

Fernanda de Souza Gomes Kehdy

**Estudo da diversidade genética humana em
populações nativas e miscigenadas através da
utilização de INDELS informativos de
ancestralidade localizados no cromossomo 5 e
microssatélites tetra-locais do cromossomo Y
*DYS464 e DYS503***

Belo Horizonte
Instituto de Ciências Biológicas
Universidade Federal de Minas Gerais
2010

Fernanda de Souza Gomes Kehdy

Estudo da diversidade genética humana em populações mundiais nativas e miscigenadas através da utilização de INDELS informativos de ancestralidade localizados no cromossomo 5 e microssatélites tetra-locais do cromossomo Y *DYS464* e *DYS503*

Tese apresentada ao Departamento de Bioquímica e Imunologia do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção de título de Doutor em Ciências. Área de concentração: Biologia Molecular

**Orientador:
Dr. Sergio Danilo Junho Pena**

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Bioquímica e Imunologia
Belo Horizonte – MG – Brasil
2010

*À minha eterna "mãezinha" Maria Cristina Lima de Castro
com imenso amor e muita saudade*

Agradecimentos

Embora uma tese seja, por sua finalidade acadêmica, um trabalho individual, há contribuições de natureza diversa que não podem deixar de ser realçadas. Por essa razão, desejo expressar os meus sinceros agradecimentos:

Em primeiro lugar, ao meu orientador Dr. Sérgio Danilo Junho Pena, por todos estes anos de convivência. Pela indiscutível competência e seriedade científica, pelos valiosos ensinamentos, críticas e constante preocupação em fazer de seus alunos bons cientistas. Pela grande capacidade de nos preparar não só para uma carreira acadêmica e sim para a vida. Pelo apoio, incentivo, paciência e disponibilidade em todas as fases que levaram à concretização deste trabalho. Pela compreensão silenciosa dos momentos difíceis pelos quais passei, permitindo que meu tempo interno fluísse, respeitosamente. Pela alegria de trabalharmos juntos. Por isso, Sérgio, a você, todo o meu carinho e minha gratidão.

Aos professores do laboratório de Genética Bioquímica Andréa Mara Macedo, Glória Regina Franco (Glorinha) e Carlos Renato Machado pela convivência sempre muito agradável, pelas discussões, críticas e sugestões em meu trabalho.

Ao professor Dr. Eduardo Tarazona (Dudu), pela paciência, carinho e receptividade me incluindo sempre nos eventos do seu laboratório. Pelos valiosíssimos ensinamentos de genética de populações dentro e fora da sala de aula. Por todos os cafés que tomamos juntos e construtivas conversas que tivemos no decorrer destes anos.

À Neuzinha, por toda a sua preocupação e eficiência em manter o laboratório organizado (o que definitivamente não é fácil!). Pelas várias conversas agradáveis que tivemos no decorrer destes anos. Por todo o carinho, paciência e compreensão em todos os meus momentos alegres e difíceis.

À Katita, por todo o carinho e amizade e por estar sempre disposta a me ajudar com um sorriso no rosto.

À Sabrina (Sabrininha) por manter tudo limpo e organizado para trabalharmos tranquilamente.

Às colegas do grupo de humanos Pricila e Danielle pelo carinho e companheirismo.

Ao queridíssimo amigo e colega Lucas Secchim Ribeiro por compartilhar intensamente comigo a segunda metade do meu doutorado. Pela indispensável ajuda na execução e também escrita deste trabalho (as tabelas...). Pelas infinitas ajudas com tudo o que “pisca” no laboratório. Por estar sempre pronto a me ajudar e parar tudo o que estivesse fazendo para me socorrer. Por todas as gargalhadas que demos e bobagens que falamos juntos. Por tornar nosso cotidiano no laboratório bem mais divertido.

À Dra. Vanessa Faria Gonçalves de Oliveira, a Van Vaaaaaaaaaaaaannnnnnnnn, minha grandessíssima amiga, por me acompanhar desde o meu primeiro dia no laboratório, por me ensinar simplesmente todas as técnicas de biologia molecular que precisei utilizar neste trabalho. Por todas as conversas que tivemos e cocas-light que tomamos juntas. Por todas as brilhantes sugestões em todas as etapas do meu doutorado e por representar o meu porto seguro dentro e fora do laboratório. Pelo apoio emocional e prático na execução e na escrita desta tese e principalmente pela verdadeira amizade. Van, essa tese é nossa!

À querida amiga Carol Furtado (a Carolzinha) por estar sempre por perto e ter se tornado, no decorrer desta caminhada, uma verdadeira amiga. Por compartilhar comigo de todos os meus momentos de angústia, sempre pronta a me escutar pacientemente. Por conseguir entender tudo o que estou sentindo apenas pelo olhar. Pela companhia alegre, engraçada e indispensável. Simplesmente pela amizade.

Ao João Pedro (Johnnyzinho) por estar sempre presente nos mostrando como a vida pode ser simples e feliz. Pela convivência bem-humorada, carinho e companheirismo. Pela sinceridade (nem precisava de tanta...) e por estar sempre pronto a ajudar com competência e simplicidade.

Ao amigo Matheus Rajão (Matheuzim) pelo carinho, amizade, companheirismo e milhões de conversas. Por toda a convivência diária agradável e os *happy hours* semanais.

Ao amigo Rodrigo Baptista, pelo carinho, amizade e companheirismo.

À Paula pelos papos agradáveis, carinho e amizade.

À todos os colegas do LGB que já passaram e que ainda estão no laboratório pela agradável convivência por todos estes anos de trabalho. Em especial, gostaria de agradecer ao Bruno Chamber, Clarice, Dani, Helder, Lucas Santana, Pedro e Rogério.

Ao Higgor pela amizade companheirismo e ajuda nos momentos de aperto.

À minha amiga Simone Santos pela amizade e por estabelecer o meu primeiro contato com o LGB.

À amiga Luciana Werneck, por estar sempre disponível não medindo esforços para me ajudar. Pela amizade, confiança e carinho.

À amiga Miriam Dutra pelo carinho, amizade e horas de conversas super produtivas.

Ao Juliano (Núcleo de Análise Genômica/UFMG), pela ajuda nos momentos de aperto com o Mega Bace.

À secretaria da pós-graduação em bioquímica e imunologia, nas pessoas de Celise e Grazielle.

À Polly pelo carinho e amizade.

Às agências financiadoras CAPES e CNPq.

À grande amiga Ana Carolina Simões Ramos, minha geneticista predileta, exemplo de competência profissional, pela amizade incondicional desde que nos conhecemos. Pela contagiante alegria e visão otimista da vida mesmo quando está tudo difícil. Por me mostrar como a vida pode ser “leve”. Pelo carinho e companheirismo independentes da distância física.

Ao meu irmão, Rafael de Souza Gomes Abdala Kehdy, por estar sempre presente e compartilhar comigo da mesma história.

À minha cunhada, Melina, pelo carinho e amizade.

À minha grande e mais antiga amiga, Regina Maria Freitas Faria, por estar sempre por perto desde os tempos de escola e representar para mim o verdadeiro retrato da amizade.

Aos meus “pais espirituais” (primos), Antônio João e Nena por estarem sempre presentes em todos os momentos de minha vida e supervalorizarem todos os meus feitos. Pelo amor, dedicação, companheirismo e principalmente, por todos os “mimos” que me fizeram e continuam fazendo.

Ao meu primo querido, Eduardo Antônio Fonseca de Torrecillas, por estar sempre presente de uma maneira silenciosa e delicada. Por conseguir, sem falar nada, demonstrar todo seu carinho de forma tão confortante.

Ao meu “irmão mais velho”, Paulo Henrique Palis Aguiar, pela antiga amizade e por proporcionar, ainda criança, meu primeiro contato com a ciência (a culpa de tudo é toda dele!).

Ao tio Odilon e tia Rosália simplesmente por existirem desde sempre em minha vida e me mostrarem o verdadeiro valor de uma amizade que se perpetua por mais de uma geração.

Ao meu marido, Leonardo de Araújo Lopes, pelo cuidado, preocupação e paciência em todos os momentos difíceis. Pelo companheirismo e delicada compreensão nos meus momentos de ausência (que não foram poucos). Pelo grande incentivo e apoio financeiro e emocional. Por toda sua admiração e amor a mim dedicados.

Aos meus queridos e amados pais, Sérgio Kehdy e Regina Maria, por inculcarem em mim, desde minha infância, o amor aos estudos. Por serem os grandes patrocinadores de todos os meus sonhos de forma financeira e emocional. Por me proporcionarem sempre um ambiente estável e estruturado em que eu pudesse estudar tranquilamente. Por estarem sempre presentes. Pelo amor e apoio incondicionais. Vocês não existem!! Amo vocês.

Índice

I - Apresentação.....	i
II - Lista de Abreviaturas e Símbolos.....	iii
III - Lista de Figuras.....	iv
IV - Lista de Tabelas.....	viii
V - Resumo.....	x
VI - Abstract.....	xiii
1 – Introdução Geral.....	1
1.1 – Marcadores moleculares e a genética de populações.....	2
1.1.1 - SNPs.....	3
1.1.2 - Microssatélites.....	4
1.1.3 – INDELS.....	5
1.1.4 - Marcadores de linhagem.....	6
1.1.5 - Marcadores biparentais.....	8
1.2 - O Projeto Diversidade do Genoma Humano.....	9
2 – Capítulo 1.....	14
2.1 - Introdução.....	15
2.1.1 – A ocupação dos diferentes continentes pelo homem moderno.....	15
2.1.2- A distribuição da variabilidade genética e estruturação das populações humanas	18
2.1.3- A colonização do continente americano.....	21
2.1.4 – O retrato molecular do Brasil.....	22
2.1.5- O genoma dos indivíduos miscigenados.....	25
2.2 – Objetivos.....	28

2.2.1 – Objetivo geral	28
2.2.2 – Objetivos específicos	29
2.3 – Material e Métodos.....	31
2.3.1 - Amostras de DNA.....	31
2.3.2 - Amplificação dos INDELS do cromossomo 5	32
2.3.2.1 - Seleção de marcadores	32
2.3.2.2 - Sistema <i>multiplex</i>	33
2.3.2.3 - Amplificação do <i>multiplex</i> A.....	33
2.3.2.4 - Amplificação do <i>multiplex</i> B.....	36
2.3.2.5 - Amplificação do <i>multiplex</i> C.....	38
2.3.2.6 - Amplificação do <i>multiplex</i> D	38
2.3.3 - Genotipagem.....	41
2.3.4 – Análise dos dados	41
2.3.4.1 – Análise das populações mundiais.....	42
2.3.4.1.1 - Parâmetros básicos das populações mundiais.....	42
2.3.4.1.2 – Análise de Variância Molecular (AMOVA).....	42
2.3.4.1.3 – Escalonamento Multidimensional (MDS) em populações mundiais.....	43
2.3.4.1.4 – Diferenciação genética entre indivíduos das populações mundiais	44
2.3.4.1.5 - Análise de desequilíbrio de ligação nas regiões mundiais	44
2.3.4.1.6 - Análise do Componente Principal (PCA) em populações mundiais	45
2.3.4.2 – Análise das populações brasileiras.....	46
2.3.4.2.1 – Parâmetros básicos das populações brasileiras	46
2.3.4.2.2 – Escalonamento Multidimensional (MDS) das populações brasileiras	47
2.3.4.2.3 - Análise de desequilíbrio de ligação nas populações brasileiras	47
2.3.4.2.4 - Análise do Componente Principal (PCA) com populações brasileiras	48
2.3.4.2.5 – Proporções de contribuição parental nas populações brasileiras	49

2.3.4.2.6 – Estrutura populacional.....	49
2.3.4.2.7 – Inferência de haplótipos.....	51
2.3.4.2.8 – Origem ancestral dos fragmentos cromossômicos das populações miscigenadas.....	52
2.3.4.2.9 – Correlação de ancestralidade entre cromossomos homólogos.....	53
2.4 – Resultados	54
2.4.1 – Indels do cromossomo 5.....	54
2.4.2 – Populações Mundiais	56
2.4.2.1 - Análise de Variância Molecular (AMOVA).....	67
2.4.2.2 – Escalonamento Multidimensional (MDS) em populações mundiais	69
2.4.2.3 – Diferenciação genética entre indivíduos das populações mundiais.....	70
2.4.2.4 – Análise de desequilíbrio de ligação nas regiões mundiais.....	74
2.4.2.5 – Análise do Componente Principal em populações mundiais.....	74
2.4.3.1 – Escalonamento Multidimensional (MDS) nas populações brasileiras e parentais	87
2.4.3.2 – Análise de desequilíbrio de ligação nas populações brasileiras.....	91
2.4.3.3 – Análise do Componente Principal (PCA) em populações brasileiras e parentais	91
2.4.3.4 – Análise de mistura genética nas populações brasileiras	97
2.4.3.5 - Estrutura populacional.....	97
2.4.3.6 – Inferência de haplótipos.....	101
2.4.3.7 - Origem ancestral dos fragmentos cromossômicos nas populações miscigenadas.....	102
2.4.3.8 – Correlação de ancestralidade entre pares homólogos	110
2.5 – Discussão.....	111
2.5.1 – INDELS do cromossomo 5 em populações mundiais.....	111
2.5.2 – INDELS do cromossomo 5 em populações brasileiras	115

2.5.2.1- Análise das proporções de mistura genética em amostras de brancos e pretos de São Paulo.....	118
2.5.2.2- Análise da origem ancestral de fragmentos cromossômicos nas populações parentais e brasileiras.....	123
2.6 – Conclusões.....	128
3 – Capítulo 2.....	130
3.1 – Introdução	131
3.1.1 – O cromossomo Y humano	131
3.1.2 – O conjunto de microssatélites <i>DYS464</i>	133
3.1.3 – O conjunto de microssatélites <i>DYS503</i>	134
3.2 – Objetivos.....	137
3.2.1 – Objetivo geral	137
3.2.2 – Objetivos específicos	138
3.3 – Material e Métodos.....	140
3.3.1 – Amostras de DNA.....	140
3.3.2 – Amplificação do microssatélite <i>DYS464</i>	141
3.3.3 – Amplificação do microssatélite <i>DYS503</i>	142
3.3.4 – Genotipagem dos microssatélites <i>DYS464</i> e <i>DYS503</i>	144
3.3.5 – Análise dos dados	144
3.3.5.1 – Microssatélite <i>DYS464</i>	144
3.3.5.1.1 – Parâmetros básicos de genética de populações.....	144
3.3.5.1.2 – Análise de estruturação populacional.....	145
3.3.5.1.2.1 – Análise de Variância Molecular (AMOVA)	145
3.3.5.1.2.2 – Escalonamento Multidimensional (MDS)	146
3.3.5.1.2.3 – Análise do Componente principal (PCA).....	146
3.3.5.1.2.4 – Cálculo das proporções de mistura nas populações miscigenadas.....	147

3.3.5.2 – Microsatélite <i>DYS503</i>	147
3.4 – Resultados	148
3.4.1 – Microsatélite <i>DYS464</i>	148
3.4.1.2 – Microsatélite <i>DYS464</i> em populações mundiais.....	151
3.4.1.2.1 – Índices de diversidade.....	151
3.4.1.2.3 – Escalonamento Multidimensional (MDS).....	161
3.4.1.2.4 – Análise do Componente Principal (PCA)	169
3.4.1.3 – Microsatélite <i>DYS464</i> em populações miscigenadas	175
3.4.1.3.1- Escalonamento Multidimensional	177
3.4.1.3.2- Análise do Componente Principal.....	181
3.4.1.3.3 – Estimativa do grau de miscigenação populacional.....	181
3.4.2 – Microsatélite <i>DYS503</i>	187
3.4.2.1- Análise de índices de diversidade	189
3.5 – Discussão.....	192
3.5.1 – Microsatélite <i>DYS464</i> em populações mundiais	192
3.5.2 – Microsatélite <i>DYS464</i> em populações miscigenadas	198
3.5.3 – Microsatélite <i>DYS503</i> em populações mundiais.....	202
3.6 – Conclusões.....	208
4 – Referências Bibliográficas.....	210
5 – Anexos.....	225
Anexo 1 – Frequências genotípicas do <i>DYS464</i> encontradas em cada população e região geográfica do painel do CEPH-HGDP.....	226
Anexo 2 – Frequências genotípicas do microsatélite <i>DYS464</i> em populações parentais e miscigenadas	260
Artigo Científico - Worldwide diversity of the Y-chromosome tetra-local microsatellite <i>DYS464</i>	265

I - Apresentação

A tese apresentada está organizada da seguinte forma: Introdução Geral seguida de duas partes independentes (capítulos 1 e 2), relacionadas aos dois temas trabalhados durante o período do doutorado. Cada capítulo apresenta Introdução, Objetivos Geral e Específicos, Material e Métodos, Resultados, Discussão e Conclusões. Em Anexo encontram-se duas tabelas que foram consideradas extensas para estarem no corpo da tese, além do artigo científico produzido durante o período do doutorado. Dois outros artigos científicos estão em processo de preparação para submissão.

A tese inicia-se com uma Introdução Geral na qual se procura fazer uma breve revisão a respeito de marcadores moleculares e suas aplicações em genética de populações. Nesta introdução, também descrevemos um rápido histórico do Projeto Diversidade do Genoma Humano (HGDP) e do painel de amostras internacionais do CEPH-HGDP que foi utilizado nos dois capítulos desta tese.

Em seguida encontram-se os dois capítulos da tese. Resumidamente, o Capítulo 1 dedica-se a duas análises distintas. Primeiro estudamos a estruturação genética em populações humanas mundiais através da utilização de polimorfismos de inserção e deleção (INDELs) informativos de ancestralidade localizados no braço longo do cromossomo 5. Em seguida, utilizamos estes marcadores para determinar a origem

ancestral de fragmentos cromossômicos em populações de indivíduos pretos e brancos da cidade de São Paulo.

No Capítulo 2 estudamos os microssatélites tetra locais do cromossomo Y, *DYS464* e *DYS503*, nas populações mundiais do painel do CEPH-HGDP e analisamos sua diversidade e filogeografia em diferentes populações humanas. Além disso, testamos a utilização das frequências alélicas do *DYS464* em estudos de ancestralidade paterna através da análise deste marcador em populações miscigenadas.

II - Lista de Abreviaturas e Símbolos

AMOVA – análise de variância molecular (*analysis of molecular variance*)

CEPH – *Centre d'Etude du Polymorphisme Humain*

DNA – Ácido desoxirribonucleico

HGDP – *Human Genome Diversity Project*

INDEL – Polimorfismo de inserção e deleção

LCL – *Lymphoblastoid Cell Line*

MDS – Escalonamento multidimensional (*Multidimensional Scaling*)

MIA – Marcador informativo de ancestralidade

mtDNA – DNA mitocondrial

NRY – região não recombinante do cromossomo Y

PCA – análise do componente principal (*principal component analysis*)

SNP – *Single Nucleotide Polymorphism*

III - Lista de Figuras

Figura 1- Localização geográfica e nome das populações do painel do CEPH-HGDP.....	13
Figura 2: – Mapa físico do cromossomo 5 com a localização dos INDELS selecionados.....	55
Figura 3: Eletroferograma produzido pelo programa <i>Fragment Profile</i> , versão 1.2, representando a genotipagem dos INDELS do cromossomo 5 para os <i>multiplexes A e B</i>	58
Figura 4: Eletroferograma produzido pelo programa <i>Fragment Profile</i> , versão 1.2, representando a genotipagem dos INDELS do cromossomo 5 para os <i>multiplexes C e D</i>	59
Figura 5 – Escalonamento Multidimensional realizado com as 7 regiões geográficas.....	71
Figura 6 – Escalonamento Multidimensional realizado com as 52 populações mundiais	72
Figura 7 – Árvores filogenéticas <i>neighbor-joining</i> relativas ao compartilhamento de alelos idênticos por estado entre os 1045 indivíduos do painel do CEPH-HGDP estudados.....	73
Figura 8– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para África	76
Figura 9– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para Oriente Médio	77
Figura 10– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Ásia Central	78
Figura 11– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para o Leste Asiático	79
Figura 12– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Oceania	80
Figura 13– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Europa	81
Figura 14 – Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para as Américas	82
Figura 15 – Gráfico de dispersão dos dois primeiros componentes principais baseados nas frequências alélicas dos INDELS do cromossomo 5 nas grandes regiões geográficas.....	83
Figura 16 - Gráfico de dispersão dos dois primeiros componentes principais baseados nas frequências alélicas dos INDELS do cromossomo 5 nas 52 populações mundiais	84

Figura 17 – Escalonamento Multidimensional realizado com as populações brasileiras e as grandes regiões geográficas parentais.....	89
Figura 18 - Escalonamento Multidimensional realizado com as populações brasileiras e as populações pertencentes às grandes regiões geográficas parentais.....	90
Figura 19– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para pretos brasileiros	92
Figura 20 – Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para pretos e brancos brasileiros	93
Figura 21 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências relativas de alelos dos INDELS do cromossomo 5 nas regiões geográficas parentais e nas populações brasileiras.....	95
Figura 22 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências relativas de alelos dos INDELS do cromossomo 5 nas populações parentais e nas populações brasileiras.....	96
Figura 23 - Estrutura populacional estimada pelo programa <i>Structure</i> das três regiões que abrangem as populações parentais da população brasileira.....	100
Figura 24: Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da África	105
Figura 25 - Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da Europa	106
Figura 26: Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da América	107
Figura 27: – Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo preto da cidade de São Paulo	108
Figura 28: – Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo branco da cidade de São Paulo.....	109
Figura 29 – Mapa do cromossomo Y com a localização dos microssatélites DYS464 e DYS503 em relação à região AZFc.....	136
Figura 30 - Eletroferograma produzido pelo programa <i>Fragment Profile</i> , versão 1.2, representando a genotipagem do conjunto de microssatélites <i>DYS464</i> pelo sequenciador MEGA BACE 1000.....	150

Figura 31 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microssatélite <i>DYS464</i> e Fst (Distância de Reynolds) como medida de distância.....	163
Figura 32 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microssatélite <i>DYS464</i> e Rst como medida de distância.....	164
Figura 33 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microssatélite <i>DYS464</i> e $\delta\mu^2$ como medida de distância.	165
Figura 34 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite <i>DYS464</i> e Fst (Distância de Reynolds) como medida de distância.....	166
Figura 35 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite <i>DYS464</i> e Rst como medida de distância.	167
Figura 36 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite <i>DYS464</i> e $\delta\mu^2$ como medida de distância.....	168
Figura 37 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências genotípicas do microssatélite <i>DYS464</i> nas 52 populações	170
Figura 38 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências alélicas do microssatélite <i>DYS464</i> nas 52 populações.....	171
Figura 39 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências genotípicas do microssatélite <i>DYS464</i> nas 7 regiões geográficas.....	173
Figura 40 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências alélicas do microssatélite <i>DYS464</i> nas 7 regiões geográficas	174
Figura 41 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite <i>DYS464</i> e Fst (distância de Reynolds) como medida de distância.....	178
Figura 42 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite <i>DYS464</i> e Rst como medida de distância.....	179
Figura 43 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite <i>DYS464</i> e $\delta\mu^2$ como medida de distância.....	180
Figura 44 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências alélicas do microssatélite <i>DYS464</i> encontradas nas populações miscigenadas e parentais.....	182

Figura 45 – Gráfico de minimização da soma dos quadrados para os pretos de São Paulo	183
Figura 46 – Gráfico de minimização da soma dos quadrados para os brancos de São Paulo.	184
Figura 47 – Gráfico de minimização da soma dos quadrados para os afro-americanos.....	185
Figura 48 – Gráfico de minimização da soma dos quadrados para os caucasianos	186
Figura 49: Eletroferograma produzido pelo programa <i>Fragment Profile</i> , versão 1.2, representando a genotipagem do conjunto de microssatélites <i>DYS503</i> pelo sequenciador MEGA BACE 1000.	188

IV - Lista de Tabelas

Tabela 1: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do <i>multiplex A</i>	35
Tabela 2: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do <i>multiplex B</i>	37
Tabela 3: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do <i>multiplex C</i>	39
Tabela 4: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do <i>multiplex D</i>	40
Tabela 5: Características dos INDELS (MIDs) selecionados. Nome, localização no cromossomo 5 e frequências alélicas encontradas por Weber <i>et al</i> (2002) em populações da África, Europa e Américas.....	57
Tabela 6 – Frequências alélicas dos 31 INDELS do cromossomo 5 em cada população e região geográfica do painel do CEPH-HGDP.....	60
Tabela 7 – Análise de variância molecular (AMOVA) de dados dos 31 INDELS localizados no braço longo do cromossomo 5 referentes às sete regiões geográficas do painel do CEPH-HGDP estudadas.....	68
Tabela 8 – Porcentagem de contribuição das populações parentais para as populações de Brancos e Pretos de São Paulo. Resultado gerado pelo programa <i>Admix 2.0</i>	99
Tabla 9 - Comparação das proporções de mistura africana, européia e ameríndia encontradas para pretos e brancos de São Paulo pelos diferentes métodos utilizados.....	120
Tabela 10: Seqüência de iniciadores e características dos microssatélites do cromossomo Y.....	143
Tabela 11 – Frequências alélicas do microssatélite <i>DYS464</i> encontradas em cada população e região geográfica do painel do CEPH-HGDP.....	155
Tabela 12 - Número de indivíduos, alelos, alelos região-específicos, genótipos, genótipos região-específicos e diversidade genotípica encontrados em cada região geográfica e no total das amostras estudadas para o microssatélite <i>DYS464</i>	158
Tabela 13- Análise de variância molecular (AMOVA) de dados do microssatélite <i>DYS464</i> referentes às sete regiões geográficas do painel do CEPH-HGDP estudadas. O intervalo de confiança foi de 95%.....	160

Tabela 14 – Frequências alélicas do <i>DYS464</i> em populações parentais e miscigenadas	176
Tabela 15: Número de indivíduos, alelos, alelos região-específicos, padrões alélicos, padrões alélicos região-específicos e diversidade de padrões alélicos encontrados em cada região geográfica e no total das amostras estudadas para o microssatélite <i>DYS503</i>	190
Tabela 16: Frequência de padrões alélicos encontradas em cada região geográfica para o conjunto de microssatélites <i>DYS503</i>	191

V - Resumo

No primeiro capítulo desta tese nós avaliamos um painel de 31 INDELS com ampla variação de frequência alélica entre África e Europa (δ médio = 0,46), em uma única região cromossômica, abrangendo 90Mb do braço longo do cromossomo 5. Primeiramente genotipamos o painel de diversidade do CEPH-HGDP (1045 indivíduos de 52 populações mundiais, distribuídas em 7 regiões geográficas) para os 31 INDELS. A análise de variância molecular (AMOVA) demonstrou que, dentro das populações, a diferença entre indivíduos equivale a aproximadamente 81% da variabilidade genética, enquanto que a diferença entre as regiões geográficas constitui cerca de 16%. Uma representação gráfica (escalonamento multidimensional) da matriz de distância baseada em F_{st} (distância de Reynolds) com as 52 populações revelou cinco claros agrupamentos correspondentes às cinco grandes regiões geográficas: África, Oceania, Leste Asiático, Américas e Eurásia (Europa, Ásia Central e Oriente Médio). Nossos resultados corroboram os estudos prévios com locos espalhados no genoma. Em seguida, a fim de inferir a origem ancestral de fragmentos cromossômicos e caracterizar o padrão de mistura da população brasileira, estudamos os 31 INDELS em 261 indivíduos não aparentados, autoclassificados de acordo com o IBGE como pretos (n=104) e brancos (n=157) da cidade de São Paulo. Utilizando as populações parentais como referência, a provável origem ancestral de cada loco foi estimada para cada cromossomo de cada indivíduo brasileiro. Os resultados mostraram que, em média, cada indivíduo preto de SP apresenta 27% da região cromossômica analisada

de origem européia, 53% de origem africana e 19% de origem ameríndia, enquanto para os brancos, estas porcentagens foram de 60%, 16% e 24%, respectivamente. Uma correlação significativa entre o número de locos de uma mesma origem ancestral entre os dois cromossomos 5 homólogos de um mesmo indivíduo foi encontrada tanto para os pretos ($p < 0,000001$) quanto para os brancos ($p < 0,000001$). Uma possível explicação para esses achados é o casamento assortativo de acordo com a cor, mas independente da ancestralidade genômica. Nossos resultados mostraram que os marcadores utilizados são capazes de discriminar as três raízes ancestrais brasileiras e que a abordagem utilizada na caracterização da população brasileira em nível cromossômico foi adequada.

No segundo capítulo da tese estudamos dois microssatélites tetra locais do cromossomo Y, *DYS464* e *DYS503*, em todas as amostras de homens do painel mundial do CEPH-HGDP. Além disso, avaliamos a utilização das frequências alélicas do *DYS464* na inferência de proporções de mistura de populações parentais em populações miscigenadas. Nos estudos das populações mundiais, para o *DYS464*, um total de 14 alelos com variações de tamanho de 9 a 23 repetições foram encontrados. Um total de 175 diferentes genótipos foram observados dos quais 90 foram encontrados em um único continente. A África foi a região que apresentou maior porcentagem de genótipos exclusivos. A diversidade genotípica foi de 0,98 na Europa, 0,97 na Ásia Central e Leste asiático, 0,95 na África, 0,94 na Oceania, 0,92 no Oriente Médio e 0,90 nas Américas. A análise de variância molecular (AMOVA) revelou baixos níveis de

estruturação genética com 88,42% da variabilidade genética dentro de populações e apenas 1,96% entre regiões geográficas. A utilização de frequências alélicas do *DYS464* na inferência de proporções de mistura em populações miscigenadas revelou valores que corroboram os dados encontrados na literatura. O microssatélite *DYS503* apresentou uma diversidade de padrões alélicos na África menor do que na maioria dos outros continentes. Nossos resultados juntos indicam que, embora o *DYS464* não nos permita fazer detalhadas inferências a respeito da filogeografia humana, ele é altamente polimórfico e pode ser uma ferramenta útil em estudos evolutivos e em estudos de populações miscigenadas. Além disso, a baixa diversidade do *DYS503* na é um dado incomum que, para ser compreendido, deve ser mais detalhadamente estudado.

VI - Abstract

In the first chapter of this thesis we evaluated a panel of 31 INDELS with wide variation in allele frequency between Africa and Europe (average $\delta = 0.46$) in a single chromosomal region, covering 90Mb of the long arm of chromosome 5. First, we genotyped the 31 INDELS in 1045 samples from 52 worldwide populations, distributed in seven geographical regions from the HGDP-CEPH Human Genome Diversity Cell Line Panel. The analysis of molecular variance (AMOVA) showed that, within populations, the difference between individuals is equivalent to approximately 81% of genetic variability, while the difference between geographic regions is about 16%. A graphical representation (multidimensional scaling) of the distance matrix based on F_{st} distance (Reynolds) with the 52 populations revealed five clear clusters corresponding to the five main geographical regions: Africa, Oceania, East Asia, Americas and Eurasia (Europe and Central Asia, Middle East). Our results corroborate previous studies with widespread chromosomal location. Then, in order to infer the ancestral origin of chromosomal fragments and characterize the admixed pattern of the Brazilian population, we studied the 31 INDELS in 261 unrelated individuals self-classified according to the IBGE as black ($n = 104$) and whites ($n = 157$) from São Paulo. Using the parental populations as reference, the probable ancestral origin of each locus was estimated for each chromosome of each Brazilian individual. The results showed that on average, each black Brazilian individual has 27% of the chromosome region analyzed from Europe, 53% of African origin and 19% of

Amerindian origin, while for whites Brazilians, these percentages were 60%, 16% and 24% respectively. A significant correlation between the number of loci from the same ancestral origin between the two homologues chromosome 5 of the same individual was found for both black ($p < 0.000001$) and white Brazilians ($p < 0.000001$). One possible explanation for these findings is assortative marriage according to color, but independent of genomic ancestry. Our results showed that the markers used in this study are able to discriminate the three ancestral Brazilian roots and the approach used in the characterization of the Brazilian population at chromosomal level was adequate.

In the second chapter of the thesis, two tetralocal Y- linked microsatellites were studied in all samples of men from the CEPH-HGDP Panel. Furthermore, we evaluated the use of allele frequencies of *DYS464* in inference of mixing ratios of admixed populations. In our studies of worldwide populations with the *DYS464*, fourteen different alleles were found with allele lengths varying from 9 to 23 repeats. A total of 175 different genotypes were observed, of which 90 appeared to be continent-specific. The region with the highest percentage of unique genotypes was Africa. Genotype diversity was 0.98 for Europe, 0.97 for Central and East Asia, 0.95 for Africa, 0.94 for Oceania, 0.92 for the Middle East and 0.90 for the Americas. A hierarchical analysis of molecular variance (AMOVA) showed low levels of worldwide genetic structure: 88.42%. The use of allele frequencies of *DYS464* to the inference of the mixing proportions in admixed populations showed values that corroborate data

found in the literature. The microsatellite *DYS503* showed lower allelic pattern diversity in Africa than in most other continents. Our results demonstrated that although *DYS464* microsatellite does not permit detailed phylogeographical inferences, it is highly polymorphic and still appears to be an informative tool for evolutionary and admixed population studies. Furthermore, the low diversity of *DYS503* is an unusual result that to be understood should be further studied

1 - Introdução Geral

1.1 – Marcadores moleculares e a genética de populações

A genética de populações humanas é a ciência que estuda a origem e as forças que determinam a evolução do homem moderno. Para entender a história e a evolução das populações humanas, a genética molecular utiliza a variabilidade genômica como uma ferramenta para inferir cenários evolutivos passados (Cavalli-Sforza, 1998). Esta variabilidade é que constitui a heterogeneidade do genoma humano, sendo as regiões ditas variáveis conhecidas como polimorfismos genéticos. Estes, quando utilizados em estudos genômicos, são denominados marcadores moleculares.

Hoje, com o advento e os avanços da engenharia genética e o seqüenciamento completo do genoma humano, diversos marcadores moleculares foram caracterizados e podem ser classificados em três grandes classes: os polimorfismos de sequência, tais como os SNPs (*Single Nucleotide Polymorphism*) (Sachidanandam *et al.*, 2001), os microssatélites (Tautz, 1989) e os polimorfismos de inserção e deleção (INDELS) (Weber *et al.*, 2002).

Além da sua classificação pela natureza molecular, como descrito acima, os marcadores moleculares podem também ser classificados de acordo com sua localização no genoma. Podem estar localizados nos cromossomos autossômicos, caracterizando os marcadores biparentais ou podem estar localizados na molécula de DNA mitocondrial (mtDNA) ou na região não recombinante (NRY) do cromossomo Y.

Estes últimos são herdados de forma uniparental constituindo, portanto, os marcadores de linhagem (matrilinhagem e patrilinhagem, respectivamente) (Underhill e Kivisild, 2007). A seguir, faremos uma breve descrição dos principais tipos de marcadores moleculares do genoma humano, incluindo tanto os classificados por sua natureza molecular (SNPs, microssatélites e INDELS) como os marcadores de linhagem (mtDNA e cromossomo Y).

1.1.1 - SNPs

Os SNPs consistem na variação de um único nucleotídeo na sequência de DNA. Os diferentes mecanismos de mutação podem gerar dois tipos de substituições de bases: as transições, quando ocorre troca de uma purina por outra purina (A \leftrightarrow G) ou de uma pirimidina por outra pirimidina (C \leftrightarrow T); e as transversões, quando ocorre troca de uma purina por uma pirimidina ou de uma pirimidina por uma purina (A \leftrightarrow C, A \leftrightarrow T, G \leftrightarrow C, G \leftrightarrow T) (Vignal *et al.*, 2002).

Embora em teoria qualquer uma das quatro bases possa ser encontrada em determinada posição de uma sequência de DNA, na prática os SNPs são, de forma geral, dialélicos. Uma das razões deste fenômeno é a baixa frequência de mutações do tipo substituição de um único nucleotídeo que originam os SNPs . Sendo assim, a probabilidade de ocorrer dois eventos de mutação em uma mesma posição de uma sequência de DNA é muito baixa (Vignal *et al.*, 2002). Portanto, considerando que

ocorra apenas um evento de mutação em cada ponto de uma sequência de DNA, apenas dois alelos são gerados: o ancestral (original) e o derivado. A baixa taxa de mutação dos SNPs faz destes marcadores bons medidores de eventos de evolução lenta, ou seja, eventos evolutivos que ocorreram há mais tempo (Sachidanandam *et al.*, 2001).

Os SNPs são altamente abundantes no genoma ocorrendo em média a cada 1000-2000 nucleotídeos quando dois cromossomos humanos são comparados (Cooper *et al.*, 1985.). Atualmente, milhões de SNPs do genoma humano já foram caracterizados e têm sido amplamente utilizados em diversos campos da genética.

1.1.2 - Microssatélites

Os microssatélites consistem em repetições *em tandem* de sequências curtas de nucleotídeos. Essas sequências repetidas podem ter de 1 a 6 nucleotídeos e constituem o motivo de repetição do microssatélite que podem classificá-lo como mono, di, tri, tetra, penta ou hexanucleotídeos (Buschiazzo e Gemmell, 2006).

Os alelos dos microssatélites correspondem ao número de repetições do motivo e, ao contrário dos SNPs que são geralmente dialélicos, os microssatélites são polimorfismos multialélicos. O grande número de alelos dos microssatélites ocorre

devido à natureza deste tipo de marcador: sequências repetidas *em tandem* tendem a ganhar e perder unidades repetidas facilmente devido a um processo conhecido como derrapagem (*slippage*) da DNA polimerase no momento da replicação do DNA (Ellegren, 2000). Este fenômeno faz dos microssatélites marcadores altamente polimórficos e com altas taxas de mutação (Zhang e Hewitt, 2003). Devido às altas taxas de mutação os microssatélites estão sujeitos a sofrer homoplasia (recorrência) e são, portanto, bons medidores de eventos evolutivos mais recentes, ou seja, marcadores de evolução rápida (Schlotterer, 2004).

Diante do exposto, os microssatélites são considerados o tipo de marcador molecular mais popular em estudos de genética e vêm sendo amplamente utilizados na genética de populações em diversas aplicações como detecção de estrutura populacional, testes de parentesco, estudos de diversidade genética, estudos de evolução recente de populações, entre outras (Zhang e Hewitt, 2003; Schlotterer, 2004).

1.1.3 - INDELS

Os INDELS são polimorfismos, normalmente dialélicos, causados pela inserção ou deleção de um ou mais nucleotídeos na sequência de DNA. Embora ocorram em uma frequência não tão alta quanto os SNPs, os INDELS vêm sendo reconhecidos como uma abundante fonte de variabilidade genética por estarem amplamente espalhados por

todo o genoma humano e corresponderem a cerca de 20% de todos os polimorfismos humanos (Mills *et al.*, 2006).

Assim como os SNPs, os INDELS são considerados marcadores estáveis, já que a chance de ocorrer dois eventos de inserção ou deleção em uma mesma posição genômica é muito baixa. Isto explica o fato dos INDELS serem, de forma geral, dialélicos e adequados para análise de eventos evolutivos mais antigos constituindo, portanto, marcadores de evolução lenta (Vali *et al.*, 2008).

Os INDELS começaram a ganhar destaque mais recentemente com a publicação do trabalho de Weber e colaboradores (Weber *et al.*, 2002) que descreveram 2.000 INDELS distribuídos por todo o genoma e estimou as freqüências alélicas destes marcadores em europeus, africanos, japoneses e nativo americanos. Desde então, os INDELS vêm sendo utilizados em diversos estudos evolutivos e populacionais.

1.1.4 - Marcadores de linhagem

Os marcadores de linhagem mtDNA e NRY apresentam características únicas que os tornam uma rica fonte de informação para estudos filogenéticos e de populações, tais como: (i) herança uniparental. O mtDNA e o cromossomo Y são transmitidos

exclusivamente de mãe para filhos e filhas e de pai para filho, respectivamente. Esta característica nos permite traçar a história genética exclusivamente materna ou paterna das populações humanas; (ii) a ausência de recombinação. Por constituírem genomas haplóides, o mtDNA e o NRY não sofrem recombinação e são, portanto, herdados em blocos. A ausência de recombinação faz com que a única fonte de variação entre duas moléculas seja a mutação. Assim, as mutações são acumuladas de forma cronológica (seqüencial) nas moléculas de mtDNA e NRY, gerando moléculas cada vez menos relacionadas entre si, até que se tornem linhagens independentes conhecidas como haplótipos. Relações entre haplótipos podem então ser feitas através de redes filogenéticas onde as mutações são classificadas em níveis hierárquicos e permitem traçar a rota que determinada linhagem percorreu servindo, portanto, como bons marcadores geográficos (Underhill e Kivisild, 2007); (iii) o mtDNA possui uma alta taxa de mutação, o que nos permite diferenciar populações separadas por um intervalo pequeno de tempo, como acontece com as populações humanas; (iv) o mtDNA aparece em um grande número de cópias em cada célula humana. Esta propriedade facilita sua obtenção para análises genéticas, sendo de especial importância para estudos envolvendo amostras de DNA antigo, o qual representa uma ferramenta chave para reconstrução genética do passado das populações humanas (Jehaes, 1999; Pena, 2002; Pakendorf e Stoneking, 2005).

Como as mutações do mtDNA e do NRY ocorrem de forma seqüencial e não sofrem recombinação, estes marcadores nos dão informações relativas a muitas gerações

atrás. Estas informações são muito úteis, pois nos permitem reconstruir a história da humanidade a partir da migração de mulheres e homens, respectivamente. No entanto, é importante lembrar que os marcadores de linhagem representam uma porção muito pequena do genoma de uma pessoa. Eles informam apenas a origem ancestral materna ou paterna de um determinado indivíduo. Desta forma, os marcadores de linhagem são pouco informativos em estudos em nível individual, mas adequados para estudos em nível populacional.

Nos últimos 20 anos, os marcadores de linhagem foram amplamente utilizados em estudos populacionais como ferramenta para responder uma série de perguntas relativas à história evolucionária humana como eventos demográficos, fluxo gênico, subestruturação e expansão populacional (Brown, 1980; Cann *et al.*, 1987; Hammer, 1994; Jobling *et al.*, 1996; Kayser *et al.*, 1997; Mathias *et al.*, 1994; Underhill *et al.*, 1997).

1.1.5 - Marcadores biparentais

Os marcadores biparentais estão localizados nos cromossomos autossômicos ou nas regiões pseudo autossômicas dos cromossomos X e Y, sendo herdados, portanto, de ambos os progenitores. Desta forma, os marcadores autossômicos são diplóides e estão sujeitos à recombinação.

Embora as filogenias traçadas por marcadores autossômicos sejam bastante complicadas devido aos eventos de recombinação, os marcadores biparentais são excelentes marcadores de individualidade. Ao contrário dos marcadores de linhagem, dependendo do número e localização dos marcadores analisados, os marcadores autossômicos podem representar uma grande proporção do genoma de um indivíduo, que são herdadas tanto do pai quanto da mãe.

Desta forma, embora aplicáveis à estudos em nível populacional, os marcadores biparentais autossômicos são ótimos marcadores para análises em nível individual.

1.2 - O Projeto Diversidade do Genoma Humano

O estudo da variabilidade genética humana é uma ferramenta de estudo de extrema importância em diversas áreas do conhecimento. Sua importância compreende desde estudos de doenças humanas e farmacogenômica até estudos da história, geografia e evolução das populações humanas.

O estudo da variabilidade genética aplicada à genética de populações começou por volta da Primeira Guerra mundial com a pesquisa da prevalência de diferentes tipos sanguíneos do sistema ABO em diferentes populações humanas. Durante o século XX, o aumento da disponibilidade de marcadores clássicos e os avanços na biologia

molecular forneceram grandes grupos de novos marcadores de DNA e os estudos da variabilidade genética humana se expandiram enormemente. No entanto, esta expansão continha informações fragmentadas, pois enquanto uns pesquisadores obtinham amostras de determinadas populações e testavam estas amostras para um grupo de marcadores, outros pesquisadores testavam amostras de outras populações para um grupo diferente de marcadores. Desta forma, os estudos de diferentes grupos de pesquisa eram incomparáveis e “incompletos” já que diferentes marcadores não eram testados nas mesmas populações.

Diante deste problema, em 1991 surgiu a idéia de se criar o Projeto Diversidade do Genoma Humano (HGDP). A proposta do HGDP era que amostras de DNA de populações dispersas por todo o mundo, ou seja, amostras representativas da espécie humana fossem coletadas e amplamente disponibilizadas para os pesquisadores interessados em estudar evolução e genética de populações humanas (Cavalli-Sforza *et al.*, 1991). Desta maneira, diferentes grupos de pesquisa poderiam utilizar as mesmas amostras de DNA e ter acesso a todas as análises realizadas por outros pesquisadores com marcadores moleculares distintos. Isto faria com que os estudos de diversos laboratórios pudessem ser comparados e complementados, permitindo análises mais apuradas acerca da diversidade genética humana.

Em 2002 a proposta do HGDP se concretizou e linhagens celulares de linfoblastos (LCL) de populações distribuídas pelo mundo todo foram coletadas por vários laboratórios (Cann *et al.*, 2002). As LCLs foram então armazenadas no Centro de Estudos de Polimorfismos Humanos (CEPH) da Fundação *Jean Dausset*, em Paris, e amostras de DNA suficientes para centenas de testes foram distribuídas para os diversos grupos de pesquisa de todo o mundo. Este conjunto de LCLs é conhecido como o painel de amostras do CEPH-HGDP (painel do CEPH-HGDP) e consiste em exatamente 1064 LCLs pertencentes a 52 populações oriundas de todas as partes do globo: África, Oriente Médio, Centro e Sul da Ásia, Leste da Ásia, Oceania, Europa e América (Figura 1).

O painel do CEPH-HGDP é considerado o mais completo painel de amostras mundiais de DNA humano existente até o momento e tem sido amplamente utilizado como uma importante fonte de estudos não só evolucionários e de genética de populações humanas, como também de estudos biomédicos.

No presente trabalho, utilizamos as amostras do painel do CEPH-HGDP em dois estudos distintos: no primeiro utilizamos INDELS autossômicos em amostras mundiais e em indivíduos brasileiros da cidade de São Paulo para estudar estrutura populacional mundial e origem ancestral de fragmentos cromossômicos de populações brasileiras. No segundo, utilizamos microssatélites localizados no

cromossomo Y para estudarmos a diversidade genética de populações mundiais e miscigenadas.

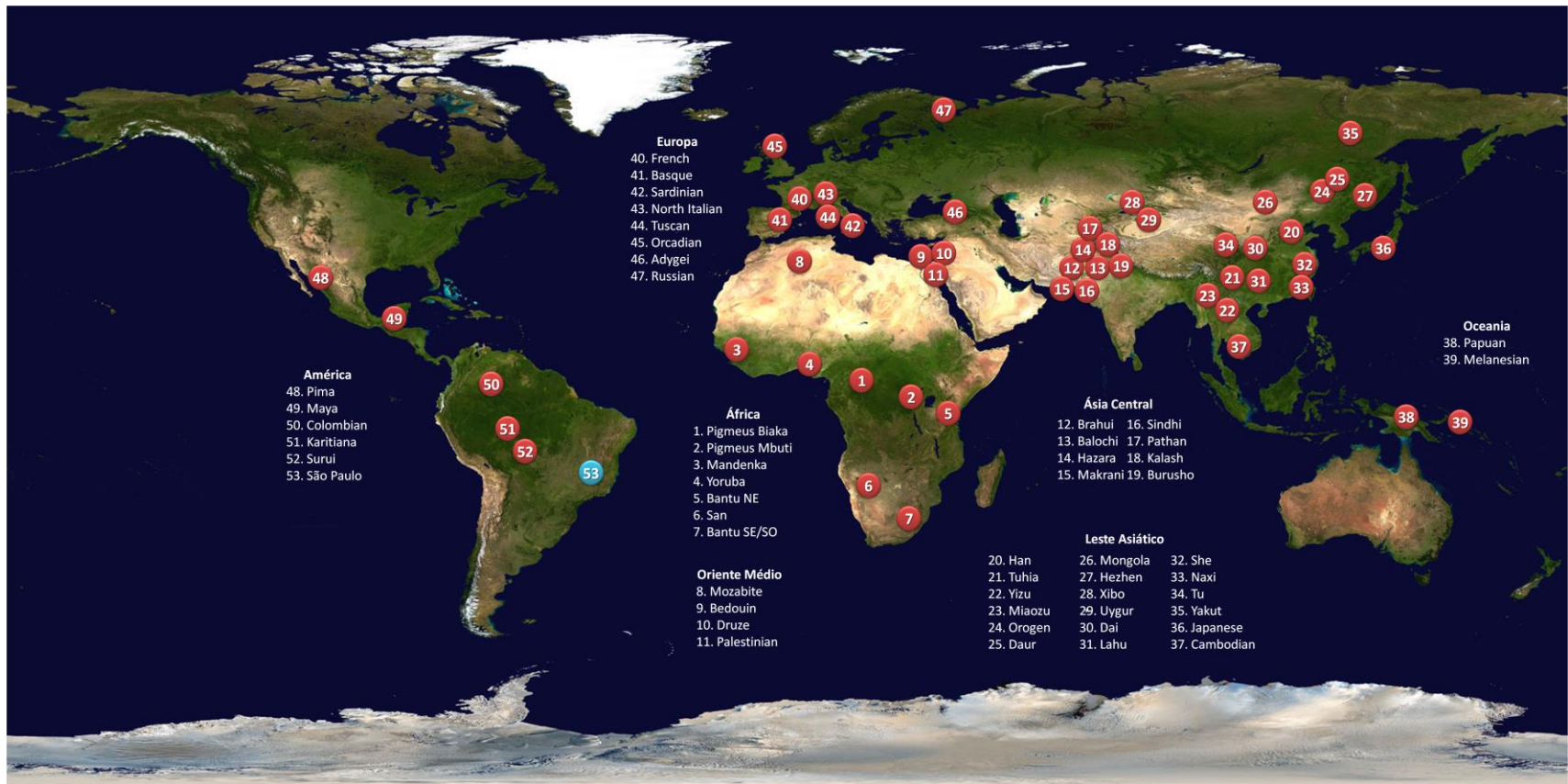


Figura 1- Localização geográfica e nome das populações do painel do CEPH-HGDP.

2 – Capítulo 1

Estrutura genética de populações humanas e análise da origem ancestral de fragmentos cromossômicos da população brasileira através do uso de INDELS informativos de ancestralidade

2.1 - Introdução

2.1.1 - A ocupação dos diferentes continentes pelo homem moderno

Evidências arqueológicas e moleculares (principalmente através do mtDNA e NRY) indicam uma origem única da humanidade moderna no nordeste africano há cerca de 200 mil anos atrás (Cann *et al.*, 1987; Underhill *et al.*, 2000). Durante aproximadamente 100 mil anos a humanidade permaneceu exclusivamente na África. Esta longa permanência da humanidade no continente africano resultou no acúmulo da grande diversidade genética, cultural, lingüística e fenotípica observadas hoje neste continente (Tishkoff e Williams, 2002).

Entretanto, há aproximadamente 80 mil anos o homem moderno migrou da África e colonizou os outros continentes (teoria “*Out of Africa*”) e, acredita-se que este fenômeno tenha ocorrido através de duas rotas. A primeira, conhecida como via costal ou rota sul, abrangeu, através da costa sul do continente asiático, o sul e o sudeste da Ásia, de onde houve uma bifurcação para o sul e para o norte do continente asiático. Através da bifurcação para o sul, o homem moderno atingiu a Oceania e a Austrália há cerca de 60-40 mil anos atrás. A expansão para o norte chegou até a China e o Japão. A segunda rota de expansão ocorreu há cerca de 40 mil anos atrás, e abrangeu a Ásia Central através do Oriente Médio. A partir da Ásia central, o homem moderno migrou em todas as direções atingindo a Europa, o

leste e o nordeste da Ásia. Há aproximadamente 15 mil anos atrás o homem moderno alcançou a América via estreito de Bering (Cavalli-Sforza e Feldman, 2003).

Estima-se que apenas cerca de 1200 pessoas deixaram a África em direção aos outros continentes (Macaulay *et al.*, 2005; Forster e Matsumura, 2005). Este pequeno número de indivíduos sugere que a colonização dos outros continentes ocorreu através de sucessivos efeitos fundadores seguidos de expansão populacional (Reed e Tishkoff, 2006).

O efeito fundador ocorre quando uma nova população é estabelecida a partir de um número pequeno de indivíduos provenientes de uma população maior original. Desta forma, como o número de indivíduos que deixa a população original é pequeno, o grupo que migra para estabelecer a nova população não representa toda a variabilidade genética da população original. Sendo assim, a principal consequência deste processo é a perda de variabilidade genética da nova população estabelecida em relação à população original.

Além disso, o pequeno grupo que migra para estabelecer a nova população pode carregar mutações que eram raras (com baixa frequência) na população original. Desta forma, as mutações raras na população original passam a ser super representadas na nova população. Por outro lado, mutações novas podem surgir

nas novas populações. Estes dois fenômenos juntos geram não só grandes diferenças de frequências alélicas entre as populações como também alelos específicos de determinadas populações ou regiões. Outro processo que cria diferenças de frequências alélicas entre populações é a deriva genética. A deriva genética é a flutuação aleatória, ou seja, independente do valor adaptativo, de frequências alélicas causadas simplesmente por uma amostragem randômica de gametas de uma geração para outra. Quando a diferença de frequências alélicas de um loco entre duas regiões geográficas é muito alta ($\sim 0,40$ ou maior), os alelos deste loco podem servir como marcos geográficos indicando a origem ancestral do loco em questão e são, portanto, chamados de marcadores informativo de ancestralidade (MIA) (Shriver *et al.*, 1997).

Portanto, à medida que o homem foi migrando através dos continentes, houve uma gradual diminuição da diversidade genética e as novas populações, agora fixadas nos diferentes continentes, permaneceram “isoladas” por um longo período de tempo (até as grandes navegações). Desta forma, as populações foram se diferenciando através do acúmulo de mutações neutras, adaptações ambientais e diferenças de frequências alélicas causadas pela deriva genética. Dado os fatos, diante de todos estes processos evolutivos, como ficou distribuída a variabilidade genética humana? Como estavam estruturadas estas populações?

2.1.2- A distribuição da variabilidade genética e estruturação das populações humanas

O primeiro estudo acerca da distribuição geográfica da diversidade humana foi realizado em 1972 por Richard Lewontin (Lewontin, 1972) que utilizou freqüências alélicas de 17 polimorfismos clássicos (determinados por eletroforese de proteínas) em amostras do mundo inteiro. As amostras estudadas foram agrupadas em oito grupos continentais: africanos, ameríndios, australianos, asiáticos do leste, asiáticos do sul, indianos, indivíduos da Oceania e caucasianos. A variabilidade foi então fracionada em três componentes: variabilidade entre continentes, entre populações dentro dos continentes e entre indivíduos dentro de populações. Os resultados mostraram que 85,4% da diversidade alélica estavam dentro das populações, 8,3% entre populações do mesmo grupo continental e somente 6,3% da diversidade alélica entre os grupos continentais.

Posteriormente, estudos utilizando polimorfismos de DNA confirmaram os resultados encontrados por Lewontin. Um deles foi o trabalho realizado por Rosenberg e colaboradores (2002) que genotiparam o painel do CEPH-HGDP (1064 indivíduos de 52 populações mundiais originadas de sete regiões geográficas) para 377 microssatélites autossômicos. A partição da variabilidade genética mostrou que 93,2% da diversidade se encontravam dentro das populações, 2,5% entre populações dentro da mesma região continental e apenas 4,3% entre as regiões continentais.

Mais recentemente, outro estudo de variabilidade genética humana foi realizado por nosso grupo (Bastos-Rodrigues *et al.*, 2006). Neste estudo, as mesmas amostras do painel do CEPH-HGDP utilizadas por Rosenberg foram genotipadas para um grupo de 40 INDELS autossômicos. Os resultados, mais uma vez, indicaram que a maior parte da variabilidade genética encontra-se dentro das populações (85,7%) e não entre populações dentro das regiões (2,3%) ou entre as regiões (12,1%).

Nestes mesmos estudos realizados por Rosenberg (Rosenberg *et al.*, 2002) e Bastos-Rodrigues (Bastos-Rodrigues *et al.*, 2006), os autores também testaram a capacidade dos marcadores analisados em detectar estrutura genética das populações sem nenhuma informação prévia a respeito da origem geográfica ou populacional das amostras. Nesta análise, a intenção era verificar se existe alguma correlação entre as regiões geográficas e a diversidade genética. Para isto, foi utilizado o programa *Structure*, no qual os indivíduos são agrupados em K grupos (Pritchard *et al.*, 2000). Este agrupamento é feito utilizando-se apenas dados genéticos, ou seja, os indivíduos são separados em subgrupos que apresentem diferentes frequências alélicas. Sendo assim, vários K s são testados e a probabilidade de cada K é calculada. Em ambos os trabalhos, o K mais provável foi o $K=5$. Isto significa que, geneticamente, os indivíduos estão estruturados em cinco grupos e, o interessante, é que os cinco grupos estimados pelos genótipos corresponderam às cinco grandes regiões geográficas: África Sub-Saariana, Leste Asiático, América, Oceania e Eurásia (composta por Ásia Central, Oriente Médio e Europa).

Mas, se a maior parte da variabilidade genética encontra-se entre indivíduos dentro de uma mesma população e não entre as regiões, como os indivíduos podem estar estruturados, geneticamente, em cinco grupos que correspondem às grandes regiões continentais?

Apesar da baixa porcentagem de variabilidade genética entre as regiões geográficas (4,3% e 12,2%), e a existência de poucos alelos região-específicos, a análise de genótipos de múltiplos locos permitiu a detecção de estruturação populacional correspondente às grandes regiões continentais. Isto ocorre porque embora a maioria dos alelos esteja amplamente distribuída por todos os continentes (o que explica uma baixa variabilidade entre as populações), as diferentes populações apresentam pequenas diferenças nas frequências alélicas. Sendo assim, quando muitos locos são analisados, estas pequenas diferenças de frequências alélicas são acumuladas e uma estruturação populacional pode ser detectada.

Portanto, de forma geral, as diferenças genéticas entre as populações humanas ocorrem mais por pequenas diferenças nas frequências alélicas do que por alelos região-específicos. No entanto, é importante ressaltar que alguns locos podem apresentar grandes diferenças de frequências alélicas entre diferentes regiões geográficas e podem, assim, servir como marcadores “diagnósticos” de origem ancestral, como o caso dos MIAs explicados anteriormente.

2.1.3- A colonização do continente americano

No processo de expansão do homem da África para outras regiões geográficas, a América foi o último continente a ser povoado. Estima-se que a entrada do homem neste continente tenha ocorrido através de uma única onda migratória partindo do leste asiático, passando pelo estreito de Bering até chegar à América do Norte. Da América do Norte a migração continuou até o extremo sul do continente originando as populações nativo-americanas, que constituíram pequenos grupos bastante isolados entre si (Cavalli-Sforza *et al.*, 1994).

Além disso, a história das Américas mostra-se ainda mais dinâmica, pois entre os séculos XV e XX, um grande número de europeus chegou às Américas, como consequência da política de expansão colonial europeia, lideradas por Portugal e Espanha. Os africanos também vieram, mas de modo forçado, através do tráfico de escravos e em maior número que os europeus (Goulart, 1975; Curtin, 1969). Estas migrações resultaram em um forte impacto negativo sobre a demografia das populações nativo-americanas e muitas destas populações terminaram dizimadas (Ribeiro, 1995; Bedoya *et al.*, 2006). Por outro lado, deste encontro de etnias surgiram as populações americanas atuais, híbridas, com diferentes graus de mistura genética, que variam de acordo com o número de populações parentais e intensidade de mistura (Sans, 2000; Bedoya *et al.*, 2006). Um exemplo de população americana atual é a população brasileira, que foi um dos focos de interesse desta tese.

2.1.4 – O retrato molecular do Brasil

O povo brasileiro é um dos mais heterogêneos do mundo. Vários estudos etnológicos sobre a origem do povo brasileiro têm enfatizado a nossa natureza triíbrida, resultado de um intenso processo de miscigenação que envolveu os europeus, os africanos e os ameríndios (Freyre, 1933; Ribeiro, 1995). Mas afinal, quanto há de africano, europeu e ameríndio na população brasileira? Como estes diferentes povos contribuíram para a composição genética do brasileiro atual?

Para responder a estas perguntas, às vésperas dos 500 anos do Brasil, nosso grupo de pesquisa iniciou um projeto intitulado “O Retrato Molecular do Brasil” para identificação da origem ancestral de linhagens mitocondriais e do cromossomo Y em uma grande amostra de brasileiros brancos de todas as regiões geográficas do Brasil (Alves-Silva *et al.*, 2000; Carvalho-Silva *et al.*, 2001). Os resultados destes estudos nos revelaram que a contribuição européia foi basicamente através dos homens, com mais de 98% dos cromossomos Y com origem neste continente, enquanto que a contribuição africana e ameríndia aconteceu, principalmente, através das mulheres (60%). Quando estendemos nossos estudos para população de indivíduos pretos brasileiros da cidade de São Paulo verificamos agora que 85% das linhagens mitocondriais eram africanas, 11,5% ameríndias e 3,5% européias, enquanto que as linhagens do cromossomo Y eram 48% africanas e 52% européias (Gonçalves *et al.*, 2008). Juntos, estes dados evidenciam a forte assimetria sexual que existiu entre o homem europeu e as mulheres africanas e ameríndias. Isto

indica um processo de exploração sexual, um fenômeno comum em sistemas coloniais.

Além disso, estes dados nos sugeriam que, no brasileiro, a cor da pele não é um bom indicativo sobre a ancestralidade do indivíduo, pelo menos para marcadores de linhagem uniparental (DNA mitocondrial e cromossomo Y). Mas, o quanto a cor da pele é indicativo da ancestralidade do brasileiro?

Para responder a esta pergunta, nosso grupo utilizou 10 marcadores autossômicos informativos de ancestralidade (MIAs), ou seja, marcadores com diferenças de frequências alélicas maiores que 0,40 entre africanos e Europeus (Parra *et al.*, 1998) e estimou o índice de ancestralidade africana (IAA) para cada indivíduo pertencente à diferentes categorias de cor de pele (brancos, pardos e pretos). Os resultados indicaram pequenas diferenças de IAA entre as diferentes categorias de cor de pele, indicando que, em nível individual, no Brasil, existe uma baixa relação entre cor da pele e ancestralidade genômica devido ao alto grau de miscigenação (Parra *et al.*, 2003).

Apesar de informativos, os trabalhos realizados com DNA mitocondrial e cromossomo Y têm como unidade de estudo **a população**, indicando a porcentagem **da população** que tem ancestralidade africana, europeia ou ameríndia. Já o trabalho realizado com os MIAs tem como unidade de estudo o

indivíduo, no entanto, a informação está apenas acerca da ancestralidade africana (e europeia, já que os MIAs eram informativos para África e Europa). Isto nos leva a mais uma pergunta: quanto há de africano, europeu e ameríndio em cada **indivíduo** brasileiro?

Esta pergunta foi respondida com mais um estudo de nosso grupo, onde 40 polimorfismos de inserção e deleção (INDELs) com alto poder de discriminação entre africanos, europeus e ameríndios, espalhados pelo genoma, foram analisados em amostras de indivíduos brancos, pardos e pretos brasileiros. Neste trabalho, a contribuição africana, europeia e ameríndia de cada indivíduo foi calculada e os valores médios destas contribuições foram estimados para cada categoria de cor de pele. Os dados mostraram que, em média, a contribuição europeia, africana e ameríndia é de respectivamente 0,78, 0,12 e 0,10 para cada indivíduo branco brasileiro, 0,62, 0,28 e 0,10 para cada indivíduo pardo e 0,37, 0,50 e 0,13 para cada indivíduo preto (Pena *et al.*, 2009). Estes resultados indicam que, embora os indivíduos pretos sejam mais misturados que os brancos, a grande maioria dos brasileiros apresenta proporções significativas (10% ou mais) de ancestralidade africana, europeia e ameríndia em seu genoma, o que caracteriza a população brasileira como uma população altamente miscigenada.

2.1.5- O genoma dos indivíduos miscigenados

Os trabalhos anteriormente citados mostram que o genoma de indivíduos miscigenados, como os brasileiros, apresenta uma mistura de alelos os quais são herdados de suas diferentes populações parentais. Mas, como estão estruturados os **cromossomos** destes indivíduos miscigenados?

Os cromossomos humanos são constituídos por centenas de milhares de blocos genômicos, de tamanhos variáveis, que se encontram em alto grau de desequilíbrio de ligação e são chamados blocos haplotípicos. O tamanho destes blocos haplotípicos variam na escala de centenas de milhares de pares de bases e depende de fatores demográficos e genômicos como, por exemplo, a existência de *hot-spots* de recombinação (Gabriel, *et al.*, 2002).

Podemos imaginar então que os cromossomos de indivíduos miscigenados são constituídos da junção de blocos haplotípicos sendo cada um proveniente de uma população parental específica (Tang *et al.*, 2006). Além disso, o processo de miscigenação gera um adicional e temporário desequilíbrio de ligação (que não é encontrado nas populações parentais), que abrange grandes extensões dos cromossomos (até 15cM), gerando longos segmentos que são herdados de maneira intacta das populações parentais (Collins-Schramn *et al.*, 2003). Como resultado, os cromossomos dos indivíduos miscigenados constituem verdadeiros mosaicos de

longos fragmentos com diferentes ancestralidades (blocos ancestrais). A distribuição dos tamanhos dos blocos ancestrais depende de fatores demográficos como o tempo de mistura das populações parentais e podem ser bastante informativos acerca da história evolutiva da população (Tang *et al.*, 2006).

Deste modo, é possível inferir a origem ancestral de um segmento cromossômico a partir da combinação da informação de múltiplos locos informativos de ancestralidade (MIAs) ao longo da região de interesse com a informação do desequilíbrio de ligação gerado pela mistura (Seldin *et al.*, 2004).

Dentro desta idéia, Seldin e colaboradores (2004) investigaram a possível utilização da população de afro-americanos em estudos de mapeamento por miscigenação (*admixture mapping*). Neste estudo, Seldin e colaboradores (2004) determinaram a origem ancestral de fragmentos cromossômicos de uma região do cromossomo 5 em afro-americanos, uma população considerada miscigenada. Para isto, Seldin utilizou 31 MIAs (15 INDELS e 16 SNPs) distribuídos em uma extensão de 60 cM na posição 100cM-161cM do cromossomo 5 e detectou um alto grau de miscigenação nos indivíduos, assim como os pontos de mudança de ancestralidade ao longo da região cromossômica estudada.

No presente trabalho, tivemos como objetivo ampliar nossos estudos relativos à miscigenação da população brasileira através de uma abordagem diferente

daquelas realizadas anteriormente por nosso grupo. Nossos estudos sobre as raízes ancestrais da população brasileira se iniciaram em nível populacional, com análise das linhagens do DNA mitocondrial e cromossomo Y. O passo seguinte foi analisar a ancestralidade genômica em nível individual, utilizando MIAs e INDELS espalhados por todo o genoma (autossômico). Aqui, nosso próximo passo é aumentar ainda mais a resolução destas análises. Nossa proposta é analisar o padrão de mistura da população brasileira, não apenas em nível individual como também em nível cromossômico (haplotípico), através da estratégia experimental desenhada por Seldin e colaboradores (2004) para determinação da origem ancestral de fragmentos ao longo de uma região do cromossomo 5. Esta abordagem nos permitirá dar um “*zoom*” no genoma do indivíduo brasileiro nos possibilitando, futuramente, medir a extensão dos blocos ancestrais e inferir a respeito do processo histórico e da dinâmica da miscigenação.

Neste trabalho nós utilizamos um painel de 31 INDELS (15 utilizados por Seldin mais 16 outros selecionados), localizados no braço longo do cromossomo 5, com ampla variação de frequência alélica entre África e Europa (MIAs) em dois estudos distintos. Primeiramente investigamos a estrutura populacional em um painel internacional de amostras, o painel do CEPH-HGDP. Posteriormente, nosso foco foi a população brasileira. Analisamos os 31 INDELS em amostras de indivíduos brancos e pretos da cidade de São Paulo, a fim de inferir a origem ancestral de fragmentos cromossômicos e caracterizar o padrão de mistura da população brasileira.

2.2 – Objetivos

2.2.1 – Objetivo geral

Este capítulo tem como objetivo geral estudar a variabilidade genética humana no painel mundial de amostras de DNA do CEPH-HGDP, através da utilização de 31 INDELS, com ampla variação de frequências alélicas entre africanos e europeus, localizados no braço longo do cromossomo 5. Em seguida, utilizar estes marcadores na caracterização, em nível cromossômico, do padrão de mistura da população brasileira, através da análise da origem ancestral de fragmentos cromossômicos de indivíduos pretos e brancos da cidade de São Paulo.

2.2.2 – Objetivos específicos

- Selecionar polimorfismos de inserção e deleção (INDELS) que sejam informativos de ancestralidade para África e Europa e que estejam presentes em uma região específica do braço longo do cromossomo 5 (q12.3-q33.3);
- Padronizar multiplexes e genotipar os INDELS selecionados para todas as amostras de DNA do painel do CEPH-HGDP e para as amostras de DNA de indivíduos pretos e brancos da cidade de São Paulo;
- Calcular frequências alélicas e fazer análise do equilíbrio de Hardy Weinberg para os INDELS em todas as populações estudadas;
- Fazer análise de desequilíbrio de ligação entre os marcadores analisados para todas as regiões mundiais estudadas e para as populações brasileiras;
- Verificar a distribuição da variabilidade genética entre as populações mundiais, através de uma hierárquica análise de variância molecular (AMOVA);
- Verificar estruturação populacional em populações mundiais, através do Escalonamento Multidimensional (MDS) e Análise dos Componentes Principais (PCA).

- Verificar, qualitativa e quantitativamente o poder discriminatório entre as populações africanas, européias e ameríndias dos marcadores selecionados e localizar as populações brasileiras em relação às parentais;

- Estimar as proporções de contribuição parental para pretos e brancos brasileiros em nível populacional e individual;

- Inferir, através de dados genotípicos dos INDELS, os haplótipos de cada indivíduo das populações da África, Europa, América, pretos e brancos de São Paulo;

- Determinar a origem ancestral de fragmentos cromossômicos de indivíduos pretos e brancos de São Paulo além dos africanos, europeus e ameríndios do painel do CEPH-HGDP;

- Investigar um possível padrão de cruzamento entre os indivíduos pretos e brancos brasileiros, ou seja, investigar a existência de casamentos assortativos;

2.3 – Material e Métodos

2.3.1 - Amostras de DNA

Para a análise de populações mundiais foi utilizado o painel internacional de diversidade humana do CEPH-HGDP (<http://www.cephb.fr/>; Cann *et al.*, 2002). Analisamos um total de 1045 amostras de DNA provenientes de 52 populações distribuídas em 7 regiões geográficas (124 amostras da África, 176 do Oriente Médio, 199 da Ásia Central, 247 do Leste Asiático, 36 da Oceania, 155 da Europa e 108 da América).

Para estudo da população brasileira, amostras de DNA de 104 indivíduos pretos e 157 indivíduos brancos da cidade de São Paulo foram analisadas. As amostras foram cedidas pelo Dr. Sérgio Bydlowski (Departamento de Hematologia, FMUSP) e os indivíduos foram autoclassificados como pretos ou brancos (denominação usada pelo IBGE), incluindo os seus pais e avós. O estudo foi anônimo e todos os indivíduos assinaram um termo de consentimento para o uso de sua amostra de DNA em trabalhos científicos. A utilização destas amostras para estudos científicos foi aprovada pela Comissão de Ética para Análise de Projetos de Pesquisa (CAPPesq) da Diretoria Clínica do Hospital das Clínicas da Faculdade de São Paulo em protocolo de pesquisa número 543/04.

2.3.2 - Amplificação dos INDELS do cromossomo 5

2.3.2.1 - Seleção de marcadores

Foram selecionados, para genotipagem, 29 polimorfismos de inserção/deleção (INDELS) bialélicos, localizados no braço longo do cromossomo 5, que apresentassem ampla variação de frequência alélica entre África e Europa (δ maior ou igual a 0,40). Os INDELS foram selecionados a partir da lista de INDELS descrita por Weber *et al* (2002) que está disponível no endereço eletrônico <http://research.marshfieldclinic.org/genetics/indels/default.asp>. Além dos 29 marcadores selecionados, incluímos em nossas análises mais dois INDELS (MID857 e MID1002) também localizados no braço longo do cromossomo 5, que já haviam sido genotipados e analisados por nosso grupo para as mesmas amostras estudadas no presente trabalho (Bastos-Rodrigues *et al.*, 2006). Assim, nossas análises foram realizadas com um total de 31 INDELS localizados no braço longo do cromossomo 5 em amostras de DNA do painel do CEPH-HGDP e em amostras de indivíduos pretos e brancos da cidade de São Paulo.

2.3.2.2 - Sistema *multiplex*

A amplificação de todos os INDELS do cromossomo 5 foi realizada através do sistema de *multiplex* onde em uma mesma reação de PCR são utilizados diversos pares de iniciadores para amplificação simultânea de diferentes marcadores.

Nossa análise englobou 29 marcadores que foram amplificados através de 4 *multiplexes* denominados A, B, C e D. Para os *multiplexes* A, B e C utilizamos 7 pares de iniciadores enquanto que para o *multiplex* D foram utilizados 8 pares.

2.3.2.3 - Amplificação do *multiplex* A

Para a reação de PCR do *multiplex* A, 1,5 μ l (~4ng) de DNA molde foram amplificados em um volume final de 10 μ l de reação. Cada reação continha 250 μ M de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,5 μ M do iniciador M13-40 marcado com a fluorescência FAM, iniciadores direto e reverso (Tabela 1) nas concentrações de 0,025 e 0,25 μ M para o marcador MID-878; 0,05 e 0,5 μ M para os marcadores MID-768, MID-1013, MID-1102 e MID-1272 e 0,1 e 1,0 μ M para os marcadores MID-719 e MID-990, 3 unidades de *Taq DNA platinum* (Invitrogen) em tampão de reação com 10mM de tris, 50mM KCl, 1,5mM MgCl₂, 0,1% de triton X-100, pH8,4. O ciclo de reação consistiu em uma desnaturação inicial a 95°C por 5 minutos, 29 ciclos a 95°C por 30 segundos, 50°C por 1 minuto e 72°C por 1 minuto.

A extensão final foi a 72°C por 10 minutos. Para todas as reações de PCR foi feito um controle branco de amplificação, onde água estéril foi acrescentada no lugar do DNA.

Tabela 1: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do *multiplex A*

MID	dbSNP ID	Iniciador direto	Iniciador reverso	INDEL	alelo menor*	alelo maior*
768	rs1160979	5' GTTTTCCCAGTCACGAC CATTACCAGTAGAGTGGGGA 3'	5' CTATGCCCTACTGGATCTAGG 3'	aag	89	92
879	rs1610963	5' GTTTTCCCAGTCACGAC GATGAAACCTCAGCAAAGA 3'	5' CATCATCACCAAAGCTTTC 3'	ataactaa	100	108
1013	rs1611095	5' GTTTTCCCAGTCACGAC CAGGATCTCATGCAGGATAC 3'	5' TGTGTTTAGCTTCCTGGAC 3'	ccag	133	137
719	rs1160938	5' GTTTTCCCAGTCACGAC CTGGAAGTCAGGAGGTTCTT 3'	5' TCCCAAGATGTCAAACTA 3'	ctc	144	147
1102	rs2067190	5' GTTTTCCCAGTCACGAC TCTCTTACCCCTTTCCTGTT 3'	5' GTATCAGAGCAATCCAGACG 3'	cttt	152	156
1272	rs2067354	5' GTTTTCCCAGTCACGAC AAGCAATTGGCATGTAATTC 3'	5' AAGGCATATTTTCAGGGAGT 3'	gctt	162	166
990	rs1611072	5' GTTTTCCCAGTCACGAC CTGGATCTGCAATATTCTGAA 3'	5' TGTCTTGGGCATATTATTTAACA 3'	tgaggcacagaaa	186	199

* Tamanho em pares de base (pb)

2.3.2.4 - Amplificação do *multiplex B*

Para a reação de PCR do *multiplex B*, 1,5µl (~4ng) de DNA molde foram amplificados em um volume final de 10ul de reação. Cada reação continha 250µM de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,5µM do iniciador M13-40 marcado com a fluorescência FAM, iniciadores direto e reverso (Tabela 2) nas concentrações de 0,025 e 0,25µM para o marcador MID-1191 e 0,05 e 0,5µM para os marcadores MID-1683, MID-883, MID-737, MID-1937, MID-739 e MID-1348, 3 unidades de *Taq DNA platinum* (Invitrogen) em tampão de reação com 10mM de Tris, 75mM KCl, 1,5mM MgCl₂, pH8,3. O ciclo de reação (*Step Down*) consistiu em uma desnaturação inicial a 95°C por 5 minutos, 5 ciclos a 95°C por 45 segundos, 60°C por 45 segundos e 72°C por 1 minuto, 5 ciclos a 95°C por 45 segundos, 58°C por 45 segundos e 72°C por 1 minuto, 5 ciclos a 95°C por 45 segundos, 56°C por 45 segundos e 72°C por 1 minuto, 5 ciclos a 95°C por 45 segundos, 54°C por 45 segundos e 72°C por 1 minuto, 5 ciclos a 95°C por 45 segundos, 52°C por 45 segundos e 72°C por 1 minuto e 25 ciclos a 95°C por 45 segundos, 50°C por 45 segundos e 72°C por 1 minuto. A extensão final foi a 72°C por 10 minutos. Para todas as reações de PCR foi feito um controle branco de amplificação, onde água estéril foi acrescentada no lugar do DNA.

Tabela 2: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do *multiplex B*

MID	dbSNP ID	Iniciador direto	Iniciador reverso	INDEL	alelo menor*	alelo maior*
1683	rs2307879	5' GTTTTCCCAGTCACGAC TTATTTTTGTTCTGCCAAC 3'	5' TTTTCCTTTTGCTGTTTCAC 3'	tta	92	96
1191	rs2067278	5' GTTTTCCCAGTCACGAC ACCCAGTTATGTGCTTTCTTT 3'	5' CTTTACTTGGGAGGGATAACC 3'	atgtcttcagctg	112	125
883	rs1610967	5' GTTTTCCCAGTCACGAC AAGGAGCTTTACAACACACG 3'	5' TTGCTAGTTATTATTGGCCG 3'	tataaatataaa	134	146
737	rs1160951	5' GTTTTCCCAGTCACGAC TTGTTTTATTTGAGATGGCA 3'	5' CTTTTTATTACTGTGCACCAAAT 3'	tat	138	141
1937	rs2308129	5' GTTTTCCCAGTCACGAC GGGTGGGAGTAGAGAGAAGA 3'	5' TCTTGACCCAGTTCATAAG 3'	agaa	151	155
739	rs1160953	5' GTTTTCCCAGTCACGAC TGTTCAGAAAATGCATCTCA 3'	5' AGGAAGAGGCTTTAAAGCAG 3'	tcta	157	161
1348	rs2307544	5' GTTTTCCCAGTCACGAC ACATTTTGTGCATAGGAGGA 3'	5' CAGTGACTTTCAGGGTCAAA 3'	aga	164	167

* Tamanho em pares de base (pb)

2.3.2.5 - Amplificação do *multiplex C*

Para a reação de PCR do *multiplex C*, 1,5µl (~4ng) de DNA molde foram amplificados em um volume final de 10µl de reação. Cada reação continha 250µM de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,25µM do iniciador M13-40 marcado com a fluorescência FAM, iniciadores direto e reverso (Tabela 3) nas concentrações de 0,0125 e 0,125µM para o MID-858; 0,025 e 0,25µM para os marcadores MID-776, MID-1838 e MID-1441; 0,05 e 0,5µM para o MID-1271 e 0,1 e 1,0 para o MID-1249, 3 unidades de *Taq DNA platinum* (Invitrogen) em tampão de reação com 10mM de Tris, 50mM KCl, 1,5mM MgCl₂, 0,1% de triton X-100, pH8,4. O ciclo de reação foi o mesmo descrito para o *multiplex B* (*Step Down*).

2.3.2.6 - Amplificação do *multiplex D*

Para a reação de PCR do *multiplex D*, 1,5µl (~4ng) de DNA molde foram amplificados em um volume final de 10µl de reação. Cada reação continha 250µM de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,25µM do iniciador M13-40 marcado com a fluorescência FAM, iniciadores direto e reverso (Tabela 4) nas concentrações de 0,025 e 0,25µM para o marcador MID-741; 0,05 e 0,5µM para os marcadores MID-942, MID-1052, MID-1004 e MID-1434 e 0,1 e 1,0 para os marcadores MID-1373, MID-971 e MID-1848, 3 unidades de *Taq DNA platinum* (Invitrogen) em tampão de reação com 10mM de Tris, 50mM KCl, 1,5mM MgCl₂,

Tabela 3: Seqüência de iniciadores e características dos INDELS do cromossomo 5 do *multiplex C*

MID	dbSNP ID	Iniciador direto	Iniciador reverso	INDEL	alelo menor*	alelo maior*
776	rs1610580	5' GTTTCCCAGTCACGAC GATTTCCCCTTTCTTCAGTC 3'	5' AGCACATTTTGTAGCTGGAG 3'	ctt	90	93
858	rs673423	5' GTTTCCCAGTCACGAC CTAGCTCTCAAGAACGCAG 3'	5' TACTGGCCTTCTGATCTCAG 3'	atatcagcgattctgaatgctgatt	114	139
1271	rs2067353	5' GTTTCCCAGTCACGAC TTTGGGCATTTCAGTATTTTG 3'	5' AGTGTCAAACCCAGTTGGTA 3'	atatt	118	123
1249	rs2067331	5' GTTTCCCAGTCACGAC TAATTTTGCATTTCAGCCTTC 3'	5' CACGGATAGAAAAACACTCAA 3'	tct	129	132
717	rs1160936	5' GTTTCCCAGTCACGAC GAGCAAATACTGAGCACCAT 3'	5' TATCACCTTGCTTTTGAAG 3'	attta	150	155
1838	rs2308034	5' GTTTCCCAGTCACGAC CCAGTTCCAGCTTCATAGTG 3'	5' AATAGAAACCGCTGAATGTG 3'	ata	160	163
1441	rs2307637	5' GTTTCCCAGTCACGAC TTTCTTCTCAGCTCTGTGCT 3'	5' ACAGACCATAGATCAAGGCA 3'	taat	175	179

*Tamanho em pares de base (pb)

Tabela 4: Sequência de iniciadores e características dos INDELS do cromossomo 5 do *multiplex D*

MID	dbSNP ID	Iniciador direto	Iniciador reverso	INDEL	alelo menor*	alelo maior*
942	rs1611025	5' GTTTTCCCAGTCACGAC CATGGCCTTAATTGCTACTG 3'	5' CACATGATCAAAAACCTGCTG 3'	ctt	110	113
1052	rs2067141	5' GTTTTCCCAGTCACGAC AAGCCTTCAGAGTCACAAAA 3'	5' CAAAAGCAGATGGCACTATT 3'	aaa	120	124
741	rs32861	5' GTTTTCCCAGTCACGAC CTGGACAGTGTAGACTGCCT 3'	5' TATCTCCCAACACAGGAAAA 3'	ctag	127	131
1004	rs1611086	5' GTTTTCCCAGTCACGAC TGCTAGCAAGTTCATTTTT 3'	5' AGTTGAGGCATCCTAACTGAT 3'	ctt	135	138
1848	rs2308044	5' GTTTTCCCAGTCACGAC TCCCAAATCACTTCATTTGT 3'	5' CTAGTTGCACTTTGGCTTTC 3'	atctttattgtaata	141	157
1373	rs2307569	5' GTTTTCCCAGTCACGAC TATGGCACAGTTATGGTCT 3'	5' GCCCTACAAGTAAACATGGA 3'	gtaactatta	144	154
1434	rs2307630	5' GTTTTCCCAGTCACGAC GCACATTGTTTTCCAGAAT 3'	5' TTAACCCAGTTAACGTTTGC 3'	agta	163	167
971	rs1611053	5' GTTTTCCCAGTCACGAC GGAAAAGTTCAGGAAAAACC 3'	5' ATATGGGTGCTCTTTTCCTT 3'	tac	181	184

*Tamanho em pares de base (pb)

0,1% de triton X-100, pH8,4. O ciclo de reação foi o mesmo descrito para o *multiplex B (Step Down)*.

2.3.3 - Genotipagem

Para a genotipagem dos INDELS do cromossomo 5 utilizamos o sequenciador MegaBACE 1000 (Molecular Dynamics). Os produtos de PCR foram diluídos em água na proporção de 1:3 e cerca de 2µl desta diluição foram misturados com solução contendo 0,25µl de ET-550-R *Size Standard* (GE Healthcare) e 7,75µl de *Tween* 0,1%. As amostras foram desnaturadas a 95°C por 3 minutos e injetadas no sequenciador a 3kV com uma voltagem de corrida de 10kV. O tempo de injeção foi de 80 segundos e 75 minutos de corrida. A análise do tamanho dos alelos foi realizada utilizando os programas *Genetic Profiler* (versão 2.2) e *Fragment Profiler* (versão 1.2), ambos da GE Healthcare.

2.3.4 – Análise dos dados

No presente trabalho analisamos INDELS do cromossomo 5 em populações mundiais e em populações brasileiras. Como as análises estatísticas aplicadas em cada um dos estudos foram realizadas separada e independentemente, elas estão separadas em

diferentes tópicos a seguir (tópicos 2.3.4.1 e 2.3.4.2), com o intuito de facilitar sua compreensão.

2.3.4.1 – Análise das populações mundiais

2.3.4.1.1 - Parâmetros básicos das populações mundiais

Parâmetros básicos de genética de populações como, frequências alélicas e análise de equilíbrio de Hardy-Weinberg foram calculadas utilizando o pacote de programas de análise de dados para genética de populações, *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000) com número de permutações de 10000. As frequências alélicas foram calculadas tanto para as grandes regiões (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América) quanto para as 52 populações separadamente, enquanto que o equilíbrio de Hardy-Weinberg foi testado para os 31 locos nas 52 populações.

2.3.4.1.2 – Análise de Variância Molecular (AMOVA)

A fim de verificar a distribuição da variabilidade genética entre as diferentes populações mundiais, realizamos uma análise de variância molecular (AMOVA) (Excoffier *et al.*, 1992) utilizando o programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000). Analisamos a distribuição da variabilidade em três condições: a) o mundo

como uma única região com 52 populações; b) o mundo dividido em 7 regiões geográficas (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América) abrangendo as 52 populações; c) cada região separadamente abrangendo suas respectivas populações.

2.3.4.1.3 - Escalonamento Multidimensional (MDS) em populações mundiais

Para verificar os padrões de semelhança entre as diferentes populações, utilizamos o programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000) para calcular matrizes de distância genética baseada em F_{st} normalizado (Distância de Reynolds) (Reynolds *et al.*, 1983) tanto para as grandes regiões (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América) quanto para as 52 populações separadamente. A fim de visualizar as distâncias entre as populações e/ou regiões, as matrizes de distância foram representadas graficamente através do escalonamento multidimensional (MDS) (Kruskal & Wish, 1978), realizado pelo programa *Statistica* (versão 7.0).

2.3.4.1.4 - Diferenciação genética entre indivíduos das populações mundiais

Além de verificar os padrões de semelhança entre as diferentes populações estudadas, analisamos também o padrão de similaridade entre os diferentes indivíduos analisados. Para isto nós construímos uma matriz de distância baseada em compartilhamento de alelos idênticos por estado (*IBS distance matrix*). Esta matriz fornece a fração de alelos idênticos por estado compartilhada por cada par de indivíduos analisados. Posteriormente, esta matriz foi graficamente representada através da construção de árvores filogenéticas *neighbor-joining*. A matriz de distância foi gerada pelo programa *Plink* (Purcell *et al.*, 2007). A árvore foi gerada através do módulo *neighbor* do programa *Phylip v.3.64* (<http://evolution.genetics.washington.edu/phylip.html>) e os ramos foram coloridos com auxílio do programa *HyperTree v.1.0.0* (<http://www.kinase.com/tools/HyperTree.html>).

2.3.4.1.5 - Análise de desequilíbrio de ligação nas regiões mundiais

Para verificar se os marcadores ao longo do cromossomo 5 se encontravam em desequilíbrio de ligação, utilizamos o programa *HaploView* (Barrett *et al.*, 2005). Além de indicar os marcadores que se estão em desequilíbrio de ligação, o programa fornece as frequências haplotípicas para os marcadores que se encontram em desequilíbrio de ligação. O *HaploView* foi aplicado separadamente em cada região

geográfica e o desequilíbrio de ligação foi considerado significativo para valores de r^2 maiores ou iguais a 0,8.

2.3.4.1.6 - Análise do Componente Principal (PCA) em populações mundiais

Para verificar estruturação genética baseada nos dados dos INDELS do cromossomo 5 realizamos a Análise do Componente Principal (PCA). A Análise do Componente Principal aplicada à genética de populações procura agrupar, baseado em dados de frequências alélicas e/ou haplotípicas, a variação genética em um número mínimo de dimensões que descrevam ao máximo a variabilidade genética. Desta forma, os resultados podem ser visualmente examinados quanto à presença ou não de estrutura genética.

A análise do componente principal foi realizada com o auxílio do programa *Statistica* (versão 7.0) para as regiões geográficas e para as 52 populações separadamente. Na PCA com as regiões geográficas foram utilizadas frequências alélicas geradas pelo *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000) e, quando os marcadores estavam em desequilíbrio de ligação em todas as regiões geográficas analisadas, frequências haplotípicas. Já na PCA realizada com as 52 populações, foram utilizadas apenas frequências alélicas, já que não foi possível a inferência de frequências haplotípicas pelo *HaploView* (Barrett *et al.*, 2005) devido ao tamanho reduzido das populações.

2.3.4.2 – Análise das populações brasileiras

Na análise das populações brasileiras utilizamos, além das amostras de indivíduos brancos e pretos da cidade de São Paulo, as populações do painel do CEPH-HGDP provenientes da África, Europa e América, que correspondem às populações parentais da população brasileira.

2.3.4.2.1 – Parâmetros básicos das populações brasileiras

Assim como para as populações mundiais, o cálculo das freqüências alélicas e análise de equilíbrio de Hardy-Weinberg para as populações brasileiras foram realizados também pelo programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000). As freqüências alélicas foram calculadas separadamente para os brancos e para os pretos e a análise de equilíbrio de Hardy-Weinberg foi realizada para os 31 marcadores em cada uma das duas populações brasileiras.

2.3.4.2.2 - Escalonamento Multidimensional (MDS) das populações brasileiras

A distância genética entre as populações parentais e as populações brasileiras foi calculada através do programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000). Nesta análise, duas matrizes de distância de Reynolds (Reynolds *et al.*, 1983) foram geradas: uma onde as populações parentais foram tratadas como grandes regiões geográficas (África, Europa e América) e outra onde as populações de cada região geográfica parental foram discriminadas. Assim, a primeira matriz continha 5 populações (África, Europa, América, brancos e pretos brasileiros), enquanto que a segunda continha 22 populações (7 populações africanas, 8 populações europeias, 5 populações ameríndias, brancos e pretos brasileiros). A representação gráfica das matrizes de distância foi realizada através do MDS (Kruskal & Wish, 1978). Para isto foi utilizado o programa *Statistica* (versão 7.0).

2.3.4.2.3 - Análise de desequilíbrio de ligação nas populações brasileiras

Assim como para as regiões mundiais, o desequilíbrio de ligação entre os marcadores localizados no braço longo do cromossomo 5 foi testado também para as populações de pretos e brancos da cidade de São Paulo. Para isto, foi utilizado o programa *HaploView* (Barrett *et al.*, 2005) já descrito no item 2.3.4.1.4. O desequilíbrio de ligação foi considerado significativo para valores de r^2 maiores ou iguais a 0,8.

2.3.4.2.4 - Análise do Componente Principal (PCA) com populações brasileiras

Para verificar se os INDELS selecionados eram capazes de distinguir as três populações parentais e localizar as populações de brancos e pretos de São Paulo em relação a África, América e Europa, realizamos uma análise do componente principal. Para isto, utilizamos o programa *Statistica* (versão 7.0), onde duas análises foram realizadas: a primeira onde as populações parentais foram tratadas como grandes regiões geográficas (África, Europa e América) e outra onde as populações de cada região geográfica parental foram discriminadas. Na primeira PCA foram utilizadas frequências alélicas geradas pelo *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000) e, quando os marcadores estavam em desequilíbrio de ligação nas 5 populações analisadas (africanos, europeus, ameríndios, brancos e pretos brasileiros), foram utilizadas frequências haplotípicas geradas pelo *HaploView* (Barrett *et al.*, 2005). Já na PCA realizada com as populações de cada região geográfica discriminadas, foram utilizadas apenas frequências alélicas, já que não foi possível a inferência de frequências haplotípicas pelo *HaploView* (Barrett *et al.*, 2005), devido ao tamanho reduzido das populações.

2.3.4.2.5 – Proporções de contribuição parental nas populações brasileiras

As proporções de contribuição parental nas populações de brancos e pretos de São Paulo foram inferidas através do programa *Admix 2.0*. (Dupanloup e Bertourelle, 2001). O programa *Admix 2.0* (Dupanloup e Bertourelle, 2001) é um programa de inferência das proporções de mistura em populações miscigenadas que leva em consideração não só a diferença de frequências alélicas entre as populações parentais como também as distâncias moleculares entre alelos.

Em nossas análises com o *Admix 2.0* (Dupanloup e Bertourelle, 2001), utilizamos as frequências alélicas absolutas das populações parentais (África, Europa e América) e miscigenadas (de indivíduos brancos e pretos de São Paulo) geradas pelo *Arlequin* (Schneider *et al.*, 2000), com tempo de mistura de 500 anos e taxa de mutação de 1×10^{-7} por ano.

2.3.4.2.6 – Estrutura populacional

A fim de verificar se os INDELS selecionados seriam capazes de distinguir as três populações parentais, inferimos a estrutura populacional das populações parentais. Para isto utilizamos o programa *Structure 2.3.3* (Pritchard *et al.*, 2000).

O programa *Structure* faz inferências sobre estrutura populacional baseado em dados genotípicos de múltiplos locos. O programa assume que existem K populações (onde o K é determinado pelo usuário) que são caracterizadas por um grupo de frequências alélicas para cada loco. O programa procura o agrupamento que melhor atenda ao equilíbrio de Hardy-Weinberg e que não apresente desequilíbrio de ligação. Assim, os indivíduos são probabilisticamente inseridos em uma dada população (modelo de não mistura) ou em mais de uma população quando seus genótipos indicam que são indivíduos miscigenados (modelo de mistura). Desta forma, no modelo de não mistura, o programa nos fornece a probabilidade de cada indivíduo pertencer a cada um dos K s inferidos pelo programa. Já no modelo de mistura, o programa nos fornece as proporções do genoma de cada indivíduo que são provenientes de cada K inferido pelo programa. Isto pode ser feito sem nenhuma informação prévia a respeito da população de origem dos indivíduos.

Em 2003, Falush implementou uma segunda versão do programa *Structure* (Falush *et al.*, 2003) acrescentando o modelo de ligação. Este modelo é uma extensão do modelo de mistura que leva em consideração o desequilíbrio de ligação criado pela miscigenação. O modelo assume que a ancestralidade de um loco não ocorre de forma independente da ancestralidade dos locos adjacentes. Isto nos permite analisar dados de marcadores localizados em um único cromossomo, sendo, portanto, o modelo adequado a ser empregado no presente estudo.

Em nossa primeira análise com o *Structure 2.3.3*, utilizamos o modelo de ligação apenas com as populações parentais (África, América e Europa) a fim de verificar se os dados dos INDELS do cromossomo 5 seriam capazes de distinguir as três populações parentais e estimar a porcentagem individual média de ancestralidade africana, europeia e ameríndia em cada população parental. Para isto realizamos 5 corridas independentes com cada K que variou de 1 a 5. Cada corrida consistiu em 50000 passos de *burn-in* seguidos de 250000 interações da Cadeia de Markov Monte Carlo. Nesta análise, não foi utilizada nenhuma informação inicial a respeito das populações de origem dos indivíduos amostrados.

2.3.4.2.7 – Inferência de haplótipos

A partir de dados genotípicos dos INDELS do cromossomos 5 dos africanos, europeus, ameríndios, brancos e pretos de São Paulo, os dois haplótipos de cada indivíduo foram estimados para cada população, através do emprego do programa *Phase* (versão 2.1) (Stephens *et al.*, 2001) para futura análise da origem ancestral de fragmentos cromossômicos. O programa *Phase* é um método estatístico de reconstrução haplotípica a partir de dados genotípicos que se baseia na amostragem Gibbs (tipo de algoritmo Cadeia de Markov Monte Carlo). O método consiste primeiro, na resolução de haplótipos **não** ambíguos que vão então constituir uma lista de haplótipos

resolvidos. Em seguida, ocorre a resolução dos haplótipos ambíguos baseado na frequência e na distância genética em relação aos haplótipos resolvidos.

O *Phase* (versão 2.1) foi empregado de modo independente em cada uma das cinco populações (africanos, europeus, ameríndios, brancos e pretos de São Paulo) e todas as corridas foram realizadas com número de interações de 100, *burn-in* de 100 e assumindo recombinação entre os locos.

2.3.4.2.8 - Origem ancestral dos fragmentos cromossômicos das populações miscigenadas

Conforme mencionado no item 2.3.4.2.6, o modelo de ligação do programa *Structure* leva em consideração o desequilíbrio de ligação criado pela miscigenação e, desta forma, nos permite analisar dados de marcadores localizados em um único cromossomo. Desta forma podemos inferir, além das proporções de ancestralidade dos indivíduos (que é possível também pelo modelo de mistura do *Structure*), a ancestralidade de um loco particular dos indivíduos amostrados. Isto é interessante pois, quando utilizamos dados em fase gamética (haplótipos) é possível visualizar a extensão da mistura dos indivíduos ao longo de cada um de seus cromossomos.

Desta forma, utilizando as populações parentais como referência, a provável origem ancestral de cada loco foi estimada para cada cromossomo de cada indivíduo das populações brasileiras. Para isto utilizamos os haplótipos das populações brasileiras e parentais inferidos pelo programa *Phase* (versão 2.1) e aplicamos o modelo de ligação do programa *Structure* (versão 2.2) utilizando as cinco populações (africanos, europeus, ameríndios, brancos e pretos brasileiros). Nesta análise foram utilizadas informações iniciais a respeito da origem geográfica apenas para os indivíduos das populações parentais (África, América, Europa) para que servissem de referência. A corrida foi realizada com $K=3$ e consistiu em 50000 passos de *burn-in* seguidos de 250000 interações da Cadeia de Markov Monte Carlo. A porcentagem média individual de ancestralidade africana, européia e ameríndia dos pretos e brancos da cidade de São Paulo também foi inferida.

2.3.4.2.9 – Correlação de ancestralidade entre cromossomos homólogos

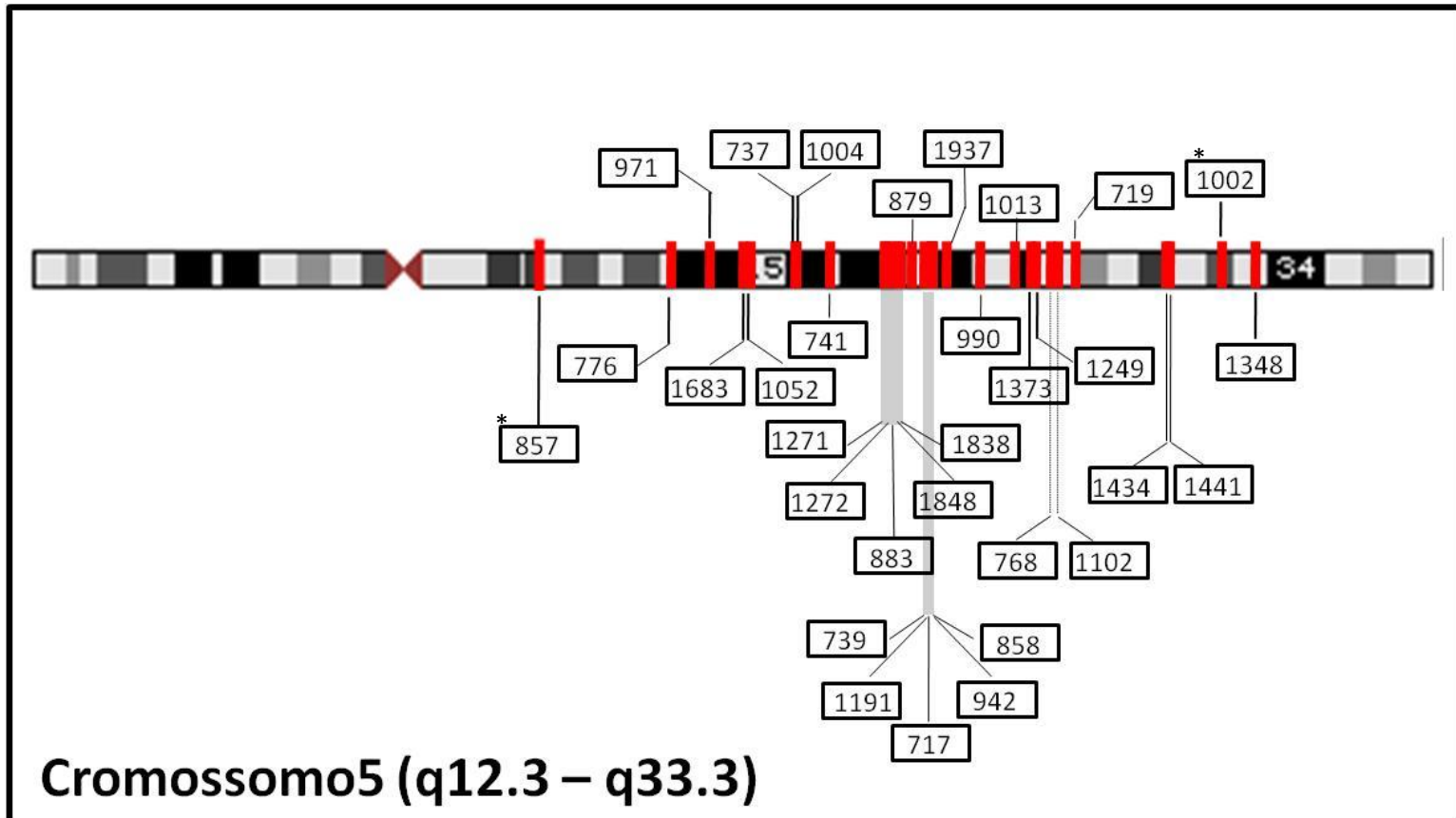
Para investigar um possível padrão de cruzamento entre os indivíduos brasileiros, realizamos uma correlação entre o número de locos de uma mesma origem ancestral entre os dois pares homólogos do cromossomos 5 de um mesmo indivíduo. Para isto, os dados dos indivíduos pretos e brancos brasileiros gerados pelo *Structure* foram submetidos a uma análise de correlação de *Spearman*, através do *website* de análises estatísticas *VassarStat* (<http://faculty.vassar.edu/lowry/VassarStats.html>).

2.4 – Resultados

2.4.1 – Indels do cromossomo 5

O presente trabalho teve como enfoque dois estudos distintos. Primeiramente investigamos a estrutura populacional em um painel internacional de amostras, o painel do CEPH-HGDP. Posteriormente, nosso foco foi o estudo das raízes ancestrais da população brasileira. No entanto, em relação à população brasileira, nossa preocupação não estava mais nas porcentagens de ancestralidade em nível populacional. Nossa idéia foi estudar o padrão de mistura em nível individual em uma resolução maior do que aquela oferecida pela análise do genoma como um todo. Nossa proposta foi tentar uma análise em nível haplotípico através da determinação da origem ancestral de fragmentos ao longo de uma região do cromossomo 5.

Para realização dos dois estudos, selecionamos 29 polimorfismos de inserção e deleção (INDELs) bialélicos, com ampla variação de freqüências alélicas entre africanos e europeus, localizados no braço longo do cromossomo 5. Um mapa do cromossomo 5 com a localização dos INDELs selecionados está representado na Figura 2. Os INDELs selecionados foram descritos por Weber e colaboradores (2002) e estão disponíveis no endereço eletrônico <http://research.marshfieldclinic.org/genetics/indels/default.asp>.



*Marcadores já tipados por nosso grupo (Bastos-Rodrigues *et al.*, 2006)

Figura 2: – Mapa físico do cromossomo 5 com a localização dos INDELs selecionados.

Algumas características dos INDELS como localização no cromossomo e frequências alélicas encontradas por Weber em africanos, europeus e ameríndios estão representados na tabela 5.

Fizemos a tipagem genotípica dos 29 INDELS em todas as populações mundiais do painel do CEPH-HGDP (1045 indivíduos distribuídos em 52 populações que abrangem 7 grandes regiões geográficas) e em duas populações brasileiras: brancos (157 indivíduos) e pretos (104 indivíduos) da cidade de São Paulo. A genotipagem dos 29 marcadores foi realizada através de quatro multiplex cujos perfis de picos estão representados nas figuras 3 e 4.

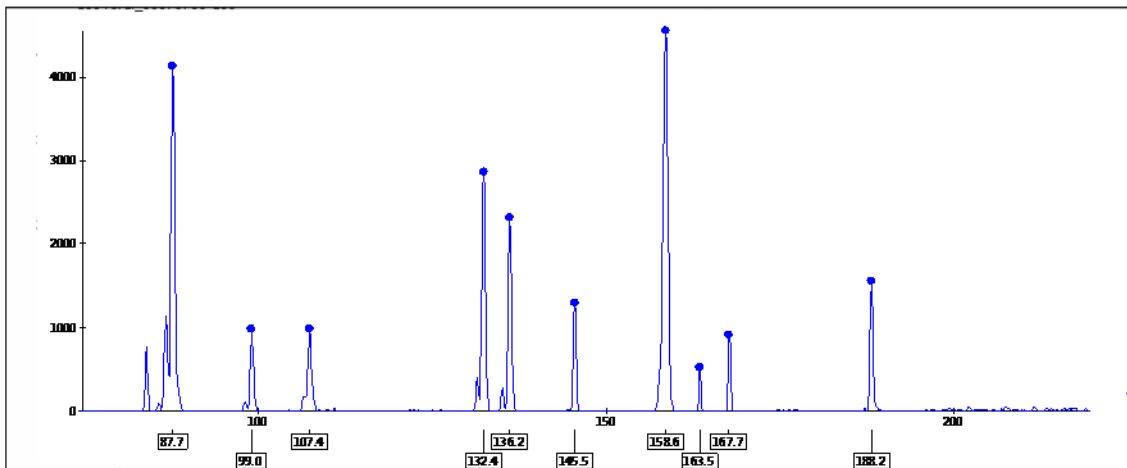
2.4.2 – Populações Mundiais

Após a genotipagem de todas as amostras pertencentes às 52 populações mundiais do painel do CEPH-HGDP, iniciamos as análises de frequências alélicas e teste de equilíbrio de Hardy-Weinberg. As frequências alélicas nas 52 populações assim como as frequências alélicas em cada região geográfica estão representadas na tabela 6. O teste de equilíbrio de Hardy-Weinberg mostrou que dentre os 1612 valores obtidos 81 ($\sim 0,05$) foram significativos em nível de 0,05, não havendo assim evidência de desvios.

Tabela 5: Características dos INDELS (MIDs) selecionados. Nome, localização no cromossomo 5 e frequências alélicas encontradas por Weber *et al* (2002) em populações da África, Europa e Américas

MID	dbSNP ID	Localização no cromossomo 5 (pb)	Frequência do alelo maior		
			África	Europa	América
857*	rs1610942	65.488.401	0,73	0,28	0,66
776	rs1610580	82.368.959	0,93	0,55	0,98
971	rs1611053	87.538.762	0,83	0,45	0,23
1683	rs2307879	91.280.252	0,82	0,19	0,52
1052	rs2067141	92.405.460	0,84	0,41	0,87
737	rs1160951	98.237.631	0,95	0,49	0,29
1004	rs1611086	98.305.629	0,30	0,71	0,95
741	rs32861	102.567.910	0,70	0,34	0,42
1271	rs2067353	109.801.765	0,88	0,04	0,12
1272	rs2067354	109.814.809	0,93	0,11	0,27
883	rs1610967	110.877.151	0,04	0,71	0,32
1848	rs2308044	111.691.965	0,20	0,70	0,80
1838	rs2308034	111.727.296	0,76	0,33	0,12
879	rs1610963	112.274.948	0,99	0,47	0,28
739	rs1160953	115.653.921	0,80	0,34	0,59
1191	rs2067278	115.788.388	0,98	0,42	0,85
717	rs1160936	115.815.319	0,93	0,37	0,83
942	rs1611025	115.927.562	0,78	0,40	0,54
858	rs673423	116.117.136	0,19	0,60	0,33
1937	rs2308129	118.150.671	0,76	0,33	0,66
990	rs1611072	122.251.997	0,00	0,27	0,00
1013	rs1611095	126.870.901	0,76	0,22	0,67
1373	rs2307569	128.655.437	0,96	0,56	0,66
1249	rs2067331	129.551.741	0,62	0,22	0,68
768	rs1160979	131.351.026	0,86	0,17	0,68
1102	rs2067190	132.124.175	0,76	0,15	0,46
719	rs1160938	134.405.689	0,91	0,23	0,71
1434	rs2307630	146.033.040	0,69	0,72	0,59
1441	rs2307637	146.187.115	0,61	0,23	0,31
1002*	rs1611084	153.636.208	0,40	0,50	0,11
1348	rs2307544	157.541.175	0,85	0,30	0,16

A)



B)

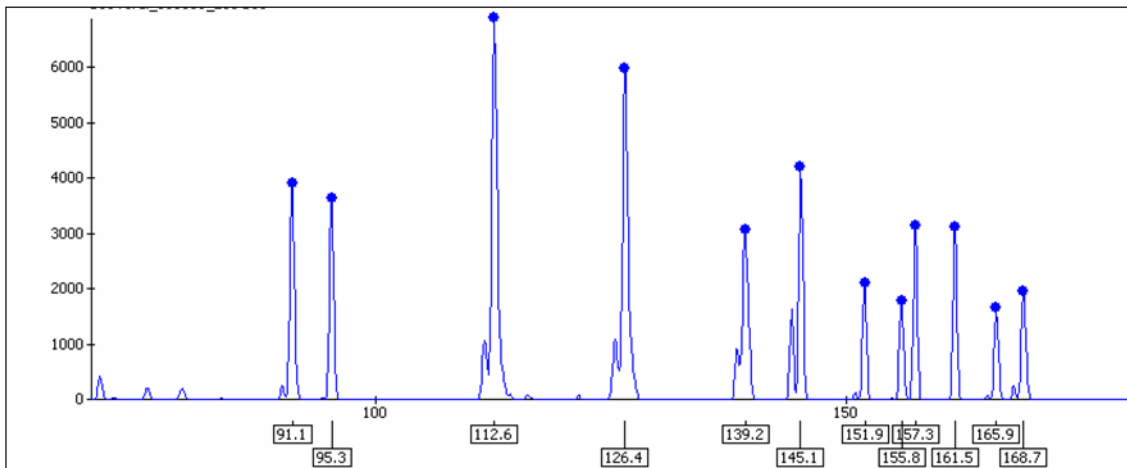
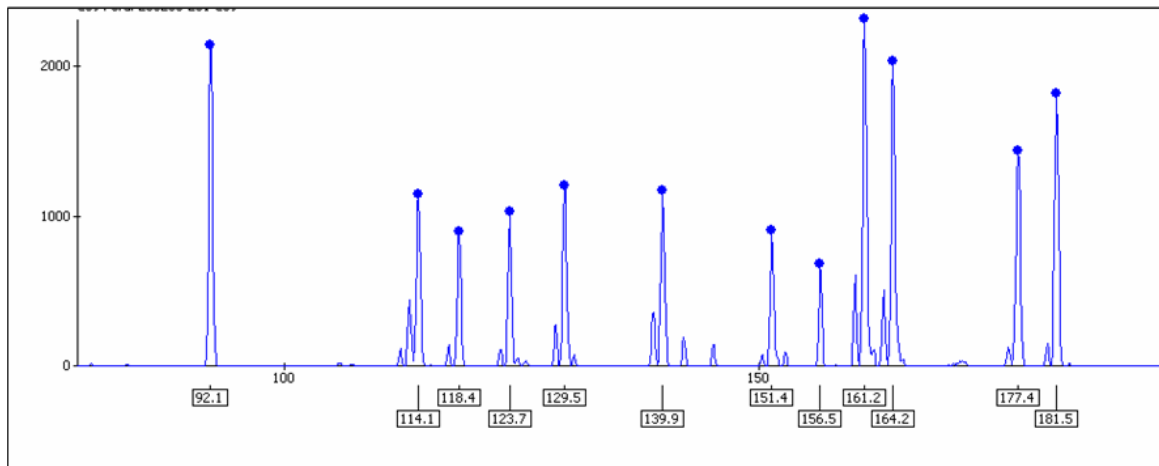


Figura 3: Eletroferograma produzido pelo programa *Fragment Profile*, versão 1.2, representando a genotipagem dos INDELS do cromossomo 5. A) Exemplo da genotipagem de indivíduo para o *Multiplex A*; B) Exemplo da genotipagem de indivíduo para o *Multiplex B*.

C)



D)

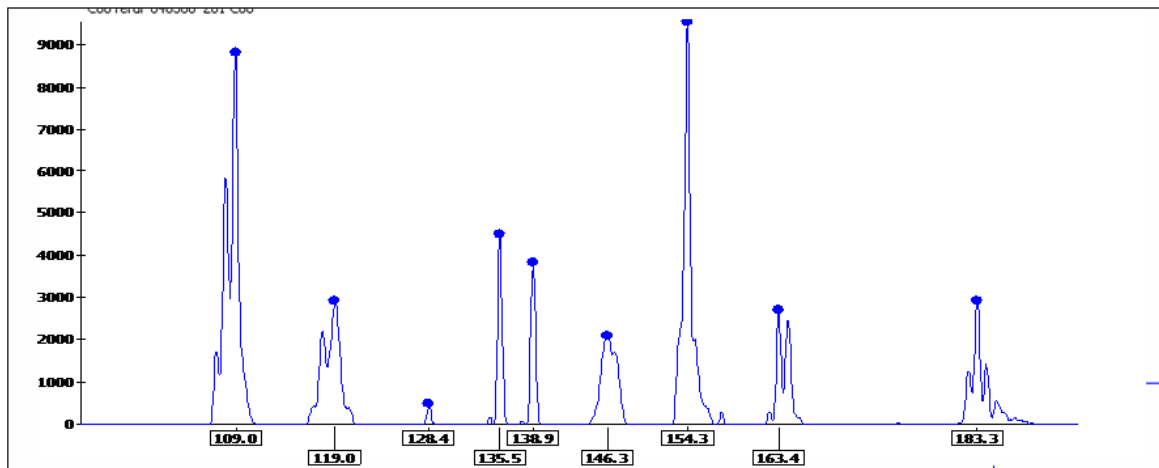


Figura 4: Eletroferograma produzido pelo programa *Fragment Profile*, versão 1.2, representando a genotipagem dos INDELS do cromossomo 5. C) Exemplo da genotipagem de indivíduo para o *Multiplex C*; D) Exemplo da genotipagem de indivíduo para o *Multiplex D*.

Tabela 6 – Frequências alélicas dos 31 INDELS do cromossomo 5 em cada população e região geográfica do painel do CEPH-HGDP

MID	dbSNP ID	Frequência do alelo maior							
		Biaka (n=33)	Mbuti (n=15)	Mandenka (n=24)	Yoruba (n=25)	Bantu NE (n=12)	San (n=7)	Bantu SE (n=8)	África (n=124)
MID-768	rs1160979	0.818	0.933	0.896	0.940	0.917	1.000	0.750	0.887
MID-879	rs1610963	0.985	0.967	0.938	0.920	1.000	0.929	1.000	0.960
MID-1013	rs1611095	0.864	0.900	0.646	0.720	0.750	0.857	0.750	0.778
MID-719	rs1160938	0.909	0.933	0.729	0.620	0.750	0.714	0.750	0.782
MID-1102	rs2067190	0.803	0.933	0.854	0.900	0.958	1.000	0.750	0.871
MID-1272	rs2067354	0.955	0.967	0.854	0.920	1.000	1.000	0.875	0.931
MID-990	rs1611072	0.015	0.000	0.021	0.000	0.042	0.000	0.188	0.024
MID-1683	rs2307879	0.818	0.967	0.896	0.860	0.750	1.000	1.000	0.875
MID-1191	rs2067278	1.000	1.000	0.938	1.000	0.958	1.000	0.938	0.980
MID-883	rs1610967	0.106	0.000	0.000	0.080	0.250	0.071	0.125	0.081
MID-737	rs1160951	1.000	1.000	0.979	1.000	1.000	1.000	1.000	0.996
MID-1937	rs2308129	0.924	0.767	0.792	0.820	0.833	1.000	0.688	0.839
MID-739	rs1160953	0.773	0.900	0.792	0.760	0.750	0.214	0.688	0.750
MID-1348	rs2307544	0.864	0.867	0.667	0.620	0.625	0.929	0.750	0.750
MID-776	rs1610580	0.939	0.967	0.896	0.940	0.833	0.857	0.813	0.911
MID-858	rs673423	0.242	0.267	0.375	0.420	0.250	0.000	0.500	0.310
MID-1271	rs2067353	0.939	0.967	0.875	0.900	1.000	1.000	0.813	0.923
MID-1249	rs2067331	0.409	0.800	0.271	0.240	0.375	0.643	0.375	0.403
MID-717	rs1160936	0.864	0.933	0.896	0.900	0.917	0.786	0.813	0.883
MID-1838	rs2308034	0.652	0.900	0.792	0.640	0.667	0.571	0.500	0.694
MID-1441	rs2307637	0.576	0.900	0.458	0.580	0.667	0.357	0.563	0.589
MID-942	rs1611025	0.924	0.700	0.896	0.920	0.958	1.000	1.000	0.903
MID-1052	rs2067141	0.924	0.900	0.792	0.860	0.917	0.429	0.750	0.843
MID-741	rs32861	0.379	0.400	0.313	0.333	0.545	0.500	0.563	0.393
MID-1004	rs1611086	0.348	0.233	0.458	0.620	0.542	0.786	0.438	0.460
MID-1848	rs2308044	0.348	0.133	0.333	0.300	0.375	0.143	0.563	0.315
MID-1373	rs2307569	0.848	0.933	0.979	1.000	0.833	1.000	0.938	0.927
MID-1434	rs2307630	0.773	0.533	0.521	0.500	0.458	1.000	0.438	0.601
MID-971	rs1611053	0.894	0.767	0.938	0.960	0.958	0.714	0.875	0.895
MID-857	rs1610942	0.697	0.633	0.604	0.440	0.625	0.929	0.750	0.629
MID-1002	rs1611084	0.455	0.433	0.771	0.780	0.667	0.714	0.563	0.621

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior				
		Mozabite (n=30)	Bedouin (n=48)	Druze (n=47)	Palestinian (n=51)	Oriente Médio (n=176)
MID-768	rs1160979	0.500	0.222	0.255	0.333	0.312
MID-879	rs1610963	0.483	0.678	0.713	0.745	0.673
MID-1013	rs1611095	0.433	0.300	0.298	0.245	0.306
MID-719	rs1160938	0.267	0.133	0.117	0.049	0.127
MID-1102	rs2067190	0.367	0.222	0.128	0.186	0.211
MID-1272	rs2067354	0.150	0.256	0.223	0.235	0.223
MID-990	rs1611072	0.233	0.318	0.298	0.340	0.304
MID-1683	rs2307879	0.383	0.239	0.255	0.265	0.276
MID-1191	rs2067278	0.467	0.598	0.447	0.637	0.546
MID-883	rs1610967	0.517	0.761	0.851	0.775	0.747
MID-737	rs1160951	0.917	0.864	0.734	0.716	0.794
MID-1937	rs2308129	0.417	0.304	0.267	0.373	0.334
MID-739	rs1160953	0.467	0.435	0.362	0.471	0.431
MID-1348	rs2307544	0.583	0.359	0.436	0.392	0.428
MID-776	rs1610580	0.793	0.792	0.713	0.657	0.731
MID-858	rs673423	0.414	0.417	0.532	0.441	0.454
MID-1271	rs2067353	0.138	0.313	0.277	0.304	0.271
MID-1249	rs2067331	0.190	0.260	0.255	0.255	0.246
MID-717	rs1160936	0.466	0.563	0.489	0.667	0.557
MID-1838	rs2308034	0.483	0.298	0.223	0.275	0.302
MID-1441	rs2307637	0.397	0.351	0.511	0.471	0.437
MID-942	rs1611025	0.650	0.449	0.424	0.598	0.524
MID-1052	rs2067141	0.367	0.295	0.359	0.373	0.349
MID-741	rs32861	0.172	0.329	0.380	0.441	0.351
MID-1004	rs1611086	0.567	0.605	0.696	0.706	0.655
MID-1848	rs2308044	0.767	0.882	0.844	0.892	0.854
MID-1373	rs2307569	0.633	0.434	0.630	0.686	0.603
MID-1434	rs2307630	0.417	0.263	0.196	0.140	0.235
MID-971	rs1611053	0.817	0.635	0.543	0.549	0.616
MID-857	rs1610942	0.417	0.323	0.330	0.304	0.335
MID-1002	rs1611084	0.583	0.385	0.479	0.569	0.497

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior								
		Brahui (n=25)	Baloshi (n=25)	Hazara (n=25)	Makran i (n=25)	Sindhi (n=25)	Pathan (n=24)	Kalash (n=25)	Burusho (n=25)	Ásia Central (n=199)
MID-768	rs1160979	0.260	0.292	0.542	0.220	0.120	0.167	0.250	0.240	0.260
MID-879	rs1610963	0.646	0.630	0.479	0.620	0.660	0.521	0.521	0.800	0.611
MID-1013	rs1611095	0.320	0.271	0.458	0.160	0.260	0.292	0.646	0.340	0.342
MID-719	rs1160938	0.100	0.188	0.208	0.140	0.160	0.083	0.167	0.060	0.138
MID-1102	rs2067190	0.140	0.167	0.354	0.080	0.080	0.083	0.104	0.180	0.148
MID-1272	rs2067354	0.160	0.125	0.271	0.360	0.240	0.292	0.396	0.280	0.265
MID-990	rs1611072	0.280	0.333	0.271	0.260	0.400	0.271	0.583	0.380	0.347
MID-1683	rs2307879	0.480	0.320	0.340	0.560	0.480	0.438	0.060	0.300	0.372
MID-1191	rs2067278	0.560	0.540	0.600	0.640	0.520	0.604	0.760	0.540	0.595
MID-883	rs1610967	0.760	0.820	0.700	0.720	0.860	0.729	0.760	0.680	0.754
MID-737	rs1160951	0.580	0.620	0.625	0.600	0.600	0.542	0.580	0.420	0.571
MID-1937	rs2308129	0.340	0.300	0.420	0.340	0.380	0.375	0.400	0.280	0.354
MID-739	rs1160953	0.400	0.500	0.540	0.340	0.320	0.333	0.340	0.600	0.422
MID-1348	rs2307544	0.460	0.320	0.180	0.300	0.360	0.271	0.440	0.420	0.344
MID-776	rs1610580	0.700	0.680	0.840	0.740	0.620	0.646	0.540	0.700	0.683
MID-858	rs673423	0.640	0.540	0.620	0.520	0.660	0.479	0.520	0.580	0.570
MID-1271	rs2067353	0.220	0.200	0.220	0.540	0.320	0.354	0.340	0.220	0.302
MID-1249	rs2067331	0.300	0.180	0.380	0.400	0.520	0.458	0.140	0.120	0.312
MID-717	rs1160936	0.520	0.640	0.700	0.720	0.480	0.583	0.740	0.560	0.618
MID-1838	rs2308034	0.240	0.200	0.220	0.300	0.380	0.229	0.300	0.160	0.254
MID-1441	rs2307637	0.460	0.560	0.667	0.600	0.540	0.604	0.420	0.380	0.528
MID-942	rs1611025	0.440	0.520	0.480	0.600	0.620	0.563	0.500	0.708	0.553
MID-1052	rs2067141	0.440	0.240	0.520	0.360	0.400	0.542	0.380	0.604	0.434
MID-741	rs32861	0.260	0.340	0.460	0.560	0.260	0.313	0.160	0.104	0.308
MID-1004	rs1611086	0.700	0.800	0.860	0.720	0.760	0.792	0.880	0.875	0.798
MID-1848	rs2308044	0.920	0.900	0.938	0.900	0.920	0.917	0.780	0.813	0.886
MID-1373	rs2307569	0.760	0.700	0.800	0.720	0.800	0.771	0.780	0.761	0.761
MID-1434	rs2307630	0.160	0.060	0.200	0.060	0.104	0.125	0.240	0.167	0.140
MID-971	rs1611053	0.460	0.580	0.420	0.340	0.580	0.542	0.620	0.583	0.515
MID-857	rs1610942	0.180	0.220	0.220	0.180	0.060	0.146	0.060	0.140	0.151
MID-1002	rs1611084	0.540	0.440	0.420	0.480	0.520	0.521	0.460	0.560	0.492

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior																		Leste Asiático (n=247)
		Han (n=44)	Tujia (n=10)	Yizu (n=10)	Miaozu (n=10)	Oroqen (n=10)	Daur (n=10)	Mongola (n=10)	Hezhen (n=9)	Xibo (n=9)	Uygur (n=10)	Dai (n=10)	Lahu (n=10)	She (n=10)	Naxi (n=10)	Tu (n=10)	Yakut (n=25)	Japanese (n=29)	Cambodian (n=11)	
MID-768	rs1160979	0.511	0.350	0.700	0.400	0.600	0.500	0.450	0.389	0.500	0.250	0.750	0.350	0.550	0.800	0.500	0.480	0.621	0.818	0.532
MID-879	rs1610963	0.682	0.200	0.600	0.500	0.600	0.450	0.450	0.611	0.500	0.600	0.800	0.350	0.750	0.800	0.700	0.560	0.500	0.500	0.575
MID-1013	rs1611095	0.534	0.600	0.650	0.750	0.650	0.700	0.550	0.722	0.556	0.450	0.700	0.850	0.750	0.700	0.450	0.440	0.603	0.455	0.593
MID-719	rs1160938	0.114	0.250	0.100	0.100	0.050	0.050	0.100	0.111	0.111	0.050	0.100	0.150	0.100	0.100	0.050	0.220	0.138	0.273	0.128
MID-1102	rs2067190	0.795	0.950	0.600	0.800	0.800	0.500	0.800	0.778	0.667	0.350	0.850	0.900	0.800	0.550	0.850	0.660	0.707	0.727	0.731
MID-1272	rs2067354	0.341	0.300	0.350	0.450	0.450	0.550	0.450	0.389	0.278	0.150	0.450	0.450	0.450	0.400	0.250	0.160	0.362	0.455	0.354
MID-990	rs1611072	0.102	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.056	0.150	0.000	0.000	0.050	0.000	0.000	0.080	0.000	0.136	0.047
MID-1683	rs2307879	0.455	0.650	0.350	0.550	0.500	0.650	0.650	0.500	0.556	0.350	0.600	0.550	0.550	0.400	0.400	0.420	0.483	0.591	0.496
MID-1191	rs2067278	0.693	0.700	0.800	0.800	0.750	0.850	0.700	0.611	0.778	0.500	0.750	0.650	0.850	0.850	0.800	0.640	0.862	0.682	0.735
MID-883	rs1610967	0.761	0.700	0.750	0.650	0.650	0.650	0.800	0.722	0.611	0.600	0.650	0.750	0.550	0.700	0.750	0.700	0.621	0.818	0.696
MID-737	rs1160951	0.341	0.500	0.350	0.300	0.350	0.150	0.300	0.111	0.556	0.350	0.150	0.500	0.300	0.350	0.250	0.400	0.414	0.591	0.356
MID-1937	rs2308129	0.625	0.600	0.850	0.550	0.350	0.750	0.650	0.667	0.611	0.450	0.550	0.389	0.750	0.500	0.600	0.400	0.603	0.636	0.581
MID-739	rs1160953	0.659	0.600	0.600	0.850	0.600	0.600	0.500	0.500	0.556	0.400	0.550	0.700	0.750	0.600	0.500	0.700	0.534	0.864	0.621
MID-1348	rs2307544	0.295	0.400	0.200	0.200	0.100	0.150	0.200	0.111	0.167	0.200	0.250	0.150	0.100	0.100	0.250	0.060	0.190	0.227	0.194
MID-776	rs1610580	0.977	1.000	0.850	0.850	0.900	0.900	0.850	0.944	0.889	0.900	0.850	1.000	1.000	0.900	0.900	0.940	0.983	1.000	0.937
MID-858	rs673423	0.523	0.500	0.700	0.650	0.450	0.500	0.650	0.500	0.611	0.600	0.400	0.400	0.400	0.350	0.550	0.520	0.483	0.364	0.508
MID-1271	rs2067353	0.386	0.300	0.350	0.500	0.450	0.550	0.450	0.389	0.278	0.150	0.500	0.500	0.450	0.400	0.250	0.180	0.379	0.591	0.379
MID-1249	rs2067331	0.341	0.200	0.100	0.200	0.050	0.000	0.350	0.111	0.056	0.300	0.350	0.600	0.400	0.300	0.100	0.160	0.310	0.500	0.261
MID-717	rs1160936	0.693	0.650	0.800	0.750	0.800	0.700	0.800	0.722	0.722	0.450	0.800	0.650	0.850	0.750	0.800	0.660	0.793	0.818	0.729
MID-1838	rs2308034	0.273	0.250	0.050	0.250	0.200	0.400	0.100	0.278	0.444	0.350	0.300	0.050	0.450	0.300	0.500	0.180	0.276	0.045	0.257
MID-1441	rs2307637	0.557	0.350	0.700	0.600	0.550	0.600	0.600	0.611	0.556	0.350	0.600	0.550	0.750	0.400	0.600	0.580	0.672	0.682	0.579
MID-942	rs1611025	0.628	0.600	0.650	0.500	0.550	0.550	0.550	0.667	0.625	0.450	0.550	0.500	0.643	0.500	0.650	0.643	0.500	0.409	0.570
MID-1052	rs2067141	0.667	0.850	0.833	0.800	0.778	0.950	0.950	0.722	0.750	0.400	0.800	0.500	0.643	0.600	0.550	0.738	0.643	0.773	0.707
MID-741	rs32861	0.440	0.500	0.300	0.350	0.250	0.350	0.550	0.389	0.313	0.300	0.700	0.800	0.786	0.833	0.500	0.575	0.648	0.136	0.489
MID-1004	rs1611086	0.869	0.900	0.950	0.950	0.750	1.000	1.000	0.889	0.875	0.750	0.800	0.600	0.571	0.550	0.750	0.690	0.821	0.818	0.814
MID-1848	rs2308044	0.863	0.950	0.900	0.800	1.000	0.800	0.900	0.722	0.938	0.850	0.900	0.500	0.857	0.444	0.889	0.900	0.804	1.000	0.840
MID-1373	rs2307569	0.905	0.950	0.850	0.900	0.950	0.750	0.950	1.000	0.750	0.700	0.950	0.900	0.929	0.950	0.750	0.929	0.911	1.000	0.896
MID-1434	rs2307630	0.427	0.600	0.450	0.167	0.200	0.278	0.400	0.333	0.063	0.125	0.600	0.600	0.357	0.222	0.450	0.619	0.333	0.318	0.389
MID-971	rs1611053	0.202	0.100	0.050	0.100	0.300	0.350	0.150	0.389	0.125	0.450	0.100	0.050	0.000	0.111	0.150	0.167	0.179	0.136	0.179
MID-857	rs1610942	0.477	0.200	0.250	0.650	0.400	0.150	0.250	0.389	0.278	0.200	0.400	0.400	0.500	0.350	0.450	0.280	0.328	0.273	0.358
MID-1002	rs1611084	0.148	0.300	0.100	0.150	0.300	0.300	0.200	0.167	0.111	0.600	0.150	0.450	0.050	0.200	0.150	0.340	0.155	0.273	0.221

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior		
		Papuan (n=17)	Melanesian (n=19)	Oceania (n=36)
MID-768	rs1160979	0.353	0.105	0.222
MID-879	rs1610963	0.294	0.553	0.431
MID-1013	rs1611095	0.000	0.184	0.097
MID-719	rs1160938	0.029	0.079	0.056
MID-1102	rs2067190	0.118	0.658	0.403
MID-1272	rs2067354	0.735	0.632	0.681
MID-990	rs1611072	0.000	0.000	0.000
MID-1683	rs2307879	0.618	0.158	0.375
MID-1191	rs2067278	0.971	0.974	0.972
MID-883	rs1610967	1.000	0.895	0.944
MID-737	rs1160951	0.294	0.500	0.403
MID-1937	rs2308129	0.265	0.237	0.250
MID-739	rs1160953	0.618	0.789	0.708
MID-1348	rs2307544	0.676	0.526	0.597
MID-776	rs1610580	0.618	0.947	0.792
MID-858	rs673423	0.500	0.342	0.417
MID-1271	rs2067353	0.735	0.632	0.681
MID-1249	rs2067331	0.294	0.105	0.194
MID-717	rs1160936	1.000	0.974	0.986
MID-1838	rs2308034	0.000	0.447	0.236
MID-1441	rs2307637	0.500	0.132	0.306
MID-942	rs1611025	0.765	0.526	0.639
MID-1052	rs2067141	0.765	0.237	0.486
MID-741	rs32861	0.265	0.105	0.181
MID-1004	rs1611086	0.971	0.974	0.972
MID-1848	rs2308044	1.000	0.947	0.972
MID-1373	rs2307569	1.000	1.000	1.000
MID-1434	rs2307630	0.536	0.579	0.561
MID-971	rs1611053	0.618	0.579	0.597
MID-857	rs1610942	0.176	0.342	0.264
MID-1002	rs1611084	0.735	0.553	0.639

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior								
		French (n=24)	Basque (n=24)	Sardinian (n=28)	Bergamo (n=13)	Tuscan (n=8)	Orcadian (n=16)	Adygei (n=17)	Russian (n=25)	Europa (n=155)
MID-768	rs1160979	0.229	0.146	0.089	0.115	0.188	0.219	0.265	0.280	0.190
MID-879	rs1610963	0.396	0.271	0.446	0.538	0.375	0.406	0.471	0.400	0.406
MID-1013	rs1611095	0.208	0.250	0.089	0.154	0.188	0.313	0.294	0.340	0.229
MID-719	rs1160938	0.125	0.167	0.268	0.154	0.063	0.031	0.088	0.180	0.152
MID-1102	rs2067190	0.146	0.208	0.071	0.077	0.188	0.125	0.147	0.260	0.155
MID-1272	rs2067354	0.063	0.104	0.125	0.077	0.125	0.063	0.118	0.200	0.113
MID-990	rs1611072	0.229	0.417	0.357	0.308	0.313	0.531	0.265	0.360	0.348
MID-1683	rs2307879	0.229	0.188	0.125	0.077	0.125	0.156	0.147	0.300	0.181
MID-1191	rs2067278	0.354	0.563	0.411	0.308	0.438	0.406	0.676	0.320	0.432
MID-883	rs1610967	0.604	0.542	0.804	0.577	0.688	0.656	0.706	0.620	0.652
MID-737	rs1160951	0.500	0.479	0.696	0.692	0.625	0.313	0.471	0.580	0.545
MID-1937	rs2308129	0.417	0.333	0.286	0.346	0.313	0.250	0.206	0.400	0.326
MID-739	rs1160953	0.375	0.396	0.375	0.308	0.500	0.281	0.676	0.300	0.390
MID-1348	rs2307544	0.354	0.229	0.339	0.192	0.188	0.313	0.206	0.320	0.284
MID-776	rs1610580	0.543	0.813	0.500	0.577	0.500	0.719	0.588	0.520	0.597
MID-858	rs673423	0.455	0.667	0.643	0.577	0.625	0.469	0.706	0.646	0.602
MID-1271	rs2067353	0.087	0.167	0.125	0.077	0.125	0.063	0.118	0.160	0.120
MID-1249	rs2067331	0.217	0.208	0.054	0.038	0.125	0.219	0.118	0.140	0.143
MID-717	rs1160936	0.348	0.563	0.518	0.385	0.375	0.406	0.676	0.400	0.468
MID-1838	rs2308034	0.283	0.354	0.214	0.192	0.375	0.219	0.235	0.260	0.263
MID-1441	rs2307637	0.239	0.146	0.393	0.154	0.313	0.344	0.324	0.300	0.279
MID-942	rs1611025	0.500	0.604	0.407	0.538	0.563	0.344	0.441	0.400	0.468
MID-1052	rs2067141	0.396	0.292	0.315	0.500	0.313	0.344	0.235	0.438	0.353
MID-741	rs32861	0.326	0.250	0.259	0.231	0.313	0.438	0.206	0.360	0.297
MID-1004	rs1611086	0.667	0.708	0.611	0.538	0.813	0.813	0.794	0.660	0.688
MID-1848	rs2308044	0.708	0.854	0.889	1.000	0.688	0.906	0.794	0.800	0.831
MID-1373	rs2307569	0.729	0.563	0.685	0.654	0.500	0.844	0.735	0.840	0.708
MID-1434	rs2307630	0.125	0.063	0.037	0.115	0.063	0.100	0.088	0.060	0.078
MID-971	rs1611053	0.458	0.417	0.481	0.423	0.563	0.375	0.471	0.500	0.458
MID-857	rs1610942	0.208	0.229	0.143	0.269	0.188	0.188	0.382	0.220	0.223
MID-1002	rs1611084	0.479	0.479	0.500	0.500	0.375	0.594	0.500	0.560	0.506

Tabela 6 – Continuação

MID	dbSNP ID	Frequência do alelo maior					America (n=108)
		Pima (n=25)	Maya (n=25)	Colombian (n=13)	Karitiana (n=24)	Surui (n=21)	
MID-768	rs1160979	0.660	0.560	0.615	0.729	0.476	0.611
MID-879	rs1610963	0.300	0.280	0.115	0.229	0.357	0.269
MID-1013	rs1611095	0.620	0.720	0.846	0.688	0.762	0.713
MID-719	rs1160938	0.400	0.300	0.500	0.896	0.595	0.537
MID-1102	rs2067190	0.500	0.520	0.692	0.375	0.643	0.528
MID-1272	rs2067354	0.340	0.400	0.846	0.208	0.000	0.319
MID-990	rs1611072	0.000	0.080	0.077	0.000	0.000	0.028
MID-1683	rs2307879	0.560	0.380	0.577	0.375	0.714	0.509
MID-1191	rs2067278	0.720	0.540	0.654	0.917	0.833	0.736
MID-883	rs1610967	0.640	0.560	0.538	0.146	0.643	0.500
MID-737	rs1160951	0.360	0.260	0.154	0.313	0.167	0.264
MID-1937	rs2308129	0.800	0.800	0.769	0.542	0.929	0.764
MID-739	rs1160953	0.640	0.480	0.769	0.729	0.476	0.606
MID-1348	rs2307544	0.080	0.020	0.000	0.063	0.190	0.074
MID-776	rs1610580	1.000	0.960	1.000	1.000	1.000	0.991
MID-858	rs673423	0.260	0.440	0.115	0.625	0.262	0.366
MID-1271	rs2067353	0.360	0.380	0.846	0.250	0.000	0.329
MID-1249	rs2067331	0.380	0.540	0.500	0.750	0.881	0.611
MID-717	rs1160936	0.800	0.620	0.654	0.938	0.738	0.759
MID-1838	rs2308034	0.140	0.160	0.192	0.021	0.000	0.097
MID-1441	rs2307637	0.580	0.320	0.346	0.313	0.238	0.366
MID-942	rs1611025	0.640	0.780	0.692	0.625	0.452	0.639
MID-1052	rs2067141	1.000	0.740	0.885	0.708	1.000	0.861
MID-741	rs32861	0.240	0.167	0.000	0.250	0.405	0.229
MID-1004	rs1611086	0.900	0.920	1.000	0.979	1.000	0.954
MID-1848	rs2308044	0.920	0.880	0.846	0.958	1.000	0.926
MID-1373	rs2307569	0.640	0.720	0.846	0.688	1.000	0.764
MID-1434	rs2307630	0.040	0.060	0.385	0.292	0.405	0.213
MID-971	rs1611053	0.220	0.220	0.385	0.354	0.071	0.241
MID-857	rs1610942	0.700	0.380	0.538	0.708	0.690	0.606
MID-1002	rs1611084	0.000	0.140	0.115	0.000	0.167	0.079

2.4.2.1 - Análise de Variância Molecular (AMOVA)

Após a genotipagem dos 29 INDELS em todo o painel de amostras mundiais do CEPH-HGDP (1045 indivíduos de 52 populações diferentes, distribuídas em 7 grandes regiões geográficas: África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e Américas), submetemos os genótipos de 31 INDELS (os 29 tipados no presente estudo mais os dois tipados anteriormente pelo nosso grupo) a uma análise de variância molecular (AMOVA) utilizando o programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000). Os resultados da análise estão representados na tabela 7. Se focarmos em cada região separadamente, a porcentagem de variação dentro das populações corresponde a mais de 94%, com exceção da Oceania e Américas que apresentaram um componente de variação menor entre indivíduos dentro das populações. Quando analisamos as 7 regiões geográficas, uma porcentagem de 16,38% da variabilidade genética é encontrada entre as grandes regiões.

Tabela 7 – Análise de variância molecular (AMOVA) de dados dos 31 INDELS localizados no braço longo do cromossomo 5 referentes às sete regiões geográficas do painel do CEPH-HGDP estudadas.

Regiões	Número de Regiões	Número de Populações	Distribuição da variação (%)		
			Dentro de Populações	Entre Populações	Entre Regiões
Mundo	1	52	82,71	17,29	
Mundo	7	52	80,72	2,90	16,38
África	1	7	94,63	5,37	
Eurásia	1	20	96,70	3,30	
Eurásia	3	20	96,16	1,94	1,90
Europa	1	8	98,84	1,16	
Oriente Médio	1	4	97,84	2,16	
Ásia Central	1	8	97,51	2,49	
Leste Asiático	1	18	97,00	3,00	
Oceania	1	2	86,47	13,53	
Américas	1	5	90,27	9,73	

2.4.2.2 – Escalonamento Multidimensional (MDS) em populações mundiais

Verificamos o poder discriminatório do nosso painel de 31 INDELS através da construção de matrizes de distância de Reynolds com as 7 regiões geográficas e com as 52 populações mundiais do painel do CEPH-HGDP. A partir das matrizes de distância, realizamos um MDS para uma representação gráfica das mesmas (Figuras 5 e 6). Com somente duas dimensões obtivemos representações gráficas muito adequadas das matrizes de distância: *stress* de 0,0037 para as regiões e *stress* de 0.0935 para as populações. No primeiro gráfico (Figura 5), onde a análise foi realizada com as grandes regiões geográficas, é possível observar a África bastante isolada das demais regiões. A Ásia Central e Oriente Médio aparecem em um mesmo ponto, próximos da Europa que juntos, constituem a Eurásia. A América, embora distante das demais regiões, apresenta menor distância em relação ao Leste asiático. Já a Oceania que também aparece distante dos demais continentes, apresenta a menor distância em relação a Ásia Central e Oriente Médio.

No segundo gráfico (Figura 6), onde a análise foi realizada com as 52 populações, podemos observar cinco agrupamentos claros que correspondem às cinco grandes regiões geográficas: África, Eurásia (Europa, Oriente Médio e Ásia Central), Leste asiático, América e Oceania. Dos cinco agrupamentos, o correspondente a África é o que se encontra mais distante. Em relação ao agrupamento que contém as populações africanas, a população San é a que aparece mais distante das demais.

As populações Uygur do Leste asiático e Hazara da Ásia Central encontram-se em posições intermediárias entre o agrupamento da Eurásia e Leste asiático. Já as populações nativo americanas aparecem mais espalhadas do que as dos demais agrupamentos, sendo a Maya a mais próxima do Leste Asiático.

2.4.2.3 – Diferenciação genética entre indivíduos das populações mundiais

Analisamos o padrão de similaridade entre os diferentes indivíduos através da construção de uma matriz de distância baseada em compartilhamento de alelos idênticos por estado (*IBS distance matrix*) que, posteriormente, foi graficamente representada através da construção de árvores filogenéticas *neighbor-joining* (Figura 7). Nesta análise construímos uma árvore onde os indivíduos foram coloridos de acordo com uma das cinco grandes regiões geográficas de onde são oriundos (Figura 7a) e outra onde os indivíduos foram coloridos de acordo com a população a qual pertencem (Figura 7b). Nossos resultados revelaram claro agrupamento entre os indivíduos pertencentes a uma mesma região geográfica (Figura 7a), o que não foi observado para indivíduos pertencentes a uma mesma população (Figura 7b).

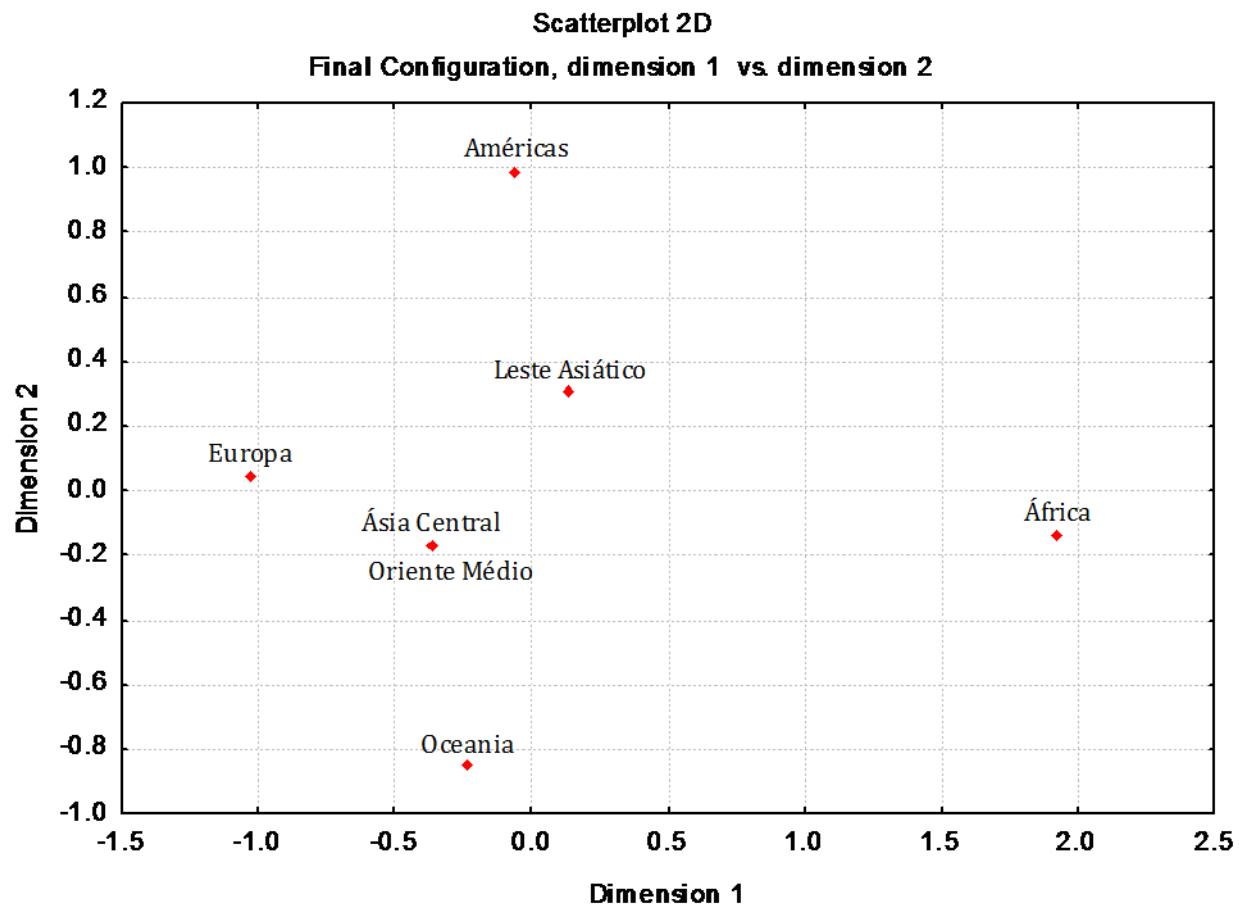


Figura 5 – Escalonamento Multidimensional realizado com as 7 regiões geográficas.

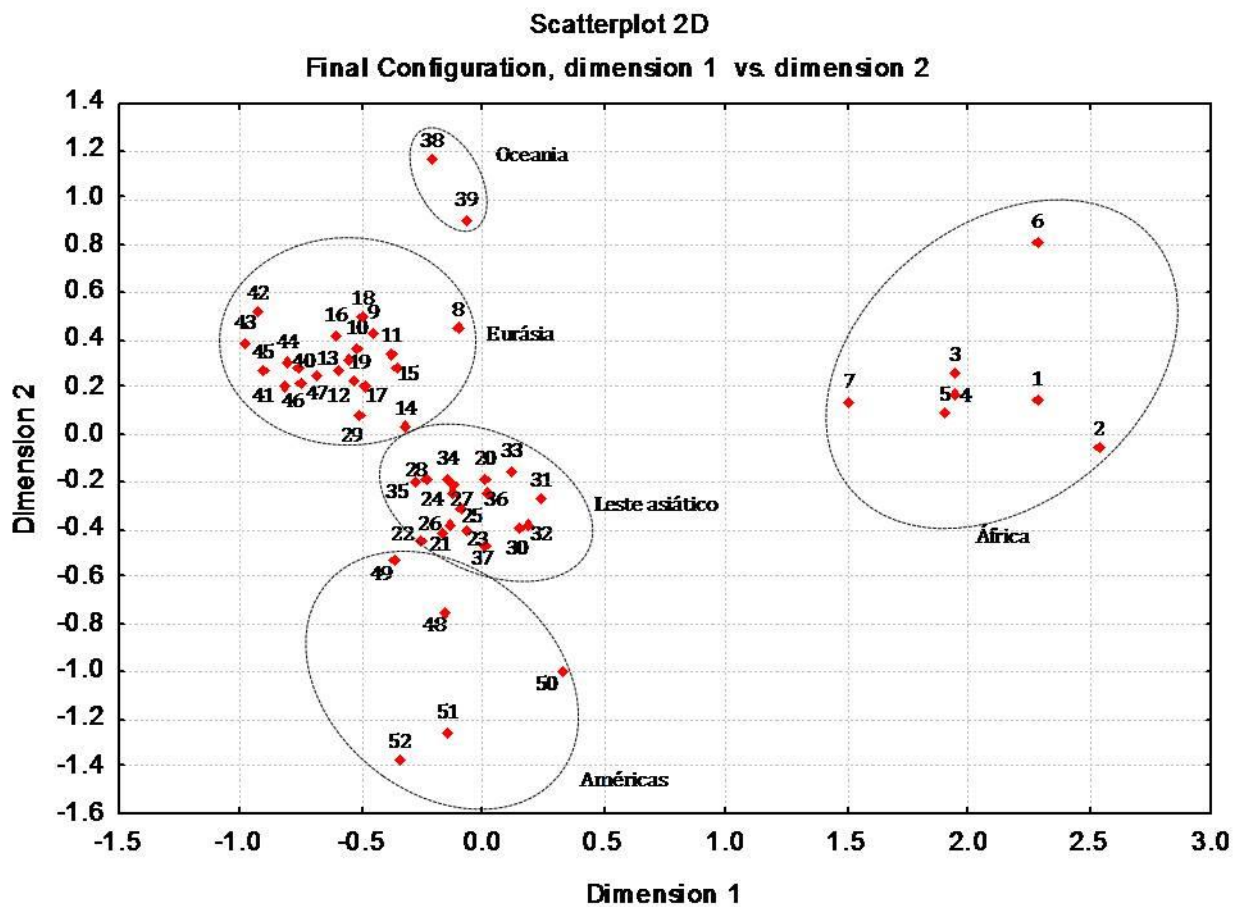


Figura 6 – Escalonamento Multidimensional realizado com as 52 populações mundiais. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

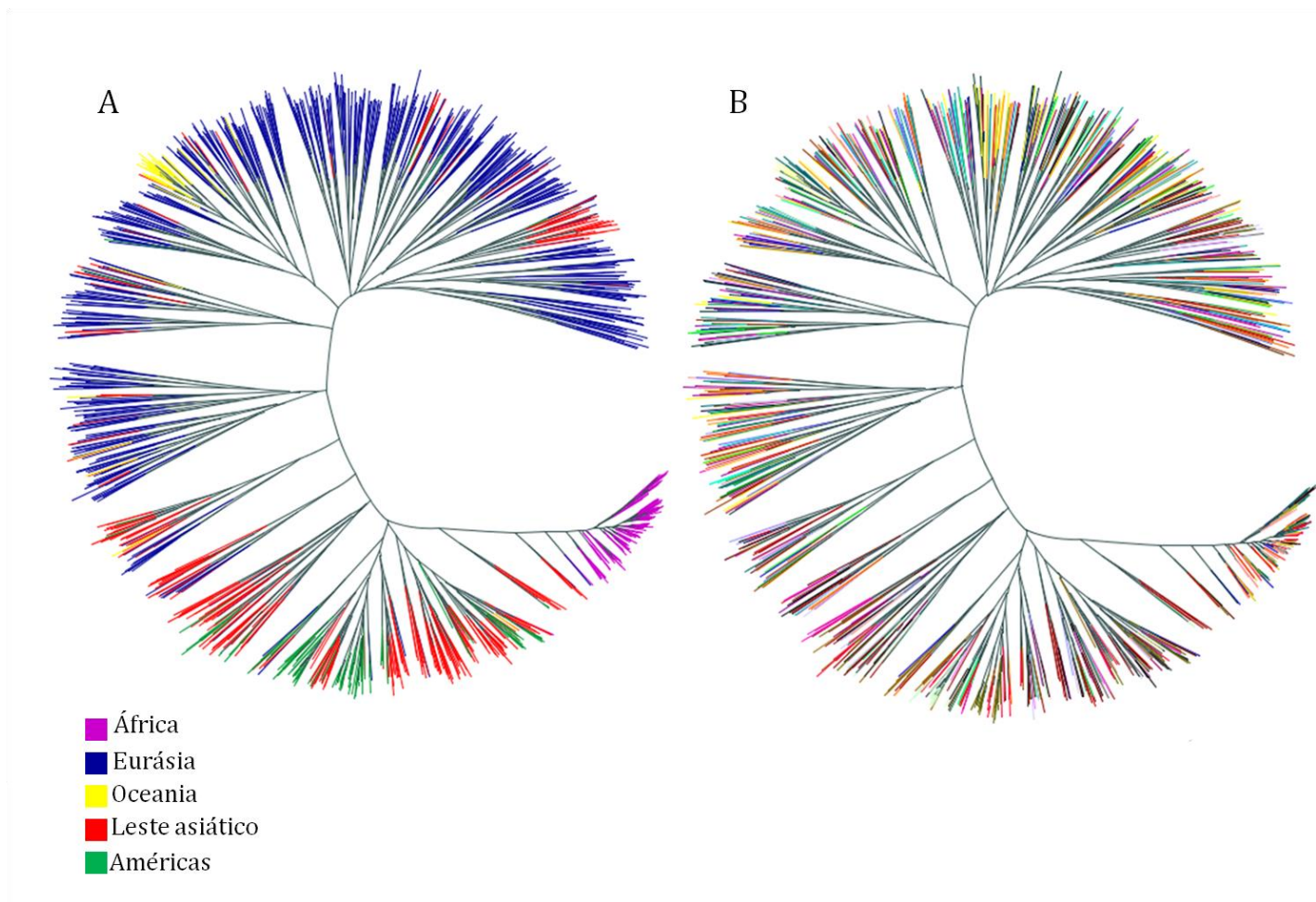


Figura 7 – Árvores filogenéticas *neighbor-joining* relativas ao compartilhamento de alelos idênticos por estado entre os 1045 indivíduos do painel do CEPH-HGDP estudados. A) Indivíduos coloridos de acordo com a região geográfica a qual pertencem; B) Indivíduos coloridos de acordo com a população a qual pertencem.

2.4.2.4 – Análise de desequilíbrio de ligação nas regiões mundiais

Através do programa *HaploView 4.1*, verificamos se os marcadores ao longo do cromossomo 5 se encontravam em desequilíbrio de ligação em cada uma das regiões geográficas estudadas. Foi observado significativo desequilíbrio de ligação ($r^2 > 0,80$) entre os marcadores MID-1271 e MID-1272 apenas no Leste asiático ($r^2 = 0,86$), Oceania ($r^2 = 1,0$), Europa ($r^2 = 0,81$) e Américas ($r^2 = 0,87$) (Figuras 8 a 14). Como nenhum marcador se mostrou em desequilíbrio de ligação com outro em todas as regiões estudadas, trabalhamos com frequências alélicas e não haplotípicas.

2.4.2.5 – Análise do Componente Principal em populações mundiais

Após a análise de desequilíbrio de ligação entre os 31 INDELS em cada uma das 7 regiões geográficas, verificamos a estruturação genética entre as populações mundiais através de duas análises do componente principal. Na primeira, utilizamos dados de frequências alélicas encontradas nas grandes regiões. Na segunda análise, utilizamos dados das frequências alélicas de cada uma das 52 populações mundiais. Nas duas análises do componente principal, os valores do primeiro e segundo componentes foram utilizados para a construção de dois gráficos de dispersão bidimensional (Figuras 15 e 16). Na análise realizada com as regiões, o total de variação genética recuperada pelos dois primeiros componentes foi de 74,29%. No gráfico (figura 15) podemos observar um agrupamento claro

correspondente a Eurásia (Ásia Central, Oriente Médio e Europa). Os demais continentes aparecem fora de agrupamentos, sendo a África, novamente, o continente mais isolado. As Américas encontram-se também distantes dos demais continentes, porém, mais próximas do Leste asiático. Este resultado indica uma próxima relação filogenética entre Leste asiático e Américas e está de acordo com a história de migração do homem moderno, já que estima-se que a entrada do homem nas Américas tenha ocorrido através de uma única onda migratória partindo do Leste asiático (Cavalli-Sforza *et al.*, 1994).

Na análise dos componentes principais realizada com as 52 populações mundiais, o total de variação genética recuperada pelos dois primeiros componentes foi de 63.33%. No gráfico (Figura 16) podemos observar quatro agrupamentos: um contendo as populações da África, um incluindo as populações da Eurásia (Ásia Central, Oriente Médio e Europa) e a população Uygur do Leste asiático, outro constituído pelas duas populações da Oceania e um último agrupamento formado pelas populações do Leste asiático e Américas. A população da Ásia Central, Hazara, se encontra fora dos agrupamentos, entre o grupo do Leste asiático e Eurásia.

África

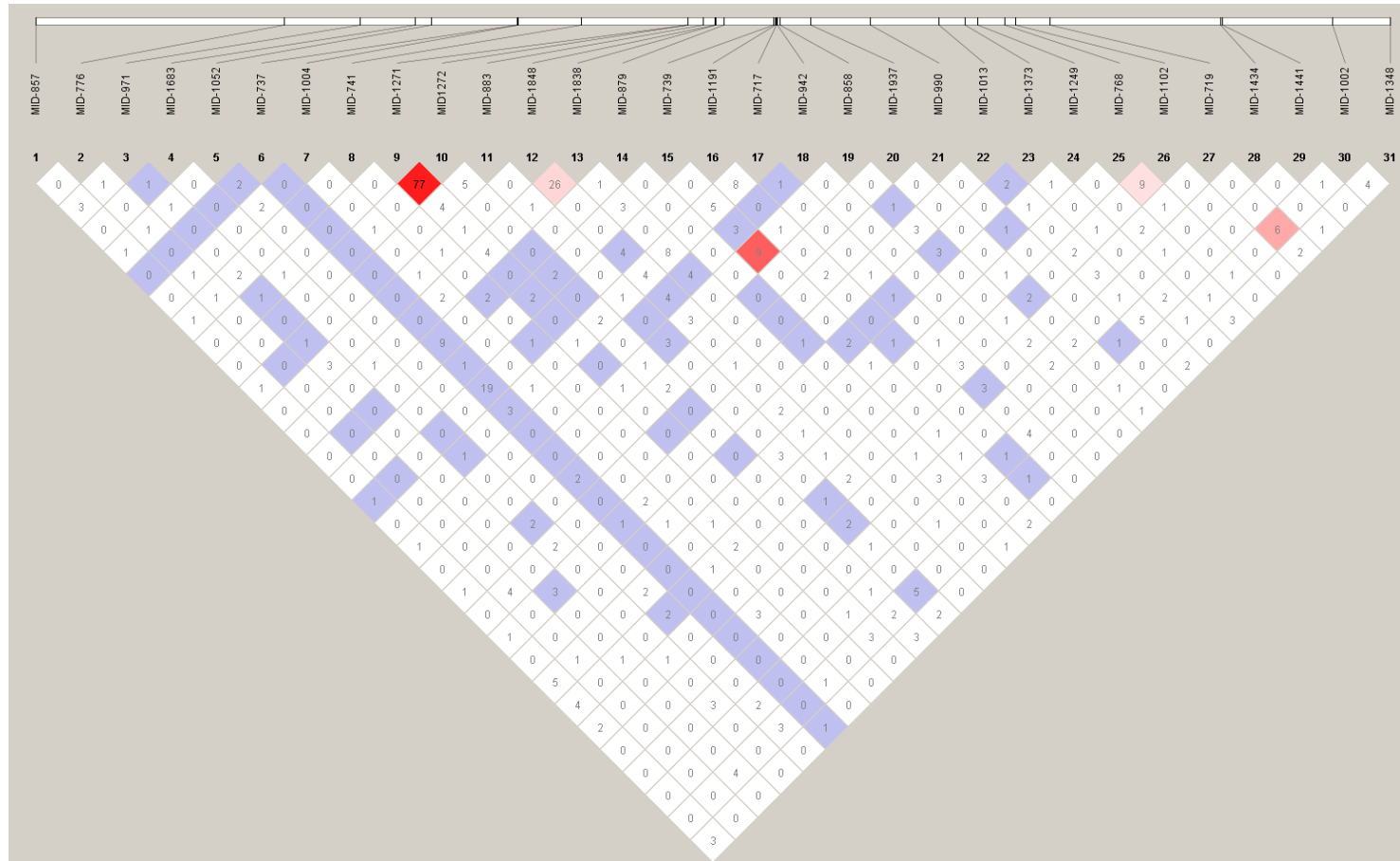


Figura 8- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para África. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $LOD < 2$, $D' < 1$; azul: $LOD < 2$, $D' = 1$; graus de rosa/vermelho: $LOD = 2$, $D' < 1$; vermelho escuro: $LOD = 2$, $D' = 1$.

Oriente Médio

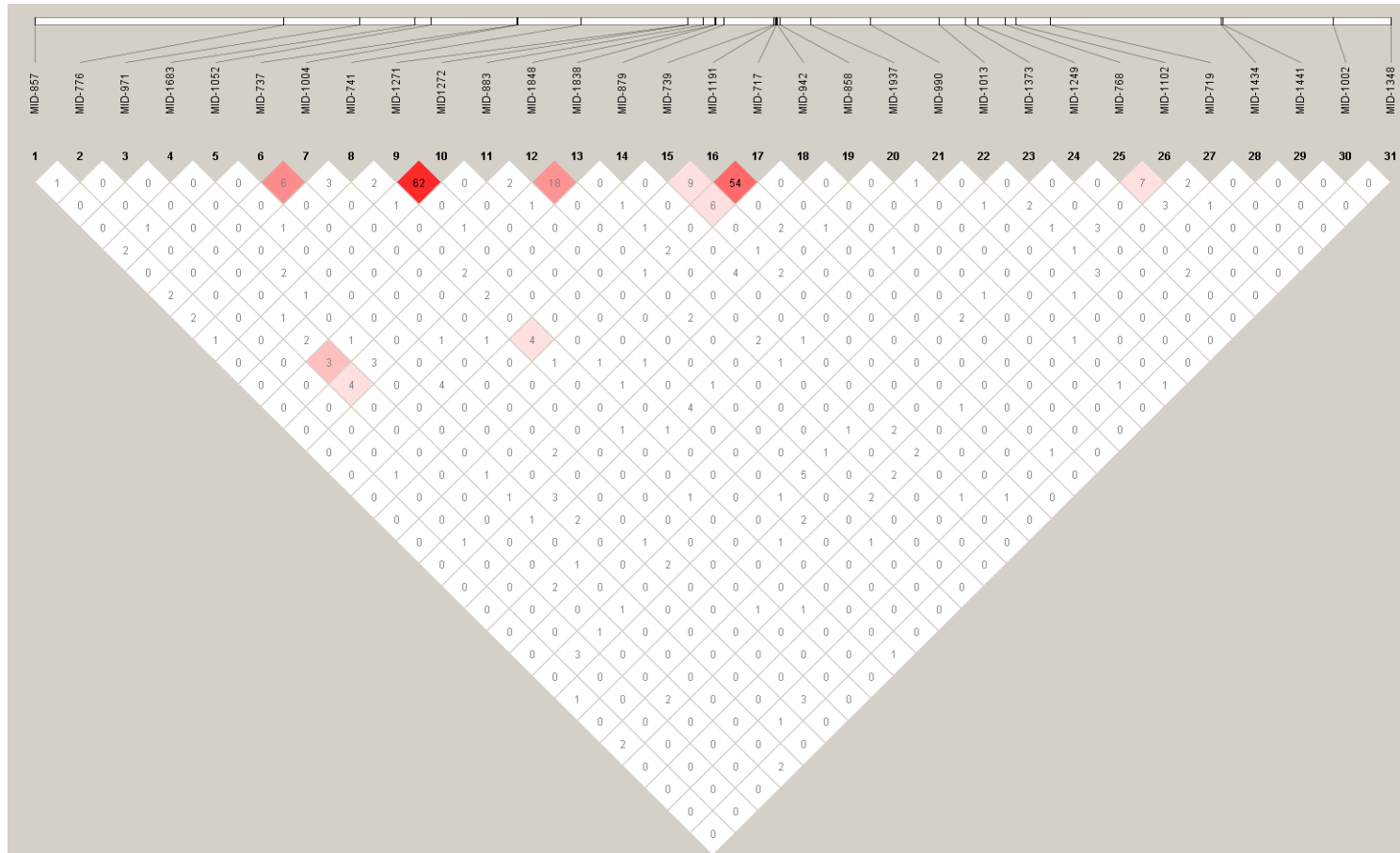


Figura 9- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para Oriente Médio. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $LOD < 2$, $D' < 1$; azul: $LOD < 2$, $D' = 1$; graus de rosa/vermelho: $LOD = 2$, $D' < 1$; vermelho escuro: $LOD = 2$, $D' = 1$.

Ásia Central

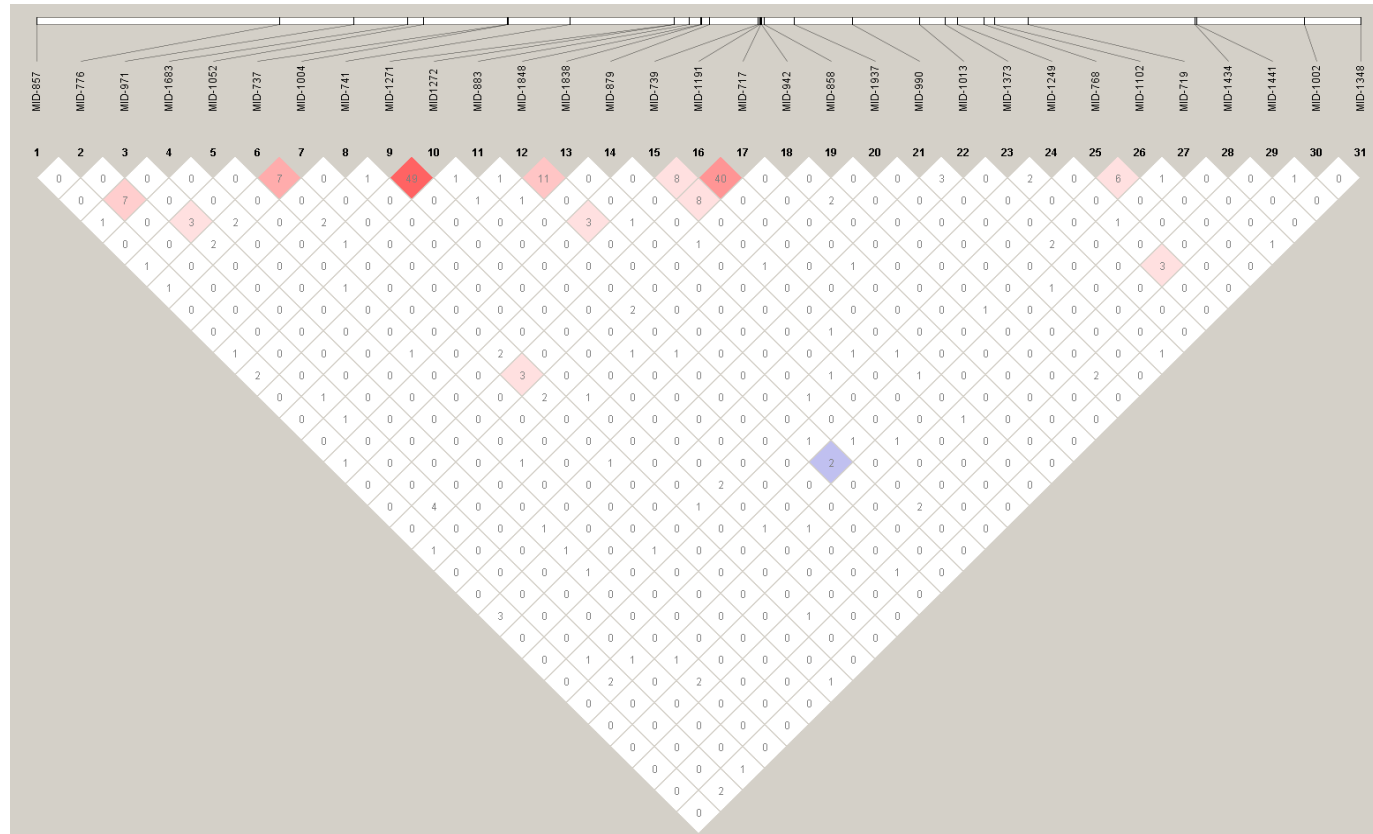


Figura 10- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Ásia Central. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $\text{LOD} < 2$, $D' < 1$; azul: $\text{LOD} < 2$, $D' = 1$; graus de rosa/vermelho: $\text{LOD} = 2$, $D' < 1$; vermelho escuro: $\text{LOD} = 2$, $D' = 1$.

Leste Asiático

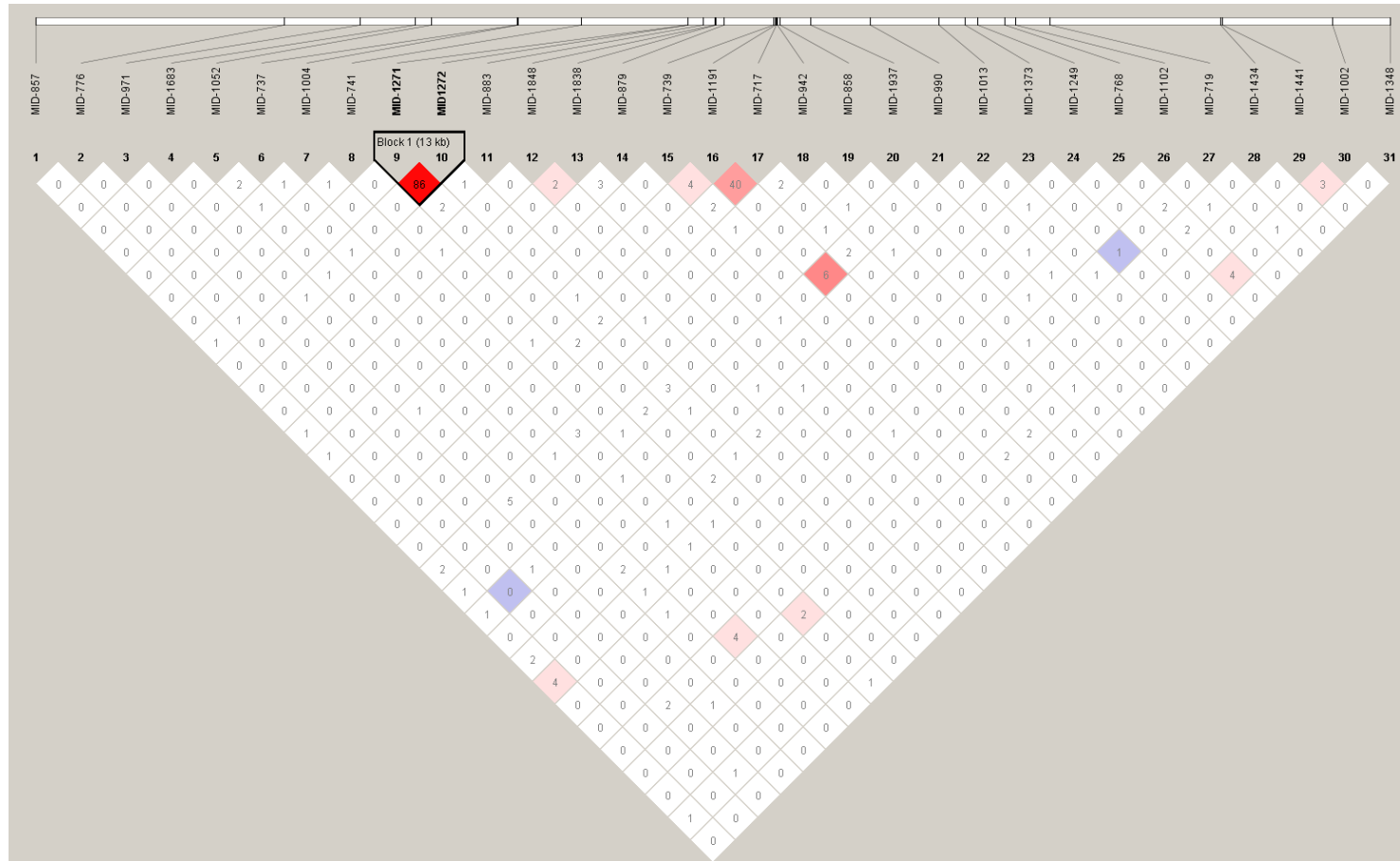


Figura 11- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para o Leste Asiático. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $LOD < 2$, $D' < 1$; azul: $LOD < 2$, $D' = 1$; graus de rosa/vermelho: $LOD = 2$, $D' < 1$; vermelho escuro: $LOD = 2$, $D' = 1$.

Oceania

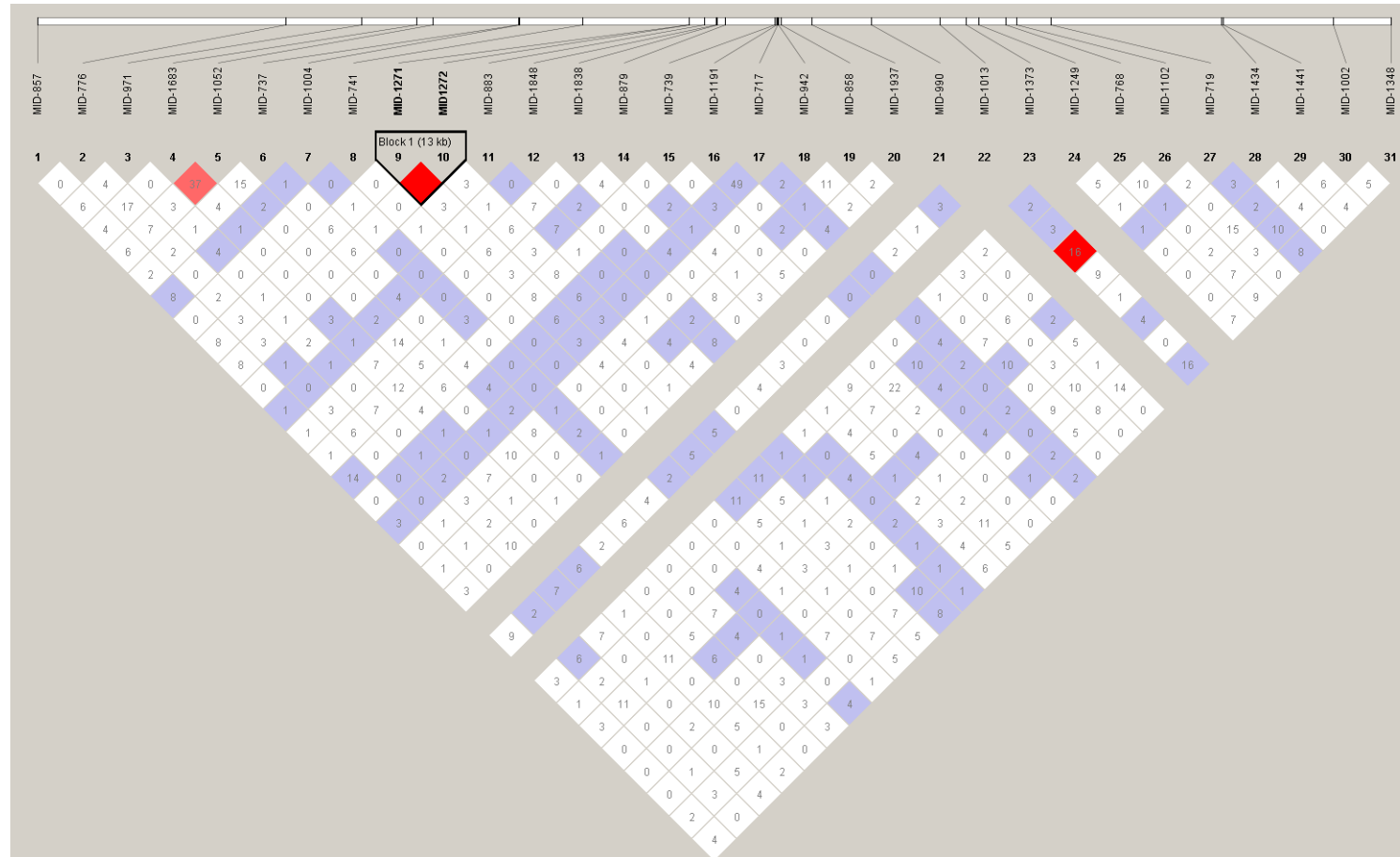


Figura 12- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Oceania. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $\text{LOD} < 2$, $D' < 1$; azul: $\text{LOD} < 2$, $D' = 1$; graus de rosa/vermelho: $\text{LOD} = 2$, $D' < 1$; vermelho escuro: $\text{LOD} = 2$, $D' = 1$.

Europa

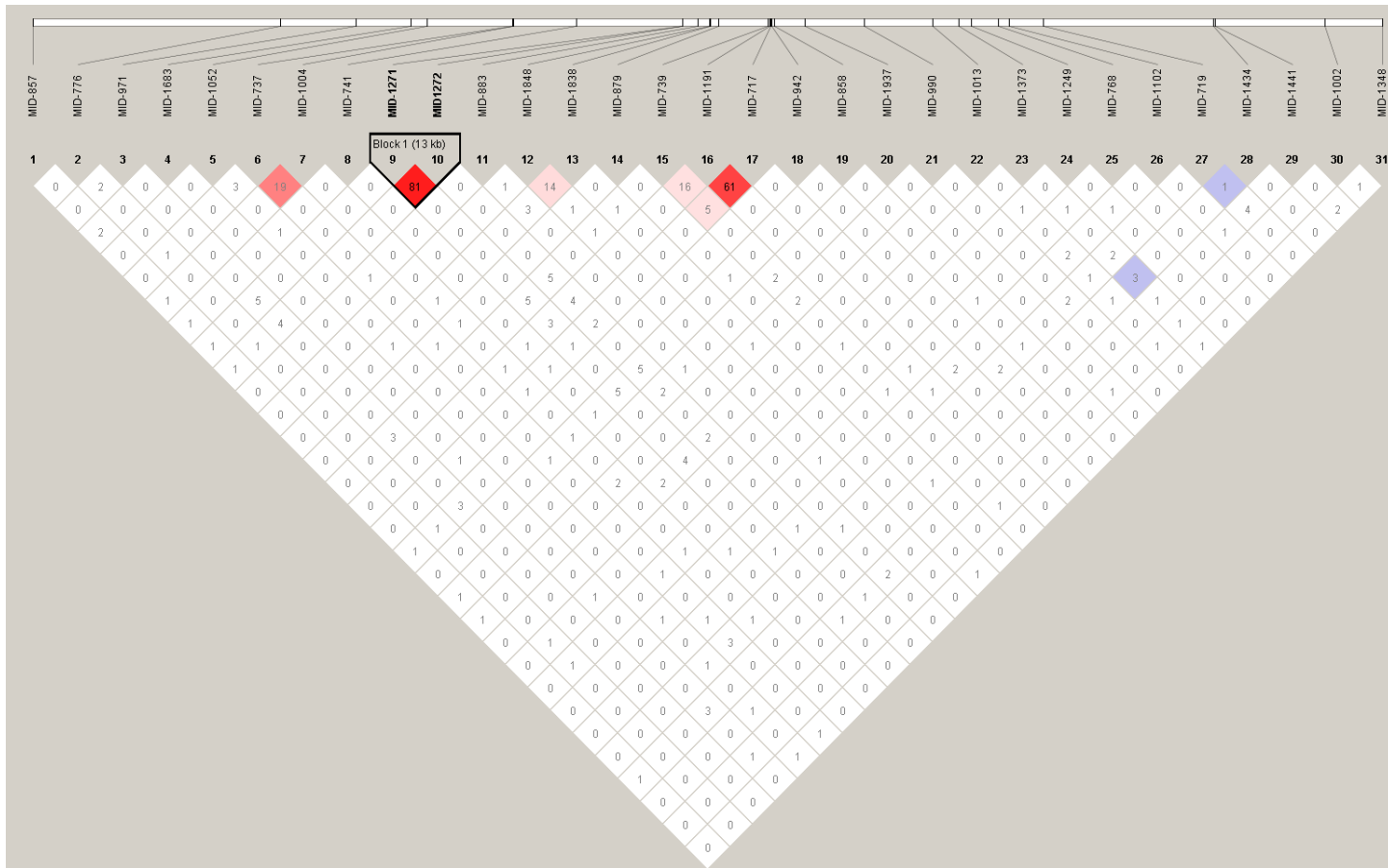


Figura 13- Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para a Europa. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $LOD < 2$, $D' < 1$; azul: $LOD < 2$, $D' = 1$; graus de rosa/vermelho: $LOD = 2$, $D' < 1$; vermelho escuro: $LOD = 2$, $D' = 1$.

Américas

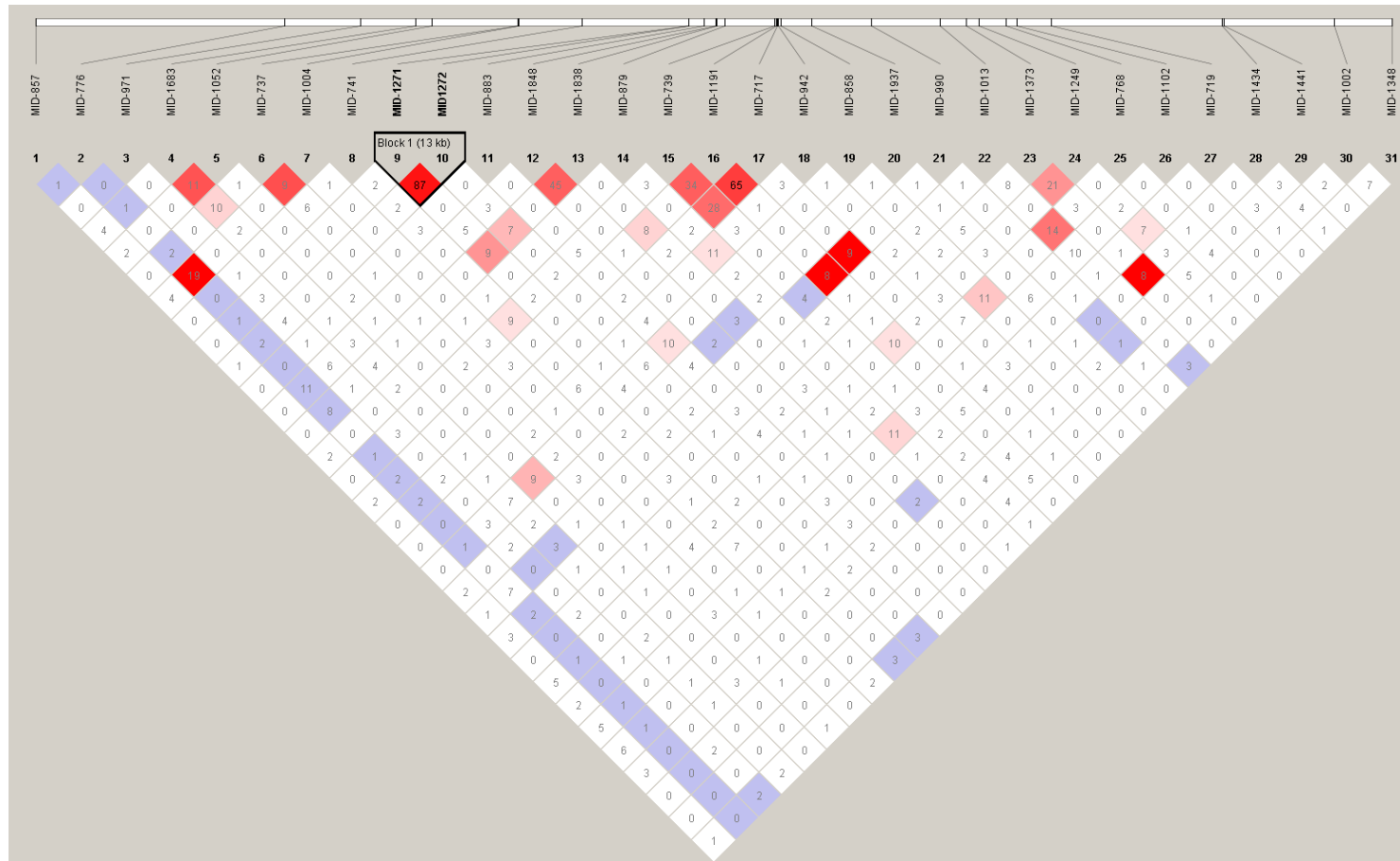


Figura 14 – Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para as Américas. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico. Branco: $LOD < 2$, $D' < 1$; azul: $LOD < 2$, $D' = 1$; graus de rosa/vermelho: $LOD = 2$, $D' < 1$; vermelho escuro: $LOD = 2$, $D' = 1$.

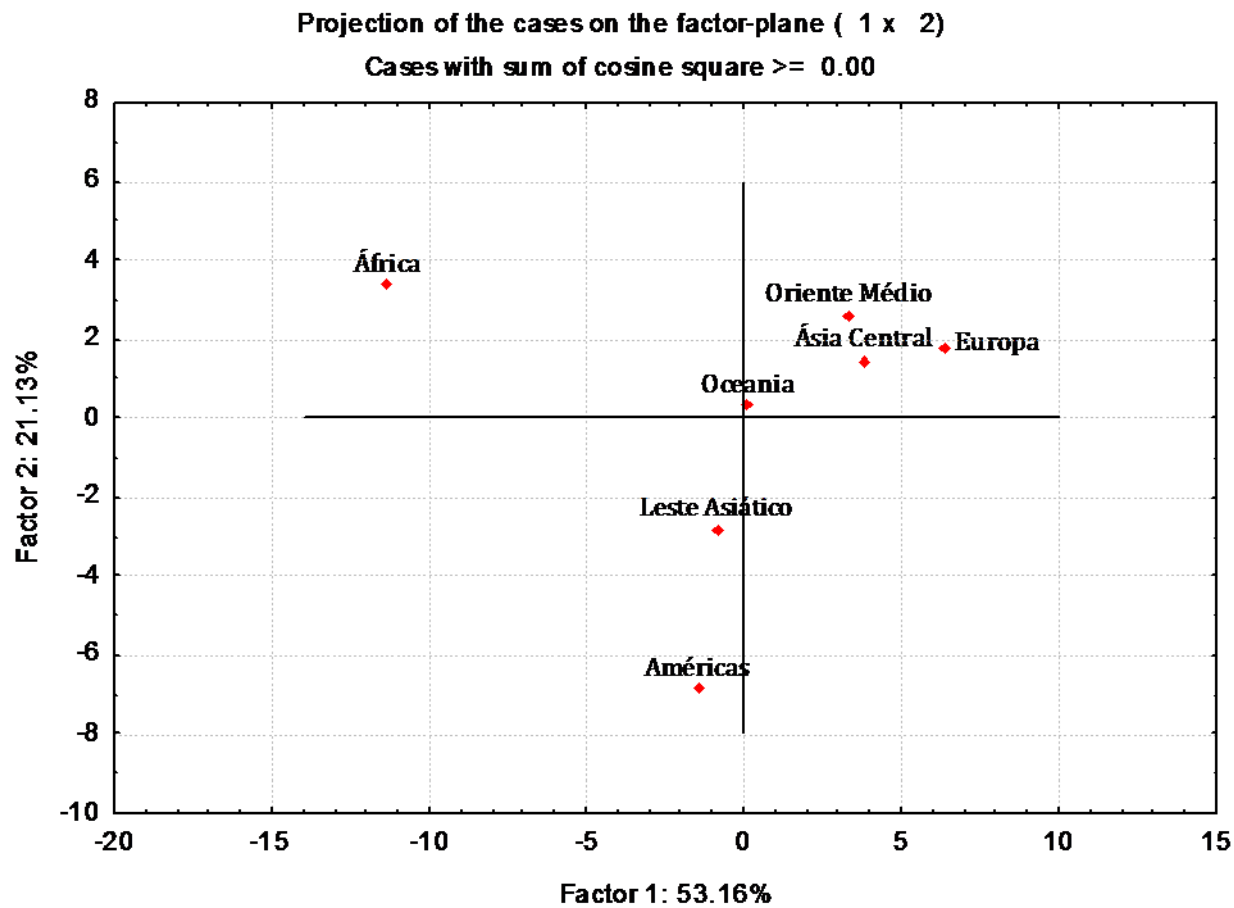


Figura 15 – Gráfico de dispersão dos dois primeiros componentes principais baseados nas freqüências alélicas dos INDELs do cromossomo 5 nas grandes regiões geográficas.

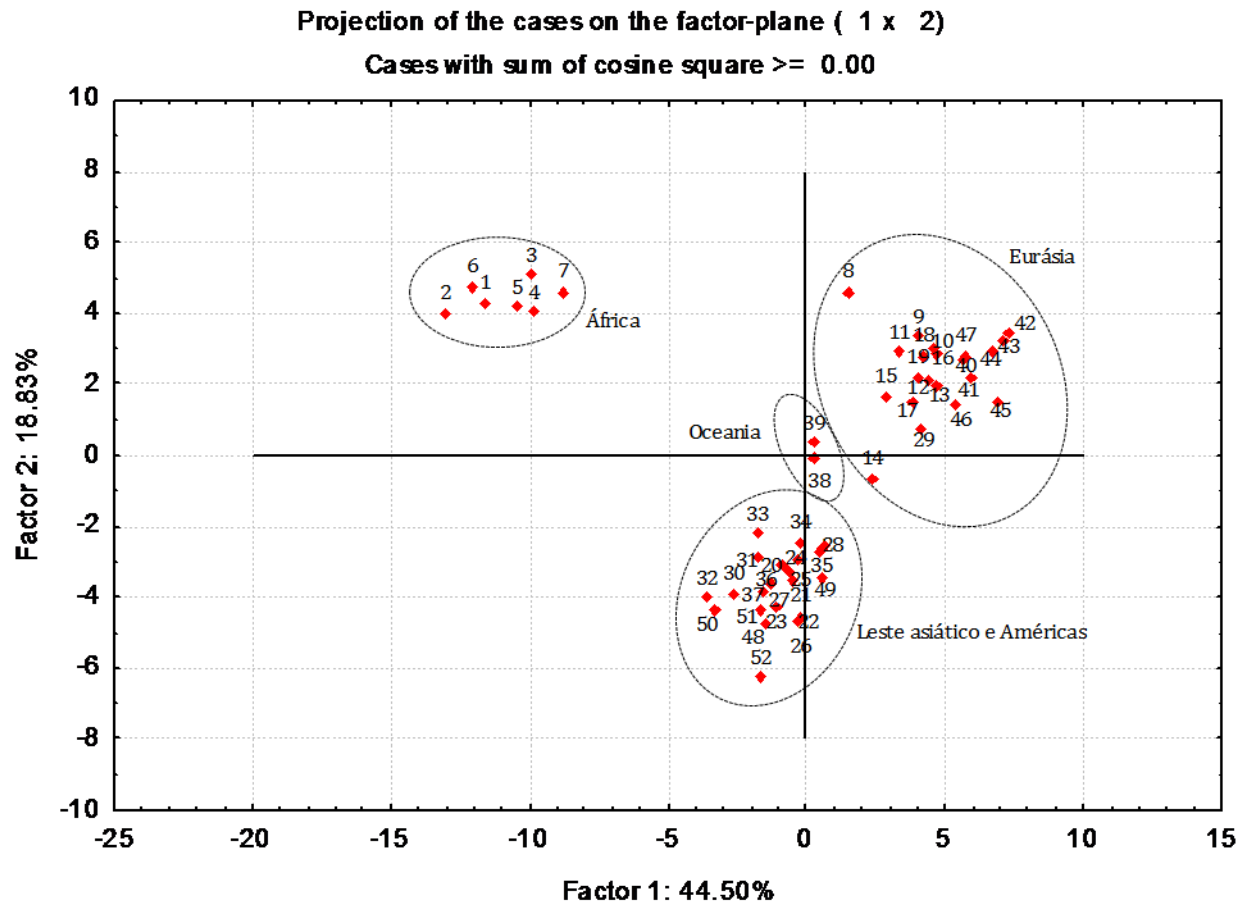


Figura 16 - Gráfico de dispersão dos dois primeiros componentes principais baseados nas frequências alélicas dos INDELS do cromossomo 5 nas 52 populações mundiais. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uyгур, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

2.4.3 – Populações Brasileiras

Além do estudo de populações mundiais, o outro enfoque do nosso trabalho foi o estudo das raízes ancestrais da população brasileira. Para isto, também genotipamos o nosso painel de 31 INDELS informativos de ancestralidade entre africanos e europeus, localizados no braço longo do cromossomo 5 (Figura 2) em duas populações brasileiras: uma de indivíduos pretos e outra de indivíduos brancos da cidade de São Paulo. Em todas as análises referentes às populações brasileiras incluímos as populações da África, Europa e Américas do painel do CEPH-HGDP, já que estas constituem as populações parentais da população brasileira.

Em seguida à genotipagem das amostras de pretos e brancos da cidade de São Paulo, iniciamos os cálculos dos parâmetros básicos de genética de populações como frequências alélicas e análise do equilíbrio de Hardy-Weinberg. As frequências alélicas de cada um dos marcadores encontradas nas populações de pretos e brancos de São Paulo, assim como na África, Europa e Américas estão representadas na tabela 7. O teste de equilíbrio de Hardy-Weinberg mostrou que entre os 713 valores obtidos 33 (0,046) foram significativos em nível de 0,05, não havendo assim evidência de desvios.

Tabela 7 – Frequências alélicas dos INDELS encontradas nas populações parentais do painel do CEPH-HGDP e nas populações brasileiras.

MID	dbSNP ID	Frequência do alelo maior				
		Africa (n=124)	Europa (n=155)	America (n=108)	Pretos SP (n=104)	Branco SP (n=157)
MID-768	rs1160979	0.887	0.190	0.611	0.625	0.245
MID-879	rs1610963	0.960	0.406	0.269	0.697	0.462
MID-1013	rs1611095	0.778	0.229	0.713	0.538	0.280
MID-719	rs1160938	0.782	0.152	0.537	0.466	0.191
MID-1102	rs2067190	0.871	0.155	0.528	0.615	0.213
MID-1272	rs2067354	0.931	0.113	0.319	0.601	0.175
MID-990	rs1611072	0.024	0.348	0.028	0.125	0.258
MID-1683	rs2307879	0.875	0.181	0.509	0.623	0.276
MID-1191	rs2067278	0.980	0.432	0.736	0.779	0.500
MID-883	rs1610967	0.081	0.652	0.500	0.382	0.578
MID-737	rs1160951	0.996	0.545	0.264	0.799	0.636
MID-1937	rs2308129	0.839	0.326	0.764	0.647	0.380
MID-739	rs1160953	0.750	0.390	0.606	0.623	0.419
MID-1348	rs2307544	0.750	0.284	0.074	0.446	0.286
MID-776	rs1610580	0.911	0.597	0.991	0.825	0.728
MID-858	rs673423	0.310	0.602	0.366	0.469	0.533
MID-1271	rs2067353	0.923	0.120	0.329	0.625	0.176
MID-1249	rs2067331	0.403	0.143	0.611	0.286	0.152
MID-717	rs1160936	0.883	0.468	0.759	0.698	0.507
MID-1838	rs2308034	0.694	0.263	0.097	0.531	0.325
MID-1441	rs2307637	0.589	0.279	0.366	0.422	0.242
MID-942	rs1611025	0.903	0.468	0.639	0.802	0.451
MID-1052	rs2067141	0.843	0.353	0.861	0.682	0.413
MID-741	rs32861	0.393	0.297	0.229	0.385	0.254
MID-1004	rs1611086	0.460	0.688	0.954	0.604	0.647
MID-1848	rs2308044	0.315	0.831	0.926	0.552	0.839
MID-1373	rs2307569	0.927	0.708	0.764	0.885	0.671
MID-1434	rs2307630	0.601	0.078	0.213	0.354	0.140
MID-971	rs1611053	0.895	0.458	0.241	0.703	0.451
MID-857	rs1610942	0.629	0.223	0.606	0.639	0.272
MID-1002	rs1611084	0.621	0.506	0.079	0.482	0.473

2.4.3.1 – Escalonamento Multidimensional (MDS) nas populações brasileiras e parentais

Após análise dos parâmetros básicos, calculamos as distâncias genéticas entre as populações parentais e as populações brasileiras através da construção de matrizes de distância de Reynolds que, posteriormente, foram submetidas a um escalonamento multidimensional. Nesta análise, dois gráficos foram produzidos: um no qual consideramos todas as populações de uma mesma região geográfica uma única população (Figura 17) e outra onde as diferentes populações pertencentes às grandes regiões geográficas foram discriminadas (Figura 18). Com somente duas dimensões, obtivemos representações gráficas muito adequadas das matrizes de distância. (*stress* de 0,0000 e 0,0655, respectivamente). Em nossa primeira análise, onde as populações parentais foram tratadas como grandes regiões geográficas, podemos observar África, Europa e América bem distantes entre si. Os pretos brasileiros aparecem em uma posição entre as três populações parentais, no entanto, mais próximos da África. Já os brancos aparecem próximos da Europa.

Em nossa segunda análise, em que as populações da África, Europa e América foram discriminadas, é possível observar três agrupamentos: um contendo as populações das Américas, outro abrangendo as populações da África e o último incluindo as populações da Europa e os brancos brasileiros. Os pretos brasileiros ficaram fora de qualquer agrupamento, em uma posição intermediária aos três grupos formados. As populações africanas San e Mbuti também não se incluíram

em nenhum agrupamento, estando a primeira acima e a segunda abaixo do grupo formado pelas populações da África. As populações nativo-americanas aparecem mais espalhadas entre si dentro do grupo do que as populações dos demais agrupamentos.

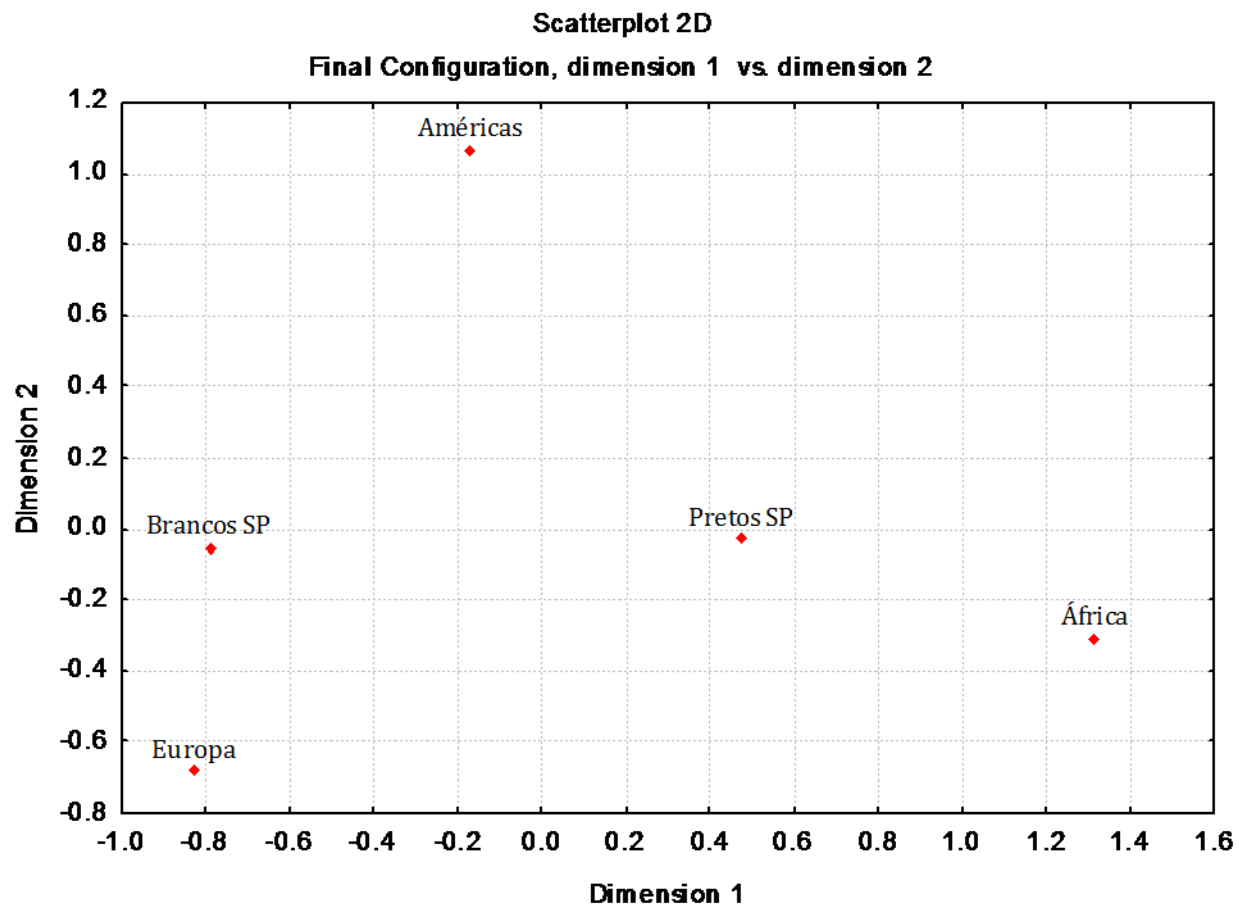


Figura 17 – Escalonamento Multidimensional realizado com as populações brasileiras e as grandes regiões geográficas parentais

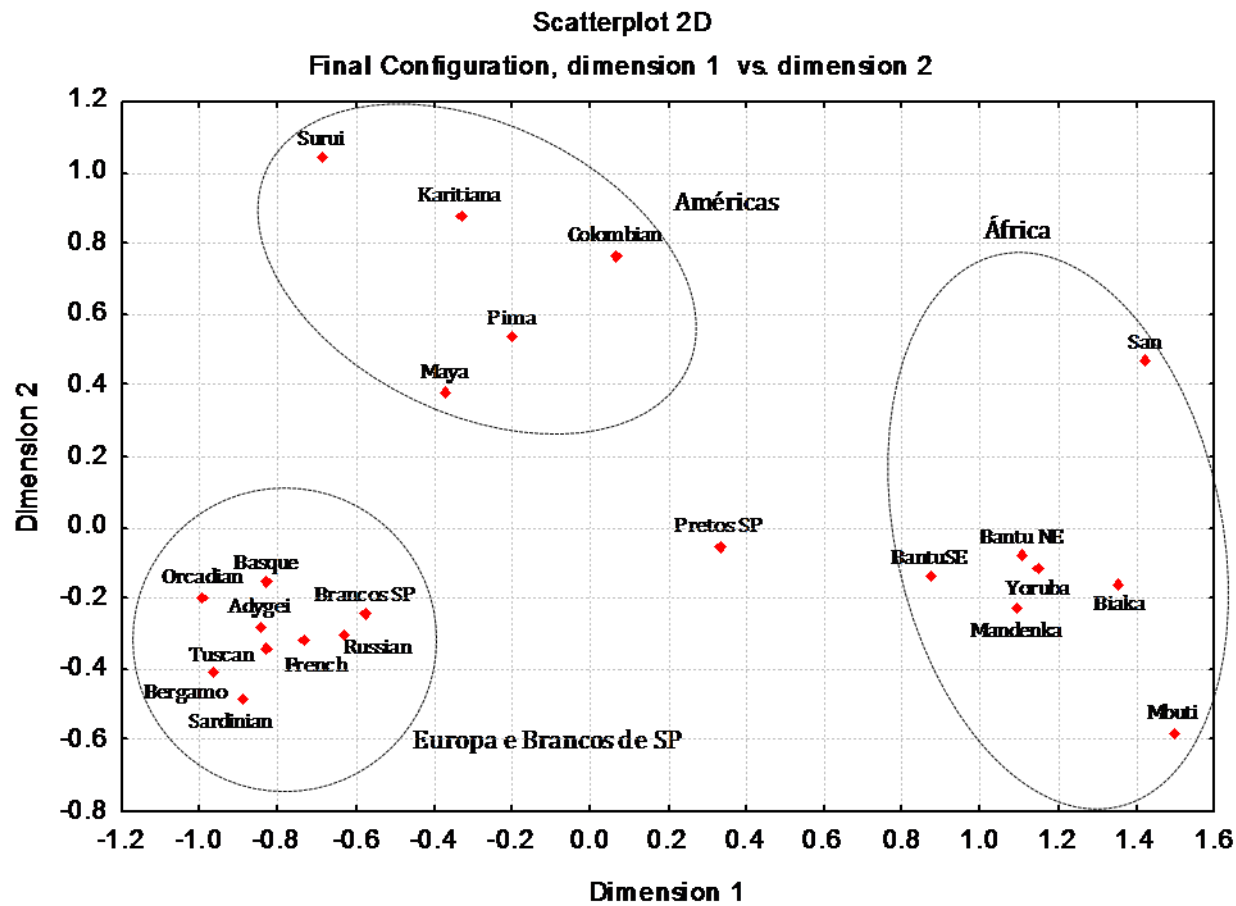


Figura 18 - Escalonamento Multidimensional realizado com as populações brasileiras e as populações pertencentes às grandes regiões geográficas parentais.

2.4.3.2 – Análise de desequilíbrio de ligação nas populações brasileiras

Assim como para as populações mundiais, verificamos através do programa *HaploView 4.1*, se os INDELS ao longo do cromossomo 5 se encontravam em desequilíbrio de ligação em cada uma das populações brasileiras (Figuras 19 e 20). Foi observado desequilíbrio de ligação significativo apenas entre os marcadores MID-1271 e MID-1272 ($r^2 = 0,95$) na população de pretos de São Paulo (Figura 19).

2.4.3.3 – Análise do Componente Principal (PCA) em populações brasileiras e parentais

Nossa próxima análise foi verificar, de modo qualitativo, se os INDELS selecionados seriam capazes de distinguir as três populações parentais e localizar as populações de brancos e pretos de São Paulo em relação a África, Europa e América. Para isto, os dados de frequências alélicas dos INDELS nas populações parentais e brasileiras foram submetidos à Análise do Componente Principal (PCA). Em primeira análise as populações parentais foram tratadas como grandes regiões geográficas (África, Europa e América), enquanto que em uma segunda análise as populações de cada região geográfica parental foram discriminadas.

Pretos de SP

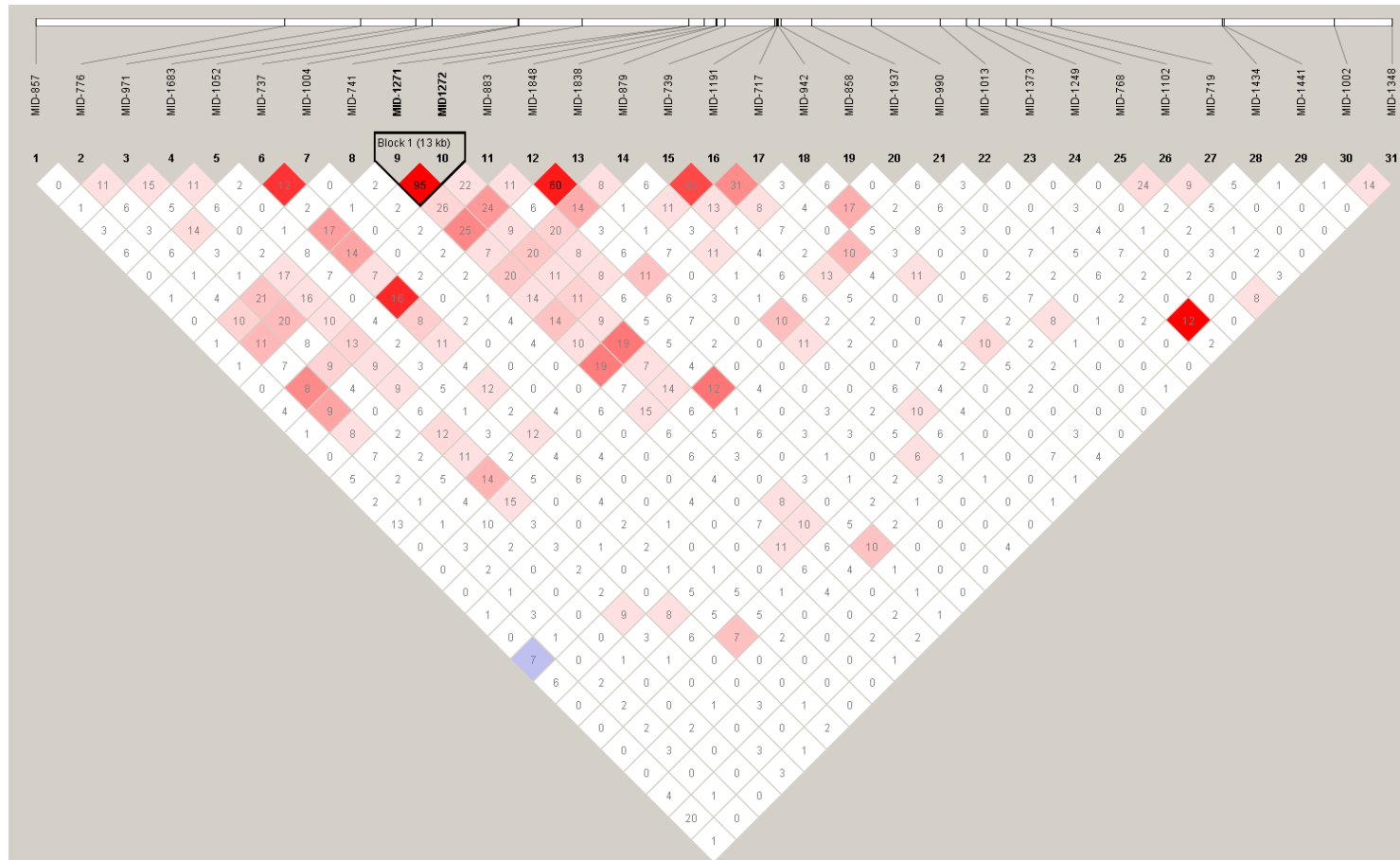


Figura 19– Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para pretos brasileiros. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico.

Brancos de SP

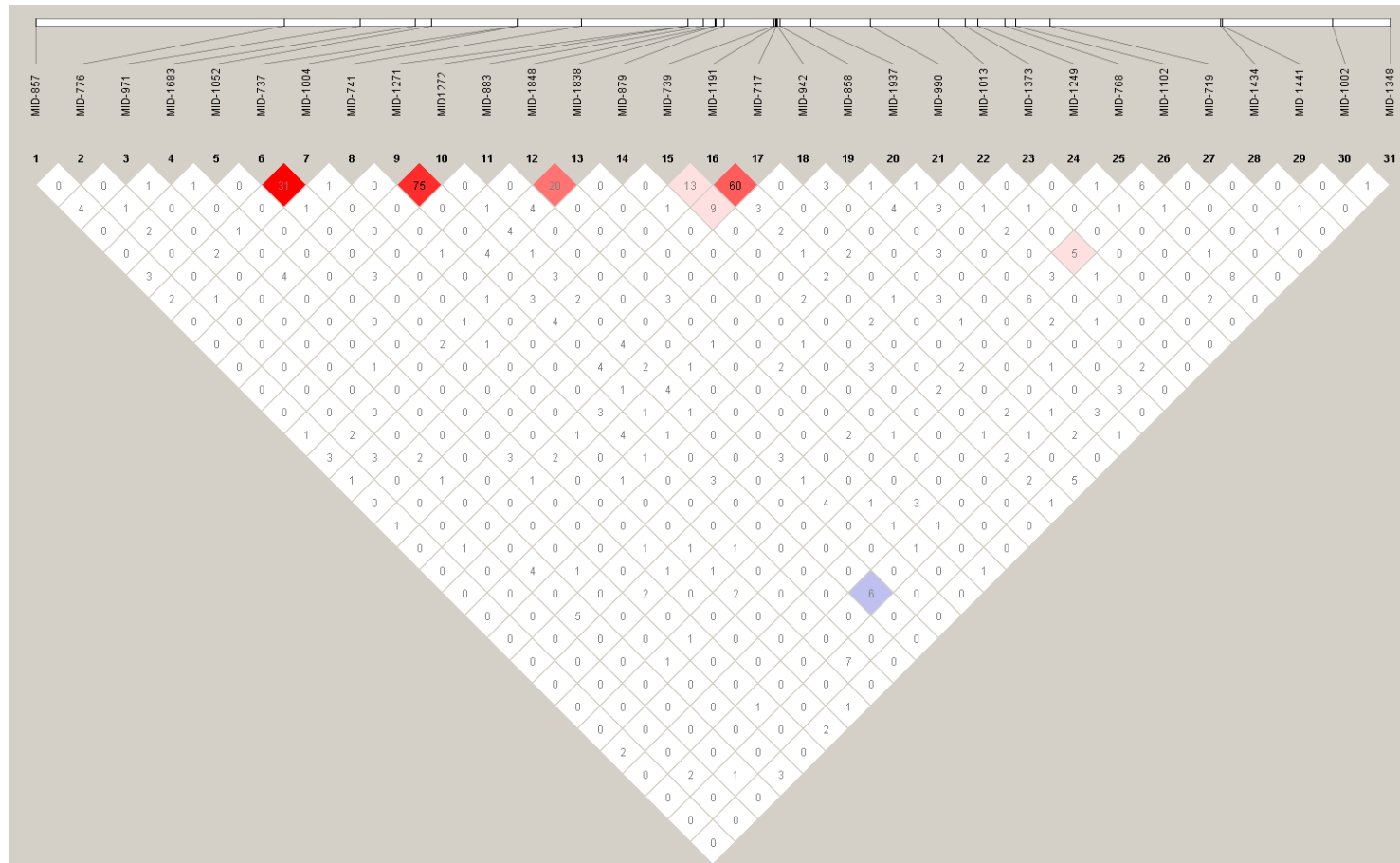


Figura 20 – Análise de desequilíbrio de ligação realizada pelo programa HaploView 4.1 para pretos e brancos brasileiros. Os valores dentro dos quadrados correspondem ao valor de r^2 que quando superior a 0,8 constitui um bloco haplotípico.

Em ambas as análises os valores do primeiro e segundo componentes foram utilizados para a construção gráficos de dispersão bidimensional (Figuras 21 e 22). O total de variação genética recuperada pelos dois primeiros componentes nas duas análises foi de 97,92% e 80,53%, respectivamente. No primeiro gráfico (Figura 21) podemos observar três agrupamentos claros: africanos e indivíduos pretos de São Paulo, europeus e indivíduos brancos de São Paulo e ameríndios. A análise também indica que os indivíduos brancos brasileiros estão mais próximos dos europeus do que os pretos dos africanos.

Na segunda análise onde as populações da África, América e Europa foram discriminadas, o resultado não se mostrou diferente. Os grupos parentais se apresentam bem definidos com os brancos completamente incluídos no grupo da Europa. Os pretos brasileiros estão fora de qualquer grupo, no entanto próximos do grupo constituído por populações africanas (Figura 22).

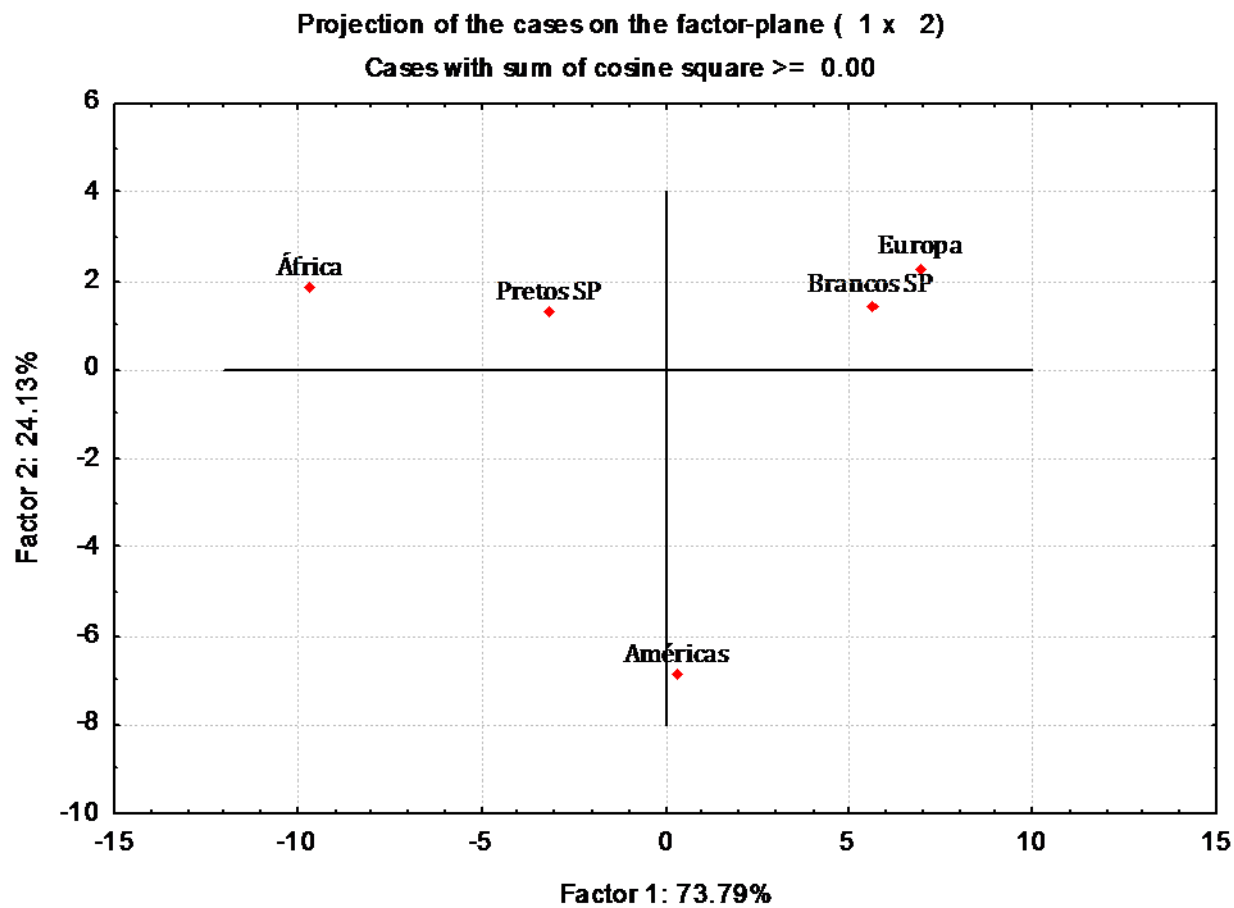


Figura 21 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências relativas de alelos dos INDELS do cromossomo 5 nas regiões geográficas parentais e nas populações brasileiras. Os valores dos componentes 1 e 2 estão mostrados no gráfico.

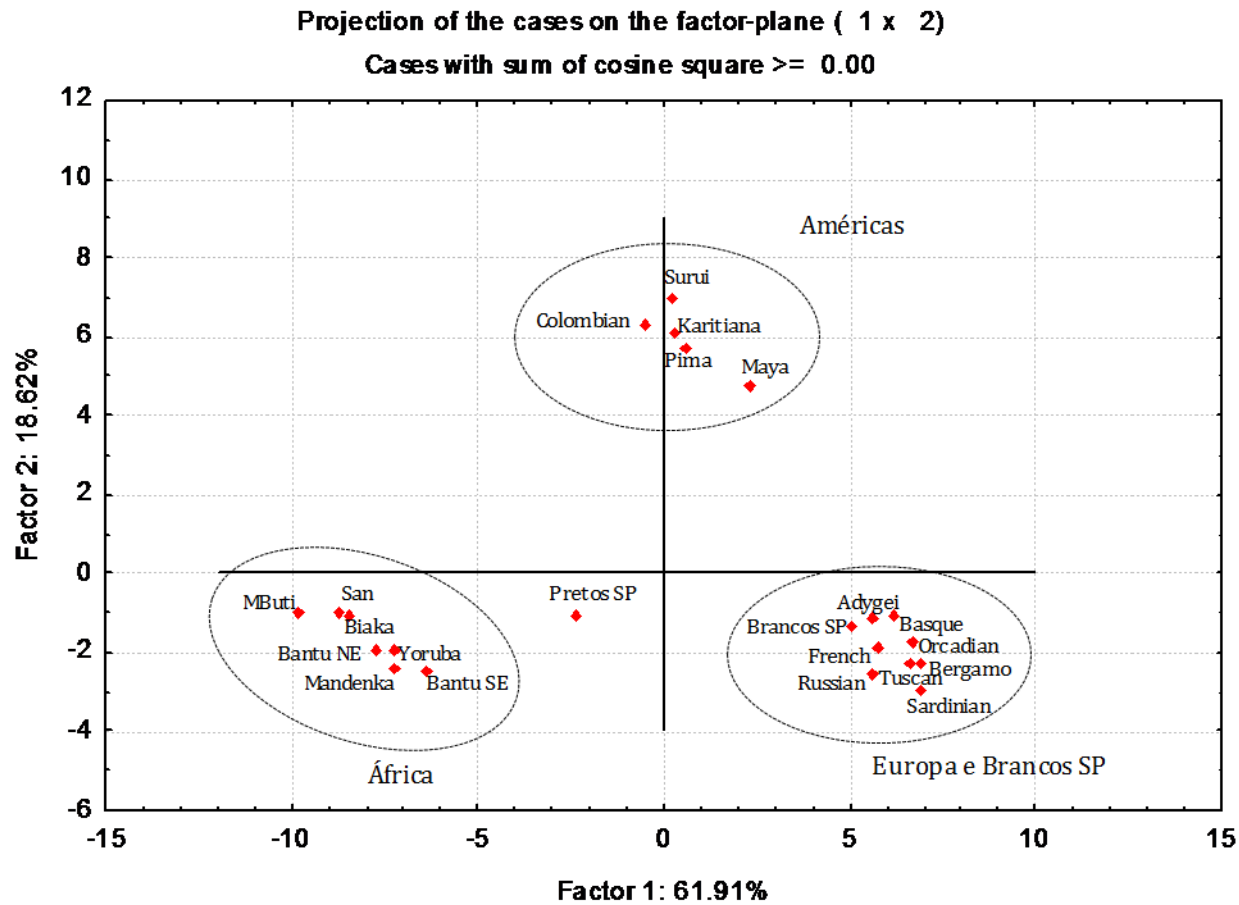


Figura 22 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas freqüências relativas de alelos dos INDELs do cromossomo 5 nas populações parentais e nas populações brasileiras. Os valores dos componentes 1 e 2 estão mostrados no gráfico.

2.4.3.4 – Análise de mistura genética nas populações brasileiras

Nosso próximo passo foi estimar de maneira quantitativa a contribuição genética das populações parentais (África, Europa e América) para a ancestralidade dos brancos e pretos de São Paulo. Para isto, utilizamos o programa *Admix 2.0* (Dupanloup & Bertourelle, 2001). Para os indivíduos brancos da cidade de São Paulo, o resultado mostrou uma contribuição africana de $6,5 \pm 1,6\%$, europeia de $86,8 \pm 2,3\%$ e ameríndia de $6,8 \pm 3,1\%$. Para os indivíduos pretos da cidade de São Paulo a contribuição africana foi de $55,1 \pm 1,9\%$, europeia de $34,0 \pm 2,4\%$ e ameríndia de $10,9 \pm 3,3\%$ (Tabela 8).

2.4.3.5 - Estrutura populacional

Nossa próxima análise tinha o objetivo de verificar se nossos marcadores seriam capazes de distinguir as três populações parentais. Para isto submetemos os dados genotípicos das amostras das populações parentais do painel do CEPH-HGDP (África, Europa e América) ao programa *Structure*. Utilizamos o modelo de ligação e não fornecemos ao programa nenhuma informação inicial a respeito da origem geográfica das amostras. Realizamos 5 corridas independentes com cada K que variou de 2 a 5. Em nossos resultados, o $K=3$ foi o mais provável (0,9999) e os 3 agrupamentos corresponderam às três populações parentais (Figura 23). Em nível individual, os africanos apresentaram em média 89,0% da região cromossômica

estudada de ancestralidade africana, os europeus 74,3% de ancestralidade europeia e os ameríndios 73,7% de ancestralidade ameríndia.

Tabela 8 – Porcentagem de contribuição das populações parentais para as populações de Brancos e Pretos de São Paulo. Resultado gerado pelo programa *Admix 2.0*.

Populações Parentais	Contribuição das populações parentais (%)	
	Brancos de São Paulo	Pretos de São Paulo
África	6,5±1,6	55,1±1,9
Europa	86,8±2,3	34,0±2,4
América	6,8±3,1	10,9±3,3

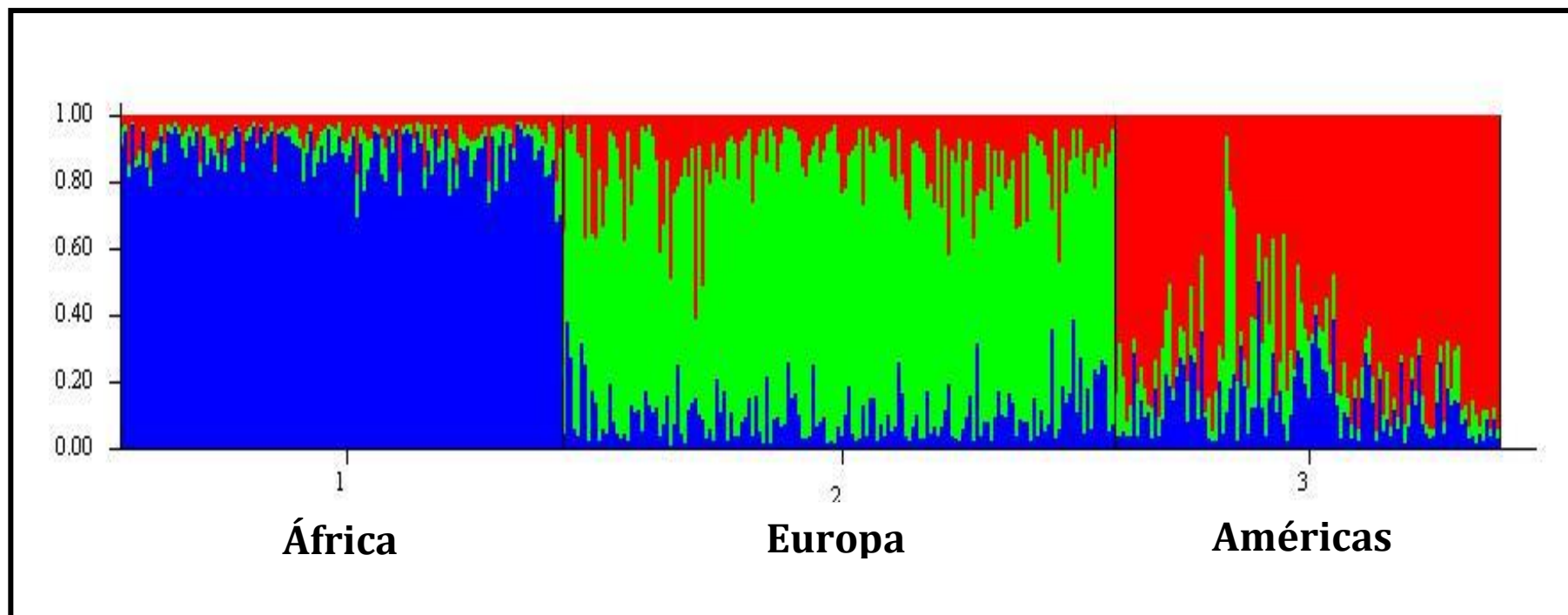


Figura 23 - Estrutura populacional estimada pelo programa *Structure* das três regiões que abrangem as populações parentais da população brasileira. O gráfico representa um total de 387 indivíduos e as cores das linhas dependem da fração de ancestralidade estimada para cada indivíduo.

2.4.3.6 – Inferência de haplótipos

Em seguida, partimos para a inferência dos pares de haplótipos de cada indivíduo das populações parentais e brasileiras. Para isso, utilizamos o programa *Phase* que infere, a partir de dados genotípicos dos indivíduos, a fase gamética em que os alelos de diferentes locos se encontram. Portanto, são inferidos dois haplótipos para cada indivíduo.

Um total de 231 diferentes haplótipos foram inferidos para a África (n=124), 302 para a Europa (n=155), 184 para a América (n=108), 309 para os brancos de São Paulo (n=157) e 203 haplótipos para amostras de pretos de São Paulo (n=104). O valor médio de certeza de inferência dos pares de haplótipos individuais foi de 0,041 para África, 0,12 para Europa, 0,21 para América, 0,08 para os pretos de São Paulo e 0,06 para amostras de brancos de São Paulo. Embora as probabilidades de inferência dos haplótipos tenham sido muito baixas para todas as populações analisadas, a probabilidade média de inferência da fase por sítio ambíguo foi relativamente alta em todas as amostras estudadas: 0,72 para África, 0,80 para Europa, 0,83 para América, 0,81 para brancos e 0,81 para pretos de São Paulo.

2.4.3.7 - Origem ancestral dos fragmentos cromossômicos nas populações miscigenadas

Nosso próximo passo foi inferir a provável origem ancestral de cada loco, para cada cromossomo de cada indivíduo das populações brasileiras. Para isto, dados haplotípicos (gerados pelo PHASE) das populações parentais e dos brancos e pretos de São Paulo foram submetidos ao programa *Structure* sob o modelo de ligação. Nesta análise, pré-identificamos as populações parentais e consideramos o $K=3$. Os resultados apresentam as porcentagens de ancestralidade africana, europeia e ameríndia para cada indivíduo amostrado e indicam as probabilidades de cada loco (em cada cromossomo) terem se originado de cada uma das populações parentais.

Através da utilização das informações de ancestralidade de cada loco de cada cromossomo de cada indivíduo inferidas pelo *Structure*, construímos gráficos que representam os pares homólogos de cada indivíduo das populações parentais e brasileiras. Para isto utilizamos a formatação condicional do programa *Excel*: quando a probabilidade de origem ancestral africana, europeia ou ameríndia de determinado loco era maior ou igual a 0,5, este loco foi colorido respectivamente de azul, verde ou vermelho e foi considerado originado da África, Europa ou América, respectivamente. Quando a probabilidade de origem ancestral africana, europeia e ameríndia de determinado loco era menor que 0,5, este foi colorido de cinza e considerado de origem ancestral indeterminada (Figuras 24 a 28).

Nossos resultados mostraram que, para as populações parentais, os segmentos cromossômicos originaram-se predominantemente de uma única população, embora fragmentos cromossômicos com outra ancestralidade também tenham sido observados (Figuras 24, 25 e 26). É possível observar uma maior homogeneidade na origem ancestral dos fragmentos cromossômicos dos africanos quando comparados com europeus e ameríndios.

Para as populações de brancos e pretos de São Paulo, foi possível identificar a origem ancestral da grande maioria dos locos e, em contraste com os resultados encontrados para as populações parentais, as populações brasileiras apresentaram fragmentos cromossômicos originados das três populações ancestrais (Figuras 27 e 28). Como esperado, nos brancos de São Paulo foi encontrado maior número de fragmentos cromossômicos de origem europeia do que africana e ameríndia. Já para os pretos de São Paulo, é possível observar muitos fragmentos cromossômicos originados da África, Europa e, um pouco menos, da América.

Através da contagem do número de locos de origem ancestral africana, europeia e ameríndia (quadrinhos azuis, verdes e vermelhos das figuras 27 e 28), fizemos uma estimativa, a nível populacional, das proporções de contribuição ancestral para os pretos e brancos brasileiros. Os resultados revelaram que para pretos de São Paulo, 63,2% dos locos em que a inferência da ancestralidade foi possível, eram de origem africana, 23,0% de origem europeia e 13,8% de origem ameríndia.

Para os brancos brasileiros estes valores foram respectivamente 11,4%, 70,0% e 18,7%.

Em nível individual, os pretos apresentaram em média 53,3% da região cromossômica estudada de ancestralidade africana, 27,2% de ancestralidade européia e 19,5% de ancestralidade ameríndia, enquanto que, para os brancos, estes valores foram de 16,0%, 60,0% e 24,0%, respectivamente.

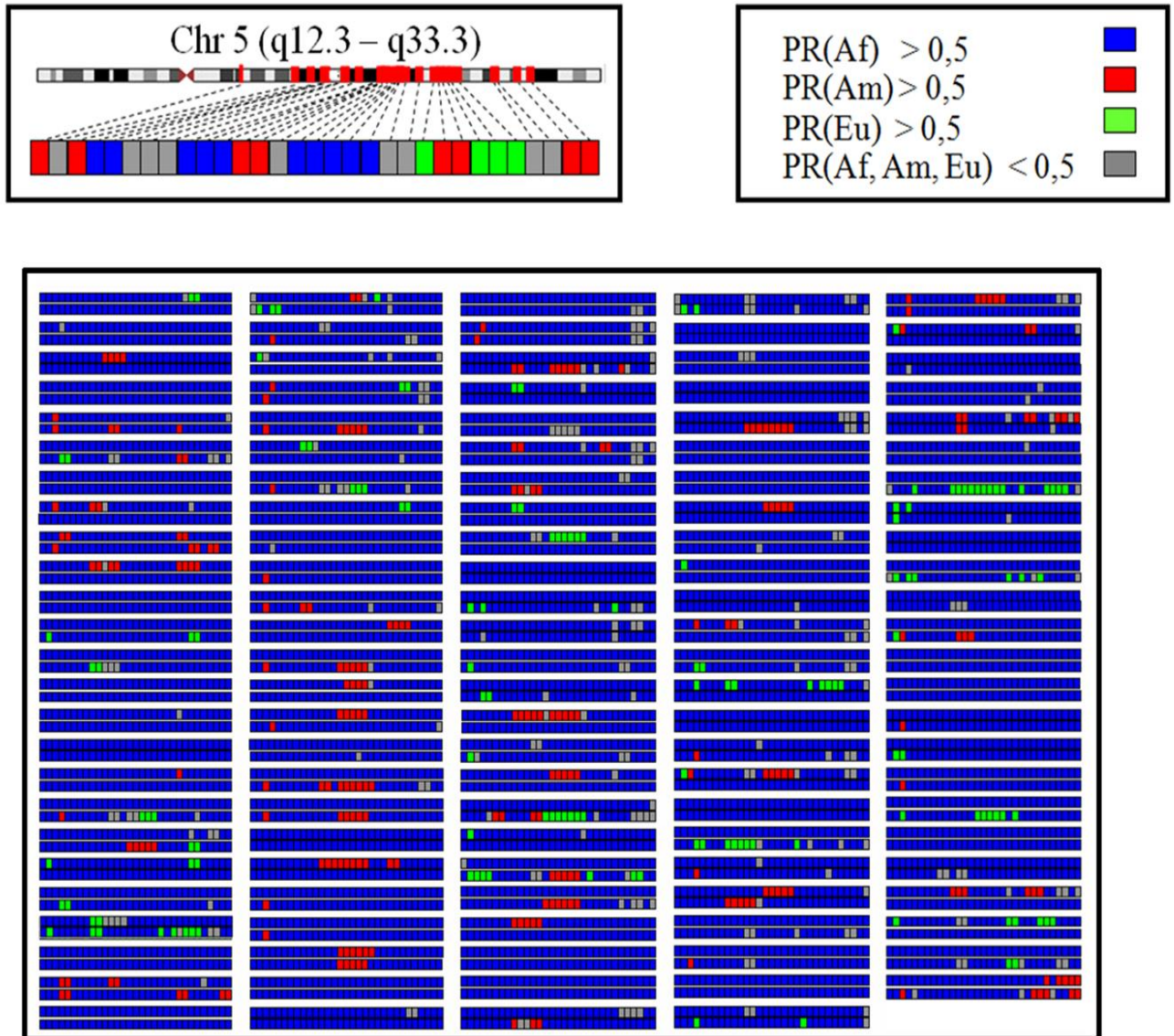


Figura 24: Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da África. Os quadrados representam os locos analisados e as cores indicam a ancestralidade de cada loco inferida pelo programa *Structure*: Azul, quando a probabilidade do loco ter ancestralidade africana > 0,5, verde, quando a probabilidade do loco ter ancestralidade europeia > 0,5, vermelho, quando a probabilidade do loco ter ancestralidade ameríndia > 0,5 e cinza quando a inferência da ancestralidade não foi possível.

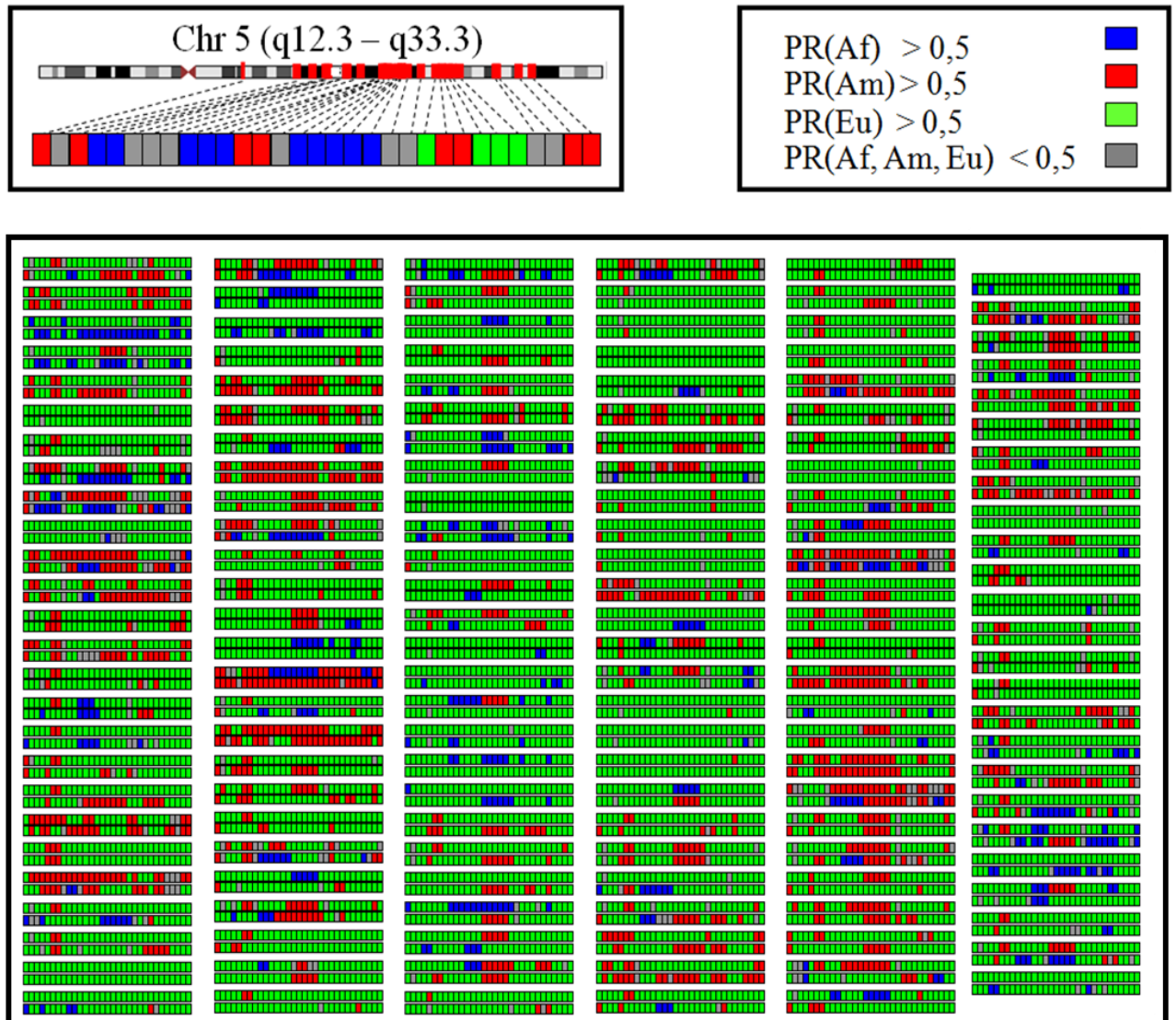


Figura 25 - Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da Europa. Os quadrados representam os locos analisados e as cores indicam a ancestralidade de cada loco inferida pelo programa *Structure*: Azul, quando a probabilidade do loco ter ancestralidade africana > 0,5, verde, quando a probabilidade do loco ter ancestralidade europeia > 0,5, vermelho, quando a probabilidade do loco ter ancestralidade ameríndia > 0,5 e cinza quando a inferência da ancestralidade não foi possível.

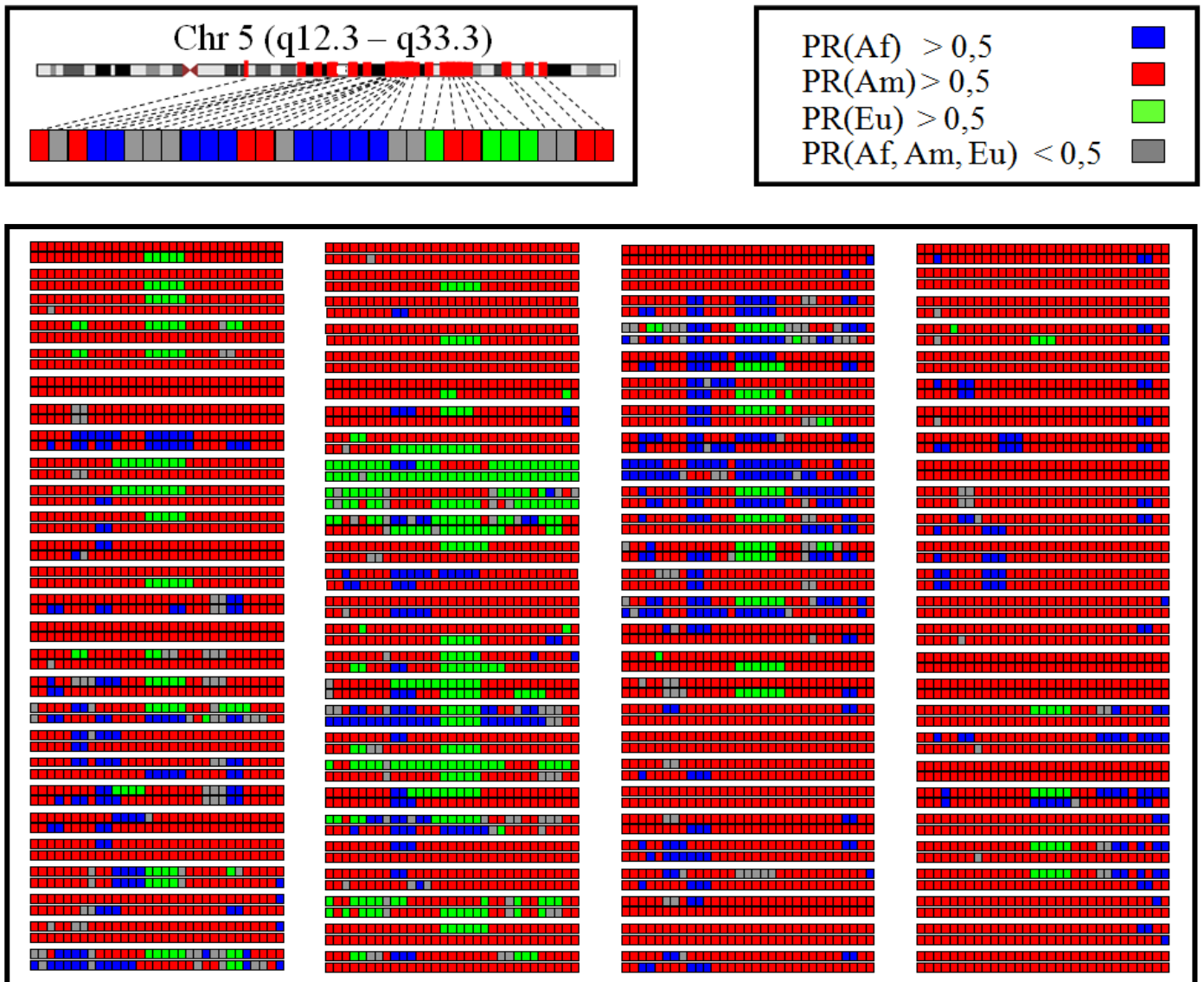


Figura 26: Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo da América. Os quadrados representam os locos analisados e as cores indicam a ancestralidade de cada loco inferida pelo programa *Structure*: Azul, quando a probabilidade do loco ter ancestralidade africana > 0,5, verde, quando a probabilidade do loco ter ancestralidade europeia > 0,5, vermelho, quando a probabilidade do loco ter ancestralidade ameríndia > 0,5 e cinza quando a inferência da ancestralidade não foi possível.

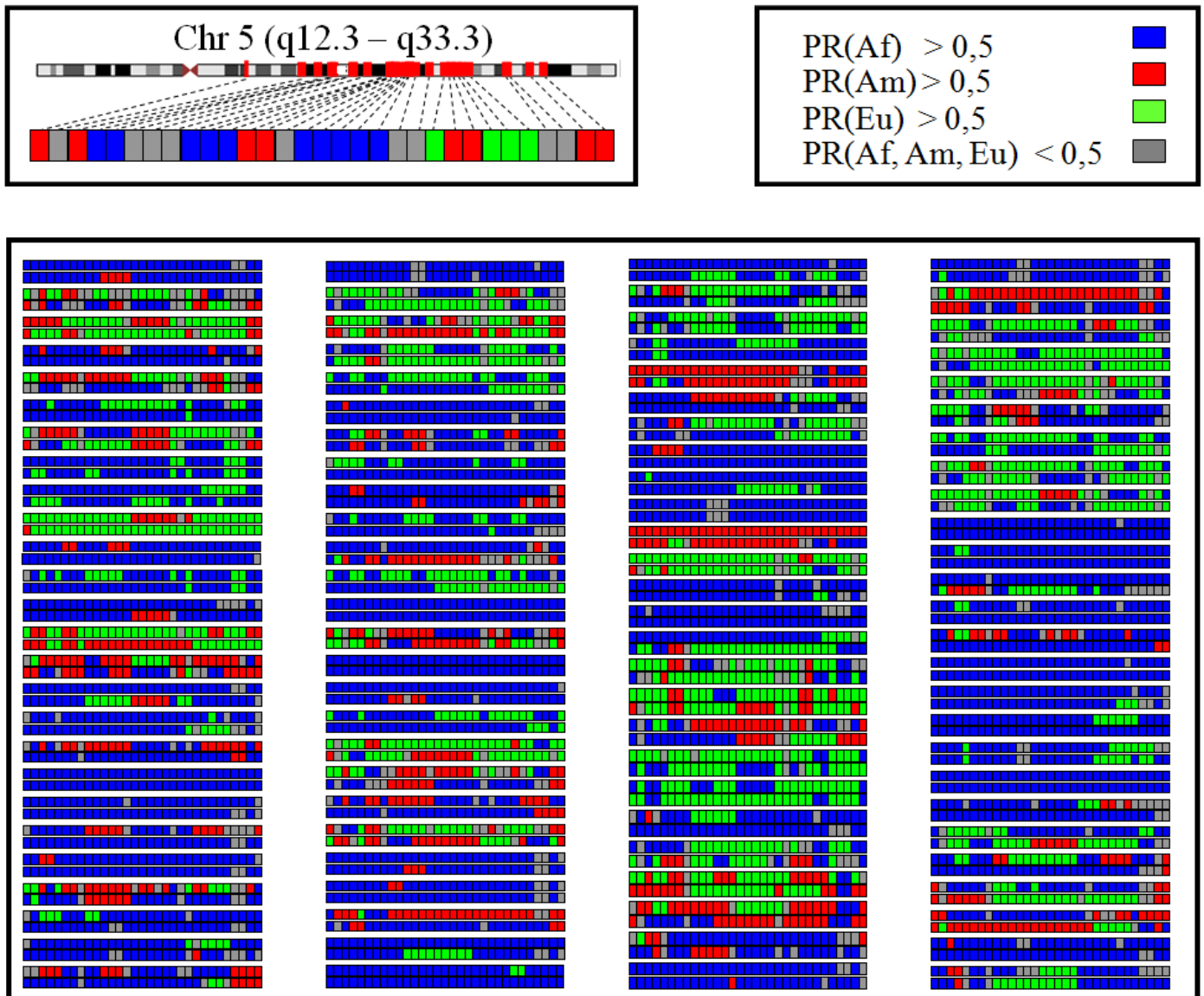


Figura 27: – Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo preto da cidade de São Paulo. Os quadrados representam os locos analisados e as cores indicam a ancestralidade de cada loco inferida pelo programa *Structure*: Azul, quando a probabilidade do loco ter ancestralidade africana > 0,5, verde, quando a probabilidade do loco ter ancestralidade europeia > 0,5, vermelho, quando a probabilidade do loco ter ancestralidade ameríndia > 0,5 e cinza quando a inferência da ancestralidade não foi possível.

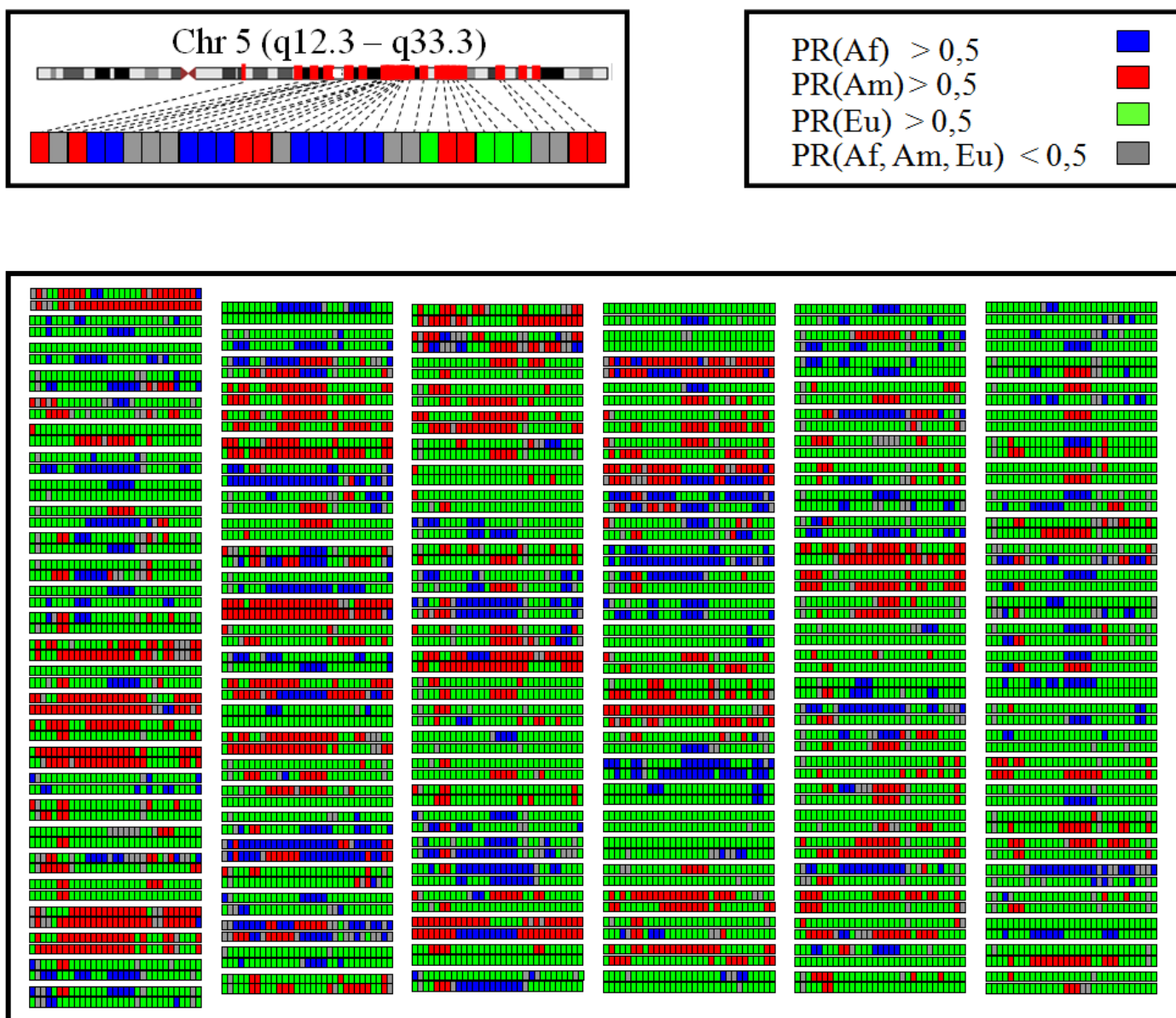


Figura 28: – Desenho esquemático dos pares de cromossomos (relativos a região q12.3q33.3 do cromossomo 5) de cada indivíduo branco da cidade de São Paulo. Os quadrados representam os locos analisados e as cores indicam a ancestralidade de cada loco inferida pelo programa *Structure*: Azul, quando a probabilidade do loco ter ancestralidade africana > 0,5, verde, quando a probabilidade do loco ter ancestralidade européia > 0,5, vermelho, quando a probabilidade do loco ter ancestralidade ameríndia > 0,5 e cinza quando a inferência da ancestralidade não foi possível.

2.4.3.8 – Correlação de ancestralidade entre pares homólogos

Em seguida à inferência da origem ancestral de cada loco, para cada cromossomo de cada indivíduo brasileiro, investigamos um possível padrão de cruzamento entre os indivíduos brasileiros. Para isto, realizamos uma correlação entre o número de locos de uma mesma origem ancestral entre os dois cromossomos 5 homólogos de um mesmo indivíduo. A correlação foi significativa tanto para os pretos ($p < 0,000001$) quanto para os brancos ($p < 0,000001$) da cidade de São Paulo.

2.5 – Discussão

No presente trabalho nós selecionamos um grupo de 31 INDELS bialélicos, localizados no braço longo do cromossomo 5, com ampla variação de frequências alélicas entre africanos e europeus e os utilizamos em dois estudos distintos. O primeiro foi realizado com populações mundiais com o objetivo de verificar estruturação populacional e o poder discriminatório dos marcadores selecionados. O segundo estudo teve como objetivo caracterizar o padrão de mistura da população brasileira através da análise destes marcadores em populações de indivíduos brancos e pretos da cidade de São Paulo.

2.5.1 – INDELS do cromossomo 5 em populações mundiais

No estudo de populações mundiais nós genotipamos os 31 INDELS selecionados em um total de 1045 amostras de DNA provenientes do painel do CEPH-HGDP, que pertencem a 52 populações mundiais distribuídas em 7 grandes regiões geográficas: África, Oriente Médio, Ásia Central, Leste asiático, Oceania, Europa e Américas.

A análise de variância molecular (AMOVA) revelou que de fato a maior parte da variabilidade genética humana encontra-se dentro de populações (80,72%, tabela 7). Este achado está de acordo com diversos estudos (Barbujani *et al.*, 1997; Jorde *et al.*, 2000; Nei e Roychoudhury, 1974; Nei e Roychoudhury, 1993) e a explicação

para este padrão está no fato da espécie humana ser uma espécie demograficamente recente com a maioria da sua variação genética oriunda de uma pequena população precursora africana que colonizou recentemente o restante do mundo (Goldstein e Chikhi, 2002).

Se focarmos em cada região separadamente, mais de 94% da variabilidade genética é encontrada dentro das populações, com exceção das Américas e Oceania que apresentaram 9,73% e 13,53% e da variabilidade genética entre populações, respectivamente (Tabela 7). Este grau de variabilidade genética entre populações relativamente alto nas Américas está de acordo com o encontrado por outros trabalhos realizados com as mesmas amostras do painel do CEPH-HGDP utilizando microssatélites e outros INDELS (Rosenberg *et al.*, 2002; Bastos-Rodrigues *et al.*, 2006). Tendo em vista a história demográfica das Américas, este resultado não é inesperado. É bem conhecido que as populações nativas americanas (principalmente as sul-americanas) constituem pequenos grupos isolados, separados geograficamente. Esta característica ocasionou fortes efeitos de deriva genética nestas populações gerando grandes variações de frequências alélicas entre as populações nativas americanas (Cavalli-Sforza *et al.*, 1994).

Para a Oceania, atribuímos o alto grau de variabilidade genética encontrado entre populações a uma questão amostral. Apenas duas populações da Oceania são representadas no painel do CEPH-HGDP e estas estão localizadas em ilhas diferentes. Isto poderia estar gerando isolamento entre as duas populações acarretando em maiores níveis de diferenciação entre elas. Nos estudos realizados

com microssatélites (Rosenberg *et al.*, 2002) e INDELS espalhados pelo genoma (Bastos-Rodrigues *et al.*, 2006) com as mesmas populações do painel do CEPH-HGDP, a Oceania também apresentou níveis mais altos de variabilidade genética entre populações quando comparados com as outras regiões geográficas (6,4% e 6,1%, respectivamente), sendo mais baixos apenas do que os valores encontrados para as Américas.

Quando analisamos os resultados da AMOVA para as sete regiões geográficas, os níveis de variabilidade genética entre as grandes regiões encontrados com nossos INDELS apresentaram valores um pouco mais altos (16,4%) do que os 12,1% encontrados para os 40 INDELS espalhados pelo genoma (Bastos-Rodrigues *et al.*, 2006). Isto pode ser explicado pelo fato dos INDELS utilizados no presente estudo apresentarem grandes variações de frequências alélicas entre africanos e europeus. Estas grandes diferenças de frequências alélicas entre África e Europa podem gerar um aumento adicional no componente de variação genética entre regiões quando comparado com os dados de Bastos-Rodrigues, 2006, cujas frequências alélicas de todos os INDELS estavam próximas de 0,5 em todas as regiões geográficas estudadas (Bastos-Rodrigues *et al.*, 2006).

A análise de diferenciação entre indivíduos baseada em compartilhamento de alelos idênticos por estado (Figura 7) mostrou claro agrupamento entre indivíduos de uma mesma região geográfica (Figura 7a). No entanto, quando os indivíduos foram identificados de acordo com a população a qual pertencem, os mesmos se mostraram bastante misturados, não havendo formação de agrupamentos claros

(Figura 7b). Estes achados indicam que embora haja grande compartilhamento de alelos entre indivíduos de uma mesma região geográfica, a semelhança entre os indivíduos não ocorre predominantemente entre indivíduos de uma mesma população. Isto pode ser um indicativo de que grande parte da variabilidade genética ocorre entre indivíduos de uma mesma população, ou seja, dentro de populações, o que está de acordo com dados encontrados pela AMOVA.

Os resultados de escalonamento multidimensional (MDS) e análise dos componentes principais realizados com as sete regiões geográficas (figuras 5 e 15) foram bem condizentes com a história de migração do homem moderno da África para os outros continentes (Cavalli-Sforza e Feldman, 2003): a África aparece mais isolada dos demais continentes, Ásia Central, Europa e Oriente Médio bem próximos entre si, constituindo a Eurásia e Américas próximas do Leste asiático.

Nas análises utilizando as 52 populações mundiais, a resolução do MDS (Figura 6) foi maior que a do PCA (Figura 16). O MDS revelou 5 agrupamentos claros correspondentes à África, Oceania, Eurásia (Europa, Oriente Médio e Ásia Central), Leste asiático e Américas. Quando analisamos o agrupamento das populações nativas americanas é possível observar que as populações encontram-se mais separadas entre si do que as populações das outras regiões geográficas. Este dado está de acordo com os resultados da AMOVA e refletem, mais uma vez, a história demográfica das Américas caracterizada por um forte efeito de deriva genética.

Nossos resultados de MDS foram muito semelhantes aos encontrados por Bastos-Rodrigues, 2006 e indicam um alto poder discriminatório dos INDELS selecionados

no presente estudo. Além disso, a semelhança dos nossos resultados obtidos com a análise de INDELs localizados no braço longo do cromossomo 5, com os de Bastos-Rodrigues, 2006 que utilizou INDELs espalhados pelo genoma, pode ser um indicativo de que a região cromossômica estudada é representativa do genoma como um todo.

2.5.2 - INDELs do cromossomo 5 em populações brasileiras

Outro enfoque desta tese foi o estudo das raízes ancestrais da população brasileira. Neste trabalho, nossa proposta foi ampliar os estudos de caracterização da população brasileira realizados por nosso grupo de pesquisa. Nossa idéia foi investigar o padrão de mistura dos brasileiros não apenas em nível populacional, mas sim em nível individual e em uma resolução maior do que a oferecida por marcadores espalhados pelo genoma. Nossa proposta foi caracterizar o padrão de mistura da população brasileira em nível haplotípico (cromossômico), através da determinação da origem ancestral de fragmentos cromossômicos de uma região específica do cromossomo 5.

Para isso genotipamos nosso painel de 31 INDELs informativos de ancestralidade entre africanos e europeus, localizados no braço longo do cromossomo 5 em 104 indivíduos pretos e 157 indivíduos brancos da cidade de São Paulo e utilizamos as

amostras de africanos, europeus e ameríndios do painel do CEPH-HGDP já genotipadas neste estudo como referência de populações parentais.

As freqüências alélicas encontradas em africanos europeus e ameríndios estão de acordo com as encontradas por Weber (2002) (Tabela 5 *versus* Tabela 7). Este achado confirma o caráter diagnóstico entre africanos e europeus dos marcadores selecionados e valida a sua utilização em estudos de caracterização da população brasileira.

Os resultados do escalonamento multidimensional (MDS) e da análise dos componentes principais (PCA) realizados com as populações brasileiras e as regiões geográficas parentais foram bastante semelhantes: as três regiões parentais (África, Europa e América) bem separadas entre si, os brancos próximos dos europeus e os pretos entre africanos e europeus, no entanto, mais próximos dos africanos (Figuras 17 e 21). Estes resultados indicam que os marcadores foram capazes de separar, ao menos em nível qualitativo, as três populações parentais. É importante salientar que os brancos aparecem mais próximos dos Europeus do que os pretos dos africanos. Isto indica que os pretos brasileiros são mais miscigenados que os brancos.

Os resultados do MDS e PCA para quando as populações de cada região foram discriminadas não foram diferentes: nas duas análises é possível observar três agrupamentos claros constituídos das populações da África, Europa e Américas. Os brancos estão incluídos no grupo dos europeus enquanto que os pretos não se

incluem em nenhum agrupamento. Mas uma vez, os resultados indicam que os pretos são mais miscigenados que os brancos.

A capacidade dos marcadores selecionados de separar as três populações parentais foi confirmada pelos dados do *Structure*. Submetemos os genótipos das três populações parentais a análise de estruturação genética e sem fornecer ao programa nenhuma informação prévia a respeito da origem geográfica das amostras, o programa as separou em três grupos que corresponderam às três raízes ancestrais brasileiras: africanos, europeus e ameríndios (Figura 23).

Nossos dados revelaram que, em nível individual, os africanos têm em média 89,0% da região cromossômica estudada de ancestralidade africana, os europeus têm 74,3 % de ancestralidade européia e os ameríndios 73,3% de ancestralidade ameríndia. Estes dados, embora confirmem o poder discriminatório dos nossos marcadores entre as três populações parentais brasileiras, indicam uma maior facilidade dos marcadores em separar africanos de não-africanos. Isto pode ser explicado pela natureza dos marcadores selecionados. Os marcadores foram selecionados baseados em suas diferenças de frequências alélicas entre africanos e europeus. Desta forma, as diferenças de frequências alélicas entre europeus e ameríndios não foram levadas em consideração e podem ser menores, gerando uma maior dificuldade em separar estas duas populações.

Nossos dados são semelhantes aos encontrados por Pena (2009), que utilizou 40 INDELS espalhados pelo genoma nas mesmas amostras de africanos, europeus e

ameríndios do painel do CEPH-HGDP. No entanto, o poder discriminatório dos 40 INDELS foi maior do que o dos 31 utilizados no presente estudo. Este fato pode ser explicado por duas razões: primeiramente, o número de marcadores utilizados no primeiro estudo é maior, o que aumenta seu poder discriminatório; em segundo lugar, enquanto utilizamos neste estudo marcadores com grandes variações de freqüências alélicas entre africanos e europeus, Pena *et al*, 2009, utilizou marcadores com freqüências alélicas próximas de 0,5 em todas as populações. Em conjunto, as pequenas diferenças de freqüências alélicas dos 40 INDELS entre várias populações podem ser mais discriminatórias do que grandes diferenças de freqüências alélicas entre apenas duas populações.

2.5.2.1- Análise das proporções de mistura genética em amostras de brancos e pretos de São Paulo

Analizamos as proporções de mistura genética em pretos e brancos brasileiros, através de três métodos distintos: o programa *Admix 2.0*, que infere as proporções de mistura de populações miscigenadas a nível populacional, o programa *Structure*, que fornece as proporções de mistura individuais de cada amostra e a contagem do número de locos de determinada origem ancestral em cada população brasileira, inferidos também pelo *Structure*. Embora não haja grandes variações entre os resultados obtidos pelos três métodos utilizados (tabela 9), existe uma maior homogeneidade da inferência de ancestralidade africana em relação às ancestralidades européia e ameríndia. Para pretos e brancos as inferências de ancestralidade africana foram respectivamente 55,1% e 6,5% pelo

Admix 2.0, 53,3% e 16,0% pelo *Structure* e 63,2% e 11,4% pela contagem de locos. Já as inferências de ancestralidade europeia e ameríndia para pretos foram de respectivamente 34,0% e 10,9% pelo *Admix 2.0*, 27,2% e 19,5% pelo *Structure* e 23,0% e 13,8% pela contagem de locos. Para os brancos estes valores foram respectivamente 86,8% e 6,8% pelo *Admix 2.0*, 60,0% e 24,0% pelo *Structure* e 70,0% e 18,7% pela contagem de locos.

A maior homogeneidade na inferência de ancestralidade africana entre os três métodos sugere uma maior sensibilidade dos marcadores selecionados em distinguir entre ancestralidade africana de não-africana. Este fato pode ser explicado pela natureza dos marcadores selecionados, que apresentam uma alta variação de frequência alélica entre africanos e europeus. Desta forma, a distinção entre ancestralidade europeia e ameríndia parece ser mais difícil, gerando divergências entre os diferentes métodos de inferência de mistura utilizados.

Tabla 9 - Comparação das proporções de mistura africana, europeia e ameríndia encontradas para pretos e brancos de São Paulo pelos diferentes métodos utilizados.

Métodos	Populações	Ancestralidade (%)		
		Africana	Europeia	Ameríndia
<i>Admix 2.0</i> (nível populacional)	Pretos SP	55,1±1,9	34,0±2,4	10,9±3,3
	Branco SP	6,5±1,6	86,8±2,3	6,8±3,1
<i>Structure</i> (nível individual)	Pretos SP	53,3 ±23,8	27,2±16,6	19,5±13,5
	Branco SP	16,0±9,9	60,0±14,8	24,0±13,5
Contagem de locos (nível populacional)	Pretos SP	63,2	23,0	13,8
	Branco SP	11,4	70,0	18,7

Em nível individual, as proporções médias de ancestralidade africana, europeia e ameríndia inferidas pelo *Structure* foram respectivamente de 53,3%, 27,2% e 19,5% para os pretos e de 16,0%, 60,0% e 24,0% para os brancos de São Paulo. Comparando estes resultados com o encontrado por Pena (2009), observamos pequenas variações nas porcentagens de ancestralidade europeia e ameríndia principalmente para os brancos brasileiros. Pena (2009) encontrou que cada indivíduo preto brasileiro tem em média 50,0% de ancestralidade africana, 37,0% de ancestralidade europeia e 13% de ancestralidade ameríndia. Para os brancos brasileiros os valores encontrados por Pena (2009) foram respectivamente 12,0%, 78,0% e 10,0%. As porcentagens individuais médias de ancestralidade africana encontradas em nosso estudo estão de acordo com as encontradas por Pena (2009) (53% *versus* 50% para os pretos e 16,0% *versus* 12,0% para os brancos). As variações entre os dois estudos estão nas porcentagens de ancestralidade europeia e ameríndia. Podemos atribuir este fato, novamente, a maior dificuldade dos nossos marcadores em diferenciar europeus de ameríndios quando comparados com os marcadores utilizados por Pena (2009). Isto pode levar a uma confusão na determinação das porcentagens de ancestralidade que criam divergências entre os diferentes estudos.

Os resultados de proporção de ancestralidade de pretos e brancos brasileiros tanto em nível populacional (inferidos pelo *Admix 2.0*) quanto em nível individual (inferidos pelo *Structure*) estão de acordo com os dados de MDS e PCA e confirmam, agora de forma quantitativa, o maior grau de miscigenação dos pretos quando comparados com os brancos.

Este fenômeno de diferentes níveis de miscigenação entre brancos e pretos brasileiros pode ser explicado por fatos históricos. Os portugueses chegaram ao Brasil em 1500 e a mistura entre portugueses e ameríndios começou em seguida à chegada dos primeiros colonizadores (Vainfas, 2000). Desta forma, os ameríndios contribuíram maciçamente para a formação inicial da população brasileira. Posteriormente, devido a conflitos com os europeus e doenças as quais os ameríndios não suportavam, as tribos passaram por um drástico declínio demográfico (Salzano e Freire-Maia, 1970). A partir da segunda metade do século XVI, os africanos começaram a ser trazidos para o Brasil para trabalhar como escravos constituindo, portanto, a terceira raiz ancestral da população brasileira. Depois de 1850 (após a abolição do tráfico de escravos) houve uma política do governo para estimular a vinda de europeus para o Brasil. Desta maneira, cerca de 5,5 milhões de imigrantes europeus chegaram ao Brasil e este fenômeno foi conhecido como “branqueamento do Brasil” (Seyferth, 1985; Santos, 2002). A idéia do governo era aumentar o número de casamentos entre europeus e pretos brasileiros para que a população tivesse a pele mais clara no decorrer das gerações. Desta forma, com a maciça chegada de europeus no Brasil na segunda metade do século XIX, houve um grande número de casamentos entre europeus e pretos brasileiros que explicam os maiores graus de mistura dos pretos brasileiros quando comparados com os brancos.

2.5.2.2- Análise da origem ancestral de fragmentos cromossômicos nas populações parentais e brasileiras

Para inferir a origem ancestral dos fragmentos cromossômicos de cada cromossomo de cada indivíduo de cada população (brasileiras e parentais), inferimos, antes de tudo, a fase gamética dos dados. Para isso utilizamos o programa *Phase*. O grau de certeza da inferência dos haplótipos para todas as populações foi baixo (0,04 para africanos, 0,12 para europeus, 0,21 para ameríndios, 0,06 para brancos brasileiros e 0,08 para pretos brasileiros). Esta incerteza se dá pela grande distância entre os marcadores. Um reflexo disto está na análise do desequilíbrio de ligação nas populações estudadas. Não foi observado desequilíbrio de ligação significativo nas populações analisadas, o que provavelmente dificultou a inferência da fase gamética. Embora o grau de incerteza da inferência dos haplótipos tenha sido alto, o grau de incerteza de inferência da fase por sítios ambíguos foi relativamente baixo: certeza média de 0,72 para africanos, 0,80 para europeus, 0,83 para ameríndios, 0,81 para brancos e pretos brasileiros. Desta maneira, consideramos válida a utilização dos haplótipos inferidos pelo *Phase* na análise da origem ancestral de fragmentos cromossômicos nas populações brasileiras.

Após a inferência de haplótipos das populações parentais e brasileiras, a ancestralidade de cada loco de cada cromossomo foi determinada para cada indivíduo através do modelo de ligação do programa *Structure*. Os resultados mostraram que a grande maioria dos locos analisados em todos os indivíduos

deriva de uma das três populações parentais, já que houve um pequeno número de locos com origem ancestral ambígua (Figuras 24 a 28).

Conforme o esperado, nas populações parentais os segmentos cromossômicos originaram-se predominantemente de uma única população, embora fragmentos cromossômicos de outra ancestralidade também tenham sido observados (Figuras 24, 25 e 26). Este achado é um bom indicativo de que os nossos marcadores foram capazes de distinguir as três origens ancestrais da população brasileira. É possível observar uma maior homogeneidade na origem ancestral dos fragmentos cromossômicos dos africanos (Figura 24) quando comparados com europeus e ameríndios. Novamente, esta aparente mistura entre europeus e ameríndios pode ser explicada pela natureza dos marcadores utilizados, que são informativos de ancestralidade específicos para África e Europa e podem ser menos informativos na diferenciação entre europeus e ameríndios.

Para as populações brasileiras, foi possível identificar a origem ancestral da grande maioria dos locos e, em contraste com os resultados encontrados para as populações parentais, as populações brasileiras apresentaram fragmentos cromossômicos originados das três populações ancestrais (Figuras 27 e 28). Como esperado, nos brancos de São Paulo foi encontrado maior número de fragmentos cromossômicos de origem européia do que africana e ameríndia. Já para os pretos de São Paulo, é possível observar muitos fragmentos cromossômicos originados da África, Europa e, um pouco menos, da América. Isto indica um maior grau de

miscigenação dos pretos quando comparado com os brancos de São Paulo e está de acordo com as porcentagens de ancestralidade encontradas em nível populacional (inferidas pelo *Admix 2.0*) e individual (inferidas pelo *Structure*).

A análise da ancestralidade dos locos ao longo da região cromossômica estudada indica que, de maneira geral, os cromossomos dos brasileiros são constituídos de fragmentos que se originaram da África ou da Europa ou da América, reforçando a natureza tri-híbrida da população brasileira. Além disso, foi possível observar que os cromossomos dos brasileiros são verdadeiros mosaicos constituídos de fragmentos de diferentes origens. Isto indica que a miscigenação ocorre não apenas em nível individual, mas também em nível cromossômico e que nossos cromossomos são constituídos de fragmentos com diferentes genealogias.

A existência de fragmentos cromossômicos com diferentes origens ancestrais nos brasileiros sugere que esta população, principalmente os pretos que apresentam um nível maior de miscigenação, pode constituir uma ferramenta útil em estudos de mapeamento por miscigenação. Estes estudos procuram identificar genes que contribuem para doenças complexas que apresentam uma grande diferença de incidência em as populações parentais. Desta maneira, a ancestralidade de fragmentos cromossômicos de indivíduos casos e controles pertencentes à população miscigenada é determinada e uma associação é feita entre a doença e as regiões cromossômicas com diferentes origens ancestrais entre os grupos caso e controle (Darvasi and Shifman, 2005). Portanto, se os cromossomos dos brasileiros

são formados por fragmentos de origem ancestral africana, europeia e ameríndia, esta população pode ser utilizada para mapear, ou seja, determinar a localização cromossômica de genes candidatos a doenças complexas se estas apresentarem grandes diferenças de incidência entre africanos, europeus e ameríndios.

Após a determinação da origem ancestral de cada loco para cada cromossomo de cada indivíduo brasileiro, investigamos um possível padrão de casamentos entre os indivíduos brasileiros através da análise de correlação entre o número de locos de uma mesma origem ancestral entre os dois cromossomos 5 homólogos de um mesmo indivíduo. A correlação foi significativa tanto para os pretos ($p < 0,000001$) quanto para os brancos ($p < 0,000001$). Já que os dois cromossomos de um mesmo indivíduo são herdados cada um de um de seus pais, uma correlação significativa entre a ancestralidade dos dois cromossomos de um mesmo indivíduo é um indicativo de casamento preferencial. Desta forma, nossos resultados indicam um provável casamento assortativo entre os brasileiros, mas não de acordo com a ancestralidade genômica, e sim de acordo com a cor da pele que estaria se refletindo na ancestralidade genômica.

Juntos, nossos resultados mostraram que os marcadores selecionados no presente estudo foram capazes de distinguir as três raízes ancestrais da população brasileira e são, portanto adequados para estudos de miscigenação desta população. Além disso, nossos dados sugerem que a estratégia experimental adotada para a determinação da origem ancestral de fragmentos cromossômicos

em brasileiros foi eficiente e permitiu a caracterização do padrão de miscigenação da população brasileira em nível cromossômico. A estratégia empregada no presente estudo também nos permitiu fazer inferências acerca do padrão de casamentos entre os brasileiros e indicou que a população brasileira é uma potencial ferramenta para estudos de mapeamento por miscigenação.

2.6 – Conclusões

A partir da análise dos 31 INDELS em populações mundiais do painel do CEPH-HGDP e populações de pretos e brancos brasileiros, podemos concluir que:

- Os INDELS selecionados no presente estudo apresentaram um alto poder discriminatório entre as 5 grandes regiões geográficas mundiais (África, Eurásia, Leste asiático, Europa e Américas) com mais de 16% da variabilidade genética total ocorrendo entre os continentes;
- Os marcadores selecionados foram capazes de distinguir entre as três populações parentais da população brasileira e são, portanto, adequados para estudo de miscigenação desta população;
- Os pretos brasileiros se mostraram mais miscigenados do que os brancos tanto em nível populacional quanto em nível individual;
- A abordagem empregada nos permitiu determinar a origem ancestral e os pontos de mudança de ancestralidade de segmentos cromossômicos nas populações brasileiras;
- A análise da ancestralidade dos locos ao longo da região cromossômica estudada indica que, de maneira geral, os locos dos brasileiros em cada cromossomo se

originam ou da África ou da Europa ou da América, reforçando a natureza tri-híbrida da população brasileira;

- Os cromossomos dos brasileiros são verdadeiros mosaicos constituídos de fragmentos de diferentes origens. Isto indica que a miscigenação ocorre não apenas em nível individual, mas também em nível cromossômico;

- Cromossomos constituídos de fragmentos de diferentes origens ancestrais indicam que os cromossomos são constituídos de fragmentos com diferentes genealogias;

- A correlação positiva encontrada entre o número de locos de uma mesma origem ancestral entre os dois cromossomos de um mesmo indivíduo na população brasileira sugere a ocorrência de casamentos preferenciais quanto a cor da pele;

3 – Capítulo 2

Estudo dos microssatélites tetralocais do cromossomo Y, DYS464 e DYS503 em populações mundiais e miscigenadas

3.1 – Introdução

3.1.1 – O cromossomo Y humano

O cromossomo Y humano reveste-se de características peculiares que fazem dele uma ferramenta muito útil em estudos forenses e evolutivos. São haplóides, paternalmente herdados e não sofrem recombinação na maior parte de sua extensão (com exceção das regiões pseudo-autossômicas) (Jobling e Tyler-Smith, 2003). Portanto, cada indivíduo do sexo masculino tem seu cromossomo Y idêntico ao de seu pai, irmãos, avô paterno, primos paternos e etc., estabelecendo assim uma patrilinearidade.

Desta forma, os marcadores moleculares polimórficos localizados no cromossomo Y são bastante úteis em estudos forenses por permitirem especificidade na identificação de DNA masculino em *pools* de DNA (Gusmão *et al.*, 2006) e possibilitarem testar paternidade de crianças do sexo masculino (Santos *et al.*, 2003). Além disso, os marcadores ligados ao cromossomo Y também permitem traçar as patrilinearidades e, a partir delas, relacionar indivíduos do sexo masculino separados por longos períodos de tempo sendo, portanto, úteis também em estudos de ancestralidade (Carvalho-Silva *et al.*, 2001, 2006) e estudos evolucionários (Chiaroni *et al.*, 2009).

Diante disto, existe um grande interesse na identificação de marcadores moleculares polimórficos do cromossomo Y e na caracterização destes marcadores em diferentes populações mundiais. Em 1992, Roewer e colaboradores descreveram o primeiro marcador polimórfico do cromossomo Y, o microssatélite *DYS19* (Roewer e Epplen, 1992). Nos 10 anos seguintes a descoberta de novos microssatélites do cromossomo Y progrediu de forma bastante lenta. Em 1997, a comunidade forense europeia estabeleceu que um grupo de nove microssatélites do cromossomo Y (compreendendo o *DYS19*, *DYS389I/II*, *DYS390*, *DYS391*, *DYS392*, *DYS393* e *DYS385a/b*) que constitui o chamado “haplótipo mínimo”, capaz de distinguir 74-90% dos indivíduos de várias localizações da Europa (de Knijff *et al.*, 1997, Kayser *et al.*, 1997). Desde então, este grupo de marcadores vem sendo amplamente utilizado em estudos forense e evolutivos com o cromossomo Y (Roewer *et al.*, 2001).

No entanto, a busca por novos marcadores do cromossomo Y continuou e com os avanços na determinação do mapa físico do deste cromossomo, o número de microssatélites descritos aumentou substancialmente. Atualmente, mais de 300 microssatélites localizados no cromossomo Y foram descritos (Hanson e Ballantyne, 2006) e entre estes, o microssatélite tetralocal *DYS464* se mostrou ser o mais polimórfico de todos (Redd *et al.*, 2002; Berger *et al.*, 2003).

3.1.2 - O conjunto de microssatélites *DYS464*

O *DYS464* foi primeiramente descrito por Redd e colaboradores em 2002 e consiste em quatro cópias idênticas (*DYS464* a, b, c e d) do motivo de repetição de tetranucleotídeos (CCTT)_n localizados na região AZFc (Kuroda-Kawaguchi *et al.*, 2001), na banda Yq11.223 do cromossomo Y (Figura 21).

A região AZFc é uma região altamente repetitiva e palindrômica localizada no braço longo do cromossomo Y (Skaletsky *et al.* 2003; Kuroda-Kawaguchi *et al.* 2001). Esta estrutura particular da região AZFc a torna bastante susceptível a rearranjos internos gerados por recombinação mitótica homóloga entre repetições específicas que pode levar à duplicações de diversos locos. Portanto, devido a sua localização na região AZFc, acredita-se que a estrutura tetralocal do *DYS464* tenha surgido a partir de duas duplicações sequenciais de um loco originalmente único, geradas por recombinação mitótica (Kuroda-Kawaguchi *et al.* 2001; Skaletsky *et al.*, 2003; Rozen *et al.*, 2003)

Além de ser considerado o microssatélite mais polimórfico de todos já descritos, a estrutura tetralocal do *DYS464* permite que com apenas um par de iniciadores os quatro sítios sejam amplificados. Assim, é possível obter informação relativa a 4 locos em uma única reação de PCR.

Embora o *DYS464* tenha sido estudado em algumas populações humanas (Redd *et al.*, 2002; Berger *et al.*, 2003; Butler e Schoske, 2005), um estudo mais completo deste marcador em populações mundiais nunca foi realizado. Portanto, diante do caráter altamente polimórfico do *DYS464* e da necessidade de se caracterizar este marcador em populações mundiais, nós genotipamos o *DYS464* em todas as amostras de homens do painel do CEPH-HGDP (Cann *et al.*, 2002) e estudamos sua diversidade e filogeografia em diferentes populações humanas.

3.1.3 – O conjunto de microssatélites *DYS503*

Em 2006, durante a anotação do mapa físico de microssatélites do cromossomo Y, Hanson e Ballantyne identificaram um novo microssatélite multi-loco, chamado *DYS503* (Hanson e Ballantyne, 2006). Assim como o *DYS464*, o *DYS503* constitui um conjunto de microssatélites também presente em 4 cópias no braço longo do cromossomo Y e apresenta o motivo de repetição (TTA)_n (Figura 29).

Com o objetivo de verificar se o *DYS503* também apresentava uma alta diversidade como o *DYS464*, Hanson e Ballantyne estenderam seus estudos e analisaram o *DYS503* em afro-americanos e caucasianos (Hanson e Ballantyne, 2007). Curiosamente, os resultados revelaram uma maior diversidade em caucasianos (0,92) do que em afro-americanos (0,75). Além disso, a diversidade do *DYS503* foi a maior já encontrada para caucasianos quando comparada com qualquer outro

microssatélite de cópia única do cromossomo Y estudado (Hanson e Ballantyne, 2007).

Este dado encontrado por Hanson e Ballantyne nos sugere que a diversidade na principal população parental dos afro-americanos, ou seja, na África, seja menor que a diversidade na principal população parental dos caucasianos, a Europa. Este dado seria incomum tendo em vista a origem da humanidade no continente africano e nos despertou curiosidade.

Desta forma, na tentativa de caracterizar e entender o padrão de diversidade do *DYS503* em populações mundiais, genotipamos o *DYS503* em todas as amostras de homens do painel do CEPH-HGDP e estudamos sua diversidade em diferentes populações humanas.

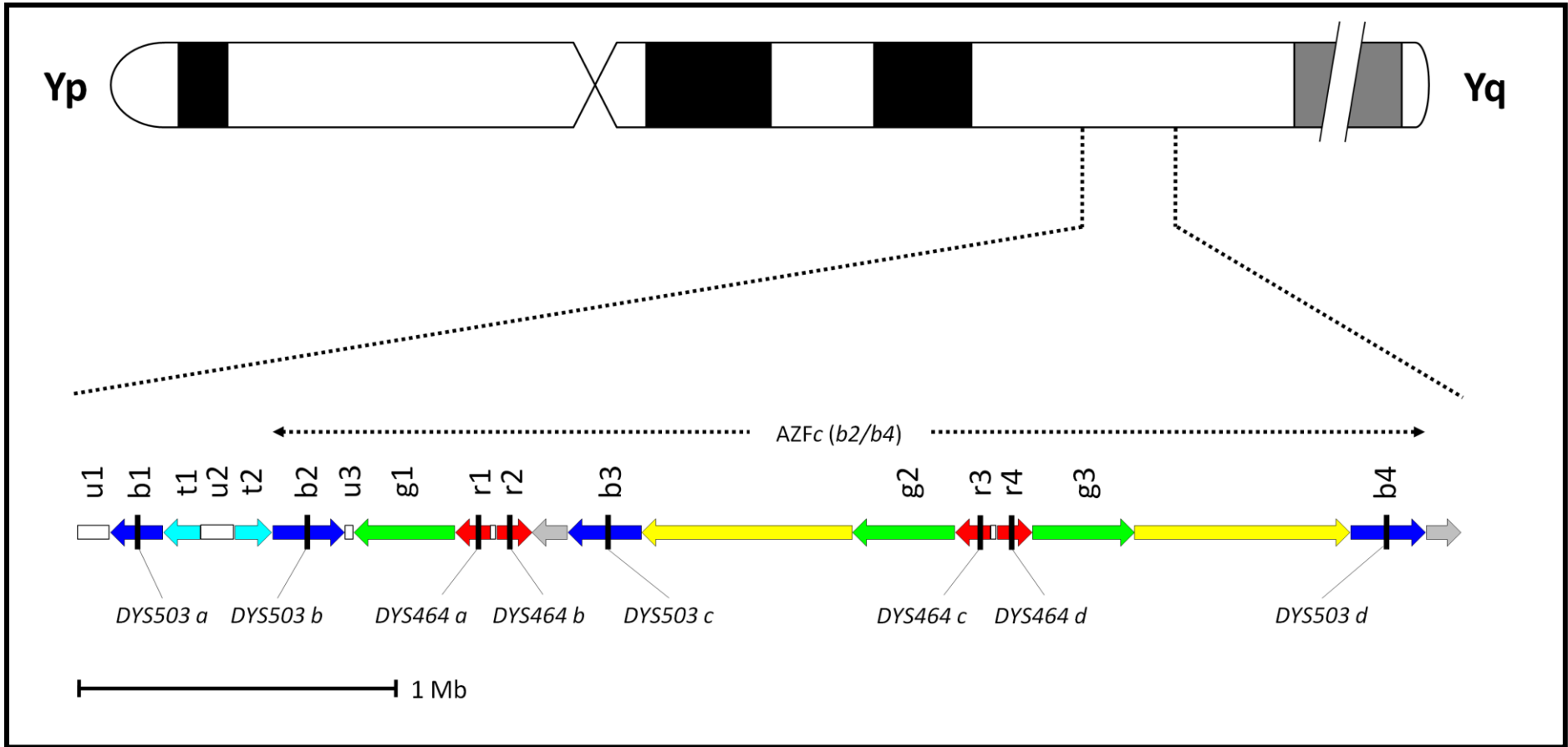


Figura 29 – Mapa do cromossomo Y com a localização dos microssatélites DYS464 e DYS503 em relação à região AZFc.

3.2 – Objetivos

3.2.1 – Objetivo geral

Estudar amplamente os microssatélites tetralocais do cromossomo Y, *DYS464* e *DYS503* em todas as amostras masculinas do painel do CEPH-HGDP e inferir a respeito de sua diversidade e filogeografia em diferentes populações humanas. Além disso, verificar a possível utilização das frequências alélicas do microssatélite *DYS464* no estudo de ancestralidade paterna em populações miscigenadas, através da análise deste marcador nas populações de pretos e brancos brasileiros, afro-americanos e caucasianos.

3.2.2 – Objetivos específicos

- Amplificar e genotipar o conjunto de microssatélites do cromossomo Y *DYS464* e *DYS503* para amostras de DNA de homens de 52 populações mundiais do painel do CEPH-HGDP distribuídas em 7 grandes regiões geográficas;
- Calcular as frequências alélicas do microssatélite *DYS464* em todas as populações e regiões geográficas estudadas;
- Determinar, para o microssatélite *DYS464*, diversidade genotípica, frequências genotípica e compartilhamento de genótipos entre as populações e regiões geográficas estudadas;
- Verificar estruturação populacional através da análise de variância molecular (AMOVA), do escalonamento multidimensional (MDS) com diferentes medidas de distância e da análise dos componentes principais (PCA) com dados de frequências genotípicas e alélicas do *DYS464*;
- Calcular, através das frequências alélicas do *DYS464*, as proporções de contribuição africana para as populações de pretos e brancos brasileiros, afro-americanos e caucasianos e comparar os valores obtidos com dados já conhecidos de ancestralidade paterna para estas populações;

- Determinar frequências e índices de diversidade de padrões alélicos para o microssatélite *DYS503* nas sete regiões geográficas mundiais estudadas;

3.3 – Material e Métodos

3.3.1 – Amostras de DNA

Amostras de DNA de todos os 677 homens não-aparentados, pertencentes às 52 populações mundiais do Painel do CEPH-HGDP, que estão distribuídas em 7 regiões geográficas (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América) do painel do CEPH-HGDP (<http://www.cephb.fr/>) (Figura 1) foram analisadas.

Além das populações mundiais do painel do CEPH-HGDP, também analisamos amostras de DNA de 124 indivíduos pretos (genotipadas anteriormente por Cláudia Carvalho) e 106 indivíduos brancos da cidade de São Paulo. Estas amostras foram cedidas pelo Dr. Sérgio Bydlowski (Departamento de Hematologia, FMUSP) e os indivíduos foram autotipificados como pretos ou brancos (denominação usada pelo IBGE), incluindo os seus pais e avós. O estudo foi anônimo e todos os indivíduos assinaram um termo de consentimento para o uso de sua amostra de DNA em trabalhos científicos. A utilização destas amostras para estudos científicos foi aprovada pela Comissão de Ética para Análise de Projetos de Pesquisa (CAPPesq) da Diretoria Clínica do Hospital das Clínicas da Faculdade de São Paulo em protocolo de pesquisa número 543/04.

3.3.2 - Amplificação do microssatélite *DYS464*

Para cada amostra de DNA, a amplificação das quatro cópias do microssatélite *DYS464* foi feita em uma única reação de PCR em que foram utilizados três iniciadores: direto (com cauda M13-40), reverso e M13-40 (5'GTTTTCCCAGTCACGAC3') marcado com fluoróforo FAM (Tabela 10).

Para a reação de PCR, 1,5µl (~ 4ng) de DNA molde foram amplificados em um volume final de 12,5µl de reação. Cada reação continha 200µM de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,01uM do iniciador direto, 0,1µM do iniciador reverse e 0,1µM do iniciador M13-40 marcado, 2 unidades de *Taq DNA polimerase* (Phoneutria) em tampão de reação com 10mM de Tris, 50mM KCl, 1,5mM MgCl₂, 0,1% de triton X-100, pH8,4. O ciclo de reação consistiu em uma desnaturação inicial a 95°C por 3 minutos, 5 ciclos a 95°C por 45 segundos, 58°C por 45 segundos e 72°C por 1 minuto e 30 ciclos a 95°C por 45 segundos, 55°C por 45 segundos e 72°C por 1 minuto. A extensão final foi a 72°C por 5 minutos. Para todas as reações de PCR foi feito um controle branco de amplificação, onde água estéril foi acrescentada no lugar do DNA.

3.3.3 – Amplificação do microssatélite *DYS503*

A amplificação do conjunto de microssatélites *DYS503* foi feita apenas para as populações mundiais do painel do CEPH-HGDP. Assim como a amplificação do microssatélite *DYS464*, a amplificação das quatro cópias do *DYS503* foi feita também em uma única reação de PCR, utilizando-se três iniciadores: direto (com cauda M13-40), reverso e M13-40 (5'GTTTTCCCAGTCACGAC3') marcado com fluorescência FAM (Tabela 10).

Para a reação de PCR, 2,0µl (~7,0ng) de DNA molde foram amplificados em um volume final de 23µl de reação. Cada reação continha 1mM de cada um dos dNTPs (dATP, dCTP, dGTP, dTTP), 0,096µM do iniciador direto, 0,96µM do iniciador reverso e 0,96µM do iniciador M13-40 marcado, 3 unidades de *Taq DNA polimerase* (Phoentria) em tampão de reação com 10mM de Tris, 25mM KCl, 3,5mM MgCl₂, pH8.3. O ciclo de reação consistiu em uma desnaturação inicial a 95°C por 11 minutos, 32 ciclos a 96°C por 30 segundos, 60°C por 1 minuto e 72°C por 1 minuto. A extensão final foi a 60°C por 30 minutos. A *Taq DNA polimerase* foi adicionada à solução após 1 minuto de início da desnaturação, em um protocolo conhecido como *Hot Start*. Para todas as reações de PCR foi feito um controle branco de amplificação, onde água estéril foi acrescentada no lugar do DNA.

Tabela 10: Seqüência de iniciadores e características dos microssatélites do cromossomo Y

Marcador do cromossomo Y	Iniciador direto	Iniciador Reverso	Repetição	Número de repetições	Tamanho do amplicon*
<i>DYS464</i>	5'- GTTTTCCCAGTCACGACTTACGAGCTTTGGGCTATG -3'	5'- CCTGGGTAACAGAGAGACTCTT -3'	[CCTT] _n	9-23	263-319
<i>DYS503</i>	5'- GTTTTCCCAGTCACGACTTTCAGCCTTTTCAATTTCCA -3'	5'- GACTGCTCAACTGCACTCCA -3'	[TTA] _n	9-18	200-227

* tamanho em pares de base (pb)

3.3.4 – Genotipagem dos microssatélites *DYS464* e *DYS503*

A genotipagem de cada conjuntos de microssatélites do cromossomo Y (*DYS464* e *DYS503*) foi realizada separadamente através do sequenciador MegaBACE 1000 (Molecular Dynamics). Cerca de 2 μ l dos produtos de PCR foram misturados com solução contendo 0,25 μ l de ET-550-R *Size Standard* (GE Healthcare) e 7,75 μ l de Tween 0,1%. As amostras foram desnaturadas a 95°C por 3 minutos e injetadas no sequenciador a 3kV com uma voltagem de corrida de 10kV. O tempo de injeção foi de 80 segundos e 75 minutos de corrida. A análise do tamanho dos alelos para o *DYS464* e *DYS503* foi realizada utilizando os programas *Genetic Profiler* (versão 2.2) e *Fragment Profiler* (versão 1.2), ambos da GE Healthcare.

3.3.5 – Análise dos dados

3.3.5.1 – Microssatélite *DYS464*

3.3.5.1.1 – Parâmetros básicos de genética de populações

O conjunto de microssatélites *DYS464* está presente em 4 cópias no braço longo do cromossomo Y, de forma que o genótipo de cada indivíduo é constituído de 1 até 4 alelos diferentes. Como os programas de análise de dados de genética de populações humanas são configurados para lidar com dados haplóides e/ou

diplóides, estes seriam incapazes de calcular frequências alélicas para o loco *DYS464*. Como uma estratégia alternativa para o cálculo das frequências alélicas para o loco *DYS464*, empregamos o programa *Autotet* (Thrall e Young, 2000) em todas as populações estudadas. Este programa utiliza um modelo desenvolvido para análise de dados em populações autotetraplóides.

Os índices de diversidade como: diversidade genotípica, frequências genotípica e compartilhamento de genótipos entre populações, foram calculados tanto para populações mundiais quanto para populações miscigenadas através do pacote de programas de análise de dados para genética de populações, *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000).

3.3.5.1.2 – Análise de estruturação populacional

3.3.5.1.2.1 – Análise de Variância Molecular (AMOVA)

Para verificar estruturação genética entre as populações mundiais do painel do CEPH-HGDP, submetemos os dados do *DYS464* a análise de variância molecular (AMOVA) (Excoffier *et al.*, 1992) através do emprego do programa *Arlequin* (versão 3.5.1) (Schneider *et al.*, 2000). A AMOVA foi realizada com 10000 permutações.

3.3.5.1.2.2 – Escalonamento Multidimensional (MDS)

No estudo de populações mundiais e miscigenadas, matrizes de distância baseadas em F_{st} (distância de Reynolds), R_{st} e $\delta\mu^2$ foram calculadas, através do programa *Arlequin* (versão 3.5.1). As matrizes foram calculadas para as 52 populações do painel do CEPH-HGDP, para as 7 regiões geográficas e para as populações miscigenadas. Posteriormente, as matrizes foram graficamente representadas através do escalonamento multidimensional (MDS), realizado a partir do programa *Statistica* (versão 7.0).

3.3.5.1.2.3 – Análise do Componente principal (PCA)

A análise do componente principal (PCA) foi empregada no estudo de populações mundiais e miscigenadas. Para a análise das populações mundiais foram utilizadas as frequências alélicas e genotípicas. O estudo foi realizado tanto para as 52 populações quanto para as 7 regiões geográficas. Na análise das populações miscigenadas, utilizamos apenas frequências alélicas.

3.3.5.1.2.4 – Cálculo das proporções de mistura nas populações miscigenadas

A partir das freqüências alélicas obtidas pelo programa *Autotet*, calculamos as proporções de mistura africana das populações de pretos e brancos brasileiros assim como de caucasianos e afro-americanos (Butler e Schoske, 2005). Para isto utilizamos o método gráfico, baseado em mínimos quadrados, descrito por Collins-Schramm e colaboradores. (2002).

3.3.5.2 – Microssatélite *DYS503*

Para o microssatélite *DYS503* calculamos as freqüências e diversidades de padrões alélicos manualmente através, respectivamente, da contagem de padrões encontrados em cada região e do emprego da fórmula $(n/n-1).(1-\sum p^2)$, onde p corresponde a freqüência de cada padrão alélico.

3.4 – Resultados

3.4.1 – Microssatélite *DYS464*

A utilização de marcadores moleculares do cromossomo Y se tornou uma poderosa ferramenta em estudos de evolução humana, especialmente por ter herança uniparental, ser haplóide e não sofrer recombinação na maior parte de sua extensão. Sendo assim, a busca por marcadores mais informativos vem ganhando grande enfoque pelo fato destes distinguirem linhagens de cromossomo Y mais eficientemente que um grupo maior de locos menos informativos. O loco *DYS464* é um conjunto de microssatélites, com motivo de repetição (CCTT)_n, presente em 4 cópias no braço longo do cromossomo Y (*DYS464a, b, c e d*). No presente estudo, analisamos o microssatélite *DYS464* em populações mundiais e miscigenadas. Para análise das populações mundiais, genotipamos o *DYS464* em 677 homens provenientes do Painel do CEPH-HGDP, abrangendo 52 populações distribuídas em 7 regiões geográficas (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América). Para o estudo das populações miscigenadas, genotipamos 106 indivíduos brancos da cidade de São Paulo. Os dados referentes à genotipagem dos indivíduos pretos brasileiros foram aproveitados de um estudo prévio realizado anteriormente por nosso grupo (dados não publicados).

Uma particularidade da genotipagem do loco *DYS464* consiste no fato de que as regiões flanqueadoras das quatro cópias são idênticas. Desta forma, com apenas um par de iniciadores e apenas uma reação de PCR, foi possível amplificar as quatro cópias do *DYS464* de um indivíduo. Os perfis dos eletroferogramas gerados pelo seqüenciador consistiram em um, dois, três ou quatro picos. Estes picos foram então convertidos em genótipos baseados no número de repetições (tamanho dos *amplicons*) e na altura/área relativa dos quatro alelos presentes em cada amostra.

A análise individual dos eletroferogramas revelou que as alturas dos picos eram consistentes com quatro amplificações independentes. Assim, foi possível realizar uma leitura em alta resolução onde a altura dos picos foi um indicador quantitativo do número de cópias dos alelos. Desta forma, para cada indivíduo, foi possível identificar os alelos e estabelecer o número de cópias de cada um deles, ou seja, foi possível estimar seus genótipos. (Figura 30).

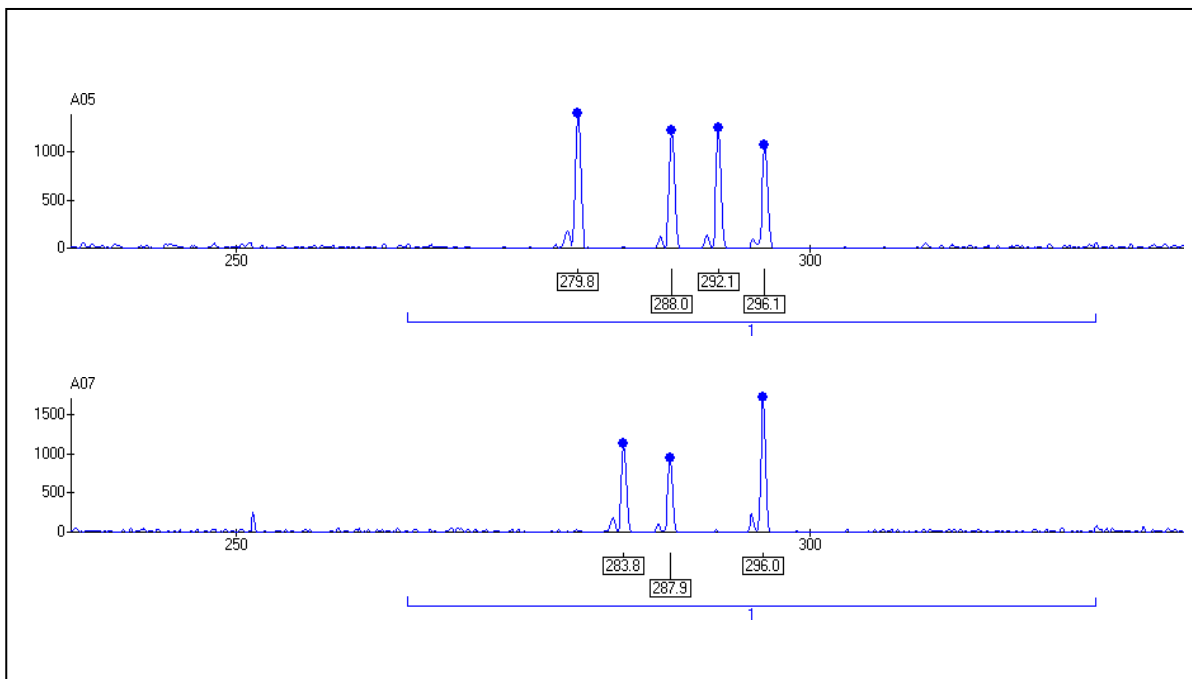


Figura 30 - Eletroferograma produzido pelo programa *Fragment Profile*, versão 1.2, representando a genotipagem do conjunto de microssatélites *DYS464* pelo sequenciador MEGA BACE 1000. A leitura dos gráficos foi realizada em alta resolução, onde a altura dos picos é um indicador quantitativo do número de cópias do alelo. Na figura A05 apresenta genótipo 13,15,16,17 e A07 apresenta o genótipo 14,15,17,17.

3.4.1.2 – Microssatélite *DYS464* em populações mundiais

3.4.1.2.1 – Índices de diversidade

Após a genotipagem do *DYS464* iniciamos as análises de frequências alélicas e dos índices de diversidade utilizando os programas *Autotet* e *Arlequin*. Em nossa análise, encontramos 14 diferentes alelos com tamanhos variando de 9 a 23 repetições, no entanto, não foi encontrado o alelo de 21 repetições em nenhuma amostra analisada.

Dentre os 14 diferentes alelos identificados, 13 (93%) foram encontrados na África, sendo 4 exclusivos deste continente (alelos 10, 20, 22 e 23). No Oriente Médio foram encontrados 8 (57%), na Ásia Central 9 (64%) e no Leste Asiático 10 (71,4%), sendo 1 (alelo 9) exclusivo desta última região geográfica. Já na Oceania, Europa e Américas foram encontrados 7 (50%), 9 (64%) e 7 (50%) alelos, respectivamente (Tabelas 11 e 12). O alelo mais frequente na África e Oceania foi o 16, no Oriente Médio foram o 14 e o 16 e na Ásia Central, Leste Asiático, Europa e América o alelo 15 (Tabela 11).

Em nossas análises identificamos um total de 175 genótipos, dos quais 49 foram encontrados na África, 53 na Europa, 29 no Oriente Médio, 72 na Ásia Central, 83 no Leste Asiático, 13 na Oceania, e 16 na América (Tabela 12).

Dos 175 genótipos, 90 foram observados em um único continente, sendo a África a região geográfica com maior porcentagem de genótipos exclusivos (20 em um total de 49 genótipos, 40%), seguida do Leste Asiático (34,9%), Ásia Central (27,7%), América (25%), Oriente Médio (24%), Oceania (23%) e Europa (15%) (Tabela 12).

Os genótipos mais frequentes encontrados em cada continente e populações foram os seguintes (Anexo 1):

- Na África, o genótipo 13-16-16-18 (20,4%). Este genótipo foi encontrado espalhado por toda a África Sub saariana, com significativas frequências entre os Biaka (43,3%), os Bantu (18,2%), os Yoruba (15,4%), os Mandenka (12,5%) e os MButi (7,7%). No entanto, este genótipo não foi observado nos San e nem fora da África.
- Na Europa, o genótipo 15-15-17-17 (8,0%). Este genótipo foi encontrado entre os Basques da França (31,3%), os Orcadian (14,3%) e os Bergamos da Itália (12,5%). O genótipo não foi observado nos French, Sardinian, Tuscan, Adygei do Caucasus, Russians e nem em outro continente que não a Europa.
- No Leste asiático, o genótipo mais frequente foi o 14-14-15-15 (11,6%). Este genótipo foi encontrado nos Yakut da Sibéria (88,9%) e nos Miaozi (28,6%), nos Naxi (12,5%) e nos Han (4,2%) da China. Fora da Ásia, o genótipo foi observado entre os Russian (12,5%) e os Brahui do Paquistão (8,0%).

- Nas Américas, o genótipo 14-15-15-16 foi o mais frequente (23,2%). Este genótipo foi observado nos Pima (64,3%) e nos Mayas (33,3%), mas não foi encontrado nos Colombians, Karitianas e Surui. O genótipo também foi observado nos Daur da Mongolia (14,3%), nos Basques da França (6,3%), nos Hazara do Paquistão (4,2%) e nos Han da China (4,2%).
- Na Ásia Central, o genótipo mais frequente foi o 12-15-15-16 (13,7%). Este genótipo foi visto entre os Brahui (16,0%), Baloshi (16,0%), Makrani (15,0%), Pathan (30,0%) e Burusho do Paquistão (23,8%), mas não foi observado na população Hazara. Fora da Ásia Central, o genótipo foi encontrado nos Bedouin de Israel (7,1%), nos Uygur da China (12,5%), nos Sardinian (6,3%), nos Adygei do Caucasus (14,3%) e nos Russians (6,3%).
- No Oriente Médio, os genótipos mais frequentes foram o 12-14-16-18 e o 14-16-16-17 (16,4% cada um deles). O primeiro foi encontrado exclusivamente na população Bedouin de Israel (46,2%), enquanto que o segundo foi observado exclusivamente entre os Mozabite da Algéria (55,0%) e os Druze de Israel (14,3%).
- Na Oceania, os genótipos mais frequentes foram o 12-12-17-17, 12-16-16-17 e o 12-17-17-17 (14,2% cada um deles). Os dois primeiros foram encontrados nos Papuan (7,7% e 15,4%, respectivamente) e nos Melanesians (25,0% e 12,5%, respectivamente) enquanto que o último foi observado somente entre os Melanesians (37,5%).

As diversidades genotípicas foram altas em todas as regiões geográficas estudadas, sendo 0,98 para os Europeus, 0,97 na Ásia Central e Leste Asiático, 0,95 na África, 0,94 na Oceania, 0,92 no Oriente Médio e 0,90 nas Américas (tabela 12).

Tabela 11 – Frequências alélicas do microsatélite *DYS464* encontradas em cada população e região geográfica do painel do CEPH-HGDP

Populações	Frequência alélica – <i>DYS464</i>													
	9	10	11	12	13	14	15	16	17	18	19	20	22	23
África														
Biaka				0.017	0.167	0.100	0.125	0.342	0.100	0.150				
Mbuti			0.019		0.192	0.192	0.096	0.154	0.096	0.212	0.038			
Mandenka				0.047	0.172	0.016	0.141	0.344	0.094	0.125		0.031	0.016	0.016
Yoruba				0.058	0.154	0.058	0.135	0.288	0.173	0.115		0.019		
Bantu NE		0.023	0.045		0.091	0.159	0.136	0.273	0.136	0.114	0.023			
San				0.143	0.214	0.214	0.214	0.179				0.036		
Bantu SE				0.031	0.125	0.125	0.125	0.313	0.188	0.063	0.031			
África Total		0.003	0.008	0.033	0.161	0.109	0.133	0.288	0.112	0.128	0.010	0.010	0.003	0.003
Oriente Médio														
Mozabite				0.038	0.050	0.213	0.050	0.463	0.163	0.025				
Bedouin			0.009	0.170	0.018	0.250	0.080	0.250	0.098	0.125				
Druze				0.071	0.143	0.286	0.125	0.196	0.161	0.018				
Palestinian				0.250	0.294	0.294	0.029	0.074	0.029	0.029				
Oriente Médio Total			0.003	0.136	0.108	0.256	0.070	0.256	0.111	0.060				
Ásia Central														
Brahui				0.180	0.100	0.220	0.230	0.190	0.060	0.020				
Baloshi			0.020	0.140	0.050	0.200	0.250	0.290	0.030	0.010	0.010			
Hazara			0.188	0.208	0.063	0.042	0.271	0.135	0.021	0.073				
Makrani				0.150	0.113	0.150	0.313	0.200	0.050	0.025				
Sindhi				0.143	0.095	0.071	0.345	0.321	0.024					

Pathan		0.025	0.150	0.075	0.163	0.350	0.163	0.050	0.025	
Kalash		0.113	0.100	0.150	0.188	0.175	0.213	0.063		
Burusho		0.125	0.125	0.1	0.15	0.35	0.125	0.025		
Ásia Central Total		0.059	0.151	0.091	0.149	0.283	0.206	0.040	0.020	0.001

Leste Asiático

Han	0.010	0.010	0.167	0.260	0.208	0.208	0.094	0.021	0.021	
Tujia		0.056	0.194	0.278	0.194	0.194	0.083			
Yizu			0.167	0.222	0.194	0.111	0.194	0.111		
Miaozu			0.107	0.179	0.321	0.214	0.179			
Oroqen		0.107	0.179	0.071	0.036	0.250	0.357			
Daur		0.179	0.036	0.143	0.143	0.250	0.143	0.071	0.036	
Mongola		0.107	0.036	0.250	0.107	0.393	0.107			
Hezhen		0.050	0.200	0.100	0.150	0.250	0.200	0.050		
Xibo		0.063	0.250	0.188	0.031	0.281	0.125	0.063		
Uygur		0.031	0.125	0.063	0.156	0.281	0.188	0.094	0.031	0.031
Dai			0.071	0.393	0.143	0.143	0.214	0.036		
Lahu		0.036		0.286		0.071	0.571	0.036		
She			0.167	0.458	0.125	0.125	0.042	0.083		
Naxi			0.125	0.156	0.156	0.531		0.031		
Tu		0.036	0.071	0.500	0.107	0.214	0.036	0.036		
Yakut			0.014	0.028	0.458	0.472	0.028			
Japanese		0.011	0.045	0.193	0.125	0.148	0.216	0.170	0.091	
Cambodian			0.250	0.125	0.250	0.250	0.083	0.042		
Leste Asiático Total	0.001	0.031	0.113	0.206	0.182	0.247	0.148	0.052	0.017	0.001

Oceania

Papuan		0.096	0.135	0.135	0.135	0.327	0.154	0.019		
--------	--	-------	-------	-------	-------	-------	-------	-------	--	--

Melanesian		0.375	0.031	0.031	0.031	0.094	0.438		
Oceania Total		0.202	0.095	0.095	0.095	0.238	0.262	0.012	
<hr/>									
Europa									
French		0.125	0.042	0.167	0.188	0.292	0.125	0.042	0.021
Basque			0.016	0.047	0.453	0.109	0.297	0.078	
Sardinian	0.125	0.063	0.031	0.344	0.234	0.109	0.063	0.031	
Bergamo	0.031		0.094	0.156	0.344	0.188	0.094	0.094	
Tuscan	0.042	0.042	0.083	0.167	0.250	0.250	0.167		
Orcadian	0.036	0.071		0.036	0.464	0.179	0.179	0.036	
Adygei		0.107	0.357	0.179	0.143	0.179	0.036		
Russian		0.125	0.047	0.188	0.281	0.313	0.031	0.016	
Europa Total	0.031	0.068	0.065	0.170	0.298	0.199	0.125	0.040	0.003
<hr/>									
América									
Pima		0.036	0.107	0.286	0.321	0.196	0.054		
Maya			0.083	0.167	0.417	0.083	0.167	0.083	
Colombian			0.100	0.400	0.250	0.100	0.100	0.050	
Karitiana		0.075	0.225	0.050	0.400		0.025	0.225	
Surui			0.182	0.068	0.455	0.045	0.182	0.068	
América Total		0.029	0.151	0.180	0.372	0.093	0.093	0.081	

Tabela 12 - Número de indivíduos, alelos, alelos região-específicos, genótipos, genótipos região-específicos e diversidade genotípica encontrados em cada região geográfica e no total das amostras estudadas para o microssatélite *DYS464*

Região geográfica	Nº indivíduos	Nº alelos	Nº de alelos região-específicos	Nº genótipos	Nº de genótipos região-específicos	Diversidade genotípica
África	98	13	4	49	20 (40%)	0,95±0,01
Oriente Médio	79	8	0	29	7 (24%)	0,92±0,01
Ásia Central	175	9	0	72	19 (26,4%)	0,97±0,01
Leste Asiático	173	10	1	83	29 (34,9%)	0,97±0,01
Oceania	21	7	0	13	3 (23%)	0,94±0,03
Europa	88	9	0	53	8 (15%)	0,98±0,01
América	43	7	0	16	4 (25%)	0,90±0,03
TOTAL	677	15	5	175	90 (51,4%)	0,99

3.4.1.2.2 – Análise de Variância Molecular (AMOVA)

A fim de verificar a distribuição da variabilidade genética entre as diferentes populações e regiões mundiais, os genótipos do *DYS464* foram submetidos a uma hierárquica análise de variância molecular (AMOVA) utilizando o programa *Arlequin*. Analisamos a distribuição da variabilidade em três condições: a) o mundo como uma única região com 52 populações; b) o mundo dividido em 7 regiões geográficas (África, Oriente Médio, Ásia Central, Leste Asiático, Oceania, Europa e América) abrangendo as 52 populações e c) cada região separadamente abrangendo suas respectivas populações. Nossos resultados revelaram baixos níveis de estruturação populacional: 88,42% da variabilidade genética foi encontrada dentro das populações, 9,62% entre populações dentro das grandes regiões geográficas e apenas 1,96% entre regiões (Tabela 13). Ao analisarmos cada região separadamente, a variabilidade genética dentro das populações foi responsável por mais de 87% de toda a variabilidade encontrada, exceto para o Oriente Médio e Américas, que apresentaram níveis mais baixos de variabilidade dentro das populações (78,22% e 66,80%, respectivamente).

Tabela 13- Análise de variância molecular (AMOVA) de dados do microssatélite *DYS464* referentes às sete regiões geográficas do painel do CEPH-HGDP estudadas. O intervalo de confiança foi de 95%.

Regiões	Número de Regiões	Número de Populações	Distribuição da variação (%)		
			Dentro de Populações	Entre Populações	Entre Regiões
Mundo	1	52	88,71	11,29	-
Mundo	7	52	88,41	9,62	1,96
África	1	7	94,55	5,45	-
Oriente Médio	1	4	78,22	21,78	-
Ásia Central	1	8	95,91	4,09	-
Leste Asiático	1	18	87,99	12,01	-
Oceania	1	2	94,90	5,10	-
Europa	1	8	97,18	2,82	-
América	1	5	66,80	33,20	-

3.4.1.2.3 – Escalonamento Multidimensional (MDS)

Para verificar o poder discriminatório e estruturação populacional do microssatélite *DYS464*, utilizamos os dados genotípicos deste marcador para construirmos matrizes de distância genética baseadas em três diferentes medidas: F_{st} (Distância de Reynolds), R_{st} e $\delta\mu^2$. As matrizes foram calculadas tanto para as 7 regiões geográficas quanto para as 52 populações discriminadamente. Em seguida, gráficos de escalonamento multidimensional foram construídos para representar as matrizes de distância calculadas.

Nossos resultados revelaram que, para as análises com as 52 populações discriminadas, não foi possível observar nenhum agrupamento claro com nenhuma das três medidas de distância utilizadas (Figuras 31, 32 e 33). Para as três medidas de distância utilizadas (F_{st} , R_{st} e $\delta\mu^2$), embora o *stress* tenha sido baixo (0,215 para F_{st} , 0,126 para R_{st} e 0,074 para $\delta\mu^2$), as populações apresentaram-se homogeneamente distribuídas nos gráficos.

Por outro lado, nas análises com as 7 regiões geográficas, observamos estruturação populacional com as três medidas de distância utilizadas (Figuras 34, 35 e 36). No gráfico baseado em F_{st} (*stress* de 0,003) é possível observar dois agrupamentos claros: um incluindo Ásia Central, Leste Asiático e Europa e outro formado por Oceania e Oriente Médio. A África e as Américas apareceram isoladas e não se

incluíram em nenhum grupo (Figura34). No gráfico baseado em R_{st} (*Stress* de 0,004) é possível observar um agrupamento contendo Oriente Médio, Oceania e Europa enquanto que os outros continentes apareceram isolados (Figura 35). No gráfico baseado em $\delta\mu^2$ (Figura 36), é possível observar dois agrupamentos: um contendo Oceania, Oriente Médio e Europa e outro contendo Ásia Central e Leste Asiático. A África e as Américas formaram grupos isolados, estando a primeira mais próxima do grupo da Europa, Oriente Médio e Oceania e as Américas mais próximas do Leste Asiático e Ásia Central.

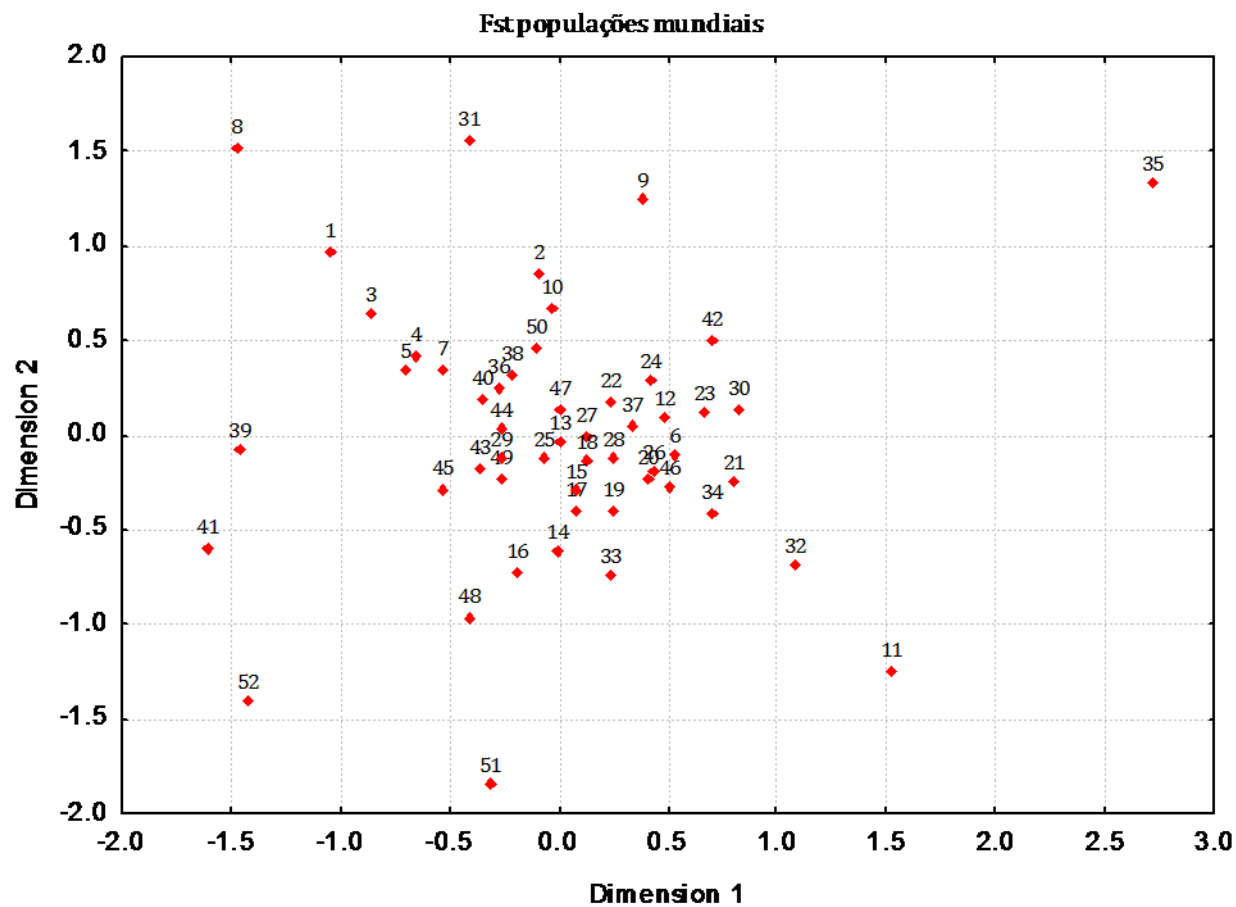


Figura 31 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microssatélite *DYS464* e Fst (Distância de Reynolds) como medida de distância. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

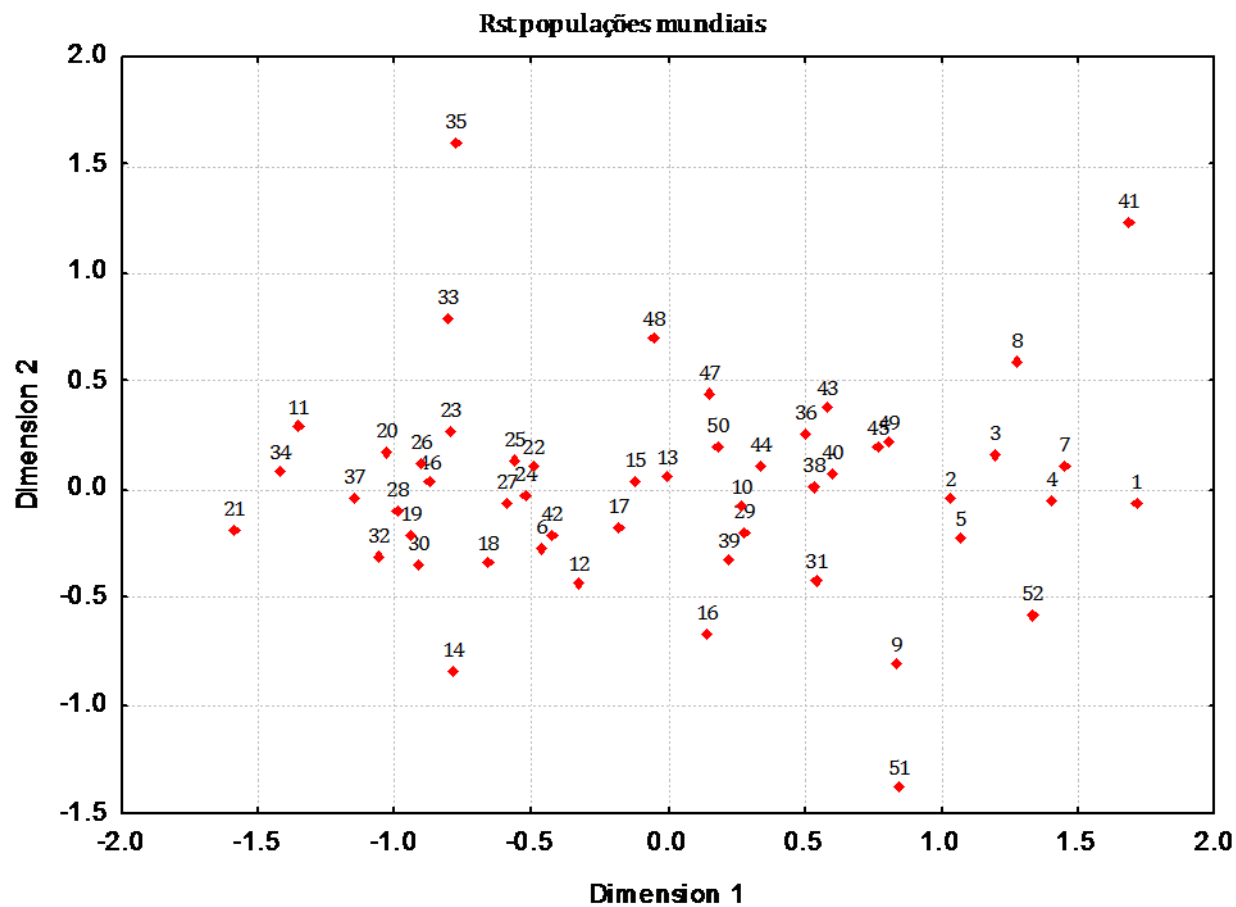


Figura 32 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microsatélite *DYS464* e Rst como medida de distância. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozi, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

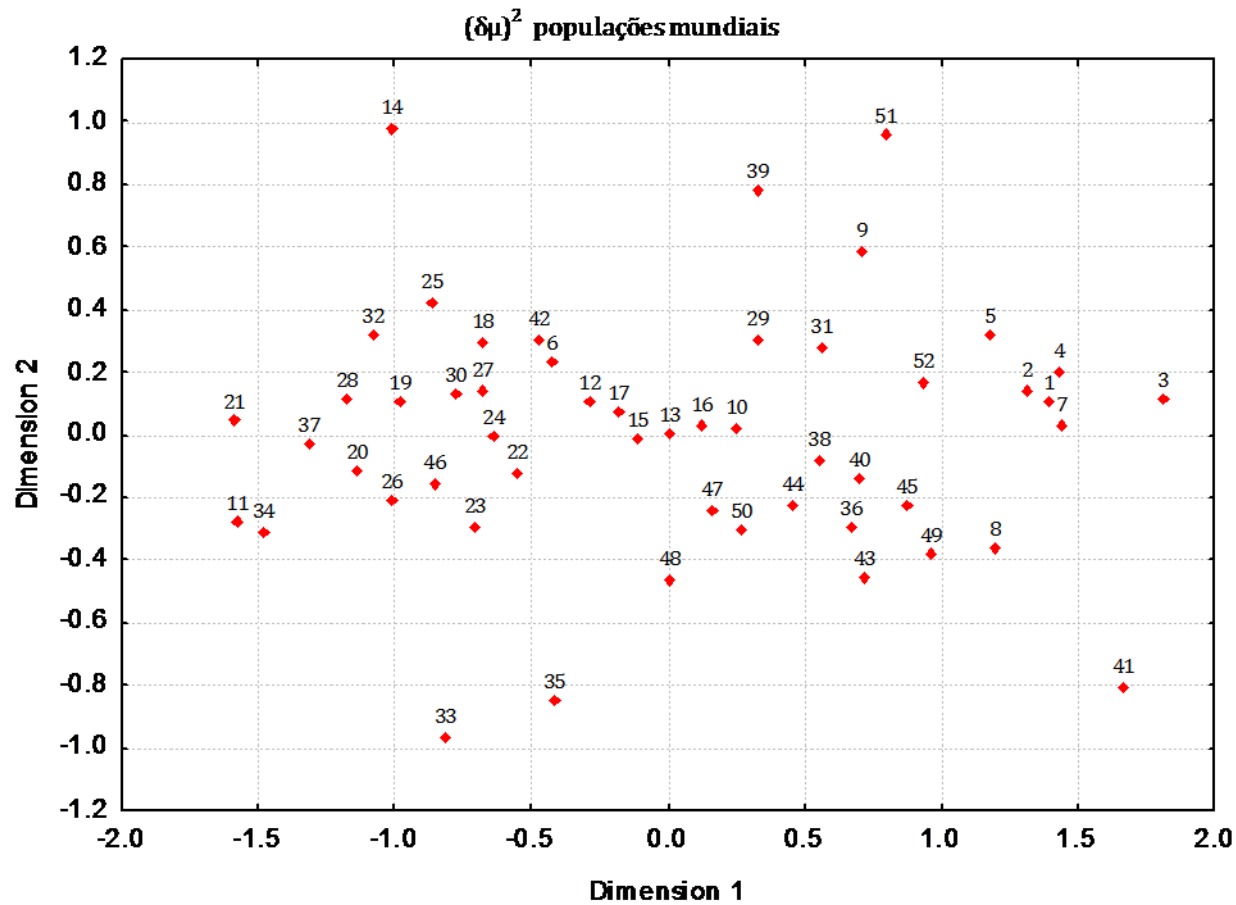


Figura 33 - Escalonamento Multidimensional realizado com as 52 populações mundiais utilizando dados genotípicos do microsatélite *DYS464* e $\delta\mu^2$ como medida de distância. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozi, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

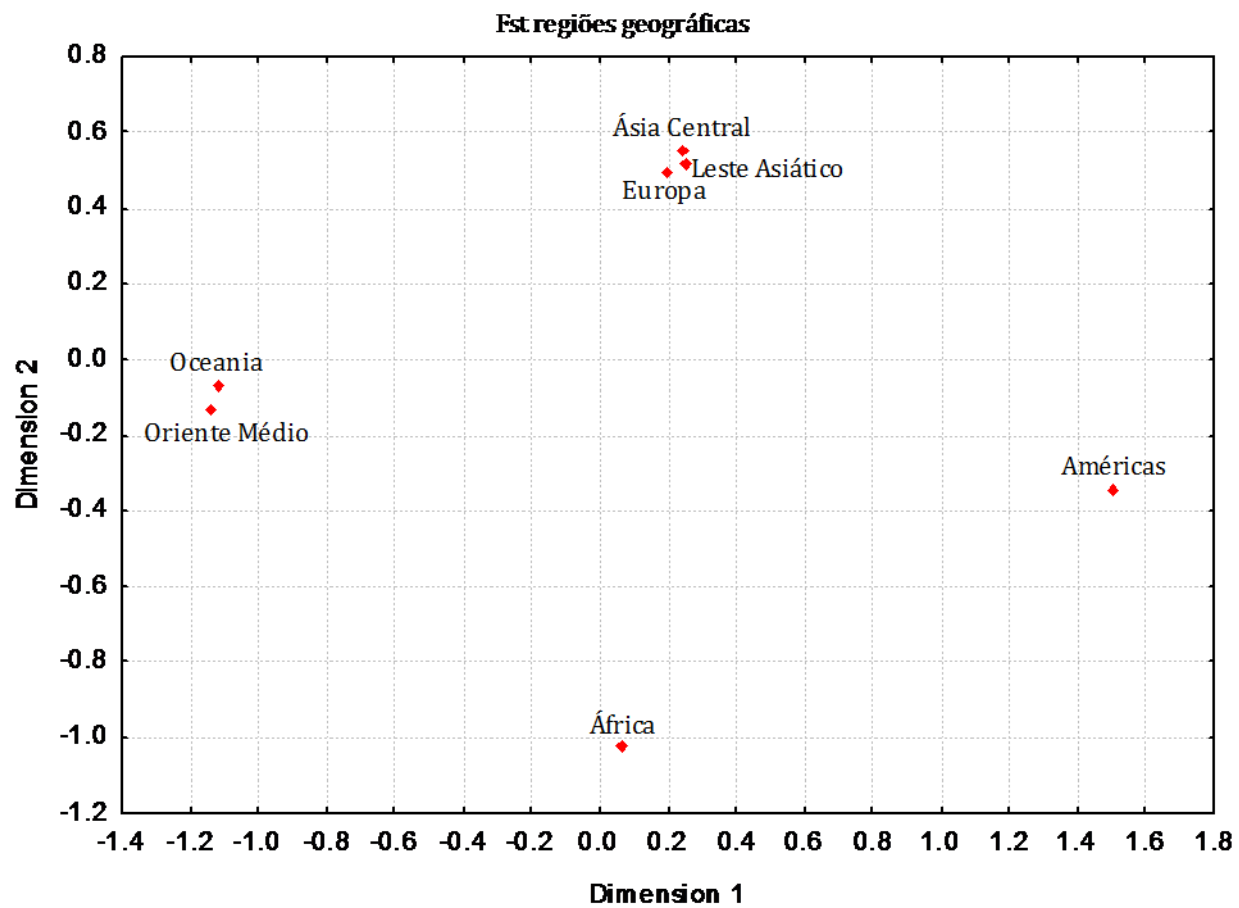


Figura 34 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite *DYS464* e *Fst* (Distância de Reynolds) como medida de distância.

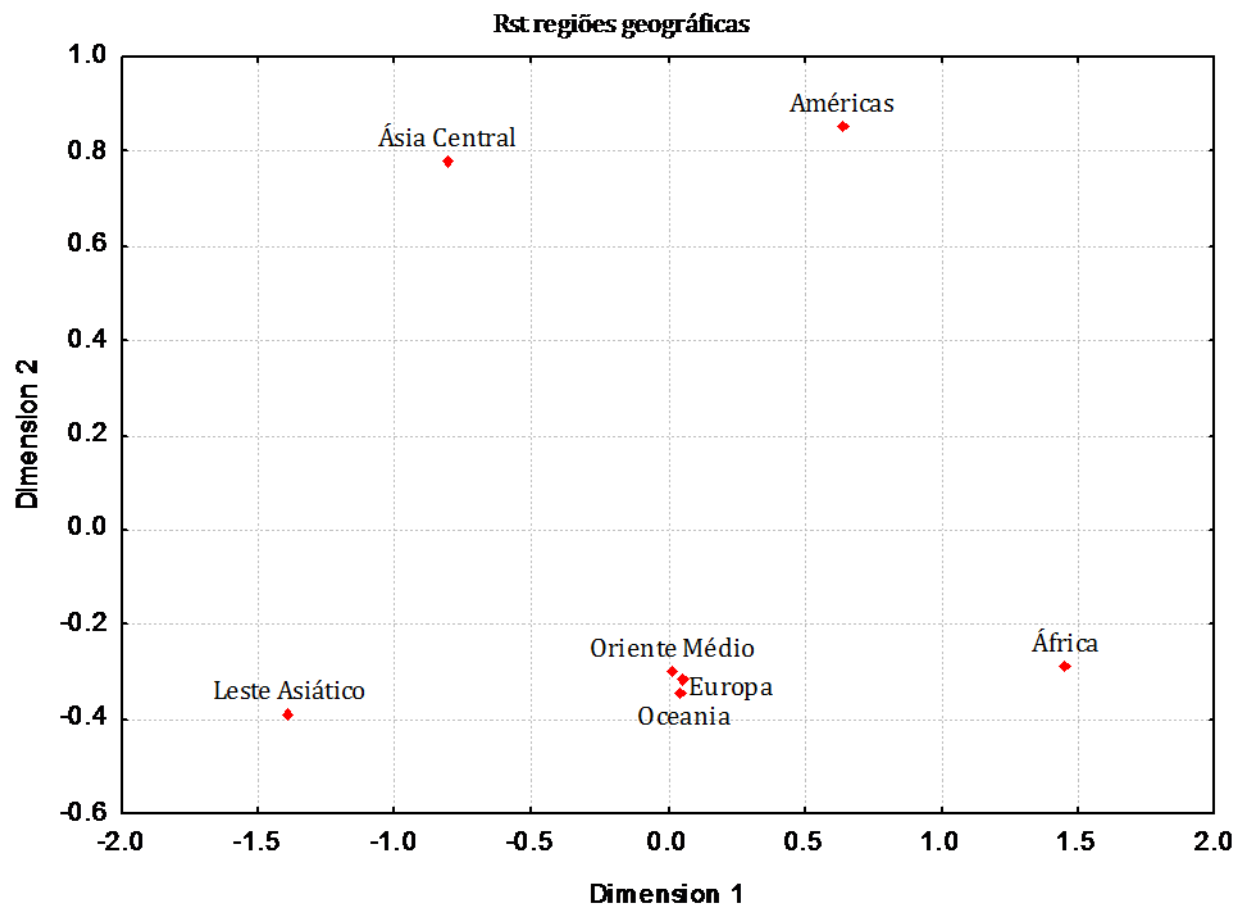


Figura 35 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite *DYS464* e Rst como medida de distância.

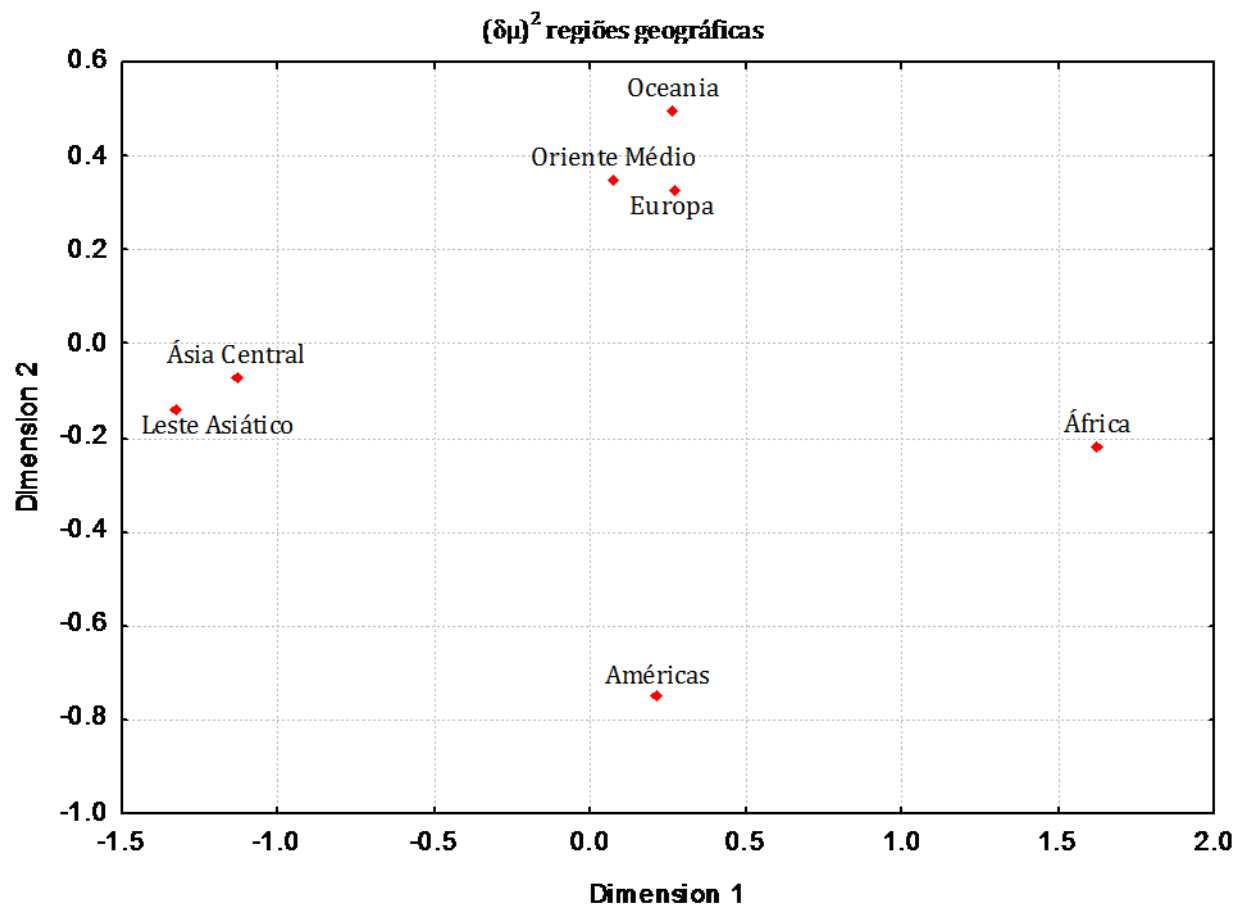


Figura 36 - Escalonamento Multidimensional realizado com as 7 regiões geográficas utilizando dados genotípicos do microssatélite *DYS464* e $\delta\mu^2$ como medida de distância.

3.4.1.2.4 – Análise do Componente Principal (PCA)

Como uma outra estratégia para verificar a estruturação populacional utilizando o microsatélite *DYS464*, realizamos análise do componente principal (PCA) e os valores do primeiro e segundo componentes foram utilizados para a construção de gráficos de dispersão bidimensional. A análise foi feita tanto para as 52 populações mundiais como para as 7 regiões geográficas utilizando dados de frequências genotípicas e alélicas.

Nossos resultados mostraram que entre as 52 populações mundiais, não foi possível observar estruturação populacional para nenhuma das análises utilizadas (dados de frequências genotípicas e alélicas). O total de variação genética recuperada pelos dois primeiros componentes nas duas análises foi de 7,5% e 34,76%, respectivamente (Figuras 37 e 38).

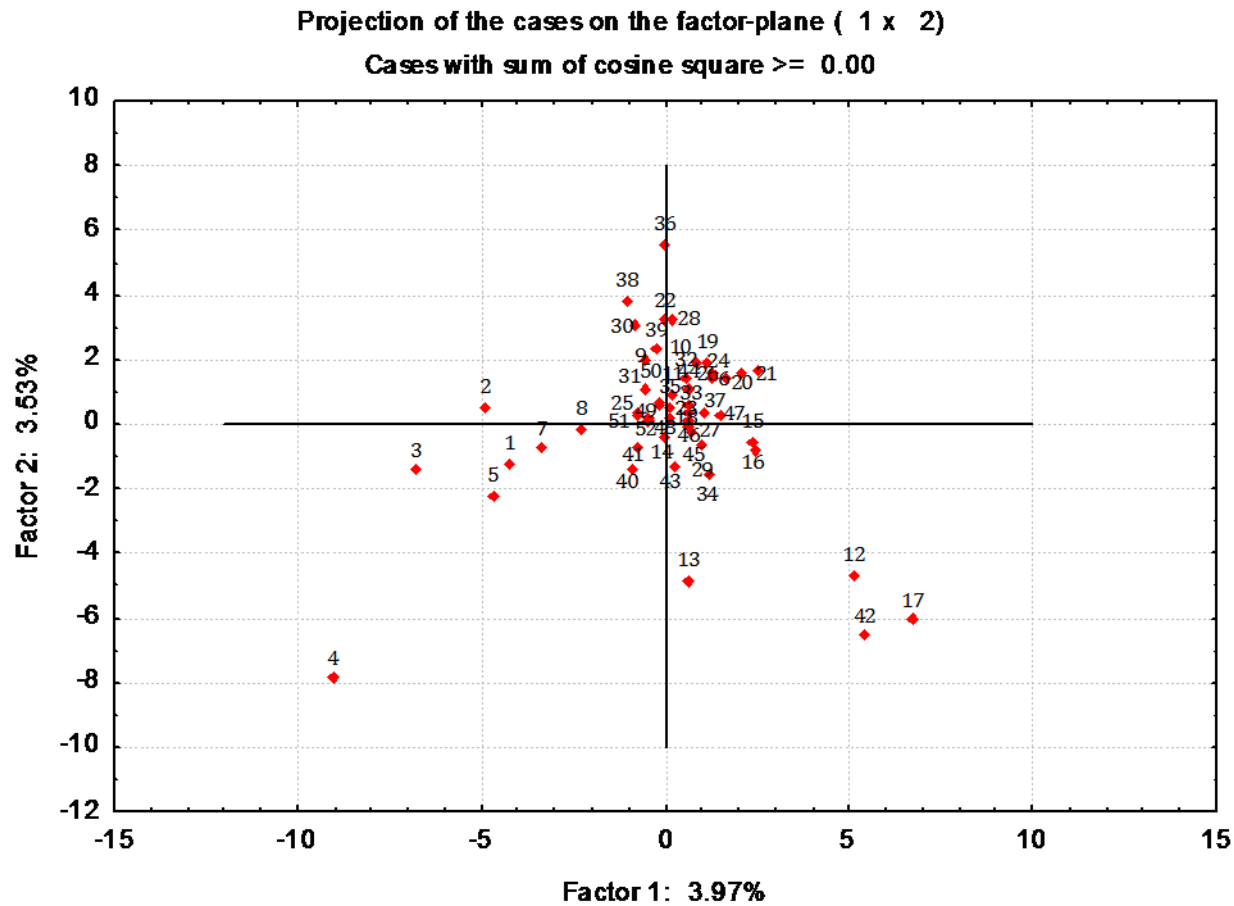


Figura 37 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências genotípicas do microsatélite *DYS464* nas 52 populações. Os valores dos componentes 1 e 2 estão mostrados no gráfico. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

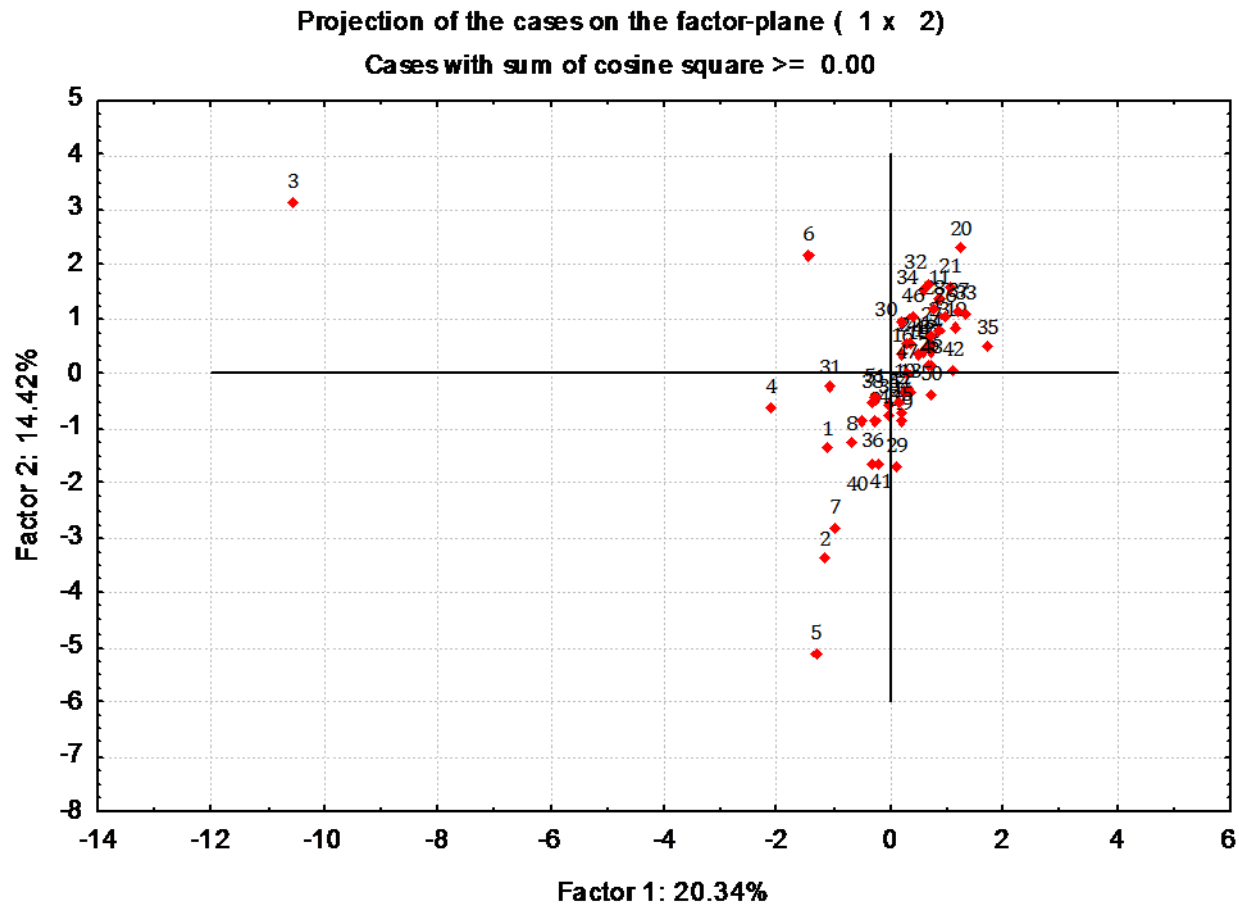


Figura 38 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências alélicas do microsatélite *DYS464* nas 52 populações. Os valores dos componentes 1 e 2 estão mostrados no gráfico. Cada ponto representa uma população numerada como a seguir: (1) Biaka Pygmies, (2) Mbuti Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu NE, (6) San, (7) Bantu SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN Malesian, (40) French, (41) French Basque, (42) Sardinian, (43) North Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui.

A análise dos componentes principais com as 7 regiões geográficas, revelou estruturação populacional tanto para dados de frequências genotípicas quanto para dados de frequências alélicas (Figuras 39 e 40). Na análise utilizando frequências genotípicas o total de variação genética recuperada pelos dois primeiros componentes foi de 47,43%. No gráfico, é possível observar um agrupamento claro incluindo Oriente Médio, Oceania e Américas enquanto que os outros continentes aparecem isolados estando o Leste Asiático o mais distante dos demais (Figura 39).

Quando dados de frequências alélicas foram utilizados, um total de 69,00% da variação genética foi recuperada pelos dois primeiros componentes. O gráfico revelou que o primeiro componente foi capaz de separar a África dos demais continentes. Além disso, é possível observar dois agrupamentos: um incluindo o Leste Asiático e as Américas e outro contendo Europa e Ásia Central. Próximo ao segundo agrupamento está o Oriente Médio e logo acima a Oceania (Figura 40).

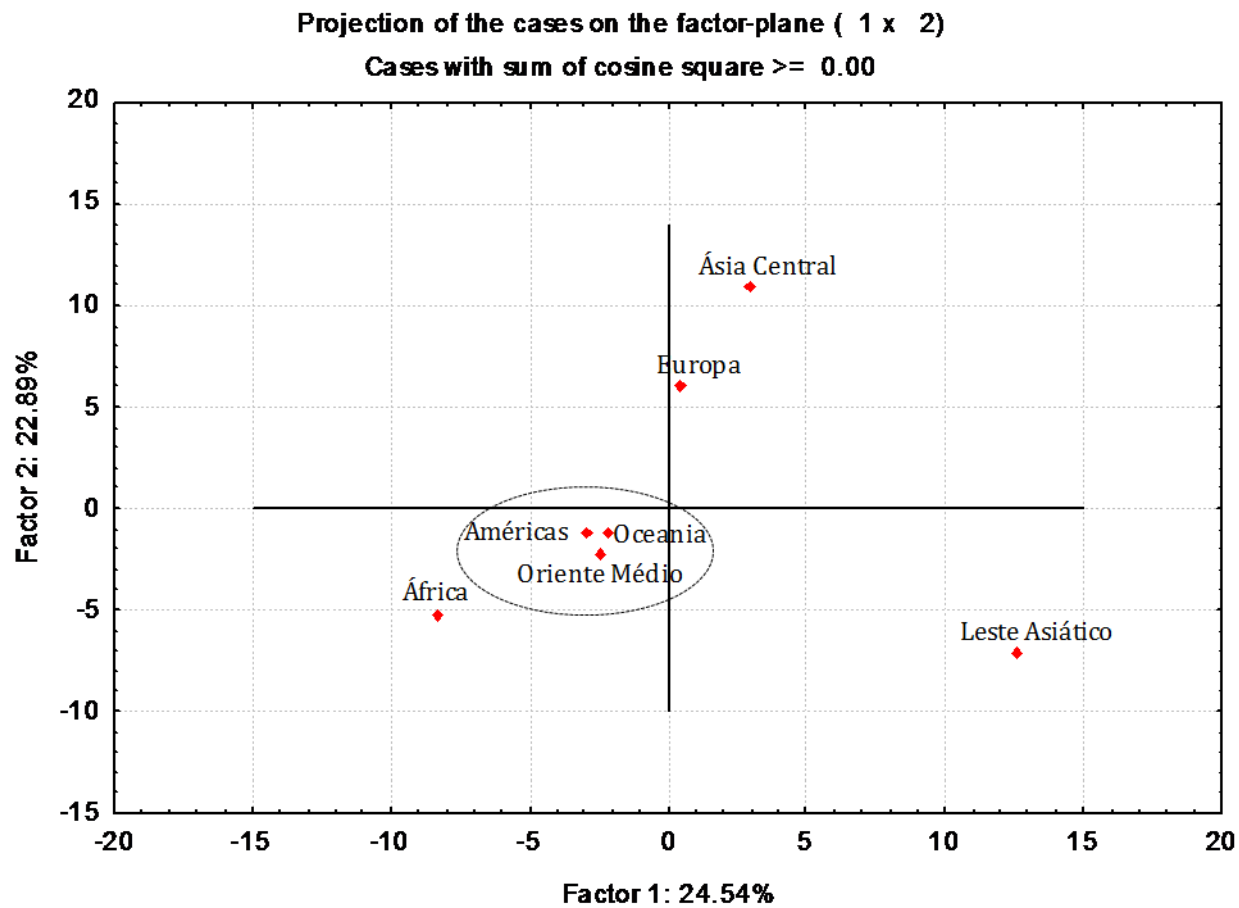


Figura 39 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências genotípicas do microssatélite *DYS464* nas 7 regiões geográficas. Os valores dos componentes 1 e 2 estão mostrados no gráfico.

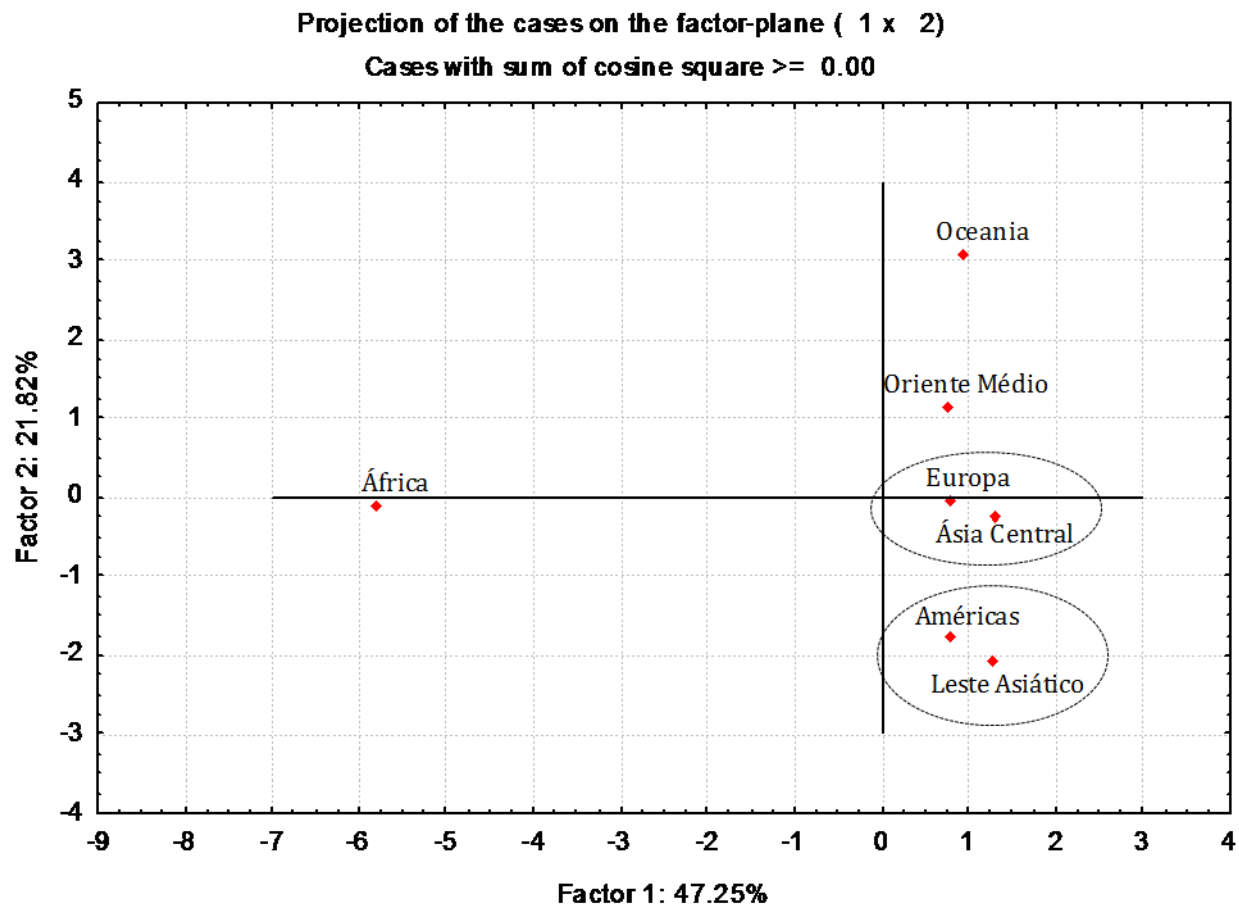


Figura 40 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências alélicas do microsatélite *DYS464* nas 7 regiões geográficas. Os valores dos componentes 1 e 2 estão mostrados no gráfico.

3.4.1.3 – Microsatélite *DYS464* em populações miscigenadas

Além de estudar o *DYS464* em populações mundiais nativas, fizemos análises do microsatélite também em populações miscigenadas. Neste estudo nosso objetivo foi, além de caracterizar as populações miscigenadas para o *DYS464*, avaliar a utilização das frequências alélicas do microsatélite como uma ferramenta para estudo de ancestralidade genética.

Para isso, utilizamos dados de tipagem do microsatélite *DYS464* de afro-americanos e caucasianos (disponíveis em Butler e Schoske, 2005), indivíduos pretos de São Paulo (Carvalho e Pena, dados não publicados) e indivíduos brancos brasileiros tipados no presente estudo. Como referência, utilizamos as populações parentais de africanos, europeus e ameríndios do painel do CEPH-HGDP. Dados de frequências alélicas (Tabela 14) e dados genotípicos (Anexo 2) foram utilizados para as análises de estruturação genética e miscigenação.

Tabela 14 – Frequências alélicas do *DYS464* em populações parentais e miscigenadas

Alelos	Frequências alélicas/Populações						
	Pretos (n=124)	Brancos (n=106)	Caucasianos (n=256)	Afro- Americanos (n=256)	Africanos (n=90)	Europeus (n=88)	Ameríndios (n=43)
9		0.002					
10					0.003		
11	0.008	0.017	0.011	0.009	0.008	0.031	
12	0.034	0.042	0.052	0.033	0.033	0.068	0.029
13	0.129	0.057	0.042	0.124	0.164	0.065	0.151
13.1			0.001	0.001			
14	0.101	0.127	0.134	0.057	0.108	0.170	0.180
14.3			0.001	0.003			
15	0.218	0.309	0.335	0.190	0.133	0.298	0.372
15.1			0.003				
15.2	0.004						
15.3			0.002				
16	0.248	0.172	0.161	0.303	0.286	0.199	0.093
17	0.181	0.191	0.212	0.175	0.106	0.125	0.093
18	0.067	0.073	0.038	0.095	0.133	0.040	0.081
19	0.010	0.005	0.006	0.009	0.008	0.003	
20		0.002	0.003	0.002	0.011		
21		0.002					
22					0.003		
23					0.003		

3.4.1.3.1- Escalonamento Multidimensional

Para verificar a distância genética entre as populações e a localização das populações miscigenadas em relação às parentais, calculamos através de dados genotípicos, matrizes de distância genética utilizando três medidas diferentes: F_{st} (distância de Reynolds), R_{st} e $\delta\mu^2$. Posteriormente, as matrizes foram graficamente representadas através do escalonamento multidimensional. Utilizando dados de F_{st} , os resultados mostraram clara separação dos brancos de São Paulo, caucasianos e europeus em relação às demais populações. Os afro-americanos e os pretos brasileiros apareceram bem próximos entre si, enquanto que os africanos e ameríndios encontraram-se isolados (Figura 41, *stress* de 0,0009). Nos gráficos utilizando medidas de R_{st} (Figura 42, *stress* de 0,0000) e $\delta\mu^2$ (Figura 43, *stress* de 0,0000), os resultados foram bem semelhantes, no entanto na segunda análise brancos de São Paulo e Caucasionos encontraram-se em um mesmo ponto, enquanto que na primeira os Caucasionos apareceram equidistantes dos brancos e pretos brasileiros (Figuras 42 e 43).

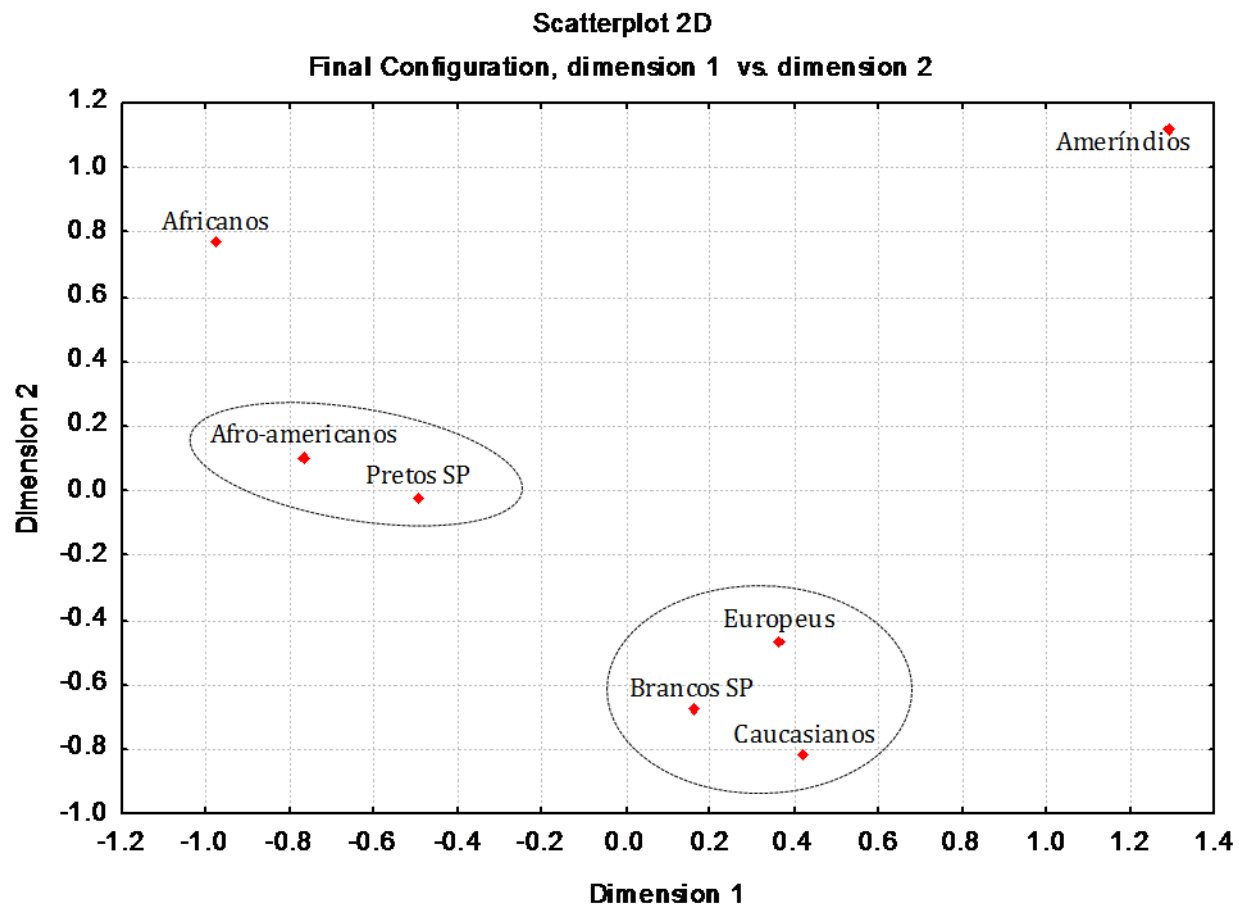


Figura 41 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite *DYS464* e *Fst* (distância de Reynolds) como medida de distância.

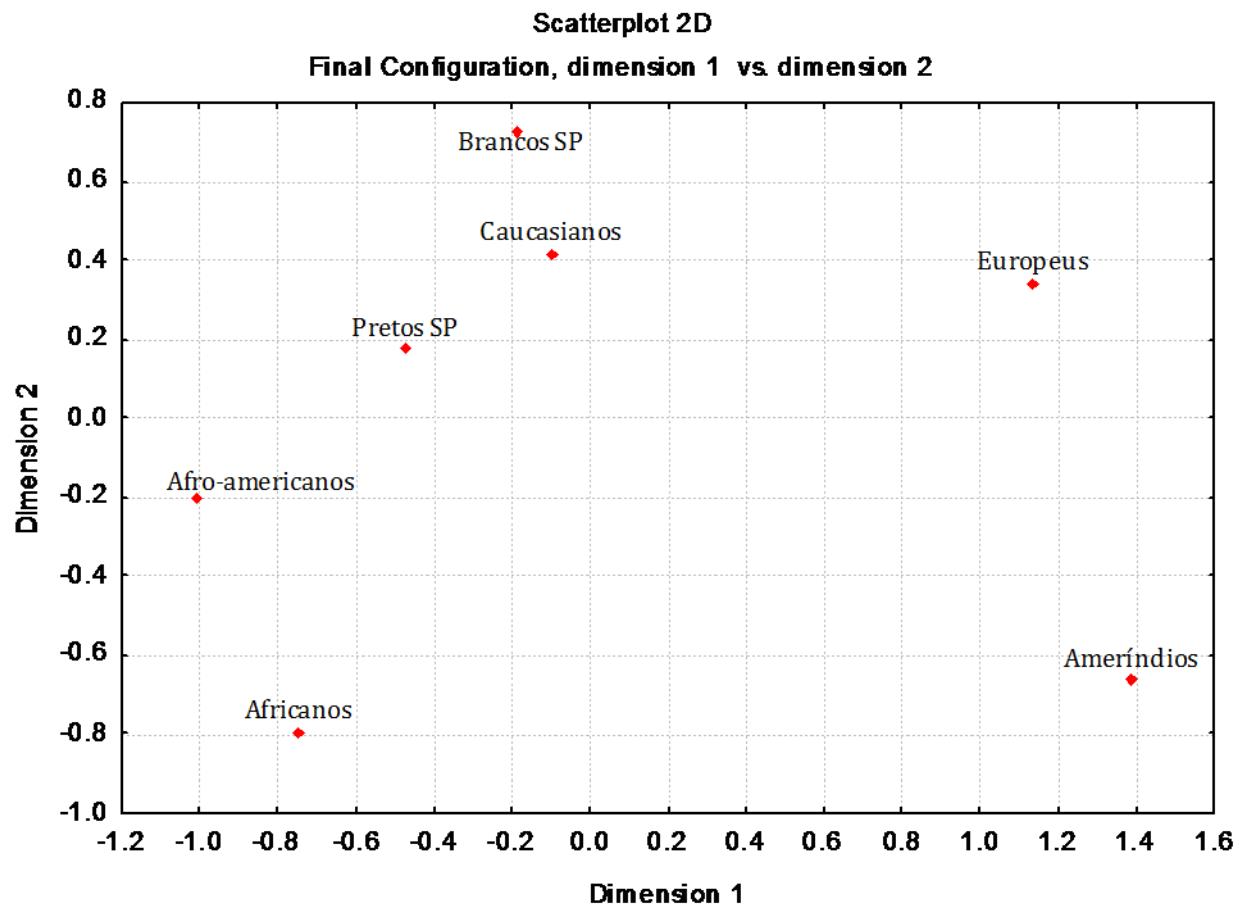


Figura 42 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite *DYS464* e Rst como medida de distância.

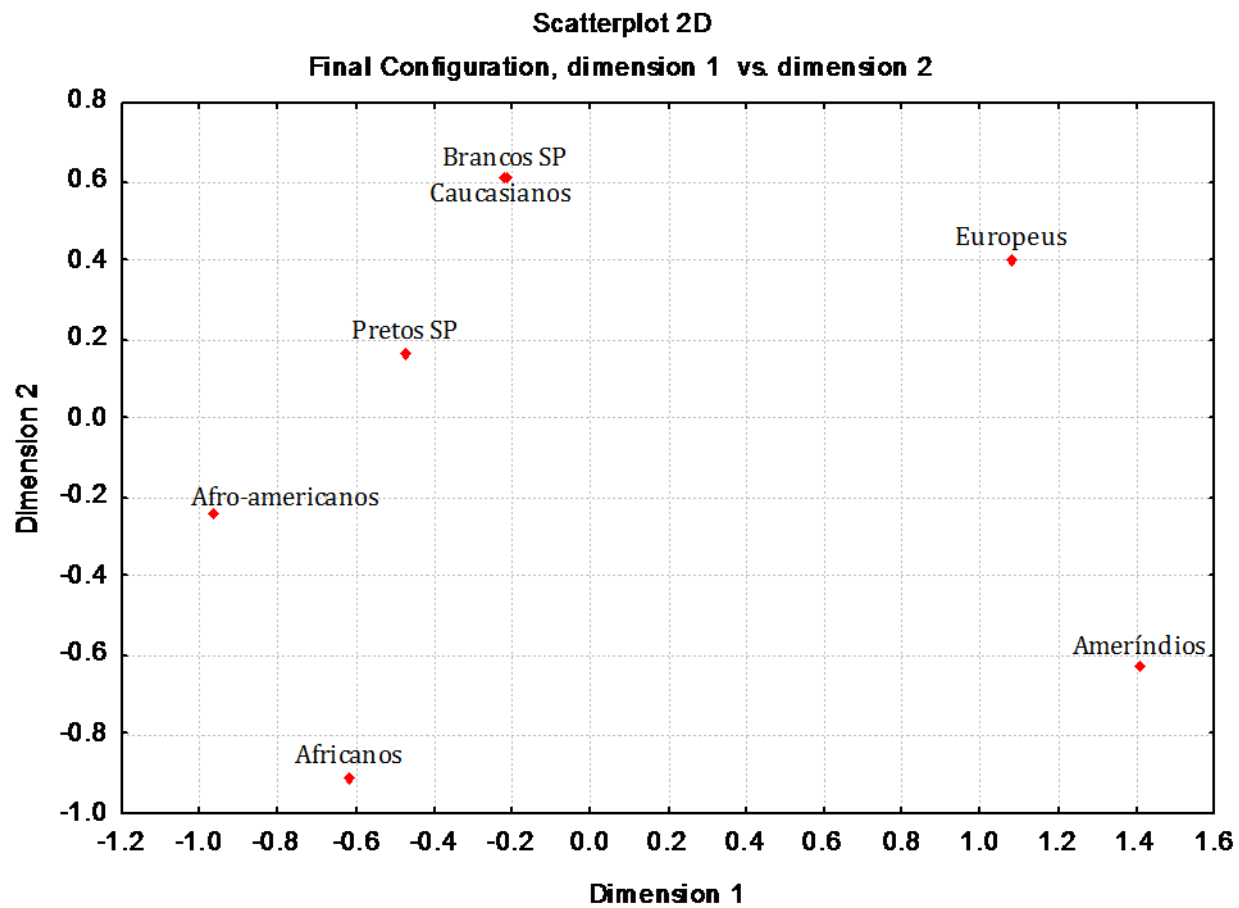


Figura 43 - Escalonamento Multidimensional realizado com as populações miscigenadas e parentais utilizando dados genotípicos do microssatélite *DYS464* e $\delta\mu^2$ como medida de distância.

3.4.1.3.2- Análise do Componente Principal

Para verificar estruturação genética entre as populações miscigenadas e parentais, dados de frequências alélicas gerados pelo programa *Arlequin* foram submetidos à análise do componente principal. Um total de 57,41% da variação genética foi recuperada pelos dois primeiros componentes. Nossos resultados revelaram uma evidente separação entre africanos e as demais populações. Além disso, é possível observar dois agrupamentos menores: um abrangendo afro-americanos e pretos brasileiros e outro englobando brancos de São Paulo e europeus (Figura 44).

3.4.1.3.3 - Estimativa do grau de miscigenação populacional

Para estimar as proporções de mistura nas populações de brancos e pretos brasileiros, afro-americanos e caucasianos, submetemos as frequências alélicas previamente calculadas ao método gráfico, descrito por Collins-Schramm e colaboradores. (2002), baseado na minimização da soma dos quadrados. Para os pretos brasileiros, o valor de proporção africana encontrada foi de 53,00%, para os brancos de 1,50%, para os afro-americanos 71,54% e para os caucasianos 3,40% (Figuras 45, 46, 47 e 48).

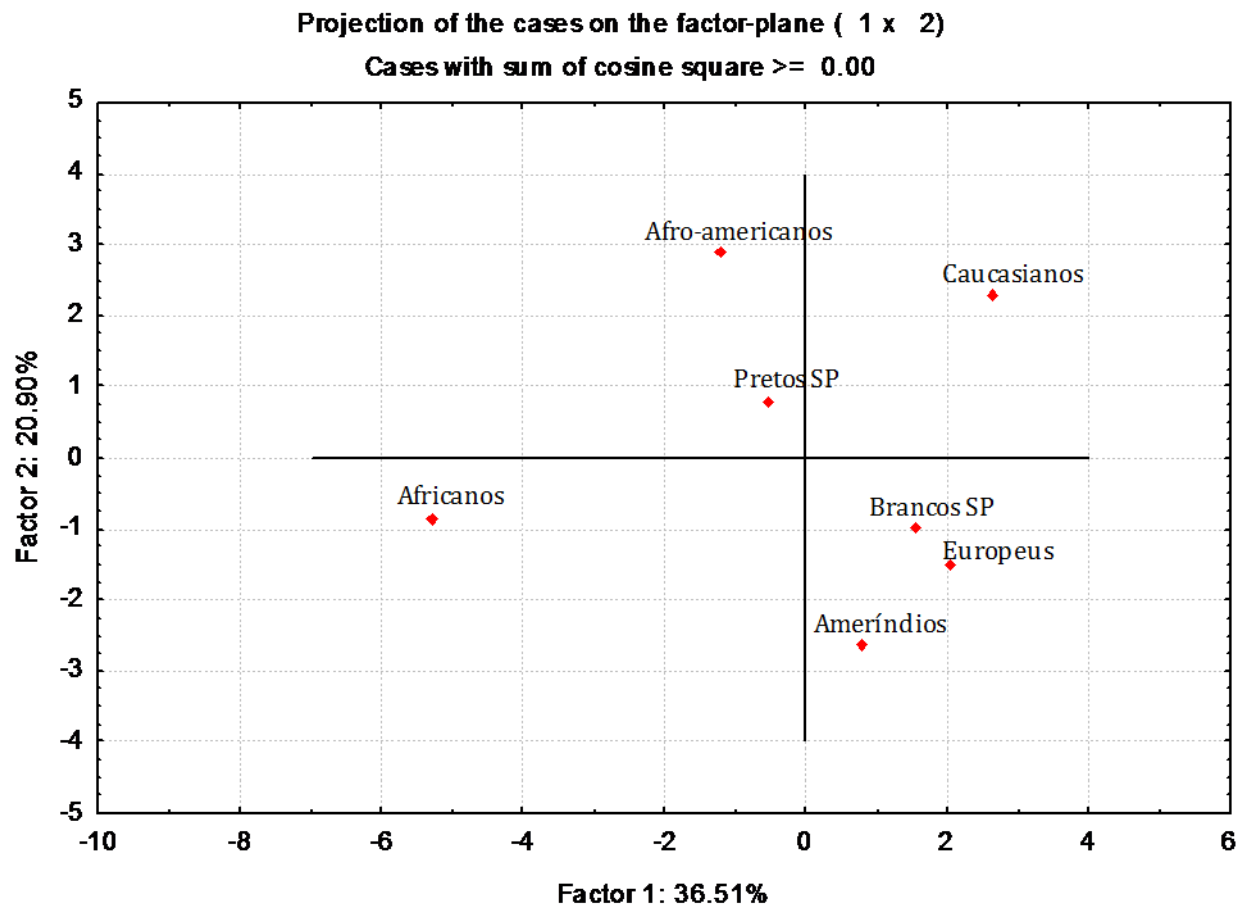


Figura 44 - Gráfico de dispersão dos primeiros dois componentes principais baseados nas frequências alélicas do microssatélite *DYS464* encontradas nas populações miscigenadas e parentais. Os valores dos componentes 1 e 2 estão mostrados no gráfico.

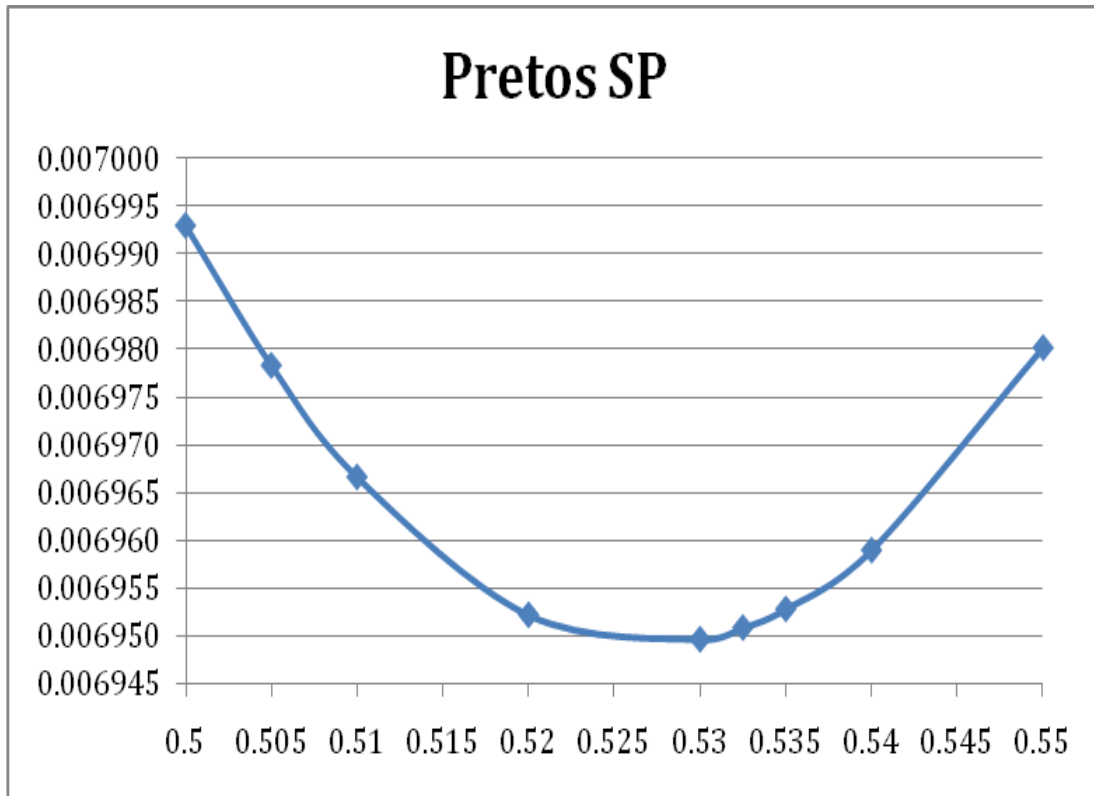


Figura 45 – Gráfico de minimização da equação $\sum [XP_i^{AF} + (1-X)P_i^{EU} - P_i^{PR}]^2$ onde X é a proporção de ancestralidade africana testada para os pretos de São Paulo, P_i^{AF} é a frequência do alelo i do DYS464 em africanos, P_i^{EU} é a frequência do alelo i em europeus e P_i^{PR} é a frequência do alelo i em pretos de São Paulo. O eixo da abscissa representa os diferentes valores de X testados e o da ordenada o valor da equação descrita acima para cada X. A melhor estimativa do valor de X é aquele que minimiza a equação.

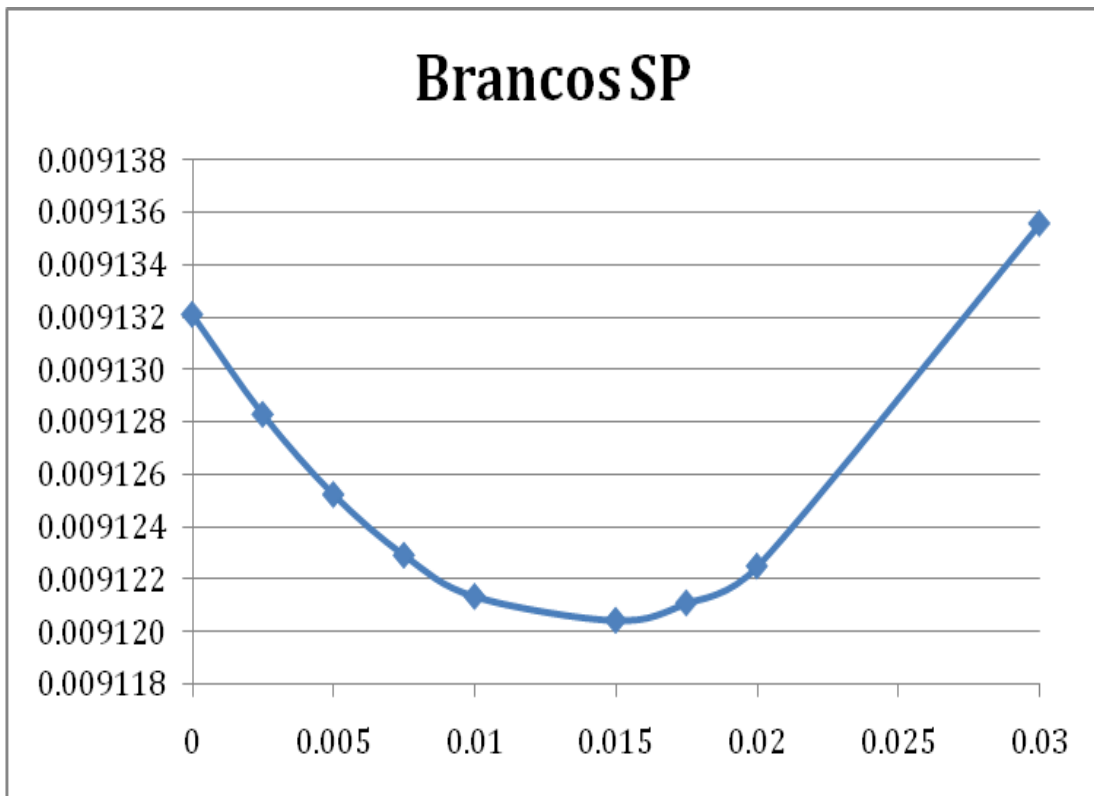


Figura 46 – Gráfico de minimização da equação $\sum [XP_i^{AF} + (1-X)P_i^{EU} - P_i^{BR}]^2$ onde X é a proporção de ancestralidade africana testada para os brancos de São Paulo, P_i^{AF} é a frequência do alelo i do DYS464 em africanos, P_i^{EU} é a frequência do alelo i em europeus e P_i^{BR} é a frequência do alelo i em brancos de São Paulo. O eixo da abscissa representa os diferentes valores de X testados e o da ordenada o valor da equação descrita acima para cada X. A melhor estimativa do valor de X é aquele que minimiza a equação.

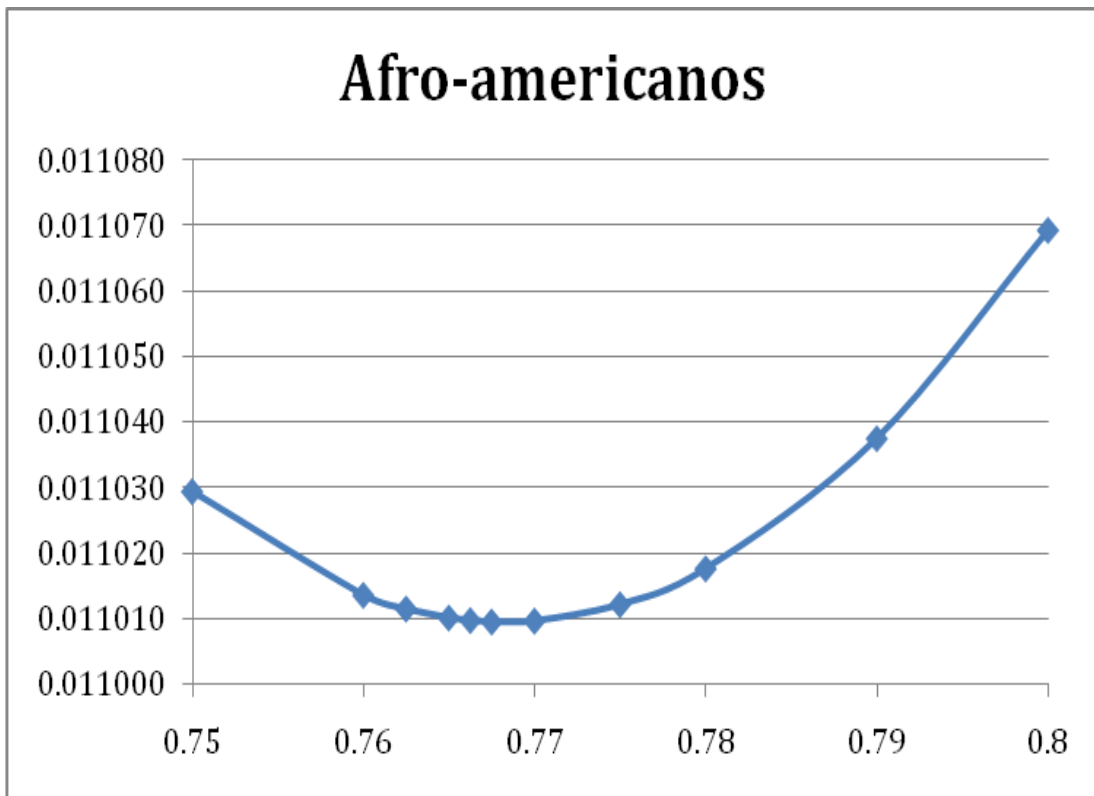


Figura 47 – Gráfico de minimização da equação $\sum [XP_i^{AF} + (1-X)P_i^{EU} - P_i^{AfAm}]^2$ onde X é a proporção de ancestralidade africana testada para os afro-americanos, P_i^{AF} é a frequência do alelo i do DYS464 em africanos, P_i^{EU} é a frequência do alelo i em europeus e P_i^{AfAm} é a frequência do alelo i em afro-americanos. O eixo da abscissa representa os diferentes valores de X testados e o da ordenada o valor da equação descrita acima para cada X. A melhor estimativa do valor de X é aquele que minimiza a equação.

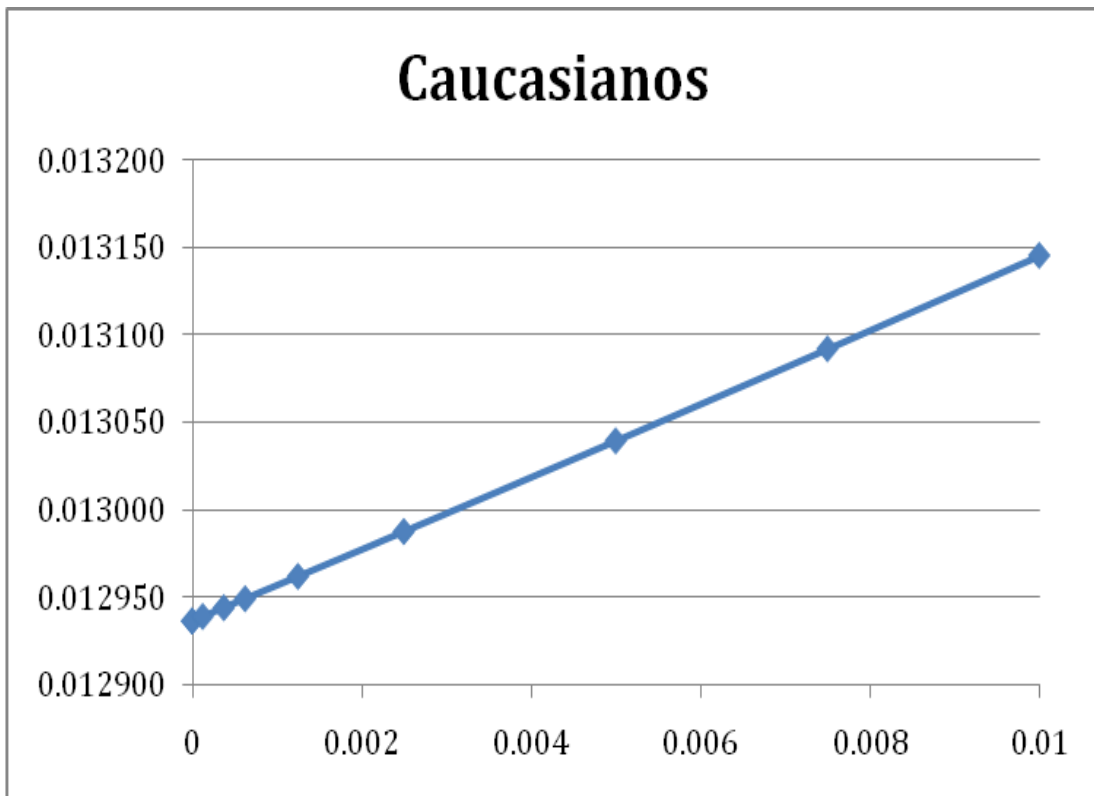


Figura 48 – Gráfico de minimização da equação $\sum [XP_i^{AF} + (1-X)P_i^{EU} - P_i^{Cau}]^2$ onde X é a proporção de ancestralidade africana testada para os caucasianos, P_i^{AF} é a frequência do alelo i do DYS464 em africanos, P_i^{EU} é a frequência do alelo i em europeus e P_i^{Cau} é a frequência do alelo i em caucasianos. O eixo da abscissa representa os diferentes valores de X testados e o da ordenada o valor da equação descrita acima para cada X. A melhor estimativa do valor de X é aquele que minimiza a equação.

3.4.2 – Microssatélite *DYS503*

O loco *DYS503*, assim como o *DYS464*, é um conjunto de microssatélites presente em 4 cópias no braço longo do cromossomo Y. Este marcador foi primeiramente descrito em 2006 por Hanson e colaboradores. que, a fim de determinar sua diversidade, estudaram o *DYS503* em caucasianos e afro-americanos. Os resultados revelaram a maior diversidade já encontrada para euro-americanos (0,92) quando comparada com qualquer outro microssatélite de cópia única do cromossomo Y. Além disso, a diversidade para afro-americanos (0,75) foi menor que a encontrada em caucasianos. Este resultado incomum nos despertou a curiosidade acerca deste marcador e, a fim de estudar a distribuição deste em populações mundiais, analisamos o *DYS503* em 660 homens provenientes do Painel do CEPH-HGDP.

Ao contrário das análises com o *DYS464*, a análise dos eletroferogramas para o *DYS503* não apresentou, em nossas análises, alturas dos picos consistentes com quatro ampliações independentes (Figura 49). Assim, não foi possível realizar uma leitura em alta resolução, de forma que, para cada indivíduo, identificamos apenas os alelos presentes e não o número de cópias de cada alelo (análise em baixa resolução). Portanto, nossos resultados foram computados não mais em genótipos e sim em padrões alélicos encontrados. Em nossa análise, encontramos 10 diferentes alelos com tamanhos variando de 9 a 18 repetições.

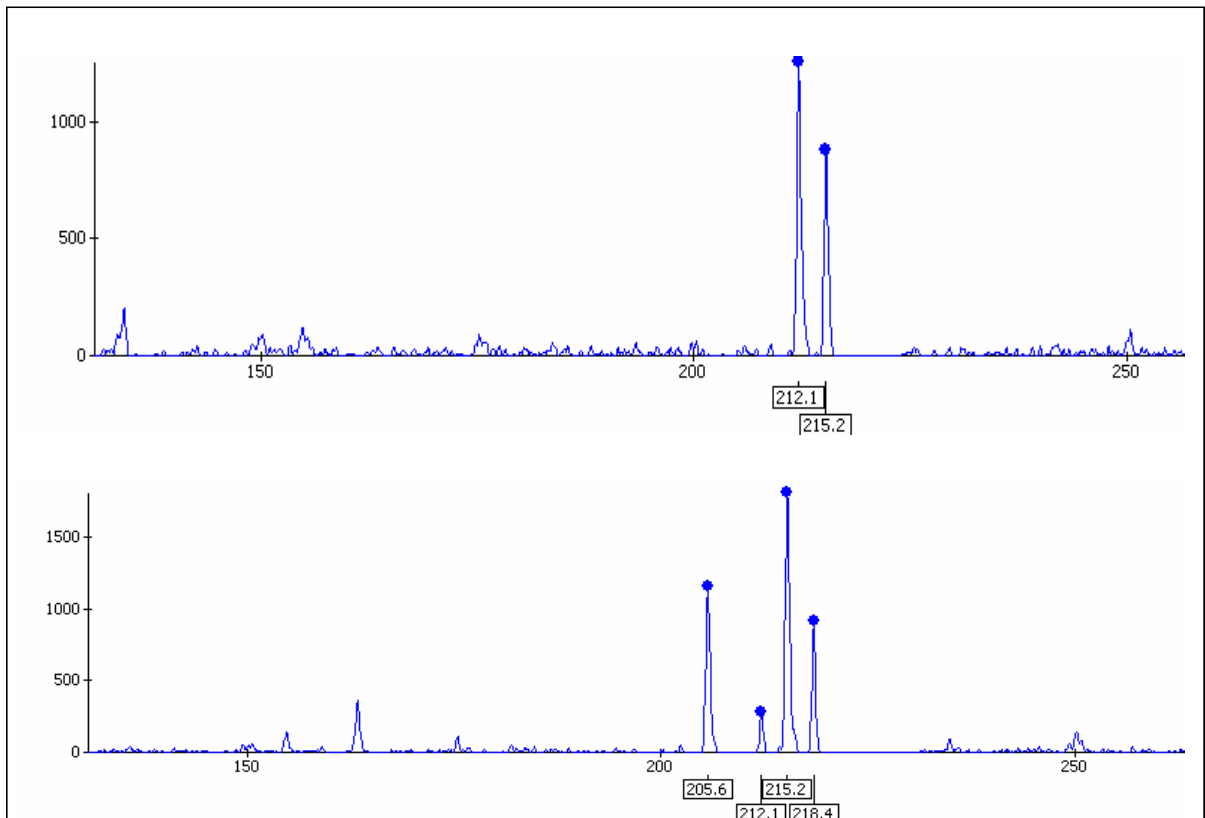


Figura 49: Eletroferograma produzido pelo programa *Fragment Profile*, versão 1.2, representando a genotipagem do conjunto de microssatélites *DYS503* pelo sequenciador MEGA BACE 1000. A leitura dos gráficos foi realizada em baixa resolução, onde foram determinados os padrões alélicos de cada indivíduo. O primeiro exemplo corresponde ao padrão alélico 13,14 e o segundo exemplo corresponde ao padrão alélico 11,13,14,15.

3.4.2.1- Análise de índices de diversidade

Após a determinação dos padrões alélicos do *DYS503*, iniciamos as análises dos índices de diversidade. Estes cálculos foram realizados manualmente sem auxílio de nenhum programa e as diversidades de padrões alélicos foram calculadas através da fórmula $(n/n-1).(1-\sum p^2)$, onde p corresponde a frequência de cada padrão alélico. Dos 10 diferentes alelos identificados, 6 (60%) foram encontrados na África, 7 (70%) no Oriente Médio, 6 (60%) na Ásia Central, 7 (70%) no Leste asiático, 4 (40%) na Oceania, 9 (90%) na Europa e 5 (50%) na América (Tabela 14). Dentre os alelos encontrados, um foi exclusivo do Leste Asiático (alelo 18) e um exclusivo da Europa (alelo 16). Como nossas análises foram feitas em baixa resolução, não foi possível calcular as frequências alélicas nas diferentes regiões.

Identificamos um total de 30 padrões alélicos, dos quais 11 foram encontrados na África, 13 no Oriente Médio, 13 na Ásia Central, 18 no Leste Asiático, 6 na Oceania, 19 na Europa e 7 na América (Tabela 15). Dos 30 padrões alélicos, 11 foram região-específicos, sendo a Europa a região geográfica com maior número (5) e maior porcentagem de padrões alélicos exclusivos (26,3%). Não foram encontrados padrões alélicos região-específicos na África e nem na Ásia Central. O padrão alélico mais frequente na África, Oriente Médio, Ásia Central, Leste Asiático, Oceania e Europa foi o 13,14 e na América foi o 12,13,14 (Tabela 16).

A diversidade de padrões alélicos foi maior na Europa (0,85) e menor na África e América (0,63 e 0,57, respectivamente).

Tabela 15: Número de indivíduos, alelos, alelos região-específicos, padrões alélicos, padrões alélicos região-específicos e diversidade de padrões alélicos encontrados em cada região geográfica e no total das amostras estudadas para o microsatélite *DYS503*.

Região geográfica	Nº indivíduos	Nº alelos	Nº de alelos região-específicos	Nº padrões alélicos	Nº de padrões alélicos região-específicos	Diversidade de padrões alélicos
África	97	6	0	11	0	0,63±0,052
Oriente Médio	73	7	0	13	2 (15,3%)	0,58±0,064
Ásia Central	172	6	0	13	0	0,73±0,027
Leste Asiático	172	7	1	18	2 (11,1%)	0,82±0,022
Oceania	18	4	0	6	1 (16,6%)	0,81±0,054
Europa	85	9	1	19	5 (26,3%)	0,85±0,024
América	43	5	0	7	1 (14,2%)	0,57±0,077
TOTAL	660	10	2	30	11	0,78

Tabela 16: Frequência de padrões alélicos encontradas em cada região geográfica para o conjunto de microssatélites *DYS503*

Padrões alélicos (<i>DYS503</i>)	África (n=97)		Oriente Médio (n=73)		Ásia Central (n=172)		Leste Asiático (n=172)		Oceania (n=18)		Europa (n=85)		América (n=43)	
	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>	<i>n</i>	<i>f</i>
13							6	0,034	4	0,222	1	0,011		
14	7	0,072	1	0,013	13	0,075	18	0,104			5	0,058		
11-13									1	0,055				
11-14	3	0,030			1	0,005	2	0,011			4	0,047		
12-13	1	0,010					10	0,058						
12-14	4	0,041	1	0,013	4	0,023	5	0,029			2	0,023		
13-14	57	0,587	46	0,630	78	0,453	63	0,366	6	0,333	25	0,294	4	0,093
13-15			1	0,013	5	0,029	4	0,023	2	0,111	1	0,011		
14-15	1	0,010	7	0,095	3	0,017	2	0,011	1	0,055				
9-11-13							1	0,005						
9-13-14	4	0,041			1	0,005								
10-12-15											1	0,011		
10-13-14			2	0,027										
11-12-14													1	0,023
11-13-14	1	0,010	1	0,013			2	0,011			7	0,082		
11-13-15			1	0,013										
11-14-15			1	0,013	3	0,017					17	0,2	1	0,023
11-14-16					0	0					2	0,023	0	0
12-13-14	13	0,134			22	0,127	17	0,098			1	0,011	27	0,627
12-13-15	3	0,030			2	0,011	2	0,011					1	0,023
12-14-15			1	0,013	2	0,011	4	0,023						
12-14-18							2	0,011						
13-14-15	3	0,030	9	0,123	35	0,203	25	0,145	4	0,222	8	0,094		
9-12-13-14											1	0,011		
10-13-14-17			1	0,013							1	0,011		
11-12-13-14							3	0,017			3	0,035	1	0,023
11-12-14-15											2	0,023		
11-13-14-15							2	0,011			1	0,011		
13-14-15-16											1	0,011		
12-13-14-15			1	0,013	3	0,017	4	0,023			2	0,023	8	0,186

3.5 – Discussão

3.5.1 – Microssatélite *DYS464* em populações mundiais

A utilização de microssatélites localizados no cromossomo Y vem sendo amplamente explorada em estudos de evolução humana já que esses marcadores são considerados polimórficos e o cromossomo Y consiste em um bloco haplóide que não sofre recombinação na maior parte de sua extensão. O marcador *DYS464* consiste em um conjunto de microssatélites presente em 4 cópias no braço longo do cromossomo Y (*DYS464a, b, c e d*) que foi considerado o mais polimórfico de todos os microssatélites já descritos (Redd *et al.*, 2002; Berger *et al.*, 2003). Embora o *DYS464* tenha sido estudado em algumas populações (Redd *et al.*, 2002; Berger *et al.*, 2003; Butler and Schoske, 2005) um estudo completo deste marcador em populações mundiais nunca havia sido realizado.

Neste trabalho, nós realizamos um amplo estudo do microssatélite *DYS464* através da análise deste marcador em 677 amostras de DNA de homens provenientes do painel do CEPH-HGDP que abrangem um total de 52 populações mundiais pertencentes à 7 grandes regiões geográficas.

Nossos resultados mostraram que, de fato, o *DYS464* é um marcador altamente polimórfico, dada a alta variabilidade (acima de 0,90, tabela 12) encontrada em todas as regiões geográficas estudadas. No entanto, as análises filogeográficas realizadas com o *DYS464* apresentaram resultados aparentemente estranhos: o genótipo mais freqüente na Ásia Central (12-15-15-16) foi também visto nos Bedouins de Israel (7,1%), Uygurs da China (12,5%), Sardinians da Itália (6,3%), Adygei da Europa (14,3%) e em russos (6,3%) (Anexo 1). Esta distribuição genotípica aparentemente “aleatória” não faz sentido evolucionário.

Estas inconsistências podem ser explicadas devido a algumas características do *DYS464* que dificultam sua utilização em inferências evolucionárias. Primeiro, como as sequências flangeadoras das quatro cópias do microssatélite são idênticas, não é possível distinguir, através do eletroferograma, qual é o alelo específico de cada loco. Por exemplo, no genótipo 12-13-14-15 não é possível identificar se o alelo 12 pertence ao loco *DYS464 a, b, c* ou *d*. Desta forma, o mesmo genótipo pode corresponder a vários haplótipos diferentes, com diferentes combinações de alelos. Desta maneira, chamamos os genótipos do *DYS464* de degenerados, fazendo uma analogia com o código genético onde o mesmo aminoácido é codificado por diferentes códons. Segundo, a estrutura palindrômica tetralocal da região onde o *DYS464* está localizado é susceptível a conversões alélicas mediadas por eventos de recombinação mitótica intracromossômica (Rozen *et al.*, 2003) o que também pode dificultar inferências evolucionárias.

Portanto, as distribuições genótípicas aparentemente aleatórias descritas acima podem fazer sentido se considerarmos que os Bedouins, Uygurs e os Sardinias apresentam, na realidade, diferentes haplótipos pertencentes a um mesmo genótipo e/ou que um ou mais eventos de conversão ocorreram levando a perda de heterozigosidade.

Por outro lado, muitas populações apresentaram altas frequências de genótipos específicos que podem, portanto, serem úteis em estudos evolucionários. O genótipo 14-14-15-15 foi visto em 88,9% dos Yakut da Sibéria (Anexo 1). Este dado sugere um evento de bottleneck ou efeito fundador no passado desta população como já sugerido por outros estudos (Pakendorf *et al.*, 2002; Khar'kov *et al.*, 2008).

A África apresentou a maior porcentagem de alelos e genótipos região-específicos (40%) (Tabela 12) o que está de acordo com a teoria de que a humanidade moderna emergiu na África e migrou para outros continentes sofrendo diversos efeitos fundadores e estreitos bottlenecks, resultando em uma diminuição da variabilidade (Yu *et al.*, 2002; Torroni *et al.*, 2006).

Entretanto, ao contrário do encontrado para a maioria dos outros marcadores do cromossomo Y (Jobling e Tyler-Smith, 2003), o *DYS464* não apresentou a maior diversidade genotípica na África. Este fato pode ser explicado pelo caráter

degenerado dos genótipos que faz com que haplótipos diferentes sejam considerados iguais, o que levaria a uma falsa diminuição da variabilidade. Além disso, estudos mostraram que o *DYS464* apresenta uma alta taxa de mutação ($2,86 \times 10^{-2}$), que corresponde a uma taxa 10 vezes maior que a média para microssatélites do cromossomo Y (Berger *et al.*, 2003). Esta alta taxa de mutação pode ter gerado homoplasia acarretando também na diminuição da variabilidade. Outro ponto importante diz respeito às amostras africanas presentes no painel do CEPH-HGDP que podem constituir apenas uma pequena porção do total da variabilidade da África.

A análise de variância molecular (AMOVA) revelou baixos níveis de estruturação genética: 88,42% da variabilidade genética foi encontrada dentro das populações, 9,62% entre populações dentro das regiões e 1,96% entre as grandes regiões geográficas (Tabela 13). Estes níveis de variabilidade genética encontrados dentro das populações são maiores quando comparados com outros estudos de populações mundiais realizados com outros marcadores do cromossomo Y: 76,8% (Kayser *et al.*, 2001) e 66,5% (Wilder *et al.*, 2004). Nós atribuímos esta diferença às altas taxas de mutação do *DYS464* que estaria gerando homoplasia e diminuindo a variabilidade entre populações.

Ainda em relação à AMOVA, se focarmos em cada região é possível observar que a variabilidade genética dentro das populações foi superior a 88% em todas as regiões geográficas estudadas, com exceção do Oriente Médio e Américas que

apresentaram respectivamente 21,78% e 33,20% da variabilidade genética entre populações (Tabela 13). Esta alta taxa de variabilidade genética entre populações encontrada nas Américas está de acordo com a história demográfica dos ameríndios, que sofreram fortes efeitos de deriva genética por constituírem pequenos grupos geograficamente separados. Isto levou a elevados níveis de variação genética entre as populações (Cavalli-Sforza *et al.*, 1994).

Os escalonamentos multidimensionais (MDS) realizados com as 52 populações mundiais com as 3 diferentes medidas de distância (F_{st} , R_{st} e $\delta\mu^2$) não mostraram estruturação populacional evidente (figuras 23, 24 e 25). Da mesma maneira, análise dos componentes principais (PCA) realizadas tanto para dados de freqüências genotípicas (figura 29) quanto alélicas (figura 30) também não foram informativos. Estes dados estão de acordo com os resultados da AMOVA que revelaram baixos níveis de estruturação genética (1,96% da variabilidade genética entre as grandes regiões).

Quando a análise foi feita para as grandes regiões geográficas, o MDS mais condizente com a história evolucionária do homem foi encontrado quando a medida de distância $\delta\mu^2$ foi utilizada: África mais isolada, América também mais distante dos outros continentes sendo o Leste Asiático o mais próximo, Europa próxima do Oriente Médio (que fazem parte da Eurásia) (Figura 28). Isto pode sugerir que a mutação do *DYS464* tem uma forte influência no processo evolutivo deste marcador, uma vez que esta medida de distância genética leva em

consideração os passos mutacionais dos microssatélites e o *DYS464* apresenta uma alta taxa de mutação (Berger *et al.*, 2003).

Nas análises dos componentes principais (PCAs) realizadas com as 7 regiões geográficas (Figuras 39 e 40), encontramos resultados mais condizentes com a história evolutiva do homem moderno quando dados de frequências alélicas foram utilizados (Figura 40): África mais distante dos demais continentes; Américas próximas do Leste Asiático; Europa e Ásia Central juntas e próximas do Oriente Médio constituindo a Eurásia e a Oceania mais isolada. Quando utilizamos dados de frequências genotípicas, os resultados do PCA se mostraram confusos: Leste Asiático isolado dos demais continentes e Oriente Médio, Américas e Oceania juntos (Figura 39). Esta confusão pode ser explicada pelo fato dos genótipos do *DYS464* serem degenerados. Assim, diferentes haplótipos são representados pelo mesmo genótipo levando a uma diminuição da resolução da análise acarretando neste resultado aparentemente ilógico. Portanto, nossos resultados sugerem que análises com o *DYS464* sejam mais adequadas em nível alélico do que em nível genotípico.

Estudos sugerem que a diferenciação de polimorfismos de microssatélites em populações humanas é regida principalmente por deriva genética, mais do que por eventos de mutação (Perez-Lezaun *et al.*, 1997). Portanto, levando-se em consideração o tamanho efetivo do cromossomo Y que é cerca de um quarto dos autossomos, a deriva genética deve ter, em teoria, um forte efeito na variação do

DYS464. No entanto, nossos resultados de AMOVA e MDS (dados de $\delta\mu^2$ foram mais adequados) indicaram que eventos de mutação têm uma forte influência na evolução do *DYS464*. Enquanto por um lado a deriva genética estaria levando a diferenciação entre as populações, as altas taxas de mutação do *DYS464* estariam produzindo homoplasia e gerando uma diminuição da variação entre populações.

Juntos, nossos resultados indicam que embora o *DYS464* apresente algumas complexidades provenientes de sua estrutura tetra-local que não permitem detalhadas inferências acerca da filogeografia humana, este conjunto de microssatélites apresenta uma alta variabilidade genética em todas as populações humanas estudadas (diversidade genotípica acima de 0,90) e pode, portanto, ser uma ferramenta útil em estudos populacionais e evolutivos.

3.5.2 - Microssatélite *DYS464* em populações miscigenadas

Além de estudar o microssatélite *DYS464* em populações mundiais nativas do painel do CEPH-HGDP, fizemos uma análise deste marcador também em populações miscigenadas. Nesta análise nosso objetivo era avaliar a utilização de frequências alélicas do *DYS464* em análises de ancestralidade populacional.

Para realização deste estudo nós analisamos o *DYS464* em populações de pretos e brancos da cidade de São Paulo e utilizamos dados disponíveis na literatura para

caucasianos e afro-americanos. Como referência de populações parentais das populações miscigenadas estudadas, utilizamos as populações africanas, européias e ameríndias do painel do CEPH-HGDP.

Nossos resultados de escalonamento multidimensional realizados com as populações miscigenadas e parentais não foram diferentes do esperado para nenhuma das 3 medidas de distância genética utilizadas (F_{st} , R_{st} e $\delta\mu^2$). Nas três análises encontramos ameríndios europeus e africanos bem separados, caucasianos e brancos brasileiros próximos entre si e africanos sempre mais próximos dos afro-americanos do que dos pretos brasileiros (Figuras 41, 42 e 43). Estes dados estão de acordo com dados de ancestralidade do cromossomo Y encontrados para brasileiros pretos, afro-americanos, brasileiros brancos e caucasianos.

Estudos onde a determinação de haplogrupos ancestrais foi realizada para pretos brasileiros revelaram que cerca de 48% dos cromossomos Y desta população apresentavam ancestralidade africana (Gonçalves *et al.*, 2008). Já a população afro-americana apresenta, para dados do cromossomo Y, um grau de miscigenação menor quando comparado com os pretos brasileiros: 28,46% de ancestralidade européia (Lind *et al.*, 2007). Se considerarmos que a população afro-americana foi formada da miscigenação entre europeus e africanos, podemos deduzir que 71,54% dos cromossomos Y de afro-americanos apresentam ancestralidade africana. Desta maneira, é de se esperar que os africanos estejam mais próximos

dos afro-americanos do que dos pretos brasileiros, conforme observado em nossas análises de MDS.

Para brancos brasileiros e caucasianos as porcentagens de ancestralidade européia encontradas para o cromossomo Y em estudos prévios foram de 98% (Carvalho-Silva *et al.*, 2001) e 96,6% (Hammer *et al.*, 2006), respectivamente. Estes dados indicam uma forte contribuição masculina européia para a formação tanto da população de brancos brasileiros como de euro-americanos e estão de acordo com os nossos resultados de MDS onde brancos brasileiros e caucasianos aparecem sempre próximos entre si (Figuras 41,42 e 43).

A análise dos componentes principais (PCA) realizada com frequências alélicas do *DYS464* apresentou resultados um pouco diferentes dos encontrados com o MDS. Dados de frequências alélicas não foram capazes de separar as populações parentais já que ameríndios e europeus aparecem próximos entre si (Figura 44). Isto pode ser explicado pelo fato do alelo mais freqüente em ambas as populações ser o mesmo (alelo 15, tabela 14). Além disso, os caucasianos aparecem distantes dos brancos brasileiros e europeus. No entanto, os brancos brasileiros encontram-se mais próximos dos europeus do que os pretos de São Paulo e afro-americanos dos africanos, evidenciando que estas duas últimas populações são mais miscigenadas do que os brancos brasileiros.

Embora as análises de MDS e PCA possam ser informativas acerca do grau de mistura das populações miscigenadas, estas análises fornecem informações mais qualitativas do que quantitativas. Por isso, procuramos validar as frequências alélicas do microssatélite *DYS464* em estudos de ancestralidade em nível quantitativo. Para isto calculamos as proporções de mistura africana nas populações miscigenadas utilizando dados de frequências alélicas do *DYS464* através do método gráfico, baseado em mínimos quadrados, descrito por Collins-Schramm e colaboradores. (2002) e comparamos nossos resultados com as proporções de ancestralidade africana já conhecidas nestas populações, determinadas pela análise de haplogrupos ancestrais do cromossomo Y.

Através da utilização de frequências alélicas do *DYS464*, uma proporção de ancestralidade africana de 53% foi encontrada para os pretos brasileiros, 1,5% para os brancos, 77% para os afro-americanos e 0% para os caucasianos. Estes valores estão bem próximos dos encontrados em estudos onde haplogrupos ancestrais do cromossomo Y foram determinados nestas populações.

Para os pretos brasileiros a determinação dos haplogrupos ancestrais do cromossomo Y foi previamente realizada por nosso grupo exatamente nas mesmas amostras utilizadas no presente estudo. Dos 124 indivíduos pretos brasileiros analisados neste estudo, 69 apresentaram haplogrupos africanos, ou seja, 55,6% das amostras analisadas apresentaram cromossomo Y de origem africana. Este dado é muito próximo do valor encontrado quando utilizamos frequências alélicas

do *DYS464* (53%). Para as demais populações miscigenadas os resultados não foram diferentes. Conforme mencionado anteriormente a proporção de cromossomos Y de origem africana em afro-americanos descrita na literatura é de 71,54% (Lind *et al.*, 2007) *versus* 77% encontrado com dados do *DYS464*; para brancos brasileiros 2,00% (Carvalho-Silva *et al.*, 2001) *versus* 1,50% encontrado com dados do *DYS464*; para caucasianos 3,40% (Hammer *et al.*, 2006) *versus* 0,00% encontrado com dados do *DYS464*.

Juntos, nossos resultados indicam que as freqüências alélicas do microssatélite *DYS464* podem ser informativas acerca da análise de ancestralidade, ao menos africana, de populações miscigenadas.

3.5.3 – Microssatélite *DYS503* em populações mundiais

Assim como o *DYS464*, o *DYS503* constitui um conjunto de microssatélites também presente em 4 cópias no braço longo do cromossomo Y. Nosso interesse acerca deste marcador surgiu pelo fato deste marcador ter apresentado diversidade genotípica em caucasianos (0,92) bem maior do que em afro-americanos (0,75) (Hanson *et al.*, 2007). Estes dados nos levaram às seguintes perguntas: Seria a diversidade em africanos menor do que a em europeus? Se sim, que processos evolutivos poderiam estar levando a este fato incomum?

Para responder estas perguntas estudamos o *DYS503* em 660 amostras de homens pertencentes a 52 populações mundiais distribuídas em 7 grandes regiões geográficas provenientes do painel do CEPH-HGDP. Em nossas análises não foi possível determinar os genótipos de cada indivíduo pois as alturas de picos presentes nos eletroferogramas não foram compatíveis com 4 cópias como o esperado. Este fato limitou bastante nossas análises de forma que só foi possível determinar quais alelos estavam presentes em cada indivíduo e não o número de cópias que cada um dos alelos apareciam. Desta maneira, todas as nossas análises foram realizadas com dados do que chamamos de padrões alélicos e não genótipos.

Nossos resultados indicaram uma diversidade de padrões alélicos maior na Europa do que em qualquer outro continente (0,85, tabela 15). A diversidade de padrões alélicos na África foi a terceira menor (0,63) sendo mais alta apenas do que os valores encontrados para as Américas (0,57) e Oriente Médio (0,58). Estes dados respondem nossa primeira pergunta e estão de acordo com o encontrado por Hanson e colaboradores, que observaram uma menor diversidade genotípica em afro-americanos quando comparada com caucasianos.

Além disso, a Europa também foi o continente que apresentou o maior número de padrões alélicos região-específicos (26,3%, Tabela 15) enquanto que a África não apresentou nenhum padrão alélico exclusivo. Estes dados reforçam ainda mais a natureza altamente diversificada deste marcador na Europa e pouco diversificada na África. Este dado é bastante incomum e contrário ao encontrado para outros

marcadores do cromossomo Y (Jobling e Tyler-Smith, 2003) que apresentam uma grande diversidade em africanos, o que é compatível com a origem do homem neste continente

Quais processos evolutivos poderiam explicar o incomum resultado de menor diversidade na África?

A primeira coisa que nos vem à cabeça para explicar a baixa diversidade do *DYS503* em africanos seria a ocorrência de seleção natural atuando sobre este loco neste continente. No entanto, levando-se em consideração que o cromossomo Y trata-se de um bloco haplóide que não sofre recombinação na maior parte de sua extensão e é, portanto, herdado em bloco, podemos excluir a possibilidade de eventual seleção natural atuando sobre este loco em africanos. Se houvesse força seletiva atuando sobre este loco, esta se refletiria em todos os outros locos do cromossomo Y africano, já que este não sofre recombinação. Sendo assim, a baixa diversidade em africanos deveria ser encontrada quando qualquer outro marcador fosse analisado, o que de fato não acontece.

Uma outra hipótese que poderia explicar a menor diversidade do *DYS503* na África é a possibilidade de na realidade o *DYS503* constituir um polimorfismo de número de cópias e não aparecer somente em quatro cópias nos africanos. Desta maneira, considerando o processo de dupla duplicação para gerar quatro cópias de um

mesmo loco (Rozen *et al.*, 2003), na África existiriam indivíduos com uma, duas e quatro cópias deste marcador. É bem conhecido que apenas um pequeno número de pessoas migrou da África para povoar os outros continentes. Se considerarmos que apenas indivíduos com quatro cópias saíram da África para os outros continentes não é de se estranhar que a diversidade na África seja menor que no restante do mundo, já que na África existiria indivíduos com uma, duas e quatro cópias enquanto que no restante do mundo apenas indivíduos com quatro cópias.

Esta hipótese seria bem compatível com os dados de frequências de padrões alélicos encontrados neste estudo. A tabela 16 indica que o padrão alélico mais freqüente na África é o 13-14 (58,7%). Dentro da nossa hipótese, isto poderia significar que este padrão alélico represente dois genótipos: 13-14 (indivíduos com apenas 2 cópias) e 13-13-14-14 (indivíduos com quatro cópias). No entanto, esta hipótese foi testada em nosso laboratório através da utilização de PCR quantitativa e os resultados não apontaram para evidência da existência de variação do número de cópias em africanos (Ribeiro, 2010).

Outro ponto que deve ser levado em consideração e que já foi mencionado no item 3.5.1 é o fato das amostras africanas do painel do CEPH-HGDP poderem não representar toda a variabilidade genética deste continente levando a uma medida subestimada da real diversidade na África.

Um outro aspecto importante que poderia explicar os resultados de baixa diversidade do *DYS503* na África seria a existência de polimorfismos de base única (SNPs) em africanos nas regiões de anelamento dos iniciadores. Isto poderia levar a um mal pareamento dos iniciadores no momento da PCR, acarretando na ausência de amplificação de determinados locos, o que levaria a uma subestimada variabilidade genética.

Uma outra possível explicação seria uma taxa de mutação diferencial do *DYS503* em africanos quando comparados com europeus. Alelos de microssatélites com números maiores de repetições apresentam uma taxa de mutação maior do que a encontrada para alelos de tamanho menor. Desta forma, se os africanos apresentarem altas frequências de alelos menores para o *DYS503*, a taxa de mutação deste loco em africanos poderia ser menor do que em europeus. Isso levaria a uma menor diversidade do marcador em africanos. No entanto, esta hipótese não pode ser testada no presente trabalho, uma vez que não conseguimos estimar o número de cópias de cada alelo do *DYS503* encontradas em cada amostra estudada, não sendo possível, portanto, calcular suas as frequências alélicas do microssatélite nas diferentes regiões geográficas.

Nossos resultados mostraram que de fato a diversidade de padrões alélicos do *DYS503* na África é menor do que na maioria dos outros continentes. Embora algumas hipóteses tenham sido levantadas para explicar este resultado incomum,

análises que testem estas possibilidades devem ser realizadas para que consigamos entender melhor os processos evolutivos que levaram a este dado.

3.6 – Conclusões

As análises dos microssatélites teralocais do cromossomo Y, *DYS464* e *DYS503* em populações mundiais do painel do CEPH-HGDP e populações miscigenadas nos levaram a concluir que:

- O microssatélite *DYS464* é altamente polimórfico, apresentando diversidade genotípica acima de 0,90 em todas as regiões geográficas estudadas;
- Algumas populações apresentaram altas frequências de genótipos específicos para o *DYS464*, o que valida sua utilização em estudos evolutivos;
- Embora a deriva genética seja a principal força evolutiva que atua sobre os microssatélites do cromossomo Y, a taxa de mutação parece exercer uma forte influência na evolução do *DYS464*;
- Embora o *DYS464* apresente algumas complexidades provenientes de sua estrutura tetra-local que não permitem detalhadas inferências acerca da filogeografia humana, este conjunto de microssatélites apresenta uma alta variabilidade genética em todas as populações humanas estudadas e pode, portanto, ser uma ferramenta útil em estudos populacionais e evolutivos.

- A utilização do programa *Autotet* para o cálculo das frequências alélicas do *DYS464* se mostrou uma estratégia útil e adequada;
- As frequências alélicas do microssatélite *DYS464* se mostraram úteis na análise de proporções ancestrais de mistura em populações miscigenadas;
- A diversidade de padrões alélicos do *DYS503* é maior em europeus e menor em africanos;
- Mais estudos com o *DYS503* são necessários para determinar os processos evolutivos que estão levando a um dado incomum de uma baixa diversidade deste marcador em africanos;

4 – Referências Bibliográficas

Referências Bibliográficas

- ALVES-SILVA, J., DA SILVA SANTOS, M., GUIMARAES, P. E., FERREIRA, A. C., BANDELT, H. J., PENA, S. D. & PRADO, V. F. 2000. The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet*, 67, 444-61.
- BARBUJANI, G., MAGAGNI, A., MINCH, E. & CAVALLI-SFORZA, L. L. 1997. An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A*, 94, 4516-9.
- BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21, 263-5.
- BASTOS-RODRIGUES, L., PIMENTA, J. R. & PENA, S. D. 2006. The genetic structure of human populations studied through short insertion-deletion polymorphisms. *Ann Hum Genet*, 70, 658-65.
- BEDOYA, G., MONTOYA, P., GARCIA, J., SOTO, I., BOURGEOIS, S., CARVAJAL, L., LABUDA, D., ALVAREZ, V., OSPINA, J., HEDRICK, P. W. & RUIZ-LINARES, A. 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proc Natl Acad Sci U S A*, 103, 7234-9.
- BERGER, B., NIEDERSTATTER, H., BRANDSTATTER, A. & PARSON, W. 2003. Molecular characterization and Austrian Caucasian population data of the multi-copy Y-chromosomal STR DYS464. *Forensic Sci Int*, 137, 221-30.
- BROWN, W. M. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci U S A*, 77, 3605-9.

- BUSCHIAZZO, E. & GEMMELL, N. J. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, 28, 1040-50.
- BUTLER, J. M. & SCHOSKE, R. 2005. U.S. population data for the multi-copy Y-STR locus DYS464. *J Forensic Sci*, 50, 975-7.
- CANN, R. L., STONEKING, M. & WILSON, A. C. 1987. Mitochondrial DNA and human evolution. *Nature*, 325, 31-6.
- CANN, H. M., 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, 261-2.
- CARVALHO-SILVA, D. R., SANTOS, F. R., ROCHA, J. & PENA, S. D. 2001. The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet*, 68, 281-6.
- CAVALLI-SFORZA, L. L., WILSON, A. C., CANTOR, C. R., COOK-DEEGAN, R. M. & KING, M. C. 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics*, 11, 490-1.
- CAVALLI-SFORZA, L. L. , MENOZZI, P., PIAZZA, A. 1994. The History and Geography of Human Genes. Princeton: *Princeton University Press*.
- CAVALLI-SFORZA, L. L. 1998. The DNA revolution in population genetics. *Trends Genet*, 14, 60-5.

- CAVALLI-SFORZA, L. L. & FELDMAN, M. W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet*, 33 Suppl, 266-75.
- CHIARONI, J., UNDERHILL, P. A. & CAVALLI-SFORZA, L. L. 2009. Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc Natl Acad Sci USA*, 106, 20174-9.
- COLLINS-SCHRAMM, H. E., CHIMA, B., OPERARIO, D. J., CRISWELL, L. A. & SELDIN, M. F. 2003. Markers informative for ancestry demonstrate consistent megabase-length linkage disequilibrium in the African American population. *Hum Genet*, 113, 211-9.
- COLLINS-SCHRAMM, H. E., PHILLIPS, C. M., OPERARIO, D. J., LEE, J. S., WEBER, J. L., HANSON, R. L., KNOWLER, W. C., COOPER, R., LI, H. & SELDIN, M. F. 2002. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am J Hum Genet*, 70, 737-50.
- COOPER, D. N., SMITH, B. A., COOKE, H. J., NIEMANN, S. & SCHMIDTKE, J. 1985. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum Genet*, 69, 201-5.
- CURTIN, P. D. 1969. *The Atlantic slave trade: a census*. Madison: University of Wisconsin Press.
- DARVASI, A. & SHIFMAN, S. 2005. The beauty of admixture. *Nat Genet*, 37, 118-9.
- DE KNIJFF, P., KAYSER, M., CAGLIA, A., CORACH, D., FRETWELL, N., GEHRIG, C., GRAZIOSI, G., HEIDORN, F., HERRMANN, S., HERZOG, B., HIDDING, M., HONDA, K., JOBLING, M., KRAWCZAK, M., LEIM, K., MEUSER, S., MEYER, E., OESTERREICH, W., PANDYA, A., PARSON, W., PENACINO, G., PEREZ-

- LEZAUN, A., PICCININI, A., PRINZ, M., ROEWER, L. & ET AL. 1997. Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med*, 110, 134-49.
- DUPANLOUP, I. & BERTORELLE, G. 2001. Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol*, 18, 672-5.
- ELLEGREN, H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet*, 16, 551-8.
- EXCOFFIER, L., SMOUSE, P. E. & QUATTRO, J. M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131, 479-91.
- FALUSH, D., STEPHENS, M. & PRITCHARD, J. K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567-87.
- FORSTER, P. & MATSUMURA, S. 2005. Evolution. Did early humans go north or south? *Science*, 308, 965-6.
- FREYRE, G. 1933. Casa-Grande & Senzala. Rio de Janeiro: Editora Record.
- GABRIEL, S. B., SCHAFFNER, S. F., NGUYEN, H., MOORE, J. M., ROY, J., BLUMENSTIEL, B., HIGGINS, J., DEFELICE, M., LOCHNER, A., FAGGART, M., LIU-CORDERO, S. N., ROTIMI, C., ADEYEMO, A., COOPER, R., WARD, R., LANDER, E. S., DALY, M. J. & ALTSHULER, D. 2002. The structure of haplotype blocks in the human genome. *Science*, 296, 2225-9.

- GOLDSTEIN, D. B. & CHIKHI, L. 2002. Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet*, 3, 129-52.
- GONCALVES, V. F., CARVALHO, C. M., BORTOLINI, M. C., BYDLOWSKI, S. P. & PENA, S. D. 2008. The phylogeography of African Brazilians. *Hum Hered*, 65, 23-32.
- GOULART, M. 1975. Escravidão africana no Brasil: das origens à extinção do tráfico. São Paulo: Alfa-ômega.
- GUSMAO, L., BUTLER, J. M., CARRACEDO, A., GILL, P., KAYSER, M., MAYR, W. R., MORLING, N., PRINZ, M., ROEWER, L., TYLER-SMITH, C. & SCHNEIDER, P. M. 2006. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Int J Legal Med*, 120, 191-200.
- HAMMER, M. F. 1994. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol*, 11, 749-61.
- HAMMER, M. F., CHAMBERLAIN, V. F., KEARNEY, V. F., STOVER, D., ZHANG, G., KARAFET, T., WALSH, B. & REDD, A. J. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int*, 164, 45-55.
- HANSON, E. K. & BALLANTYNE, J. 2006. Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Leg Med (Tokyo)*, 8, 110-20.
- HANSON, E. K. & BALLANTYNE, J. 2007. Population data for a novel, highly discriminating tetra-local Y-chromosome short tandem repeat: DYS503. *J Forensic Sci*, 52, 498-9.

JEHAES, E. 1999. Optimisation of methods and procedures for the analysis of mtDNA sequences and their application in molecular archaeological and historical finds. Leuven: Leuven university press.

JOBLING, M. A., SAMARA, V., PANDYA, A., FRETWELL, N., BERNASCONI, B., MITCHELL, R. J., GERELSAIKHAN, T., DASHNYAM, B., SAJANTILA, A., SALO, P. J., NAKAHORI, Y., DISTECHE, C. M., THANGARAJ, K., SINGH, L., CRAWFORD, M. H. & TYLER-SMITH, C. 1996. Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Hum Mol Genet*, 5, 1767-75.

JOBLING, M. A. & TYLER-SMITH, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, 4, 598-612.

JORDE, L. B., WATKINS, W. S., BAMSHAD, M. J., DIXON, M. E., RICKER, C. E., SEIELSTAD, M. T. & BATZER, M. A. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet*, 66, 979-88.

KAYSER, M., CAGLIA, A., CORACH, D., FRETWELL, N., GEHRIG, C., GRAZIOSI, G., HEIDORN, F., HERRMANN, S., HERZOG, B., HIDDING, M., HONDA, K., JOBLING, M., KRAWCZAK, M., LEIM, K., MEUSER, S., MEYER, E., OESTERREICH, W., PANDYA, A., PARSON, W., PENACINO, G., PEREZ-LEZAUN, A., PICCININI, A., PRINZ, M., SCHMITT, C., ROEWER, L. & ET AL. 1997. Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med*, 110, 125-33, 141-9.

KAYSER, M., KRAWCZAK, M., EXCOFFIER, L., DIELTJES, P., CORACH, D., PASCALI, V., GEHRIG, C., BERNINI, L. F., JESPERSEN, J., BAKKER, E., ROEWER, L. & DE KNIJFF, P. 2001. An extensive analysis of Y-chromosomal microsatellite

haplotypes in globally dispersed human populations. *Am J Hum Genet*, 68, 990-1018.

KHAR'KOV, V. N., STEPANOV, V. A., MEDVEDEV, O. F., SPIRIDONOVA, M. G., MAKSIMOVA, N. R., NOGOVITSYNA, A. N. & PUZYREV, V. P. 2008. [The origin of Yakuts: analysis of Y-chromosome haplotypes]. *Mol Biol (Mosk)*, 42, 226-37.

KRUSKAL, J. B. & WISH, M. 1978. *Multidimensional Scaling*. SAGE Publications, Nova York.

KURODA-KAWAGUCHI, T., SKALETSKY, H., BROWN, L. G., MINX, P. J., CORDUM, H. S., WATERSTON, R. H., WILSON, R. K., SILBER, S., OATES, R., ROZEN, S. & PAGE, D. C. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet*, 29, 279-86.

LEWONTIN, R. C. 1972. The Apportionment of Human Diversity. *Evolutionary Biology*, vol. 6, pp. 391-398

LIND, J. M., HUTCHESON-DILKS, H. B., WILLIAMS, S. M., MOORE, J. H., ESSEX, M., RUIZ-PESINI, E., WALLACE, D. C., TISHKOFF, S. A., O'BRIEN, S. J. & SMITH, M. W. 2007. Elevated male European and female African contributions to the genomes of African American individuals. *Hum Genet*, 120, 713-22.

MACAULAY, V., HILL, C., ACHILLI, A., RENGO, C., CLARKE, D., MEEHAN, W., BLACKBURN, J., SEMINO, O., SCOZZARI, R., CRUCIANI, F., TAHA, A., SHAARI, N. K., RAJA, J. M., ISMAIL, P., ZAINUDDIN, Z., GOODWIN, W., BULBECK, D., BANDELT, H. J., OPPENHEIMER, S., TORRONI, A. & RICHARDS, M. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308, 1034-6.

- MATHIAS, N., BAYES, M. & TYLER-SMITH, C. 1994. Highly informative compound haplotypes for the human Y chromosome. *Hum Mol Genet*, 3, 115-23.
- MILLS, R. E., LUTTIG, C. T., LARKINS, C. E., BEAUCHAMP, A., TSUI, C., PITTARD, W. S. & DEVINE, S. E. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*, 16, 1182-90.
- NEI, M. & ROYCHOUDHURY, A. K. 1974. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am J Hum Genet*, 26, 421-43.
- NEI, M. & ROYCHOUDHURY, A. K. 1993. Evolutionary relationships of human populations on a global scale. *Mol Biol Evol*, 10, 927-43.
- PAKENDORF, B., MORAR, B., TARSKAIA, L. A., KAYSER, M., SOODYALL, H., RODEWALD, A. & STONEKING, M. 2002. Y-chromosomal evidence for a strong reduction in male population size of Yakuts. *Hum Genet*, 110, 198-200.
- PAKENDORF, B. & STONEKING, M. 2005. Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet*, 6, 165-83.
- PARRA, E. J., MARCINI, A., AKEY, J., MARTINSON, J., BATZER, M. A., COOPER, R., FORRESTER, T., ALLISON, D. B., DEKA, R., FERRELL, R. E. & SHRIVER, M. D. 1998. Estimating African American admixture proportions by use of population-specific alleles. *Am J Hum Genet*, 63, 1839-51.
- PARRA, F. C., AMADO, R. C., LAMBERTUCCI, J. R., ROCHA, J., ANTUNES, C. M. & PENA, S. D. 2003. Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci U S A*, 100, 177-82.

- PENA, S. D. J. (org.). 2002. *Homo brasilis: Aspectos Genéticos, Linguísticos, Históricos e Socioantropológicos da Formação do Povo Brasileiro*. Ribeirão Preto: FUNPEC.
- PENA, S. D., BASTOS-RODRIGUES, L., PIMENTA, J. R. & BYDLOWSKI, S. P. 2009. DNA tests probe the genomic ancestry of Brazilians. *Braz J Med Biol Res*, 42, 870-6.
- PEREZ-LEZAUN, A., CALAFELL, F., SEIELSTAD, M., MATEU, E., COMAS, D., BOSCH, E. & BERTRANPETIT, J. 1997. Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol*, 45, 265-70.
- PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155, 945-59.
- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J., SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. Sep;81(3):559-75. Epub 2007 Jul 25.
- REDD, A. J., AGELLON, A. B., KEARNEY, V. A., CONTRERAS, V. A., KARAFET, T., PARK, H., DE KNIJFF, P., BUTLER, J. M. & HAMMER, M. F. 2002. Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int*, 130, 97-111.
- REED, F. A. & TISHKOFF, S. A. 2006. African human diversity, origins and migrations. *Curr Opin Genet Dev*, 16, 597-605.
- RIBEIRO, D. 1995. O povo brasileiro. A formação e o sentido do Brasil. São Paulo: Editora Scwarcz LTDA.

- REYNOLDS, J., WEIR, B. S. & COCKERHAM, C. C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105, 767-79.
- RIBEIRO, L. S. Avaliação da possível variação do número de cópias do microssatélite tetralocal DYS503 em populações da África. Belo Horizonte: UFMG, 2010. 101 p. Dissertação (Mestrado) – Programa de Pós-Graduação em Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.
- ROEWER, L. & EPPLIN, J. T. 1992. Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. *Forensic Sci Int*, 53, 163-71.
- ROEWER, L., KRAWCZAK, M., WILLUWEIT, S., NAGY, M., ALVES, C., AMORIM, A., ANSLINGER, K., AUGUSTIN, C., BETZ, A., BOSCH, E., CAGLIA, A., CARRACEDO, A., CORACH, D., DEKAIRELLE, A. F., DOBOSZ, T., DUPUY, B. M., FUREDI, S., GEHRIG, C., GUSMAO, L., HENKE, J., HENKE, L., HIDDING, M., HOHOFF, C., HOSTE, B., JOBLING, M. A., KARGEL, H. J., DE KNIJFF, P., LESSIG, R., LIEBEHERR, E., LORENTE, M., MARTINEZ-JARRETA, B., NIEVAS, P., NOWAK, M., PARSON, W., PASCALI, V. L., PENACINO, G., PLOSKI, R., ROLF, B., SALA, A., SCHMIDT, U., SCHMITT, C., SCHNEIDER, P. M., SZIBOR, R., TEIFEL-GREDING, J. & KAYSER, M. 2001. Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int*, 118, 106-13.
- 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, 2381-5.
- ROZEN, S., SKALETSKY, H., MARSZALEK, J. D., MINX, P. J., CORDUM, H. S., WATERSTON, R. H., WILSON, R. K. & PAGE, D. C. 2003. Abundant gene

conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, 423, 873-6.

SACHIDANANDAM, R., WEISSMAN, D., SCHMIDT, S. C., KAKOL, J. M., STEIN, L. D., MARTH, G., SHERRY, S., MULLIKIN, J. C., MORTIMORE, B. J., WILLEY, D. L., HUNT, S. E., COLE, C. G., COGGILL, P. C., RICE, C. M., NING, Z., ROGERS, J., BENTLEY, D. R., KWOK, P. Y., MARDIS, E. R., YEH, R. T., SCHULTZ, B., COOK, L., DAVENPORT, R., DANTE, M., FULTON, L., HILLIER, L., WATERSTON, R. H., MCPHERSON, J. D., GILMAN, B., SCHAFFNER, S., VAN ET TEN, W. J., REICH, D., HIGGINS, J., DALY, M. J., BLUMENSTIEL, B., BALDWIN, J., STANGETHOMANN, N., ZODY, M. C., LINTON, L., LANDER, E. S. & ALTSHULER, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409, 928-33.

SALZANO, F. M., FREIRE-MAIA, N. 1970. Problems in human biology: a study of brazilian populations. Detroit: Wayne State University Press.

SANS, M. 2000. Admixture studies in Latin America: from the 20th to the 21st century. *Hum Biol*, 72, 155-77.

SANTOS, A. S. 2002. Historical roots of the "Whitening" of Brazil. *Lat Am Perspect*; 28: 61-82.

SANTOS, F. R., EPPLER, J. T., PENA, S. D. J. 2003. Testing deficiency paternity cases with an Y-linked tetranucleotide repeat polymorphism. *In: DNA Fingerprinting: State of the Science* (Pena, SDJ, Chakraborty, R, Eppler, JT and Jeffreys, AJ, eds.) Birkhäuser Verlag, Basel, 261-266.

SCHLOTTERER, C. 2004. The evolution of molecular markers--just a matter of fashion? *Nat Rev Genet*, 5, 63-9.

- SCHNEIDER, S., ROESSLI, D., EXCOFFIER, L. 2000. Arlequin ver. 2000: A software for population genetic data analysis. *Genetics and Biometry Laboratory, University of Geneva*.
- SELDIN, M. F., MORII, T., COLLINS-SCHRAMM, H. E., CHIMA, B., KITTLES, R., CRISWELL, L. A. & LI, H. 2004. Putative ancestral origins of chromosomal segments in individual african americans: implications for admixture mapping. *Genome Res*, 14, 1076-84.
- SEYFERTH, G. 1985. A antropologia e a teoria do branqueamento de raça no Brasil: a tese de João Batista de Lacerda. *Revista do Museu Paulista*; 30: 81-98.
- SHRIVER, M. D., SMITH, M. W., JIN, L., MARCINI, A., AKEY, J. M., DEKA, R. & FERRELL, R. E. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet*, 60, 957-64.
- SKALETSKY, H., KURODA-KAWAGUCHI, T., MINX, P. J., CORDUM, H. S., HILLIER, L., BROWN, L. G., REPPING, S., PYNTIKOVA, T., ALI, J., BIERI, T., CHINWALLA, A., DELEHAUNTY, A., DELEHAUNTY, K., DU, H., FEWELL, G., FULTON, L., FULTON, R., GRAVES, T., HOU, S. F., LATRIELLE, P., LEONARD, S., MARDIS, E., MAUPIN, R., MCPHERSON, J., MINER, T., NASH, W., NGUYEN, C., OZERSKY, P., PEPIN, K., ROCK, S., ROHLFING, T., SCOTT, K., SCHULTZ, B., STRONG, C., TINWOLLAM, A., YANG, S. P., WATERSTON, R. H., WILSON, R. K., ROZEN, S. & PAGE, D. C. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423, 825-37.
- STEPHENS, M., SMITH, N. J. & DONNELLY, P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68, 978-89.
- TANG, H., CORAM, M., WANG, P., ZHU, X. & RISCH, N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet*, 79, 1-12.

- TAUTZ, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res*, 17, 6463-71.
- THRALL, P. H. & YOUNG, A. 2000. AUTOTET: a program for analysis of autotetraploid genotypic data. *J Hered*, 91, 348-9.
- TISHKOFF, S. A. & WILLIAMS, S. M. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet*, 3, 611-21.
- TORRONI, A., ACHILLI, A., MACAULAY, V., RICHARDS, M. & BANDELT, H. J. 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet*, 22, 339-45.
- UNDERHILL, P. A., JIN, L., LIN, A. A., MEHDI, S. Q., JENKINS, T., VOLLRATH, D., DAVIS, R. W., CAVALLI-SFORZA, L. L. & OEFNER, P. J. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res*, 7, 996-1005.
- UNDERHILL, P. A. & KIVISILD, T. 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet*, 41, 539-64.
- UNDERHILL, P. A., SHEN, P., LIN, A. A., JIN, L., PASSARINO, G., YANG, W. H., KAUFFMAN, E., BONNE-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. *Nat Genet*, 26, 358-61.
- VAINFAS, R. 2000. História indígena: 500 anos de despovoamento. Rio de Janeiro: Instituto de Geografia Estatística – IBGE; p 37-59

- VALI, U., BRANDSTROM, M., JOHANSSON, M. & ELLEGREN, H. 2008. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet*, 9, 8.
- VIGNAL, A., MILAN, D., SANCRISTOBAL, M. & EGGEN, A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol*, 34, 275-305.
- WEBER, J. L., DAVID, D., HEIL, J., FAN, Y., ZHAO, C. & MARTH, G. 2002. Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet*, 71, 854-62.
- WILDER, J. A., KINGAN, S. B., MOBASHER, Z., PILKINGTON, M. M. & HAMMER, M. F. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat Genet*, 36, 1122-5.
- YU, N., CHEN, F. C., OTA, S., JORDE, L. B., PAMILO, P., PATTHY, L., RAMSAY, M., JENKINS, T., SHYUE, S. K. & LI, W. H. 2002. Larger genetic differences within africans than between Africans and Eurasians. *Genetics*, 161, 269-74.
- ZHANG, D. X. & HEWITT, G. M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol Ecol*, 12, 563-84.

5 - Anexos

Anexo 1 – Frequências genótípicas do *DYS464* encontradas em cada população e região geográfica do painel do CEPH-HGDP

Genótipo	Frequências genótípicas							
	Biaka (n=33)	Mbuti (n=15)	Mandenka (n=24)	Yoruba (n=25)	Bantu NE (n=12)	San (n=7)	Bantu SE (n=8)	África (n=98)
9-12-15-16								
10-11-14-16					0.091			0.01
11-11-11-15								
11-11-11-16								
11-11-12-16								
11-11-14-15								
11-12-12-12								
11-12-12-15								
11-12-12-16								
11-12-12-17								
11-12-12-18								
11-12-13-15								
11-12-13-16								
11-12-14-15								
11-12-16-18								
11-13-13-14								
11-13-15-16								
11-14-14-14								
11-14-14-15								
11-14-14-16								
11-14-15-15								
11-14-15-16								
11-14-15-18								
11-14-16-18								
11-14-17-17		0.077						0.01
11-15-15-16								
11-15-16-17								
11-16-16-16								
11-16-17-17					0.091			0.01
12-12-12-12								
12-12-12-13								
12-12-12-15								
12-12-13-13								
12-12-13-14								
12-12-14-14								
12-12-14-15								
12-12-14-16								
12-12-15-15								
12-12-15-18				0.077				0.01
12-12-16-16								

12-12-16-17								
12-12-17-17								
12-13-13-13								
12-13-13-14								
12-13-13-15								
12-13-13-16								
12-13-14-14								
12-13-14-15								
12-13-14-16								
12-13-14-17								
12-13-15-16						0.286		0.02
12-13-15-17								
12-13-15-18								
12-13-16-16								
12-14-14-14								
12-14-14-15								
12-14-14-16								
12-14-15-15								
12-14-15-16						0.143		0.01
12-14-15-17							0.125	0.01
12-14-16-16								
12-14-16-17								
12-14-16-18								
12-14-17-17								
12-15-15-15								
12-15-15-16								
12-15-15-17								
12-15-15-18			0.063					0.01
12-15-16-16						0.143		0.01
12-15-16-17				0.077				0.01
12-15-16-18	0.067		0.125					0.04
12-16-16-16								
12-16-16-17								
12-16-17-17								
12-17-17-17								
13-13-13-13								
13-13-13-14								
13-13-13-15								
13-13-13-16								
13-13-13-17								
13-13-14-14								
13-13-14-15								
13-13-14-16								
13-13-14-17								
13-13-15-15						0.143		0.01
13-13-15-16								

13-13-15-17								
13-13-16-16								
13-13-16-17			0.063					0.01
13-13-16-18				0.077				0.01
13-13-17-17								
13-13-17-18								
13-13-18-18	0.033							0.01
13-14-14-14						0.143		0.01
13-14-14-15								
13-14-14-16		0.231						0.03
13-14-14-17								
13-14-14-20						0.143		0.01
13-14-15-15								
13-14-15-16								
13-14-15-17								
13-14-15-18		0.077						0.01
13-14-15-19		0.154						0.02
13-14-16-16								
13-14-16-17								
13-14-16-18								
13-14-17-17								
13-14-17-18								
13-15-15-15								
13-15-15-16								
13-15-15-17								
13-15-15-18	0.033							0.01
13-15-16-16			0.063				0.125	0.02
13-15-16-17	0.033			0.077			0.250	0.04
13-15-16-18		0.154	0.125	0.077	0.091			0.061
13-15-17-18								
13-15-17-20				0.077				0.01
13-16-16-16			0.125	0.077				0.03
13-16-16-17	0.100		0.063					0.04
13-16-16-18	0.433	0.077	0.125	0.154	0.182			0.204
13-16-16-19					0.091		0.125	0.02
13-16-17-17			0.063					0.01
13-16-17-18		0.077						0.01
14-14-14-14	0.033							0.01
14-14-14-15								
14-14-14-16								
14-14-14-17								
14-14-15-15								
14-14-15-16					0.091			0.01
14-14-15-17	0.100							0.03
14-14-15-18								
14-14-16-16				0.077				0.01

14-14-16-17								
14-14-16-18								
14-14-16-19								
14-14-17-17								
14-15-15-15								
14-15-15-16								
14-15-15-17	0.067							0.02
14-15-15-18								
14-15-16-16								
14-15-16-17								
14-15-16-18					0.182			0.02
14-15-17-17					0.182			0.02
14-15-17-18								
14-16-16-16							0.125	0.01
14-16-16-17								
14-16-16-18			0.063					0.01
14-16-17-17				0.077				0.01
14-16-17-18							0.125	0.01
14-17-17-18							0.125	0.01
15-15-15-15								
15-15-15-16								
15-15-15-17								
15-15-15-18								
15-15-16-16								
15-15-16-17								
15-15-16-18								
15-15-16-19								
15-15-17-17			0.063					0.01
15-15-17-18								
15-15-17-19								
15-16-16-16								
15-16-16-17	0.100							0.03
15-16-16-18								
15-16-17-17								
15-16-17-18				0.077				0.01
15-17-17-17				0.077				0.01
16-16-16-16								
16-16-16-17								
16-16-17-17								
16-16-18-18								
16-17-17-17								
17-18-18-18		0.154						0.02
20-20-22-23			0.063					0.01

Anexo 1 – Continuação

Genótipos	Frequências genotípicas				
	Mozabite (n=30)	Bedouin (n=48)	Druze (n=47)	Palestinian (n=51)	Oriente Médio (n=79)
9-12-15-16					
10-11-14-16					
11-11-11-15					
11-11-11-16					
11-11-12-16					
11-11-14-15					
11-12-12-12					
11-12-12-15					
11-12-12-16					
11-12-12-17					
11-12-12-18					
11-12-13-15					
11-12-13-16					
11-12-14-15					
11-12-16-18					
11-13-13-14					
11-13-15-16					
11-14-14-14					
11-14-14-15					
11-14-14-16					
11-14-15-15					
11-14-15-16					
11-14-15-18					
11-14-16-18		0.036			0.012
11-14-17-17					
11-15-15-16					
11-15-16-17					
11-16-16-16					
11-16-17-17					
12-12-12-12					
12-12-12-13					
12-12-12-15					
12-12-13-13				0.118	0.025
12-12-13-14					
12-12-14-14	0.050				0.012
12-12-14-15					
12-12-14-16					
12-12-15-15					
12-12-15-18					
12-12-16-16					
12-12-16-17					

12-12-17-17					
12-13-13-13					
12-13-13-14			0.235		0.05
12-13-13-15	0.050				0.012
12-13-13-16					
12-13-14-14		0.036	0.471		0.113
12-13-14-15					
12-13-14-16					
12-13-14-17			0.071		0.012
12-13-15-16			0.071		0.012
12-13-15-17					
12-13-15-18					
12-13-16-16					
12-14-14-14					
12-14-14-15					
12-14-14-16					
12-14-15-15			0.143		0.025
12-14-15-16					
12-14-15-17		0.107			0.037
12-14-16-16					
12-14-16-17					
12-14-16-18		0.464			0.164
12-14-17-17					
12-15-15-15					
12-15-15-16		0.071			0.025
12-15-15-17					
12-15-15-18					
12-15-16-16					
12-15-16-17					
12-15-16-18					
12-16-16-16					
12-16-16-17					
12-16-17-17			0.059		0.012
12-17-17-17					
13-13-13-13					
13-13-13-14					
13-13-13-15					
13-13-13-16					
13-13-13-17					
13-13-14-14					
13-13-14-15					
13-13-14-16					
13-13-14-17					
13-13-15-15					
13-13-15-16					
13-13-15-17					

13-13-16-16					
13-13-16-17					
13-13-16-18					
13-13-17-17					
13-13-17-18					
13-13-18-18					
13-14-14-14			0.071		0.012
13-14-14-15					
13-14-14-16					
13-14-14-17			0.071		0.012
13-14-14-20					
13-14-15-15					
13-14-15-16					
13-14-15-17					
13-14-15-18					
13-14-15-19					
13-14-16-16					
13-14-16-17		0.036	0.286		0.063
13-14-16-18					
13-14-17-17					
13-14-17-18					
13-15-15-15					
13-15-15-16					
13-15-15-17					
13-15-15-18					
13-15-16-16					
13-15-16-17	0.050				0.012
13-15-16-18	0.050				0.012
13-15-17-18					
13-15-17-20					
13-16-16-16					
13-16-16-17					
13-16-16-18					
13-16-16-19					
13-16-17-17					
13-16-17-18					
14-14-14-14					
14-14-14-15					
14-14-14-16					
14-14-14-17					
14-14-15-15					
14-14-15-16					
14-14-15-17					
14-14-15-18					
14-14-16-16		0.107			0.037
14-14-16-17			0.071		0.012

14-14-16-18					
14-14-16-19					
14-14-17-17		0.036			0.012
14-15-15-15					
14-15-15-16					
14-15-15-17					
14-15-15-18					
14-15-16-16					
14-15-16-17					
14-15-16-18					
14-15-17-17					
14-15-17-18					
14-16-16-16	0.150				0.037
14-16-16-17	0.550		0.143		0.164
14-16-16-18	0.050				0.012
14-16-17-17					
14-16-17-18					
14-17-17-18					
15-15-15-15					
15-15-15-16					
15-15-15-17					
15-15-15-18					
15-15-16-16					
15-15-16-17					
15-15-16-18			0.071		0.012
15-15-16-19					
15-15-17-17					
15-15-17-18					
15-15-17-19					
15-16-16-16					
15-16-16-17	0.050				0.012
15-16-16-18				0.118	0.025
15-16-17-17		0.071			0.025
15-16-17-18					
15-17-17-17					
16-16-16-16					
16-16-16-17		0.036			0.012
16-16-17-17					
16-16-18-18					
16-17-17-17					
17-18-18-18					
20-20-22-23					

Anexo 1 – Continuação

Genótipo	Frequências genotípicas								Ásia Central (n=175)
	Brahui (n=25)	Baloshi (n=25)	Hazara (n=25)	Makrani (n=25)	Sindhi (n=25)	Pathan (n=24)	Kalash (n=25)	Burusho (n=25)	
9-12-15-16									
10-11-14-16									
11-11-11-15			0.042						0.005
11-11-11-16									
11-11-12-16			0.250			0.050			0.04
11-11-14-15								0.200	0.022
11-12-12-12									
11-12-12-15									
11-12-12-16									
11-12-12-17									
11-12-12-18									
11-12-13-15			0.042					0.050	0.011
11-12-13-16									
11-12-14-15									
11-12-16-18			0.042						0.005
11-13-13-14							0.200		0.022
11-13-15-16									
11-14-14-14									
11-14-14-15			0.042						0.005
11-14-14-16									
11-14-15-15		0.040							0.005
11-14-15-16							0.200	0.050	0.028
11-14-15-18									
11-14-16-18									
11-14-17-17									
11-15-15-16							0.050		0.005
11-15-16-17									
11-16-16-16		0.040							0.005
11-16-17-17									
12-12-12-12									
12-12-12-13									
12-12-12-15				0.050					0.005
12-12-13-13	0.040								0.005
12-12-13-14									
12-12-14-14									
12-12-14-15		0.080						0.100	0.022
12-12-14-16							0.200	0.050	0.028
12-12-15-15									
12-12-15-18									
12-12-16-16									
12-12-16-17									

12-12-17-17									
12-13-13-13									
12-13-13-14									
12-13-13-15			0.083						0.011
12-13-13-16		0.040							0.005
12-13-14-14									
12-13-14-15	0.040	0.040				0.050			0.017
12-13-14-16	0.040	0.040		0.050					0.017
12-13-14-17									
12-13-15-16				0.050		0.050			0.011
12-13-15-17									
12-13-15-18									
12-13-16-16	0.040								0.005
12-14-14-14	0.040								0.005
12-14-14-15	0.040					0.050			0.011
12-14-14-16	0.040				0.048				0.011
12-14-15-15				0.100	0.048			0.050	0.022
12-14-15-16	0.080		0.042	0.050	0.048	0.050			0.034
12-14-15-17									
12-14-16-16									
12-14-16-17	0.040	0.040							0.011
12-14-16-18									
12-14-17-17	0.080								0.011
12-15-15-15	0.040					0.050			0.011
12-15-15-16	0.160	0.160		0.150	0.238	0.300		0.100	0.137
12-15-15-17		0.040	0.083	0.050					0.022
12-15-15-18			0.250						0.034
12-15-16-16			0.042		0.095				0.017
12-15-16-17		0.040							0.005
12-15-16-18									
12-16-16-16					0.095				0.011
12-16-16-17									
12-16-17-17									
12-17-17-17									
13-13-13-13									
13-13-13-14									
13-13-13-15									
13-13-13-16									
13-13-13-17									
13-13-14-14									
13-13-14-15	0.040			0.050	0.048				0.017
13-13-14-16									
13-13-14-17								0.050	0.005
13-13-15-15								0.050	0.005
13-13-15-16					0.048		0.050		0.011
13-13-15-17					0.048			0.050	0.011

13-13-16-16									
13-13-16-17									
13-13-16-18									
13-13-17-17				0.050					0.005
13-13-17-18									
13-13-18-18									
13-14-14-14						0.050			0.005
13-14-14-15									
13-14-14-16									
13-14-14-17									
13-14-14-20									
13-14-15-15						0.050			0.005
13-14-15-16	0.080			0.050			0.050		0.022
13-14-15-17									
13-14-15-18									
13-14-15-19									
13-14-16-16				0.050					0.005
13-14-16-17									
13-14-16-18	0.040					0.050			0.011
13-14-17-17									
13-14-17-18						0.050			0.005
13-15-15-15								0.050	0.005
13-15-15-16					0.095				0.011
13-15-15-17									
13-15-15-18									
13-15-16-16		0.040	0.042						0.011
13-15-16-17							0.050		0.005
13-15-16-18									
13-15-17-18				0.050					0.005
13-15-17-20									
13-16-16-16									
13-16-16-17									
13-16-16-18									
13-16-16-19									
13-16-17-17									
13-16-17-18									
14-14-14-14									
14-14-14-15									
14-14-14-16		0.120		0.050					0.022
14-14-14-17						0.050			0.005
14-14-15-15	0.080								0.011
14-14-15-16		0.080							0.011
14-14-15-17									
14-14-15-18									
14-14-16-16				0.050					0.005
14-14-16-17									

14-14-16-18									
14-14-16-19									
14-14-17-17									
14-15-15-15									
14-15-15-16			0.042						0.005
14-15-15-17									
14-15-15-18									
14-15-16-16							0.100		0.011
14-15-16-17						0.100			0.011
14-15-16-18		0.040							0.005
14-15-17-17									
14-15-17-18									
14-16-16-16					0.048				0.005
14-16-16-17									
14-16-16-18									
14-16-17-17									
14-16-17-18	0.040								0.005
14-17-17-18									
15-15-15-15							0.050		0.005
15-15-15-16					0.048	0.050			0.011
15-15-15-17						0.050			0.005
15-15-15-18				0.050					0.005
15-15-16-16				0.100	0.048			0.050	0.022
15-15-16-17					0.048		0.050		0.011
15-15-16-18									
15-15-16-19		0.040							0.005
15-15-17-17									
15-15-17-18									
15-15-17-19									
15-16-16-16	0.040	0.120							0.022
15-16-16-17						0.050	0.050		0.011
15-16-16-18									
15-16-17-17									
15-16-17-18									
15-17-17-17									
16-16-16-16									
16-16-16-17									
16-16-17-17									
16-16-18-18									
16-17-17-17									
17-18-18-18									
20-20-22-23									

Genótipos	Frequência genotípica																		Leste Asiático (n=173)
	Han (n=44)	Tujia (n=10)	Yizu (n=10)	Miaozu (n=10)	Oroqen (n=10)	Daur (n=10)	Mongola (n=10)	Hezhen (n=9)	Xibo (n=9)	Uygur (n=10)	Dai (n=10)	Lahu (n=10)	She (n=10)	Naxi (n=10)	Tu (n=10)	Yakut (n=25)	Japanese (n=29)	Cambodian (n=11)	
9-12-15-16	0.042																		0.005
10-11-14-16																			
11-11-11-15																			
11-11-11-16					0.143	0.143	0.143												0.017
11-11-12-16						0.143													0.005
11-11-14-15																			
11-12-12-12									0.125										0.005
11-12-12-15																	0.043		0.005
11-12-12-16										0.125									0.005
11-12-12-17								0.200											0.005
11-12-12-18	0.042																		0.005
11-12-13-15																			
11-12-13-16									0.125										0.005
11-12-14-15		0.111																	0.005
11-12-16-18																			
11-13-13-14																			
11-13-15-16												0.143							0.005
11-14-14-14																			
11-14-14-15																			
11-14-14-16																			
11-14-15-15															0.143				0.005
11-14-15-16																			
11-14-15-18																			
11-14-16-18																			
11-14-17-17																			
11-15-15-16		0.111																	0.005

11-15-16-17																			
11-16-16-16																			
11-16-17-17																			
12-12-12-12	0.042																		0.005
12-12-12-13			0.111														0.200		0.011
12-12-12-15																			
12-12-13-13													0.250						0.011
12-12-13-14	0.042																		0.005
12-12-14-14																			
12-12-14-15		0.111															0.200		0.011
12-12-14-16																			
12-12-15-15										0.143									0.005
12-12-15-18																			
12-12-16-16		0.111	0.111		0.143														0.017
12-12-16-17									0.125										0.005
12-12-17-17																			
12-13-13-13	0.042	0.111			0.143														0.017
12-13-13-14																			
12-13-13-15													0.143						0.005
12-13-13-16														0.143					0.005
12-13-14-14			0.111														0.043		0.011
12-13-14-15	0.167	0.111					0.143	0.200											0.04
12-13-14-16																	0.200		0.005
12-13-14-17													0.143						0.005
12-13-15-16	0.042				0.143								0.143			0.056			0.023
12-13-15-17													0.143						0.005
12-13-15-18																			
12-13-16-16					0.143				0.125										0.011

12-14-14-14																			
12-14-14-15																			
12-14-14-16							0.200												0.005
12-14-15-15					0.143														0.005
12-14-15-16				0.286												0.043			0.017
12-14-15-17								0.125											0.005
12-14-16-16	0.042																		0.005
12-14-16-17																			
12-14-16-18																			
12-14-17-17																			
12-15-15-15																			
12-15-15-16									0.125										0.005
12-15-15-17									0.125					0.143					0.011
12-15-15-18																			
12-15-16-16																			
12-15-16-17																			
12-15-16-18																			
12-16-16-16																			
12-16-16-17																			
12-16-17-17																			
12-17-17-17																			
13-13-13-13														0.286					0.011
13-13-13-14	0.083					0.143						0.286		0.143					0.034
13-13-13-15		0.111					0.143									0.043			0.017
13-13-13-16										0.143									0.005
13-13-13-17										0.143									0.005
13-13-14-14	0.042																		0.005
13-13-14-15		0.111																	0.005

13-13-14-16			0.143						0.125										0.011
13-13-14-17															0.043			0.005	
13-13-15-15	0.042							0.250							0.087			0.028	
13-13-15-16																			
13-13-15-17	0.042		0.111								0.143				0.043			0.023	
13-13-16-16										0.143								0.005	
13-13-16-17																			
13-13-16-18																			
13-13-17-17																			
13-13-17-18	0.042											0.143						0.011	
13-13-18-18																			
13-14-14-14	0.042	0.111																0.011	
13-14-14-15													0.125					0.005	
13-14-14-16	0.042		0.111							0.143								0.017	
13-14-14-17						0.143												0.005	
13-14-14-20																			
13-14-15-15						0.143							0.143			0.200		0.017	
13-14-15-16			0.222							0.286				0.056				0.028	
13-14-15-17			0.111															0.005	
13-14-15-18																			
13-14-15-19																			
13-14-16-16						0.143												0.005	
13-14-16-17															0.130			0.017	
13-14-16-18																			
13-14-17-17																			
13-14-17-18																			
13-15-15-15						0.143												0.005	
13-15-15-16																			

13-15-15-17																			
13-15-15-18																			
13-15-16-16																			
13-15-16-17																0.087			0.011
13-15-16-18																			
13-15-17-18																			
13-15-17-20																			
13-16-16-16							0.200				0.714								0.034
13-16-16-17																			
13-16-16-18																			
13-16-16-19																			
13-16-17-17																			
13-16-17-18																			
14-14-14-14																			
14-14-14-15																			
14-14-14-16																			
14-14-14-17																			
14-14-15-15	0.042			0.286									0.125		0.889				0.115
14-14-15-16																			
14-14-15-17																			
14-14-15-18																			
14-14-16-16	0.042			0.143															0.011
14-14-16-17																	0.200		0.005
14-14-16-18																0.043			0.005
14-14-16-19									0.125										0.005
14-14-17-17																			
14-15-15-15																0.043			0.005
14-15-15-16	0.042						0.143												0.011

14-15-15-17														0.125					0.005
14-15-15-18									0.125										0.005
14-15-16-16																			
14-15-16-17									0.125								0.043		0.011
14-15-16-18																	0.043		0.005
14-15-17-17																			
14-15-17-18																			
14-16-16-16																			
14-16-16-17																			
14-16-16-18																			
14-16-17-17																			
14-16-17-18																			
14-17-17-18																			
15-15-15-15	0.042					0.143	0.143	0.200	0.125					0.375					0.046
15-15-15-16	0.042																		0.005
15-15-15-17																			
15-15-15-18																			
15-15-16-16					0.286														0.011
15-15-16-17									0.125										0.005
15-15-16-18																			
15-15-16-19																			
15-15-17-17																			
15-15-17-18																			
15-15-17-19																			
15-16-16-16																			
15-16-16-17																			
15-16-16-18																			
15-16-17-17																	0.043		0.005

15-16-17-18						0.143													0.005
15-17-17-17																			
16-16-16-16																			
16-16-16-17																			
16-16-17-17			0.111														0.043		0.011
16-16-18-18																	0.130		0.017
16-17-17-17																	0.043		0.005
17-18-18-18																			
20-20-22-23																			

Anexo 1 – Continuação

Genótipos	Frequências genotípicas		
	Papuan (n=17)	Melanesian (n=19)	Oceania (n=21)
9-12-15-16			
10-11-14-16			
11-11-11-15			
11-11-11-16			
11-11-12-16			
11-11-14-15			
11-12-12-12			
11-12-12-15			
11-12-12-16			
11-12-12-17			
11-12-12-18			
11-12-13-15			
11-12-13-16			
11-12-14-15			
11-12-16-18			
11-13-13-14			
11-13-15-16			
11-14-14-14			
11-14-14-15			
11-14-14-16			
11-14-15-15			
11-14-15-16			
11-14-15-18			
11-14-16-18			
11-14-17-17			
11-15-15-16			
11-15-16-17			
11-16-16-16			
11-16-17-17			
12-12-12-12		0.125	0.047
12-12-12-13			
12-12-12-15			
12-12-13-13			
12-12-13-14			
12-12-14-14			
12-12-14-15			

12-12-14-16			
12-12-15-15			
12-12-15-18			
12-12-16-16			
12-12-16-17			
12-12-17-17	0.077	0.250	0.142
12-13-13-13			
12-13-13-14			
12-13-13-15			
12-13-13-16			
12-13-14-14			
12-13-14-15			
12-13-14-16			
12-13-14-17			
12-13-15-16			
12-13-15-17			
12-13-15-18			
12-13-16-16			
12-14-14-14			
12-14-14-15			
12-14-14-16			
12-14-15-15			
12-14-15-16			
12-14-15-17			
12-14-16-16			
12-14-16-17			
12-14-16-18			
12-14-17-17			
12-15-15-15			
12-15-15-16			
12-15-15-17			
12-15-15-18			
12-15-16-16			
12-15-16-17			
12-15-16-18	0.077		0.047
12-16-16-16			
12-16-16-17	0.154	0.125	0.142
12-16-17-17			
12-17-17-17		0.375	0.142
13-13-13-13			
13-13-13-14			
13-13-13-15			
13-13-13-16			

13-13-13-17			
13-13-14-14			
13-13-14-15			
13-13-14-16			
13-13-14-17	0.077		0.047
13-13-15-15			
13-13-15-16			
13-13-15-17	0.154		0.095
13-13-16-16			
13-13-16-17			
13-13-16-18			
13-13-17-17			
13-13-17-18			
13-13-18-18			
13-14-14-14			
13-14-14-15			
13-14-14-16			
13-14-14-17			
13-14-14-20			
13-14-15-15			
13-14-15-16		0.125	0.047
13-14-15-17			
13-14-15-18			
13-14-15-19			
13-14-16-16			
13-14-16-17	0.077		0.047
13-14-16-18			
13-14-17-17			
13-14-17-18			
13-15-15-15			
13-15-15-16			
13-15-15-17			
13-15-15-18			
13-15-16-16			
13-15-16-17			
13-15-16-18			
13-15-17-18			
13-15-17-20			
13-16-16-16			
13-16-16-17			
13-16-16-18			
13-16-16-19			
13-16-17-17			

13-16-17-18			
14-14-14-14			
14-14-14-15			
14-14-14-16			
14-14-14-17			
14-14-15-15			
14-14-15-16			
14-14-15-17			
14-14-15-18			
14-14-16-16	0.154		0.095
14-14-16-17			
14-14-16-18			
14-14-16-19			
14-14-17-17			
14-15-15-15	0.077		0.047
14-15-15-16			
14-15-15-17			
14-15-15-18			
14-15-16-16			
14-15-16-17			
14-15-16-18			
14-15-17-17			
14-15-17-18			
14-16-16-16			
14-16-16-17			
14-16-16-18			
14-16-17-17			
14-16-17-18			
14-17-17-18			
15-15-15-15			
15-15-15-16			
15-15-15-17			
15-15-15-18			
15-15-16-16			
15-15-16-17			
15-15-16-18			
15-15-16-19			
15-15-17-17			
15-15-17-18			
15-15-17-19			
15-16-16-16	0.077		0.047
15-16-16-17			
15-16-16-18			

15-16-17-17			
15-16-17-18			
15-17-17-17			
16-16-16-16	0.077		0.047
16-16-16-17			
16-16-17-17			
16-16-18-18			
16-17-17-17			
17-18-18-18			
20-20-22-23			

Anexo 1 – Continuação

Genótipos	Frequências genotípicas								
	French (n=24)	Basque (n=24)	Sardinian (n=28)	Bergamo (n=13)	Tuscan (n=8)	Orcadian (n=16)	Adygei (n=17)	Russian (n=25)	Europa (n=88)
9-12-15-16									
10-11-14-16									
11-11-11-15									
11-11-11-16									
11-11-12-16									
11-11-14-15									
11-12-12-12									
11-12-12-15									
11-12-12-16									
11-12-12-17									
11-12-12-18									
11-12-13-15									
11-12-13-16									
11-12-14-15									
11-12-16-18									
11-13-13-14									
11-13-15-16									
11-14-14-14			0.063						0.011
11-14-14-15			0.250						0.045
11-14-14-16			0.063						0.011
11-14-15-15			0.063						0.011
11-14-15-16			0.063						0.011
11-14-15-18				0.125					0.011
11-14-16-18									
11-14-17-17									
11-15-15-16					0.167				0.011
11-15-16-17						0.143			0.011
11-16-16-16									
11-16-17-17									
12-12-12-12									
12-12-12-13									
12-12-12-15	0.083								0.011
12-12-13-13									
12-12-13-14									
12-12-14-14									
12-12-14-15									
12-12-14-16									

12-12-15-15									
12-12-15-18									
12-12-16-16									
12-12-16-17									
12-12-17-17									
12-13-13-13									
12-13-13-14			0.063						0.011
12-13-13-15									
12-13-13-16									
12-13-14-14					0.167				0.011
12-13-14-15									
12-13-14-16	0.083								0.011
12-13-14-17									
12-13-15-16						0.143			0.011
12-13-15-17									
12-13-15-18									
12-13-16-16							0.063		0.011
12-14-14-14									
12-14-14-15			0.063						0.011
12-14-14-16	0.083								0.011
12-14-15-15							0.063		0.011
12-14-15-16					0.143	0.143	0.063		0.034
12-14-15-17									
12-14-16-16	0.083								0.011
12-14-16-17			0.063				0.063		0.022
12-14-16-18									
12-14-17-17									
12-15-15-15									
12-15-15-16			0.063			0.143	0.063		0.034
12-15-15-17									
12-15-15-18									
12-15-16-16							0.188		0.034
12-15-16-17									
12-15-16-18									
12-16-16-16					0.143				0.011
12-16-16-17									
12-16-17-17									
12-17-17-17									
13-13-13-13									
13-13-13-14				0.125		0.286			0.034
13-13-13-15									
13-13-13-16									
13-13-13-17									

13-13-14-14							0.143	0.063	0.022
13-13-14-15									
13-13-14-16									
13-13-14-17									
13-13-15-15									
13-13-15-16									
13-13-15-17									
13-13-16-16									
13-13-16-17									
13-13-16-18									
13-13-17-17									
13-13-17-18									
13-13-18-18									
13-14-14-14									
13-14-14-15									
13-14-14-16									
13-14-14-17									
13-14-14-20									
13-14-15-15									
13-14-15-16									
13-14-15-17									
13-14-15-18									
13-14-15-19									
13-14-16-16					0.167				0.011
13-14-16-17									
13-14-16-18									
13-14-17-17									
13-14-17-18									
13-15-15-15									
13-15-15-16		0.063							0.011
13-15-15-17									
13-15-15-18									
13-15-16-16									
13-15-16-17									
13-15-16-18	0.083								0.011
13-15-17-18									
13-15-17-20									
13-16-16-16									
13-16-16-17							0.143		0.011
13-16-16-18									
13-16-16-19									
13-16-17-17									
13-16-17-18									

14-14-14-14									
14-14-14-15			0.063						0.011
14-14-14-16									
14-14-14-17									
14-14-15-15							0.125		0.022
14-14-15-16				0.125					0.011
14-14-15-17									
14-14-15-18									
14-14-16-16	0.083						0.063		0.022
14-14-16-17									
14-14-16-18									
14-14-16-19									
14-14-17-17									
14-15-15-15									
14-15-15-16		0.063							
14-15-15-17		0.063							0.011
14-15-15-18									
14-15-16-16									0.011
14-15-16-17	0.083	0.063			0.167				0.034
14-15-16-18									
14-15-17-17									
14-15-17-18				0.125					0.011
14-16-16-16									
14-16-16-17							0.063		0.011
14-16-16-18									
14-16-17-17									
14-16-17-18	0.083								0.011
14-17-17-18									
15-15-15-15						0.143		0.063	0.022
15-15-15-16									
15-15-15-17				0.125		0.143			0.022
15-15-15-18									
15-15-16-16		0.063		0.125					0.022
15-15-16-17	0.083	0.063	0.063		0.167				0.045
15-15-16-18							0.063		0.011
15-15-16-19									
15-15-17-17		0.188	0.063			0.143			0.056
15-15-17-18		0.313		0.125		0.143			0.079
15-15-17-19	0.083								0.011
15-16-16-16				0.125					0.011
15-16-16-17	0.167								0.022
15-16-16-18									
15-16-17-17		0.063			0.167				0.022

15-16-17-18									
15-17-17-17		0.063							0.011
16-16-16-16							0.063		0.011
16-16-16-17									
16-16-17-17									
16-16-18-18			0.063						0.011
16-17-17-17									
17-18-18-18									
20-20-22-23									

Anexo 1 – Continuação

Genótipos	Frequências genotípicas					
	Pima (n=25)	Maya (n=25)	Colombian (n=13)	Karitiana (n=24)	Surui (n=21)	América (n=43)
9-12-15-16						
10-11-14-16						
11-11-11-15						
11-11-11-16						
11-11-12-16						
11-11-14-15						
11-12-12-12						
11-12-12-15						
11-12-12-16						
11-12-12-17						
11-12-12-18						
11-12-13-15						
11-12-13-16						
11-12-14-15						
11-12-16-18						
11-13-13-14						
11-13-15-16						
11-14-14-14						
11-14-14-15						
11-14-14-16						
11-14-15-15						
11-14-15-16						
11-14-15-18						
11-14-16-18						
11-14-17-17						
11-15-15-16						
11-15-16-17						
11-16-16-16						
11-16-17-17						
12-12-12-12						
12-12-12-13						
12-12-12-15						
12-12-13-13						
12-12-13-14						
12-12-14-14						
12-12-14-15						
12-12-14-16						
12-12-15-15						

12-12-15-18						
12-12-16-16						
12-12-16-17						
12-12-17-17						
12-13-13-13						
12-13-13-14	0.143					0.046
12-13-13-15						
12-13-13-16						
12-13-14-14						
12-13-14-15						
12-13-14-16						
12-13-14-17						
12-13-15-16						
12-13-15-17						
12-13-15-18				0.300		0.069
12-13-16-16						
12-14-14-14						
12-14-14-15						
12-14-14-16						
12-14-15-15						
12-14-15-16						
12-14-15-17						
12-14-16-16						
12-14-16-17						
12-14-16-18						
12-14-17-17						
12-15-15-15						
12-15-15-16						
12-15-15-17						
12-15-15-18						
12-15-16-16						
12-15-16-17						
12-15-16-18						
12-16-16-16						
12-16-16-17						
12-16-17-17						
12-17-17-17						
13-13-13-13						
13-13-13-14						
13-13-13-15						
13-13-13-16						
13-13-13-17						
13-13-14-14						

13-13-14-15						
13-13-14-16						
13-13-14-17						
13-13-15-15						
13-13-15-16						
13-13-15-17						
13-13-16-16						
13-13-16-17						
13-13-16-18						
13-13-17-17						
13-13-17-18						
13-13-18-18						
13-14-14-14						
13-14-14-15			0.400			0.046
13-14-14-16						
13-14-14-17						
13-14-14-20						
13-14-15-15						
13-14-15-16						
13-14-15-17						
13-14-15-18		0.333				0.023
13-14-15-19						
13-14-16-16						
13-14-16-17	0.071					0.023
13-14-16-18						
13-14-17-17	0.071					0.023
13-14-17-18						
13-15-15-15						
13-15-15-16						
13-15-15-17					0.545	0.139
13-15-15-18				0.600		0.139
13-15-16-16						
13-15-16-17					0.182	0.046
13-15-16-18						
13-15-17-18						
13-15-17-20						
13-16-16-16						
13-16-16-17						
13-16-16-18						
13-16-16-19						
13-16-17-17						
13-16-17-18						
14-14-14-14						

14-14-14-15						
14-14-14-16	0.071					0.023
14-14-14-17						
14-14-15-15						
14-14-15-16						
14-14-15-17				0.100		0.023
14-14-15-18			0.200			0.023
14-14-16-16						
14-14-16-17						
14-14-16-18						
14-14-16-19						
14-14-17-17						
14-15-15-15						
14-15-15-16	0.643	0.333				0.232
14-15-15-17						
14-15-15-18					0.273	0.069
14-15-16-16						
14-15-16-17			0.400			0.046
14-15-16-18						
14-15-17-17						
14-15-17-18						
14-16-16-16						
14-16-16-17						
14-16-16-18						
14-16-17-17						
14-16-17-18						
14-17-17-18						
15-15-15-15						
15-15-15-16						
15-15-15-17						
15-15-15-18						
15-15-16-16						
15-15-16-17						
15-15-16-18						
15-15-16-19						
15-15-17-17		0.333				0.023
15-15-17-18						
15-15-17-19						
15-16-16-16						
15-16-16-17						
15-16-16-18						
15-16-17-17						
15-16-17-18						

15-17-17-17						
16-16-16-16						
16-16-16-17						
16-16-17-17						
16-16-18-18						
16-17-17-17						
17-18-18-18						
20-20-22-23						

Anexo 2 – Frequências genótípicas do microssatélite *DYS464* em populações parentais e miscigenadas

Genótipos	Frequências genótípicas						
	Pretos (n=124)	Branco (n=106)	Afro- Americanos (n=256)	Caucasianos (n=256)	Africanos (n=90)	Europeus (n=88)	Ameríndios (n=43)
10 11 14 16					0.011		
11 11 14 15			0.004				
11 11 14 16			0.004				
11 11 16 16			0.004				
11 12 14 15	0.008						
11 12 15 16		0.009					
11 13 14 15	0.008	0.009		0.004			
11 13 15 15				0.004			
11 13 15 17		0.009					
11 13 16 16		0.019					
11 14 14 14	0.016	0.009		0.012		0.011	
11 14 14 15						0.045	
11 14 14 16				0.012		0.011	
11 14 15 15				0.008		0.011	
11 14 15 16						0.011	
11 14 15 18						0.011	
11 14 17 17					0.011		
11 15 15 15		0.009					
11 15 15 16						0.011	
11 15 15 18				0.004			
11 15 16 17			0.004			0.011	
11 16 16 18			0.004				
11 16 17 17					0.011		
11 16 18 18			0.004				
12 12 12 15						0.011	
12 12 14 14		0.009					
12 12 14 15				0.004			
12 12 15 15	0.008						
12 12 15 16	0.008			0.004			
12 12 15 18					0.011		
12 12 16 16	0.008						
12 13 13 14				0.016		0.011	0.047
12 13 14 14	0.008	0.028	0.004	0.004		0.011	
12 13 14 15		0.009	0.004	0.004			

12 13 14 16				0.004		0.011	
12 13 15 15	0.008						
12 13 15 16				0.004	0.022	0.011	
12 13 15 17			0.004				
12 13 15 18							0.070
12 13 15.1 16				0.004			
12 13 16 16		0.019				0.011	
12 13 16 17	0.008						
12 14 14 14	0.008	0.009		0.004			
12 14 14 15			0.004	0.008		0.011	
12 14 14 16		0.009	0.004	0.004		0.011	
12 14 14 17				0.008			
12 14 15 15		0.009	0.004	0.020		0.011	
12 14 15 16		0.019	0.012	0.059	0.011	0.034	
12 14 15 17			0.004				
12 14 16 16	0.008			0.004		0.011	
12 14 16 17		0.009				0.023	
12 14 16 18			0.008				
12 14 17 18	0.008						
12 15 15 15			0.008	0.008			
12 15 15 16		0.009	0.004	0.020		0.034	
12 15 15 18			0.004		0.011		
12 15 16 16		0.019		0.016	0.011	0.034	
12 15 16 17			0.016	0.008	0.011		
12 15 16 18	0.008		0.023		0.044		
12 15 17 17			0.004				
12 15 17 18			0.004				
12 15 17 19			0.004				
12 16 16 16	0.008		0.008			0.011	
12 16 16 17			0.012				
12 16 17 17	0.008						
12 16 17 19	0.008						
13 13 13 14				0.004		0.034	
13 13 13 15			0.004	0.004			
13 13 14 14				0.004		0.023	
13 13 14 15		0.009					
13 13 14 16	0.008						
13 13 14 17			0.004				
13 13 15 15					0.011		
13 13 15 16				0.004			

13 13 16 16	0.008			0.004			
13 13 16 17	0.008		0.004		0.011		
13 13 16 18					0.011		
13 13 17 19	0.008						
13 13 18 18		0.019			0.011		
13 13.1 15 16			0.004	0.004			
13 14 14 14	0.016			0.004	0.011		
13 14 14 15	0.008			0.004			0.047
13 14 14 16					0.033		
13 14 14 20					0.011		
13 14 15 15		0.009		0.004			
13 14 15 16	0.008		0.004				
13 14 15 17	0.008		0.004				
13 14 15 18					0.011		0.023
13 14 15 19					0.022		
13 14 16 16			0.004			0.011	
13 14 16 17	0.008		0.004				0.023
13 14 17 17	0.016	0.009					0.023
13 14.3 16 17			0.004				
13 15 15 16	0.008					0.011	
13 15 15 17	0.008			0.012			0.140
13 15 15 18		0.009	0.008	0.004	0.011		0.140
13 15 15 19				0.008			
13 15 15 20		0.009					
13 15 15.1 15.1				0.004			
13 15 15.2 16	0.008						
13 15 16 16	0.024	0.009	0.035		0.011		
13 15 16 17	0.032		0.016	0.008	0.022		0.047
13 15 16 18	0.024		0.035		0.067	0.011	
13 15 17 17	0.024		0.004				
13 15 17 18	0.008	0.009	0.012				
13 15 17 20					0.011		
13 15 18 18			0.004				
13 16 16 16	0.016		0.020		0.033		
13 16 16 17	0.048	0.009	0.109		0.044	0.011	
13 16 16 18	0.065	0.009	0.082	0.004	0.222		
13 16 16 19	0.008		0.016		0.011		
13 16 16 20			0.004				
13 16 17 17	0.016		0.035	0.008	0.011		
13 16 17 18	0.057		0.031		0.011		

13 16 17 19			0.004				
13 16 18 18			0.012				
13 16 18 19	0.008						
13 17 18 18			0.008				
14 14 14 14		0.009		0.012	0.011		
14 14 14 15		0.009	0.004	0.012		0.011	
14 14 14 16							0.023
14 14 15 15	0.016	0.009		0.016		0.023	
14 14 15 16	0.008			0.008	0.011	0.011	
14 14 15 17		0.009		0.008	0.033		0.023
14 14 15 18			0.004				0.023
14 14 16 16	0.008				0.011	0.023	
14 14 16 17			0.008				
14 14 16 18			0.004				
14 14 16 20			0.004				
14 14 17 17				0.008			
14 14 18 18		0.009					
14 14.3 15 17				0.004			
14 14.3 16 18			0.004				
14 15 15 16	0.024	0.019	0.008	0.004		0.011	0.233
14 15 15 17		0.009	0.008		0.022	0.011	
14 15 15 18		0.009					0.070
14 15 16 16	0.008	0.028	0.004	0.004			
14 15 16 17	0.024	0.009	0.004	0.016		0.034	0.047
14 15 16 18	0.016	0.009	0.012		0.022		
14 15 17 17	0.008	0.009	0.016	0.047	0.022		
14 15 17 18		0.019	0.008	0.016		0.011	
14 15 17 19	0.008						
14 15 18 18				0.004			
14 15.3 16 16				0.004			
14 16 16 16	0.008						
14 16 16 17	0.008	0.019	0.008	0.004		0.011	
14 16 16 18			0.004		0.011		
14 16 17 17	0.008	0.019		0.008	0.011		
14 16 17 18	0.008	0.009	0.016	0.004		0.011	
14 17 18 19			0.004				
14.3 15 16 17			0.004				
15 15 15 15	0.008	0.009		0.004		0.023	
15 15 15 16		0.009	0.008	0.008			
15 15 15 17	0.024		0.012	0.012		0.023	

15 15 15 18		0.009					
15 15 15.3 16				0.004			
15 15 16 16	0.032	0.009	0.027	0.016		0.023	
15 15 16 17	0.032	0.076	0.035	0.098		0.045	
15 15 16 18		0.028	0.008	0.008		0.011	
15 15 16 19				0.004			
15 15 16 20				0.004			
15 15 17 17	0.089	0.132	0.043	0.148	0.011	0.057	0.023
15 15 17 18		0.038	0.020	0.031		0.080	
15 15 17 19		0.009		0.004		0.011	
15 15 18 18			0.004	0.004			
15 15 18 19			0.004				
15 15.2 16 16	0.008						
15 16 16 16			0.012	0.008		0.011	
15 16 16 17	0.008	0.009	0.020	0.027	0.033	0.023	
15 16 16 18		0.019		0.016			
15 16 16 20				0.004			
15 16 17 17	0.008	0.047	0.023	0.031		0.023	
15 16 17 18	0.016	0.038	0.004	0.027	0.011		
15 16 17 19				0.004			
15 16 18 18		0.009		0.004			
15 16 18 20				0.004			
15 17 17 17		0.009			0.011	0.011	
15 17 17 18	0.008		0.008	0.004			
15 17 18 19				0.004			
16 16 16 16				0.004		0.011	
16 16 16 17			0.012				
16 16 16 18				0.004			
16 16 16 19			0.004				
16 16 17 17	0.008		0.004	0.008			
16 16 17 18		0.009					
16 16 18 18	0.008					0.011	
16 17 17 17			0.004				
16 18 18 18	0.008						
17 17 17 17				0.008			
17 17 18 18			0.004				
17 18 18 18					0.022		
17 19 21 22		0.009					
20 20 22 23					0.011		

Artigo Científico

Worldwide diversity of the Y-chromosome tetra- local microsatellite DYS464

Genetics and Molecular Research 9 (3): 1525-1534 (2010)

Worldwide diversity of the Y-chromosome tetra-local microsatellite *DYS464*

F.S.G. Kehdy and S.D.J. Pena

Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

Corresponding author: S.D.J. Pena
E-mail: spena@dcc.ufmg.br

Genet. Mol. Res. 9 (3): 1525-1534 (2010)
Received June 6, 2010
Accepted July 12, 2010
Published August 3, 2010
DOI 10.4238/vol9-3gmr968

ABSTRACT. Of all DNA markers on the human Y-chromosome, the tetra-local Y-linked microsatellite *DYS464* is the most polymorphic. We genotyped *DYS464* in 677 male samples collected worldwide, maintained in the HGDP-CEPH Human Genome Diversity Cell Line Panel. Fourteen different alleles were found, with allele lengths varying from 9 to 23 repeats. One hundred and seventy-five different genotypes were detected, of which 90 appeared to be continent-specific. The region with the highest percentage of unique genotypes was Africa. Genotype diversity was 0.98 for Europe, 0.97 for Central and East Asia, 0.95 for Africa, 0.94 for Oceania, 0.92 for the Middle East, and 0.90 for the Americas. A hierarchical analysis of molecular variance showed low levels of worldwide genetic structure; 88.42% of the genetic variance was found within populations, 9.62% between populations within regions and 1.96% between regions. Since the four *DYS464* repeats are identical, one cannot assign each peak in the electropherogram to a specific locus. Thus, the same genotype may correspond to several haplotypes, with different permutations of alleles. Consequently, genotypes are degenerate, which limits phylogeographical analyses. Yet, because of its high variability, *DYS464* still constitutes an informative tool for population and evolutionary studies.

Key words: Y-chromosome; Worldwide populations; Microsatellites; DNA; Population genetics; *DYS464*

INTRODUCTION

Y-linked loci are haploid and paternally inherited and with the exception of genes in the pseudo-autosomal regions, there is no recombination (reviewed by Jobling and Tyler-Smith, 2003). Thus, Y-chromosomal markers are transmitted together as haplotypes. Therefore, each male individual has the same Y-chromosome haplotype as his father, brothers, paternal grandfather, paternal uncles, etc., thus establishing a patrilineage.

These characteristics render Y-linked polymorphisms extremely useful as genetic tools for paternity testing (Santos et al., 2003), forensic medicine (Gusmão et al., 2006), ancestry studies (Carvalho-Silva et al., 2001, 2006), and human evolutionary genetics (Chiaroni et al., 2009).

More than 300 microsatellite loci have been described on the Y-chromosome (Hanson and Ballantyne, 2006). Among them, the most polymorphic is the tetralocal microsatellite *DYS464* (Redd et al., 2002; Berger et al., 2003).

DYS464 consists of four identical copies (a, b, c, and d) of the tetranucleotide repeat (CCTT)_n on the palindromic AZFc region (Kuroda-Kawaguchi et al., 2001) in band Yq11.223 of the Y-chromosome (Redd et al., 2002) (Figure 1). It is believed that the tetra-local structure evolved from two sequential duplications of an original monomer, promoted by mitotic recombination events (Kuroda-Kawaguchi et al., 2001; Skaletsky et al., 2003; Rozen et al., 2003).

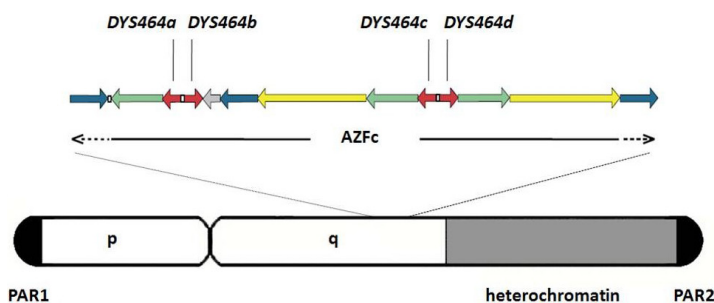


Figure 1. Diagram of the human Y-chromosome, indicating the AZFc (azoospermia factor c) region with its highly palindromic structure, which harbors the four loci of the tetra-local microsatellite *DYS464*. On the Y-chromosome are shown the two pseudo-autosomal regions (PAR1 and PAR2), the short (p) and long (q) arms and the large heterochromatic region. The figure was drawn based on data from Kuroda-Kawaguchi et al. (2001).

DYS464 genotypes are established in a straightforward fashion when four peaks of identical area are present in electropherograms. However, since one cannot assign each peak to a specific locus, the same genotype (combination of alleles - order is ignored) may correspond to different haplotypes (permutations of alleles - order is significant). In this fashion, genotypes are degenerate. Let us take, for instance, the situation in which we identify four peaks, corresponding to four different alleles, say, 13, 14, 15, and 16. This is scored as genotype 13-14-15-16, which could in fact be any of 12 different haplotypes that contain these four alleles in different permutations, i.e., 13-14-16-15, 13-16-15-14, etc. Moreover, when less than four peaks are seen, the relative peak areas have to be used to estimate the number of copies of a particular allele (Butler and Schoske, 2005).

Although *DYS464* has been studied in some human populations (Redd et al., 2002; Berger et al., 2003; Butler and Schoske, 2005), a more complete study of this marker in worldwide human population has not yet been reported. In this article, we describe our results on the typing of *DYS464* in all male samples from HGDP-CEPH Human Genome

Diversity Cell Line Panel (Cann et al., 2002) and discuss its diversity and phylogeography in different human populations.

MATERIAL AND METHODS

Samples

DNA samples from all 677 unrelated males from HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al., 2002; <http://www.cephb.fr/fr/hgdp/diversity.php>) were analyzed in this study. The individuals were sampled across all five continents and assigned to 52 different populations from seven regional groups (Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, and Americas).

DNA typing

For polymerase chain reaction (PCR) amplification of *DYS464*, the following primers were used: *DYS464*-F: 5'-TTACGAGCTTTGGGCTATG-3' with a tail of the M13-40 17-oligonucleotide GTTTTCCCAGTCACGAC and *DYS464*-R: 5'-CCTGGGTAACAGAGAGACTCTT-3'.

PCR was performed using 2 U *Taq* DNA polymerase (Phonectria, Belo Horizonte, Brazil) and 200 μ M dNTPs in 10 mM Tris-HCl buffer, pH 8.4, with 50 mM KCl, 1.5 mM MgCl₂, 0.1% Triton. For fragment separation, we used a MegaBACE 1000 DNA sequencer (GE Healthcare) followed by analyses using Genetic Profiler (version 2.2) and Fragment Profiler (version 1.2) programs (GE Healthcare).

Electropherogram profiles for *DYS464* consisted of one, two, three, or four peaks, which were converted to genotypes based on the estimated repeat number and relative areas of the four alleles present in each sample (Butler and Schoske, 2005). We always assumed that four loci were present, unless there was a compelling reason to believe differently.

Data analysis

The genetic structure of populations and basic population parameters including analyses of molecular variance (AMOVA), genotype diversity, genotype frequency, and genotype sharing were calculated using the Arlequin 2.0 software (Schneider et al., 2000) with 10,000 steps in the Markov chain.

RESULTS

We genotyped *DYS464* in all 677 male samples of the HGDP-CEPH panel, representing 52 different populations from seven regional groups worldwide (Sub-Saharan Africa, Europe, Middle East, Central Asia, East Asia, Oceania, and Americas). Fourteen different alleles were found, having lengths varying from 9 to 23 repeats. However, allele 21 was not found in any sample.

A total of 175 genotypes were identified: 49 in Africa, 53 in Europe, 29 in the Middle East, 72 in Central Asia, 83 in East Asia, 13 in Oceania, and 16 in Americas (Tables 1 and 2). It is important to remember that each genotype could correspond to many different haplotypes containing the same alleles in different orders.

Table 1. Frequency of 175 different *DYS464* genotypes in 677 male samples from the HGDP-CEPH Human Genome Diversity Cell Line Panel belonging to seven geographical groups (Africa, Middle East, Central/South Asia, East Asia, Oceania, Europe, and Americas).

Genotype	Africa (N = 98)	Middle East (N = 79)	Central Asia (N = 175)	East Asia (N = 173)	Oceania (N = 21)	Europe (N = 88)	Americas (N = 43)
9-12-15-16				0.005			
10-11-14-16	0.010						
11-11-11-15			0.005				
11-11-11-16				0.017			
11-11-12-16			0.040	0.005			
11-11-14-15			0.022				
11-12-12-12				0.005			
11-12-12-15				0.005			
11-12-12-16				0.005			
11-12-12-17				0.005			
11-12-12-18				0.005			
11-12-13-15			0.011				
11-12-13-16				0.005			
11-12-14-15				0.005			
11-12-16-18			0.005				
11-13-13-14			0.022				
11-13-15-16				0.005			
11-14-14-14						0.011	
11-14-14-15			0.005			0.045	
11-14-14-16						0.011	
11-14-15-15			0.005	0.005		0.011	
11-14-15-16			0.028			0.011	
11-14-15-18						0.011	
11-14-16-18		0.012					
11-14-17-17	0.010						
11-15-15-16			0.005	0.005		0.011	
11-15-16-17						0.011	
11-16-16-16			0.005				
11-16-17-17	0.010						
12-12-12-12				0.005	0.047		
12-12-12-13				0.011			
12-12-12-15			0.005			0.011	
12-12-13-13		0.025	0.005	0.011			
12-12-13-14				0.005			
12-12-14-14		0.012					
12-12-14-15			0.022	0.011			
12-12-14-16			0.028				
12-12-15-15				0.005			
12-12-15-18	0.010						
12-12-16-16				0.017			
12-12-16-17				0.005			
12-12-17-17					0.142		
12-13-13-13				0.017			
12-13-13-14		0.050				0.011	0.046
12-13-13-15		0.012	0.011	0.005			
12-13-13-16			0.005	0.005			
12-13-14-14		0.113		0.011		0.011	
12-13-14-15			0.017	0.040			
12-13-14-16			0.017	0.005		0.011	
12-13-14-17		0.012		0.005			
12-13-15-16	0.020	0.012	0.011	0.023		0.011	
12-13-15-17				0.005			
12-13-15-18							0.069
12-13-16-16			0.005	0.011		0.011	
12-14-14-14			0.005				

Continued on next page

Table 1. Continued.

Genotype	Africa (N = 98)	Middle East (N = 79)	Central Asia (N = 175)	East Asia (N = 173)	Oceania (N = 21)	Europe (N = 88)	Americas (N = 43)
12-14-14-15			0.011			0.011	
12-14-14-16			0.011	0.005		0.011	
12-14-15-15		0.025	0.022	0.005		0.011	
12-14-15-16	0.010		0.034	0.017		0.034	
12-14-15-17	0.010	0.037		0.005			
12-14-16-16				0.005		0.011	
12-14-16-17			0.011			0.022	
12-14-16-18		0.164					
12-14-17-17			0.011				
12-15-15-15			0.011				
12-15-15-16		0.025	0.137	0.005		0.034	
12-15-15-17			0.022	0.011			
12-15-15-18	0.010		0.034				
12-15-16-16	0.010		0.017			0.034	
12-15-16-17	0.010		0.005				
12-15-16-18	0.040				0.047		
12-16-16-16			0.011			0.011	
12-16-16-17					0.142		
12-16-17-17		0.012					
12-17-17-17					0.142		
13-13-13-13				0.011			
13-13-13-14				0.034		0.034	
13-13-13-15				0.017			
13-13-13-16				0.005			
13-13-13-17				0.005			
13-13-14-14				0.005		0.022	
13-13-14-15			0.017	0.005			
13-13-14-16				0.011			
13-13-14-17			0.005	0.005	0.047		
13-13-15-15	0.010		0.005	0.028			
13-13-15-16			0.011				
13-13-15-17			0.011	0.023	0.095		
13-13-16-16				0.005			
13-13-16-17	0.010						
13-13-16-18	0.010						
13-13-17-17			0.005				
13-13-17-18				0.011			
13-13-18-18	0.010						
13-14-14-14	0.010	0.012	0.005	0.011			
13-14-14-15				0.005			0.046
13-14-14-16	0.030			0.017			
13-14-14-17		0.012		0.005			
13-14-14-20	0.010						
13-14-15-15			0.005	0.017			
13-14-15-16			0.022	0.028	0.047		
13-14-15-17				0.005			
13-14-15-18	0.010						0.023
13-14-15-19	0.020						
13-14-16-16			0.005	0.005		0.011	
13-14-16-17		0.063		0.017	0.047		0.023
13-14-16-18			0.011				
13-14-17-17							0.023
13-14-17-18			0.005				
13-15-15-15			0.005	0.005			
13-15-15-16			0.011			0.011	
13-15-15-17							0.139
13-15-15-18	0.010						0.139
13-15-16-16	0.020		0.011				
13-15-16-17	0.040	0.012	0.005	0.011			0.046

Continued on next page

Table 1. Continued.

Genotype	Africa (N = 98)	Middle East (N = 79)	Central Asia (N = 175)	East Asia (N = 173)	Oceania (N = 21)	Europe (N = 88)	Americas (N = 43)
13-15-16-18	0.061	0.012				0.011	
13-15-17-18			0.005				
13-15-17-20	0.010						
13-16-16-16	0.030			0.034			
13-16-16-17	0.040					0.011	
13-16-16-18	0.204						
13-16-16-19	0.020						
13-16-17-17	0.010						
13-16-17-18	0.010						
14-14-14-14	0.010						
14-14-14-15						0.011	
14-14-14-16			0.022				0.023
14-14-14-17			0.005				
14-14-15-15			0.011	0.115		0.022	
14-14-15-16	0.010		0.011			0.011	
14-14-15-17	0.030						0.023
14-14-15-18							0.023
14-14-16-16	0.010	0.037	0.005	0.011	0.095	0.022	
14-14-16-17		0.012		0.005			
14-14-16-18				0.005			
14-14-16-19				0.005			
14-14-17-17		0.012					
14-15-15-15				0.005	0.047		
14-15-15-16			0.005	0.011			0.232
14-15-15-17	0.020			0.005		0.011	
14-15-15-18				0.005			0.069
14-15-16-16			0.011			0.011	
14-15-16-17			0.011	0.011		0.034	0.046
14-15-16-18	0.020		0.005	0.005			
14-15-17-17	0.020						
14-15-17-18						0.011	
14-16-16-16	0.010	0.037	0.005				
14-16-16-17		0.164				0.011	
14-16-16-18	0.010	0.012					
14-16-17-17	0.010						
14-16-17-18	0.010		0.005			0.011	
14-17-17-18	0.010						
15-15-15-15			0.005	0.046		0.022	
15-15-15-16			0.011	0.005			
15-15-15-17			0.005			0.022	
15-15-15-18			0.005				
15-15-16-16			0.022	0.011		0.022	
15-15-16-17			0.011	0.005		0.045	
15-15-16-18		0.012				0.011	
15-15-16-19			0.005				
15-15-17-17	0.010					0.056	0.023
15-15-17-18						0.079	
15-15-17-19						0.011	
15-16-16-16			0.022		0.047	0.011	
15-16-16-17	0.030	0.012	0.011			0.022	
15-16-16-18		0.025					
15-16-17-17		0.025		0.005		0.022	
15-16-17-18	0.010			0.005			
15-17-17-17	0.010					0.011	
16-16-16-16					0.047	0.011	
16-16-16-17		0.012					
16-16-17-17				0.011			
16-16-18-18				0.017		0.011	
16-17-17-17				0.005			
17-18-18-18	0.020						
20-20-22-23	0.010						

Table 2. Data on the *DYS464* genotypes encountered in this study.

Region	Number of individuals	Number of genotypes	Number of continent-specific genotypes	Genotype diversity (means \pm SE)
Africa	98	49	20 (40%)	0.95 \pm 0.01
Middle East	79	29	7 (24%)	0.92 \pm 0.01
Central Asia	175	72	19 (26.4%)	0.97 \pm 0.01
East Asia	173	83	29 (34.9%)	0.97 \pm 0.01
Oceania	21	13	3 (23%)	0.94 \pm 0.03
Europe	88	53	8 (15%)	0.98 \pm 0.01
Americas	43	16	4 (25%)	0.90 \pm 0.03
Total	677	175	90 (51.4%)	0.99

Even so, there was considerable geographical differentiation in genotype distribution. Among the 175 genotypes scored, 90 were observed in a single continent. The region with the highest percentage of unique genotypes was Africa (20 in a total of 49 genotypes, i.e., 40.8%).

The most frequent genotypes seen in each continent were the following:

- In Africa, 13-16-16-18 (20.4%) - this genotype seemed widespread in Sub-Saharan Africa, seen in significant frequencies among the Biaka (43.3%), Bantu (18.2%), Yoruba (15.4%), Mandenka (12.5%), and Mbuti (7.7%), but not the San. This genotype was not seen outside Africa.
- In Europe, 15-15-17-18 (8.0%) - this genotype was seen among the Basques of France (31.3%), Orcadian (14.3%) and Bergamo of Italy (12.5%). It was not seen among the French, Sardinian, Tuscan, Adygei of Caucasus, and Russian peoples. This genotype was not seen outside Europe.
- In East Asia, 14-14-15-15 (11.6%) - this genotype was seen among the Yakut of Siberia (88.9%) and the Miaozi (28.6%), Naxi (12.5%) and Han (4.2%) of China. Outside East Asia it was seen among the Russian (12.5%) and the Brahui in Pakistan (8.0%).
- In the Americas, 14-15-15-16 (23.2%) - this genotype was seen among the Pimas (64.3%) and Mayans (33.3%), but not in Colombians (Piapoco and Curripaco), Kari-tiana and Surui. Outside the Americas, this genotype was seen in the Daur of Mongolia (14.3%), Basques of France (6.3%), Hazara of Pakistan (4.2%), and Han of China (4.2%).
- In Central Asia, 12-15-15-16 (13.7%) - the genotype was seen among the Bahui (16.0%), Baloshi (16.0%), Makrani (15.0%), Pathan (30.0%), and Burusho (23.8%) of Pakistan, but not in the Hazara. Outside Central Asia, it was seen in Bedouin of Israel (7.1%), Uygur of China (12.5%), Sardinians of Italy (6.3%), Adygei of the Caucasus (14.3%), and Russians (6.3%).
- In the Middle East, there was a tie for genotypes 12-14-16-18 and 14-16-16-17 (16.4% each). The former was seen exclusively among the Bedouin of Israel (46.2%), while the latter was seen exclusively among the Mozabite of Algeria (55.0%) and Druze of Israel (14.3%).
- In Oceania, there was a tie for three genotypes: 12-12-17-17, 12-16-16-17 and 12-17-17-17 (14.2% each). The first two were seen in both Papuans (7.7 and 15.4%, respectively) and Melanesians (25.0 and 12.5%, respectively), while the last was observed only among Melanesians (37.5%).

The genotype diversity was 0.98 for Europe, 0.97 for Central and East Asia, 0.95 for Africa, 0.94 for Oceania, 0.92 for the Middle East, and 0.90 for the Americas (Table 2).

The genotypes were submitted to a hierarchical AMOVA using the Arlequin program (Schneider et al., 2000), and the results of this analysis are displayed in Table 3. Our analysis showed low levels of worldwide genetic structure: 88.42% of the genetic variance was found

within populations, 9.62% between populations within regions and 1.96% between regions. Focusing on each region, the within-population component was responsible for more than 87% of the genetic variance, except for the Middle East people and the Americas, which exhibited, lower within-population variance and 21.78 and 33.20% of the variance between populations within continents, respectively.

Table 3. Analysis of molecular variance (AMOVA) for *DYS464*.

Region	Number of regions	Number of populations	Variance components (%)		
			Within populations	Between populations	Between regions
World	1	52	88.71	11.29	-
World	7	52	88.42	9.62	1.96
Africa	1	7	94.55	5.45	-
Middle East	1	4	78.22	21.78	-
Central Asia	1	8	95.91	4.09	-
East Asia	1	18	87.99	12.01	-
Oceania	1	2	94.90	5.10	-
Europe	1	8	97.18	2.82	-
Americas	1	5	66.80	33.20	-

DISCUSSION

DYS464 is indeed highly polymorphic, but it has some characteristics that render the establishment of its evolutionary history difficult. First of all, the four copies are identical and one cannot assign each peak in the electropherogram to a specific locus. Thus, the same genotype may correspond to several haplotypes with different permutation of alleles. In this fashion, genotypes are degenerate. Second, *DYS464* has a high mutation rate, estimated at 2.86×10^{-2} , approximately 10 times higher than the average mutation rate for Y-chromosome microsatellites (Berger et al., 2003). Third, the tetralocal palindromic structure of *DYS464* renders it susceptible to allelic conversion mediated by intrachromosomal mitotic recombination events (Rozen et al., 2003).

Considering these complexities, it is no surprise that the phylogeography of *DYS464* may occasionally appear strange. For instance, the most common genotype in Central Asia, 12-15-15-16, was also seen in the Bedouins of Israel (7.1%), Uygurs of China (12.5%), Sardinians of Italy (6.3%), Adygei of the Caucasus (14.3%), and Russians (6.3%) and in no other populations. This "saltatory" distribution does not make evolutionary sense, unless we consider the possibility that Bedouins, Uygurs and Sardinians could have different haplotypes of the same four loci in different order and/or that one or more conversion events occurred leading to loss of heterozygosity (note the presence of two alleles 15).

On the other hand, many populations exhibit high frequencies of specific genotypes that may prove to be useful in evolutionary studies. For instance, the fact that genotype 14-14-15-15 was seen in 88.9% of the Yakut of Siberia, suggests that the male component of this population could have experienced a population bottleneck or a founder effect. Research done by other groups supports this notion (Pakendorf et al., 2002; Khar'kov et al., 2008).

Africa had the highest percentage of unique genotypes. This is compatible with the view that modern man emerged in Africa and migrated from that continent to populate all other areas of Earth (reviewed in Pena, 2007) and that the movement of modern humanity out-of-Africa was associated with a population size bottleneck and reduction of variability (Yu et al., 2002; Torroni et al., 2006).

However, contrary to most other Y-chromosomal loci studied (Jobling and Tyler-Smith, 2003), *DYS464* did not display its largest genotypic variability in Africa. One probably cannot and should not assign too much significance to this observation, especially if one considers the genotypic degeneracy of *DYS464*, its high mutation rate and also the fact the African samples present in the HGDP-CEPH Human Genome Diversity Cell Line Panel may only be a small portion of the total African variability.

A hierarchical AMOVA was performed, and the results revealed low levels of worldwide genetic structure - the within-population component was responsible for 88.42% of the genetic variability, with 9.62% between populations within regions and 1.96% between regions. These relative levels of within-population variance are higher than those observed in other worldwide surveys with Y-chromosome markers: 76.8% by Kayser et al. (2001) and 66.5% by Wilder et al. (2004). We believe that this can be attributed to the high mutation rate of *DYS464*.

It has been suggested that the differentiation of microsatellite polymorphisms in human populations is driven more by genetic drift than by mutation pressure (Perez-Lezaun et al., 1997). Since the effective population size of the Y-chromosome is around one-quarter of that in autosomes, in theory, genetic drift should have a deep impact on *DYS464* variation. On the other hand, the high mutation rate of *DYS464* will counteract the effects of drift, producing homoplasmy and decreasing the between-population variance.

In conclusion, our results demonstrate that the *DYS464* microsatellite on the human Y-chromosome shows indeed very high variability in geographically widespread human populations, with genotype diversities above 0.90 in all continents studied. Even though it only provides the tale of a single “gene” and despite that its tetra-local structure often does not permit detailed phylogeographical inferences, *DYS464* still appears to be an informative tool for population and evolutionary studies.

ACKNOWLEDGMENTS

Research supported by CAPES and CNPq of Brazil.

REFERENCES

- Berger B, Niederstatter H, Brandstatter A and Parson W (2003). Molecular characterization and Austrian Caucasian population data of the multi-copy Y-chromosomal STR *DYS464*. *Forensic Sci. Int.* 137: 221-230.
- Butler JM and Schoske R (2005). U.S. population data for the multi-copy Y-STR locus *DYS464*. *J. Forensic Sci.* 50: 975-977.
- Cann HM, de Toma C, Cazes L, Legrand MF, et al. (2002). A human genome diversity cell line panel. *Science* 296: 261-262.
- Carvalho-Silva DR, Santos FR, Rocha J and Pena SD (2001). The phylogeography of Brazilian Y-chromosome lineages. *Am. J. Hum. Genet.* 68: 281-286.
- Carvalho-Silva DR, Tarazona-Santos E, Rocha J, Pena SD, et al. (2006). Y-chromosome diversity in Brazilians: switching perspectives from slow to fast evolving markers. *Genetica* 126: 251-260.
- Chiaroni J, Underhill PA and Cavalli-Sforza LL (2009). Y-chromosome diversity, human expansion, drift, and cultural evolution. *Proc. Natl. Acad. Sci. U. S. A.* 106: 20174-20179.
- Gusmão L, Butler JM, Carracedo A, Gill P, et al. (2006). DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Int. J. Legal Med.* 120: 191-200.
- Hanson EK and Ballantyne J (2006). Comprehensive annotated STR physical map of the human Y-chromosome: Forensic implications. *Leg. Med.* 8: 110-120.
- Jobling MA and Tyler-Smith C (2003). The human Y-chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* 4: 598-612.

- Kayser M, Krawczak M, Excoffier L, Dieltjes P, et al. (2001). An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am. J. Hum. Genet.* 68: 990-1018.
- Khar'kov VN, Stepanov VA, Medvedev OF, Spiridonova MG, et al. (2008). The origin of Yakuts: analysis of Y-chromosome haplotypes. *Mol. Biol.* 42: 226-237.
- Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, et al. (2001). The AZFc region of the Y-chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* 29: 279-286.
- Pakendorf B, Morar B, Tarskaia LA, Kayser M, et al. (2002). Y-chromosomal evidence for a strong reduction in male population size of Yakuts. *Hum. Genet.* 110: 198-200.
- Pena SDJ (2007). The Evolution and Structure of Human Genetic Diversity. In: Pharmacogenomics in Admixed Populations (Suarez-Kurtz G, ed.). Landes Bioscience, Austin, 1-11. Available at [http://www.eurekah.com/chapter/3190]. Accessed...
- Perez-Lezaun A, Calafell F, Mateu E, Comas D, et al. (1997). Microsatellite variation and the differentiation of modern humans. *Hum. Genet.* 99: 1-7.
- Redd AJ, Agellon AB, Kearney VA, Contreras VA, et al. (2002). Forensic value of 14 novel STRs on the human Y-chromosome. *Forensic Sci. Int.* 130: 97-111.
- Rozen S, Skaletsky H, Marszalek JD, Minx PJ, et al. (2003). Abundant gene conversion between arms of palindromes in human and ape Y-chromosomes. *Nature* 423: 873-876.
- Santos FR, Epplen JT and Pena SDJ (2003). Testing Deficiency Paternity Cases With an Y-linked Tetranucleotide Repeat Polymorphism. In: DNA Fingerprinting: State of the Science (Pena SDJ, Chakraborty R, Epplen JT and Jeffreys AJ, eds.). Birkhäuser Verlag, Basel, 261-266.
- Schneider S, Roessli D and Excoffier L (2000). Arlequin Ver 2.000: A Software for Population Genetics Data Analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, et al. (2003). The male-specific region of the human Y-chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825-837.
- Torrioni A, Achilli A, Macaulay V, Richards M, et al. (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22: 339-345.
- Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, et al. (2004). Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat. Genet.* 36: 1122-1125.
- Yu N, Chen FC, Ota S, Jorde LB, et al. (2002). Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* 161: 269-274.