin 6, hepatocytes can switch on transcription from exon IA transcriptional initiation sites (6), these transcripts represent only a minor fraction of the overall PI mRNA production (6). It is therefore unlikely that this mechanism might compensate for the lack of correctly spliced mRNA transcribed from exon IC. In this context, the lack of detectable plasma isoelectric focusing PI bands associated with $Q0_{porto}$ is likely to be caused by the absence of splice alternatives that could buffer the effects of the IVS1C+1G→A mutation in the liver, where the bulk of PI synthesis occurs.

Taken together, our results suggest that the characterization of naturally occurring mutations will continually provide further insights into the diversity of mechanisms leading to PI deficiency.

Acknowledgements

We thank the patients and their families for consent and collaboration in this study. We are especially grateful to Drs. Raquel Seruca and Carla Oliveira for helpful suggestions and to two anonymous reviewers for comments on the manuscript. This work was supported in part by Conselho de Prevenção do Tabagismo and POCTI. Susana Seixas is supported by grant BD/13885/97 from Praxis XXI

References

1. Long GL, Chandra T, Woo SLC, Davie EW, Kurachi K. Complete sequence of the cDNA for human α1-antitrypsin and the gene for the S variant. Biochemistry 1984: 23: 4828-4837.
2. Perlino E, Cortese R, Ciliberto G. The human α1-antitrypsin gene is transcribed from two different promoters in macrophages and hepatocytes. EMBO J 1987: 6: 2767-2771.
3. Crystal RG. The α1-antitrypsin gene and its deficiency states. Trends Genet 1989: 5: 411-417.
4. Perlmutter DH, Cole FS, Killbridge P, Rossing TH, Colten HR. Expression of the alpha-1-proteinase inhibitor gene in human monocytes and macrophages. Proc Natl Acad Sci USA 1985: 82: 795-799.
5. du Bois RM, Bernaudin JF, Paakko P, Hubbard R, Takahashi H, Ferrans V, Crystal RG. Human neutrophils express the alpha 1-antitrypsin gene and produce alpha 1-antitrypsin. Blood 1991: 77: 2724-2730.
6. Hafeez W, Ciliberto G, Perlmutter DH. Constitutive and modulated expression of the human α1-antitrypsin gene. Different transcriptional initiation sites used in three different cell types. J Clin Invest 1992: 89: 1214-1222.
7. Blanco I, Fernández E, Bustillo EF. Alpha-1-antitrypsin PI phenotypes S and Z in Europe: an analysis of the published surveys. Clin Genet 2001: 60: 31-41.
8. Brantly M. Alpha-1-antitrypsin genotypes and phenotypes. In: Crystal RG, ed. Alpha-1-antitrypsin deficiency: biology, pathogenesis, clinical manifestations and therapy, Vol. 88. New York: Marcel Dekker Inc., 1996: 45-59.
9. Crystal RG. α1-Antitrypsin deficiency, emphysema and liver disease: genetic basis and strategies for therapy. J Clin Invest 1990: 85: 1343-1352.
10. Lee JH, Brantly M. Molecular mechanisms of alpha-1-antitrypsin null alleles. Respir Med 2000: 94 (Suppl. C): S7-S11.
11. Laubach VE, Ryan WJ, Brantly M. Characterization of a human α1-antitrypsin null allele involving aberrant mRNA splicing. Hum Mol Genet 1993: 2: 1001-1005.
12. Curiel D, Brantly M, Curiel E, Stier L, Crystal RG. α1-Antitrypsin deficiency caused by the α1-antitrypsin $null_{mattawa}$ gene. An insertion mutation rendering the α1-antitrypsin gene incapable of producing α1-antitrypsin. J Clin Invest 1989: 83: 1144-1152.
13. Rocha J, Pinto D, Santos MT, Amorim A, Amil-Dias J, Cardoso-Rodrigues F, Aguiar A. Analysis of the allelic diversity of a (CA)n repeat polymorphism among α1-antitrypsin gene products from northern Portugal. Hum Genet 1997: 99:194-198.
14. Dry PJ. Rapid detection of alpha-1-antitrypsin deficiency by analysis of a PCR-induced *Taq*I restriction site. Hum Genet 1991: 87: 742-744.
15. Miller SA, Dykes DD, Polesky HF. A simple procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988: 16: 1215.
16. Graham A, Kalsheker NA, Bamforth FJ, Newton CR, Markham AF. Molecular characterization of two alpha-1-antitrypsin deficiency variants: proteinase inhibitor (Pi) Nullnewport (Gly115→Ser) and (Pi) Zwrexham (Ser-19 →Leu). Hum Genet 1990: 85: 537-540.

17. Seixas S, Garcia O, Amorim A, Rocha J. A novel alpha-1-antitrypsin R281del variant found in a population sample from the Basque Country. Hum Mutat 2000: 15:121-122.
18. Rollini P, Fournier REK. Differential regulation of gene activity and chromatin structure within the human serpin gene cluster at 14q32.1 in macrophage microcell hybrids. Nucleic Acids Res 2000: 28: 1767-1777.
19. Cox DW, Levison H. Emphysema of early onset associated with a complete deficiency of alpha-1-antitrypsin (null homozygotes). Am Rev Respir Dis 1988: 137: 371-375.
20. Krawczack M, Reitsma PH, Cooper DN. The mutational demography of protein C deficiency. Hum Genet 1995: 96: 142-146.
21. Cooper DN, Krawczack M, Antonarakis SE. The nature and mechanisms of human gene mutation. In: Scriver CH, Beaudet AL, Sly WS, Valle D, eds. The metabolic and molecular bases of inherited disease, Vol. 1. New York: McGraw-Hill, 1995: 259-291.
22. Fraizer GC, Siewersten M, Harrold TR, Cox DW. Deletion/frameshift mutation in the alpha 1-antitrypsin null allele, PI*Q0bolton. Hum Genet 1989: 83: 377-382.
23. Faber J-P, Poller W, Weidinger S, Kirchgesser M, Schwaab R, Bidlingmaier F, Olek K. Identification and DNA sequence analysis of 15 new α1-antitrypsin variants, including two PI*Q0 alleles and one deficient PI*M allele. Am J Hum Genet 1994: 55: 1113-1121.
24. Brantly M, Lee JH, Hildesheim J, Uhm GS, Prakash UBS, Staats BA, Crystal RG. α1-Antitrypsin gene mutation hot spot associated with the formation of a retained and degraded null variant. Am J Respir Cell Mol Biol 1997: 16: 225-231.

3. Referências Bibliográficas

Albarrán, C., Garcia, O., Alonso, A., Deka, R., Martín, P., Trovoada, M. J., Amorim, A., Sancho, M. (1998). Patterns of haplotype variation at the D1S80 locus and a flank sequence polymorphism in African and non-African populations. *Prog Forensic Genet* **7**:401-403.

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

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

Brantly, M. (1996). Alpha-1-antitrypsin genotypes and phenotypes. In: Crystal, R. G. (ed) *Alpha-1-antitrypsin deficiency: biology, pathogenesis, clinical manifestations and therapy.* Marcel Dekker Inc, New York, pp. 45-59.

Brantly, M., Nukiwa, T., Crystal, R. G. (1988). Molecular basis of alpha-1-antitrypsin deficiency. *Am J Med* **84**:13-31.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62**:1408-1415.

Buard, J., Brenner, C., Jeffreys, A. J. (2002). Evolutionary fate of an unstable human minisatellite deduced from sperm-mutation spectra of individual alleles. *Am J Hum Genet* **70**:1038-1043.

Byth, B. C., Billingsley, G. D., Cox, D. W. (1994). Physical and genetic mapping of the serpin gene cluster at 14q32.1: allelic association and a unique haplotype associated with alpha-1-antitrypsin deficiency. *Am J Hum Genet* **55**:126-133.

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

Calafell, F., Shuster, A., Speed, W. C., Kidd, J. R., Kidd, K. K. (1998). Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet* **6**:38-49.

Carlson, J. A., Rogers, B. B., Sifers, R. N., Finegold, M. J., Clift, S. M., DeMayo, F. J., Bullock, D. W., Woo, S. L. (1989). Accumulation of PiZ alpha-1-antitrypsin causes liver damage in transgenic mice. *J Clin Invest* **83**:1183-1190.

Carrell, R. W., Lomas, D. A. (2002). Alpha-1-antitrypsin deficiency--a model for conformational diseases. *N Engl J Med* **346**:45-53.

Carvalho-Silva, D. R., Santos, F. R., Hutz, M. H., Salzano, F. M., Pena, S. D. (1999). Divergent human Y-chromosome microsatellite evolution rates. *J Mol Evol* **49**:204-214.

Cavalli-Sforza, L. L., Menozzi, P., Piazza, A. (1994). *The history and geography of human genes*. Princeton University Press, Princeton, New Jersey.

Charlesworth, D., Charlesworth, B., Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**:1619-1632.

Cooper, D. N., Krawczak, M, Antonarakis, S. E. (1995). The Nature and Mechanisms of Human Gene Mutation. In: Sriver, C. R., Beaudet, A. L., Sly, W., Valle, D. (eds). *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill, New York, pp. 259-291.

Cooper, D. N., Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* **83**:181-188.

Cooper, D. N., Krawczak, M. (1990). The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet* **85**:55-74.

Cox, D. W. (1995). Alpha-1-antitrypsin deficiency. In: Sriver, C. R., Beaudet, A. L., Sly, W., Valle, D. (eds). *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill, New York, pp. 4125-4158

Cox, D. W., Levison, H. (1988). Emphysema of early onset associated with a complete deficiency of alpha-1-antitrypsin (null homozygotes). *Am Rev Respir Dis* **137**:371-375.

Cox, D. W., Smyth, S. (1983). Risk for liver disease in adults with alpha-1-antitrypsin deficiency. *Am J Med* **74**:221-227.

Cox, D. W., Woo, S. L., Mansfield, T. (1985). DNA restriction fragments associated with alpha-1-antitrypsin indicate a single origin for deficiency allele PI Z. *Nature* **316**:79-81.

Crawford, M. H. (1998). *The Origins of Native Americans: evidence from anthropological genetics*. Cambridge University Press, Cambridge.

Creighton, T. E. (1992). *Proteins: Structures and Molecular Properties*. W. H. Freeman Company, New York.

Crystal, R. G. (1990). Alpha-1-antitrypsin deficiency, emphysema, and liver disease. Genetic basis and strategies for therapy. *J Clin Invest* **85**:1343-1352.

Curiel, D., Brantly, M., Curiel, E., Stier, L., Crystal, R. G. (1989b). Alpha-1-antitrypsin deficiency caused by the alpha-1-antitrypsin Nullmattawa gene. An insertion mutation rendering the alpha-1-antitrypsin gene incapable of producing alpha-1-antitrypsin. *J Clin Invest* **83**:1144-1152.

Curiel, D. T., Holmes, M. D., Okayama, H., Brantly, M. L., Vogelmeier, C., Travis, W. D., Stier, L. E., Perks, W. H., Crystal, R. G. (1989a). Molecular basis of the liver and lung disease associated with the alpha-1-antitrypsin deficiency allele Mmalton. *J Biol Chem* **264**:13938-13945.

Dafforn, T. R., Mahadeva, R., Elliott, P. R., Sivasothy, P., Lomas, D. A. (1999). A kinetic mechanism for the polymerization of alpha-1-antitrypsin. *J Biol Chem* **274**:9548-9555.

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.

Dillehay, T. D. (1999). The Late Pleistocene Cultures of South America. *Evol Anthropol* **7**:206-216.

Dycaico, M. J., Grant, S. G., Felts, K., Nichols, W. S., Geller, S. A., Hager, J. H., Pollard, A. J., Kohler, S. W., Short, H. P., Jirik, F. R. (1988). Neonatal hepatitis induced by alpha-1-antitrypsin: a transgenic mouse model. *Science* **242**:1409-1412.

Elliott, P. R., Lomas, D. A., Carrell, R. W., Abrahams, J. P. (1996). Inhibitory conformation of the reactive loop of alpha-1-antitrypsin. *Nat Struct Biol* **3**:676-681.

Eriksson, S., Carlson, J., Velez, R. (1986). Risk of cirrhosis and primary liver cancer in alpha-1-antitrypsin deficiency. *N Engl J Med* **314**:736-739.

Faber, J. P., Poller, W., Weidinger, S., Kirchgesser, M., Schwaab, R., Bidlingmaier, F., Olek, K. (1994). Identification and DNA sequence analysis of 15 new alpha-1-antitrypsin variants, including two PI*Q0 alleles and one deficient PI*M allele. *Am J Hum Genet* **55**:1113-1121.

Garza, J. C., Slatkin, M., Freimer, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol Biol Evol* **12**:594-603.

Goldstein, D. B., Pollock, D. D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered* **88**:335-342.

Goodwin, R. L., Baumann, H., Berger, F. G. (1996). Patterns of divergence during evolution of alpha-1-proteinase inhibitors in mammals. *Mol Biol Evol* **13**:346-358.

Guo, S. W., Xiong, M. (1997). Estimating the age of mutant disease alleles based on linkage disequilibrium. *Hum Hered* **47**:315-337.

Harpending, H. C., Eller, E (2000). Human Diversity and its history. In: Kato, M. (ed) *The Biology of Biodiversity*. Springer-Verlag, Tokyo.

Hildesheim, J., Kinsley, G., Bissell, M., Pierce, J., Brantly, M. (1993). Genetic diversity from a limited repertoire of mutations on different common allelic backgrounds: alpha-1-antitrypsin deficiency variant Pduarte. *Hum Mutat* **2**:221-228.

Hill, R. E., Hastie, N. D. (1987). Accelerated evolution in the reactive centre regions of serine protease inhibitors. *Nature* **326**:96-99.

Hofker, M. H., Nukiwa, T., van Paassen, H. M., Nelen, M., Kramps, J. A., Klasen, E. C., Frants, R. R., Crystal, R. G. (1989). A Pro--Leu substitution in codon 369 of the alpha-1-antitrypsin deficiency variant PI MHeerlen. *Hum Genet* **81**:264-268.

Horai, S., Kondo, R., Nakagawa-Hattori, Y., Hayashi, S., Sonoda, S., Tajima, K. (1993). Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Mol Biol Evol* **10**:23-47.

Huber, R., Carrell, R. W. (1989). Implications of the three-dimensional structure of alpha-1-antitrypsin for structure and function of serpins. *Biochemistry* **28**:8951-8966.

Hurles, M. E., Veitia, R., Arroyo, E., Armenteros, M., Bertranpetit, J., Perez-Lezaun, A., Bosch, E., Shlumukova, M., Cambon-Thomsen, A., McElreavey, K., Lopez, D. M., Rohl, 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.

Hutchison, D. C. (1998). Alpha-1-antitrypsin deficiency in Europe: geographical distribution of Pi types S and Z. *Respir Med* **92**:367-377.

Irving, J. A., Pike, R. N., Lesk, A. M., Whisstock, J. C. (2000). Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res* **10**:1845-1864.

Jardi, R., Rodriguez-Frias, F., Lopez-Talavera, J. C., Miravitlles, M., Cotrina, M., Costa, X., Pascual, C., Vidal, R. (2000). Characterization of the new alpha-1-antitrypsin-deficient PI M-type allele, PI M(vall d'hebron) (Pro(369)-->Ser). *Hum Hered* **50**:320-321.

Jin, L., Macaubas, C., Hallmayer, J., Kimura, A., Mignot, E. (1996). Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc Natl Acad Sci U S A* **93**:15285-15288.

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 U S A* **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.

Kalsheker, N., Hayes, K., Weidinger, S., Graham, A. (1992). What is Pi (proteinase inhibitor) null or PiQO?: a problem highlighted by the alpha-1-antitrypsin Mheerlen mutation. *J Med Genet* **29**:27-29.

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

Krawczak, M., Cooper, D. N. (1991). Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum Genet* **86**:425-441.

Krawczak, M., Cooper, D. N. (1996). Single base-pair substitutions in pathology and evolution: two sides to the same coin. *Hum Mutat* **8**:23-31.

Krawczak, M., Reitsma, P. H., Cooper, D. N. (1995). The mutational demography of protein C deficiency. *Hum Genet* **96**:142-146.

Kunkel, T. A. (1985). The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* **260**:5787-5796.

Lee, J. H., Brantly, M. (2000). Molecular mechanisms of alpha-1-antitrypsin null alleles. *Respir Med* **94 Suppl C**:S7-11.

Lewontin, R. C., Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**:175-195.

Lomas, D. A., Evans, D. L., Finch, J. T., Carrell, R. W. (1992). The mechanism of Z alpha-1-antitrypsin accumulation in the liver. *Nature* **357**:605-607.

Long, G. L., Chandra, T., Woo, S. L., Davie, E. W., Kurachi, K. (1984). Complete sequence of the cDNA for human alpha-1-antitrypsin and the gene for the S variant. *Biochemistry* **23**:4828-4837.

Luiselli, D., Simoni, L., Tarazona-Santos, E., Pastor, S., Pettener, D. (2000). Genetic structure of Quechua-speakers of the Central Andes and geographic patterns of gene frequencies in South Amerindian populations. *Am J Phys Anthropol* **113**:5-17.

Mahadeva, R., Chang, W. S., Dafforn, T. R., Oakley, D. J., Foreman, R. C., Calvin, J., Wight, D. G., Lomas, D. A. (1999). Heteropolymerization of S, I, and Z alpha-1-antitrypsin and liver cirrhosis. *J Clin Invest* **103**:999-1006.

Massi, G. (1996). Pathogenesis and pathology of liver disease associated with alpha-1-antitrypsin deficiency. *Chest* **110**:251S-255S.

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

Mesa, N. R., Mondragon, M. C., Soto, I. D., Parra, M. V., Duque, C., Ortiz-Barrientos, D., Garcia, L. F., Velez, I. D., Bravo, M. L., Munera, 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.

Miller, M. P., Kumar, S. (2001). Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* **10**:2319-2328.

Mountain, J. L., Cavalli-Sforza, L. L. (1994). Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc Natl Acad Sci U S A* **91**:6515-6519.

Nauta, M. J., Weissing, F. J. (1996). Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* **143**:1021-1032.

Nukiwa, T., Ogushi, F., Crystal, R. G. (1996a). Alpha-1-antitrypsin gene evolution. In: Crystal, R. G. (ed). *Alpha-1-antitrypsin deficiency. Biology, pathogenesis, clinical manifestations, therapy.* Marcel Dekker Inc, New York, pp. 33-43.

Nukiwa, T., Seyama, K., Kira, S. (1996b). The Prevalence of a1AT Deficiency Outside the United States and Europe. In: Crystal, R. G. (ed). *Alpha-1-antitrypsin deficiency. Biology, pathogenesis, clinical manifestations, therapy.* Marcel Dekker Inc, New York, pp. 293-300.

Owen, M. C., Brennan, S. O., Lewis, J. H., Carrell, R. W. (1983). Mutation of antitrypsin to antithrombin. alpha-1-antitrypsin Pittsburgh (358 Met leads to Arg), a fatal bleeding disorder. *N Engl J Med* **309**:694-698.

Pena, S. D., Santos, F. R., Bianchi, N. O., Bravi, C. M., Carnese, F. R., Rothhammer, F., Gerelsaikhan, T., Munkhtuja, B., Oyunsuren, T. (1995). A major founder Y-chromosome haplotype in Amerindians. *Nat Genet* **11**:15-16.

Poller, W., Merklein, F., Schneider-Rasp, S., Haack, A., Fechner, H., Wang, H., Anagnostopoulos, I., Weidinger, S. (1999). Molecular characterisation of the defective alpha-

1-antitrypsin alleles PI Mwurzburg (Pro369Ser), Mheerlen (Pro369Leu), and Q0lisbon (Thr68Ile). *Eur J Hum Genet* **7**:321-331.

Potempa, J., Korzus, E., Travis, J. (1994). The serpin superfamily of proteinase inhibitors: structure, function, and regulation. *J Biol Chem* **269**:15957-15960.

Prata, M. J., Amorim, A., Gusmao, L., Trovoada, M. J. (1996). Population genetics of the STRs TPO,TH01 and VWFA31/A in São Tomé e Príncipe. *Adv Forensic Haemogenet* **7** :604-606.

Rannala, B., Bertorelle, G. (2001). Using linked markers to infer the age of a mutation. *Hum Mutat* **18**:87-100.

Reich, D. E. Goldstein, D. B. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proc Natl Acad Sci U S A* **95**:8119-8123.

Relethford, J. H. (1997). Mutation rate and excess African heterozygosity. *Hum Biol* **69**:785-792.

Relethford, J. H. (2001). *Genetics and the Search for Modern Human Origins.* Wiley-Liss, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto.

Rocha, J., Seixas, S., Lopes, A. I., Silva, L., Salgueiro, C., Salazar-de-Sousa, J., and Batista, A. (1999). Human gene mutation report. *Hum Genet* **104**:114.

Rogers, A. R., Jorde, L. B. (1995). Genetic evidence on modern human origins. *Hum Biol* **67**:1-36.

Rogers, A. R., Jorde, L. B. (1996). Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* **58**:1033-1041.

Rollini, P., Fournier, R. E. (1997). A 370-kb cosmid contig of the serpin gene cluster on human chromosome 14q32.1: molecular linkage of the genes encoding alpha-1-antichymotrypsin, protein C inhibitor, kallistatin, alpha-1-antitrypsin, and corticosteroid-binding globulin. *Genomics* **46**:409-415.

Ruiz-Linares, A., Ortiz-Barrientos, D., Figueroa, M., Mesa, N., Munera, J. G., Bedoya, G., Velez, I. D., Garcia, L. F., Perez-Lezaun, A., Bertranpetit, J., Feldman, M. W., Goldstein, D. B. (1999). Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. *Proc Natl Acad Sci U S A* **96**:6312-6317.

Salzet, M., Vieau, D., Stefano, G. B. (1999). Serpins: an evolutionarily conserved survival strategy. *Immunol Today* **20**:541-544.

Santos, F. R., Rodriguez-Delfin, L., Pena, S. D., Moore, J., Weiss, K. M. (1996). North and South Amerindians may have the same major founder Y chromosome haplotype. *Am J Hum Genet* **58**:1369-1370.

Schlötterer, C. (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**:753-763.

Schlötterer, C., Wiehe, T. (1999). Microsatellites, a neutral marker to infer selective sweeps. In: Goldstein, D. B., Schlötterer C. (eds). *Microsatellites. Evolution and aplications*. Oxford University Press, Oxford, pp. 238-248.

Serre, J. L., Simon-Bouy, B., Mornet, E., Jaume-Roig, B., Balassopoulou, A., Schwartz, M., Taillandier, A., Boue, J., Boue, A. (1990). Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Hum Genet* **84**:449-454.

Seyama, K., Nukiwa, T., Takabe, K., Takahashi, H., Miyake, K., Kira, S. (1991). Siiyama (serine 53 (TCC) to phenylalanine 53 (TTC)). A new alpha-1-antitrypsin-deficient variant with mutation on a predicted conserved residue of the serpin backbone. *J Biol Chem* **266**:12627-12632.

Slatkin, M. (1995). Hitchhiking and associative overdominance at a microsatellite locus. *Mol Biol Evol* **12**:473-480.

Slatkin, M., Rannala, B. (2000). Estimating allele age. *Annu Rev Genomics Hum Genet* **1:225-49.**:225-249.

Stein, P. E., Carrell, R. W. (1995). What do dysfunctional serpins tell us about molecular mobility and disease? *Nat Struct Biol* **2**:96-113.

Sveger, T. (1976). Liver disease in alpha-1-antitrypsin deficiency detected by screening of 200,000 infants. *N Engl J Med* **294**:1316-1321.

Tarazona-Santos, E., Carvalho-Silva, D. R., Pettener, D., Luiselli, D., De Stefano, G. F., Labarga, C. M., Rickards, O., Tyler-Smith, C., Pena, S. D., Santos, F. R. (2001). Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. *Am J Hum Genet* **68**:1485-1496.

Teckman, J. H., Burrows, J., Hidvegi, T., Schmidt, B., Hale, P. D., Perlmutter, D. H. (2001). The proteasome participates in degradation of mutant alpha-1-antitrypsin Z in the endoplasmic reticulum of hepatoma-derived hepatocytes. *J Biol Chem* **276**:44865-44872.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A* **97**:7360-7365.

Torroni, A., Bandelt, H. J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M. L., Bonne-Tamir, B., Scozzari, R. (1998). mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* **62**:1137-1152.

Torroni, A., Bandelt, H. J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martinez-Cabrera, V., Villems, R., Kivisild, T., Metspalu, E., Parik, J., Tolk, H. V., Tambets, K., Forster, P., Karger, B., Francalacci, P., Rudan, P., Janicijevic, B., Rickards, O., Savontaus, M. L., Huoponen, K., Laitinen, V., Koivumaki, S., Sykes, B., Hickey, E., Novelletto, A., Moral, P., Sellitto, D., Coppa, A., Al Zaheri, N., Santachiara-Benerecetti, A. S., Semino, O., Scozzari, R. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* **69**:844-852.

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

Watkins, W. S., Ricker, C. E., Bamshad, M. J., Carroll, M. L., Nguyen, S. V., Batzer, M. A., Harpending, H. C., Rogers, A. R., Jorde, L. B. (2001). Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *Am J Hum Genet* **68**:738-752.

Werb, Z., Banda, M. J., McKerrow, J. H., Sandhaus, R. A. (1982). Elastases and elastin degradation. *J Invest Dermatol* **79 Suppl 1:**:154s-159s.

WHO (1997). Alpha-1-antitrypsin deficiency: memorandum from a WHO meeting. *Bull World Health Organ* **75**:397-415.

Wu, Y., Whitman, I., Molmenti, E., Moore, K., Hippenmeyer, P., Perlmutter, D. H. (1994). A lag in intracellular degradation of mutant alpha-1-antitrypsin correlates with the liver disease phenotype in homozygous PiZZ alpha-1-antitrypsin deficiency. *Proc Natl Acad Sci U S A* **91**:9014-9018.

Xu, X., Peng, M., Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24**:396-399.

PARTE III

Grupo Sanguíneo Duffy

1. Introdução

O grupo sanguíneo Duffy tem um polimorfismo genético que é essencialmente determinado por três alelos: FY*A, FY*B e FY*O. Os produtos génicos dos alelos FY*A e FY*B são glicoproteínas transmembranares que funcionam como receptores de quimiocinas pro-inflamatórias e determinam as especificidades antigénicas deste grupo nos eritrócitos. O alelo FY*O, pelo contrário, não está associado à expressão de qualquer antigénio eritrocitário característico, embora produza glicoproteínas com a estrutura do antigénio B nas membranas de outros tipos de células (Tournamille *et al.*, 1995a; Hadley e Peiper, 1997). Para além da sua função de receptores de quimiocinas, os antigénios FY são utilizados pelos merozoítos de *Plasmodium vivax* como locais de ligação durante a invasão dos eritrócitos e a sua ausência confere aos homozigóticos FY*O/FY*O resistência total à forma de malária causada por este parasita (Miller *et al.*, 1976; Hadley e Peiper, 1997).

A análise da sequência do gene FY e a sua comparação com sequências ortólogas de primatas não-humanos revelaram que FY*B é o alelo ancestral a partir do qual se originaram FY*A e FY*O através de uma única mutação (Figura III.1) (Chaudhuri *et al.*, 1995). No caso de FY*A, há uma substituição Asp44Gly que não interfere com a expressão ou com a função da proteína (Tournamille *et al.*, 1995b). Em FY*O, há uma mutação T–46C na região promotora do gene, que elimina um local de ligação ao factor de transcrição GATA1 e suprime a expressão de antigénios na linha eritróide (Tournamille *et al.*, 1995a).

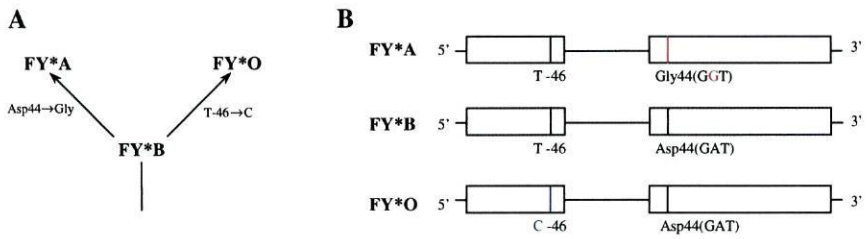
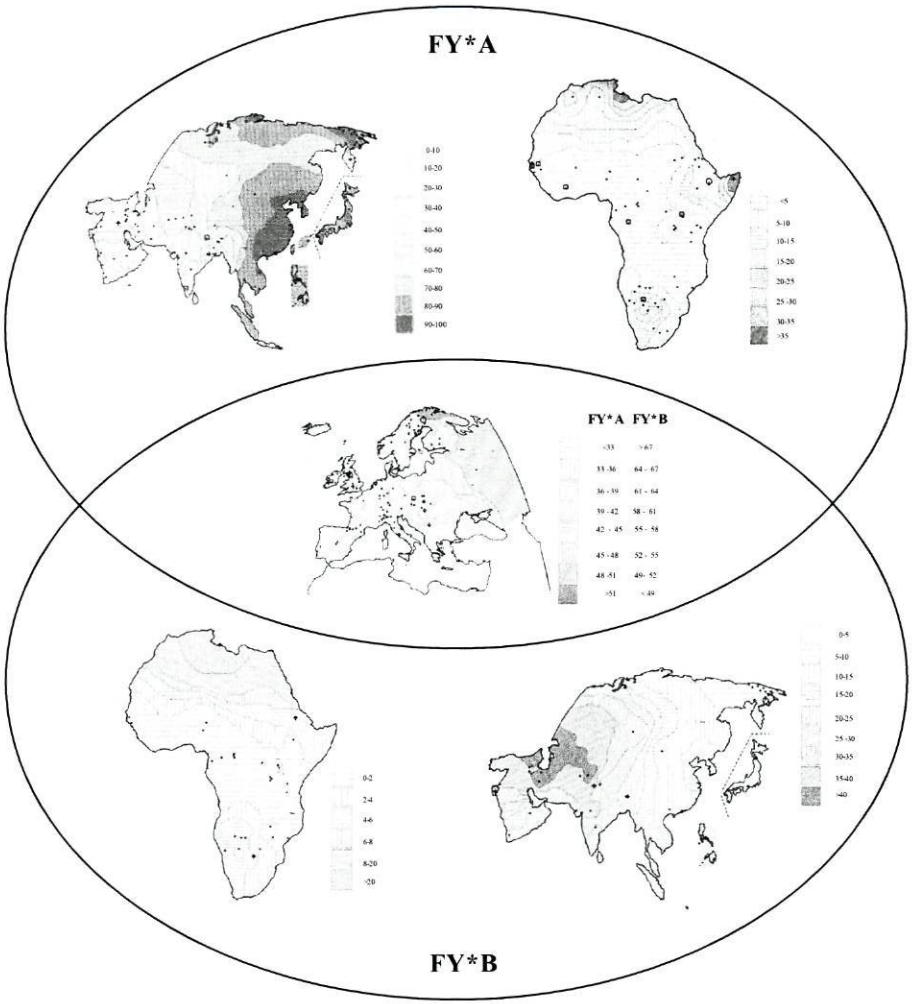


Figura III.1- A- Filogenia dos alelos do grupo sanguíneo Duffy. **B-** Representação esquemática do gene FY. Os exões são representados por rectângulos e o intrão por uma linha horizontal. As posições relativas das substituições nucleotídicas que estão na base do polimorfismo são indicadas por linhas verticais.

Ao contrário da generalidade dos polimorfismos autossómicos humanos em que os alelos mais frequentes de cada *locus* tendem a ser partilhados por populações muito diferentes, a distribuição das frequências dos alelos do grupo sanguíneo Duffy caracteriza-se por uma acentuada segregação geográfica e tem um nível de diferenciação interpopulacional que é claramente superior à média. Na Eurásia, praticamente só existem os alelos FY*A e FY*B, enquanto que em África o alelo FY*O atinge frequências superiores a 90% num grande número de populações da região sub-sahariana de onde o parasita *P. vivax* está virtualmente ausente (Figura III.2).

Devido às características particulares desta distribuição geográfica e ao seu envolvimento na resistência à infecção por *P. vivax,* o grupo Duffy é um excelente modelo para o estudo da influência da selecção natural na variabilidade genética de regiões funcionalmente relevantes do genoma, à semelhança de outros *loci* que codificam proteínas importantes para a estrutura ou metabolismo do eritrócito, como a hemoglobina (cadeias α e β), a desidrogenase da glucose-6-fosfato ou a proteína banda-3 em que também há mutações que conferem protecção contra a malária (Jarolim *et al.*, 1991; Cooke e Hill, 2001).

O padrão de diversidade do *locus* FY tem sido explicado, basicamente, por duas hipóteses alternativas. De acordo com a primeira hipótese, a fixação de FY*O em África resultou directamente da resistência à malária causada por *P. vivax* e conduziu à extinção do parasita por falta de hospedeiros susceptíveis (Figura III.3 A) (Cavalli-Sforza *et al.*, 1994). De acordo com a segunda hipótese, a fixação de FY*O não esteve relacionada com a resistência a *P. vivax*, embora tenha contribuído para impedir, posteriormente, a sua penetração em muitas regiões africanas (Figura III.3 B) (Livingstone, 1984). Entre os argumentos favoráveis a esta hipótese, conta-se a relativa benignidade da infecção por *P. vivax*, e a provável origem extra-africana da transferência do parasita para a nossa espécie, sugerida pela sua adaptação a climas temperados e pela sua homologia com *P. cynomolgi*, que infecta Primatas não-humanos na Índia e no sudeste asiático (Livingstone, 1984). A verificar-se este modelo, a distribuição das frequências de FY poderia ter resultado da acção de um agente selectivo diferente de *P. vivax* ou de factores demográficos relacionados com movimentos migratórios interpopulacionais.



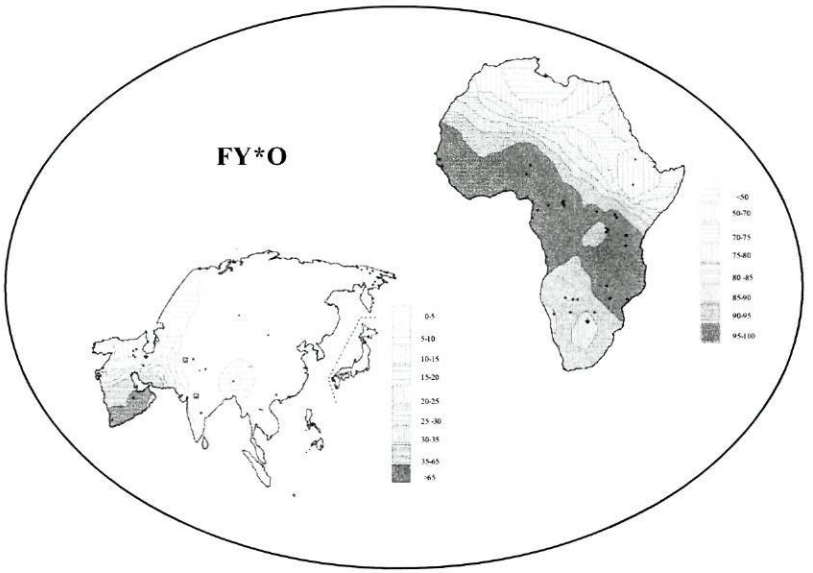


Figura III.2- Distribuição das frequências dos alelos do grupo sanguíneo Duffy (imagem modificada de Cavalli-Sforza *et al.*, 1994).

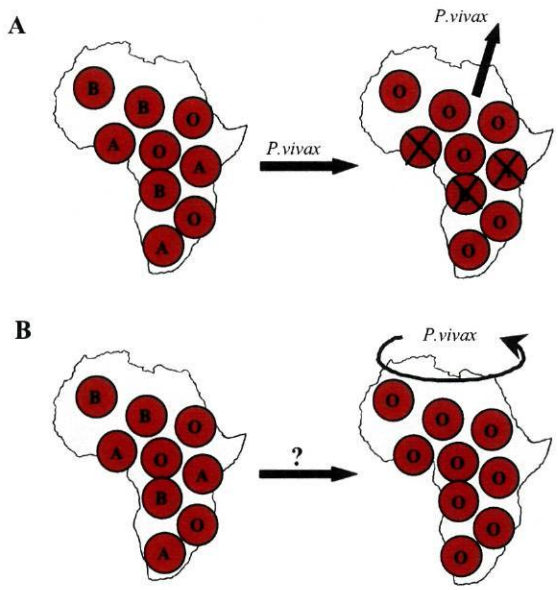


Figura III.3- Representação esquemática das duas hipóteses alternativas que procuram explicar o actual padrão de frequência de FY*O no continente africano. **A**- Hipótese 1: a pressão selectiva de *P. vivax* terá conduzido à fixação de FY*O e à posterior eliminação do parasita por falta de hospedeiros susceptíveis. **B**- Hipótese 2: a fixação de FY*O terá resultado de um acontecimento populacional fortuito ou da pressão selectiva por outro agente patogénico que terá impedido a entrada de *P. vivax* em África.

Recentemente, Hamblin e Di Rienzo (2000) e Hamblin *et al.* (2002), procuraram avaliar o impacto da selecção no *locus* FY através da análise da variação de sequências nucleotídicas adjacentes numa amostra da população italiana e em várias populações africanas. Na amostra italiana apenas se encontraram alelos FY*A e FY*B com frequências de 0,41 e 0,59, respectivamente. As genealogias das linhagens associadas a estes alelos mostraram uma frequência elevada de sequências derivadas e mutações homoplásicas que poderão ter sido transferidas por recombinação para os diferentes haplótipos (Figura III.4). No alelo FY*B observou-se uma predominância de linhagens derivadas particularmente marcada, sem que se tenha detectado qualquer haplótipo que não diferisse da sequência primitiva em pelo menos

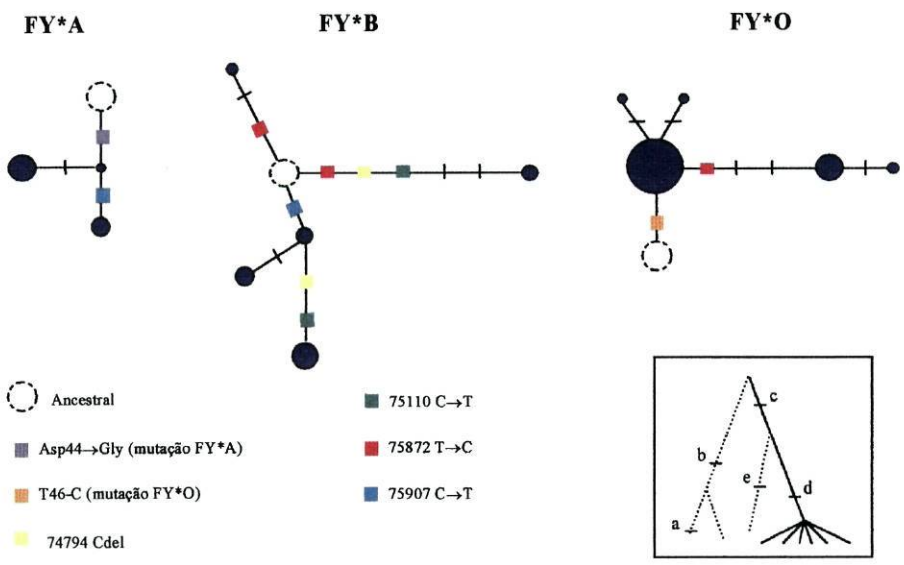


Figura III.4 - Genealogia das linhagens de FY*A, FY*B e FY*O definidas com base na variação de sequência observada por Hamblin e Di Rienzo (2000) numa região de 1.9 kb adjacente ao *locus* FY. As diferentes linhagens, representadas por círculos de áreas proporcionais à sua frequência, diferem da sequência ancestral pela acumulação de mutações nos ramos da genealogia. As mutações dos alelos FY*A e FY*O e as mutações homoplásicas estão assinaladas com quadrados de várias cores. As mutações específicas de cada linhagem estão assinaladas com o símbolo |. A figura enquadrada foi modificada de Fay e Wu (2000) e representa a genealogia esperada para uma região genómica sujeita a selecção positiva. No decorrer de um episódio selectivo espera-se que as linhagens portadoras das mutações neutras (a, b e c) sejam perdidas e que a linhagem portadora da mutação com vantagem selectiva (d) seja fixada.

um passo mutacional (Figura III.4) (Hamblin e Di Rienzo, 2000). Nas populações africanas, onde o alelo FY*O está fixado, observaram-se níveis de diversidade nucleotídica cerca de duas a três vezes menores do que na população europeia (Hamblin e Di Rienzo, 2000). Este resultado está de acordo com o que seria de esperar num modelo simples de substituição selectiva e indica que o intervalo de tempo necessário à fixação do alelo FY*O foi suficientemente curto para impedir a acumulação de variação nucleotídica (Hamblin e Di Rienzo, 2000). No entanto, contrariamente às expectativas teóricas, em vez de se detectar um único haplótipo predominante nas linhagens associadas ao alelo FY*O, observou-se uma genealogia mais complexa, com dois troncos de frequência apreciável que podem ter sido originados antes da hipotética pressão selectiva favorável à fixação daquele alelo (Figura III.4) (Hamblin e Di Rienzo, 2000; Kreitman, 2000). Consequentemente,

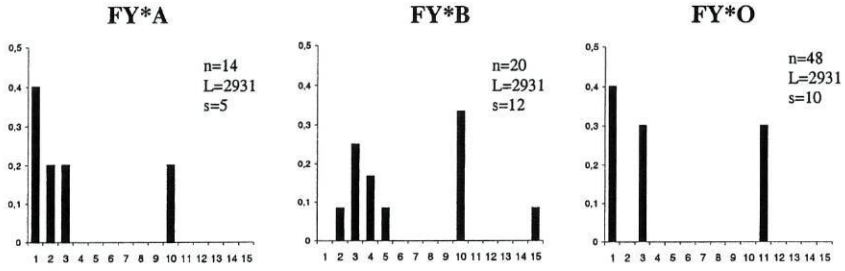


Figura III.5- Espectros de frequência dos locais polimórficos associados aos alelos FY*A, FY*B e FY*O, detectados em duas regiões de 1kb e 1,9kb, respectivamente a 5' do gene FY (Hamblin e Di Rienzo, 2000). n - numero de sequências, L - tamanho da sequência em pares de bases e s- número de locais polimórficos.

o espectro de frequências dos vários locais polimórficos encontrados não revelou a predominância de variantes raras que normalmente é causada pela fixação selectiva de uma mutação logo após a sua origem (Figura III.5).

No trabalho que a seguir se apresenta, procurou-se estudar a evolução do polimorfismo do grupo Duffy com uma abordagem complementar, em que se analisou a variação genética acumulada pelos alelos FY*A, FY*B e FY*O num microssatélite situado na região abrangida pelos estudos anteriores de diversidade nucleotídica (Figura III.6). Este tipo de análise, em muitos aspectos semelhante ao usado no estudo da α1-antitripsina, permitiu combinar um sistema genético hipervariável com as linhagens mais estáveis definidas pelas mutações pontuais, tendo-se obtido informação adicional que não poderia ter sido directamente inferida apenas com o estudo da variação das sequências nucleotídicas. O essencial dos resultados obtidos está condensado no artigo 7, ao qual se segue um comentário em que se discutem algumas das suas principais implicações.

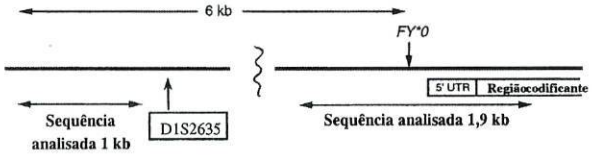


Figura III.6- Esquema da região do *locus* FY em que se representam as duas regiões previamente estudadas por Hamblin e Di Rienzo (2000) e a posição do microssatélite D1S2635 analisado no presente trabalho.

2. Resultados e Discussão

Artigo 7

Seixas, S., Ferrand, N., Rocha, J. (2002). Microsatellite Variation and Evolution of the Human Duffy Blood Group Polymorphism. *Mol Biol Evol* **19**: 1802-1806.

Microsatellite Variation and Evolution of the Human Duffy Blood Group Polymorphism

Susana Seixas, [*†] *Nuno Ferrand,* [†‡] *and Jorge Rocha* [*†]

[*]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal; [†]Faculdade de Ciências, Universidade do Porto, Portugal; [‡]Centro de Estudos de Ciência Animal (CECA), ICETA, Universidade do Porto, Portugal

The human Duffy blood group (FY) antigens are transmembrane glycoproteins that function as receptors for chemokines and for the malarial parasite *Plasmodium vivax* (reviewed in Hadley and Peiper, 1997). Most of FY antigenic variation is determined by three common alleles (FY*A, FY*B and FY*O) in a gene located on chromosome 1q21-q22. DNA sequence characterization of these alleles and interspecies comparisons with the orthologous genes from non-human primates have shown that FY*A and FY*O are derived variants, each resulting from a single mutation in an ancestral FY*B background (Chaudhuri et al. 1995; Tournamille et al. 1995). While the FY*A gene product is a functional protein with a Asp44Gly substitution, FY*O has a T-46C promoter mutation that disrupts a binding site for the GATA1 erythroid transcription factor leading to a tissue-specific loss of expression of FY antigens in red blood cells (Tournamille et al. 1995). In contrast to most human autosomal polymorphisms where common alleles tend to be shared by different, geographically distant populations, the distribution of FY alleles is peculiar: FY*O has reached near-fixation over a vast area of sub-Saharan Africa, while FY*A and FY*B are the only alleles present across Eurasia and the Americas. This peculiarity, together with the observation that homozygous individuals for the FY*O allele are completely resistant to *P. vivax* malaria (Miller et al 1976), has led to the concept that the observed pattern of allele frequencies has been driven by positive selection. According to this model, selection by vivax malaria led to the replacement of FY*A and FY*B by the advantageous FY*O allele in west and central Africa and to the extinction of *P. vivax* by lack of susceptible hosts. Alternatively, since no significant mortality is associated with *P. vivax* and an Asian origin of the parasite is conceivable, it is possible that it was the prior fixation of FY*O that has prevented vivax malaria from becoming endemic in Africa (Livingstone 1984). If this hypothesis is correct, different scenarios may account for the present distribution of FY alleles, including selection by an unknown agent other than *P. vivax* or the possibility of an entirely fortuitous event linked to the dynamics of population movements within and out of Africa. In order to search for a signature of natural selection at the FY locus, the patterns of DNA sequence variation linked to the three FY common alleles have been recently characterized (Hamblin and Di Rienzo 2000; Hamblin, Thompson and Di Rienzo, 2002). Consistent with the expectations of models of directional selection, the level of DNA sequence variation associated with FY*O was found to be significantly reduced. However, the observation that the FY*O mutation occurs in two divergent haplotypes with intermediate frequencies in most samples from Sub-Saharan Africa indicated that the signature of selection may be more complex than predicted by a simple selective sweep.

We have approached the evolutionary history of the FY polymorphism by studying the distribution of the faster-mutating D1S2635 microsatellite polymorphism within more-stable lineages carried by FY*A, FY*B and FY*O alleles.

The FY*A and FY*B alleles were sampled from a total of 123 Portuguese individuals (FY*A=0.35; FY*B=0.62; FY*O=0.03). The FY*O alleles were sampled from 141 individuals from the island of São Tomé (FY*A=0.03; FY*B=0.07; FY*O=0.90). This previously uninhabited island, located 300 km off the coast of Gabon, started to be peopled by the end of the 15th century with slaves imported by Portuguese colonists from the adjacent coasts of the Gulf of Guinea and the Congo-Angola area. As a consequence of this settlement pattern, the population of São Tomé has retained the high levels of genetic diversity that are generally observed in the African mainland and has an estimated European admixture of 11% (Tomás et al. 2002). Identification of FY alleles was done by using previously described PCR-RFLP methods (Tournamille et al. 1995). The D1S2635 microsatellite (GenBank accession number Z52215; Table 1) was typed by PCR amplification with fluorescently labeled primers (GDB: 603410; http://www.gdb.org/) followed by separation of amplification products in an ABI 310 DNA sequencer. Two flanking polymorphic sites described by Hamblin and Di Rienzo (2000) in the region around the FY locus were additionally characterized (nucleotide positions as in BAC bk134P22; GenBank accession number AL35403; Table 1): (i) a C→T transition in position 70628, which has previously been found to be always associated with a 69596 T→C transition in a C-T haplotype shared by both FY*O and non-FY*O

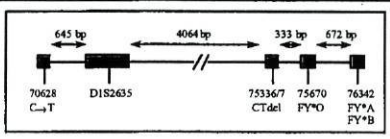
Key words: human Duffy blood group, microsatellite variation, malarial selection.

Address for correspondence and reprints: Jorge Rocha, Instituto de Patologia e Imunologia Molecular, Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal. E-mail:jrocha@ipatimup.pt

Mol. Biol. Evol. 19(10):000-000. 2002

Table 1

Distribution of Haplotypes Defined by Sequence Variation at the D1S2635 Microsatellite and Flanking Polymorphisms Among FY alleles. Location of the Different Polymorphic Regions is Shown in Inset.



Sequence (5'→3')					Frequency[a]		
					Portugal		São Tomé
Position 70628	Microsatellite D1S2635	Positions 75336/7	Haplotype		FY*A N=81	FY*B N=143	FY*O N=222
C	--$(CA)_5$CGCA TG CGTG CA AG CA $(CA)_{n-1}$ CTTCTCTTT CTT--	CT	H1		0.55	0.06	0.44
T[b]	--$(CA)_5$CGCA TG CGTG CA AG CA $(CA)_{n-1}$ CTTCTCTTT CTT--	CT	H2		0.07	0.16	0.03
C	--$(CA)_5$CGCA TG CGTG CA AG CG $(CA)_n$ CTTCTCTTT CTT--	CT	H3		0.29	0.19	0.10
C	--$(CA)_5$CGCA TG CGTG CA AG CA $(CA)_{n-1}$ CTTCTCTTT -TT--	CT	H4		0.05	0.35	-
T	--$(CA)_5$CGCA TG CGTG CA AG CA $(CA)_{n-1}$ CTTCTCTTT -TT--	CT	H5		0.04	0.24	-
C	--$(CA)_5$CGCA TG CGTG CA AG CA $(CA)_{n-1}$ CTTCTCTTT CTT--	del	H6		-	-	0.19
C	--$(CA)_5$CGCA TG CGTG CA AG $(CG)_3$ $(CA)_n$ CTTCTCTTT CTT--	CT	H7		-	-	0.24
C	--$(CA)_5$CG[TG CGCG CGAG CA $(CA)_{10}$ CTTCTCTTT CTT--	CT	Chimpanzee[c]				
C	--$(CA)_5$CG[CGTG CACA CGCG CA $(CA)_{10}$ CTTCTCTTT CTT--	CT	Gorilla				

[a] Haplotypes linked to FY*A, FY*B and FY*O alleles were determined in a subset of the sample used for allele frequency estimation
[b] Derived variants are shaded
[c] The region of the putative ancestral CpG stretch where interspecific mutations had accumulated is boxed

alleles; (ii) a CT deletion at nucleotides 75336/7 that defines one of the two major common lineages associated with FY*O. Both polymorphisms were typed after PCR amplification of DNA fragments containing each of the corresponding positions. The 70628 C/T variation was detected by *Sty*I restriction enzyme digestion. Length variation at nucleotides 75336/7 was scored by electrophoretic separation of amplification products in 12% polyacrylamide gels. Microsatellite alleles were sequenced in both directions from PCR products cloned into a pCR4 plasmid vector with the TOPO TA cloning kit (Invitrogen) using the ABI Prism Big Dye Dydeoxy Terminator Cycle sequencing kit. Sequencing products were analyzed in an ABI 377 automatic DNA sequencer. Human sequences were compared with the homologous regions from one chimpanzee (*Pan troglodytes*) and two gorilla (*Gorilla gorilla*) specimens. Allele frequencies at the individual loci were calculated by direct gene counting. Maximum-likelihood haplotype frequencies were estimated using the expectation-maximization algorithm implemented in the ARLEQUIN package (Schneider et al. 1997). Unbiased estimates of heterozygosity were calculated according to Nei (1987). Significant differences among heterozygosity estimates were tested by comparing the corresponding 95% confidence intervals established by 10000 bootstrap simulations with the GENETIX software (Belkhir et al. 1998).

The D1S2635 microsatellite allele frequency distribution within the FY alleles is depicted in figure 1A. Although the FY*O distribution presents a decreased variance in allele length and a significantly lower heterozygosity than FY*B, there is no drastic reduction in diversity levels, suggesting that the signature of directional selection in microsatellite variation is not as evident as at the DNA sequence level (Hamblin and Di Rienzo 2000; Hamblin, Thompson and Di Rienzo 2002). However, the three microsatellite distributions still have noticeable differences in shape, with FY*O presenting a more smoothly peaked pattern than both non-FY*O alleles. Most interestingly, it was found that the D1S2635 allele size changes within FY*A and FY*B were not always in increments of 2bp. In order to examine the cause of this heterogeneity, 17 D1S2635 alleles from Portugal and 24 alleles from São Tomé were sequenced. Sequence information has revealed four types of D1S2635 lineages that were defined by different flanking polymorphisms (Table 1). The lack of regularity in D1S2635 allele size increments was found to be due to the deletion of a single C within a run of five T's located in the 3' flanking region of the (CA)n repeat. In addition, two further D1S2635 lineages were defined by the presence of one or three CG dinucleotides immediately before the variable (CA)n repeat. The presence of a single CG dinucleotide could have resulted from a CA→CG misincorporation promoted by slipped mispairing in the (CA)n repeated motif. Reiteration of this process has probably led to the occurrence of a (CG)3 motif, although no intermediate sequences bearing a (CG)2 were found. All interspecies differences in the microsatellite sequence structure between humans, chimpanzee and gorillas were found to be the likely result of mutation accumulation in an ancestral CpG

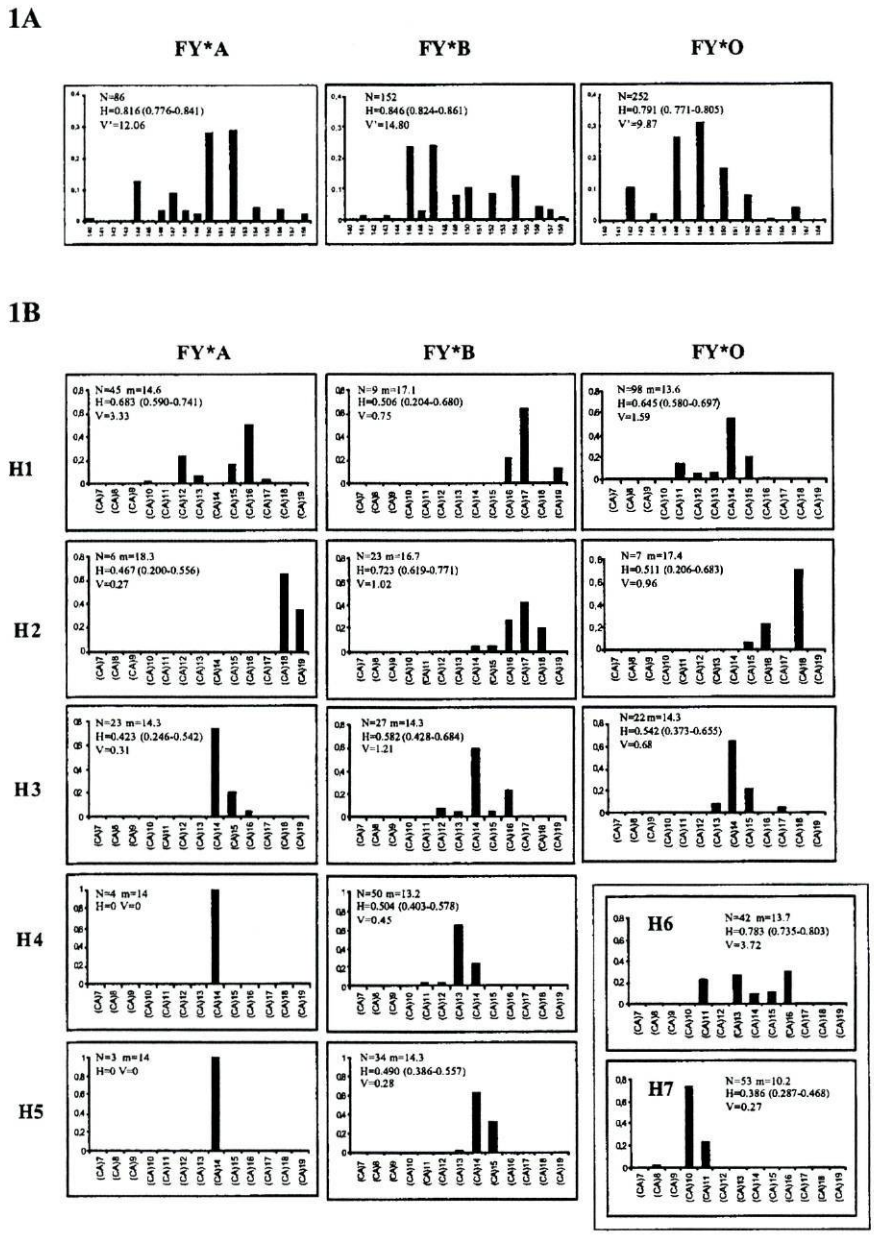


FIG 1.- A. D1S2635 microsatellite allele frequency distributions within FY alleles. B. Distribution of D1S2635 (CA)n variation within the seven haplotypes (H1-H7) carried by FY*A, FY*B and FY*O alleles. N= number of chromosomes, V'= variance in allele length, V= variance in repeat number, m= average repeat number.

dinucleotide stretch lying between an invariable (CA)5 array and a polymorphic (CA)n repeated block (Table 1). The joint distribution of the four types of D1S2635 flanking sequences and the polymorphisms at nucleotides 70628 and 75336/7 was further analyzed in an extended survey, which led to the definition of seven distinct haplotypes (Table 1). In this survey, the D1S2635 indel could be directly typed by the observation of even (insertion) or odd (deletion) numbers of base pairs in the different microsatellite alleles. The presence and number of CG dinucleotides adjacent to the variable (CA)n repeat could be assessed through the digestion of D1S2635 alleles of known length with the *Hin*6I restriction enzyme that recognizes the G!CGC sequence. For example, if two alleles have the same size but

...AAGCG(CA)10... and ...AAGCGCGCG(CA)8... sequences, *Hin*6I digestion will produce in both cases a fragment of constant size, corresponding to the 5' portion of the D1S2635 amplimer, and additional CG (CA)10... and CG (CA)8... fragments, which can be easily discriminated upon polyacrylamide gel electrophoresis. In the absence of CG dinucleotides, microsatellite alleles will remain uncut.

The seven haplotypes had different distributions among the FY alleles (Table1). Haplotype H1, which shows the highest similarity with non-human primates and is likely to be the ancestral lineage, was found to be shared by all three alleles and is most frequent among FY*A and FY*O. Haplotypes H4 and H5 were found exclusively in non-FY*O alleles, while H6 and H7 were detected only among FY*O. Within FY*O exclusive lineages, haplotype H6, defined by the 75336/7 CT deletion, has a 19% frequency and corresponds to one of the two major FY*O haplotype branches previously found in 23% of FY*O alleles from five sub-Saharan African populations (Hamblin and Di Rienzo 2000). Haplotype H7 adds an additional common FY*O sublineage and strengths the disagreement between the observed haplotype structure and the skew of the frequency spectrum toward rare alleles that should be expected under a simple selective sweep. These patterns of lineage sorting along with either FY*O or non-FY*O alleles are consistent with the present geographic distribution of the FY variants and agree with previous observations on sequence variation linked to the FY locus (Li et al. 1997; Hamblin and Di Rienzo, 2000; Hamblin, Thompson and Di Rienzo 2002). However, two derived lineages were found to be shared by all FY alleles. Haplotype H2, is defined by the 70628C→T transition, that was previously found in 6% of FY*O alleles from other African populations and in 50% of Italian FY*B alleles (Hamblin and Di Rienzo, 2000). Haplotype H3 adds further evidence for gene conversion or recombination-driven sequence transfer between FY*O and non-FY*O alleles. This sharing of derived haplotypes between alleles that currently occupy distinct geographical areas shows that both FY*O and non-FY*O lineages were indeed present in the same ancestral population before the fixation of FY*O in Africa.

Figure 1B presents the distributions of the faster-mutating D1S2635 alleles within the haplotypes carried by FY*A, FY*B and FY*O. Derived haplotypes that are shared by at least two FY variants (H2, H3 H4 and H5) have modal (CA)n alleles with the same, or a very similar, number of repeats in each FY allele, thus providing additional evidence for lineage spread through recombination or gene conversion. Analysis of the (CA)n repeat variation within the lineages defined by sequence polymorphisms allows the comparison of diversity levels accumulated since the origin of each haplotype and provides information on the relative antiquity of different lineages that cannot be directly inferred from sequence data alone. Haplotype H6, which corresponds to one of the two major FY*O haplotypes previously described, has been found to be characterized by the joint occurrence in absolute linkage disequilibrium of the 75336/7 CT deletion together with two additional mutations: 75082A→G and 75872 T→C (Hamblin and Di Rienzo 2000). In spite of its derived sequence structure, this haplotype has the highest (CA)n repeat heterozygosity and is likely to be the oldest lineage linked to FY*O. On the contrary, the FY*O-linked haplotype H1, which would be included in the other major FY*O haplotype branch defined by Hamblin and Di Rienzo (2000), is associated with lower levels of (CA)n diversity although it bears a more primitive sequence structure. Taking this evidence into account, it is probable that FY*O has arisen in Africa by two independent mutational events. According to this hypothesis, a first FY*O mutation is likely to have occurred long after the origin of FY*B in a derived background carrying haplotype H6 that has been lost from currently sampled populations. More recently, a second FY*O mutation occurring in a less derived FY*B chromosome, here represented by haplotype H1, would have given rise to a second FY*O major branch to which haplotypes H2, H3 and H7 are connected. Alternatively, gene conversion could have occurred between a FY*O-linked haplotype H6 and a FY*B-linked haplotype H1, but the recurrence of the FY*O mutation is further supported by the finding of a recent independent GATA1 T-46C transition in a FY*A allelic background with a 2% frequency in *P. vivax* endemic region of Papua New Guinea (Zimmerman et al 1999). In any case, the high levels of (CA)n variation within haplotype H6 and the sharing of the H1 haplotype both by FY*O and non-FY*O haplotypes, indicate that the two major FY*O branches had arisen before the action of positive selection, as previously noted (Hamblin and Di Rienzo 2000). Since, under the recurrent mutation scenario, the two FY*O mutations could have had different geographical origins, it is conceivable that they could provide replication evidence for selection driven independent increase in FY*O allele frequencies. Further studies of FY haplotypes and (CA)n variation in an extended panel of African populations, including those with remnant FY*A and FY*B alleles, will be necessary to identify the major paths of spread of FY*O and to confirm this hypothesis.

We have also attempted to estimate an upper limit to the date of fixation of FY*O by using the (CA)n variation to infer the age of FY*O-linked

H2 and H3 derived haplotypes, which were found to be shared with non-FY*O alleles (Figure 1B). The age of the most recent common ancestor of each haplotype was approximated by simulating the overtime decay in the frequency of the microsatellite allele originally associated with each lineage under the stepwise-mutation-model in a population of infinite size (Seixas et al 2001). Assuming a 0.001 mutation rate at the microsatellite locus (Weber and Wong 1993), a rough calculation of the time necessary for the ancestral (CA)14 allele to reach its current 65% frequency within the 22 FY*O-linked H3 haplotypes was estimated at 490 generations. A minimum 170-1060 generations support interval was calculated as ±2 x the SD of the binomial distribution with parameters n=22 and p=0.65 (Goldstein et al 1999). If a generation time of 30 years is assumed (Tremblay and Vézina 2000), our calculations would imply that non-FY*O alleles were still not replaced by FY*O as late as 14700 (5100-31800) years ago, in West Africa. Similar calculations using the (CA)n variation within the FY*O-linked haplotype H2 led to a 11100 years estimate, but since only 7 FY*O chromosomes were found to have this lineage the estimation is associated with a very wide uninformative interval (0-44000 years). A more recent coalescent time of 9300 years (4350-15750) was calculated for the less diverse haplotype H7, which is exclusive to FY*O and might have arisen after the fixation of this allele. Under our set of assumptions, these estimates point to a more recent date for replacement of FY*A and FY*B alleles in Africa than a previous 33000 years (6500-97200) calculation based on single nucleotide polymorphisms (Hamblin and Di Rienzo, 2000). While absolute age estimates are strongly dominated by uncertainties about relevant parameters such as mutation rates, we note that our calculations place the fixation date of FY*O closer to the origins of agriculture and to the concomitant spreading of malaria as a generalized selective pressure (Livingstone, 1984). This would imply that the FY polymorphism may have become subject to malarial selection only shortly before known *P. falciparum* protective mutations (Tishkoff et al. 2001; Currat et al. 2002) and that *P. vivax* might have been indeed the selective agent that promoted FY*O fixation.

Supplementary Material

The GenBank accession numbers of the D1S2635 microsatellite sequences referred to are as follows: AF515840 (included in haplotypes H1, H2 and H6); AF515841 (included in haplotype H3); AF515842 (included in haplotypes H4 and H5); AF515843 (included in haplotype H7); AF515844 (Chimpanzee); AF515845 (Gorilla).

Acknowledgements

We thank the encouragement and suggestions of Drs. Sarah Tishkoff and James Harris. This work was partially supported by POCTI. Field work in São Tomé was supported by Instituto de Cooperação Científica e Tecnológica Internacional (ICCTI). Susana Seixas is supported by grant BD/13885/97 from Praxis XXI.

LITERATURE CITED

Belkhir, K., P. Borsa, J. Goudet, L. Chikhi and F. Bonhomme.1998. Genetix, logiciel sous WindowsTM pour la génétique des populations. Laboratoire Génome et populations, CNRS UPR 9060, Université de Montpellier II, Montpellier, France.

Chaudhuri, A., J. Polyakova, V. Zbrzezna and A. O. Pogo. 1995. The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythoid tissues in Duffy -negative individuals. Blood **85**: 615-621.

Currat, M., G. Trabuchet, D. Rees, P. Perrin, R. M. Harding, J. B. Clegg, A. Langaney and L. Excoffier. 2002. Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β^S Senegal mutation. Am. J. Hum. Genet. **70**: 207-223.

Goldstein, D.B., D. E. Reich, N. Breidman, S. Usher, U. Seligsohn and H. Peretz.1999. Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am. J. Hum. Genet. **64**: 1071-1075.

Hadley, T. J. and S.C. Peiper. 1997. From malaria to chemokine receptor: the emerging physiologic role of the Duffy Blood group antigen . Blood **89**: 3077-3091.

Hamblin, M.T. and A. Di Rienzo.2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66**: 1669-1679.

Hamblin, M.T., E. E. Thompson and A. Di Rienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70**: 369-383.

Li, J., S. S. Iwamoto, N. Sugimoto, H. Okuda and E. Kajii.1997. Dinucleotide repeat in the 3' flanking region provides a clue to the molecular evolution of the Duffy gene. Hum. Genet. **99**: 573-577.

Livingstone, F.B. 1984. The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum. Biol. **56**: 413-425.

Miller, L.H., S.J. Mason, D. F. Clyde and M. H. McGiniss.1976. The resistance factor to *Plasmodium vivax* in blacks: the Duffy-blood-group genotype, FyFy. New Engl. J. Med. **295**: 302-304.

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

Schneider, S., J-M Kueffer, D. Roessli and L. Excoffier.1997. Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and

Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland.

Seixas S., O. Garcia, M.J. Trovoada, M.T. Santos, A. Amorim and J Rocha. 2001. Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the α1-antitrypsin polymorphism. Hum. Genet. **108**: 20-30.

Tishkoff, S., R. Varkonyi, N. Cahinhinan et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293**: 455-462.

Tomás, G., L. Seco, S. Seixas, P. Faustino, J. Lavinha and J. Rocha. 2002. The peopling of São Tomé: origins of slave settlers and admixture with the Portuguese. Hum. Biol. **74**:397-411..

Tournamille, C., Y. Colin, J. P. Cartron and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat. Genet. **10**: 224-228.

Tremblay M. and H. Vézina. 2000. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am. J. Hum. Genet. **66**: 651-658.

Weber, J.L and C.Wong. 1993. Mutation of short tandem repeats. Hum Mol Genet **2**: 1123-1128.

Zimmerman, P.A., I. Woolley, G.L. Masinde et al. (11 co-authors). 1999. Emergence of FY*A^{null} in a *Plasmodium vivax*-endemic region of Papua New Guinea. Proc. Natl. Acad. Sci. USA. **96**: 13973-13977.

NARUYA SAITOU, reviewing editor

Accepted May 9, 2002

2.1. Comentário

2.1.1 Níveis de diversidade dos alelos FY*A, FY*B e FY*O

Um dos principais aspectos dos dados de Hamblin e Di Rienzo (2000) e Hamblin *et al.* (2002) sobre a variação nucleotídica associada ao *locus* Duffy, é a redução de diversidade que se observou nas populações africanas, onde o alelo FY*O está fixado, relativamente à única população europeia analisada. Este padrão é oposto ao que se observa na maioria dos *loci* em que não houve enviezamento amostral (ver secção 2.1.1.1 - Parte II) (Rogers e Jorde, 1996; Relethford, 1997) nos quais normalmente se regista maior diversidade genética em populações africanas (Rogers e Jorde, 1995; Relethford, 2001). Tendo em conta que, durante os processos de selecção positiva, há uma perda das linhagens associadas aos alelos que não são seleccionados e que a frequência de um alelo favorecido aumenta rapidamente sem que haja tempo para acumular variação nucleotídica nas regiões adjacentes, a menor diversidade registada nas populações africanas foi considerada o resultado da fixação selectiva do alelo FY*O (Hamblin e Di Rienzo, 2000).

Os resultados agora obtidos com o microssatélite D1S2635 não mostram, contudo, diferenças drásticas nas diversidades associadas aos alelos FY*A, FY*B e FY*O (Figura 1A do artigo 7). Uma das possíveis razões para esta discrepância, é a diferença entre a taxa de mutação pontual e a taxa de inserção/delecção dos motivos repetitivos dos microssatélites, que têm valores da ordem dos 10^{-10}-10^{-9}/bp/geração e 10^{-3}/*locus*/geração, respectivamente (Weber e Wong, 1993; Harding *et al.*, 1997; Harris e Hey, 1999). Com uma disparidade tão grande, é natural que os microssatélites recuperem muito mais rapidamente a diversidade que se perdeu durante o hipotético episódio selectivo e possam convergir para distribuições de equilíbrio em que as diferenças de variabilidade entre linhagens já não se consigam notar (Slatkin, 1995; Schlötterer e Wiehe, 1999).

A análise das relações genealógicas entre as linhagens dos diferentes alelos sugere ainda uma explicação adicional. Como já se referiu, na genealogia das linhagens associadas ao alelo FY*B há uma frequência elevada de sequências derivadas (Figura III.4), provavelmente devido à perda de haplótipos primitivos causada pela maior deriva genética nas populações euro-asiáticas (Relethford, 2001). Este predomínio de haplótipos derivados, separados por vários passos mutacionais,

aumenta a heterogeneidade das sequências comparadas, mas tende a reduzir a diversidade do microssatélite porque de cada vez que surge uma nova mutação dá-se a eliminação completa da variação alélica neste *locus* (ver secção 2.1.1.2.2 - Parte II). Nestas condições, é possível que a semelhança entre os níveis de diversidade do microssatélite nos alelos do *locus* FY também possa resultar, pelo menos em parte, da ausência de linhagens primitivas associadas a FY*B.

2.1.2 Recombinação entre linhagens associadas aos alelos FY*A, FY*B e FY*O

Se a actual distribuição de FY*A, FY*B e FY*O fosse inteiramente determinada por movimentos migratórios, seria de esperar que a mutação T-46C característica de FY*O tivesse ocorrido logo após a origem do alelo ancestral FY*B e tivesse coincidido com a separação entre as populações africanas do sul do Sahara e os restantes grupos humanos. Nestas condições, os alelos FY*B e FY*O deveriam ter uma idade semelhante e não teriam tido tempo para coexistir na mesma população ancestral, pelo que não se esperaria encontrar sinais de recombinação ou conversão génica entre os haplótipos que lhes estão associados. A evidência disponível mostra, contudo, que só muito dificilmente se poderá ter registado este cenário. Em primeiro lugar, as diferenças significativas na diversidade nucleotídica associada aos alelos FY*O e FY*B observadas por Hamblin e Di Rienzo (2000) indicam claramente que terá decorrido um intervalo de tempo considerável entre a origem dos dois alelos. Em segundo lugar, os resultados agora apresentados (artigo 7) mostram que há dois haplótipos derivados (H2 e H3) que são partilhados pelos alelos FY*A, FY*B e FY*O devido a recombinação ou conversão génica. O haplótipo H2 é definido pela mutação 70628C→T, previamente descrita por Hamblin e Di Rienzo (2000), e está presente em 7% dos alelos FY*A, 16% dos alelos FY*B e 3% dos alelos FY*O (Tabela 1, artigo 7). O haplótipo H3 é definido pela presença de um dinucleotídeo CG na extremidade 5' do motivo repetitivo (CA)n do microssatélite D1S2635 e encontra-se associado a 29% dos alelos FY*A, 19% dos alelos FY*B e 10% dos alelos FY*O (Tabela 1, artigo 7). A hipótese de que a partilha destes haplótipos se deve a recombinação ou conversão génica é reforçada pelas distribuições das frequências do elemento

repetitivo do microssatélite, nas quais as linhagens derivadas associadas a mais do que uma variante do *locus* FY têm alelos modais com número muito próximo de repetições CA (Figura 1B, artigo 7). Quando os polimorfismos das regiões flanqueantes e a variação no elemento repetitivo (CA)n são usados simultaneamente na reconstrução das relações entre as linhagens definidas no artigo 7, obtém-se uma rede de conexões que evidencia a importância da recombinação/conversão génica na genealogia dessas linhagens (Figura III.7). A maior partilha de haplótipos entre os alelos FY*A e FY*B é um reflexo da sua distribuição geográfica conjunta. A transferência de haplótipos entre FY*B e FY*O indica que estes alelos estiveram presentes na mesma população ancestral e que houve, de facto, um processo de fixação de FY*O em África com eliminação da variação alélica preexistente.

Estes resultados ilustram a utilidade da recombinação para demonstrar a existência de contactos passados entre populações ou linhagens com diferentes áreas actuais de distribuição.

2.1.3 Possível recorrência da mutação T-46C (FY*O) em África

Como já se referiu, a presença de dois haplótipos relativamente frequentes, separados por vários passos mutacionais, é uma característica das linhagens associadas

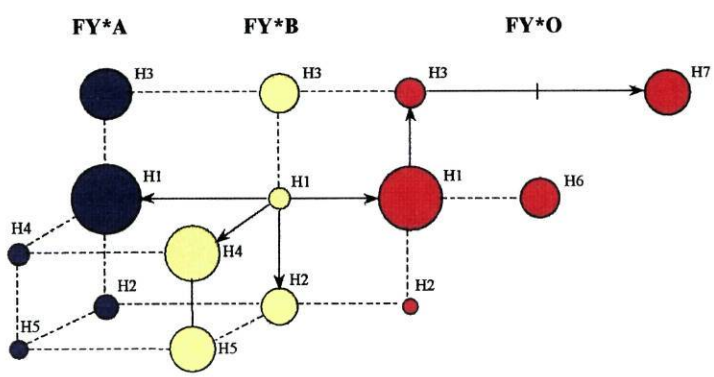


Figura III.7- Relação entre os haplótipos definidos pela variação da sequência no microssatélite D1S2635, associada aos alelos FY*A, FY*B e FY*O (ver Tabela 1 artigo 7). A área dos círculos é proporcional à frequência dos haplótipos. Os eventos mutacionais são indicados por setas e os processos de recombinação/conversão génica por linhas descontínuas. Os haplótipos ligados a FY*A, FY*B e FY*O estão a cores azul, amarelo e vermelho, respectivamente.

a FY*O que não é compatível com um modelo simplificado de fixação selectiva (Figura III.4). No artigo 7 esses dois haplótipos correspondem, por um lado, ao conjunto das linhagens H1-H3 e H7 e, por outro, à linhagem H6. Esta linhagem, aqui definida apenas pela delecção do dinucleotídeo CT nas posições flanqueantes 75336 e 75337, tem mais duas mutações (75082A→G e 75872T→C) e, segundo Hamblin e Di Rienzo (2000), poderia ter sido originada pelos seguintes processos alternativos que o estudo da diversidade nucleotídica não permitiu diferenciar: 1) ganho da mutação 75872T→C por recombinação com uma linhagem associada ao alelo FY*B, seguida de acumulação das restantes mutações na linhagem FY*O (Figura III.8 A); 2) acumulação de mutações numa linhagem FY*B, actualmente extinta, seguida de a) conversão génica (Figura III.8 B) ou b) recorrência da mutação T-46C característica de FY*O (Figura III.8 C). A análise da diversidade acumulada no elemento repetitivo (CA)n do microssatélite D1S2635 mostra que os níveis de variação mais elevados se observam na linhagem H6 e permite estabelecer uma distinção entre estas alternativas. Se a primeira hipótese se tivesse verificado, seria de esperar que, ao contrário do que se observou, a diversidade registada no elemento repetitivo (CA)n fosse menor na linhagem H6 do que na linhagem H1 associada a FY*O, devido às sucessivas eliminações da variação alélica provocadas pela acumulação de mutações (Figura III.8 A e D). Por outro lado, em caso de conversão génica entre uma linhagem H6 associada a FY*B e uma linhagem H1 associada a FY*O (Figura III.8 B), também seria de esperar que o haplótipo H1 fosse mais variável do que o haplótipo H6 que veio a ficar ligado ao alelo FY*O (Figura III.8 D). Pelo contrário, se tiver havido recorrência mutacional é perfeitamente possível que o haplótipo H6 seja o mais antigo e que, por isso, registe os níveis mais elevados de diversidade no microssatélite D1S2635 (Figura III.8 C e D).

Embora a posição -46 não esteja associada a nenhum dos principais motivos hipervariáveis discutidos na secção 2.2.2 - Parte II, a recente observação da mutação T-46C num alelo base FY*A na Nova Guiné (Zimmerman *et al.*, 1999) satisfaz dois critérios independentes de hipermutabilidade (ver tabela II.7) e apoia a hipótese de recorrência mutacional. Se esta hipótese se confirmar, poderá haver no *locus* FY uma situação semelhante à da β-globina, em que o alelo S (β6Val→Glu), associado à protecção relativamente às formas graves de malária provocadas por *P. falciparum*,

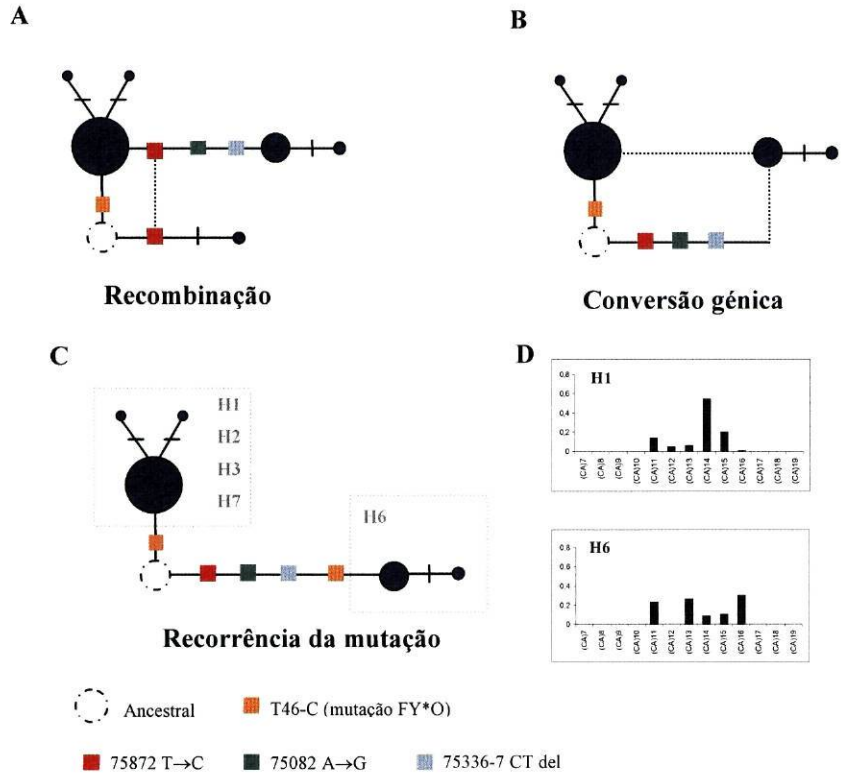


Figura III.8- Representação esquemática dos processos alternativos que podem ter estado na origem das principais linhagens de FY*O. **A**- Recombinação com uma linhagem associada a FY*B, aquisição da mutação 75872T→C e acumulação posterior de novas mutações. **B**- Conversão génica a partir de uma de linhagem de FY*B actualmente extinta. **C**- Recorrência da mutação T46-C numa linhagem de FY*B actualmente extinta. **D**- Distribuições das frequências dos alelos do motivo repetitivo (CA)n do microssatélite D1S2635 nas linhagens mais variáveis correspondentes aos dois ramos principais da genealogia de FY*O.

está ligado a cinco haplótipos muito distintos que atingem frequências elevadas em diferentes populações e resultam da provável recorrência da substituição β6Val→Glu em quatro regiões africanas e na região indo-arábica (Nagel e Ranney, 1990). Este tipo de distribuição mostra que o aumento de frequência de um alelo pode ser replicado de forma independente em regiões que partilham o mesmo tipo de pressão selectiva e é considerado uma das evidências mais importantes da influência da selecção (Flint *et al.*, 1993). Deve, no entanto, ter-se em conta que, ao contrário da β-globina, os dois haplótipos de FY*O não parecem ter uma segregação geográfica suficientemente vincada para permitir a identificação dos respectivos centros de origem, embora a

informação reunida por Hamblin e Di Rienzo (2000), sugira que a linhagem associada a H6 é especialmente menos frequente em populações da África Ocidental (Tabela III.1). Futuramente, será necessário aumentar o número de populações estudadas para conhecer com mais detalhe a distribuição dos dois haplótipos de FY*O e compreender melhor o seu significado.

2.1.4 Datação e contexto da fixação de FY*O

A questão da datação dos alelos que conferem protecção contra a malária está directamente ligada à identificação dos factores que permitiram a emergência da doença e a sua propagação. Nas mutações protectoras que têm consequências patológicas esta relação é mais evidente. Sem a contrapartida da pressão selectiva da malária, só muito dificilmente se poderiam atingir as frequências apreciáveis que caracterizam essas mutações, pelo que a estimativa do tempo necessário para que as actuais frequências sejam alcançadas pode dar uma ideia aproximada da altura em que houve um aumento de mortalidade atribuível à infecção. Por exemplo, as recentes estimativas de 6 357 anos (3 840-11 760) e 1 350 a 2 100 anos para o início do favorecimento selectivo do alelo A^- da deficiência da desidrogenase da glucose-6-fosfato e do alelo S da hemoglobina, respectivamente, indicam que, em África, a generalização da mortalidade devida a *P. falciparum* foi relativamente recente (Tishkoff *et al.*, 2001; Currat *et al.*, 2002). Estes resultados estão de acordo com o que seria de esperar com base nas características epidemiológicas da infecção, que sugerem que a dispersão de *P. falciparum* não teria sido possível sem as condições

Tabela III.1- Frequência da linhagem correspondente ao haplótipo H6 em diferentes populações africanas.

Populações[a]	Frequências
Mandinka (Gambia)	0/10 (0%)
Hausa (Camarões)	0/10 (0%)
Beti (Camarões)	3/10 (30%)
Luo (Quénia)	2/8 (25%)
Pigmeus Mbuti (Rep. Central Africana)	6/10 (60%)

[a] Dados de Hamblin e Di Rienzo (2000.)

criadas pelo estabelecimento da agricultura na floresta húmida africana, há menos de 6000 anos (Livingstone, 1984; Coluzzi, 1994; de Zulueta, 1994; Coluzzi, 1999). A elevada virulência do parasita e a sua curta sobrevivência no hospedeiro implicariam um risco importante de extinção se a sua transmissão dependesse exclusivamente de grupos dispersos de caçadores-recolectores, na sua maioria nómadas, com densidades populacionais insuficientes para sustentar uma endemia (Coluzzi, 1999).

Os recentes estudos de regiões não codificantes do genoma de *P. falciparum* também apoiam a hipótese de uma origem relativamente recente das suas estirpes contemporâneas, cujos níveis de diversidade nucleotídica são compatíveis com uma expansão rápida iniciada há apenas 3 200 a 7 700 anos a partir de um ancestral comum (Rich *et al.*, 1998; Volkman *et al.*, 2001).

A relação entre a fixação do alelo FY*O e a malária provocada por *P. vivax* é, contudo, mais complexa e menos bem estudada. Em primeiro lugar, a ausência de consequências clínicas da mutação T-46C pode levar a que, na altura do seu favorecimento selectivo, o alelo já possa ter atingido uma frequência suficientemente elevada para impedir a identificação clara do momento em que se iniciou a influência da selecção. Em segundo lugar, a virulência mais atenuada de *P. vivax* pode ter sido insuficiente para que este tenha exercido uma pressão selectiva capaz de levar à fixação de FY*O (Hill e Motulsky, 1999). Por último, e como já foi referido, a homologia com um dos parasitas de Primatas não humanos do sul da Ásia (*P. cynomolgi*) pode significar que *P. vivax* se originou nesta região e que não chegou a penetrar no continente africano durante o processo de fixação de FY*O (Livingstone, 1984).

Segundo Hamblin e Di Rienzo (2000), o limite superior da fixação de FY*O é dado pela data aproximada dos movimentos migratórios do Homem Moderno para fora de África, há cerca de 100 000 anos, pois de contrário seria difícil explicar a quase completa ausência do alelo na Eurásia e nas Américas. Assumindo que duas mutações observadas apenas uma vez em duas sequências associadas a FY*O poderiam ter ocorrido depois da fixação do alelo, estes autores estimaram em 33 000 anos (6 500-97 200) a data dessa fixação, o que dá um intervalo máximo de 67 000 anos para a extensão de um possível episódio selectivo. Apesar desta datação afastar a fixação de FY*O da introdução da agricultura e das alterações ecológicas que

favoreceram a transmissão da malária, a possibilidade de *P. vivax* ter períodos de incubação de 6 a 9 meses e uma capacidade de recidiva máxima de 3 anos podem ter favorecido a sua dispersão nas comunidades de caçadores-recolectores (Hamblin e Di Rienzo, 2000). Alternativamente, dada a baixa virulência associada a este tipo de comportamento epidemiológico (Coluzzi, 1999), foi sugerido que a subida de frequência de FY*O pode ter sido provocada por outro agente patogénico que dependesse igualmente da ligação às quimiocinas dos eritrócitos para completar o ciclo de vida (Hill e Motulsky, 1999; Hamblin e Di Rienzo, 2000). Esta situação seria análoga à observada com o receptor de quimiocinas CC-5 (CCR-5), usado pelo vírus da imunodeficiência humana HIV-1 para invadir os macrófagos. Este receptor tem uma mutação inactivadora que confere aos homozigóticos resistência à síndrome da imunodeficiência adquirida (SIDA) e é exclusiva das populações europeias, onde atinge uma frequência média próxima de 0,10 (Stephens *et al.*, 1998). Segundo Stephens *et al.* (1998), a mutação terá sido favorecida por um agente infeccioso diferente do HIV-1, há cerca de 700 anos, muito antes do início da epidemia de SIDA.

Em qualquer caso, a antiguidade da data proposta por Hamblin e Di Rienzo (2000) para a fixação de FY*O implica uma grande capacidade de difusão de um agente virulento em populações de caçadores-recolectores espalhadas numa área vasta do continente africano.

No artigo 7 procurou-se datar a fixação de FY*O através de uma abordagem baseada na variação acumulada pelo motivo repetitivo (CA)n do microssatélite D1S2635 em linhagens derivadas que fossem partilhadas com pelo menos outro alelo. Como a partilha de linhagens se deve à transferência de sequências entre alelos devido a recombinação ou conversão génica, o processo utilizado permitiu estimar um limite superior do início da fixação através do cálculo da data mais recente em que os alelos FY*O e não-FY*O teriam coexistido numa população ancestral. A datação de 14 700 anos (5 100-31 800) assim obtida indica que a fixação de FY*O em África poderá ter sido muito mais recente do que os 33 000 anos propostos por Hamblin e Di Rienzo (2000). As duas possíveis mutações independentes T-46C, que teriam originado as linhagens H1 e H6 de FY*O, são muito anteriores ao episódio de fixação, com idades estimadas em 23 850 (18 000 - 30 900) anos e 53 700 anos (38 700 - 60 000), respectivamente, tendo em conta os níveis de diversidade do microssatélite.

Teoricamente, é possível obter a quase fixação de um alelo recessivo num intervalo de 14 700 anos usando diferentes combinações de valores da frequência inicial desse alelo e do coeficiente de selecção *s* (Figura III.9). Por exemplo, se a frequência inicial for 0,05, obtém-se uma frequência maior do que 0,90 em cerca de 500 gerações com um valor de s=0,05 (Figura III.9), correspondente a cerca de 1/3 do coeficiente de selecção negativo associado ao genótipo AA no *locus* da cadeia β da hemoglobina (Bodmer e Cavalli-Sforza, 1976; Hartl e Clark, 1989). Para valores mais elevados da frequência inicial, ou no caso de os heterozigóticos também terem algum tipo de protecção (Michon *et al.*, 2001), os coeficientes de selecção poderão ser menores (Figura III.9). Embora a mortalidade devida a *P. vivax* não seja conhecida com rigor (Mendis *et al.*, 2001), é provável que actualmente não atinja valores suficientemente elevados para que aos genótipos não protegidos correspondam coeficientes de selecção necessários para promover a fixação de FY*O. No entanto, não se pode excluir a hipótese de a intensidade patogénica do parasita se ter atenuado recentemente, após um período de maior virulência no passado (Ebert, 1999; Hill e Motulsky, 1999). Além disso, alguns dados indirectos mostram que, em condições

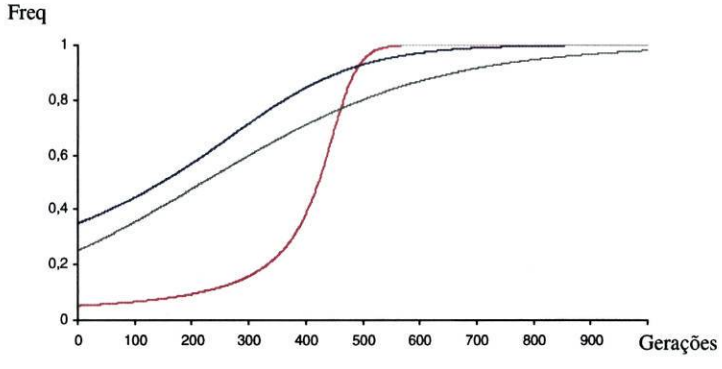


Figura III.9- Simulação da fixação do alelo FY*O com diferentes combinações de valores de frequência inicial (p_0) e de coeficiente de selecção (s). Curva vermelha: p_0 = 0,05 e s=0,05. Curva azul: p_0 = 0,35 e s=0,01. Curva verde: p_0=0,25, s_1=0,01 e s_2=0,005, em que s_1e s_2 correspondem aos coeficientes de selecção dos homozigóticos e heterozigóticos, respectivamente. Nas simulações correspondentes às curvas vermelha e azul considera-se que a selecção apenas favorece os homozigóticos. Na simulação correspondente à curva verde atribui-se aos heterozigóticos um coeficiente de selecção positivo equivalente à metade do assumido para os homozigóticos.

desfavoráveis, a mortalidade por *P. vivax* pode ser importante. Por exemplo, calcula-se que o risco de morte em doentes sifilíticos submetidos a malarioterapia por infecção com *P. vivax* alcançasse os 5% (O'Leary, 1927) e há descrições que indicam que o parasita pode ter causado mortalidades elevadas em Inglaterra no século XIX (James, 1929; Zimmerman *et al.*, 1999). Nestas condições, o argumento da relativa benignidade actual de *P. vivax* face a *P. falciparum* pode ser insuficiente para não o considerar o agente da fixação de FY*O.

Ao contrário da datação proposta por Hamblin e Di Rienzo (2000), a marcação de um limite superior do início da fixação há 14 700 anos implica que pelo menos parte do episódio selectivo pode ter decorrido durante os períodos de desenvolvimento e expansão da agricultura em África. Esta aproximação à agricultura permite compreender melhor a actual distribuição de FY*O, uma vez que sem uma ligação aos movimentos populacionais provocados pela transição demográfica e ecológica associada à produção de alimentos seria difícil que o alelo atingisse frequências elevadas numa zona tão extensa do continente africano.

Os resultados agora obtidos permitem imaginar um cenário que combina a informação genética, arqueológica e linguística num modelo explicativo que pode vir a ser testado posteriormente. Segundo este modelo, a criação de condições para a sustentação de endemias de malária terá começado no período do Paleolítico africano tardio que precedeu a transição para a agricultura. Neste período, iniciado há cerca de 12 000 a 10 000 anos, houve alterações climáticas e demográficas caracterizadas por um aumento da pluviosidade e pela progressiva sedentarização de comunidades mais numerosas, em especial nas proximidades dos cursos de água e lagos interiores (Iliffe, 1995; Curtin *et al.*, 1995; Tishkoff *et al.*, 2001). É possível que também tenham começado então a ser criadas as condições para a reprodução e o desenvolvimento de estirpes mais antropofílicas do vector *Anopheles gambiae* que aumentaram a eficácia da transmissão da infecção (Coluzzi, 1999; Tishkoff *et al.*, 2001). Estas condições poderão ter facilitado o início da pressão selectiva provocada por *P. vivax* e outros agentes patogénicos, que foi posteriormente amplificada pela introdução da agricultura florestal na África Ocidental, há cerca de 5 000 a 6 000 anos (Rushdi e Faure, 1986), originando a subida da frequência de FY*O nas regiões onde se falam as línguas não-bantus do grupo níger-congolês (Figura III.10). Mais recentemente, há cerca de 2 000

-3 000 anos, os movimentos populacionais associados à expansão Bantu e à difusão da agricultura tropical a partir das margens do rio Benué terão facilitado a dispersão da malária e do alelo FY*O e causado a sua fixação nas regiões ao sul do equador, assim se explicando a sobreposição notável entre as zonas onde a frequência deste alelo é mais elevada e a área ocupada pelas línguas níger-congolesas (Figura III.10). Curiosamente, a distribuição dos alelos S da β-globina ligados a diferentes haplótipos, mostra um padrão semelhante para a malária provocada por *P. falciparum* (Figura III.11). Neste caso, a fase inicial de infestação na África Ocidental é sugerida pela elevação da frequência da mutação protectora pelo menos três vezes de forma independente nas regiões centradas no Senegal, Benim e Congo-Angola (Figura III.11). A fase de expansão através da área situada ao sul do equador é documentada pela distribuição do hapótipo "Bantu" (Figura III.11) (Nagel e Ranney, 1990). A baixa diversidade observada nos diferentes haplótipos ligados ao alelo S, a que correspondem datações máximas de 2 000 a 3 000 anos (Currat *et al.*, 2002), indica, no entanto, que este processo foi mais recente e mais rápido do que o do *locus* FY.

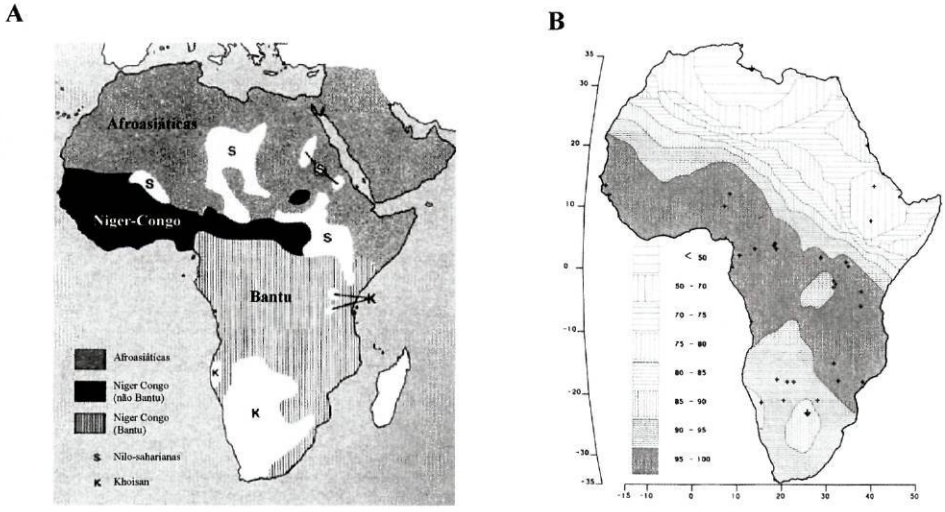


Figura III.10- Distribuição dos principais grupos linguísticos (**A**) e da frequência do alelo FY*O (**B**) no continente africano (imagem modificada de Diamond, 1999 e imagem de Cavalli-Sforza *et al.*, 1994).

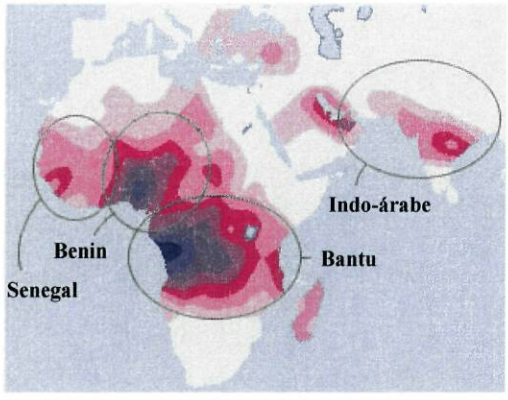


Figura III.11- Distribuição dos principais haplótipos associados ao alelo S da β-globina.

Embora o modelo agora descrito esteja assente na noção de que foi *P. vivax* que conduziu à fixação do alelo FY*O, é possível que a mesma sequência de acontecimentos tenha sido desencadeada por outro tipo de agente infeccioso, em relação ao qual a ausência dos receptores de quimiocinas do grupo Duffy possa ter conferido protecção. No entanto, a correlação com a distribuição dos alelos S da β-globina, embora desfasada no tempo pelo menos nos momentos iniciais, reforça a hipótese de que houve uma série de acontecimentos associados ao fim do Paleolítico na África tropical que proporcionou a dispersão de diferentes parasitas da malária e que pode ter começado por favorecer *P. vivax*. Esta hipótese também é apoiada pela observação da espécie *P. schwetzi*, semelhante a *P. vivax*, em chimpanzés e gorilas da Serra Leoa ao Congo, o que sugere a possibilidade de uma origem africana do parasita humano e contraria a ideia de uma dispersão centrada na Ásia a partir de *P. cynomolgi* (Livingstone, 1984).

3. Referências Bibliográficas

Bodmer, W. F., Cavalli-Sforza, L. L. (1976). *Genetics, Evolution and Man*, W. H. Freeman and Company, San Francisco.

Cavalli-Sforza, L. L., Menozzi, P. , Piazza, A. (1994). *The history and geography of human genes.* Princeton University Press, Princeton, New Jersey.

Chaudhuri, A., Polyakova, J., Zbrzezna, V., Pogo, A. O. (1995). The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. *Blood* **85**:615-621.

Coluzzi, M. (1994). Malaria and the Afrotropical ecosystems: impact of man-made environmental changes. *Parassitologia* **36**:223-227.

Coluzzi, M. (1999). The clay feet of the malaria giant and its African roots: hypotheses and inferences about origin, spread and control of Plasmodium falciparum. *Parassitologia* **41**:277-283.

Cooke, G. S. Hill, A. V. (2001). Genetics of susceptibility to human infectious disease. *Nat Rev Genet* **2**:967-977.

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney, A., Excoffier, L. (2002). Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *Am J Hum Genet* **70**:207-223.

Curtin, P., Feierman, S, Thompson, L., Vasina, J (1995). *African History: From Earliest Times to Independence,* Longman, London, New York.

de Zulueta, J. (1994). Malaria and ecosystems: from prehistory to posteradication. *Parassitologia* **36**:7-15.

Diamond, J. (1999). *Guns, Germs, and Steel. The Fates of Human Societies*. W. W. Norton and Company, New York.

Ebert, D. (1999). The evolution and expression of parasite virulence. In: Stearns, S. C. (ed). *Evolution in Health and Disease.* Oxford University Press, Oxford, pp. 161-172.

Fay, J. C., Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405-1413.

Flint, J., Harding, R. M., Clegg, J. B., Boyce, A. J. (1993). Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. *Hum Genet* **91**:91-117.

Hadley, T. J., Peiper, S. C. (1997). From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. *Blood* **89**:3077-3091.

Hamblin, M. T., Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* **66**:1669-1679.

Hamblin, M. T., Thompson, E. E., Di Rienzo, A. (2002). Complex signatures of natural selection at the duffy blood group locus. *Am J Hum Genet* **70**:369-383.

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.

Harris, E. E., Hey, J. (1999). X chromosome evidence for ancient human histories. *Proc Natl Acad Sci U S A* **96**:3320-3324.

Hartl, D. L., Clark, A. G. (1989). *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts.

Hill, A. V. S, Motulsky, A. G. (1999). Genetic variation and human disease: the role of natural selection. In: Stearns, S. C. (ed). *Evolution in Health and Disease.* Oxford University Press, Oxford, pp. 50-61.

Iliffe, J. (1995). *Os africanos: história de um continente*. Terramar, Lisboa.

James, S. P. (1929). The disappearance of malaria from England. *Proceedings of the Royal Society for Medicine* **23**:71-85.

Jarolim, P., Palek, J., Amato, D., Hassan, K., Sapak, P., Nurse, G. T., Rubin, H. L., Zhai, S., Sahr, K. E., Liu, S. C. (1991). Deletion in erythrocyte band 3 gene in malaria-resistant Southeast Asian ovalocytosis. *Proc Natl Acad Sci U S A* **88**:11022-11026.

Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* **1**:539-559.

Livingstone, F. B. (1984). The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. *Hum Biol* **56**:413-425.

Mendis, K., Sina, B. J., Marchesini, P., Carter, R. (2001). The neglected burden of Plasmodium vivax malaria. *Am J Trop Med Hyg* **64**:97-106.

Michon, P., Woolley, I., Wood, E. M., Kastens, W., Zimmerman, P. A., Adams, J. H. (2001). Duffy-null promoter heterozygosity reduces DARC expression and abrogates adhesion of the P. vivax ligand required for blood-stage infection. *FEBS Lett* **495**:111-114.

Miller, L. H., Mason, S. J., Clyde, D. F., McGinniss, M. H. (1976). The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* **295**:302-304.

Nagel, R. L., Ranney, H. M. (1990). Genetic epidemiology of structural mutations of the beta-globin gene. *Semin Hematol* **27**:342-359.

O'Leary, P. A. (1927). Treatment of neurosyphilis by malaria: report on the three years' observation of the first one hundred patients treated. *JAMA* **89**:95-100.

Relethford, J. H. (1997). Mutation rate and excess African heterozygosity. *Hum Biol* **69**:785-792.

Relethford, J. H. (2001). *Genetics and the Search for Modern Human Origins*. Wiley-Liss, New York, Chichester, Weinheim, Brisbane, Singapore, Toronto.

Rich, S. M., Licht, M. C., Hudson, R. R., Ayala, F. J. (1998). Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of Plasmodium falciparum. *Proc Natl Acad Sci U S A* **95**:4425-4430.

Rogers, A. R., Jorde, L. B. (1995). Genetic evidence on modern human origins. *Hum Biol* **67**:1-36.

Rogers, A. R., Jorde, L. B. (1996). Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* **58** :1033-1041.

Rushdi, S., Faure, H (1986). Le cadre chronologique des phases pluviales et glaciaires de l'Afrique. In: Ki-Zerbo, J. (ed) *Histoire générale de l'Afrique. I - Méthodologie et préhistoire africaine.* UNESCO/ Edicef/Présence Africaine, Paris, pp. 204-229.

Schlötterer, C., Wiehe, T. (1999). Microsatellites, a neutral marker to infer selective sweeps. In: Goldstein, D. B., Schlötterer, C. (eds). *Microsatellites. Evolution and aplications*, Oxford University Press, Oxford, pp. 238-248.

Slatkin, M. (1995). Hitchhiking and associative overdominance at a microsatellite locus. *Mol Biol Evol* **12**:473-480.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., Winkler, C., Huttley, G. A., Allikmets, R., Schriml, L., Gerrard, B., Malasky, M., Ramos, M. D., Morlot, S., Tzetis, M., Oddoux, C., di Giovine, F. S., Nasioulas, G., Chandler, D., Aseev, M., Hanson, M., Kalaydjieva, L., Glavac, D., Gasparini, P., Dean, M. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet* **62**:1507-1515.

Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M., Clark, A. G. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**:455-462.

Tournamille, C., Colin, Y., Cartron, J. P., Le Van, K. C. (1995a). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**:224-228.

Tournamille, C., Le Van, K. C., Gane, P., Cartron, J. P., Colin, Y. (1995b). Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism. *Hum Genet* **95**:407-410.

Volkman, S. K., Barry, A. E., Lyons, E. J., Nielsen, K. M., Thomas, S. M., Choi, M., Thakore, S. S., Day, K. P., Wirth, D. F., Hartl, D. L. (2001). Recent origin of Plasmodium falciparum from a single progenitor. *Science* **293**:482-484.

Weber, J. L., Wong, C. (1993). Mutation of human short tandem repeats. *Hum Mol Genet* **2**:1123-1128.

Zimmerman, P. A., Woolley, I., Masinde, G. L., Miller, S. M., McNamara, D. T., Hazlett, F., Mgone, C. S., Alpers, M. P., Genton, B., Boatin, B. A., and Kazura, J. W. (1999). Emergence of FY*A(null) in a Plasmodium vivax-endemic region of Papua New Guinea. *Proc Natl Acad Sci U S A* **96**:13973-13977.

CONCLUSÕES

1. Apolipoproteína E (APOE)

As distribuições das frequências dos alelos comuns da APOE (*2, *3 e *4) em Portugal e no País Basco enquadram-se no intervalo de variação registado na Europa e confirmam a tendência para o declínio gradual das frequências de APOE*4 no sul do continente. Na ilha de São Tomé, as frequências alélicas são significativamente diferentes das observadas nas populações ibéricas e incluem-se na gama de valores encontrada em África, onde, apesar da diversidade interpopulacional, as frequências de APOE*4 são em média mais elevadas do que na Europa. No entanto, a distribuição de frequências não é homogénea e observaram-se diferenças significativas entre as amostras provenientes de localidades situadas no norte e no sul da ilha. Esta diferenciação indica que o isolamento das comunidades angolares no sul de São Tomé, postulado com base em evidências entnolinguísticas, pode ter provocado uma subestruturação genética acentuada que deverá ser avaliada com um maior número de marcadores.

A análise da diversidade haplotípica associada ao *locus* da APOE permitiu verificar que os níveis de desequilíbrio gamético das amostras de São Tomé são significativamente mais baixos do que os da população portuguesa. Este resultado está de acordo com a tendência observada nas populações africanas para outras regiões do genoma e pode ter sido provocado pelo estrangulamento de efectivo populacional que, segundo o modelo unirregional de origem do Homem Moderno, acompanhou a dispersão das populações humanas para fora de África. Do ponto de vista epidemiológico, os níveis mais baixos de desequilíbrio gamético podem implicar uma menor capacidade de detectar associações entre a variação alélica da APOE e diferentes patologias nas populações africanas.

A comparação da diversidade haplotípica associada aos alelos da APOE demonstrou que, tanto em São Tomé como em Portugal, os níveis mais elevados de desequilíbrio gamético são observados no alelo APOE*2, enquanto os níveis mais baixos estão associados a APOE*4. Este resultado indica que APOE*4 é o alelo mais antigo e que a diversidade neste *locus* foi gerada pela ordem 4→3→2. A datação

absoluta destes alelos, feita com base na intensidade do desequilíbrio gamético nas amostras africanas, conduziu a estimativas de 109 200 anos, 575 100 anos e mais de 1 300 000 anos, para APOE*2, APOE*3 e APOE*4, respectivamente. Estes resultados indicam que a diversidade alélica da APOE é anterior à divergência entre populações humanas ocorrida há cerca de 100 000 anos, segundo o modelo unirregional. Por outro lado, a idade dos alelos derivados APOE*2 e APOE*3 parece ser suficientemente antiga para que tenham podido alcançar as suas actuais frequências médias em condições de neutralidade selectiva.

2. α1-antitripsina (PI)

2.1 Diversidade genética e história natural do polimorfismo

As distribuições das frequências dos alelos comuns da α1-antitripsina em Portugal, País Basco, São Tomé e numa amostra de ameríndios Quechua confirmam que, neste *locus*, há uma divergência acentuada entre populações de diferentes continentes. Estes resultados, juntamente com os de outras populações, permitem reconstruir padrões de afinidade semelhantes aos obtidos com a análise conjunta de outros polimorfismos e mostram que o *locus* da PI capta algumas das características mais evidentes da história evolutiva das populações humanas.

O estudo dos padrões de diversidade haplotípica revelou que há diferenças substanciais nas distribuições das frequências dos microssatélites ligados ao *locus* da α1-antitripsina, quer a nível interpopulacional, quer a nível interalélico. A análise destas distribuições permitiu testar e demonstrar a vantagem da utilização de microssatélites no registo da diversidade acumulada pelas mutações pontuais dos alelos da α1-antitripsina durante o seu percurso evolutivo. Por outro lado, o facto de, com esta abordagem, se poderem estudar distribuições de microssatélites ancoradas em linhagens com diferentes idades criou uma oportunidade para analisar alguns aspectos da dinâmica evolutiva deste tipo de marcadores.

A comparação das distribuições dos microssatélites observadas em Portugal e no País Basco ilustra a influência da história populacional na variação haplotípica associada aos alelos da PI, tendo-se verificado que a população Basca regista níveis de diversidade mais baixos, de acordo com o que seria de esperar do seu maior isolamento e exposição aos efeitos da deriva genética. Os padrões observados nos Quechua dos Andes centrais peruanos também se caracterizam por uma redução significativa da diversidade dos microssatélites e enquadram-se na tendência para a diminuição da variabilidade genética nos ameríndios registada com outros *loci*. Neste caso, as distribuições das frequências do microssatélite PI(TG)n sugerem que poderá ter havido um estrangulamento de efectivo suficientemente intenso para eliminar completamente a variação preexistente. Assumindo que a variação actualmente observada foi recuperada a partir da expansão de uma população fundadora, calculou-se que o povoamento da região andina de onde provém a amostra Quechua terá ocorrido há cerca de 9 375 anos, em concordância com a evidência paleo-ecológica disponível e com datações baseadas noutros marcadores.

Na população de São Tomé, apesar de se terem registado os níveis mais elevados de heterozigotia no microssatélite PCI(TG)n localizado a maior distância do gene da PI, verificou-se uma redução acentuada da diversidade do *locus* PI(TG)n na linhagem definida pelo alelo ancestral M1Ala213. Esta redução está em desacordo com a história do povoamento da ilha, em que a colonização com escravos de diferentes proveniências permitiu reter os níveis elevados de diversidade observados na generalidade das populações africanas. Uma das causas possíveis para esta dissociação é o favorecimento selectivo de uma mutação localizada numa região adjacente ao microssatélite, que pode incluir ou não o gene da PI. Alternativamente, a redução de diversidade pode ter sido originada por enviezamentos nos processos mutacionais do microssatélite que tenham levado à convergência para uma distribuição de estacionária. A pesquisa adicional de variação de sequência na região do gene da PI poderá contribuir para esclarecer qual é a hipótese mais plausível.

A comparação das distribuições dos microssatélites associadas aos diferentes alelos da α1-antitripsina demonstrou que no *locus* PI(TG)n, situado a apenas 7,1 kb do gene

da PI, as diferenças observadas correspondem, em grande parte, a distintos estados de recuperação mutacional de diversidade, que reflectem a antiguidade relativa das mutações características dos vários produtos génicos. No microssatélite PCI(TG)n, localizado a uma distância muito maior do gene da PI (200 kb), observaram-se níveis de diferenciação interalélica consideravelmente mais baixos que já não permitiram discriminar a antiguidade relativa das mutações, a não ser no caso dos alelos mais recentes PI*S e PI*Z. Esta discrepância entre o conteúdo informativo de microssatélites situados a diferentes distâncias do *locus* PI mostra bem como a recombinação pode levar rapidamente à perda dos sinais de diferenciação filogenética resultantes da acumulação de eventos mutacionais.

Com base nos métodos desenvolvidos para obter datações absolutas a partir das distribuições de microssatélites, foi possível estimar em 3 630 - 7 440 anos e 8 370 - 16 335 anos as idades das mutações que originaram os alelos PI*Z e PI*S, respectivamente. Ambas as estimativas são compatíveis com o confinamento dos alelos S e Z ao continente europeu, embora a origem e dispersão de cada um tenha estado associada a movimentos demográficos diferentes. No caso de PI*Z, a dispersão terá ocorrido durante a expansão populacional associada ao período pós-Neolítico. O facto de não se registar maior diversidade haplotípica em amostras do norte da Europa, indica que a maior frequência de PI*Z na Escandinávia se terá devido a efeitos da deriva genética e não à origem do alelo nessa região. No caso de PI*S, é possível que a mutação correspondente tenha tido origem na Península Ibérica, na transição do final do Paleolítico para o Mesolítico, estando a sua difusão associada a prováveis movimentos de reexpansão populacional a partir de refúgios glaciares localizados na região peninsular.

2.2 Espectro mutacional

A sequenciação de DNA de 36 variantes raras da α1-antitripsina, permitiu identificar 14 mutações distribuídas por 15 alelos diferentes nas populações de São Tomé, País Basco e Portugal. Apesar da sua relativa raridade, estes alelos, quando considerados colectivamente, são suficientemente comuns para pôr em causa os métodos de diagnóstico de deficiência que se baseiam apenas na detecção dos alelos S

e Z. A experiência adquirida, indica que, em caso de suspeita de défice de α1-antitripsina, continua a ser preferível iniciar o diagnóstico com métodos de focagem isoeléctrica e, sempre que se justifique, pesquisar por PCR-RFLP pelo menos as mutações Pro369Leu e Pro369Ser.

Das 14 mutações identificadas, 5 foram pela primeira vez observadas no decurso deste trabalho: IVS1C+1G→A, Ser47Arg, Arg281del, Pro362His e Pro369Ser. Estas mutações ilustram a utilidade do estudo dos espectros mutacionais para a localização das regiões funcionalmente mais importantes da proteína e para o conhecimento dos mecanismos que provocam a sua deficiência. As mutações Ser47Arg, Arg281del e Pro362His não estão associadas a qualquer défice da proteína circulante. A mutação IVS1C+1G→A, que causa a ausência total de α1-antitripsina no plasma, é o terceiro exemplo conhecido de uma alteração do processamento do mRNA no gene da PI e revela as consequências patogénicas da restrição das alternativas desse processamento durante a síntese da proteína nos hepatócitos. A substituição Pro369Ser mostra que há alterações no resíduo 369 que podem causar uma importante acumulação intrahepática da proteína, com formação de grânulos visíveis, à semelhança do que acontece com as mutações Phe52del, Ser53Phe e Glu342Lys, previamente descritas.

A análise haplotípica dos alelos identificados por sequenciação permitiu verificar que a maioria das mutações raras são relativamente homogéneas, o que indica que a sua ocorrência em diferentes populações pode ter resultado de simples difusão por fluxo génico. No entanto, a diversidade haplotípica do alelo PI*I, sugere a possibilidade de a mutação Arg39Cys ter ocorrido pelo menos duas vezes de forma independente. O mesmo se passa com a mutação Leu353framStop376 da variante Q0ourém que foi encontrada num alelo base diferente do anteriormente observado. A caracterização dos haplótipos associados a PI*M4 e PI*T forneceu evidência sugestiva de que estes alelos terão sido originados por recombinação intragénica.

A revisão das propriedades gerais do espectro mutacional demostrou que, à semelhança de outros *loci*, as mutações no gene da PI não se distribuem ao acaso e tendem a ocorrer preferencialmente em motivos hipermutáveis. As inserções e

delecções concentram-se na vizinhança do motivo TG(A/G)(A/G)(G/T)(A/C), ou tendem a ocorrer em regiões repetitivas que favorecem desalinhamentos durante a replicação do DNA. As substituições nucleotídicas dão-se preferencialmente no motivo CpG e originam um excesso de transições e a subrepresentação deste dinucleótido.

O estudo das relações filogenéticas entre alelos de sequência conhecida levou à identificação de nove posições homoplásicas, correspondentes aos codões 51/52, 101, 115, 148, 256, 342, 353, 362 e 369, que de acordo com um conjunto de critérios agora definido podem constituir pontos especialmente hipermutáveis do gene da α1-antitripsina.

A análise dos padrões de conservação dos aminoácidos da PI em sequências ortólogas de diferentes espécies, confirmou que há uma clara associação entre as consequências de uma mutação e o grau de conservação evolutiva dos resíduos em que ocorre. Nas comparações com sequências parálogas, os níveis de conservação observados foram mais baixos, embora tenha permanecido a tendência para as mutações patogénicas estarem sobrerepresentadas nas regiões mais conservadas das proteínas da família SERPIN.

3. Grupo sanguíneo Duffy (FY)

O estudo da variação genética acumulada pelos alelos FY*A, FY*B e FY*O no microssatélite D1S2635, situado a cerca de 4 kb do *locus* do grupo sanguíneo Duffy, permitiu combinar um sistema genético hipervariável com as linhagens mais estáveis definidas pelas mutações pontuais que originaram esses alelos, tendo-se obtido informação adicional que não poderia ter sido directamente inferida com o estudo da variação das sequências nucleotídicas.

Através da sequenciação das regiões flanqueantes do microssatélite, definiram-se quatro linhagens que, juntamente com duas posições polimórficas adjacentes, constituem um total de sete haplótipos localizados na região 5' do gene FY. A

verificação de que há haplótipos derivados que são partilhados pelos alelos FY*A e FY*B da população portuguesa e por alelos FY*O da população de São Tomé indica que os três alelos coexistiram numa população ancestral e que a sua actual distribuição resultou da fixação de FY*O em África, com eliminação de FY*A e FY*B. Estes resultados ilustram a utilidade da recombinação para demonstrar a existência de contactos passados entre populações ou linhagens com diferentes áreas actuais de distribuição.

Ao contrário de estudos anteriores, baseados na análise da variação de sequência, não se observaram diferenças drásticas entre os níveis de diversidade acumulada pelo motivo repetitivo (CA)n do microssatélite D1S2635 nos três alelos do grupo FY. Este resultado pode-se dever aos efeitos combinados da elevada taxa de inserção/delecção dos motivos repetitivos dos microssatélites e da perda das linhagens primitivas associadas aos alelos FY*A e FY*B actualmente presentes na Europa.

A observação de que os níveis de variação do motivo repetitivo (CA)n nas duas principais linhagens associadas ao alelo FY*O são significativamente diferentes favorece a hipótese de que estas linhagens terão resultado da recorrência da mutação T-46C, característica de FY*O, em dois alelos base FY*B distintos. A confirmar-se esta hipótese, o aumento da frequência de FY*O poderá ter ocorrido pelo menos duas vezes de forma independente, o que constitui evidência adicional para a acção da selecção na fixação deste alelo. O estudo mais aprofundado da distribuição das duas linhagens de FY*O em várias regiões de África poderá contribuir para a elucidação dos respectivos padrões de dispersão naquele continente.

A análise comparativa dos níveis de variação do motivo repetitivo (CA)n permitiu verificar que é possível que o processo de fixação de FY*O em África não tenha sido anterior a 14 700 anos, tendo ocorrido bastante depois da origem das duas principais linhagens associadas aquele alelo. Este limite superior é consideravelmente mais recente do que as estimativas anteriores de 33 000 anos para a data do episódio de fixação de FY*O.

Com base na datação agora proposta pode pensar-se num cenário em que o início da pressão selectiva terá estado associado às transições ecológicas e demográficas do fim do Paleolítico que teriam facilitado a transmissão da malária provocada por *Plasmodium vivax* na África Ocidental. Posteriormente, a introdução da agricultura tropical poderá ter criado condições ainda mais favoráveis de propagação da doença, contribuindo para a elevação da frequência de FY*O e para a sua dispersão pela área sub-sahariana do continente através dos movimentos populacionais associados à expansão Bantu.