Juliana Chevitarese

DETERMINAÇÃO DA ESTRUTURA GENÉTICA DAS POPULAÇÕES HUMANAS E INFERÊNCIA DOS FATORES EVOLUTIVOS QUE CONTRIBUÍRAM PARA A SUA FORMAÇÃO

Dissertação apresentada ao Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do grau de Mestre em Genética.
Orientador: Prof. Eduardo M. Tarazona Santos

Belo Horizonte, MG - Brasil
Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral
Setembro - 2009

Agradecimentos

Agradeço ao Professor Eduardo Tarazona pela orientação e confiança, apoio e incentivo. Ao Dr. Stephen J. Chanock (NCI) pela colaboração na geração dos dados analisados. Ao colega de trabalho Giordano, cuja contribuição foi essencial para o desenvolvimento desse projeto. À equipe do LDGH, pelo companheirismo.

À minha família e amigos, pelo apoio incondicional e compreensão nas horas mais difíceis. E ao Júnior, sempre presente, por toda a força e carinho.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS

ACP - Análise de Componentes Principais

CEPH – *Centre d'Etude du Polymorphisme Humain* (Centro de Estudo do Polimorfismo Humano)

CP – Componente Principal

DNA – Ácido Desoxirribonucléico

EHW – Equilíbrio de Hardy Weinberg

GPL – *General Public Licence* (Licença Pública Geral)

HGDP – *Human Genome Diversity Project* (Projeto da Diversidade do Genoma Humano)

HGP – *Human Genome Project* (Projeto Genoma Humano)

LCL – Linhagem Celular de Linfoblastos

PCR – *Polymerase chain reaction* (Reação em Cadeia da Polimerase)

SNP – *Single Nucleotide Polymorphism* (Polimorfismo de Base Única)

ANEXOS

RESUMO

A diversidade genética é moldada tanto por fatores demográficos quanto biológicos e o seu estudo traz implicações para a compreensão da história humana. Nesse trabalho, através de métodos clássicos de genética de populações - estatísticas-*F* de Wright, teste do equilíbrio de Hardy Weinberg e Análise de Componentes Principais (ACP) - investigamos os fatores evolutivos que possam ter contribuído na formação dos padrões de variabilidade genética atuais. Foram estudados dados de genotipagem de 1256 SNPs de interesse biomédico, distribuídos em genes de cromossomos autossômicos. Foram analisados 1198 indivíduos pertencentes às 52 populações do painel de amostras do *Human Genome Diversity Project* (HGDP), às quatro populações do projeto *SNP500Cancer* e a populações nativo-americanas do Peru e Equador. Devido à maior heterogeneidade presente nas amostras populacionais do *SNP500Cancer*, estas apresentaram valores de heterozigosidade esperada observados relativamente maiores do que os encontrados para as amostras populacionais do HGDP. A adição das quatro populações nativo-americanas às amostras do HGDP, para a realização das análises, permitiu a identificação de um padrão de diferenciação leste/oeste na América do Sul. Valores elevados da estatística F_{IS}, geralmente negligenciada nos estudos recentes de genética de populações, evidenciaram que o endocruzamento pode ser um fator evolutivo importante nas populações humanas, que não deveria ser ignorado. Quando analisamos grupos populacionais regionais, o valor da estatística F_{IT} foi determinado primordialmente pela diversidade inter-populacional (F_{ST}), refletindo o efeito *Wahlund*. Os gráficos gerados nas análises de ACP também permitiram a realização de inferências plausíveis acerca de eventos importantes de migração e de miscigenação das populações humanas. Assim, nosso estudo mostrou que as metodologias clássicas permanecem como importantes ferramentas para descrever a distribuição da variação genética entre e dentro das populações humanas.

ABSTRACT

Demographic and biological factors are important in shaping the human genetic diversity and their study has implications for the understanding of human history. In this work, through traditional methods of population genetics - *F*-statistics of Wright, Hardy Weinberg equilibrium test and Principal Component Analysis (PCA) - we investigated the evolutionary factors that have contributed in shaping the patterns of genetic variability observed nowadays. We studied data from 1256 autosomal SNPs of biomedical interest, in 1198 individuals belonging to 52 populations of the panel of samples of the Human Genome Diversity Project (HGDP), to the four populations of the project *SNP500Cancer* and to Native American populations of Peru and Ecuador. The greater heterogeneity present in the sample population of *SNP500Cancer* result in higher values of expected heterozygosity than those found to the population samples of HGDP. The addition of four Native American populations to samples HGDP, to perform the analysis, allowed the identification of a pattern of diferentiation east/west in South America. Large values of the statistical F_{IS}, usually neglected in recent studies of population genetics, showed that inbreeding may be an important factor of evolution in human populations, which should not be ignored. When we analyzed regional population groups, the statistical value of F_{IT} was determined primarily by inter-population diversity (F_{ST}), reflecting the Wahlund effect. The graphs generated by PCA analysis also allowed the creation of plausible inferences about important events of migration and mixing of human populations. Thus, our study showed that the classical methodologies remain important tools to describe the distribution of genetic variation within and between human populations.

I - Introdução

I - 1. Estruturação genética nas populações humanas

Com a conclusão do sequenciamento do genoma humano (*International Human Genome Sequencing Consortium*, 2004), passou a haver um interesse crescente na identificação em alta resolução da diversidade genômica individual e populacional (Chakravarti, 1999; Weiss & Clark, 2002). Para tanto, foram desenvolvidos estudos que investigam como os padrões de variação genética são criados e mantidos nas diferentes populações da espécie humana (Li *et al.* 2008; Bastos-Rodrigues *et al.* 2006; Tishkoff *et al*. 2009).

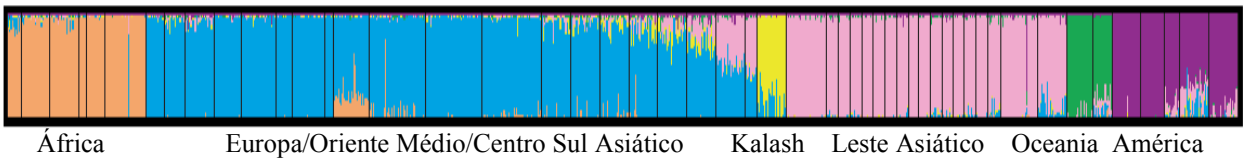
A espécie humana é sabidamente estruturada geneticamente, uma vez que apresenta diferenças na variação genética entre suas populações constituintes (Pritchard *et al.* 2000; Bastos-Rodrigues *et al.* 2006; Hammer *et al.* 2003; Rosenberg *et al.* 2002). Essa estruturação está relacionada a fatores como geografia, cultura, religião, comportamento, etnia e aparência física. Esses fatores muitas vezes contribuem para determinar afastamentos da condição de panmixia, diminuindo o fluxo gênico entre grupos e permitindo que a deriva genética, a seleção e, até mesmo, a mutação atuem levando à diferenciação genética. Entretanto, a maioria dos estudos que investigam a variação humana inicia-se pela amostragem de "populações" humanas pré-definidas com base na cultura e geografia, que poderiam não refletir as relações genéticas subjacentes (Foster & Sharp 2002).

Em um estudo recente, ao analisar esse problema, Rosenberg *et al.* (2002) investigaram a estruturação genética humana sem considerar informações prévias, relativas às populações a que pertenciam os indivíduos participantes. Nesse estudo, utilizando-se unicamente dados genéticos, foram identificados seis grupos populacionais principais, cinco dos quais são correspondentes às cinco maiores regiões geográficas do globo (África; Europa/Oriente Médio/Centro Sul Asiático; Leste Asiático; Oceania; e América), além de terem sido identificados subgrupos que, na maioria das vezes, correspondem às populações já conhecidas devido à identidade cultural, tais como as populações nativo-americanas Suruí, Karitiana, Piapoco e Curripaco, Maia e Pima (Figura 1). Dessa forma, mostrou-se que a classificação em populações pré-definidas é informativa em relação ao padrão de variação genética entre e dentro dos grupos humanos. Foi demonstrado também que em várias populações, como Uigures (China) e Hazara (Paquistão), os indivíduos apresentaram componentes genéticos de mais de um dos cinco grupos identificados, porém, presentes em

proporções similares na maioria dos indivíduos (Rosenberg *et al.* 2002). Entretanto, nesse mesmo estudo, Rosenberg *et al.* (2002) sugerem que, em populações recentemente miscigenadas, pode haver uma diferença substancial entre os componentes de ancestralidade presentes em cada um dos seus indivíduos. Isso também é aludido em Thomas & Witte (2002) que falam da dificuldade em se determinar os componentes de miscigenação étnica individual presentes em populações miscigenadas como a hispânica.

A.

K = 6



B.

K = 5

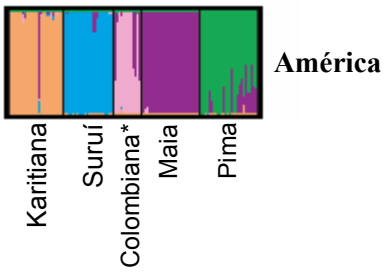


Figura 1. Estimativa da estruturação genética das populações humanas. Cada indivíduo está representado por uma linha vertical, dividida em *K* segmentos coloridos que representam as frações de componentes genéticos estimadas em *K* grupos. As linhas verticais pretas separam indivíduos de populações diferentes. **A.** Separação das populações humanas em seis grupos populacionais principais, cinco dos quais correspondem às maiores regiões geográficas mundiais e um correspondente à população paquistanesa Kalash; **B.** No continente americano, os grupos inferidos correspondem a populações nativo-americanas previamente identificadas.

* Colombiana: refere-se às populações Piapoco e Curripaco, amostradas conjuntamente.

(Figura modificada de Rosenberg *et al.* 2002)

Além disso, há situações em que os pesquisadores lidam com estruturação populacional críptica, ou seja, situações em que a estrutura populacional é difícil de detectar através de caracteres visíveis, mas pode ser relevante em termos genéticos (Pritchard *et al.* 2000). Isso é especialmente relevante em estudos de associação, em que o mapeamento é

utilizado para encontrar genes associados a doenças. Nesses casos, a presença de estruturação populacional pode levar a associações espúrias, pois se a incidência da doença investigada for maior entre indivíduos de uma das subpopulações crípticas, qualquer alelo que tenha freqüência mais alta nessa subpopulação poderá ser erroneamente associado à doença, independentemente de ser ele o alelo causal (Ewens & Spielman, 1995; Devlin & Roeder, 1999; Devlin *et al.* 2001a; Devlin *et al.* 2001b; Pritchard & Donnelly 2001). O problema de estruturação críptica também está presente no contexto do uso de DNA *fingerprinting* ("impressões digitais de DNA") na genética forense, onde é importante acessar o grau de estruturação populacional para estimar a probabilidade de se incorrer em erros nas identificações realizadas (Balding & Nichols 1994; Balding & Nichols 1995; Foreman *et al.* 1997; Roeder *et al.* 1998). Ainda, quando a estruturação populacional não é evidente, uma amostragem dessa população pode consistir em um grupo heterogêneo de sub-amostras das subpopulações presentes. Wahlund (1928) demonstrou que, quando há diferença nas freqüências alélicas entre essas subpopulações, haverá uma deficiência de heterozigotos e um excesso de homozigotos em relação ao esperado no equilíbrio de Hardy-Weinberg, mesmo se cada sub-amostra apresentar genótipos dentro das proporções de Hardy-Weinberg. Esse efeito ficou conhecido como efeito *Wahlund* e já foi reconhecido em estudos com nativo-americanos (Hedrick & Black, 1997).

Desse modo, um desafio central ao se analisar qualquer conjunto de dados genéticos é verificar se há alguma evidência de que as amostras vêm de populações estruturadas. O programa *Structure* (Pritchard *et al.* 2000; Falush *et al.* 2003; Falush *et al.* 2007), que utiliza um algoritmo para agrupar indivíduos em grupos, baseando-se em dados genotípicos de vários *loci*, tem sido aplicado na investigação desse problema, uma vez que permite identificar a estruturação populacional críptica, detectar imigrantes ou indivíduos miscigenados e inferir miscigenações populacionais históricas (Rosenberg *et al.* 2002; Bastos-Rodrigues *et al.* 2006; Auton *et al.* 2009; Wang *et al.* 2007). Entretanto, para conjunto de dados muito grandes, como os utilizados em estudos de associação em larga escala genômica, a utilização do *Structure* pode se tornar operacionalmente impraticável, devido ao intensivo trabalho computacional (Price *et al.* 2006; Patterson *et al.* 2006). Além disso, uma vez que essa metodologia adota a classificação dos indivíduos em grupos discretos, ela é bastante sensível ao número de grupos estimados, de modo que pode não refletir a história real das populações, pois: (1) ainda não está bem definido se a variação genética humana está distribuída de forma contínua, como um gradiente, ou descontínua; (2) e, neste último caso, há dificuldades em se

estimar corretamente o número de grupos a ser utilizado para cada análise (Serre & Paabo, 2004; Setakis *et al.* 2006).

Outro método bastante utilizado na investigação da estruturação das populações humanas é a análise de componentes principais (ACP), introduzida nos estudos de genética de populações pelo grupo de Cavalli-Sforza (Cavalli-Sforza & Edwards, 1964; Menozzi, 1978; Cavalli-Sforza & Menozzi, 1994). A ACP é uma abordagem especialmente útil para fins gráficos, e pode ser aplicada na análise de vários tipos de dados genéticos, como SNPs (Li *et al.* 2008; Shriver *et al.* 2005; Reich *et al.* 2008), microssatélites (Chakraborty & Jin, 1993), frequências haplotípicas (Capelli *et al.* 2006; Lovell *et al.* 2005) e dados de distribuição de polimorfismos de inserções Alu (Stoneking *et al.* 1997). Diferentemente do programa *Structure*, a ACP aplicada a estudos de genética de populações não busca primordialmente classificar cada indivíduo em grupos discretos, mas sim localizar cada indivíduo em um espaço multidimensional contínuo, no qual cada dimensão sintetiza componentes de variabilidade que estão correlacionados entre sim, sendo que cada componente tende a ser independente em relação aos demais. Além disso, a ACP pode ser realizada de forma bastante rápida, mesmo para conjuntos de dados genéticos bastante grandes (Patterson *et al.* 2006; Price *et al.* 2006).

Graficamente, a ACP pode ser descrita como a movimentação de eixos em um espaço multidimensional, de tal forma que estes eixos representem novas variáveis, chamadas componentes principais (CPs). Os CPs são combinações lineares das variáveis originais e são gerados buscando-se maximizar a correlação entre essas variáveis. Eles são ordenados decrescentemente, de acordo com a porcentagem da variância presente no conjunto de dados que representem. Normalmente, os primeiros 2-4 CPs são suficientes para representar a maior parte da variabilidade genética presente em um conjunto de dados (Cavalli-Sforza & Menozzi, 1994). A ACP é, assim, um método de redução da dimensionalidade, pois resume a informação de padrões independentes presentes em uma matriz de dados, simplificando os dados multivariados com uma perda de informação mínima e mensurável.

Quando a ACP é aplicada a uma matriz de genótipos individuais e o seu resultado é apresentado como um diagrama cartesiano, em que os dois primeiros CPs são utilizados como coordenadas e no qual cada indivíduo é representado como um ponto, indivíduos geneticamente mais similares tenderão a estarem mais próximos nesse espaço bidimensional. Assim, o diagrama dos dois primeiros CPs pode refletir como os fluxos gênicos contribuíram para o estabelecimento de padrões de variação genética similares, de modo que Cavalli-Sforza & Menozzi (1994) demonstraram que essa representação muitas vezes irá se assemelhar à

situação encontrada no mapa geográfico, pois populações próximas geograficamente tendem a ser mais similares geneticamente. Por outro lado, caso haja estruturação populacional na amostra analisada, esta estruturação também poderá ser detectada no diagrama da ACP, pois os indivíduos de cada subpopulação se apresentarão mais próximos entre si no espaço bidimensional, diferenciando-se dos indivíduos de outras subpopulações presentes na amostra.

I - 2. Medidas de variabilidade genética

O padrão de diversidade genética das populações humanas modernas é o resultado da combinação de diversos fatores evolutivos: (1) demográficos (atuam em todo o genoma), como flutuações no tamanho efetivo populacional, sub-estruturação, endogamia e fluxo gênico; (2) fatores ligados ao acaso, como a deriva genética; e (3) fatores gene-específicos, como mutações, taxas de recombinação e pressões seletivas (Tishkoff & Verrelli 2003; Jorde *et al.* 2001). Entretanto, para entender a influência desses fatores evolutivos na genética das populações, é necessário que se saiba descrever e quantificar a variação genética dentro de uma população e entre as populações.

Um dos princípios mais importantes para os estudos de variação genética foi proposto por Hardy (1908) e Weinberg (1908) e ficou conhecido como princípio Hardy-Weinberg. Esse princípio estipula que após uma geração de cruzamentos ao acaso, as freqüências genotípicas de cada *locus* podem ser representadas por uma função binomial (se o locus é bialélico) ou multinomial (se multialélico) das freqüências alélicas. Por exemplo, para um *locus* com dois alelos *A* e *a* com freqüências *p* e *q* respetivamente:

$$(p+q)^2 = p^2 + 2pq + q^2$$

Onde:

- p^2 é a frequência esperada do genótipo AA;
- $2pq$ é a frequência esperada do genótipo Aa;
- q^2 a frequência esperada do genótipo aa.

Também por esse princípio, na ausência de fatores evolutivos que alteram as freqüências alélicas - como seleção, deriva genética, fluxo gênico e mutação - e mantendo-se os cruzamentos ao acaso, as freqüências genotípicas, conhecidas como proporções de Hardy-Weinberg, não irão se alterar ao longo do tempo. Freqüentemente, refere-se às proporções de Hardy-Weinberg como freqüências do equilíbrio de Hardy-Weinberg (EHW), porém, esse é

um tipo de equilíbrio especial, pois: (1) se as proporções genotípicas são perturbadas sem que se alterem as freqüências alélicas, elas retornarão às frequências do EHW após uma geração de cruzamentos ao acaso; (2) mas se as freqüências alélicas são alteradas, os genótipos estarão presentes em uma nova configuração de proporções de Hardy-Weinberg, determinadas pelas novas freqüências alélicas.

Nos estudos de genética de populações, é uma prática comum testar se as frequências genotípicas observadas estão de acordo com as esperadas na hipótese de EHW, sendo que a expectativa de equilíbrio geralmente é atendida em estudos com populações humanas (Witherspoon *et al.* 2006; Bastos-Rodrigues *et al.* 2006; Tishkoff *et al.* 1998). Embora teoricamente a adaptação ao EHW indique que os fatores evolutivos não estejam atuando, sua interpretação mais apropriada é a de que os testes estatísticos utilizados não são suficientemente poderosos para capturar afastamentos muito pequenos do EHW. Desse modo, para capturar desvios do EHW, é necessário que se estude muitos *loci*, ou que o efeito dos fatores evolutivos seja suficientemente forte. Os desvios das freqüências genotípicas esperadas no EHW podem sugerir a ocorrência de fatores evolutivos como endogamia e estruturação populacional, ou podem apresentar uma explicação de ordem técnica (e não evolutiva) sendo decorrentes de erros de genotipagem (Wittke-Thompson *et al.* 2005; Cox & Kraft, 2006).

A quantidade de heterozigosidade ($h = 2pq$) esperada no EHW, para um dado *locus*, é uma medida de variação genética bastante utilizada na investigação da variação genética populacional (Nei & Roychoudhury, 1974; Nei, 1987). Geralmente, como os estudos são realizados com um grande número de *loci*, utiliza-se a média da heterozigosidade por *locus* (H), definida como a média de h ao longo de todos os *loci* estudados. Valores de H elevados indicam alta diversidade e são esperados em populações antigas ou recém miscigenadas. Em estudos com populações humanas: (1) os maiores valores de H são encontrados para as populações africanas (Rosenberg *et al.* 2002; Li *et al.* 2008; Jorde *et al.* 2000); (2) os menores valores são encontrados para populações que têm passado por uma série de efeitos fundador na sua história demográfica, como nativas da América e Oceania (Rosenberg *et al.* 2000; Li *et al.* 2008); e (3) populações miscigenadas da Ásia central usualmente apresentam valores altos de H (Li *et al.* 2008).

Outra abordagem clássica adotada nos estudos de genética de populações, comumente utilizada para se estimar a distribuição da variação genética em uma população subdividida, foi desenvolvida por Wright (1951; 1956) e ficou conhecida por estatísticas-*F*. Essa abordagem envolve três coeficientes F de correlação (F_{ST}, F_{IS} e F_{IT}), usados para alocar a

variação genética em três níveis: populacional (*T*), subpopulacional (*S*) e individual (*I*). Esses três coeficientes (ou mais, se houver subdivisões adicionais) se inter-relacionam de modo que F_{ST} é uma medida de diferenciação genética entre as subpopulações e é sempre positivo. F_{IS} e F_{IT} são medidas de desvio das proporções de Hardy-Weinberg dentro das subpopulações e na população total, respectivamente, onde valores positivos indicam uma deficiência de heterozigotos e valores negativos indicam um excesso de heterozigotos em relação ao esperado sob EHW. Mais especificamente: F_{ST} é a correlação entre gametas amostrados aleatoriamente de subpopulações diferentes em relação à população total (e, consequentemente, é positivo e diferente de zero quando o efeito Wahlund determina um excesso de homozigotos na população total); F_{IS} é a média, ao longo de todas as subpopulações, da correlação entre os gametas que se unem para formar os indivíduos, em relação aos gametas das subpopulações a que esses indivíduos pertencem; e F_{IT} é a correlação entre os gametas que se unem formando indivíduos, em relação àqueles da população total (Wright, 1965) e depende do efeito conjunto de F_{IS} e F_{ST}. As três estatísticas-*F* se relacionam através da seguinte fórmula:

$$(1-F_{IT})=(1-F_{IS})(1-F_{ST})$$

Desde sua criação, as estatísticas-*F* têm sido bastante utilizadas por geneticistas de populações na interpretação de dados de populações naturais (Weir & Cockerham 1984; Robertson & Hill, 1984; Long, 1985; Slatkin & Barton, 1989; Kidd *et al.* 2004; Weir et al 2005; Bastos-Rodrigues *et al.* 2006). Embora as estatísticas-*F* sejam úteis para descrever a estrutura genética das populações, os modelos ideais utilizados para inferir a ação de fatores evolutivos, a partir das estatísticas-*F*, são pouco realistas (Whitlock & McCauley, 1999). Isso é particularmente verdade em pesquisas de genética da conservação, em que as populações das espécies de interesse são pequenas ou recentemente diminuíram de tamanho, sofreram fragmentação ou, de algum outro modo, foram perturbadas demograficamente: esses são exatamente os tipos de situações demográficas que podem criar um viés nas estimativas de migração baseadas nas estatísticas-*F* (Whitlock & McCauley, 1999; Kinnison *et al.* 2002). Apesar dessas limitações, as estatísticas-*F* de Wright têm permanecido como o parâmetro padrão utilizado para descrever a distribuição primária da variação genética entre subpopulações pré-definidas (Neigel 2002; Weir & Hill 2002).

Muitos estudos já demonstraram que as diferenças entre as populações humanas representam apenas uma pequena parcela do total da variância genética encontrada, sendo comumente estimado um F_{ST} que varia de 0,10 a 0,15 para as populações humanas (Lewontin

1972; Nei & Roychoudhury, 1982; Latter, 1980; Bowcock *et al.*, 1991; Barbujani *et al.* 1997; Jorde *et al.* 2000; Watkins *et al.* 2003; Altshuler *et al.* 2005; Rosenberg *et al.* 2005). Em 1972, em um estudo clássico, Lewontin realizou uma estimativa da distribuição dos componentes de variância genética em humanos nos níveis individual, populacional e regional, através de estudos com grupos sanguíneos, variantes protéicas e isoenzimas. Lewontin (1972) obteve a seguinte distribuição: 85,4% da variância está dentro das populações; 8,3% está entre as populações e dentro dos continentes; e 6,3% está entre os continentes. Posteriormente, utilizando marcadores de DNA, outros autores obtiveram resultados similares (Barbujani *et al.* 1997; Barbujani & Di Benedetto, 2001; Excoffier & Hamilton, 2003), enquanto Rosenberg *et al.* (2002) observaram que 93-95% da variância genética está dentro das populações e que apenas 3,6% está entre as maiores regiões geográficas do globo. Esses estudos levaram ao importante corolário de que a espécie humana possui um baixo nível de estruturação geográfica que não é compatível com a existência de raças (Templeton, 1999; Bamshad *et al.* 2004).

I - 3. Origem africana do *Homo sapiens*

Em meados do século XX, estudos feitos em proteínas já indicavam o potencial de dados genéticos em fornecer informações sobre a história e geografia das populações humanas (Race & Sanger 1975; Pauling *et al.* 1949). Posteriormente, os esforços para reconstruir a origem e dispersão do homem moderno focavam em sequências de DNA obtidas de compartimentos haplóides do genoma: o DNA mitocondrial e a região não-recombinante do cromossomo Y (Stringer & Andrews, 1988; Cann *et al.* 1987; Excoffier 2002; Harpending *et al.* 1998; Hammer *et al.* 1997). Atualmente, a crescente sofisticação das ferramentas computacionais somada à abundância de dados de dois outros compartimentos genômicos - cromossomo X (Takahata *et al.* 2001; Garrigan *et al.* 2005; Hammer *et al.* 2004) e autossomos (Harding *et al.* 1997; Zhao *et al.* 2000; Yu *et al.* 2000) - estão produzindo novas perspectivas no estudo da história evolutiva humana.

Os estudos realizados até hoje, permitem que se tirem algumas conclusões gerais sobre a origem e evolução dos *Homo sapiens* (Garrigan & Hammer, 2006). Em primeiro lugar, estudos paleoantropológicos e genéticos sugerem que o homem anatomicamente moderno surgiu na África equatorial há aproximadamente 200 mil anos atrás e de lá se espalhou pelo resto do globo nos últimos 100 mil anos, substituindo completamente as outras formas arcaicas de *Homo*, enquanto avançava por todo o Velho Mundo (Cann *et al.* 1987; Excoffier

2002; Harpending & Cochran, 2002). Esse modelo de evolução proposto, conhecido como "Modelo fora da África", é freqüentemente confrontado com o modelo "Multi-regional", que sugere que os humanos anatomicamente modernos evoluíram de formas arcaicas de *Homo* em várias localidades do Velho Mundo. Por este modelo, a homogeneidade genética vista hoje seria resultado de fluxo gênico e da seleção natural (Wolpoff *et al.* 1994). Entretanto, embora alguns autores ainda o defendam (Wolpoff *et al.* 2000; Thorne & Wolpoff 2003; Templeton 1997; Hawks *et al.* 2000), as evidências genéticas claramente favorecem o "Modelo fora da África".

Dentre as evidências que sustentam a origem unicamente africana, está o excesso de diversidade genética encontrado na África em relação às demais regiões mundiais, em vários compartimentos do genoma, como o DNA mitocondrial (Vigilant *et al.* 1991), cromossomo Y (Underhill *et al.* 2000) e DNA autossômico (Rosenberg *et al.* 2002; Rosenberg *et al.* 2005; Ramachandran *et al.* 2005). Esse padrão de diversidade é compatível com um cenário no qual uma pequena fração da população africana, contendo um subconjunto da variação genética da África, deixou a África para colonizar o resto do mundo há cerca de 100 mil anos atrás. Além disso, uma segunda evidência para a origem única africana vem de estudos em que foi possível determinar o estado ancestral de uma série de polimorfismos, por exemplo, através de comparações com um grupo externo como os chimpanzés. Nesses estudos, a raiz da árvore filogenética humana usualmente recai mais próxima ou dentro das populações africanas (Underhill *et al.* 2000; Ingman *et al.* 2000; Thomson *et al.* 2000). Em terceiro lugar, os valores de desequilíbrio de ligação, que podem refletir as idades dos haplótipos dentro das populações, geralmente são mais baixos em populações africanas, quando comparadas com não-africanas, sugerindo que as africanas são populações mais antigas (Jorde *et al.* 2000; Kidd *et al.* 2000; Tishkoff *et al.* 1998; Tishkoff *et al.* 2000). Por último, os estudos sobre variação haplotípica mostram que haplótipos encontrados fora da África usualmente são subconjuntos da grande coleção de haplótipos encontrados na África (Underhill *et al.* 2000; Wooding *et al.* 2002; Tishkoff *et al.* 2000).

I - 4. Projetos de investigação da variação genética em humanos

Até recentemente, as pesquisas voltadas para a genética de populações humanas apresentaram-se, em grande parte, como um esforço fragmentado (Cavalli-Sfrorza, 2005). De fato, apenas quando o Projeto Genoma Humano (*Human Genome Project* - HGP; Watson, 1990) estava em plena atividade, foi levantada a idéia de um estudo sistemático e em larga

escala das variações do genoma humano (Cavalli-Sforza, 1990). Mais especificamente, notou-se que amostras renováveis de populações bem escolhidas ao longo de todo mundo, para as quais qualquer parte do genoma poderia ser examinada, poderiam facilitar enormemente os estudos genéticos de geografia e história da espécie humana. Nascia, assim, a idéia do *Human Genome Diversity Project* (HGDP).

O HGDP é um projeto internacional que procura entender a diversidade e a história das populações humanas modernas (Cann *et al.* 2002). Laboratórios de várias regiões do mundo têm contribuído com linhagens celulares autóctones, de diferentes populações, para uma coleção mantida pela *Foundation Jean Dausset - Centre d'Etude du Polymorphisme Humain* (CEPH), em Paris, França. Atualmente, há no CEPH um total de 1063 linhagens celulares de linfoblastos (LCLs) cultivadas, de 1050 indivíduos, distribuídos em 51 populações mundiais. Optou-se por culturas celulares por razões de exatidão e capacidade de renovação, evitando-se assim erros potenciais de troca de nucleotídeos e esgotameno das amostras, advindos da utilização de métodos *in vitro* como a PCR (Cavalli-Sforza, 2005). Todas as amostras armazenadas no CEPH seguem os critérios do comitê ético do HGDP, incluindo a coleta apenas mediante consentimento informado (*Committee on Human Genome Diversity - National Research Council*, 1997). As informações de cada LCL resumem-se ao sexo do indivíduo e à sua população e origem geográfica. O DNA dessas células está disponível para pesquisas sem fins lucrativos, essencialmente a preço de custo. As amostras do CEPH já foram dstribuídas para mais de 100 pesquisadores de todo o mundo e estudos de grande impacto foram realizados a partir dessas amostras (e.g. Rosenberg *et al.* 2002; Zhivotovsky *et al.* 2003; Gonzalez-Neira *et al.* 2004; Rosenberg *et al.* 2005; Serre & Paabo 2004; Sabeti *et al.* 2007).

Além do HGDP, outro projeto que busca caracterizar a variação genética humana é o *SNP500Cancer* (Packer *et al.* 2004). O *SNP500Cancer*, iniciativa do *National Cancer Institute - Cancer Genome Anatomy Project* (Strausberg *et al.* 2000), é especificamente voltado para gerar recursos que contribuam na identificação e caracterização de polimorfismos de base única (SNPs) e de outras formas de variação genética potencialmente importantes em estudos de epidemiologia molecular de doenças complexas como o câncer (Packer *et al.* 2004). No *SNP500Cancer*, são estudados amostras dos genomas de 102 indivíduos anônimos, obtidas dos repositórios celulares do *Coriell Institute for Medical Research* (Camden, Nova Jérsei, Estados Unidos da América). Essas amostras representam os quatro grupos étnicos mais representativos da população dos Estados Unidos, de acordo com as declarações de ascendência dos indivíduos amostrados: 24 africanos/afro-americanos, 31

caucasianos, 24 asiáticos e 23 hispânicos, sendo que estes últimos são os indivíduos de ancestralidade latino-americana, que não se auto-identificam como nativo-, afro- ou euro-americanos.

No presente estudo, utilizamos amostras desses dois projetos, HGDP e *SNP500Cancer*, além de amostras de quatro populações nativo-americanas, para calcular estimadores clássicos da diversidade inter- e intra-populacional, nas diversas populações humanas distribuídas por todo o mundo.

II - Objetivos

II - 1. Objetivo geral

Determinar a estrutura genética das populações humanas a partir de dados de SNPs em genes de interesse biomédico e discutir os fatores evolutivos que possam ter contribuído para defini-la.

II - 2. Objetivos Específicos

1. Calcular as freqüências alélicas e genotípicas para 1256 SNP das amostras de cada uma das 52 populações estudadas do HGDP, de quatro populações nativo-americanas estudadas por nosso laboratório (Cayapa, Quechua, San Martín e Matsiguenga), dos nove grupos populacionais formados a partir dessas 56 populações, e das quatro populações do Projeto *SNP500Cancer*;

2. Aplicar o teste do Equilíbrio de Hardy-Weinberg para cada um dos SNPs, dentro das amostras populacionais;

3. Estimar a variabilidade intra-populational utilizando a heterozigosidade esperada para cada SNP, em cada uma das amostras dessas populações e grupos populacionais;

4. Calcular as estatísticas-*F* (F_{ST}, F_{IS} e F_{IT}) para determinar a estrutura genética dentro e entre as populações e grupos populacionais amostrados;

5. Verificar, através da Análise de Componentes Principais (ACP), a estrutura genética das populações e grupos amostrais, estudados isolados e conjuntamente.

III - Materiais e Métodos

III - 1. Amostragem

Os dados de genotipagem utilizados correspondem a 1256 SNPs, distribuídos em 390 genes de cromossomos autossômicos, do *Cancer SNP Panel* - plataforma *Illumina GoldenGate*® (Fan *et al*. 2003; Shen *et al*. 2005; Figura 2; Anexo I). Ao todo, foram analisados 1198 indivíduos. Desses, 1029 são indivíduos autóctones de diferentes ascendências, pertencentes às 52 populações do painel de amostras do *Human Genome Diversity Project* - HGDP (Cann *et al*. 2002; Anexo II); 102 pertencem às quatro populações (Caucasiana, Hispânica, Asiática e Afro-americana) do projeto *SNP500Cancer* (Packer *et al*. 2004); e 67 são nativo-americanos do Peru e Equador, cujas amostras estão disponíveis em nosso laboratório (Figura 3). Este último conjunto de amostras é composto por: (a) 22 indivíduos Quechua dos Andes peruanos; (b) 17 indivíduos Quechua/Matsiguenga de San Martín (Peru); (c) 21 indivíduos Matsiguengas de Monte Carmelo (Peru); e (d) 7 Cayapas do Equador. As amostras (b), (c) e (d) foram coletadas na região localizada entre as montanhas andinas (oeste) e a região amazônica (leste). Todos os dados analisados foram gerados em colaboração com o Dr. Stephen J. Chanock (*National Cancer Institute* - NCI) e encontram-se no banco de dados *DIVERGENOMEdb*, a plataforma de bioinformática em desenvolvimento no nosso laboratório (Magalhães *et al*. em preparação).

Para a realização das análises, em um primeiro momento, os dados de genotipagem dos representantes dos quatro grupos étnicos do *SNP500Cancer* foram mantidos isolados. Os demais indivíduos foram amostrados de duas formas: em suas respectivas populações isoladas e em grupos populacionais formados de acordo com a origem geográfica das amostras. Para a formação desses grupos, as amostras de chineses Han do norte e sul da China, assim como em Xing *et al*. (2009) e Tishkoff *et al*. (2009), foram combinadas em uma só, enquanto os Bantus africanos do nordeste foram analisados separadamente dos Bantus do sudeste e sudoeste, do mesmo modo que em Bastos-Rodrigues *et al*. (2006) e Rosenberg *et al* (2005). Os norte-africanos Mozabites foram analisados juntamente com as populações do Oriente Médio, baseando-se em estudos que relatam a semelhança genética entre essas populações (Rosenberg *et al*. 2002; Nievergelt *et al*. 2007; Li *et al*. 2008; Tishkoff *et al*. 2009) e a suposta origem dos Mozabites no Oriente Médio (Foster & Romano, 2007; Hellenthal *et al*. 2008).

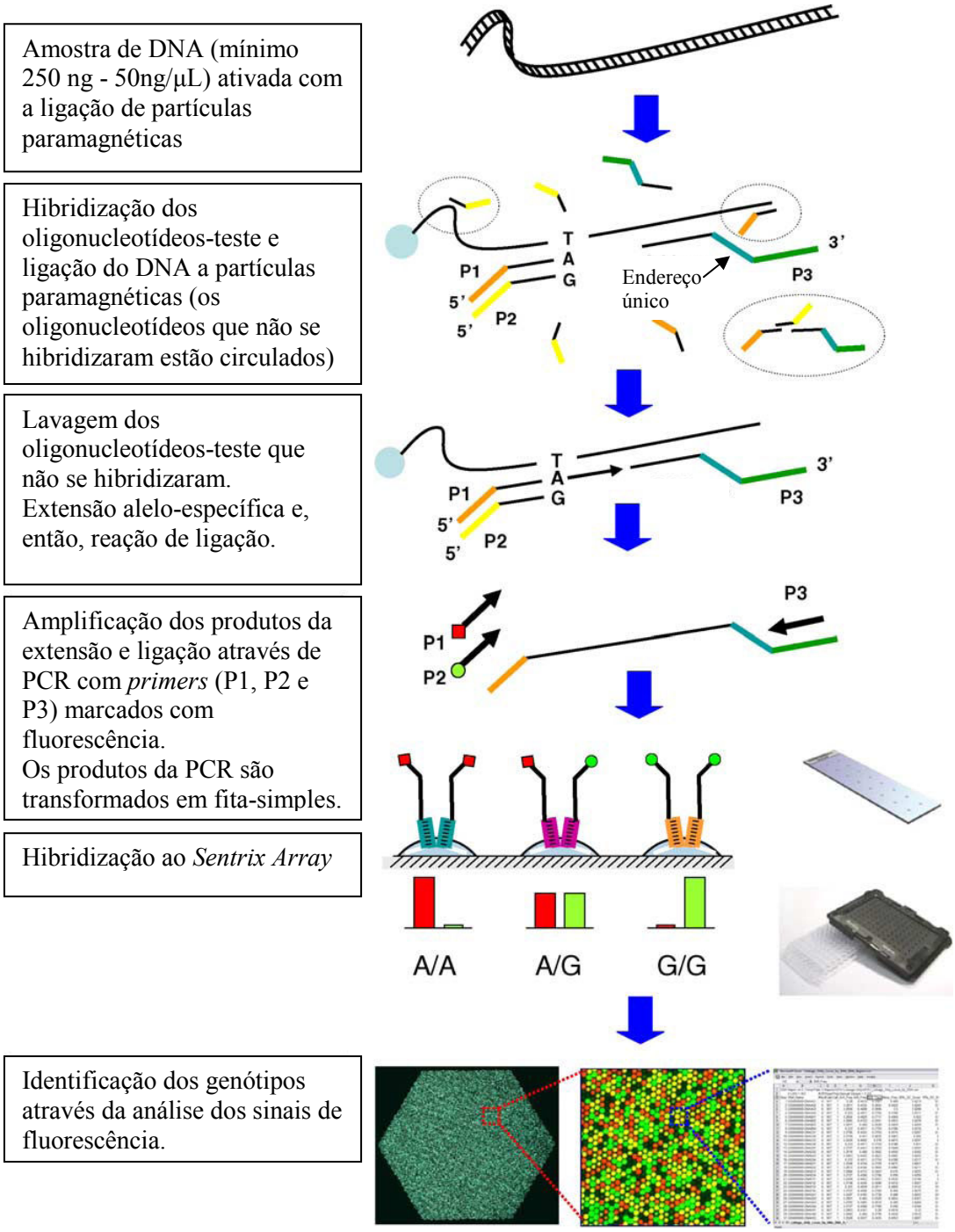


Figura 2. Ilustração do método de genotipagem *Illumina GoldenGate*®. São utilizados três oligonucleotídeos-teste, todos contendo regiões de complementaridade genômica e sítios de iniciadores universais de PCR. Dois deles são específicos a cada alelo do SNP; o terceiro hibridiza-se de uma a 20 bases de distância do SNP e contém sequência de "endereço único", complementar a um tipo particular de pérola da matriz *Sentrix Array*. Após a hibridização, são realizadas lavagens para diminuir o ruído na genotipagem e é adicionada uma polimerase que realiza a extensão do oligonucleotídeo que contém o alelo específico até o oligonucleotídeo que carrega a sequência de "endereço único". Em seguida, ocorre a ligação pela DNA ligase e o produto dessa ligação serve de molde para a amplificação por PCR, utlizando apenas três iniciadores universais (P1, P2 e P3) marcados com fluorescência. Os produtos da PCR, transformados em fita simples, são hibridizados a seu tipo de pérola complementar na matriz, através de sua sequência de "endereço único", permitindo a identificação dos genótipos através do *BeadArray Reader* (Barker *et al*. 2003), um programa especializado na análise da fluorescência emitida. (Figura modificado de Shen *et al*. 2005).

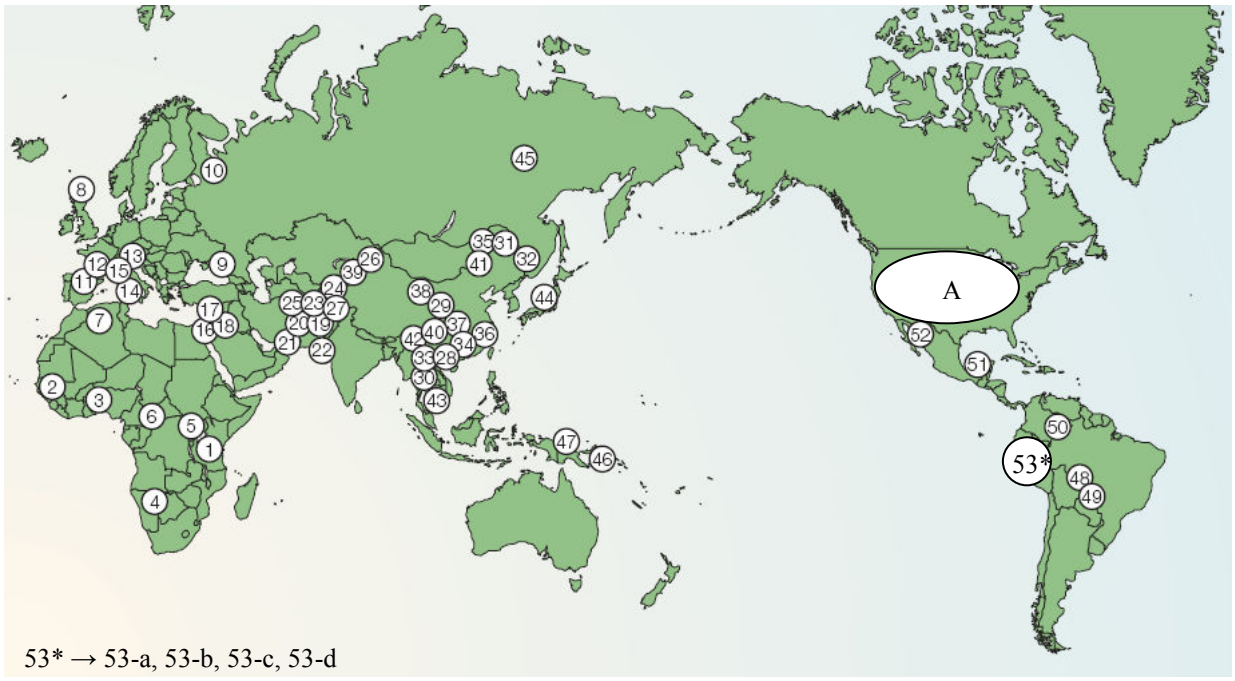


Figura 3: Distribuição geográfica aproximada das populações do HGDP-CEPH, do *SNP500Cancer* e das 4 populações Nativo Americanas do Peru e Equador de nosso laboratório, cujos dados estão disponíveis para esse estudo.

1 Bantu NE e SE/SO; **2** Mandenka; **3** Ioruba; **4** San; **5** Pigmeu Mbuti; **6** Pigmeu Biaka; **7** Mozabite; **8** Orcadiana; **9** Adygei; **10** Russa NO; **11** Francesa Basca; **12** Francesa; **13** Bérgamo; **14** Sardenha; **15** Toscana; **16** Beduína; **17** Drusa; **18** Palestina; **19** Balochi; **20** Brahui; **21** Makrani; **22** Sindhi; **23** Pathan; **24** Burusho; **25** Hazara; **26** Uigur; **27** Kalash; **28** Han (China S); **29** Han (China N); **30** Dai. **31** Daur; **32** Hezhen; **33** Lahu; **34** Miaozu (Miao); **35** Oroqen; **36** She; **37** Tujia; **38** Tu; **39** Xibo; **40** Yizu (Yi); **41** Mongólia; **42** Naxi; **43** Camboja; **44** Japonesa; **45** Yakut; **46** Melanésia; **47** Papua; **48** Karitiana; **49** Suruí; **50** Piapoco e Curripaco; **51** Maia; **52** Pima; **53*** Populações Nativas do Peru e Equador (**53-a** Cayapa, **53-b** Quechua, **53-c** San Martín e **53-d** Matsiguenga); **A** Populações do *SNP500Cancer* (Ascendência Caucasiana; Ascendência Asiática; Hispânicos e Afro-americanos).
(Figura modificada de Cavalli-Sforza, 2005)

A primeira forma de análise envolveu 56 amostras populacionais: quatro populações nativo-americanas do nosso laboratório e 52 do HGDP. Em seguida, foram analisadas as quatro populações do *SNP500Cancer*. Por fim, foram analisados os nove grupos populacionais formados por amostras do HGDP e nativo-americanas do nosso laboratório.

III - 2. A plataforma R

A plataforma R (*R Development Core Team*, 2008; http://www.r-project.org) é um ambiente e uma poderosa linguagem de programação voltada para a manipulação de dados estatísticos, modelagem e visualização de gráficos. R oferece recursos similares aos fornecidos pela linguagem S (Becker *et al*. 1998) e por outros programas de análise estatística, como o *MatLab*® e *Mathematica*®. Porém, apresenta a grande vantagem de ser um software livre *GNU* que utiliza a licença aderida GPL (*General Public Licence*).

R, assim como S, é uma plataforma projetada em torno de uma verdadeira linguagem de computação bem desenvolvida, simples e eficaz, que inclui condições, loops e ferramentas de entrada e saída de arquivos. Isso permite aos seus usuários criar novas funções, adicionando funcionalidade ao programa. Grande parte do seu sistema é escrito no próprio dialeto R, o que torna mais fácil a compreensão das escolhas algorítmicas realizadas. A R é altamente extensível através de pacotes. Muitos deles foram desenvolvidos através do projeto *Bioconductor*, que visa ao desenvolvimento de pacotes de programas que forneçam ferramentas de análise e compreensão de dados genômicos. Esses pacotes encontram-se disponíveis no repositório do projeto, no sítio http://www.bioconductor.org/.

Através da plataforma R é possível realizar uma ampla e variada gama de análises estatísticas, como, por exemplo, modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais e métodos multivariados. Os gráficos gerados são bem delineados e de alta qualidade para impressão, e há a possibilidade de inclusão de fórmulas e símbolos matemáticos quando necessário. Devido a todas essas vantagens a R tem sido cada vez mais usada em análises de dados biológicos (e. g. Todd *et al*. 2007; Schmid *et al*. 2006; Aldrich *et al*. 2008; Stranger *et al*. 2007).

Nesse projeto, na estimativa dos parâmetros de diversidade, utilizou-se um conjunto de pacotes disponíveis na plataforma R para análises de genética de populações: *The Genetics Package* (Warnes *et al*. 2003), *Package 'adegenet'* (Jombart & Solymos 2008), *The ade4 Package* (Chessel *et al*. 2004) e *HIERFSTAT* (Goudet, 2005). Para tanto, foram desenvolvidos *scripts* na linguagem R que automatizaram os cálculos para as combinações de SNPs e

populações ou grupos populacionais estudados. Para a produção de gráficos foram utilizadas ferramentas básicas da plataforma R e, quando necessário, utilizou-se o pacote R *Lattice* (Sarkar 2002).

Para gerar o arquivo de entrada (*input file*) para o programa R, foi utilizado um algoritmo escrito em linguagem *Perl*, desenvolvido pelo aluno de doutorado Wagner Magalhães (Programa de Doutorado em Bioinformática - UFMG). Através desse algoritmo, foi possível criar uma matriz de genótipos (SNPs), nas colunas, por indivíduos, nas linhas, a partir dos dados obtidos no *DIVERGENOMEdb*, para cada população e grupo populacional estudados.

III - 3. Freqüências alélicas e teste do equilíbrio de Hardy-Weinberg

As freqüências alélicas e genotípicas foram calculadas através das funções R *summary* e *genotype*, esta última pertencente ao pacote R *The Genetics Package* (Warnes *et al*. 2003). A função *genotype* cria um objeto com os genótipos dos *loci* analisados, a partir de um vetor contendo os alelos dos SNPs para cada um dos indivíduos amostrados. A função *summary*, por sua vez, é uma função R genérica usada para produzir resumos dos objetos resultantes da aplicação de outras funções. Ao aplicá-la ao objeto resultante da função *genotype* foram obtidas, entre outros argumentos, as freqüências alélicas e genotípicas para os SNPs dentro das amostras das populações e grupos populacionais.

Para testar o equilíbrio de Hardy-Weinberg (EHW) das proporções genotípicas de cada SNP, dentro das populações estudadas, foi empregado o teste exato, utilizando-se a função *HWE.exact* do pacote R *The Genetics Package* (Warnes *et al*. 2003). O teste exato estima a probabilidade de se obter a distribuição observada (ou uma distribuição ainda mais diferente da esperada no EHW), quando a hipótese nula, isto é, a hipótese de que as freqüências genotípicas estão em EHW, é verdadeira.

Na análise dos valores de significância (P) obtidos com o teste, foi adotada a seguinte correção: se todos os SNPs que usamos fossem independentes, ou seja, não houvesse desequilíbrio de ligação entre eles, por se tratar de uma comparação múltipla, a correção de *Bonferroni* α/N poderia ser adotada. A partir dessa correção, o patamar do valor de P, que delimita a região de rejeição da hipótese nula, seria reduzido, diminuindo-se, assim, a probabilidade do erro tipo I (em que se rejeita a hipótese nula quando ela é, de fato, verdadeira). Nesse caso, como temos 56 populações e, para cada uma delas, há 1256 SNPs

que estão sendo testados para o EHW (56 x 1256 comparações múltiplas), ao se fixar α=5%, teríamos:

$$\alpha/N = 0{,}05/1256x56 = 7{,}11 \times 10^{-7}$$

Assim, uma possibilidade seria adotar 10^{-7} como o patamar superior do valor P. Entretanto, os 1256 SNPs considerados nas análises não são completamente independentes. De fato, muitos deles estão no mesmo gene e podem apresentar desequilíbrio de ligação. Dessa forma, um conjunto de SNPs no mesmo gene poderia estar fora do EHW porque compartilham uma história evolutiva, e não por um erro tipo I. Essa ausência de independência entre os SNPs implicaria que, ao adotarmos a estratégia conservadora da correção de *Bonferroni*, aumentaríamos a possibilidade de incorrermos em erro tipo II (em que não rejeitamos a hipótese nula quando ela é falsa). Por este motivo, optamos por elevar esse patamar para 10^{-4}, para diminuir a possibilidade da ocorrência desse tipo de erro, apesar de reconhecermos que através dessa estratégia poderíamos obter resultados falsos positivos (erro tipo I). Por raciocínio análogo, também foi adotado o patamar 10^{-4} na análise dos valores P encontrados para os SNPs das populações do *SNP500Cancer*.

III - 4. Heterozigosidade esperada

A heterozigosidade esperada (*Hu*) também foi obtida através aplicação da função R *summary* ao resultado da função *genotype*. Para o seu cálculo, o programa utilizou a seguinte fórmula:

$$H_u = \left(1 - \sum_{i=1}^{k} \hat{p}_i^2\right) \frac{2n}{2n-1}$$

Onde:

- $\hat{p}_i$ é a freqüência alélica estimada para o alelo "i" (i.e. a proporção observada na amostra);
- *n* é o número de indivíduos amostrados.

A partir de todos os valores de heterozigosidade encontrados para os *loci*, dentro de cada amostra das populações e dos grupos populacionais, foi calculada uma média por amostra.

III - 5. Estatísticas-*F* (F_{ST}, F_{IS}, F_{IS})

No cálculo das estatísticas-*F* (F_{ST}, F_{IS}, F_{IS}), utilizou-se os pacotes R *Package 'adegenet'* (Jombart & Solymos, 2008) e *HIERFSTAT* (Goudet, 2005). Através da função *df2genind* do *Package 'adegenet'*, uma matriz de SNPs por indivíduos, de cada de população ou grupo populacional amostrado, foi transformada em um objeto no formato *genind*. Esse é um formato usado em R para armazenar genótipos individuais. Após obter os dados nesse formato, foi acionado o pacote *HIERFSTAT*, no qual está implementado o algoritmo de Yang (1998), que calcula as estatísticas-*F*, com base nas seguintes definições:

- σ^2_i é o componente de variância para o nível *i*;
- $\sigma^2_{\sum i} = \sum_{k=1}^{i} \sigma_k^2$ é a soma dos componentes de variância do nível hierárquico mais baixo até o nível *i*;
- $\sigma^2_{i(j)} = \sum_{k=(j+1)}^{i} \sigma_k^2$ é a soma dos componentes de variância a partir de um nível acima do nível hierárquico *j* até o nível *i*;

A estatística-*F* entre os níveis hierárquicos *j* e *i* é determinada por:

$$F_{ji} = \frac{\sigma^2 i(j)}{\sigma^2_{\sum i}}$$

Em nossas análises, foram utilizados quatro níveis hierárquicos: indivíduos (nível 1), populações (nível 2), grupos populacionais (nível 3) e mundo (nível 4), relativo ao conjunto resultante da união das amostras de todas as populações. Dentro dos grupos populacionais, adotando-se o nível 3 como o nível hierárquico mais alto, foram calculadas as estatísticas-*F* F_{ST}, F_{IS}, F_{IS} (respectivamente, F_{23}, F_{12} e F_{13}). Quando se considerou o nível 4 como o nível hierárquico mais elevado, foram calculadas as estatísticas F_{ST} (F_{24}) e F_{CT} (F_{34}), onde o sub-índice *c* refere-se aos grupos populacionais.

Nas amostras das populações do HGDP, dos nativos americanos do nosso laboratório e do *SNP500Cancer*, calculou-se o coeficiente de endocruzamento F_{IS} (F_{13}).

III - 6. Análise dos Componentes Principais (ACP)

A ACP foi utilizada para investigar a estruturação genética dentro dos grupos e populações analisados. Essa análise consiste em uma transformação linear de "n" variáveis

originais em “n” variáveis novas, chamadas componentes principais (CPs), de modo que a primeira variável nova computada seja responsável pela maior variação possível existente no conjunto de dados, a segunda pela maior variação possível restante e assim por diante até que toda a variação do conjunto tenha sido explicada. A ACP é, assim, uma técnica de transformação de variáveis: se cada variável medida pode ser considerada como um eixo de variabilidade, estando usualmente correlacionada com outras variáveis, esta análise transforma os dados, de modo a descrever a mesma variabilidade total existente, com o mesmo número de eixos originais, porém, não mais correlacionados entre si.

Para a sua realização, os dados a serem transformados utilizados foram matrizes de genótipos individuais. Foram empregados os pacotes R *Package ‘adegenet’* (Jombart & Solymos, 2008) e *The ade4 Package* (Chessel *et al*. 2004). Visando-se diminuir a distorção na estrutura dos *eigenvalues*, causada pelo desequilíbrio de ligação entre os SNPs analisados no mesmo gene, para essas análises utilizou-se apenas um SNP de cada um dos genes disponíveis no estudo, para um total de 390 SNPs (Anexo I).

As análises de ACP foram realizadas nas matrizes de genótipos de indivíduos de cada um dos nove grupos populacionais, formados por amostras das populações do HGDP e nativo-americanas geradas em nosso laboratório, sendo cada indivíduo do grupo identificado com a sua respectiva população. Também foram realizadas análises do tipo ACP para todas essas as populações conjuntamente e para o conjunto amostral formado pelas quatro populações do *SNP500Cancer*. Por fim, a ACP foi utilizada para pesquisar a influência genética das populações parentais na formação da população miscigenada hispânica.

IV - Resultados

IV - 1. Freqüências alélicas e genotípicas e Teste do Equilíbrio de Hardy-Weinberg

A partir do cálculo das freqüências alélicas em cada população, foi possível a aplicação do teste do EHW. A população que apresentou mais SNPs para os quais se rejeitou o EHW foi a francesa (20 SNPs), seguida das populações do centro-sul asiático Sindhi (16 SNPs) e Pathan (10), e da nativo-americana Quechua (10 SNPs; Tabela 1). Nos três primeiros casos a rejeição deveu-se, na grande maioria das vezes, a um excesso na freqüência de genótipos homozigotos em relação à esperada. Na população Quechua, situada nos Andes Centrais do Peru, a maior parte das ocorrências de rejeição do EHW deveu-se a um excesso na freqüência de heterozigotos em relação à esperada.

Entra as populações do *SNP500Cancer*, a população asiática apresentou o maior número de SNPs com freqüências genotípicas fora do EHW (Tabela 2).

IV - 2. Variabilidade intra-populacional: Heterozigosidade esperada

A média da heterozigosidade esperada sob a hipótese de EHW para todos os *loci* foi utilizada como estimativa de diversidade intra-populacional. As populações nativo-americanas e da Oceania apresentaram a menor diversidade (Gráfico 1), sendo que a população Suruí apresentou o valor mais baixo dentre todas as populações analisadas (Tabela 1). Com exceção de Matsiguenga, as demais populações do oeste da América do Sul - Cayapa, San Martín e Quechua - apresentaram valores de heterozigosidade esperada mais altos do que os encontrados para as populações do leste - Piapoco e Curripaco, Karitiana e Suruí. O continente africano (leste e oeste) também apresentou níveis baixos de diversidade intra-populacional (Tabela 1; Gráfico 1). As populações do leste asiático apresentaram níveis intermediários de diversidade intra-populacional. Dentro deste grupo, a população Uigur apresentou uma maior diversidade (Tabela 1; Gráfico 1).

As populações do centro-sul asiático, da Europa e do Oriente Médio apresentaram os maiores níveis de diversidade intra-populacional. Em particular, as populações paquistanesas do centro-sul asiático, Sindhi, Makrani e Brahui, e as populações européias, Rússia e Adygei, foram as que apresentaram os valores mais elevados dentre todas as populações mundiais estudadas.

Tabela 1: Distribuição das populações estudadas (HGDP e nativo-americanos do nosso laboratório) ao longo de seus respectivos grupos e parâmetros de estimativa da diversidade intra-populacional.

Grupo	População[1]	N[2]	Het. Esp.[3]	*Ranking* Het. Esp.[4]	F_{IS}	*Ranking* F_{IS}[5]	SNPs fora do EHW[6]	Hom.[7]	Het.[8]
Leste Africano	Pigmeu Biaka (6)	31	0,271	13	-0,012	14	2	1	1
Leste Africano	Pigmeu Mbuti (5)	13	0,246	7	0,011	36	zero	–	–
Leste Africano	Bantu NE (1)	11	0,292	19	-0,006	18	zero	–	–
Leste Africano	Bantu SE e SO (1)	8	0,284	17	0,025	46	zero	–	–
Oeste Africano	Mandenka (2)	24	0,281	15	-0,001	22	2	2	–
Oeste Africano	Ioruba (3)	25	0,282	16	0,019	42	zero	–	–
Oeste Africano	San (4)	7	0,228	2	-0,015	11	zero	–	–
Oriente Médio	Mozabite (7)	30	0,336	38	0,006	29	2	1	1
Oriente Médio	Beduína (16)	48	0,340	40	0,049	55	zero	–	–
Oriente Médio	Drusa (17)	47	0,344	47	0,023	45	3	3	–
Oriente Médio	Palestina (18)	49	0,348	51	0,009	34	7	6	1
Europa	Francesa (12)	29	0,342	43	0,025	49	20	17	3
Europa	Francesa Basca (11)	24	0,341	42	-0,023	9	3	3	–
Europa	Sardenha (14)	28	0,339	39	-0,001	23	3	1	2
Europa	Bérgamo (13)	13	0,346	50	0,027	50	zero	–	–
Europa	Toscana (15)	8	0,345	49	-0,001	24	zero	–	–
Europa	Orcadiana (8)	16	0,331	37	0,025	48	1	1	–
Europa	Adygei (9)	15	0,353	56	0,018	41	1	1	–
Europa	Russa NO (10)	25	0,351	55	-0,008	16	6	4	2
Centro Sul Asiático	Brahui (20)	25	0,350	54	0,017	40	3	2	1
Centro Sul Asiático	Balochi (19)	25	0,341	41	0,046	52	2	2	–
Centro Sul Asiático	Hazara (25)	25	0,342	45	0,021	44	zero	–	–
Centro Sul Asiático	Makrani (21)	25	0,349	53	0,046	53	zero	–	–
Centro Sul Asiático	Sindhi (22)	24	0,349	52	0,041	51	16	16	–
Centro Sul Asiático	Pathan (23)	24	0,343	46	0,047	54	10	10	–
Centro Sul Asiático	Kalash (27)	25	0,320	36	-0,007	17	1	1	–
Centro Sul Asiático	Burusho (24)	25	0,345	48	0,011	37	1	1	–
Leste Asiático	Camboja (43)	10	0,317	35	0,009	33	zero	–	–
Leste Asiático	Han (28 e 29)	39	0,299	23	0,025	47	1	1	–
Leste Asiático	Tujia (37)	10	0,307	30	0,012	38	zero	–	–
Leste Asiático	Yizu-Yi (40)	10	0,303	29	0,002	27	zero	–	–
Leste Asiático	Miaozu-Miao (34)	10	0,301	27	-0,025	6	zero	–	–
Leste Asiático	Oroqen (35)	10	0,300	25	-0,029	5	zero	–	–
Leste Asiático	Daur (31)	10	0,296	20	0,007	30	zero	–	–
Leste Asiático	Mongólia (41)	10	0,314	34	0,009	32	zero	–	–
Leste Asiático	Hezhen (32)	9	0,300	26	-0,024	7	zero	–	–
Leste Asiático	Xibo (39)	9	0,307	31	0,01	35	zero	–	–
Leste Asiático	Uigur (26)	10	0,342	44	0,008	31	zero	–	–
Leste Asiático	Dai (30)	10	0,303	28	-0,003	20	zero	–	–
Leste Asiático	Lahu (33)	10	0,289	18	-0,016	10	zero	–	–

Grupo	População[1]	N[2]	Het. Esp.[3]	*Ranking* Het. Esp.[4]	F_{IS}	*Ranking* F_{IS}[5]	SNPs fora do EHW[6]	Hom.[7]	Het.[8]
Leste Asiático	She (36)	10	0,300	24	-0,041	4	zero	–	–
Leste Asiático	Naxi (42)	10	0,298	21	0,004	28	zero	–	–
Leste Asiático	Tu (38)	10	0,312	32	-0,002	21	zero	–	–
Leste Asiático	Yakut (45)	25	0,313	33	-0,004	19	1	1	–
Leste Asiático	Japonesa (44)	30	0,299	22	0,002	26	zero	–	–
Oceania	Papua (47)	17	0,250	9	0,002	25	1	1	–
Oceania	Melanésia (46)	15	0,267	10	-0,045	3	1	1	–
América Central	Pima (52)	24	0,246	6	0,02	43	2	1	1
América Central	Maia (51)	25	0,278	14	-0,013	13	3	3	–
América do Sul	Piapoco e Curripaco (50)	13	0,244	5	-0,054	2	zero	–	–
América do Sul	Karitiana (48)	23	0,232	4	-0,023	8	1	1	–
América do Sul	Suruí (49)	21	0,203	1	-0,102	1	zero	–	–
América do Sul	Cayapa* (53-a)	7	0,246	8	0,017	39	zero	–	–
América do Sul	Quechua* (53-b)	22	0,270	12	-0,015	12	10	2	8
América do Sul	San Martín* (53-c)	17	0,270	11	0,065	56	1	1	–
América do Sul	Matsiguenga* (53-d)	21	0,231	3	-0,01	15	1	1	–

* Amostras populacionais disponíveis em nosso laboratório;
[1] Número entre parêntesis referentes à localização geográfica da população no mapa da Figura 3;
[2] N: Número de indivíduos da população;
[3] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de *Hardy Weinberg*;
[4] *Ranking* Het. Esp.: Classificação em relação aos valores da média da heterozigosidade esperada (1 mínimo, 56 máximo);
[5] Ranking F_{IS}: Classificação em relação aos valores de F_{IS} (1 mínimo, 56 máximo);
[6] SNPs fora do EHW: Número de SNPs para os quais foi rejeitada a hipótese de equilíbrio de *Hardy Weinberg* ($p<10^{-4}$);
[7] Hom e [8] Het: Número de SNPs para os quais a rejeição da hipótese de equilíbrio de *Hardy Weinberg* se deveu a um excesso na freqüência de homozigotos ou heterozigotos, respectivamente, em relação à esperada sob a hipótese nula.

Tabela 2: Populações estudadas do *SNP500Cancer* e parâmetros de estimativa da diversidade intra-populacional.

População	N[1]	Het. Esp.[2]	*Ranking* Het. Esp.[3]	F_{IS}	*Ranking* F_{IS}[4]	SNPs fora do EHW[5]	Hom.[6]	Het.[7]
África	24	0,310	1	0,041	2	1	1	–
Europa	31	0,350	3	0,001	1	2	2	–
Ásia	24	0,324	2	0,091	4	9	9	–
Hispânicos	23	0,350	4	0,048	3	1	–	1

[1] N: Número de indivíduos da população;
[2] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de *Hardy Weinberg*;
[3] *Ranking* Het. Esp.: Classificação em relação aos valores da média da heterozigosidade esperada (1 mínimo, 4 máximo);
[4] Ranking F_{IS}: Classificação em relação aos valores de F_{IS} (1 mínimo, 4 máximo);
[5] SNPs fora do EHW: Número de SNPs para os quais foi rejeitada a hipótese de equilíbrio de *Hardy Weinberg* ($p<10^{-4}$);
[6] Hom e [7] Het: Número de SNPs para os quais a rejeição da hipótese de equilíbrio de *Hardy Weinberg* se deveu a um excesso na freqüência de homozigotos ou heterozigotos, respectivamente, em relação à esperada sob a hipótese nula.

A análise da heterozigosidade média encontrada nos grupos populacionais do HGDP e nas populações do *SNP500Cancer* refletiu os resultados das análises das populações isoladas (Tabelas 2 e 3). A população miscigenada hispânica, por sua vez, do *SNP500Cancer* apresentou um valor elevado de diversidade intra-populacional (Tabela 2).

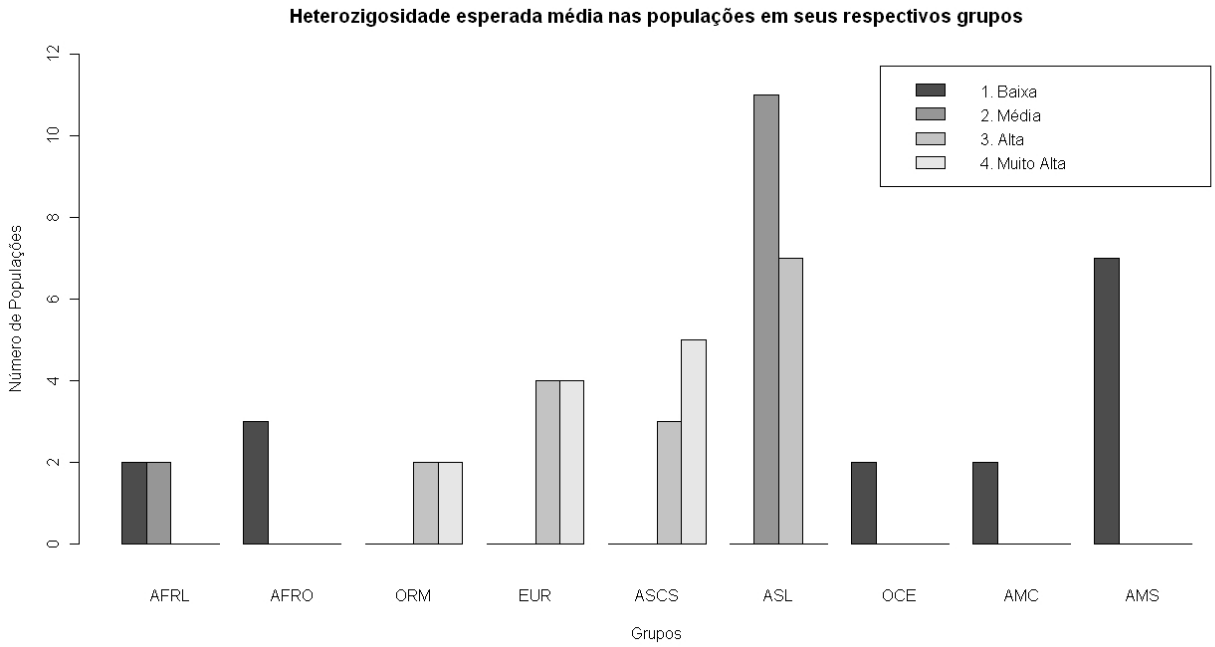


Gráfico 1: Classes* da heterozigosidade esperada média (Het. Esp.) sob a hipótese de EHW, dentro das populações distribuídas ao longo de seus grupos.
*As classes para os valores da Het. Esp. foram definidas de acordo com os quartis:
1. Baixa - Primeiro quartil: Het. Esp. ≤ 0,2825
2. Média - Segundo quartil: 0,2825 < Het. Esp. ≤ 0,3030
3. Alta - Terceiro quartil: 0,3030 < Het. Esp. ≤ 0,3420
4. Muito Alta - Quarto quartil: Het. Esp. > 0,3420
AFRL: Leste Africano; AFRO: Oeste Africano; ORM: Oriente Médio; EUR: Europa; ASCS: Centro Sul Asiático; ASL: Leste Asiático; OCE: Oceania; AMC: América Central; AMS: América do Sul.

IV - 3. Estatísticas-*F* (F_{ST}, F_{IS} e F_{IT})

Dentro das populações do HGDP e nativo-americanas, os menores valores estimados para o coeficiente de endocruzamento (F_{IS}), associados a baixos níveis de heterozigosidades, foram encontradas nas populações nativo-americanas Suruí e Piapoco-Curripaco e na população da Oceania Melanésia (Tabela 1; Gráfico 2). Para população nativo-americana San Martin, que também apresentou um valor baixo de heterozigosidade média, foi encontrado o valor mais elevado de F_{IS} dentre as 56 populações mundiais (0.065), ao lado dos beduínos do Oriente Médio (0,049).

A população francesa, e as populações do centro-sul asiático Sindhi e Pathan, que apresentaram o maior número de SNPs com excesso de homizigotos em relação ao EHW, apresentaram, conforme o esperado, valores altos de F_{IS} (respectivamente 0,025, 0,041 e 0,047). Os nativo-americanos Quéchuas, que apresentaram 10 SNPs com excesso de heterozigotos, apresentaram coerentemente um valor de F_{IS} negativo (-0,015). Dentro do conjunto formado por todas essas populações mundiais (do HGDP e nativo-americanas do nosso laboratório) o valor encontrado para o F_{ST} foi 0,121.

Dentro dos grupos populacionais do HGDP, foram calculadas as estatísticas-*F* F_{ST}, F_{IS} e F_{IT} (Tabela 3; Gráfico 3). Os grupos da América Central, Oceania e América do Sul apresentaram os maiores valores de F_{ST} e F_{IT}, sugerindo uma maior estruturação entre as populações (efeito *Wahlund*) nestas regiões. Os grupos do leste e do oeste africano apresentaram valores intermediários de F_{ST} e F_{IT}, enquanto os grupos do centro-sul e leste asiático, Europa e o Oriente Médio apresentaram os menores valores para essas estatísticas. Para o F_{IS}, os valores mais baixos foram encontrados para os grupos da América do Sul e Oceania. Esses grupos contêm as três populações (Suruí, Piapoco e Curripaco e Melanésia) que apresentaram os valores mais baixos de F_{IS} quando analisadas isoladamente. Os grupos do centro-sul asiático e do Oriente Médio, também refletindo os resultados para as populações isoladas, apresentaram os valores mais altos de F_{IS}. O F_{CT} encontrado quando se analisou todos os grupos conjuntamente foi 0,107.

Em relação às populações do *SNP500Cancer*, a população de ascendência asiática, que apresentou nove SNPs com excesso de homozigotos em relação à EHW, apresentou o maior valor de F_{IS}. O segundo maior valor ocorreu na população miscigenada hispânica, seguida da população de ascendência africana, enquanto a população de ascendência européia apresentou F_{IS} próximo a zero (Tabela 1).

Tabela 3: Grupos populacionais estudados e parâmetros de estimativa da diversidade interna aos grupos.

Grupos Populacionais	N[1]	Het. Esp.[2]	F_{ST}	F_{IS}	F_{IT}
Leste Africano	63	0,318	0,051	-0,006	0,046
Oeste Africano	56	0,289	0,055	0,007	0,061
Oriente Médio	174	0,349	0,018	0,023	0,040
Europa	158	0,349	0,013	0,005	0,017
Centro Sul Asiático	198	0,354	0,019	0,039	0,057
Leste Asiático	242	0,310	0,020	0,003	0,023
Oceania	32	0,272	0,101	-0,019	0,084
América Central	49	0,275	0,078	-0,003	0,075
América do Sul	124	0,299	0,135	-0,019	0,118

[1] N: Número de indivíduos;

[2] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de *Hardy Weinberg;*

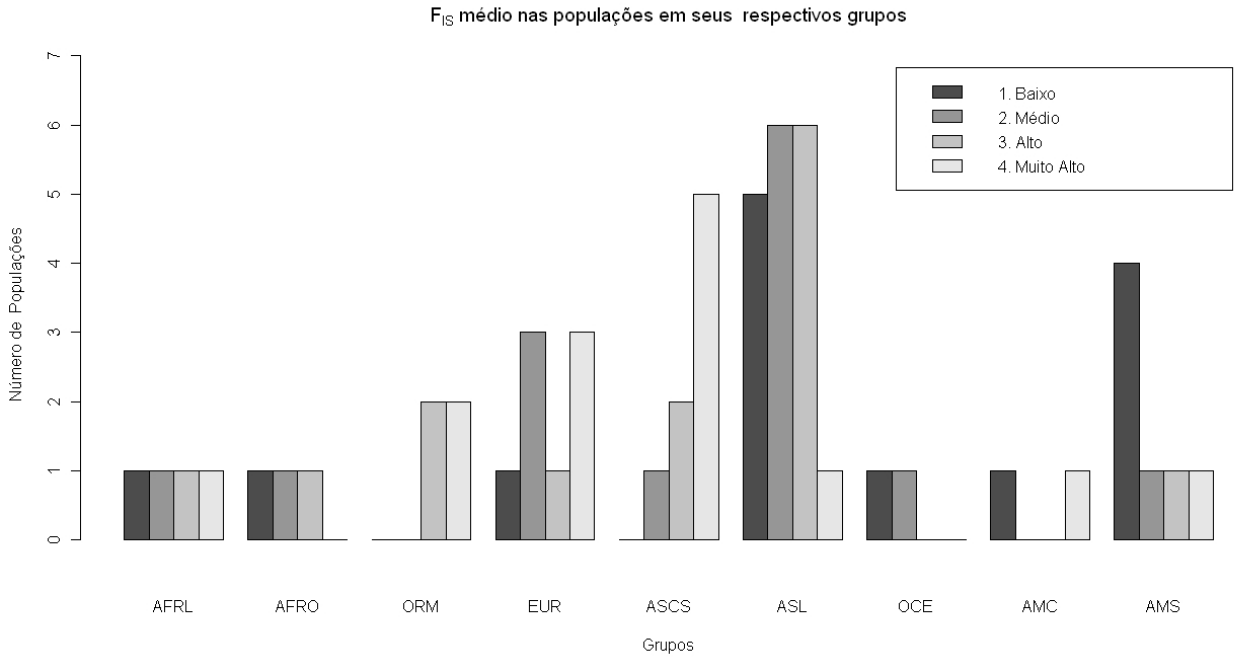


Gráfico 2: Classes* do coeficiente de endocruzamento, F_{IS}, dentro das populações, distribuídas ao longo de seus grupos.

*As classes para os valores do F_{IS} foram definidas de acordo com os quartis:

1. Baixo - Primeiro quartil: $F_{IS} \leq$ -0105
2. Médio - Segundo quartil: -0,0105 < $F_{IS} \leq$ 0,0050
3. Alto - Terceiro quartil: 0,0050 < $F_{IS} \leq$ 0,0192
4. Muito Alto - Quarto quartil: F_{IS} > 0,0192

AFRL: Leste Africano; AFRO: Oeste Africano; ORM: Oriente Médio; EUR: Europa; ASCS: Centro Sul Asiático; ASL: Leste Asiático; OCE: Oceania; AMC: América Central; AMS: América do Sul.

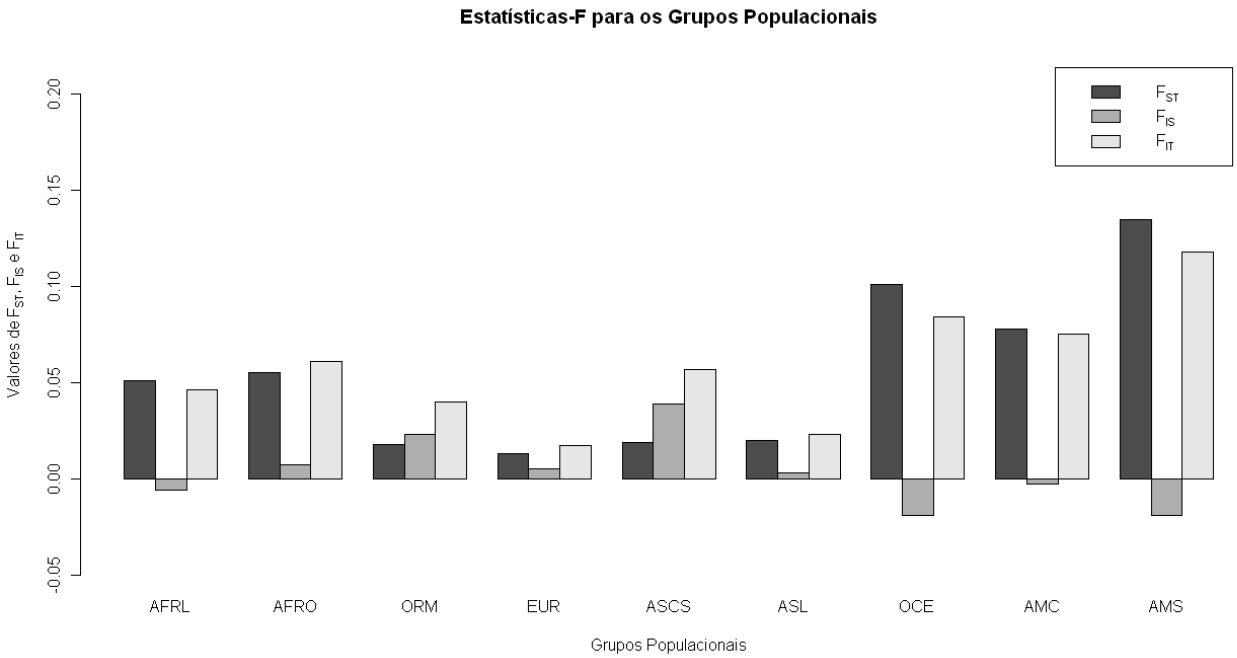


Gráfico 3: Valores das estatísticas-F (F_{ST}, F_{IS} e F_{IT}) dentro dos grupos populacionais.
AFRL: Leste Africano; AFRO: Oeste Africano; ORM: Oriente Médio; EUR: Europa; ASCS: Centro Sul Asiático; ASL: Leste Asiático; OCE: Oceania; AMC: América Central; AMS: América do Sul.

IV - 4. Análise de componentes principais (ACP)

Análises de ACP foram realizadas nas matrizes de genótipos dos grupos populacionais, buscando-se relacionar os resultados dessa análise com os valores de F_{ST} estimados, uma vez que ambos (ACP e F_{ST}) permitem investigar a estruturação genética. Posteriormente foi analisado, através da ACP, o conjunto formado por todos os grupos populacionais. Por último, a ACP foi utilizada para investigar a influência das populações parentais na formação da população miscigenada hispânica do *SNP500Cancer*.

Nos grupos da Oceania e da América Central, que apresentaram valores altos de F_{ST}, o primeiro componente principal (CP1) diferenciou as populações que os compõem (Figuras 4 e 5). Na América do Sul, as populações do leste (Piapoco e Curripaco, Karitiana e Suruí) foram diferenciadas pelos dois primeiros componentes principais (CP1 e CP2), evidenciando a grande contribuição dessas populações para o elevado F_{ST} (o mais alto observado entre os grupos), (Figura 6). Na ACP realizada com todos os indivíduos nativo-americanos, manteve-se esse padrão de diferenciação leste/oeste, com os Piapoco e Curripaco mais próximos das populações da América Central, o que é coerente com sua posição geográfica no norte de América do Sul (Anexo III).

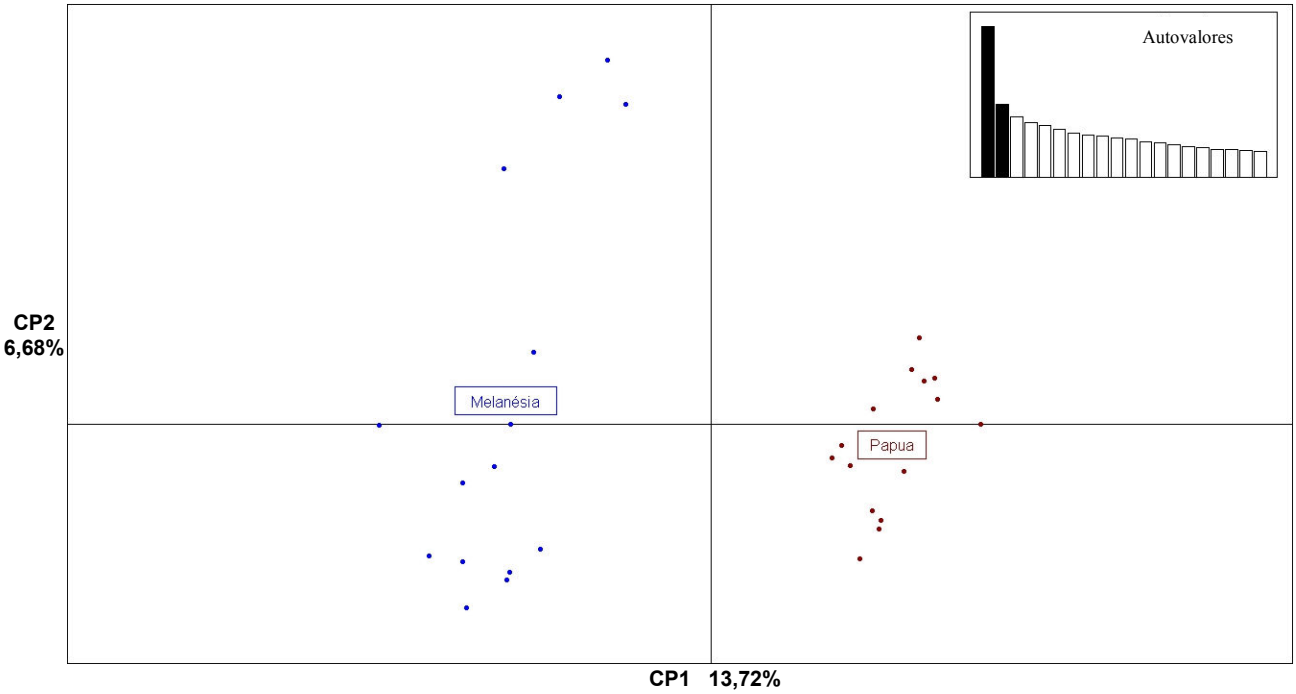


Figura 4: Análise de Componentes Principais na matriz de genótipos do grupo da Oceania. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais.
CP1: Primeiro componente principal; CP2: Segundo componente principal.

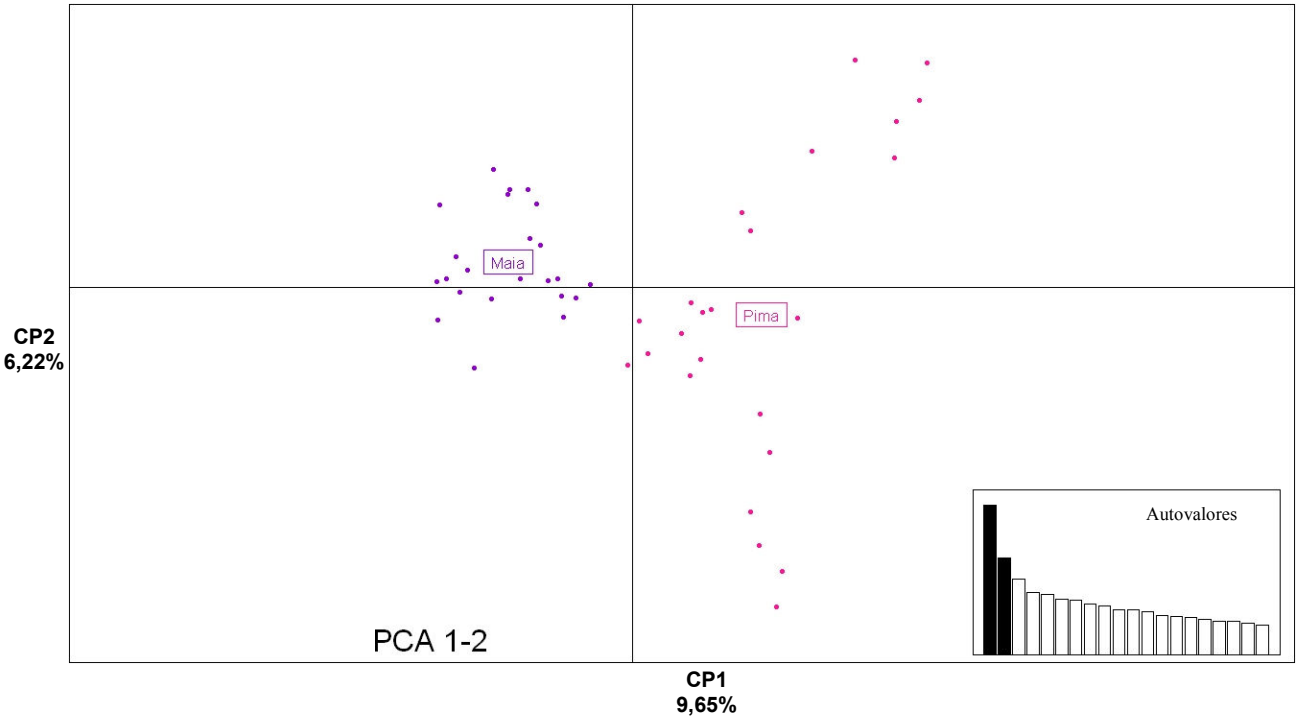


Figura 5: Análise de Componentes Principais na matriz de genótipos do grupo da América Central. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

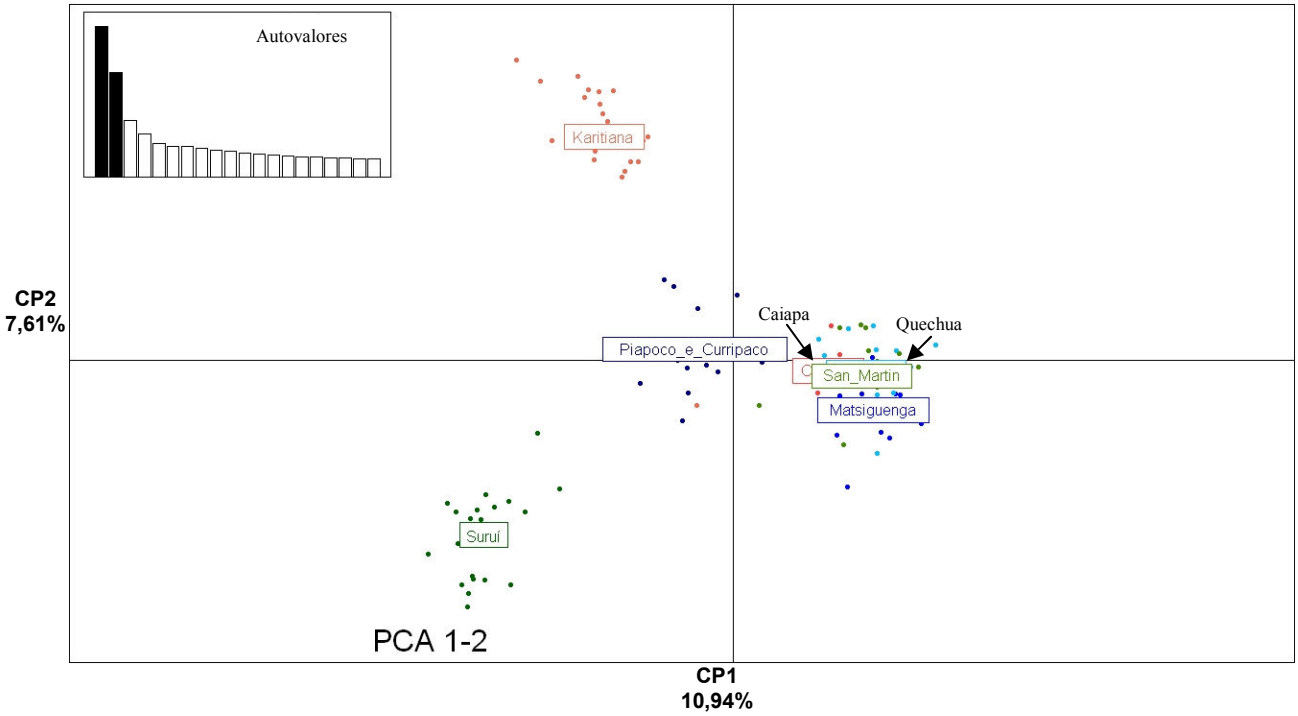


Figura 6: Análise de Componentes Principais na matriz de genótipos do grupo da América do Sul. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

No grupo do leste africano, que apresentou valor de F_{ST} intermediário, o CP1 diferenciou as populações Pigméias Mbuti e Biaka e os Bantus, porém, estes últimos não foram diferenciados em nordeste e sudeste e sudoeste (Figura 7). No grupo do oeste africano, com um F_{ST} também intermediário, é evidente a grande contribuição dos San ao valor de F_{ST} (CP1; Figura 8). A ACP envolvendo os dois grupos africanos não evidenciou a separação entre estes grupos (Anexo IV).

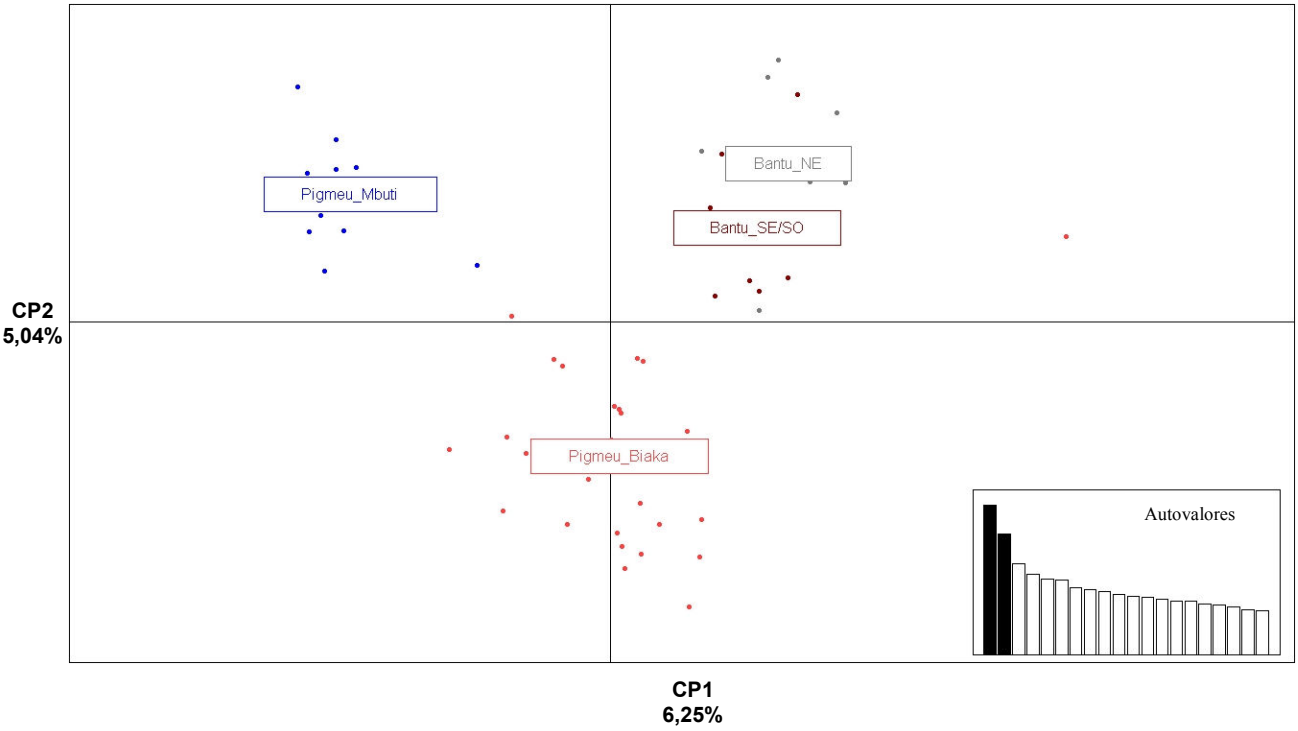


Figura 7: Análise de Componentes Principais na matriz de genótipos do grupo do Leste Africano. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

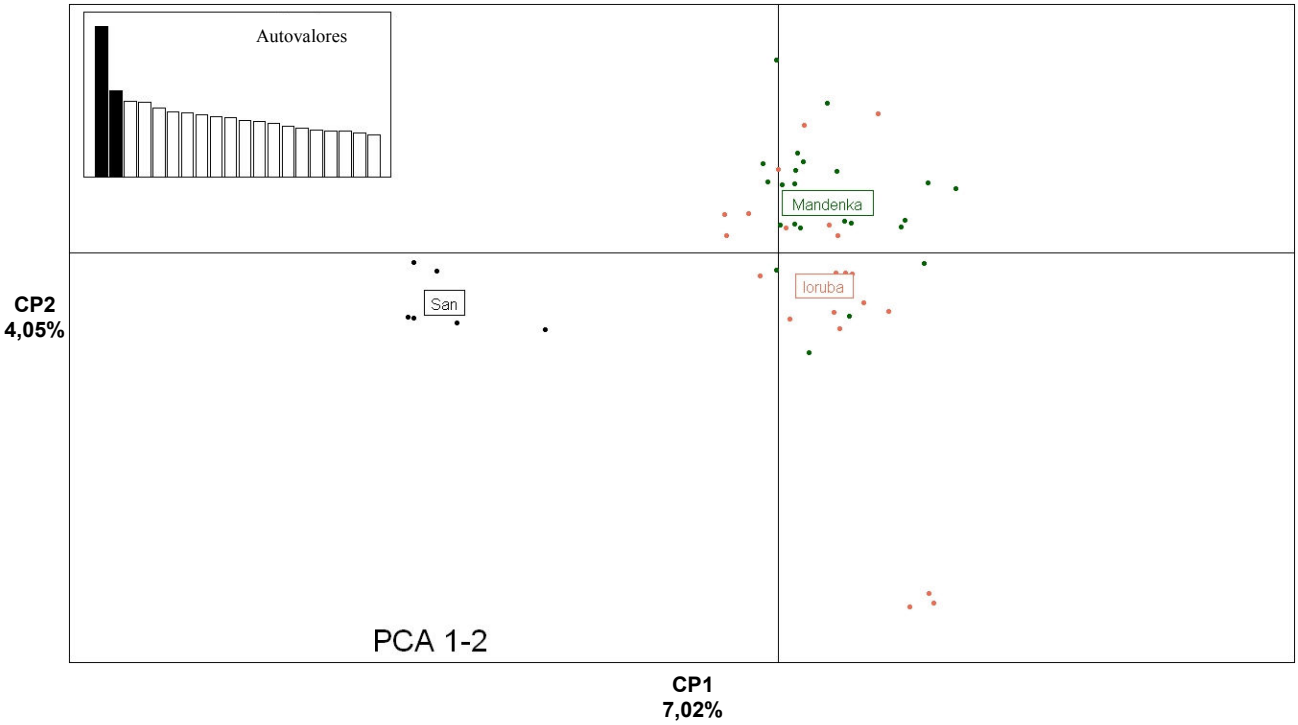


Figura 8: Análise de Componentes Principais na matriz de genótipos do grupo do Oeste Africano. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

Em geral, os grupos euro-asiáticos apresentaram valores baixos de F_{ST}. Na ACP, isto se refletiu em uma separação menos nítida entre os indivíduos pertencentes a populações diferentes no espaço das componentes principais. Porém, é possível observar uma relativa diferenciação de algumas poucoas populações: Kalash (CP1) e Hazara (CP2) no Centro-Sul Asiático (Figura 9); e Yakut e Uigur no Leste Asiático (CP1; Figura 10). Na ACP realizada com todos os indivíduos do centro-sul e leste asiático, embora seja evidente a diferença entre esses dois grupos populacionais (Anexo V), devida em parte à descontinuidade da amostragem no painel do HGDP (Figura 3), ocorreu a sobreposição entre indivíduos da população Hazara (centro-sul) e Uigur (leste), refletindo a proximidade geográfica de Uigur com o grupo do centro-sul asiático.

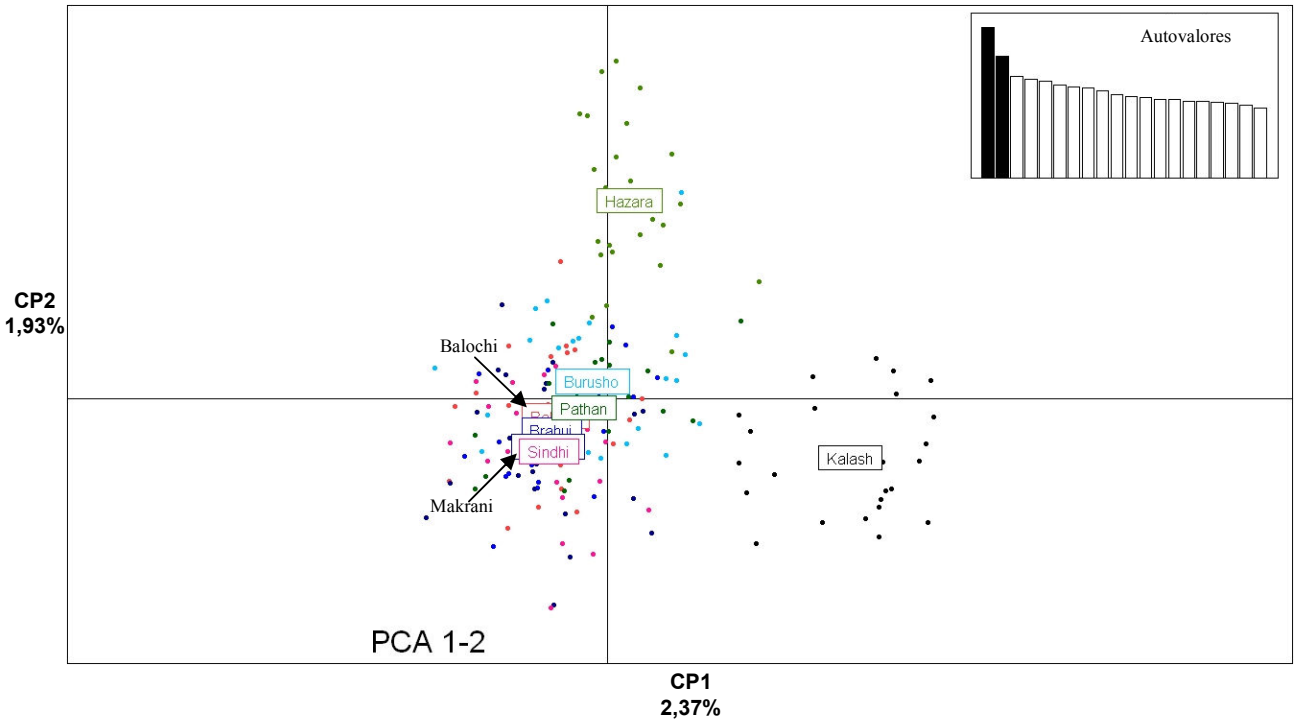


Figura 9: Análise de Componentes Principais na matriz de genótipos do grupo do Centro Sul Asiático. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

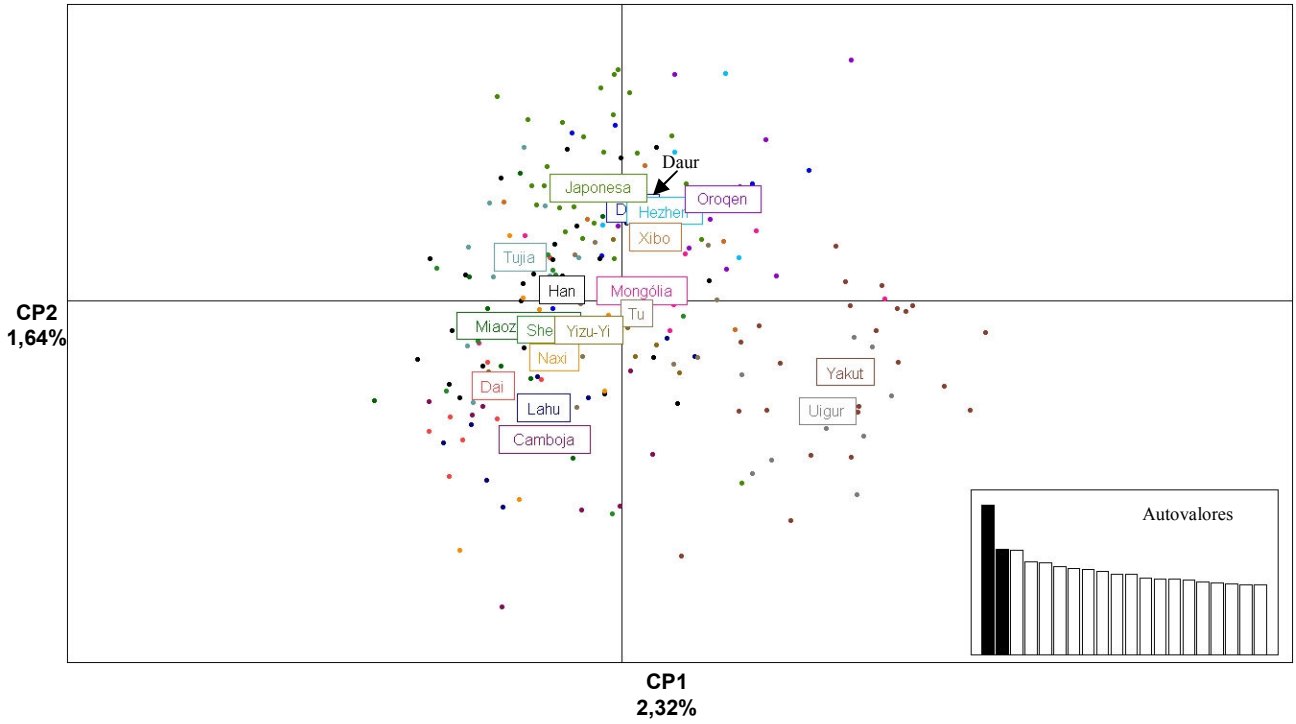


Figura 10: Análise de Componentes Principais na matriz de genótipos do grupo do Leste Asiático. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

No grupo do Oriente Médio, o CP1 evidenciou a maior diferenciação dos Mozabites (Figura 11). Na Europa, as populações européias mostraram-se homogêneas, com exceção de Sardenha e Rússia (Figuras 12).

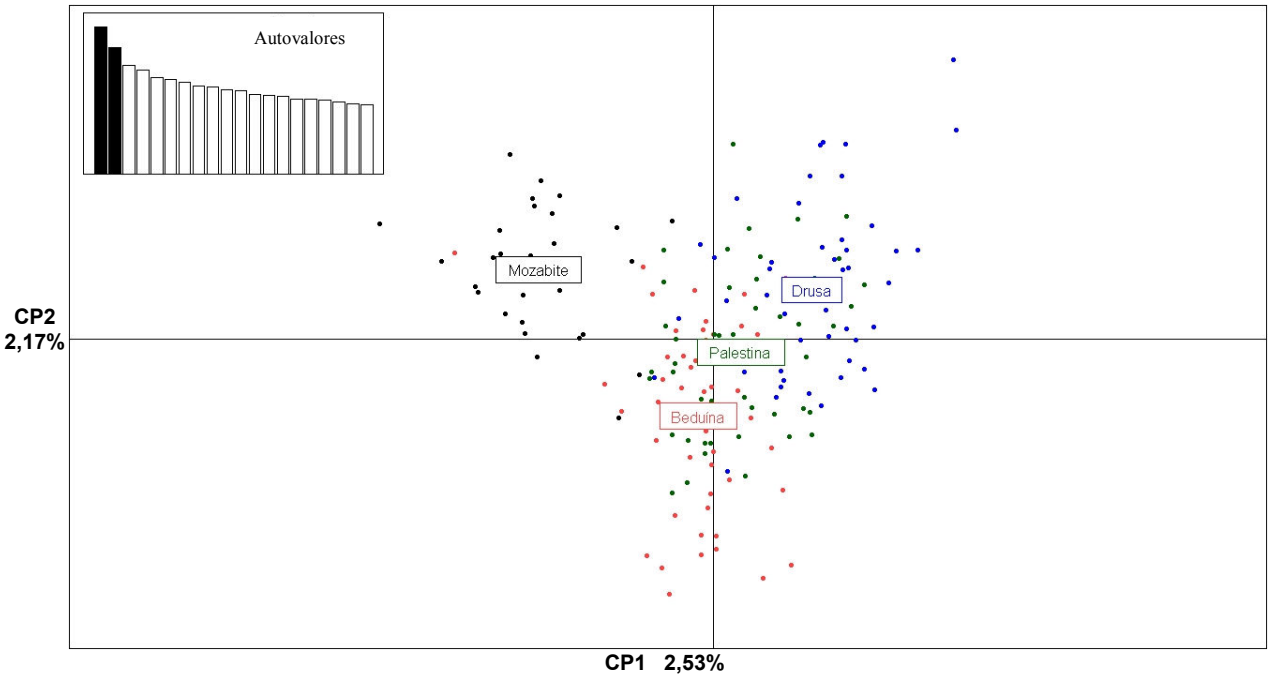


Figura 11: Análise de Componentes Principais na matriz de genótipos do grupo do Oriente Médio. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

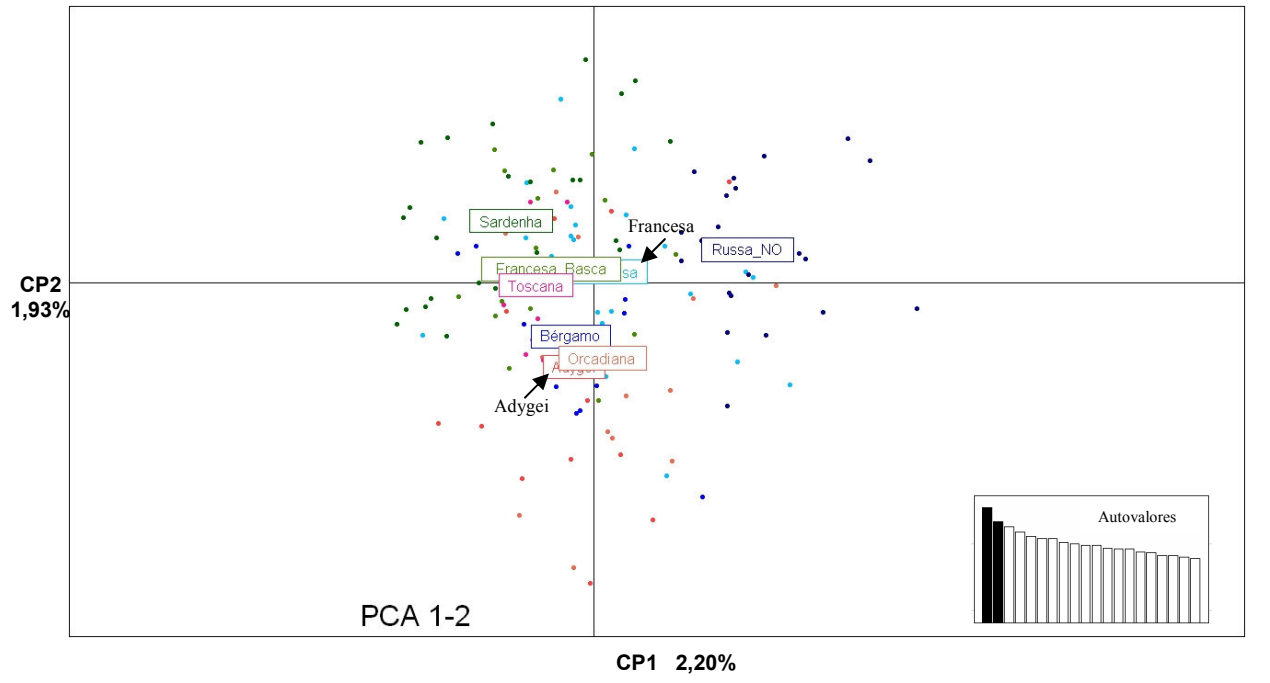
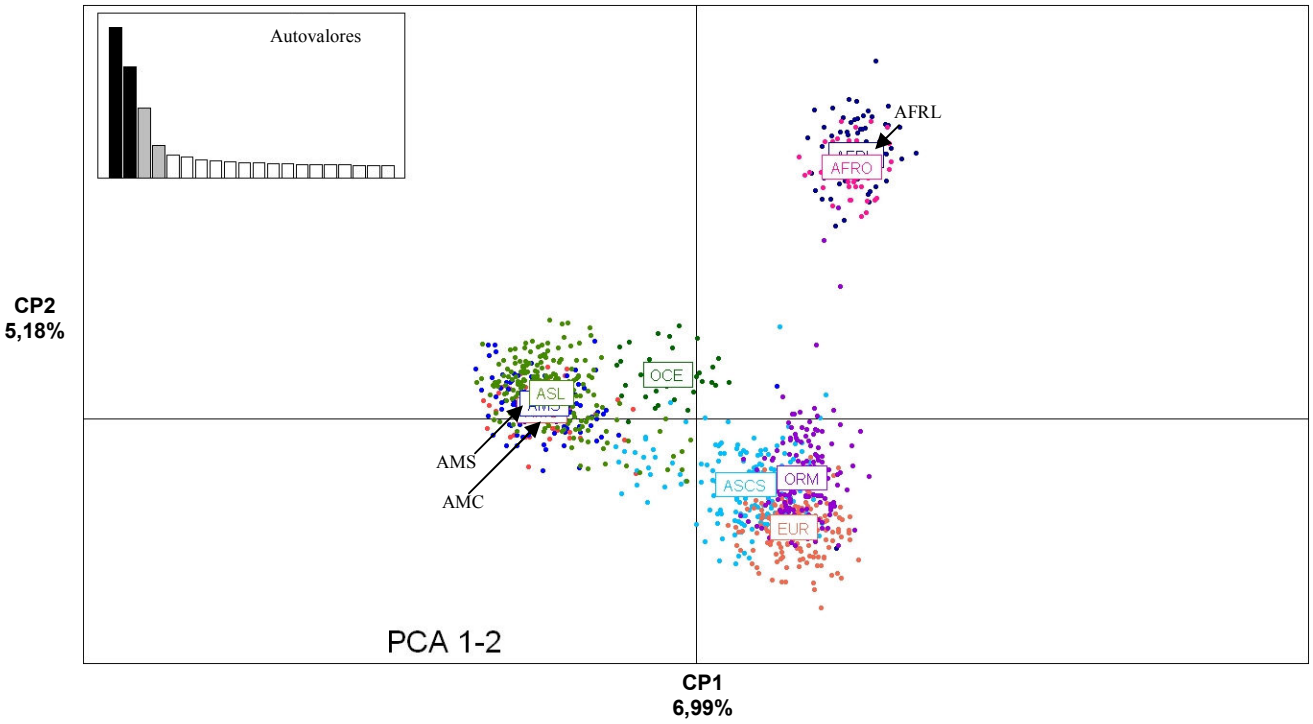


Figura 12: Análise de Componentes Principais na matriz de genótipos do grupo da Europa. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.

Na ACP de todos os indivíduos estudados, o CP1 descreveu uma continuidade leste-oeste na Eurásia, incorporando os nativo-americanos (próximos aos asiáticos do leste) e os africanos (próximos ao grupo formado por Europa, Oriente Médio e centro-sul asiático). O CP2 evidenciou a diferença entre as populações africanas e não-africanas (Figura 13-A). O CP3 e o CP4 distinguiram, respectivamente, os grupos de nativo-americanos e o da Oceania dos demais grupos (Figuras 13-B e 13-C).

A.



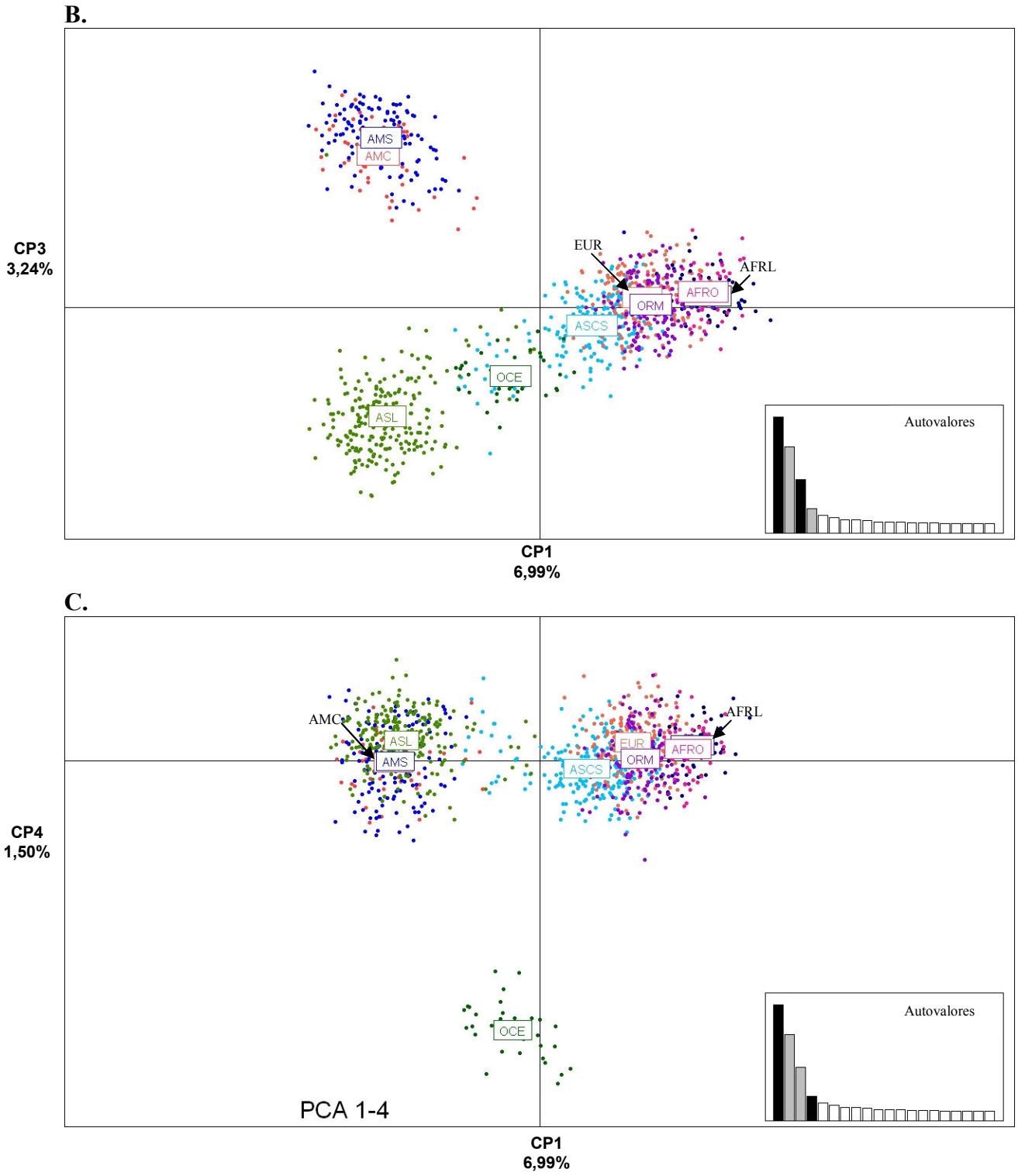


Figura 13: Análise de Componentes Principais na matriz de genótipos de todos os nove grupos populacionais estudados. A. CP1 e CP2; **B.** CP1 e CP3; **C.** CP1 e CP4. Indivíduos do mesmo grupo populacional estão representados por pontos de mesma cor. Valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal; CP3: Terceiro componente principal; CP4: Quarto componente principal. AFRL: Leste Africano; AFRO: Oeste Africano; ORM: Oriente Médio; EUR: Europa; ASCS: Centro Sul Asiático; ASL: Leste Asiático; OCE: Oceania; AMC: América Central; AMS: América do Sul.

Ao se aplicar a ACP à matriz de genótipos formada pelas quatro populações do *SNP500Cancer*, o CP1 separou os indivíduos afro-americanos dos demais e evidenciou a contribuição européia e asiática na formação da população miscigenada hispânica (Figura 14). Na ACP realizada com os grupos populacionais do HGDP, americanos, africanos e europeu, juntamente com a população hispânica, do *SNP500Cancer*, para verificar a influência das populações parentais sobre essa população miscigenada, os hispânicos localizaram-se entre essas populações parentais no espaço das componentes principais (Anexo VI).

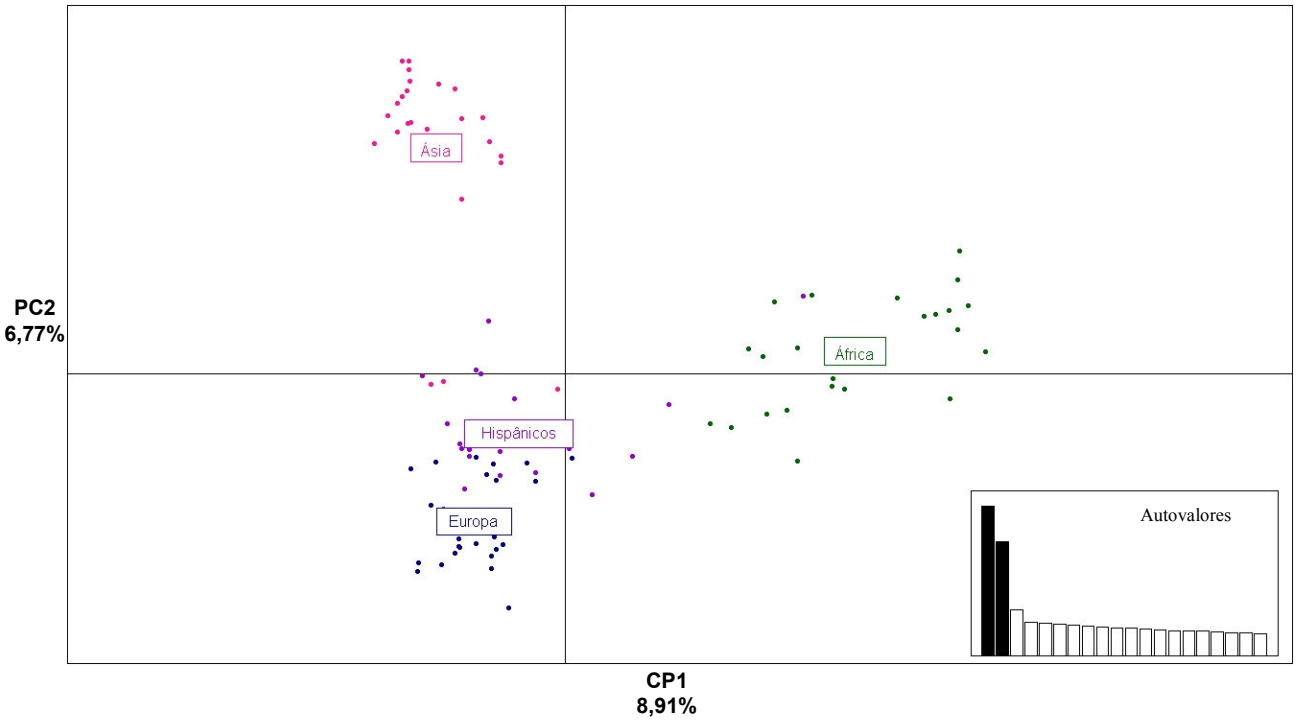


Figura 14: Análise de Componentes Principais na matriz de genótipos formada pelas quatro populações do *SNP500Cancer*. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais.
CP1: Primeiro componente principal; CP2: Segundo componente principal.

V - Discussão

V - 1. Análise populacional: Desvios do EHW - Heterozigosidade esperada e F_{IS}

Em nosso trabalho, além de aplicarmos o teste do EHW para cada SNP individualmente, estimamos o valor de F_{IS} que, ao sintetizar o desvio do EHW dentro da população considerando todos os seus SNPs, reflete os fatores evolutivos que atuam sobre todo o genoma, como a história demográfica da população e os padrões de acasalamento. Essa estatística-*F*, clássica nos estudos de genética de populações, tem sido negligenciada por trabalhos recentes de investigação da estrutura genética humana que utilizam o programa *Structure*, pois este assume que os grupos populacionais em que os indivíduos são distribuídos estão em EHW (Pritchard *et al.* 2000).

A população francesa (HGDP) apresentou vários SNPs com excesso de homozigotos em relação ao EHW e um valor coerentemente elevado para a estimativa F_{IS} (Tabela 1). Os indivíduos franceses estudados são de diferentes regiões da França, segundo informações disponíveis na página eletrônica do projeto (http://www.cephb.fr/en/hgdp/). Devido a esse padrão de amostragem, é provável que o valor de F_{IS} observado deva-se à ocorrência do efeito *Wahlund*, sendo, assim, determinado pelo F_{ST} entre as subpopulações dessas regiões. Provavelmente, as diferenças nas freqüências alélicas são ainda mais acentuadas nos SNPs para os quais houve rejeição do EHW. Na ocorrência do efeito *Wahlund*, o F_{IS} seria mais apropriadamente chamado de F_{IT}, por não se tratar de uma única subpopulação. A população francesa apresentou uma alta diversidade intra-populacional, o que também pode estar relacionado ao padrão de amostragem adotado (Tabela 1). A heterogeneidade da amostra francesa do HGDP também foi notada por Rosenberg *et al.* (2002) que, através do programa *Structure*, verificaram que os franceses apresentam uma mistura de componentes genéticos de diversas populações. A presença de subpopulações crípticas pode também explicar o elevado F_{IS} da população nativo-americana San Martin (Tabela 1). A amostra analisada é heterogênea, pois foi coletada em Pangoa, pequena cidade na encosta oriental dos Andes, habitada por Quechuas e Nomatsiguengas (Fuselli *et al.* 2003).

As populações paquistanesas do centro-sul asiático, Sindhi e Pathan, também apresentaram valores elevados de *Fis* e diversos SNPs com excesso de homozigotos em relação ao EHW (Tabela 1). Os Sindhis, que habitam o sudeste do Paquistão, apresentam uma origem étnica mista (Ansari, 1996; Burton, 1851), enquanto os Pathans habitam o noroeste paquistanês e apresentam herança indefinida: judaica (Ahmad, 1952) ou grega (Caroe, 1992;

Bellew, 1979). Essas populações, assim como as populações Makrani, Balochi e Brahui, para as quais também foram observados valores elevados de F_{IS} (Tabela 1), são de regiões predominantemente mulçumanas do Paquistão, em que os casamentos consangüíneos são comuns (Ayub & Chris Tyler-Smith, 2009; Hussain, 2005; Hussain & Bittles, 1999; Hussain & Bittles, 1998; Yaqoob *et al.* 1998; Bittles *et al.* 1993; Qidwai *et al.* 2003). Assim, conforme esperado, o grupo e as populações do centro-sul asiático apresentaram valores elevados de F_{IS} (Tablelas 1 e 3; Gráficos 2 e 3).

A população nativo-americana Quechua apresentou um número elevado de SNPs com excesso de heterozigotos em relação ao esperado no EHW e um coerente valor de F_{IS} negativo (Tabela 1). Pettener *et al.* (1998), ao investigarem os sobrenomes e a estrutura genética dos Quechuas dos Andes peruanos, encontraram indicativos de uma forte rejeição a casamentos consaguíneos, o que poderia explicar os nossos resultados. Para as populações tribais nativo-americanas, Piapoco e Curripaco (Amazônia colombiana), Suruí e Karitiana (Amazônia brasileira) e Matsiguenga (Monte Carmelo, Peru), e para indígenas da Melanésia, na Oceania, também foram encontrados valores negativos de F_{IS} (Tabela 1). Neel & Ward (1972), ao elucidarem as situações de F_{IS} negativos, destacaram situações que geralmente ocorrem com populações tribais como essas: são populações pequenas, há migração de poucos indivíduos entre as populações e, em muitas delas, há níveis de fertilidade diferenciais entre homens e mulheres, devido a fatores culturais como a poliginia (Hern, 1994; Ingman & Gyllensten, 2003).

A população Beduína apresentou valor elevado de F_{IS}, porém, não apresentou SNPs com freqüências genotípicas fora das proporções de EHW. (Tabela 1). A amostra de beduínos analisada foi coletada em Israel, e se trata dos beduínos de Negev, população em que os casamentos consangüíneos são frequentes (Kisch, 2008; Thein, 2007; Raz & Atar, 2005; Raz *et al.* 2003; Randolph & Coult, 1968). Além disso, na análise populacional realizada em Li *et al.* (2008), partindo-se da mesma amostra do HGDP, os beduínos puderam ser divididos em dois subgrupos, um deles similar aos palestinos. Em nossa análise de ACP, embora essa divisão não tenha sido tão evidente, de fato parte dos beduínos apresentaram-se sobrepostos aos palestinos no eixo do CP1 (Figura 11). Assim, é possível que o tanto a endogamia quanto o efeito *Wahlund* estejam contribuindo para o alto valor de F_{IS}.

Entre as populações do *SNP500Cancer*, a população asiática foi a que apresentou o maior número de SNPs com freqüências genotípicas fora do EHW e o valor de F_{IS} mais elevado, coerentemente com o excesso de homozigotos encontrado (Tabela 2). Os indivíduos do projeto *SNP500Cancer* não são autóctones, apresentando origens étnicas distintas (Packer

et al. 2004). Assim, é bastante provável que esses resultados devam-se, primordialmente, ao efeito *Wahlund*. A heterogeneidade presente nas amostras populacionais do *SNP500Cancer* reflete-se, também, nos altos valores de heterozigosidade esperada observados, relativamente maiores do que os encontrados para as amostras populacionais do HGDP (Tabelas 1 e 2).

V - 2. Diversidade intra-populacional e nos grupos: Heterozigosidade esperada

As populações nativo-americanas (América do Sul e Central) e indígenas da Oceania (Tabela 1; Gráfico 1), assim como os grupos populacionais formados por essas amostras (Tabela 3) apresentaram diversidade genética menor do que as demais regiões estudadas. Resultados semelhantes foram encontrados em outros estudos que também estudaram amostras do HGDP (Rosenberg *et al.* 2002; Li *et al.* 2008; Wang *et al.* 2007). Esse resultado é esperado pelo modelo de evolução humana que sustenta a origem do homem moderno unicamente na África, o "Modelo fora da África": no cenário proposto, teria ocorrido uma rápida expansão do homem anatomicamente moderno para fora da África, acompanhada de uma série de reduções de variadas amplitudes nos tamanhos populacionais, que teriam reduzido a diversidade genética na medida em que as populações se distanciavam da África (Cann *et al.* 1987; Excoffier 2002; Harpending & Cochran, 2002). Esse declínio da diversidade com o aumento da distância da África tem sido observado em vários estudos genéticos (Liu *et al.* 2006; Prugnolle *et al.* 2005; Ramachandran *et al.* 2005), e também é corroborado por evidências arqueológicas e antropológicas (Mellars, 2006).

Entre os nativo-americanos da América do Sul, as populações do leste, Piapoco e Curripaco, Karitiana e Suruí, apresentaram valores de heterozigosidade esperada menores do que as populações do oeste, Cayapa, Martín e Quechua (Tabela 1). Apenas a população Matsiguenga, de Monte Carmelo, no oeste da América do Sul, apresentou valor de heterozigosidade esperada tão baixo quanto o encontrado nas populações do leste. Provavelmente, esse resultado ocorreu devido ao padrão de amostragem adotado na comunidade Monte Carmelo, que incluiu pais e filhos, diminuindo assim a diversidade genética da amostra. A diferença de diversidade genética entre o leste e o oeste da América já havia sido observada em outros estudos (Tarazona-Santos *et al.* 2001; Fuselli *et al.* 2003; Lewis *et al.* 2004; Wang *et al.* 2007) e é atribuída à ação diferencial da deriva genética e do fluxo gênico entre as populações do leste e oeste, conforme o modelo proposto por Tarazona-Santos *et al.* (2001). Esse modelo, baseado primeiramente em dados do cromossomo Y (Tarazona-Santos *et al.* 2001) e DNA mitocondrial (Fuselli *et al.* 2003; Lewis *et al.* 2004),

propõe que: (1) as populações do oeste do continente, associadas à região dos Andes, possuiriam maiores tamanhos efetivos populacionais e maiores níveis de fluxo gênico entre elas, levando a uma maior diversidade genética intra-populacional; enquanto (2) as populações do leste (localizadas na região Amazônica, no Planalto Central brasileiro e na planície do Chaco) possuiriam taxas mais elevadas de deriva genética e níveis mais baixos de fluxo gênico (Tarazona-Santos *et al.* 2001). A ACP realizada para o grupo da América do Sul (Figura 6) é coerente com esse modelo: através dos dois primeiros componentes principais, as populações do leste apresentaram-se bastante diferenciadas entre si e das populações do oeste, enquanto estas se apresentaram sobrepostas e indiferenciadas.

Além disso, devido ao próprio padrão de colonização da América do Sul - que se iniciou no oeste e foi seguido de sub-amostragens das populações do oeste para formar as do leste - as populações do leste apresentariam diversidade genética progressivamente menor, à medida que as migrações se interiorizaram na América (Wang *et al.* 2007). Assim, espera-se que as populações do interior amazônico apresentem heterozigosidade relativamente baixa, como de fato foi observado para Karitiana e Suruí nesse e em outros estudos (Fuselli *et al.* 2003; Harpending & Eller, 1999; Li *et al.* 2008; Hutz *et al.* 2002; Rosenberg *et al.* 2002; Wang *et al.* 2007). Conforme esperado, na ACP realizada para o grupo da América do Sul, essas populações apresentaram-se ainda mais diferenciadas das demais (Figura 6).

Entre as populações da América, a população Maia apresentou o valor mais elevado de diversidade intrapopulacional (Tabela 1). Esse resultado, também observado em outros trabalhos (Rosenberg *et al.* 2002, Harpending & Eller, 1999 e Conrad *et al.* 2006), provavelmente deve-se aos eventos de miscigenação européia e africana (Conrad *et al.* 2006; Zhivotovsky *et al.* 2003; Hellenthal *et al.* 2008), ocorridos em decorrência das migrações pós-Colombianas.

No leste e oeste africanos (Tabela 1; Gráfico 1), bem como na população africana do *SNP500Cancer* (Tabela 2) e nos grupos populacionais da África (Tabela 3), foram encontrados valores baixos de heterozigosidade esperada. Esse resultado contrasta com o geralmente observado em estudos que utilizaram outros conjuntos de marcadores nas amostras do HGDP (Rosenberg *et al.* 2002; Li *et al.* 2008). Uma preocupação potencial acerca de trabalhos que utilizam SNPs é que os seus resultados podem ser enviesados, devido aos procedimentos adotados na escolha dos SNPs que serão genotipados (Clark *et al.* 2005; Mountain & Cavalli-Sforza, 1994; Nielsen, 2004; Rogers & Jorde, 1996). Nesse trabalho, os SNPs foram escolhidos de acordo com dois critérios: (1) foram selecionados dentre SNPs do projeto *SNP500Cancer* e, assim, há um viés em direção a SNPs não-sinônimos e presentes na

região promotora de genes relacionados ao câncer (Packer *et al.* 2004); e (2) eram SNPs bem conhecidos na literatura, por apresentarem grande possibilidade de estar associados a doenças complexas. Ao adotarmos esses critérios, foi criado um viés em direção a SNPs mais freqüentes em populações não-africanas, especialmente as européias, uma vez que a grande maioria dos SNPs é descoberta nessas populações (Rogers & Jorde, 1996). Desse modo, a representação inadequada das amostras africanas nos painéis usados para descoberta de SNPs pode explicar o baixo valor de heterozigosidade esperada observado nas populações desse continente, ao mesmo tempo em que explica os altos valores de heterozigosidade encontrado para as populações européias, tanto do HGDP (Tabela 1), quanto do *SNP500Cancer* (Tabela 2). Esse mesmo viés foi observado em Conrad *et al.* (2006) que, assim como no presente estudo (Tabela 1; Tabela 3), encontraram os maiores valores de heterozigosidade esperada nas amostras da Europa, Oriente Médio e centro-sul asiático. Por outro lado, estudos em que não há viés na escolha dos marcadores identificaram consistentemente as populações africanas como tendo os maiores valores de diversidade genética (Rosenberg *et al.* 2002; Ramachandran *et al.* 2005; Bowcock *et al.* 1994; Crawford *et al.* 2004; Stephens *et al.* 2001), conforme esperado pelo "Modelo fora da África" (Cann *et al.* 1987; Excoffier 2002; Harpending & Cochran, 2002).

As populações européias Adygei (região do Cáucaso, na Rússia) e Rússia (noroeste russo) apresentaram os valores de heterozigosidade esperada mais elevados (Tabela 1). Embora esse resultado deva-se em parte ao viés na amostragem, o estudo de Li *et al.* (2008) demonstrou que Adygei apresenta componente genético significativo da Ásia centro-sul, enquanto Rússia apresenta contribuições menores do centro-sul e leste asiáticos e da Améria.

As populações do grupo leste asiático, formado predominantemente por populações chinesas (Figura 3), apresentaram valores intermediários a altos de heterozigosidade esperada (Tabela 1; Gráfico 1). Como o leste asiático é uma região de fluxo gênico considerável (Chu *et al.* 1998), os níveis de diversidade intra-populacional encontrados são esperados. Dentre as populações desse grupo, Uigur se destacou por apresentar o mais alto valor de heterozigosidade esperada. Os Uigures correspondem a uma população do extremo oeste da China, cuja alta diversidade intra-populacional está relacionada à existência de miscigenação européia e do leste e do oeste asiáticos, conforme diversas evidências antropológicas (Ai *et al.* 1993) e genéticas - DNA mitocondrial (Yao *et al.* 2004), cromossomo Y (Wells *et al.* 2001), inserções *Alu* (Xiao *et al.* 2002) e SNPs (Xu *et al.* 2008; Xu & Jin, 2008).

O grupo formado pelas populações do centro-sul da Ásia apresentou o maior valor de heterozigosidade esperada encontrado (Tabela 3). Os habitantes dessa região geralmente

resultam de eventos de miscigenação entre populações diferenciadas - do leste asiático, Europa, Oriente Médio e do próprio centro-sul asiático - o que produz uma alta diversidade genética (Comas *et al.* 2004; Comas *et al.* 1998; Pérez-Lezaun *et al.* 1999; Wells *et al.* 2001; Karafet *et al.* 2001). Em um estudo recente, Li *et al.* (2008) encontraram componentes genéticos do leste asiático na composição dos indivíduos das populações do centro-sul asiático Burusho, Sindhi e Pathan. Na análise com o programa *Structure*, realizada por Rosenberg *et al.* (2002), também com amostras do HGDP, as populações centro-sul asiáticas Balochi, Makrani, Pathan, Sindhi e Brahui apresentaram componentes genéticos de diversas outras populações, distribuídos em diferentes proporções ao longo dos seus indivíduos. Diferentes componentes genéticos, em proporções individuais distintas, levam a uma alta diversidade intra-populacional, como foi observado para essas populações (Tabela 1; Gráfico 1).

Também devido à miscigenação, a população hispânica (*SNP500Cancer*) apresentou valor elevado de heterozigosidade esperada (Tabela 2). O grau com que cada população parental - nativo-americana, africana e européia - contribui na formação dos hispânicos é bastante variável, sendo influenciado pelo passado histórico e por características locais das populações, como a densidade das populações nativo-americanas na época em que os imigrantes chegaram à América e a quantidade de imigração européia e africana em regiões específicas (Sans 2000; Salzano & Bortolini, 2002; Price *et al.* 2007; Burchard *et al.* 2005; Wang *et al.* 2008). Assim, os hispânicos são bastante heterogêneos ao longo de sua distribuição, de modo que cada amostra dessa população se caracteriza por proporções distintas de componentes genéticos parentais. Ao investigarmos a contribuição parental na amostra hispânica do *SNP500Cancer* através da ACP, O CP1 e o CP2 descreveram semelhanças entre hispânicos e europeus, sugerindo a predominância da ancestralidade européia nessa amostra (Figura 14; Anexo VI). Também foi demonstrada grande influência das populações asiáticas (Figura 14) e nativo-americana (Anexo VI). A ancestralidade asiática provavelmente se deu através dos nativo-americanos, que têm origem no leste asiático (Karafet *et al.* 1999; Forster *et al.* 1996; Pena *et al.* 1995). Apenas o CP2 descreveu semelhanças entre hispânicos e afro-descendentes (Figuras 14; Anexo VI), provavelmente relacionadas à vinda de negros para a América, no período pós-Colombiano, que mais tarde participariam na formação da população hispânica.

V - 3. Análise dos grupos populacionais: Estruturação populacional e estatísticas-*F*

Os maiores valores de F_{ST} foram encontrados para os grupos das populações nativo-americanas (América do Sul e Central) e da Oceania. A grande diferenciação dentro desses grupos é esperada, pois suas populações são relativamente isoladas umas das outras, ocasionando altos níveis de deriva genética (Cavalli-Sforza *et al.* 1994). Outros estudos também encontraram, para América e Oceania, as maiores porcentagens de componentes de variância entre as populações e, correspondentemente, as menores porcentagens para os componentes intra-populacionais (Wang *et al.* 2007; Bastos-Rodrigues *et al.* 2006; Rosenberg *et al.* 2002). As ACPs realizadas para os grupos da Oceania (Figura 4) e da América Central (Figura 5) refletiram os valores de F_{ST}, enquanto a ACP realizada para a América do Sul (Figura 6) demonstrou a maior contribuição das populações do leste para o valor de F_{ST} encontrado.

Os valores encontrados para a estatística-*F* F_{IT}, nessas regiões, também foram os mais elevados, corroborando a previsão de *Wahlund* (1928), posteriormente demonstrada em termos de estatísticas-*F* por Neel & Ward (1972): se há diferenças entre as populações ($F_{ST} \neq 0$), o valor de F_{IT} será positivo. De fato, o valor de F_{IT} estimado foi positivo para todos os grupos analisados e, de um modo geral, foi mais ou menos elevado de acordo com o grau de diferenciação entre as populações do grupo analisado (Tabela 3; Gráfico 3).

Os valores de F_{IS} encontrados para os grupos da América do Sul e Oceania foram negativos (Tabela 3; Gráfico 3), refletindo o padrão proposto por Neel & Ward (1972) para populações tribais relativamente isoladas. Na América Central, o valor de F_{IS} aproxima-se de zero, indicando uma aproximação do EHW nas populações nativas Pima e Maia.

Na ACP realizada para todas as amostras de nativo-americanos (América do Sul e América Central), o CP1 descreveu a variância genética em três subconjuntos: oeste da América do Sul, América Central e leste da América do Sul (Anexo III).

Nos grupos africanos, os valores altos de F_{ST} (Tabela 3; Gráfico 3) corroboram a grande estruturação encontrada no continente africano por Tishkoff *et al.* (2009). Para o oeste africano, o valor de F_{IT} encontrado foi mais elevado do que para o leste, indicando um efeito *Wahlund* mais acentuado. Pela ACP realizada com indivíduos do oeste africano, percebemos que isso provavelmente deve-se à grande diferenciação de San em relação às populações Mandenka e Ioruba (Figura 8). De fato, a ACP realizada para os dois grupos africanos conjuntamente não corroborou a divisão leste/oeste que realizamos (Anexo IV): San (oeste), população caçador-coletora, mostrou-se mais próxima às outras duas populações caçador-

coletoras, Pigmeu Mbuti e Pigmeu Biaka (ambas do leste), conforme a disposição do CP1; além disso, o grupo formado por essas três populações pôde ser separado daquele formado pelas quatro populações de línguas bantas (Bantu NE e Bantu SE e SO, do leste; Mandenka e Ioruba, do oeste), que desenvolveram agricultura cerca de 3000 anos atrás. Esses resultados estão de acordo com os encontrados por Li *et al.* (2008).

Os grupos asiáticos apresentaram valores baixos de F_{ST} (Tabela 3; Gráfico 3), assim como encontrado em outros estudos (Li *et al.* 2008; Rosenberg *et al.* 2002; Bastos-Rodrigues *et al.* 2006). Na ACP realizada para o centro-sul asiático, Kalash e Hazara mostraram-se diferenciadas das demais populações paquistanesas (Figura 9). A diferenciação de Kalash, já demonstrada em outras análises (Rosenberg *et al.* 2002; Bastos-Rodrigues *et al.* 2006; Li *et al.* 2008; Zhivotovsky *et al.* 2003), provavelmente está relacionada a uma acentuada deriva genética (Ayub & Tyler-Smith, 2009), devido ao isolamento geográfico - habitam os vales montanhosos Hindu Kush, no norte do Paquistão - e religioso - não são islâmicos, ao contrário das demais populações paquistanesas estudadas (Parkes, 1990). Quanto à população Hazara, provavelmente sua diferenciação deve-se à grande ancestralidade leste asiática presente nessa população (Ayub & Tyler-Smith, 2009; Li *et al.* 2008). De fato, na ACP realizada com os dois grupos asiáticos conjuntamente (Anexo V), o CP1 descreveu a população Hazara como intermediária entre o leste e o centro-sul asiático, enquanto o CP2 diferenciou Kalash do conjunto formado por esses dois grupos asiáticos.

No grupo leste asiático, formado predominantemente por populações chinesas, o baixo valor de F_{ST} (Gráfico 3; Tabela 3) está relacionado à existência de fluxo gênico considerável entre as populações, já demonstrado por Chu *et al.* (1998). Como consequência, na ACP realizada para esse grupo, a maioria das populações mostraram-se indiferenciadas (Figura 10). Apenas as populações Yakut (Sibéria) e Uigur (China) mostraram-se diferenciadas em relação às demais, o que é esperado devido às localizações geográficas dessas duas populações (Figura 3). Uigures e Yakutes apresentam relações genéticas com o centro asiático (Tian *et al.* 2008; Xu *et al.* 2008; Xu & Jin, 2008) e pertenem a uma mesma família lingüística, a Altaica, o que explica o agrupamento desses indivíduos na ACP (Figura 10), também encontrado por Chu *et al.* (1998). Em decorrência do padrão miscigenado e da localização geográfica da população Uigur, a sua posição no grupo do leste asiático ou do centro-sul asiático é controversa (Li *et al.* 2008). Na ACP que realizamos com todos os indivíduos asiáticos (Anexo V), essa população apresentou uma posição intermediária a esses dois grupos e mostrou bastante semelhança com os Hazaras (Paquistão), conforme encontrado em outras análises genéticas recentes (Li *et al.* 2008; Bastos-Rodrigues *et al.* 2006; Zhivotovsky *et al.*

2003). Embora essas populações não sejam tão próximas geograficamente (Figura 3) e sejam de famílias lingüísticas distintas (Altaica e Indo-europeu), é provável que a posição das mesmas, agrupadas entre os grupos asiáticos, reflita a ancestralidade mongol comum (Zhivotovsky *et al.* 2003).

No grupo do Oriente Médio, o F_{ST} também foi baixo (Tabela 3; Gráfico 3). Na análise da ACP (Figura 11), percebe-se um contínuo na disposição dos indivíduos das populações Beduína, Palestina e Drusa, enquanto a maior contribuição ao F_{ST} e ao F_{IT} parece ser dada pela diferenciação da população norte-africana Mozabite, diferenciada desse contínuo pelo CP1. A posição diferenciada dos Mozabites é esperada, pois, além da contribuição genética do Oriente Médio, os Mozabites apresentam componentes genéticos da Europa, da África e do centro-sul asiático (Rosenberg *et al.* 2002; Li *et al.* 2008). De fato, grande parte da ancestralidade africana presente nas populações não-africanas remete aos Mozabites (Hellenthal *et al.* 2008). Entretanto, esse padrão de variabilidade intermediário - africano/não-africano - pode ser reflexo de uma migração dos Mozabites de volta à África, vindos do Oriente Médio e da Europa (Foster & Romano, 2007; Olivieri *et al.* 2006; Cavalli-Sforza *et al.* 1994). Corroborando essa hipótese, os Mozabites constituem a última população africana formada, e Hellenthal *et al.* (2008) encontraram evidências genéticas de que realmente teria ocorrido um efeito fundador na história dessa população, não compartilhado pelas demais populações africanas.

O grupo que apresentou os valores de F_{ST} e F_{IT} mais baixos foi o europeu (Tabela 3; Gráfico 3). Na ACP realizada para esse grupo, Sardenha e Rússia apresentaram-se um pouco mais diferenciadas, refletindo a localização geográfica dessas amostras (Figuras 3 e 12). É provável que a baixa diferenciação populacional encontrada na Europa deva-se, em parte, a um viés na escolha dos SNPs aqui analisados: como esses SNPs foram descobertos principalmente em populações européias e apresentam freqüências alélicas elevadas nessas populações (próximas a 0,5) a margem para que ocorra diferenciação fica reduzida em relação àquelas populações em que as freqüências alélicas dos SNPs variam com amplitudes maiores.

V - 4. Análise mundial: Estatísticas-F e estruturação genética

Na ACP realizada para todos os grupos analisados, a maior parte da variação (CP1) está relacionada à diferença entre leste e oeste na Eurásia (Figura 13). Os grupos nativos da América Central e América do Sul localizaram-se próximos ao grupo do leste asiático, refletindo a origem histórica dos nativo-americanos. Os africanos locarizaram-se na

extremidade oposta, próximos à Europa, ao Oriente Médio e à Ásia centro-sul, que provavelmente foram as primeiras regiões povoadas após o homem anatomicamente moderno se expandir para fora da África (Cavalli-Sforza *et al.* 1994).

O CP2 descreveu o contraste entre os grupos africanos subsaarianos e os não-africanos (Figura 13-A). Geralmente, em estudos que realizam análise global de variância genética, esse contraste é observado pelo CP1 (Li *et al.* 2008; Tishkoff *et al.* 2009). Isso é o esperado, pois, ao considerarmos que a maior diversidade genética humana está na África (Liu *et al.* 2006; Prugnolle *et al.* 2005; Ramachandran *et al.* 2005), espera-se que a maior parte da variância encontrada nas populações humanas esteja relacionada à diferenciação entre africanos/não-africanos. Entretanto, como no nosso estudo os SNPs foram escolhidos de forma enviesada, com tendência para aqueles mais freqüentes na Eurásia, a maior parte da diversidade genética que encontramos está nessa região (Tabela 3). Os grupos América (Central e do Sul) e Oceania foram diferenciados, respectivamente, pelo PC3 e PC4 (Figuras 13-B e 13-C), o que é esperado, uma vez que estes grupos distinguem-se dos demais devido aos baixos valores de diversidade genética.

Nossos resultados corroboram a ausência de bases genéticas que sustentem a existência de raças humanas (Templeton, 1999; Bamshad *et al.* 2004), pois as diferenças entre os grupos estudados (F_{CT} = 0,107) - definidos primordialmente a partir de definições étnico-culturais e da localização geográfica - bem como as diferenças entre as populações humanas (F_{ST} = 0,121), representaram apenas pequenas parcelas da diversidade genética global presente na nossa espécie. O valor estimado para F_{ST} está de acordo com os geralmente encontrados na literatura, entre 0,10 e 0,15 (Lewontin 1972; Nei & Roychoudhury, 1982; Latter, 1980; Bowcock *et al*., 1991; Barbujani *et al.* 1997; Jorde *et al.* 2000; Watkins *et al.* 2003; Altshuler *et al.* 2005; Rosenberg *et al.* 2005). O F_{CT} também está próximo aos valores dos componentes de variância genética entre regiões 9.6%, 8.9% e 12,1%, observados, respectivamente, por Watkins *et al.* (2003), Romualdi *et al.* (2002) e Bastos-Rodrgues *et al.* (2006).

Entretanto, esses valores de F_{CT} são bem mais elevados do que o observado (3,6%) por Rosenberg *et al.* (2002), ao estudar microssatélites do tipo STR (repetições curtas em tandem) em amostras do HGDP. Excoffier & Hamilton (2003) atribuíram essa diferença a um artefato devido ao modelo de mutação de microssatélites adotado por Rosenberg *et al.* (2002), que não considera a possibilidade de homoplasia. De fato, segundo alguns estudos, não considerar a homoplasia pode levar à subestimação de componentes de variância genética entre grupos populacionais (Flint *et al.* 1999; Romualdi *et al.* 2002). Rosenberg *et al.* (2003), em resposta,

defenderam o modelo adotado e atribuíram o baixo valor obtido ao esquema de amostragem do HGDP. Entretanto, no presente estudo, assim como em Bastos-Rodrigues *et al.* (2006), embora também tenham sido utilizadas amostras do HGDP, foram encontrados valores mais elevados do que o observado por Rosenberg *et al.* (2002). Desse modo, pode-se supor que a subestimação no valor observado por Rosenberg *et al.* (2002) deva-se mais ao modelo de mutação adotado para os microssatélites, do que ao esquema de amostragem do HGDP.

VI - Conclusão

No presente estudo, adicionamos amostras de quatro populações nativo-americanas ao analisarmos as amostras do HGDP. Por sua vez, as amostras do HGDP se distinguem das do *International HapMap Project* (The International HapMap Consortium, 2003) - outro projeto que estuda a diversidade genética humana em larga escala - por incluírem populações nativo-americanas e da Oceania, assim como uma densa amostragem do Oriente Médio e do centro sul asiático. A caracterização de todas essas amostras adicionais é importante para o estudo dos fatores evolutivos e da incidência de doenças tanto nessas populações, como também naquelas que compartilham ancestralidade com essas, devido a eventos de migração recentes. Este é o caso da população hispânica, aqui analisada a partir da amostra do *SNP500Cancer*. Embora a amostragem do HGDP não seja uma amostra aleatória de todas as populações mundiais, uma vez que algumas regiões (como China e Paquistão) são mais densamente amostradas do que outras (como a África, América e Oceania), a riqueza amostral desse projeto permitiu que se encontrassem relações entre variação genética e fatores locais, como cultura e religião, além de indicar a presença de eventos de miscigenação recente.

A utilização de medidas clássicas de variação genética - como as estatísticas-*F* de Wright, o teste do equilíbrio de Hardy-Weinberg e a estimativa da heterozigosidade esperada - proporcionou a realização de inferências relevantes acerca de quais fatores evolutivos influenciam a estrutura da variação genética dentro e entre as populações. Os valores de F_{IS}, superiores a 0,05 em algumas populações, evidenciaram que o endocruzamento pode ser um fator evolutivo importante nas populações humanas, e que não deveria ser ignorado em estudos sobre a estrutura genética das populações. A aplicação da também clássica metodologia ACP, na investigação da variação genética humana, forneceu evidências de eventos de migração e de miscigenação importantes, que permitiram a realização de inferências históricas plausíveis. Os dados genéticos complementam e corroboram os dados da arqueologia, lingüística e geografia, levando a uma compreensão mais detalhada da atuação das forças evolutivas em diferentes regiões do mundo e, assim, ao conhecimento da história das populações humanas.

VII - Referências bibliográficas:

AHMAD A. K. N. (1952) Jesus in heaven on earth. The Civil and Military Gazette Ltd, Lahore, Pakistan.

AI Q., XIAO H., ZHAO J., XU Y., SHI F. (1993) A survey on physical characteristics of Uigur Nationality. *ACTA Anthropologica Sinica* 12:357-365.

ALDRICH M. C., SELVIN S., HANSEN H. M., BARCELLOS L. F., WRENSCH M. R., SISON J. D., QUESENBERRY C. P., KITTLES R. A., SILVA G., BUFFLER P. A., SELDIN M. F., WIENCKE J. K. (2008) Comparison of statistical methods for estimating genetic admixture in a lung cancer study of African Americans and Latinos. *Am. J. Epidemiol.* 168:1035-1046.

ALTSHULER D., BROOKS L.D., CHAKRAVARTI A., COLLINS, F.S., DALY, M.J., DONNELLY, P., *International HapMap Consortium* (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.

ANSARI S. S. A. (1996) The Musalman Races Found in Sindh, Baluchistan and Afghanistan. Indus Publications, Karachi, Pakistan.

AUTON A., BRYC K, BOYKO A. R., LOHMUELLER K. E., NOVEMBRE J., REYNOLDS A., INDAP A., WRIGHT M. H., DEGENHARDT J. D., GUTENKUNST R. N., KING K. S., NELSON M. R., BUSTAMANTE C. D. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* 19:795-803.

AYUB Q., TYLER-SMITH C. (2009) Genetic variation in South Asia: assessing the influences of geography, language and ethnicity for understanding history and disease risk. *Briefings in Functional Genomics and Proteomics* June:1-10.

BALDING D. J., NICHOLS R. A. (1994) DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* 64:125–140.

BALDING D. J., NICHOLS R. A. (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96:3–12.

BAMSHAD M., WOODING S., SALISBURY B. A., STEPHENS J. C. (2004) Deconstructing the relationship between genetics and race. *Nature Rev. Genet.* 5:598-609.

BARBUJANI G., DI BENEDETTO G. (2001) Genetic variances within and between human groups. In: Genes, Fossils and Behaviour (eds. P. Donnelly & R. A. Foley), pp. 63–77. IOS press, Amsterdam.

BARBUJANI G., MAGAGNI A., MINCH E., CAVALLI-SFORZA L. L. (1997) An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci.* 94:4516–4519.

BARKER D. L., THERIAULT G., CHE D., DICKINSON T. SHEN R., KAIN R. (2003) Self-assembled random arrays: High-performance imaging and genomics applications on a high-density microarray platform. *Proc. SPIE* 4966:1-11.

BASTOS-RODRIGUES L., PIMENTA J. R., PENA S. D. J. (2006) The Genetic Structure of Human Populations Studied Through Short Insertion-Deletion Polymorphisms. *Annals of Human Genetics* 70:658–665.

BECKER R., CHAMBERS J., WILKS A. (1998) The New S Language: A Programming Environment for Data Analysis and Graphics. Pacific Grove, Ca.: Wadsworth & Brooks/Cole.

BELLEW H. W. (1979) The races of Afghanistan. Sang-e-Meel Publications, Lahore, Pakistan.

BITTLES A. H., GRANT J. C., SHAMI S. A. (1993) Consanguinity as a determinant of reproductive behavior and mortality in Pakistan. *Int. J. Epidemiol.* 22:463-7.

BOWCOCK A. M., KIDD J. R., MOUNTAIN J. L., HEBERT J. M., CAROTENUTO L., KIDD K. K., CAVALLI-SFORZA L. L. (1991) Drift, Admixture, and Selection in Human Evolution: A Study with DNA Polymorphisms. *Proc. Natl. Acad. Sci.* 85:839–843.

BOWCOCK A.M., RUIZ-LINARES A., TOMFOHRDE J., MINCH E., KIDD J. R., CAVALLI-SFORZA L. L. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455-457.

BURCHARD E. G., BORRELL L. N., CHOUDHRY S., NAQVI M., TSAI H. J., RODRIGUEZ-SANTANA J. R., CHAPELA R., ROGERS S. D., MEI R., RODRIGUEZ-CINTRON W., ARENA J. F., KITTLES R., PEREZ-STABLE E. J., ZIV E., RISCH N. (2005) Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am. J. Public. Health* 95:2161-2168.

BURTON R. F. (1851) Sindh and the races that inhabit the valley of the Indus. WH Allen and Co Ltd, London, England.

CANN H., DE TOMA C., CAZES L., LEGRAND M-F., MOREL V., PIOUFFRE L., BODMER J., BODMER W. F., BONNE-TAMIR B., CAMBON-THOMSEN A., CHEN Z., CHU J., CARCASSI C., CONTU L., DU R., EXCOFFIER L., FERRARA G. B., FRIEDLAENDER J. S., GROOT H., GURWITZ D., JENKINS T., HERRERA R. J., HUANG X., KIDD J., KIDD K. K., LANGANEY A., LIN A. A., MEHDI S. Q., PARHAM P., PIAZZA A., PISTILLO M. P., QIAN Y., SHU Q., XU J., ZHU S., WEBER J. L., GREELY H. T., FELDMAN M. W., THOMAS G., DAUSSET J., CAVALLI-SFORZA L. L. (2002) A Human Genome Diversity Cell Line Panel. *Science* 296:261-262.

CANN R. L., STONEKING M., WILSON, A. C. (1987) Mitochondrial DNA and human evolution. *Nature* 325:31–36.

CAPELLI C., REDHEAD N., ROMANO V., CALI F., LEFRANC G., DELAGUE V., MEGARBANE A., FELICE A. E., PASCALI V. L., NEOPHYTOU P. I., POULLI Z., NOVELLETTO A., MALASPINA P., TERRENATO L., BEREBBI A., FELLOUS M., THOMAS M. G., GOLDSTEIN D. B. (2006) Population structure in the Mediterranean basin: A Y chromosome perspective. *Ann. Hum. Genet.* 70: 207–225.

CAROE O. (1958) The Pathans. Oxford University Press, Karachi, Pakistan.

CAVALLI-SFORZA L. L. (1990) How can one study individual variation for three billion nucleotides of the human genome? *The American Journal of Human Genetics* 46:649-651.

CAVALLI-SFORZA L. L. & EDWARDS A. W. F. (1964) Analysis of human evolution. *Proc. 11th Int. Congr. Genet.* 2:923-933.

CAVALLI-SFORZA L. L., MENOZZI P., PIAZZA, A. (1994) The History and Geography of Human Genes. Princeton Univ. Press, New Jersey.

CHAKRABORTY R., JIN L. (1993) A unified approach to study hypervariable polymorphisms: Statistical considerations of determining relatedness and population distances. In: Pena S., Jeffreys A., Epplen J., Chakraborty R., editors. *DNA fingerprinting, current state of the science.* Basel: Birkhauser. pp. 153–175.

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

CHESSEL D., DUFOUR A. B., THIOULOUSE J. (2004) The ade4 package - I: One-table methods. *R News* 4:5-10.

CHU J. Y., HUANG W., KUANG S. Q., WANG J. M., XU J. J., CHU Z. T., YANG Z. Q., LIN K. Q., LI P., WU M., GENG Z. C., TAN C. C., DU R. F., JIN L. (1998) Genetic relationship of populations in China. *Proc Natl Acad Sci* 95:11763-11768.

CLARK A. G., HUBISZ M. J., BUSTAMANTE C. D., WILLIAMSON S. H., NIELSEN R. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496-1502.

COMAS D., PLAZA S., WELLS R. S., YULDASEVA N., LAO O., CALAFELL F., BERTRANPETIT J. (2004) Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *European Journal of Human Genetics* 12:495-504.

COMAS D, CALAFELL F, MATEU E, PÉREZ-LEZAUN A., BOSCH E., MARTÍNEZ-ARIAS R., CLARIMÓN J., FACCHINI F., FIORI G., LUISELLI D., PETTENER D., BERTRANPETIT J. (1998) Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am. J. Hum. Genet.* 63:1824-1838.

Committee on Human Genome Diversity (1997) National Research Council. Evaluating Human Genetic Diversity US National Academy of Sciences, Washington DC.

CONRAD D. F., JAKOBSSON M., COOP G., WEN X., WALL J. D., ROSENBERG N. A., PRITCHARD J. K. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics* 38:1251-1260.

COX D. G., KRAFT P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.* 61:10-14.

CRAWFORD D.C., CARLSON C. S., RIEDER M. J., CARRINGTON D. P., YI Q., SMITH J. D., EBERLE M. A., KRUGLYAK L., NICKERSON D. A. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* 74:610-622.

DEVLIN B., ROEDER K. (1999) Genomic control for association studies. *Biometrics* 55:997–1004.

DEVLIN B., ROEDER K., BACANU S-A. (2001a) Unbiased methods for population based association studies. *Genet. Epidemiol*. 21:273–284.

DEVLIN B., ROEDER K., WASSERMAN L. (2001b) Genomic control, a new approach to genetic-based association studies. *Theor. Pop. Biol.* 60:155–166.

EWENS W. J., SPIELMAN R. S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 57:455–464.

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

EXCOFFIER L., HAMILTON G. (2003) Comment on "Genetic structure of human populations." *Science* 300:1877b.

FALUSH D., STEPHENS M., PRITCHARD J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587.

FALUSH D., STEPHENS M., PRITCHARD J. K. (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* 7:574–578.

FAN J. B., OLIPHANT A., SHEN R., KERMANI B. G., GARCIA F., GUNDERSON K. L., HANSEN M., STEEMERS F., BUTLER S. L., DELOUKAS P., GALVER L., HUNT S., MCBRIDE C., BIBIKOVA M., RUBANO T., CHEN J., WICKHAM E., DOUCET D., CHANG W., CAMPBELL D., ZHANG B., KRUGLYAK S., BENTLEY D., HAAS J., RIGAULT P., ZHOU L., STUELPNAGEL J., CHEE M. S. (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol.* 68:69-78.

FIRASAT S, KHALIQ S, MOHYUDDIN A, PAPAIOANNOU M., TYLER-SMITH C., UNDERHILL P. A., AYUB Q. (2007) Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. *Eur. J. Hum. Genet.* 15:121-126.

FLINT J., BOND J., REES D. C., BOYCE A. J., ROBERTS-THOMSON J. M., EXCOFFIER L., CLEGG J. B., BEAUMONT M. A., NICHOLS R. A., HARDING R. M. (1999) Minisatellite mutational processes reduce Fst. estimates. *Hum. Genet.* 6, 567-576.

FOREMAN L. A., SMITH A. F. M, EVETT I. W. (1997) Bayesian analysis of DNA profiling data in forensic identification applications. *Journal of the Royal Statistical Society - Series A* 160:429–469.

FORSTER P., HARDING R., TORRONI A., BANDELT H. J. (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–945.

FOSTER M. W., SHARP R. R. (2002) Race, Ethnicity, and Genomics: Social Classifications as Proxies of Biological Heterogeneity. *Genome Res.* 12:844-850.

FOSTER P., ROMANO V. (2007) Timing of a back-migration into Africa. *Science* 316:50-51.

FUSELLI S., TARAZONA-SANTOS E., DUPANLOUP I., SOTO A., LUISELLI D., PETTENER D. (2003) Mitochondrial DNA diversity in South America and the genetic history of Andean Highlanders. *Mol. Biol. Evol.* 20:1682-1691.

GARRIGAN D., HAMMER M. F. (2006) Reconstructing human origins in the genomic era. *Nature Reviews Genetics* 7:669-680.

GARRIGAN D., MOBASHER Z., SEVERSON T., WILDER J. A., HAMMER M. F. (2005) Evidence for archaic Asian ancestry on the human X chromosome. *Mol. Biol. Evol.* 22:189–192.

GONZALEZ-NEIRA A., CALAFELL F., NAVARRO A., LAO O., CANN H., COMAS D., BERTRANPETIT J. (2004) Geographic stratification of linkage disequilibrium: a worldwide population study in a region of chromosome 22. *Hum. Genomics* 1:399-409.

GOUDET J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* 5:184-186.

HAMMER M. F., BLACKMER F., GARRIGAN D., NACHMAN M. W., WILDER J. A. (2003) Human population structure and its effects on sampling Y-chromosome variation. *Genetics* 164:1495–1509.

HAMMER M. F., GARRIGAN D., WOOD E., WILDER J. A., MOBASHER Z., BIGHAM A., KRENZ J. G., NACHMAN M. W. (2004) Heterogeneous patterns of variation among

multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. *Genetics* 167:1841–1853.

HAMMER M. F., SPURDLE A. B., KARAFET T., BONNER M. R., WOOD E. T., NOVELLETTO A., MALASPINA P., MITCHELL R. J., HORAI S., JENKINS T., ZEGURA S. L. (1997) The Geographic Distribution of Human Y Chromosome Variation. *Genetics* 145:787-805.

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.

HARDY G. H. (1908) Mendelian proportions in a mixed population. *Science* 28:49-50.

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.* 95:1961–1967.

HARPENDING H., COCHRAN G. (2002) In our genes. *Proceedings of the National Academy of Sciences* 99:10-12.

HARPENDING H. C., ELLER E. (1999) Human diversity and its history. In The Biology of biodiversity. Ed. Kato M, Takahata N. Tokyo: Springer-Verlag, 1999:301-314.

HAWKS J., HUNLEY K., LEE S. H., WOLPOFF M. (2000) Population bottlenecks and Pleistocene human evolution. *Molecular Biology and Evolution* 17:2:22.

HEDRICK P. W., BLACK F. L. (1997) HLA and Mate Selection: No Evidence in South Amerindians. *Am. J. Hum. Genet.* 61:505–511.

HELLENTHAL G., AUTON A., FALUSH D. (2008) Inferring Human Colonization History Using a Copying Model. *PLoS Genetics* 4(5):e1000078.

HERN W. M. (1994) Cultural change, polygyny, and fertility among the Shipibo of the Peruvian Amazon. *South American Indian Studies* 4:77-86.

HUSSAIN R. (2005) The effect of religious, cultural and social identity on population genetic structure among Muslims in Pakistan. *Annals of Human Biology* 32:145-153.

HUSSAIN R., BITTLES A. H. (1998) The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *J. Biosoc. Sci.* 30:261-275.

HUSSAIN R., BITTLES A. H. (1999) Consanguineous marriage and differentials in age at marriage, contraceptive use and fertility in Pakistan. *J. Biosoc. Sci.* 31:121-138.

HUTZ M. H., CALLEGARI-JACQUES S. M., ALMEIDA S. E. M., ARMBORST T., SALZANO F. M. (2002) Low levels of STRP variability are not universal in American Indians. *Human Biology* 74:791-806.

INGMAN M., GYLLENSTEN U. (2003) Mitochondrial Genome Variation and Evolutionary History of Australian and New Guinean Aborigines. *Genome Res.* 13:1600-1606.

INGMAN M., KAESSMANN H., PAABO S., GYLLENSTEN U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.

JOMBART T., SOLYMOS P. (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

JORDE L. B., WATKINS W. S., BAMSHAD M. J. (2001) Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* 10:2199–207.

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. *The American Journal of Human Genetics* 66:979-988.

KARAFET T., XU L., DU R., WANG W., FENG S., WELLS R. S., REDD A. J., ZEGURA S. L., HAMMER M. F. (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am. J . Hum. Genet.* 69:615-628.

KARAFET T. M., ZEGURA S. L., POSUKH O., OSIPOVA L., BERGEN A., LONG J., GOLDMAN D., KLITZ W., HARIHARA S., DE KNIJFF P., WIEBE V., GRIFFITHS R. C., TEMPLETON A. R., HAMMER M. F. (1999) Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* 64:817-831.

KIDD K. K., PAKSTIS A. J., SPEED W. C., KIDD J. R. (2004) Understanding human DNA sequence variation. *J. Hered.* 95:406–20.

KIDD J. R., PAKSTIS A. J., ZHAO H, LU R. B., OKONOFUA F. E., ODUNSI A., GRIGORENKO E., TAMIR B. B., FRIEDLAENDER J., SCHULZ L. O., PARNAS J., KIDD K. K. (2000) Haplotypes and Linkage Disequilibrium at the Phenylalanine Hydroxylase Locus, PAH, in a Global Representation of Populations. *The American Journal of Human Genetics* 66:1882-1899.

KINNISON M. T., BENTZEN P., UNWIN M. J., QUINN T. P. (2002) Reconstructing recent divergence: evaluating nonequilibrium population structure in New Zealand chinook salmon. *Mol. Ecol.* 11:739–754.

KISCH S. (2008) "Deaf Discourse": The Social Construction of Deafness in a Bedouin Community. *Medical Anthropology: Cross-Cultural Studies in Health and Illness* 27: 283-313.

LATTER B. D. H. (1980) Genetic differences within and between populations of the major human subgroups. *Am. Naturalist* 116:220-237.

LEWIS C. M., JR., TITO R. Y., LIZARRAGA B., STONE A. C. (2005) Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru. *Am J Phys Anthropol* 127:351-360.

LEWONTIN R. C. (1972) The apportionment of human diversity. *Evol. Biol.* 6:381–398.

LI J. Z., ABSHER D. M., TANG H., SOUTHWICK A. M., CASTO A. M., RAMACHANDRAN S., CANN H. M., BARSH G. S., FELDMAN M., CAVALLI-SFORZA L. L., MYERS R. M. (2008) Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* 319:1100–1104.

LIU H., PRUGNOLLE F., MANICA A., BALLOUX F. (2006) A geographically explicit genetic model of worldwide human-settlement history. *The American Journal of Human Genetics* 79:230-237.

LOVELL A., MOREAU C., YOTOVA V., XIAO F., BOURGEOIS S., GEHL D., BERTRANPETIT J., SCHURR E., LABUDA D. (2005) Ethiopia: Between Sub-Saharan Africa and western Eurasia. *Ann. Hum. Genet.* 69:275–287.

LONG J. C. (1986) The allelic correlation structure of Gainj-And Kalam-Speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* 112:629-647.

MAGALHÃES W. C. S., LOBÃO M. I., CAMPOS A. A. F., TARAZONA-SANTOS E. *DIVERGENOMEdb*: a bioinformatics tool to assist the analysis of genetic variation. Em preparação.

MELLARS P. (2006) Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* 313:796–800.

MENOZZI P, PIAZZA A, CAVALLI-SFORZA L. L. (1978) Synthetic maps of human gene frequencies in Europeans. *Science* 201: 786–792.

MOUNTAIN J. L., CAVALLI-SFORZA L. L. (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc. Natl. Acad. Sci.* 91:6515-6519.

NEEL J. V., WARD R. H. (1972) The genetic structure of a tribal population, the Yanomana Indians. VI. Analysis by F-Statistics, including a comparison with the Makiritare and Xavante. *Genetics* 72:639-666.

NEI M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583-590.

NEI M., ROYCHOUDHURY A. K. (1974) Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379-390.

NEI M., ROYCHOUDHURY A. K. (1982) Genetic relationship and evolution of human races. *Evolutionary biology* 14:1-59.

NEIGEL J. E. (2002) Is F_{ST} obsolete? *Conserv. Genet.* 3:167–173.

NIELSEN R. (2004) Population genetic analysis of ascertained SNP data. *Hum. Genomics* 1:218–224.

NIEVERGELT C. M., LIBIGER O., SCHORK N. J. (2007) Generalized Analysis of Molecular Variance. *PLoS Genet.* 3(4):e51.

OLIVIERI A., ACHILLI A., PALA M., BATTAGLIA V., FORNARINO S., AL-ZAHERY N., SCOZZARI R., CRUCIANI F., BEHAR D. M., DUGOUJON J-M., COUDRAY C., SANTACHIARA-BENERECETTI A. S., SEMINO O., BANDELT H-J., TORRONI A. (2006) The mtDNA legacy of the Levantine Early Upper Palaeolithic in Africa. *Science* 314:1767–1770.

PACKER B. R., YEAGER M., STAATS B. WELCH R., CRENSHAW A., KILEY M., ECKERT A., BEERMAN M., MILLER E., BERGEN A., ROTHMAN N., STRAUSBERG R., CHANOCK S. J. (2004) SNP500Cancer: a public resource for sequence validation and

assay development for genetic variation in candidate genes. *Nucleic Acids Res*. Database issue - 32:D617-D621.

PARKES P. (1990) Kalasha Rites of Spring: Backstage of a `Disappearing World' Film. *Anthropology Today* 6:11-13.

PATTERSON N., PRICE A.L., REICH D. (2006) Population structure and eigenanalysis. *PLoS Genet.* 2(12):e190.

PAULING L., ITANO A. H., SINGER S. J., WELLS I. C. (1949) Sickle cell anemia, a molecular disease. *Science* 110:543–548.

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.

PÉREZ-LEZAUN A., CALAFELL F., COMAS D., MATEU E., BOSCH E., MARTÍNEZ-ARIAS R., CLARIMÓN J., FIORI G., LUISELLI D., FACCHINI F., PETTENER D., BERTRANPETIT J. (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y chromosome short tandem repeats and mtDNA. *Am. J. Hum. Genet.* 65:208-219.

PETTENER D., PASTOR S., TARAZONA-SANTOS E. (1998) Surnames and genetic structure of a high-altitude Quechua community from the Ichu River Valley, Peruvian Central Andes, 1825-1914. *Hum. Biol.* 70:865-87.

PRICE A. L., PATTERSON N. J., PLENGE R. M., WEINBLATT M. E., SHADICK N. A., REICH D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38: 904–909.

PRICE A. L., PATTERSON N., YU F., COX D. R., WALISZEWSKA A., MCDONALD G. J., TANDON A., SCHIRMER C., NEUBAUER J., BEDOYA G., DUQUE C., VILLEGAS A., BORTOLINI M. C., SALZANO F. M., GALLO C., MAZZOTTI G., TELLO-RUIZ M., RIBA L., AGUILAR-SALINAS C. A., CANIZALES-QUINTEROS S., MENJIVAR M.,

KLITZ W., HENDERSON B., HAIMAN C. A., WINKLER C., TUSIE-LUNA T., RUIZ-LINARES A., REICH D. (2007) A genomewide admixture map for latino populations. *Am. J. Hum. Genet.* 80:1024-1036.

PRITCHARD J. K., DONNELLY P. (2001) Case-control studies of association in structured or admixed populations. *Theor. Pop. Biol.* 60:227–237.

PRITCHARD J. K., STEPHENS M. & DONNELLY P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

PRUGNOLLE F., MANICA A., BALLOUX F. (2005) Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15:R159–R160.

QAMAR R., AYUB Q., MOHYUDDIN A., HELGASON A., MAZHAR K., MANSOOR A., ZERJAL T., TYLER-SMITH C., MEHDI Q. (2002) Y-Chromosomal DNA Variation in Pakistan *Am. J. Hum. Genet.* 70:1107-1124.

QIDWAI W., SYED I. A., KHAN F. M. (2003) Prevalence and perceptions about consanguineous marriages among patients presenting to family physicians, in 2001 at a Teaching Hospital in Karachi, Pakistan. *Asia Pacific Family Medicine* 2:27–31.

R Development Core Team (2008) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Áustria.

RACE R. R., SANGER R. (1975) Blood Groups in Man. Blackwell Scientific Publishers, Oxford.

RAMACHANDRAN S., DESHPANDE O., ROSEMAN C. C., ROSENBERG N. A., FELDMAN M. W., CAVALLI-SFORZA L. L. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences* 102:15942-15947.

RANDOLPH R. R., COULT A. D. (1968) A Computer Analysis of Bedouin Marriage. *Southwestern Journal of Anthropology* 24:83-99.

RAZ A. E., ATAR M. (2005) Perceptions of Cousin Marriage Among Young Bedouin Adults in Israel. *Marriage & Family Review* 37:27-46.

RAZ A. E., ATAR M., RODNAY M., SHOHAM-VARDI I., CARMI R. (2003) Between Acculturation and Ambivalence: Knowledge of Genetics and Attitudes towards Genetic Testing in a Consanguineous Bedouin Community. *Community Genet.* 6:88-95.

REICH D., PRICE A. L., PATTERSON N. (2008) Principal component analysis of genetic data. *Nat. Genet.* 40:491–492.

ROBERTSON A., HILL W. G. (1984) Deviations from the Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* 107:703-718.

ROEDER K., ESCOBAR M., KADANE J. B., BALAZS I. (1998) Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika* 85:269–287.

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

ROMUALDI, C., BALDING, D., NASIDZE, I. S., RISCH G., ROBICHAUX M., SHERRY S. T., STONEKING M., BATZER M. A., BARBUJANI G. (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* 12:602-612.

ROSENBERG N. A., MAHAJAN S., RAMACHANDRAN S., ZHAO C., PRITCHARD J. K., FELDMAN M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1:660-671.

ROSENBERG N. A., PRITCHARD J. K., WEBER J. L., CANN H. M., KIDD K. K., ZHIVOTOVSKY L. A., FELDMAN M. W. (2002) Genetic structure of human populations. *Science* 298:2381–2385.

SABETI P. C., VARILLY P., FRY B., LOHMUELLER J., HOSTETTER E., COTSAPAS C., XIE X., BYRNE E. H., MCCARROLL S. A., GAUDET R., SCHAFFNER S. F., LANDER E. S., *International HapMap Consortium et al.* (2007). Genome-wide detection and chracterization of positive selection in human populations. *Nature* 449:913-918.

SALZANO F. M., BORTOLINI M. C. (2002) The evolution and genetics of Latin American populations. Cambridge University Press, Cambridge, United Kingdom.

SANS M. (2000) Admixture studies in Latin America: from the 20th to the 21st century. *Hum. Biol.* 72:155-177.

SARKAR D. (2002) Lattice. *R News* 2:19-23.

SCHMID K. J., TÖRJÉK O., MEYER R., SCHMUTHS H., HOFFMANN M. H., ALTMANN T. (2006) Evidence for a large-scale population structure of Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers. *Theor. Appl. Genet.* 112:1104-1114.

SERRE D., PAABO S. (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res.*14:1679-85.

SETAKIS E., STIRNADEL H., BALDING D. J. (2006) Logistic regression protects against population structure in genetic association studies. *Genome Res.* 16:290-296.

SHEN R., FAN J. B., CAMPBELL D., CHANG W., CHEN J., DOUCET D., YEAKLEY J., BIBIKOVA M., WICKHAM GARCIA E., MCBRIDE C., STEEMERS F., GARCIA F., KERMANI B. G., GUNDERSON K., OLIPHANT A. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutat. Res.* 573:70-82.

SHRIVER M., MEI R., PARRA E., SONPAR V., HALDER I., TISHKOFF S. A., SCHURR T. G. ZHADANOV S. I., OSIPOVA L. P., BRUTSAERT T. D., FRIEDLAENDER J., JORDE L. J., WATKINS W. S., BAMSHAD M. J., GUTIERREZ G., LOI H., MATSUZAKI H., KITTLES R. A., ARGYROPOULOS G., FERNANDEZ J. R., AKEY J. M., JONES K.

W. (2005) Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Human Genomics* 2:81–89.

SLATKIN M., BARTON N. H. (1989) Methods for estimating gene flow. *Evolution* 43:1349–1368.

STEPHENS, J.C., SCHNEIDER J. A., TANGUAY D. A., CHOI J., ACHARYA T., STANLEY S. E., JIANG R., MESSER C. J., CHEW A., HAN J-H., DUAN J., CARR J. L., LEE M. S., KOSHY B., KUMAR A. M., ZHANG G., NEWELL W. R., WINDEMUTH A., XU C., KALBFLEISCH T. S., SHANER S. L., ARNOLD K., SCHULZ V., DRYSDALE C. M., NANDABALAN K., JUDSON R. S., RUAÑO G., VOVIS G. F. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293:489-493.

STONEKING M., FONTIUS J.J., CLIFFORD S.L., SOODYALL H., ARCOT S.S., SAHA N., JENKINS T., TAHIR M. A., DEININGER P. L., BATZER M. A. (1997) *Alu* insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Res.* 7:1061–1071.

STRANGER B. E., NICA A. C., FORREST M. S., DIMAS A., BIRD C. P., BEAZLEY C., INGLE C. E., DUNNING M., FLICEK P., KOLLER D., MONTGOMERY S., TAVARÉ S., DELOUKAS P., DERMITZAKIS E. T. (2007) Population genomics of human gene expression. *Nat. Genet.* 39:1217-24.

STRAUSBERG R. L., BUETOW K. H., EMMERT-BUCK M. R., KLAUSNER R. D. (2000) The Cancer Genome Anatomy Project: building an annotated gene index. *Trends in Genetics* 16:103-106.

STRINGER C. B., ANDREWS P. (1988) Genetic and fossil evidence for the origin of modern humans. *Science* 239:1263–1268.

TAKAHATA N., LEE S. H., SATTA Y. (2001) Testing multiregionality of modern human origins. *Mol. Biol. Evol.* 18:172–183.

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.

TEMPLETON A. R. (1997) Out of Africa? What do genes tell us? *Current Opinion in Genetics & Development* 7:841-847.

TEMPLETON A. R. (1999) Human races: A genetic and evolutionary perspective. *Am. Anthropol.* 100:632–650.

The International HapMap Consortium (2003) The International HapMap Project *Nature* 426:789-795.

THOMAS D. C., WITTE J. S. (2002) Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? *Cancer Epidemiol. Biomark. Prev.* 11:505-512.

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. *Proceedings of the National Academy of Sciences* 97:7360-7365.

THORNE A., WOLPOFF M. (2003) The multiregional evolution of humans. *Scientific American* Special Editions.

TIAN C., KOSOY R., LEE A., RANSOM M., BELMONT J. W., GREGERSEN P. K., SELDIN M. F. (2008) Analysis of East Asia genetic substructure using genome-wide SNP arrays *PLoS ONE* 3(12):e3862.

TISHKOFF S. A., GOLDMAN A., CALAFELL F., SPEED W. C., DEINARD A. S., BONNE-TAMIR B., KIDD J. R., PAKSTIS A. J., JENKINS T., KIDD K. K. (1998) A Global Haplotype Analysis of the Myotonic Dystrophy Locus: Implications for the Evolution

of Modern Humans and for the Origin of Myotonic Dystrophy Mutations. *The American Journal of Human Genetics* 62:1389-1402.

TISHKOFF S. A., PAKSTIS A. J., STONEKING M., KIDD J.R., DESTRO-BISOL G., SANJANTILA A., LU R. B., DEINARD A. S., SIRUGO G., JENKINS T., KIDD K. K., CLARK A. G. (2000) Short Tandem-Repeat Polymorphism/Alu Haplotype Variation at the PLAT Locus: Implications for Modern Human Origins. *The American Journal of Human Genetics* 67: 901-925.

TISHKOFF S. A., REED F. A., FRIEDLAENDER F. R., EHRET C., RANCIARO A., FROMENT A., HIRBO J. B., AWOMOYI A. A., BODO J. M., DOUMBO O., IBRAHIM M., JUMA A. T., KOTZE M. J., LEMA G., MOORE J. H., MORTENSEN H., NYAMBO T. B., OMAR S. A., POWELL K., PRETORIUS G. S., SMITH M. W., THERA M. A., WAMBEBE C., WEBER J. L., WILLIAMS S. M. (2009) The genetic structure and history of Africans and African Americans. *Science* 324:1035-1044.

TISHKOFF S. A., VERRELLI B. C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* 4:293-340.

TODD J. A., WALKER N. M., COOPER J. D., SMYTH D. J., DOWNES K., PLAGNOL V., BAILEY R., NEJENTSEV S., FIELD S. F., PAYNE F., LOWE C. E., SZESZKO J. S. *et al.* (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* 39:857-864.

UNDERHILL P. A., SHEN P., LIN A. A., JIN L., PASSARINO G., YANG W. H., KAUFFMAN E., BONNÉ-TAMIR B., BERTRANPETIT J., FRANCALACCI P., IBRAHIM M., JENKINS T., KIDD J. R., MEHDI S. Q., SEIELSTAD M. T., WELLS R. S., PIAZZA A., DAVIS R. W., FELDMAN M. W., CAVALLI-SFORZA L. L., OEFNER P. J. (2000) Y chromosome sequence variation and the history of human populations. *Nature Genetics* 26:358–361.

VIGILANT L, STONEKING M, HARPENDING H, HAWKES K, WILSON AC. (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507.

WAHLUND, S. (1928) The combination of populations and the appearance of correlation examined from the standpoint of the study of heredity. *Hereditas* 11:65–106.

WANG S., LEWIS C.M., JAKOBSSON M., RAMACHANDRAN S., RAY N., BEDOYA G., ROJAS W., PARRA M. V., MOLINA J. A., GALLO C., MAZZOTTI G., POLETTI G., HILL K., HURTADO A. M., LABUDA D., KLITZ W., BARRANTES R., BORTOLINI M. C., SALZANO F. M., PETZL-ERLER M. L., TSUNETO L. T., LLOP E., ROTHHAMMER F., EXCOFFIER L., FELDMAN M. W., ROSENBERG N. A., RUIZ-LINARES A. (2007) Genetic variation and population structure in Native Americans. *PLoS Genet.* 3:e185.

WANG S., RAY N., ROJAS W., PARRA M. V., BEDOYA G., GALLO C., POLETTI G., MAZZOTTI G. HILL K., HURTADO A. M., CAMRENA B., NICOLINI H., KLITZ W., BARRANTES R., MOLINA J. A., FREIMER N., BORTOLINI M. C., SALZANO F. M., PETZL-ERLER M. L., TSUNETO L. T., DIPIERRI J. E., ALFARO E. I., BAILLIET G., BIANCHI N. L., LLOP E., ROTHAMMER F., EXCOFFIER L., RUIZ-LINARES A. (2008) Geographic patterns of genome admixture in Latin American mestizos. *PLoS Genet.* 4(3):e1000037.

WARNES G. R., GORJANC G., LEISCH F., MAN M. (2003) The genetics package: Population Genetics. *R News* 3:9-13.

WATKINS W. S., ROGERS A. R., OSTLER C. T., WOODING S., BAMSHAD M. J., BRASSINGTON A-M. E., CARROLL M. L, NGUYEN S. V., WALKER J. A., PRASAD B.V. R., REDDY P. G., DAS P. K., BATZER M. A., JORDE L. B. (2003) Genetic variation among world populations: Inferences from 100 *Alu* insertion polymorphisms. *Genome Res.* 13:1607-1618.

WATSON J. D. (1990) The human genome project: past, present, and future. *Science* 248:44-49.

WEINBERG W. (1908) U¨ ber den Nachweis der Vererbung beim Menschen. Jh. Ver. vaterl. Naturk. *Wu¨rttemb.* 64:369–382.

WEIR B. S., CARDON L. R., ANDERSON A. D., NIELSEN D. M., HILL W. G. (2005) Measures of human population structure show heterogeneity among genomic regions. 315:1468–1476.

WEIR B. S., COCKERHAM C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

WEIR B. S., HILL W. G. (2002) Estimating F-statistics. *Annu. Rev. Gen.* 36:721–750.

WEISS, K. M., CLARK A. G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18:19–24.

WELLS R. S., YULDASHEVA N., RUZIBAKIEV R., UNDERHILL P. A., EVSEEVA I., BLUE-SMITH J., JIN L., SU B., PITCHAPPAN R., SHANMUGALAKSHMI S., BALAKRISHNAN K., READ M., PEARSON N. M., ZERJAL T., WEBSTER M. T., ZHOLOSHVILI I., JAMARJASHVILI E., GAMBAROV S., NIKBIN B., DOSTIEV A., AKNAZAROV O., ZALLOUA P., TSOY I., KITAEV M., MIRRAKHIMOV M., CHARIEV A., BODMER W. F. (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl. Acad. Sci.* 98:10244-10249.

WHITLOCK M., MCCAULEY D. E. (1999) Indirect estimation of gene flow and migration: FST≠1/(4Nm+1). *Heredity* 82:117–125.

WITHERSPOON D. J., MARCHANI E. E., WATKINS E. S., OSTLER C. T., WOODING S. P., ANDERS B. A., FOWLKES J. D., BOISSINOT S., FURANO A. V., RAY D. A., ROGERS A. R.,. BATZER M. A, JORDE L. B. (2006) Human Population Genetic Structure and Diversity Inferred from Polymorphic L1 (LINE-1) and *Alu* Insertions. *Hum. Hered.* 62:30-46.

WITTKE-THOMPSON J. K., PLUZHNIKOV A., COX N. J. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *The American Journal of Human Genetics* 76:967-986.

WOLPOFF M. H., HAWKS J., CASPARI R. (2000) Multiregional, not multiple origins. *Am. J. Phys. Anthropol* 112:129–136.

WOLPOFF M., THORNE A. G., SMITH F. H., FRAYER D. W., POPE G. G. (1994) Multiregional evolution: a world-wide source for modern human populations. In Origins of anatomically modern humans. *New York Plenum Press* 175–200.

WOODING S. P., WATKINS W. S., BAMSHAD M. J., DUNN D. M., WEISS R. B., JORDE L. B. (2002) DNA Sequence Variation in a 3.7-kb Noncoding Sequence 5' of the CYP1A2 Gene: Implications for Human Population History and Natural Selection. *The American Journal of Human Genetics* 71:528–542.

WRIGHT S. (1951) The genetical structure of populations. *Ann. Eugen.* 15: 323–354.

WRIGHT S. (1956) Gene and Organism. *American Naturalist* 87:5-18.

WRIGHT S. (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* 19:395-420.

XIAO F. X., YANG J. F., CASSIMAN J. J., DECORTE R. (2002) Diversity at eight polymorphic Alu insertion loci in Chinese populations shows evidence for European admixture in an ethnic minority population from northwest China. *Hum. Biol.* 74:555-568.

XING J, WATKINS WS, WITHERSPOON DJ, ZHANG Y, GUTHERY SL, THARA R, MOWRY BJ, BULAYEVA K, WEISS RB, JORDE LB. (2009) Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Res.* 19:815-25.

XU S., HUANG W., QIAN J., JIN L. (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am. J. Hum. Genet.* 82:883-894.

XU S., JIN L. (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *Am J Hum Genet.* 83:322-336.

YANG R. C. (1998) Estimating hierarchical f-statistics. *Evolution* 52:950-956.

YAO Y. G., KONG Q. P., WANG C. Y., ZHU C. L., ZHANG Y. P. (2004) Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in china. *Mol. Biol. Evol.* 21:2265-2280.

YAQOOB M., CNATTINGIUS S., JALIL F., ZAMAN S., ISELIUS L., GUSTAVSON K. H. (1998) Risk factors for mortality in young children living under various socio-economic conditions in Lahore, Pakistan: with particular reference to inbreeding. *Clin. Genetics* 54:426-34.

YU N., ZHAO Z., FU Y-X, SAMBUUGHIN N., RAMSAY M., JENKINS T., LESKINEN E., PATTHY L., JORDE L. B., KUROMORI T., LI W-H. (2001) Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* 18:214–222.

ZHAO Z., JIN L., FU, Y-X, RAMSAY M., JENKINS T., LESKINEN E., PAMILO P., TREXLER M., PATTHY L., JORDE L. B., RAMOS-ONSINS S., YU N., LI W-H. (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Natl Acad. Sci.* 97:11354–11358.

ZHIVOTOVSKY L. A., ROSENBERG N. A., FELDMAN M. W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* 72:1171–1186.

Sítios eletrônicos acessados:

Centre d'Etude du Polymorphisme Humain (2009) Human Genome Diversity Population Groups. Versão 24/07/2008. Disponível em http://www.cephb.fr/en/hgdp/. Acesso: 10/08/2009.

The R Project for Statistical Computing (2009) About R. Disponível em http://www.r-project.org/. Acesso: 07/08/2009.

BIOCONDUCTOR (2009) Open source software for bioinformatics. Disponível em http://www.bioconductor.org/. Acesso: 22/08/2009.

ANEXO I

Lista de SNPs analisados nesse estudo e os genes a que pertencem. Os SNPs utilizados na realização da ACP estão marcados com asteriscos.

Gene	SNP	Gene	SNP	Gene	SNP
ABCA1	ABCA1_04*	**AKR1C4**	AKR1C4_01*	**ARHGDIB**	rs2075267*
	ABCA1_12	**AKT1**	AKT1_15*		ARHGDIB_01
	ABCA1_15	**ALAD**	ALAD_01*		ARHGDIB_03
	ABCA1_17		ALAD_03	**ARNT**	rs1889740*
	ABCA1_26		ALAD_10		rs2228099
	ABCA1_31	**ALDH1L1**	ALDH1L1_01*		rs2256355
ABCA5	ABCA5_01*		ALDH1L1_03		rs2864873
ABCA6	ABCA6_01*		ALDH1L1_06		rs7517566
	ABCA6_05	**ALDH2**	ALDH2_08*		ARNT_01
ABCA7	ABCA7_05*	**ALOX12**	ALOX12_02*		ARNT_05
	ABCA7_06	**ALOX15**	ALOX15_02*		ARNT_06
ABCB1	ABCB1_01*		ALOX15_12		ARNT_10
	ABCB1_09	**ALOX5**	ALOX5_02*	**ARVCF**	rs2240716*
	ABCB1_12		ALOX5_06	**ATM**	rs170548*
ABCB11	ABCB11_02*		ALOX5_10		rs3092993
	ABCB11_08		ALOX5_12		rs664143
ABCC2	ABCC2_01*		ALOX5_15		ATM_01
	ABCC2_02		ALOX5_26		ATM_02
	ABCC2_03		ALOX5_28		ATM_03
	ABCC2_10	**AMACR**	AMACR_01*		ATM_06
ABCC4	ABCC4_04*		AMACR_02		ATM_27
	ABCC4_07		AMACR_03	**ATP1B2**	ATP1B2_01*
ABCG8	ABCG8_01*		AMACR_05		ATP1B2_04
	ABCG8_02		AMACR_08		ATP1B2_13
	ABCG8_06		AMACR_09	**AURKA**	AURKA_02*
ADH1C	ADH1C_01*		AMACR_17		AURKA_03
	ADH1C_15	**ANKK1**	ANKK1_01*		AURKA_04
	ADH1C_16	**APAF1**	APAF1_03*		AURKA_06
	ADH1C_18		APAF1_04		AURKA_08
AHR	rs2074113*		APAF1_07		AURKA_15
	rs7796976		APAF1_09		AURKA_16
	AHR_01	**APC**	APC_03*	**AXIN2**	rs11867417*
AHRR	AHRR_02*		APC_09		rs11868547
	AHRR_10		APC_13		rs2240308
AKR1A1	AKR1A1_02*		APC_19		rs3923087
AKR1C3	AKR1C3_01*		APC_26		rs4128941
	AKR1C3_08	**APEX1**	APEX1_03*		rs4541111
	AKR1C3_11		APEX1_09		rs7210356
	AKR1C3_17		APEX1_16	**B9D2**	B9D2_03*
	AKR1C3_19	**APOA2**	APOA2_02*	**BAK1**	BAK1_05*
	AKR1C3_21		APOA2_04		BAK1_06
	AKR1C3_24		APOA2_06		BAK1_07
	AKR1C3_26	**APOA4**	APOA4_02*	**BARD1**	BARD1_02*
	AKR1C3_28		APOA4_07		BARD1_04
	AKR1C3_29	**APOB**	APOB_01*		BARD1_11
	AKR1C3_30		APOB_04		BARD1_18
	AKR1C3_31		APOB_07		BARD1_22
	AKR1C3_33		APOB_08	**BAX**	BAX_03*
	AKR1C3_35		APOB_21		BAX_05
	AKR1C3_36	**APOE**	APOE_03*	**BCL2L1**	BCL2L1_01*

Gene	SNP	Gene	SNP	Gene	SNP
BCL2L1	BCL2L1_02	**BRIP1**	BRIP1_03	**CCDC97**	CCDC97_03*
	BCL2L1_03		BRIP1_05	**CCL5**	CCL5_03*
BCL6	BCL6_05*		BRIP1_09		CCL5_04
	BCL6_06		BRIP1_15	**CCNA2**	CCNA2_01*
	BCL6_07	**C11orf65**	C11orf65_02*		CCNA2_06
	BCL6_09	**CALCR**	CALCR_01*		CCNA2_12
	BCL6_11		CALCR_03	**CCND1**	CCND1_01*
BCR	BCR_01*	**CARD15**	CARD15_04*		CCND1_02
	BCR_02		CARD15_05		CCND1_03
BHMT	BHMT_01*		CARD15_09	**CCND3**	CCND3_01*
	BHMT_02		CARD15_10		CCND3_02
	BHMT_04		CARD15_19	**CCNH**	CCNH_01*
BIC	BIC_01*	**CASP10**	CASP10_02*		CCNH_04
	BIC_04	**CASP3**	CASP3_02*	**CCR2**	CCR2_01*
	BIC_07		CASP3_07		CCR2_02
	BIC_10		CASP3_08		CCR2_06
	BIC_11		CASP3_09	**CCR3**	CCR3_01*
	BIC_15	**CASP8**	CASP8_06*		CCR3_05
	BIC_21		CASP8_07	**CD14**	CD14_03*
	BIC_32		CASP8_22	**CD4**	CD4_03*
	BIC_33	**CASP9**	CASP9_01*	**CD40**	CD40_01*
	BIC_34		CASP9_03		CD40_03
BIRC2	BIRC2_01*		CASP9_27	**CD80**	CD80_01*
BIRC3	BIRC3_02*	**CASR**	CASR_01*		CD80_02
	BIRC3_03		CASR_05		CD80_04
BLM	BLM_02*		CASR_06	**CD81**	CD81_04*
	BLM_03		CASR_07		CD81_06
	BLM_05		CASR_09	**CD86**	CD86_02*
	BLM_06		CASR_11		CD86_03
	BLM_16		CASR_15	**CDC25A**	CDC25A_04*
	BLM_22	**CAT**	rs1049982*	**CDC25B**	CDC25B_06*
	BLM_25		CAT_02	**CDC25C**	CDC25C_01*
BPI	rs1131847*		CAT_03	**CDH1**	CDH1_06*
BRCA1	BRCA1_01*		CAT_05		CDH1_09
	BRCA1_05		CAT_06	**CDK4**	CDK4_01*
	BRCA1_06		CAT_07	**CDK5**	CDK5_08*
	BRCA1_18	**CAV1**	CAV1_02*		CDK5_16
	BRCA1_20		CAV1_05	**CDK7**	CDK7_01*
	BRCA1_21		CAV1_07	**CDKN1B**	CDKN1B_04*
	BRCA1_26		CAV1_09	**CDKN1C**	CDKN1C_09*
	BRCA1_32		CAV1_19	**CDKN2A**	CDKN2A_03*
BRCA2	BRCA2_01*		CAV1_23		CDKN2A_09
	BRCA2_02		CAV1_29		CDKN2A_11
	BRCA2_03	**CBR1**	CBR1_01*		CDKN2A_12
	BRCA2_04		CBR1_10		CDKN2A_13
	BRCA2_06		CBR1_11		CDKN2A_14
	BRCA2_25	**CBR3**	CBR3_01*		CDKN2A_16
	BRCA2_32	**CBS**	CBS_01*		CDKN2A_18
BRIP1	BRIP1_01*		CBS_03		CDKN2A_19
	BRIP1_02		CBS_07		CDKN2A_20

Gene	SNP	Gene	SNP	Gene	SNP
CETP	CETP_08*	**CTNNB1**	rs11564437	**CYP1B1**	CYP1B1_28
	CETP_21		rs11564452		CYP1B1_31
	CETP_23		rs11564465	**CYP24A1**	CYP24A1_01*
CFH	CFH_01*		rs1798794		CYP24A1_03
	CFH_03		rs1880481		CYP24A1_05
	CFH_05		rs2371452		CYP24A1_08
	CFH_06		rs2953	**CYP2C19**	CYP2C19_03*
	CFH_07		rs3864004		CYP2C19_08
CG018	CG018_03*		rs4135385	**CYP2D6**	CYP2D6_65*
CGA	CGA_02*		rs4533622	**CYP2E1**	CYP2E1_02*
	CGA_03		rs5743395		CYP2E1_31
	CGA_05		rs9813198	**CYP3A4**	CYP3A4_57*
	CGA_06		CTNNB1_21	**CYP3A7**	CYP3A7_01*
CHEK1	CHEK1_01*	**CTSB**	CTSB_03*	**CYP7B1**	CYP7B1_01*
	CHEK1_02	**CTSH**	CTSH_01*		CYP7B1_02
	CHEK1_03	**CX3CR1**	CX3CR1_01*		CYP7B1_03
COASY	COASY_01*		CX3CR1_02		CYP7B1_06
COL18A1	COL18A1_01*	**CYP17A1**	CYP17A1_01*	**DHDH**	DHDH_02*
	COL18A1_02		CYP17A1_08		DHDH_03
	COL18A1_03		CYP17A1_10	**DHFR**	DHFR_07*
COMT	COMT_01*		CYP17A1_11		DHFR_11
	COMT_03		CYP17A1_12		DHFR_18
	COMT_16		CYP17A1_13	**DIO1**	DIO1_01*
	COMT_29	**CYP19A1**	CYP19A1_01*		DIO1_05
CRP	CRP_02*		CYP19A1_04	**DNAJC18**	DNAJC18_01*
	CRP_03		CYP19A1_06	**DRD1**	DRD1_02*
CSF1R	CSF1R_02*		CYP19A1_08	**DRD2**	DRD2_01*
	CSF1R_03		CYP19A1_09		DRD2_03
	CSF1R_05		CYP19A1_14	**DRD4**	DRD4_07*
CSF2	CSF2_02*		CYP19A1_15		DRD4_15
CSF3	CSF3_02*		CYP19A1_16	**EDN1**	EDN1_01*
	CSF3_06		CYP19A1_27		EDN1_02
CSTF1	CSTF1_08*		CYP19A1_29	**EFNB3**	EFNB3_01*
	CSTF1_10		CYP19A1_30		EFNB3_02
	CSTF1_21		CYP19A1_34	**EGF**	EGF_02*
	CSTF1_22		CYP19A1_36		EGF_04
CTH	CTH_01*		CYP19A1_37		EGF_08
	CTH_03		CYP19A1_38	**EGFR**	EGFR_03*
	CTH_07		CYP19A1_39		EGFR_04
	CTH_10		CYP19A1_40		EGFR_05
	CTH_13		CYP19A1_41	**ENG**	ENG_06*
	CTH_14	**CYP1A1**	CYP1A1_14*	**ENPP1**	ENPP1_04*
CTLA4	CTLA4_01*		CYP1A1_15	**EPHX1**	EPHX1_01*
	CTLA4_07		CYP1A1_78		EPHX1_06
	CTLA4_10		CYP1A1_81		EPHX1_10
	CTLA4_16		CYP1A1_91		EPHX1_12
	CTLA4_17	**CYP1B1**	CYP1B1_07*		EPHX1_14
	CTLA4_19		CYP1B1_08		EPHX1_15
	CTLA4_25		CYP1B1_18		EPHX1_17
CTNNB1	rs11129895*		CYP1B1_27		EPHX1_18

Gene	SNP	Gene	SNP	Gene	SNP
EPHX2	EPHX2_04	**FBXW7**	FBXW7_05	**GHR**	GHR_45
ERBB2	ERBB2_03*		FBXW7_44		GHR_46
ERCC1	ERCC1_05*	**FLJ45983**	FLJ45983_03*		GHR_47
	ERCC1_06		FLJ45983_16		GHR_50
	ERCC1_30	**FOS**	FOS_02*		GHR_77
ERCC2	ERCC2_03*		FOS_06		GHR_79
	ERCC2_09		FOS_08		GHR_90
ERCC3	ERCC3_02*	**FOXA1**	FOXA1_41*	**GPX1**	GPX1_06*
	ERCC3_04	**FOXC1**	FOXC1_02*		GPX1_28
ERCC4	ERCC4_01*		FOXC1_06	**GPX2**	GPX2_02*
	ERCC4_15		FOXC1_07		GPX2_07
ERCC5	ERCC5_01*		FOXC1_13		GPX2_09
	ERCC5_02		FOXC1_22		GPX2_13
	ERCC5_05		FOXC1_23		GPX2_14
ERCC6	ERCC6_04*	**FUT2**	FUT2_05*		GPX2_16
	ERCC6_12	**FZD7**	FZD7_06*		GPX2_17
ESR1	ESR1_01*		FZD7_10		GPX2_18
	ESR1_07		FZD7_15		GPX2_19
	ESR1_08		FZD7_16		GPX2_21
	ESR1_13		FZD7_17	**GPX3**	GPX3_04*
	ESR1_14		FZD7_20		GPX3_16
	ESR1_17	**GATA3**	GATA3_10*		GPX3_18
	ESR1_30		GATA3_21		GPX3_21
	ESR1_31		GATA3_23		GPX3_25
	ESR1_34		GATA3_25		GPX3_28
ESR2	ESR2_02*		GATA3_27	**GPX4**	GPX4_06*
	ESR2_05		GATA3_28		GPX4_08
EXO1	EXO1_01*		GATA3_29		GPX4_09
	EXO1_02		GATA3_46		GPX4_12
FAM82A	FAM82A_01*		GATA3_68	**GSK3B**	rs10934500*
	FAM82A_02		GATA3_76		rs10934503
	FAM82A_08	**GC**	GC_02*		rs1154597
FANCA	FANCA_02*	**GDF15**	GDF15_01*		rs12630592
	FANCA_03		GDF15_02		rs1381841
	FANCA_12	**GGH**	GGH_01*		rs1574154
	FANCA_16		GGH_02		rs16830683
	FANCA_22	**GHR**	GHR_01*		rs16830689
	FANCA_25		GHR_03		rs1719888
	FANCA_28		GHR_110		rs1719889
	FANCA_34		GHR_113		rs1719895
	FANCA_35		GHR_16		rs17204605
	FANCA_37		GHR_21		rs17204878
	FANCA_39		GHR_214		rs1732170
FAZ	FAS_01*		GHR_27		rs17810235
	FAS_04		GHR_28		rs17810302
	FAS_09		GHR_29		rs17810676
FASLG	FASLG_01*		GHR_30		rs1870931
FBXW7	FBXW7_01*		GHR_31		rs2319398
	FBXW7_02		GHR_33		rs2873950
	FBXW7_04		GHR_34		rs334535

Gene	SNP	Gene	SNP	Gene	SNP
GSK3B	rs334555	HSD17B4	HSD17B4_19	IGF2	IGF2_16
	rs334559		HSD17B4_21		IGF2_22
	rs3732361	HSD3B1	HSD3B1_18*	IGF2AS	IGF2AS_01*
	rs3755557		HSD3B1_22		IGF2AS_03
	rs4072520		HSD3B1_23		IGF2AS_04
	rs4624596		HSD3B1_24	IGF2R	IGF2R_02*
	rs4688046		HSD3B1_25		IGF2R_03
	rs4688047		HSD3B1_26		IGF2R_04
	rs6438553	HSD3B2	HSD3B2_07*		IGF2R_05
	rs6770314		HSD3B2_14		IGF2R_07
	rs6779828		HSD3B2_19		IGF2R_11
	rs6781942		HSD3B2_25	IGFALS	IGFALS_84*
	rs7617372	HSPB8	HSPB8_01*		IGFALS_91
	rs7620750	HTR1B	HTR1B_02*	IGFBP1	IGFBP1_01*
	rs9851174		HTR1B_07	IGFBP2	IGFBP2_25*
	rs9873477	HTR1D	HTR1D_01*		IGFBP2_26
	rs9878473		HTR1D_03		IGFBP2_29
GSTA4	GSTA4_01*		HTR1D_04	IGFBP3	IGFBP3_04*
	GSTA4_02	HUS1	HUS1_01*	IGFBP5	IGFBP5_05*
	GSTA4_04		HUS1_05		IGFBP5_10
	GSTA4_07	ICAM1	ICAM1_06*	IGFBP6	IGFBP6_17*
GSTM3	GSTM3_01*		ICAM1_15		IGFBP6_18
	GSTM3_05		ICAM1_16		IGFBP6_19
	GSTM3_06		ICAM1_19	IL10	IL10_01*
GSTP1	GSTP1_01*	IFNAR2	IFNAR2_01*		IL10_03
	GSTP1_02		IFNAR2_06		IL10_05
GSTZ1	GSTZ1_02*		IFNAR2_10		IL10_06
	GSTZ1_03	IFNG	IFNG_07*		IL10_07
HADHA	HADHA_01*	IFNGR1	IFNGR1_01*		IL10_13
	HADHA_05		IFNGR1_05		IL10_17
	HADHA_10	IFNGR2	IFNGR2_03*	IL10RA	IL10RA_02*
HAO2	HAO2_01*	IGF1	IGF1_04*		IL10RA_08
HFE	HFE_01*		IGF1_11	IL12A	IL12A_09*
	HFE_07		IGF1_15	IL12B	IL12B_04*
	HFE_08		IGF1_16		IL12B_11
HIF1AN	HIF1AN_02*		IGF1_22	IL13	IL13_01*
HMGCR	HMGCR_01*		IGF1_24		IL13_02
	HMGCR_02		IGF1_27		IL13_03
HSD17B1	HSD17B1_06*		IGF1_44		IL13_06
	HSD17B1_10		IGF1_46	IL15	IL15_01*
HSD17B2	HSD17B2_01*	IGF1R	IGF1R_01*		IL15_02
	HSD17B2_02		IGF1R_05		IL15_06
HSD17B4	HSD17B4_01*		IGF1R_06		IL15_07
	HSD17B4_03		IGF1R_12		IL15_10
	HSD17B4_08		IGF1R_18	IL15RA	IL15RA_02*
	HSD17B4_10		IGF1R_26		IL15RA_04
	HSD17B4_15		IGF1R_27		IL15RA_05
	HSD17B4_16	IGF2	IGF2_02*		IL15RA_06
	HSD17B4_17		IGF2_03	IL1A	IL1A_01*
	HSD17B4_18		IGF2_09		IL1A_04

Gene	SNP	Gene	SNP	Gene	SNP
IL1B	IL1B_02*	IRS1	IRS1_08	LIPC	LIPC_06
	IL1B_03	JAK3	JAK3_01*		LIPC_08
	IL1B_08		JAK3_02		LIPC_09
	IL1B_12		JAK3_12		LIPC_17
IL1RN	IL1RN_02*	JTV1	rs2009115*		LIPC_23
	IL1RN_04	KRAS	rs10505980*		LIPC_25
	IL1RN_05		rs10842515		LIPC_37
IL2	IL2_01*		rs10842518	LITAF	LITAF_01*
	IL2_03		rs11047902		LITAF_02
IL3	IL3_01*		rs11047918	LMO2	LMO2_01*
IL4	IL4_01*		rs1137196		LMO2_04
	IL4_02		rs1137282		LMO2_08
	IL4_03		rs12226937	LMOD1	LMOD1_03*
	IL4_10		rs12228277	LOC391073	LOC391073_01*
	IL4_11		rs13096	LOC646837	LOC646837_05*
IL4R	IL4R_02*		rs17329025	LOC727797	LOC727797_05*
	IL4R_03		rs17329424		LOC727797_06
	IL4R_05		rs17388148		LOC727797_11
	IL4R_07		rs2970532	LPL	LPL_01*
	IL4R_10		rs4246229		LPL_03
	IL4R_24		rs4368021		LPL_04
	IL4R_27		rs4623993		LPL_05
IL6	IL6_01*		rs6487461		LPL_06
	IL6_04		rs712		LPL_08
IL6R	IL6R_04*		rs7133640		LPL_09
IL7R	IL7R_01*		rs7973746	LRP5	LRP5_01*
	IL7R_08		rs9266		LRP5_04
IL8	IL8_01*	KRT23	KRT23_03*		LRP5_06
	IL8_05	LCAT	LCAT_03*		LRP5_07
	IL8_11		LCAT_05		LRP5_15
IL8RA	IL8RA_04*	LDLR	LDLR_01*	LRP6	LRP6_02*
INSR	INSR_01*		LDLR_03		LRP6_03
	INSR_05		LDLR_08	LTA	LTA_01*
	INSR_06		LDLR_12		LTA_05
	INSR_07		LDLR_18	MASP1	rs1001073*
	INSR_11	LEP	LEP_01*		rs12635264
	INSR_13	LEPR	LEPR_01*		rs13089330
	INSR_19		LEPR_03		rs13094773
	INSR_28		LEPR_04		rs1533593
	INSR_30		LEPR_08		rs3105782
	INSR_51	LIG1	LIG1_01*		rs3733001
	INSR_59		LIG1_02		rs3864099
	INSR_61		LIG1_03		rs4376034
IRF1	IRF1_03*		LIG1_18		rs696405
	IRF1_05		LIG1_29		rs698079
IRF3	IRF3_01*	LIG3	LIG3_08*		rs698090
	IRF3_02	LIG4	LIG4_01*		rs698105
	IRF3_12	LIPC	LIPC_01*		rs710459
IRS1	IRS1_03*		LIPC_02		rs7609662
	IRS1_04		LIPC_04		MASP1_01

Gene	SNP	Gene	SNP	Gene	SNP
MATR3	MATR3_01*	**MSH2**	MSH2_03	**NBN**	NBN_04
MBD2	MBD2_01*	**MSH3**	MSH3_02*		NBN_13
	MBD2_02		MSH3_03	**NCF2**	NCF2_03*
	MBD2_03		MSH3_07		NCF2_04
	MBD2_04		MSH3_09		NCF2_05
MBD4	MBD4_02*		MSH3_12	**NCOA3**	NCOA3_01*
MBL2	MBL2_03*		MSH3_29		NCOA3_02
	MBL2_06	**MSH6**	MSH6_01*		NCOA3_04
	MBL2_09		MSH6_04	**NFKB1**	NFKB1_01*
	MBL2_11	**MSR1**	MSR1_01*		NFKB1_02
	MBL2_12		MSR1_02		NFKB1_09
	MBL2_27	**MTHFD2**	MTHFD2_01*		NFKB1_14
	MBL2_30	**MTHFR**	rs1801133*		NFKB1_21
	MBL2_38		MTHFR_02		NFKB1_33
	MBL2_44		MTHFR_03	**NFKBIE**	NFKBIE_01*
	MBL2_46		MTHFR_07		NFKBIE_02
	MBL2_65	**MTR**	rs1805087*		NFKBIE_03
MDM2	MDM2_01*		MTR_01		NFKBIE_08
MEST	MEST_03*		MTR_05	**NICN1**	NICN1_01*
MET	MET_01*		MTR_06	**NINJ1**	NINJ1_01*
	MET_04	**MTRR**	MTRR_05*		NINJ1_03
	MET_13		MTRR_07	**NOS2A**	NOS2A_02*
	MET_26		MTRR_10		NOS2A_07
METTL1	METTL1_01*		MTRR_11	**NOS3**	rs3918226*
MGMT	MGMT_03*		MTRR_19		NOS3_01
	MGMT_06		MTRR_22	**NPAT**	NPAT_01*
	MGMT_12	**MX1**	MX1_01*	**NQO1**	NQO1_07*
	MGMT_19		MX1_03		NQO1_08
MLH1	MLH1_02*		MX1_04		NQO1_15
	MLH1_05		MX1_07	**NR1H4**	NR1H4_05*
MMP1	MMP1_01*		MX1_08		NR1H4_18
	MMP1_03		MX1_10	**NUBP2**	NUBP2_01*
	MMP1_05		MX1_11	**OCA2**	OCA2_03*
	MMP1_09		MX1_22		OCA2_07
MPDU1	MPDU1_01*		MX1_28		OCA2_23
MPO	MPO_04*	**MYBL2**	MYBL2_03*	**OGG1**	OGG1_12*
MSH2	rs17036577*		MYBL2_06		OGG1_13
	rs17036614		MYBL2_09	**OPRD1**	OPRD1_03*
	rs1863332		MYBL2_19		OPRD1_05
	rs1981928		MYBL2_30	**OPRM1**	OPRM1_01*
	rs2042649		MYBL2_31		OPRM1_02
	rs2303428		MYBL2_36		OPRM1_03
	rs3771281		MYBL2_46		OPRM1_23
	rs3821227	**MYC**	MYC_02*	**P2RX7**	P2RX7_10*
	rs4608577	**MYNN**	MYNN_01*	**PAK6**	PAK6_13*
	rs4952887	**MYO5A**	MYO5A_01*		PAK6_14
	rs6544991		MYO5A_06		PAK6_16
	rs7585925		MYO5A_07		PAK6_19
	rs7602094	**NBN**	NBN_01*		PAK6_24
	rs7607076		NBN_02		PAK6_43

Gene	SNP	Gene	SNP	Gene	SNP
PARP1	PARP1_01*	PMS1	rs1233258	RAB15	RAB15_02*
	PARP1_06		rs1233284		RAB15_03
	PARP1_10		rs1233288		RAB15_04
	PARP1_12		rs1233291	RAC1	RAC1_03*
	PARP1_13		rs1233297	RAD23B	RAD23B_02*
	PARP1_14		rs1233299		RAD23B_03
PARP4	PARP4_01*		rs1233302		RAD23B_04
	PARP4_03		rs12618262		RAD23B_05
	PARP4_17		rs256550	RAD51	rs11852786*
	PARP4_19		rs256552		rs1801320
	PARP4_23		rs256563		rs2304579
PCNA	PCNA_06*		rs256564		rs2412546
	PCNA_07		rs256567		rs2412547
	PCNA_10		rs5742926		rs2619679
PCTP	PCTP_01*		rs5742938		rs2619681
	PCTP_03		rs5743030		rs4144242
PGR	PGR_01*		rs5743072		rs4924496
	PGR_05		rs5743112		RAD51_01
	PGR_07		rs5743116	RAD52	RAD52_01*
	PGR_11	PMS2	rs2345060*		RAD52_07
	PGR_12		rs3735295	RAD54L	RAD54L_04*
	PGR_14		rs6463524	RAG1	RAG1_01*
	PGR_15	POLB	POLB_05*	RB1CC1	RB1CC1_10*
	PGR_16		POLB_08		RB1CC1_24
	PGR_17		POLB_16		RB1CC1_40
	PGR_18	POLD1	POLD1_13*		RB1CC1_50
	PGR_20	POT1	POT1_02*	RERG	RERG_03*
	PGR_21		POT1_03		RERG_10
	PGR_23		POT1_05		RERG_24
	PGR_24		POT1_07		RERG_29
	PGR_26		POT1_09		RERG_30
	PGR_27		POT1_10		RERG_31
	PGR_28		POT1_11		RERG_33
PHB	PHB_02*		POT1_18		RERG_36
PIM1	PIM1_01*		POT1_37		RERG_37
	PIM1_03	PPARG	PPARG_06*		RERG_41
	PIM1_17		PPARG_07		RERG_44
	PIM1_25		PPARG_11		RERG_47
PIN1	PIN1_01*	PPP1R13L	rs6966*	RET	RET_01*
	PIN1_02	PTEN	PTEN_01*		RET_02
	PIN1_16		PTEN_10	RGS17	RGS17_01*
	PIN1_17	PTGS1	PTGS1_02*		RGS17_03
	PIN1_21	PTGS2	PTGS2_05*	RGS5	RGS5_01*
PLA2G2A	PLA2G2A_03*		PTGS2_08	RGS6	RGS6_02*
PLA2G6	PLA2G6_02*		PTGS2_19		RGS6_04
	PLA2G6_08		PTGS2_33		RGS6_05
	PLA2G6_10		PTGS2_44	RNASEL	RNASEL_01*
	PLA2G6_12	PTH	PTH_01*		RNASEL_02
PLK1	PLK1_15*		PTH_03	ROS1	ROS1_03*
PMS1	rs1233255*		PTH_04		ROS1_04

Gene	SNP	Gene	SNP	Gene	SNP
ROS1	ROS1_12	SLC23A2	SLC23A2_48	TERT	TERT_14
	ROS1_14	SLC2A1	SLC2A1_01*		TERT_15
	ROS1_15	SLC2A4	SLC2A4_02*		TERT_21
	ROS1_18	SLC30A1	SLC30A1_01*		TERT_54
	ROS1_20	SLC30A4	SLC30A4_01*	TFF1	TFF1_01*
RXRA	RXRA_01*	SLC39A2	SLC39A2_05*	TFF3	TFF3_02*
	RXRA_03		SLC39A2_07	TFRC	TFRC_01*
RXRB	RXRB_02*		SLC39A2_10	TGFB1	TGFB1_03*
	RXRB_11	SLC4A2	SLC4A2_01*	TGFBR1	TGFBR1_01*
SAT2	SAT2_01*		SLC4A2_02		TGFBR1_03
	SAT2_03		SLC4A2_04		TGFBR1_04
SCARB1	SCARB1_01*	SLC6A18	SLC6A18_13*	TGM1	TGM1_01*
	SCARB1_02	SLC6A3	SLC6A3_03*		TGM1_02
	SCARB1_03		SLC6A3_05	TLR2	TLR2_04*
	SCARB1_08		SLC6A3_10		TLR2_05
	SCARB1_09		SLC6A3_14		TLR2_06
SCUBE2	SCUBE2_02*	SOAT2	SOAT2_01*	TNF	TNF_02*
	SCUBE2_03		SOAT2_09		TNF_09
	SCUBE2_13		SOAT2_21		TNF_12
SEC14L2	SEC14L2_01*	SOD1	SOD1_01*		TNF_13
	SEC14L2_04	SOD2	SOD2_01*	TNFRSF10A	TNFRSF10A_02*
	SEC14L2_05		SOD2_06		TNFRSF10A_06
SELE	SELE_01*	SOD3	SOD3_05*	TNFRSF1A	TNFRSF1A_02*
SEP15	SEP15_02*	SRA1	SRA1_03*	TNIP1	TNIP1_02*
	SEP15_04		SRA1_04	TNKS	TNKS_01*
SEPP1	SEPP1_01*		SRA1_05		TNKS_03
	SEPP1_02	SSTR3	SSTR3_01*		TNKS_05
SEPT2	SEPT2_01*		SSTR3_03		TNKS_110
SFTPD	SFTPD_01*	STAT1	STAT1_01*		TNKS_124
	SFTPD_03	STK11	STK11_03*		TNKS_13
SHBG	SHBG_01*	SULT1A2	SULT1A2_09*		TNKS_15
	SHBG_05	TCTA	TCTA_02*		TNKS_18
	SHBG_12		TCTA_04		TNKS_20
	SHBG_13	TEP1	TEP1_01*		TNKS_22
SLAMF1	SLAMF1_02*		TEP1_02		TNKS_23
	SLAMF1_03		TEP1_03		TNKS_26
	SLAMF1_04		TEP1_08		TNKS_33
SLC19A1	SLC19A1_01*		TEP1_10		TNKS_34
	SLC19A1_05		TEP1_11		TNKS_35
SLC23A1	SLC23A1_05*	TERF1	TERF1_01*		TNKS_36
	SLC23A1_09		TERF1_02		TNKS_38
	SLC23A1_18		TERF1_04		TNKS_46
	SLC23A1_20		TERF1_06		TNKS_64
SLC23A2	SLC23A2_01*		TERF1_27		TNKS_76
	SLC23A2_02	TERF2	TERF2_01*	TP53	TP53_09*
	SLC23A2_03		TERF2_03		TP53_11
	SLC23A2_05		TERF2_14		TP53_14
	SLC23A2_25	TERT	TERT_02*		TP53_66
	SLC23A2_31		TERT_03		TP53_69
	SLC23A2_33		TERT_08	TP53I3	TP53I3_03*

Gene	SNP	Gene	SNP	Gene	SNP
TP53I3	TP53I3_10	**TXNRD2**	TXNRD2_83	**XBP1**	XBP1_02
	TP53I3_12		TXNRD2_88		XBP1_09
	TP53I3_13	**TYMS**	TYMS_01*		XBP1_10
	TP53I3_18		TYMS_05	**XPA**	rs1800975*
TP73L	TP73L_03*		TYMS_10	**XPC**	XPC_01*
	TP73L_13	**TYR**	TYR_02*		XPC_03
	TP73L_15		TYR_08		XPC_08
	TP73L_16	**UCP3**	UCP3_01*	**XRCC1**	XRCC1_01*
	TP73L_17		UCP3_02	**XRCC3**	XRCC3_03*
	TP73L_26	**UGT1A1**	rs1042640*		XRCC3_04
	TP73L_28	**VCAM1**	VCAM1_02*	**XRCC4**	XRCC4_01*
	TP73L_46		VCAM1_05		XRCC4_04
	TP73L_47		VCAM1_38		XRCC4_05
	TP73L_52	**VDR**	VDR_07*		XRCC4_07
TSG101	TSG101_07*		VDR_12		XRCC4_10
	TSG101_28	**VEGF**	VEGF_04*	**XRCC5**	XRCC5_02*
	TSG101_30		VEGF_05		XRCC5_12
	TSG101_33		VEGF_19		XRCC5_14
	TSG101_36	**VIL2**	VIL2_02*		XRCC5_17
	TSG101_40		VIL2_03		XRCC5_19
TSPO	TSPO_03*	**WDR79**	WDR79_06*	**ZFPM1**	ZFPM1_07*
	TSPO_05		WDR79_08	**ZNF230**	ZNF230_01*
	TSPO_09		WDR79_09	**ZNF350**	ZNF350_04*
WRN	WRN_04*		WDR79_11		ZNF350_06
	WRN_07	**WRN**	WRN_01*		ZNF350_08
	WRN_08		WRN_03		
TXNRD2	TXNRD2_76*	**XBP1**	XBP1_01*		

ANEXO II

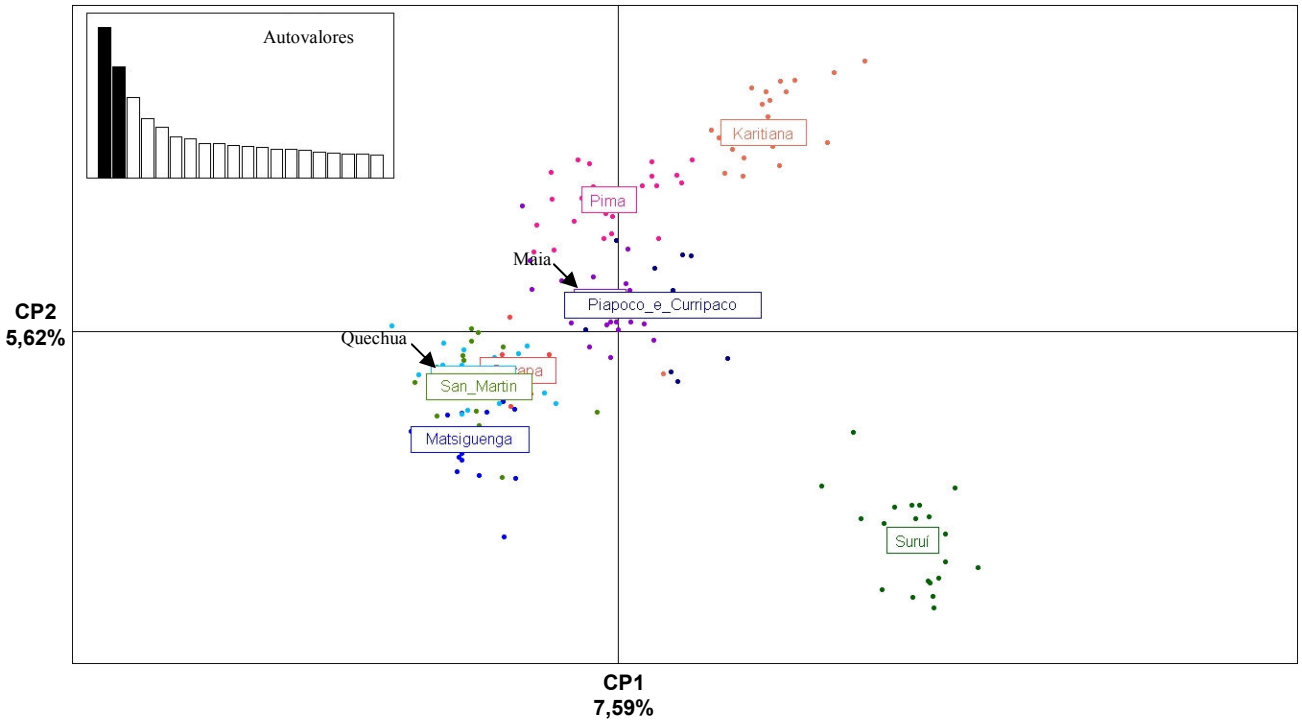
Populações do painel de amostras do *Human Genome Diversity Project* – HGDP, região gegráfica amostrada e suas respectivas coordenadas geográficas. (Tabela modificada de Cann *et al.* 2002)

Região Geográfica	Origem da Amostra	Coordenadas Geográficas	População
África Subsaariana	República Centro-Africana	4°N, 17°L	Pigmeu Biaka
África Subsaariana	República Democrática do Congo	1°N, 29°L	Pigmeu Mbuti
África Subsaariana	Senegal	12°N, 12°O	Mandenka
África Subsaariana	Nigéria	6-10°N, 2-8°L	Ioruba
África Subsaariana	Namíbia	21°S, 20°L	San
África Subsaariana	Kenia	3°S, 37°L	Bantu NE
África Subsaariana	Bantu SE	29°S, 30°L	Bantu SE Pedi
África Subsaariana	Bantu SE	29°S, 29°L	Bantu Sotho
África Subsaariana	Bantu SE	28°S, 24°L	Bantu SE Tswana
África Subsaariana	Bantu SE	28°S, 31°L	Bantu SE Zulu
África Subsaariana	Bantu SO	22°S, 19°L	Bantu SO Herero
África Subsaariana	Bantu SO	19°S, 18°L	Bantu SO Ovambo
Norte Africano	Argélia	32°N, 3°L	Mozabite
Oriente Médio	Israel (Negev)	31°N, 35°L	Beduína
Oriente Médio	Israel (Monte Carmelo)	32°N, 35°L	Drusa
Oriente Médio	Israel (Região Central)	32°N, 35°L	Palestina
Ásia	Paquistão	30-31°N, 66-67°L	Brahui
Ásia	Paquistão	30-31°N, 66-67°L	Balochi
Ásia	Paquistão	33-34°N, 70°L	Hazara
Ásia	Paquistão	26°N, 62-66°L	Makrani
Ásia	Paquistão	24-27°N, 68-70°L	Sindhi
Ásia	Paquistão	32-35°N, 69-72°L	Pathan
Ásia	Paquistão	35-37°N, 71-72°L	Kalash
Ásia	Paquistão	36-37°N, 73-75°L	Burusho
Ásia	China	26-39°N, 108-120°L	Han
Ásia	China	29°N, 109°L	Tujia
Ásia	China	28°N, 103°L	Yizu (Yi)
Ásia	China	28°N, 109°L	Miaozu (Miao)
Ásia	China	48-53°N, 122-131°L	Oroqen
Ásia	China	48-49°N, 124°L	Daur
Ásia	China	48-49°N, 118-120°L	Mongólia
Ásia	China	47-48°N, 132-135°L	Hezhen
Ásia	China	43-44°N, 81-82°L	Xibo
Ásia	China	44°N, 81°L	Uigur
Ásia	China	21°N, 100°L	Dai
Ásia	China	22°N, 100°L	Lahu
Ásia	China	27°N, 119°L	She
Ásia	China	26°N, 100°L	Naxi
Ásia	China	36°N, 101°L	Tu
Ásia	Sibéria	62-64°N, 129-130°L	Yakut
Ásia	Japão	38°N, 138°L	Japonesa
Ásia	Camboja	12°N, 105°L	Camboja
Oceania	Nova Guiné	4°S, 143°L	Papua
Oceania	Bougainville	6°S, 155°L	Melanésia
Europa	França	46°N, 2°L	Francesa (várias regiões)

Região Geográfica	Origem da Amostra	Coordenadas Geográficas	População
Europa	França	43°N, 0°	Francesa Basca
Europa	Itália	40°N, 9°L	Sardenha
Europa	Itália	46°N, 10°L	Bérgamo
Europa	Itália	43°N, 11°L	Toscana
Europa	Ilhas Órcades	59°N, 3°O	Orcadiana
Europa	Cáucaso Russo	44°N, 39°L	Adygei
Europa	Rússia	61°N, 39-41°L	Russa NO
América	México	29°N, 108°O	Pima
América	México	19°N, 91°O	Maia
América	Colômbia	3°N, 68°O	Piapoco e Curripaco
América	Brasil	10°S, 63°O	Karitiana
América	Brasil	11°S, 62°O	Suruí

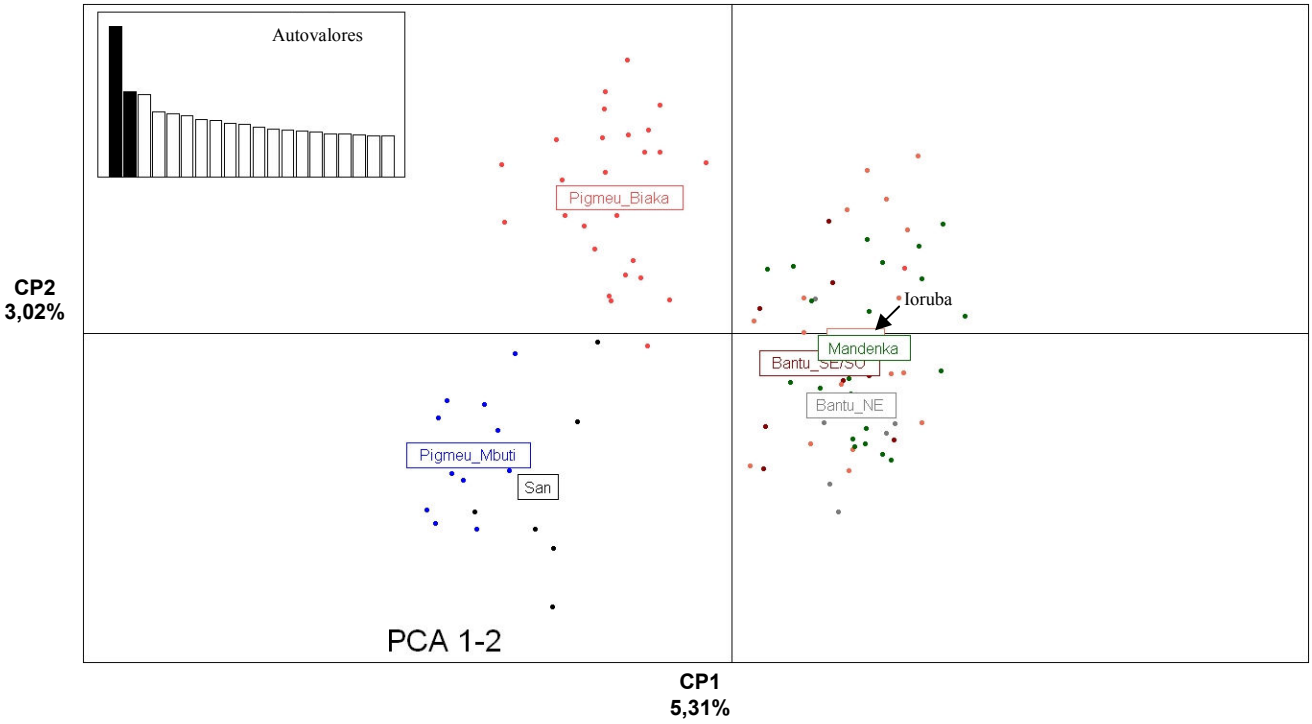
ANEXO III

Estruturação populacional no continente americano. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.



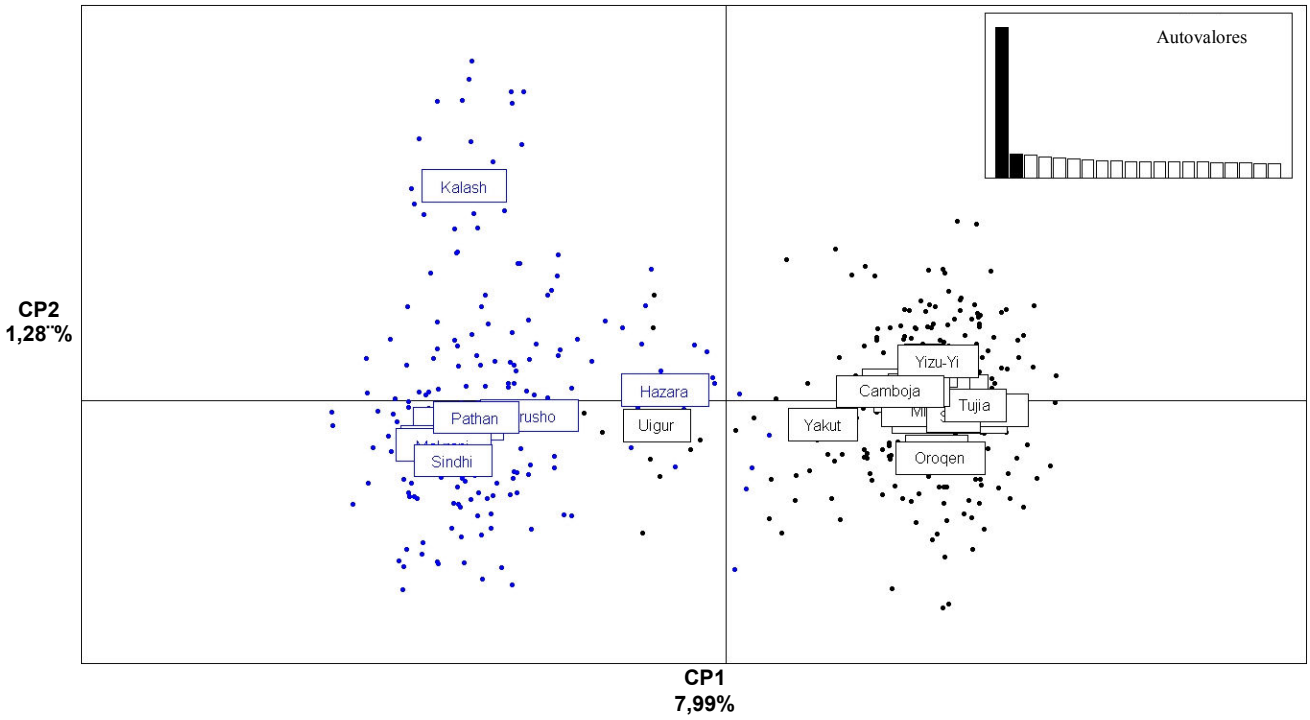
ANEXO IV

Estruturação populacional no continente africano. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.



ANEXO V

Estruturação populacional no continente asiático. Indivíduos do mesmo grupo populacional estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal.



ANEXO VI

Influência das populações parentais na formação da população hispânica. Indivíduos de mesma população estão representados por pontos de mesma cor. Os valores em porcentagem indicam a parcela de variância explicada pelos respectivos componentes principais. CP1: Primeiro componente principal; CP2: Segundo componente principal. AFRL: Leste Africano; AFRO: Oeste Africano; EUR: Europa; AMC: América Central; AMS: América do Sul; HISP: Hispânicos.