UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

DISSERTAÇÃO DE MESTRADO

# Identificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e aplicações biomédicas.

ORIENTADO: Giordano Bruno Soares Souza
ORIENTADOR: Eduardo Martín Tarazona Santos

BELO HORIZONTE
Janeiro – 2010

Giordano Bruno Soares Souza

# Identificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e aplicações biomédicas.

Dissertação apresentada ao Programa de Pós-Graduação de Genética da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção de título de Mestre em Genética.

Orientador: Eduardo Martin Tarazona Santos

Belo Horizonte
2010

043 Souza, Giordano Bruno Soares.
Identificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e aplicações biomédicas [manuscrito] / Giordano Bruno Soares Souza. – 2010.

107 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Eduardo Martín Tarazona Santos.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Variação Genética**.** 3. Características da População. 4. Evolução humana. 5. Seleção Genética. I. Santos, Eduardo Martín Tarazona. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575

Ficha catalográfica elaborada pela bibliotecária Rosilene Moreira Coelho de Sá – CRB 6 – 2726

Pós-Graduação em Genética
Departamento de Biologia Geral, ICB
Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 - C.P. 486 - Pampulha - 31270-901 - Belo Horizonte - MG
e-mail: pg-gen@icb.ufmg.br FAX: (+31) - 3409-2570

# "Indentificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e implicações biomédicas."

**Giordano Bruno Soares Souza**

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Eduardo M. Tarazona Santos - Orientador
UFMG

Profa. Dr. Cleusa Graça da Fonseca
UFMG

Profa. Dra. Valeria Sandrim
Santa Casa de Belo Horizonte

Belo Horizonte, 20 de janeiro de 2010.

## AGRADECIMENTOS

Ao meu orientador Prof. Eduardo Martin Tarazona Santos, pela oportunidade de realizar esse trabalho, pelo incentivo constante, pelos ensejos a mim apresentados em diversas ocasiões e pela paciência e compreensão.

Aos colegas do Laboratório de Diversidade Genética Humana, pelo apoio e companheirismo. Aos integrantes do *National Cancer Institute*, na figura do Dr. Stephen Chanock, pelos dados disponibilizados para análise.

E em especial, às pessoas mais importantes nesses dois anos de caminhada: meus pais, pelo exemplo constante e apoio; meus avós pelo carinho e incentivo; minha irmã, que apesar da distância continua tão próxima e querida; aos amigos pelo alento e disponibilidade e à Camila, companheira nos momentos mais difíceis e a quem admiro tanto.

**SUMÁRIO**

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE QUADROS

## LISTA DE ABREVIATURAS

ACP – Análise de Componentes Principais

AMOVA – Análise de Variância Molecular

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

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

DL – Desequilíbrio de Ligação

DLM – Desequilíbrio de Ligação gerado por Miscigenação

DNA – Ácido Desoxirribonucleico

E-HW – Equilíbrio de Hardy-Weinberg

GC – *Genomic Control* (Controle Genômico)

GO – *Gene Ontology*

$H_E$ – Heterozigosidade Esperada

HGDP – *Human Genome Diversity Project* (Projeto de Diversidade Genética Humana)

MIAs – Marcadores Informativos de Ancestralidade

NCI – *National Cancer Institute*

NIH – *National Institute of Health*

PCR – Reação em cadeia da polimerase

RAO – *Recent Africa Origin* (Origem Recente Africana)

SA – Structured Association (Associação Estruturada)

SGDB – Sistema de Gerenciamento de Bancos de Dados

SQL – *Structured Query Language*

SNP – Single Nucleotide Polymorphism (Polimorfismo de Única Base)

## ANEXOS

## RESUMO

A variabilidade genética está associada à diferenciação fenotípica encontrada entre as populações humanas. Ainda que grande parte dessa variabilidade se dê entre indivíduos da mesma população e a interação com o ambiente seja fundamental na determinação do fenótipo, o estudo das variantes que se encontram estruturadas geneticamente na população humana é essencial em dois importantes campos da biologia: evolução e medicina. O presente trabalho objetiva identificar polimorfismos genéticos que possam contribuir para a ocorrência de falso-positivos em estudos de associação caso-controle ou que possam estar sob seleção natural. Para tal, analisamos 1442 SNPs, distribuídos por 411 genes importantes em imunidade, carcinogênese e farmacogenética. A amostragem compreende 1198 indivíduos provenientes do HGDP (Projeto de Diversidade Genética Humana), SNP500Cancer e populações nativas do Peru e Equador e pertencentes a 60 populações de várias localidades geográficas. Através da Análise de Variância Molecular identificamos 196 polimorfismos localizados em 111 genes que podem causar resultados espúrios em estudos de associação caso-controle realizados em populações miscigenadas tri-híbridas compostas por populações parentais européias, nativo-americanas e do oeste africano. Identificamos ainda diversos polimorfismos que podem causar o mesmo efeito em populações miscigenadas de composição di-híbrida. Evidenciamos ainda a existência de 36 SNPs com alta diferenciação entre populações do Leste da Ásia e Nativo-Americanos, que podem estar sob seleção positiva ou devem sua diferenciação ao *allele surfing*. Evidenciamos através da análise de heterozigosidade, SNPs candidatos aos três tipos de seleção: balanceadora, purificadora e positiva. A análise combinada de distintas estatísticas genéticas em diferentes populações pode auxiliar não apenas no desenho de estudos de associação, como pode elucidar os mecanismos de prevalência das doenças humanas e auxiliar na caracterização da história evolutiva das populações.

## ABSTRACT

Genetic diversity is related to phenotypic differentiation among human populations. Even though most of this variation occurs among individuals and interaction with environment plays key role in phenotypic determination, the study of variants that are genetic structured in human populations is crucial in two major fields of biology: evolution and medicine. Investigate loci with differentiated allelic frequency in human populations allows improvement of case-control association studies in admixed populations and discover of genetic polymorphisms that have experienced natural selection. In this context, we studied hierarchical genetic structure of 1442 SNPs located in 411 genes related to immune response, carcinogenesis and pharmacogenetics. This genetic characterization was made for 1198 individuals from 60 worldwide populations belonging HGDP, SNP500Cancer and Native-American populations of Ecuador and Peru. The following results emerge from Analysis of Molecular Variance approach: we identified 196 polymorphisms allocated in 111 genes that can lead to spurious association in case-controls studies performed in tri-hybrid populations like Brazilian population. In this context, many genetic markers were recognized in alternative models of bi-hybrid populations formed from European, Native-American and West African populations. We show 36 SNPs highly differentiated among East Asians and Amerindians that could have suffered positive selection or allele surfing. Using data from loci heterozigosity we appointed some loci possibly under one of three types of natural selection: positive, purifying and balancing. The combined use of distinct genetic statistics in different populations can improve not only design of epidemiological studies as can clarify aspects of prevalence in human diseases and to tells a little about human history.

# 1 INTRODUÇÃO

## 1.1 Estrutura Genética das Populações Humanas

Projetos de diversidade genômica humana têm como objetivo analisar a variabilidade genética entre indivíduos e populações de modo a entender as origens e o processo de evolução do *Homo sapiens* (Cann, 1998). O potencial dos dados genéticos para prover informações sobre a história e geografia das populações humanas já era conhecido a partir do estudo de proteínas, ainda no começo do século XX. Entretanto, ainda recentemente, a coleta de dados permanecia um esforço fragmentado (Cavalli-Sforza, 2005). Nos últimos anos, painéis de indivíduos de diferentes partes do globo têm sido construídos permitindo melhor caracterização da variabilidade genética humana.

No presente estudo, utilizamos três painéis de indivíduos de diferentes regiões geográficas, sendo que dois desses pertencem a grandes projetos de descrição da variabilidade humana: HGDP e SNP500Cancer; o terceiro painel compreende amostras de nativo-americanos do Laboratório de Diversidade Genética Humana.

O Projeto de Diversidade Genômica Humana (HGDP – *Human Genome Diversity Project*) tem como objetivos: coletar e preservar amostras biológicas de várias populações, conservando e disponibilizando painéis de amostras de DNA a pesquisadores de todo o mundo; proporcionar recursos para o estudo da diversidade genética humana; e compartilhar os resultados das tipagens realizadas (Cann, 1998). O HGDP objetiva ainda entender como e quando os padrões de diversidade observados foram formados provendo informações que podem ser valiosas em várias áreas da pesquisa biomédica. A fundação Jean Dauseet, em Paris, estoca linhagens celulares de linfoblastos de 1056 indivíduos provenientes de 52 populações distribuídas por todo o globo. Diferentes metodologias têm sido propostas para a investigação da diversidade genética humana. Contudo, alguns dos principais estudos acerca da variabilidade genética humana são os estudos de variabilidade genética baseados em gene e análises de cladogramas populacionais. O primeiro tipo de estudo é baseado no cálculo das distâncias genéticas entre populações a partir de freqüências gênicas e a média entre vários loci. O segundo tipo de estudo é baseado

na comparação par-a-par das diferenças de freqüências alélicas entre populações, permitindo a descrição da divergência entre populações (Cavalli-Sforza. 2005).

O banco de dados SNP500Cancer provê informações de seqüenciamento e genotipagem para SNPs candidatos a associação com doenças complexas, como o câncer. Esse banco de dados faz parte do Projeto de Anatomia Genômica do Câncer do *National Cancer Institute* – NCI, dos EUA. SNP500Cancer estuda amostras de DNA de 102 indivíduos dos repositórios celulares *Coriell Institute of Medical Research*, sendo esses sujeitos representantes de quatro etnicidades: Afro-americanos, Caucasianos, Hispânicos e Asiáticos (Packer *et al*., 2006).

O trabalho de caracterização genética das populações do CEPH-HGDP, SNP500Cancer e Nativo-Americanos foi realizado pela mestre Juliana Chevitarese (2009) em sua dissertação de mestrado. As mesmas populações e loci foram utilizados no presente estudo e por isso, a descrição da estrutura genética desse conjunto de dados é equivalente para os dois estudos.

Os maiores valores encontrados de $F_{ST}$ e $F_{IT}$ são observados nos grupos América Central, Oceania e América do Sul. Tais valores refletem estruturação entre as populações desses grupos, o efeito Wahlund. Os grupos do Leste e Oeste Africano apresentam valores intermediários enquanto os grupos correspondentes à Eurásia apresentam os menores valores para essas estatísticas. Os valores de $F_{IS}$ mais baixos encontrados correspondem aos continentes América do Sul e Oceania (Chevitarese, 2009) (Tabela 1).

| **Tabela 1: Grupos populacionais estudados e parâmetros de estimativa da diversidade interna aos grupos.** | | | | | |
|---|---|---|---|---|---|
| **Grupos Populacionais** | N[1] | Het. Esp.[2] | $F_{ST}$ | $F_{IS}$ | $F_{IT}$ |
| Leste Africano | 63 | 0,318 | 0,051 | -0,006 | 0,046 |
| Oeste Africano | 56 | 0,289 | 0,055 | 0,007 | 0,061 |
| Oriente Médio | 174 | 0,349 | 0,018 | 0,023 | 0,040 |
| Europa | 158 | 0,349 | 0,013 | 0,005 | 0,017 |
| Centro Sul Asiático | 198 | 0,354 | 0,019 | 0,039 | 0,057 |
| Leste Asiático | 242 | 0,310 | 0,020 | 0,003 | 0,023 |
| Oceania | 32 | 0,272 | 0,101 | -0,019 | 0,084 |
| América Central | 49 | 0,275 | 0,078 | -0,003 | 0,075 |
| América do Sul | 124 | 0,299 | 0,135 | -0,019 | 0,118 |
| [1] N: Número de indivíduos; [2] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de Hardy Weinberg; Modificada de Chevitarese (2009). | | | | | |

Em relação às populações do SNP500Cancer, os maiores valores de $F_{IS}$ foram encontrados para a população asiática, enquanto o menor valor encontrado corresponde aos Euro-descendentes. A heterozigosidade esperada nesse painel de indivíduos não apresentou grandes diferenças. O maior valor observado para Hispânicos reflete o impacto da miscigenação na variabilidade de uma população (Tabela 2).

| **Tabela 2: Populações estudadas do *SNP500Cancer* e parâmetros de estimativa da diversidade intra-populacional.** | | | | | | |
|---|---|---|---|---|---|---|
| **População** | N[1] | Het. Esp.[2] | *Ranking* Het. Esp.[3] | $F_{IS}$ | *Ranking* $F_{IS}$[4] | SNPs fora do HWE[5] |
| África | 24 | 0,310 | 1 | 0,041 | 2 | 1 |
| Europa | 31 | 0,350 | 3 | 0,001 | 1 | 2 |
| Ásia | 24 | 0,324 | 2 | 0,091 | 4 | 9 |
| Hispânicos | 23 | 0,350 | 4 | 0,048 | 3 | 1 |
| [1] N: Número de indivíduos da população; [2] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de *Hardy Weinberg*; [3] *Ranking* Het. Esp.: Classificação em relação aos valores da média da heterozigosidade esperada (1 mínimo, 4 máximo); [4] Ranking $F_{IS}$: Classificação em relação aos valores de $F_{IS}$ (1 mínimo, 4 máximo); [5] SNPs fora do HWE: Número de SNPs para os quais foi rejeitada a hipótese de equilíbrio de *Hardy Weinberg* ($p<10^{-4}$); Modificada de Chevitarese (2009). | | | | | | |

Chevitarese (2009) observou que os menores valores do Coeficiente de Endocruzamento ocorreram em populações da América do Sul, Suruí e Piapoco-Curripaco; e da Oceania, Melanésia. Os maiores valores para essa estatística foram relatados para as populações San Martín e Beduínos. Entre as estimativas de heterozigosidade esperada, os maiores valores de $H_E$ estão associados às populações do Oriente Médio, Centro Sul Asiático e Europa. Os menores valores referem-se às populações nativo-americanas e da Oceania. E em geral, as populações do Oeste Sul-Americano, apresentaram maiores valores de heterozigosidade esperada que as populações do Leste Sul-Americano (Tabela 3).

**Tabela 3: Distribuição das populações estudadas (HGDP e nativo-americanos do nosso laboratório) ao longo de seus respectivos grupos e parâmetros de estimativa da diversidade intrapopulacional.**

| Grupo | População[1] | N[2] | Het. Esp.[3] | *Ranking* Het. Esp.[4] | $F_{IS}$ | *Ranking* $F_{IS}$[5] | SNPs fora do HWE[6] |
|---|---|---|---|---|---|---|---|
| Leste Africano | Pigmeu Biaka (6) | 31 | 0,271 | 13 | -0,012 | 14 | 2 |
| Leste Africano | Pigmeu Mbuti (5) | 13 | 0,246 | 7 | 0,011 | 36 | zero |
| Leste Africano | Bantu NE (1) | 11 | 0,292 | 19 | -0,006 | 18 | zero |
| Leste Africano | Bantu SE e SO (1) | 8 | 0,284 | 17 | 0,025 | 46 | zero |
| Oeste Africano | Mandenka (2) | 24 | 0,281 | 15 | -0,001 | 22 | 2 |
| Oeste Africano | Ioruba (3) | 25 | 0,282 | 16 | 0,019 | 42 | zero |
| Oeste Africano | San (4) | 7 | 0,228 | 2 | -0,015 | 11 | zero |
| Oriente Médio | Mozabite (7) | 30 | 0,336 | 38 | 0,006 | 29 | 2 |
| Oriente Médio | Beduína (16) | 48 | 0,340 | 40 | 0,049 | 55 | zero |
| Oriente Médio | Drusa (17) | 47 | 0,344 | 47 | 0,023 | 45 | 3 |
| Oriente Médio | Palestina (18) | 49 | 0,348 | 51 | 0,009 | 34 | 7 |
| Europa | Francesa (12) | 29 | 0,342 | 43 | 0,025 | 49 | 20 |
| Europa | Basca (11) | 24 | 0,341 | 42 | -0,023 | 9 | 3 |
| Europa | Sardenha (14) | 28 | 0,339 | 39 | -0,001 | 23 | 3 |
| Europa | Bérgamo (13) | 13 | 0,346 | 50 | 0,027 | 50 | zero |
| Europa | Toscana (15) | 8 | 0,345 | 49 | -0,001 | 24 | zero |
| Europa | Orcadiana (8) | 16 | 0,331 | 37 | 0,025 | 48 | 1 |
| Europa | Adygei (9) | 15 | 0,353 | 56 | 0,018 | 41 | 1 |
| Europa | Russa NO (10) | 25 | 0,351 | 55 | -0,008 | 16 | 6 |
| Centro Sul Asiático | Brahui (20) | 25 | 0,350 | 54 | 0,017 | 40 | 3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Centro Sul Asiático | Balochi (19) | 25 | 0,341 | 41 | 0,046 | 52 | 2 |
| Centro Sul Asiático | Hazara (25) | 25 | 0,342 | 45 | 0,021 | 44 | zero |
| Centro Sul Asiático | Makrani (21) | 25 | 0,349 | 53 | 0,046 | 53 | zero |
| Centro Sul Asiático | Sindhi (22) | 24 | 0,349 | 52 | 0,041 | 51 | 16 |
| Centro Sul Asiático | Pathan (23) | 24 | 0,343 | 46 | 0,047 | 54 | 10 |
| Centro Sul Asiático | Kalash (27) | 25 | 0,320 | 36 | -0,007 | 17 | 1 |
| Centro Sul Asiático | Burusho (24) | 25 | 0,345 | 48 | 0,011 | 37 | 1 |
| Leste Asiático | Camboja (43) | 10 | 0,317 | 35 | 0,009 | 33 | zero |
| Leste Asiático | Han (28 e 29) | 39 | 0,299 | 23 | 0,025 | 47 | 1 |
| Leste Asiático | Tujia (37) | 10 | 0,307 | 30 | 0,012 | 38 | zero |
| Leste Asiático | Yizu-Yi (40) | 10 | 0,303 | 29 | 0,002 | 27 | zero |
| Leste Asiático | Miaozu-Miao (34) | 10 | 0,301 | 27 | -0,025 | 6 | zero |
| Leste Asiático | Oroqen (35) | 10 | 0,300 | 25 | -0,029 | 5 | zero |
| Leste Asiático | Daur (31) | 10 | 0,296 | 20 | 0,007 | 30 | zero |
| Leste Asiático | Mongólia (41) | 10 | 0,314 | 34 | 0,009 | 32 | zero |
| Leste Asiático | Hezhen (32) | 9 | 0,300 | 26 | -0,024 | 7 | zero |
| Leste Asiático | Xibo (39) | 9 | 0,307 | 31 | 0,01 | 35 | zero |
| Leste Asiático | Uigur (26) | 10 | 0,342 | 44 | 0,008 | 31 | zero |
| Leste Asiático | Daí (30) | 10 | 0,303 | 28 | -0,003 | 20 | zero |
| Leste Asiático | Lahu (33) | 10 | 0,289 | 18 | -0,016 | 10 | zero |
| Leste Asiático | She (36) | 10 | 0,300 | 24 | -0,041 | 4 | zero |
| Leste Asiático | Naxi (42) | 10 | 0,298 | 21 | 0,004 | 28 | zero |
| Leste Asiático | Tu (38) | 10 | 0,312 | 32 | -0,002 | 21 | zero |
| Leste Asiático | Yakut (45) | 25 | 0,313 | 33 | -0,004 | 19 | 1 |
| Leste Asiático | Japonesa (44) | 30 | 0,299 | 22 | 0,002 | 26 | zero |
| Oceania | Papua (47) | 17 | 0,250 | 9 | 0,002 | 25 | 1 |
| Oceania | Melanésia (46) | 15 | 0,267 | 10 | -0,045 | 3 | 1 |
| América Central | Pima (52) | 24 | 0,246 | 6 | 0,02 | 43 | 2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| América Central | Maia (51) | 25 | 0,278 | 14 | -0,013 | 13 | 3 |
| América do Sul | Piapoco e Curripaco (50) | 13 | 0,244 | 5 | -0,054 | 2 | zero |
| América do Sul | Karitiana (48) | 23 | 0,232 | 4 | -0,023 | 8 | 1 |
| América do Sul | Suruí (49) | 21 | 0,203 | 1 | -0,102 | 1 | zero |
| América do Sul | Cayapa* (53-a) | 7 | 0,246 | 8 | 0,017 | 39 | zero |
| América do Sul | Quechua* (53-b) | 22 | 0,270 | 12 | -0,015 | 12 | 10 |
| América do Sul | San Martín* (53-c) | 17 | 0,270 | 11 | 0,065 | 56 | 1 |
| América do Sul | Matsiguenga* (53-d) | 21 | 0,231 | 3 | -0,01 | 15 | 1 |

* Amostras populacionais disponíveis em nosso laboratório; [1] Número entre parêntesis referentes à localização geográfica da população no mapa das Figuras 4 e 5; [2] N: Número de indivíduos da população; [3] Het. Esp.: Média da heterozigosidade esperada para todos os loci, sob a hipótese de equilíbrio de *Hardy Weinberg*; [4] *Ranking* Het. Esp.: Classificação em relação aos valores da média da heterozigosidade esperada (1 mínimo, 56 máximo); [5] Ranking $F_{IS}$: Classificação em relação aos valores de $F_{IS}$ (1 mínimo, 56 máximo); [6] SNPs fora do HWE: Número de SNPs para os quais foi rejeitada a hipótese de equilíbrio de *Hardy Weinberg* ($p<10^{-4}$); Modificada de Chevitarese (2009).

O valor de diferenciação entre populações ($F_{ST}$) nesse conjunto de dados alcança o valor de 0,121, número próximo ao observado por outros autores (Nei; Roychoudhury, 1982; Barbujani *et al*., 1997; Rosenberg *et al*., 2005).

## 1.2 Diversidade Genética e Povoamento das Américas

O processo histórico de povoamento da América, iniciado no Pleistoceno (Bonatto; Salzano, 1997), ainda é controverso em relação ao número de levas migratórias, à rota seguida pelas primeiras populações, à idade dos primeiros assentamentos no continente e *pool* gênico dos colonizadores (Corella *et al*., 2007). Atualmente a hipótese mais aceita é de que apenas uma leva migratória tenha contribuído efetivamente para a formação do pool gênico das populações nativo-americanas e a rota de colonização tenha ocorrido a partir da costa (Wang *et al*., 2007). Entretanto, há consenso em relação a certos eventos evolutivos que atuaram sobre as populações pioneiras na exploração do continente, tal como a deriva (efeito

de gargalo) que imprimiu sua assinatura no genoma das populações nativo-americanas, reduzindo a diversidade genética dentro dos diversos grupos populacionais.

**Tabela 4: Comparação entre os níveis de $H_E$ e $F_{ST}$ entre diversas populações humanas**

| Região Geográfica | Nº de populações | Heterozigosidade (*Pooled*) | Heterozigosidade (Média) | $F_{ST}$(X100) |
|---|---|---|---|---|
| **Mundial** | 78 | 0,7400 | 0,6850 | 7,1 |
| **África** | 7 | 0,7740 | 0,7540 | 3,0 |
| **Europa** | 8 | 0,7320 | 0,7280 | 0,8 |
| **Oriente Médio** | 4 | 0,7400 | 0,7330 | 1,4 |
| **Centro Sul Asiático** | 9 | 0,7380 | 0,7300 | 1,3 |
| **Leste Asiático** | 19 | 0,7140 | 0,7040 | 1,4 |
| **Oceania** | 2 | 0,6900 | 0,6880 | 6,4 |
| **América** | 29 | 0,6760 | 0,6230 | 8,1 |
| **América do Norte** | 3 | 0,6970 | 0,6840 | 3,4 |
| **América Central** | 8 | 0,6690 | 0,6380 | 5,5 |
| **Oeste da América do Sul** | 10 | 0,6720 | 0,6350 | 5,7 |
| **Leste da América do Sul** | 8 | 0,6390 | 0,5710 | 14,7 |

Níveis de heterozigosidade e $F_{ST}$ estimados utilizando microssatélites. Os valores de heterozigosidade são significativamente menores em populações nativo-americanas. $F_{ST}$ multiplicado por 100 vezes por razões de conveniência. Modificada de: Wang *et al*., 2007

Os primeiros trabalhos buscando a caracterização da variabilidade genética em populações ameríndias utilizavam-se de marcadores como grupos sanguíneos (Rodriguez *et al*., 1963, Long *et al*., 1991, O'Rourke *et al*., 1992), HLA I (Parham; Ohta, 1996) e HLA II (Erlich *et al*., 1997). Ao final da década de 90, os polimorfismos de uma única base (SNP – *Single Nucleotide Polymorphism*) passaram a ser amplamente utilizados em análises genômicas de larga escala, análises genéticas de doenças complexas e em estudos globais de genética de populações (Brookes, 1999). Isso se deve ao fato dos SNPs estarem distribuídos ao longo do genoma, sendo capazes de

fornecer informações independentes de diversas regiões gênicas, sobre genes específicos de interesse biomédico e em alguns casos, de atuar como marcadores informativos de ancestralidade (MIAs). A limitação deste enfoque ocorre devido aos elevados custos da genotipagem em estudos populacionais: os dados existentes sobre as populações humanas encontram-se reduzidos a um pequeno número de indivíduos. Entretanto, custos menores de genotipagem têm permitido a disseminação de estudos envolvendo a amostragem de grande número de indivíduos (Nelson *et al*., 2008).

Atualmente, grandes quantidades de dados sobre a variabilidade humana têm sido geradas devido às novas técnicas de genotipagem e seqüenciamento baseadas em microarranjos. Concomitantemente há o surgimento de grandes bases de dados públicas que armazenam e permitem o acesso gratuito de pesquisadores a essas informações, tais como SNP500Cancer (Packer *et al*., 2004; 2006) , International HapMap Project (*International HapMap Consortium*, 2003; 2005; Frazer *et al*., 2007), Seattle SNPs (http://pga.mbt.washington.edu; Crawford; Akey; Nickerson, 2005) além do desenvolvimento de ferramentas de análises robustas que permitem descrições confiáveis da estrutura genética e inferências sobre os processos evolutivos que a determinaram. Entretanto, as populações nativo-americanas são usualmente negligenciadas em estudos globais de caracterização genética.

A identificação de genes e variantes causais para doenças complexas representa um importante passo em direção à elucidação dos mecanismos genéticos envolvidos na patogênese de doenças complexas, e em alguns casos, na melhora no tratamento, diagnóstico e prevenção de doenças. Quando se aborda a variabilidade genética existente por trás das doenças comuns, deve-se considerar que vários fatores interagem no desencadeamento da patogenia, inclusos fatores ambientais e múltiplas variantes genéticas (*International HapMap Consortium*, 2005).

Quanto à estruturação genética, as populações da América do Sul apresentam altos valores de correlação entre a distância geográfica e a heterozigosidade esperada em comparação a outras populações não-africanas e os menores valores de heterozigosidade, que decrescem linearmente a partir da África. Essas observações também demonstram que as populações nativo-americanas apresentam os menores

índices de diversidade genética e as populações africanas, os maiores. Tais dados corroboraram a hipótese de origem humana a partir da África e que a diáspora e expansão da população humana se deu através de sucessivos efeitos fundadores (Ramachandran *et al*., 2005).

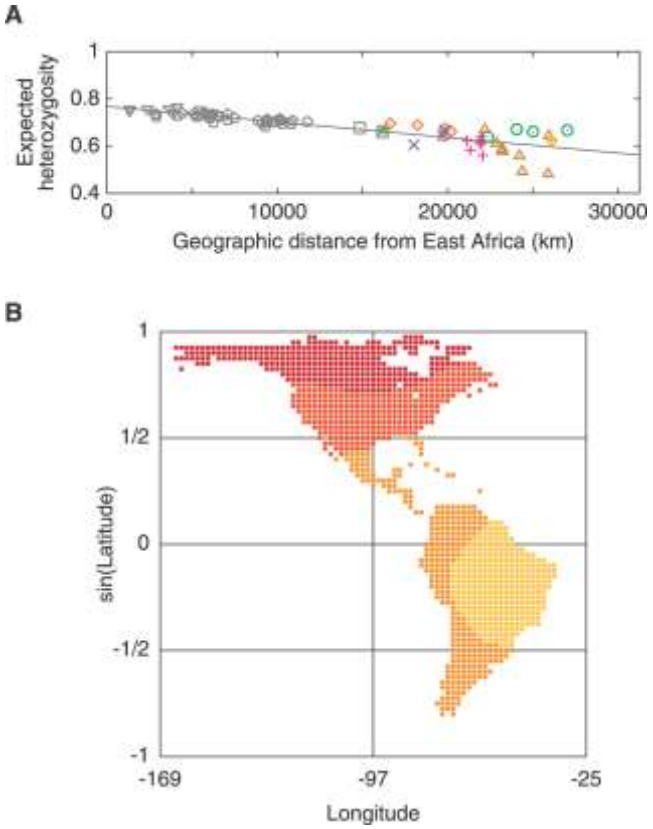



**Figura 1.**

A) Relação entre a distância geográfica a partir do Leste da África e o declínio da heterozigosidade. Quadrados e triângulos cinzas, representam, respectivamente, populações da Oceania e África Sub-Saariana. As populações não-americanas restantes são marcadas por pentágonos cinzas. Populações Nativo-americanas são representadas por formas geométricas coloridas.

B) Demonstração do declínio da heterozigosidade apenas para populações ameríndias. É possível observar o declínio no sentido Norte-Sul.

Em: Wang *et al*., 2007

Mesmo que a proporção do componente de variância interpopulacional seja pequena e existam poucos alelos “privados”, análises genotípicas com um suficiente número de loci (por exemplo, 377 microssatélites) permitem a inferência da ancestralidade genética sem a necessidade de informações sobre a localização dos indivíduos amostrados (Rosenberg *et al*., 2002).

### *1.2.1 Leste Asiático e a Contribuição ao Pool Gênico dos Nativo-Americanos*

Devido ao efeito fundador originado pela colonização do novo continente, a América, espera-se que grande parte dos polimorfismos com alta variância em suas freqüências alélicas tenha suas dessemelhanças devidas a efeitos aleatórios, como a deriva e demográficos, tais como expansão populacional, endogamia ou exogamia

(Mulligan *et al*., 2004)

Além disso, o efeito fundador experimentado pelas primeiras populações nativo-americanas pode ter impacto direto na saúde das populações nativo-americanas pioneiras. Isso porque os alelos em baixa freqüência tendem a se perder durante um efeito de gargalo populacional, por isso estima-se que essas primeiras populações sofressem com poucos distúrbios mendelianos. Entretanto, muitas doenças comuns estão aumentando sua prevalência em populações nativo-americanas, levando a crer que a perda de alelos raros não altera significativamente a genética de doenças complexas (Mulligan *et al*., 2004).

Habitualmente situa-se a população Yakut como a população do painel do CEPH mais próxima tanto geneticamente, quanto geograficamente das populações indígenas da América (Li *et al*., 2008). Entretanto, estudos demonstram que a colonização da parte oriental da Rússia pelos Yakuts é bem posterior a chegada do *Homo sapiens* à América. Estima-se que os Yakuts passaram a colonizar a atual região onde vivem, apenas no século XIII. Sendo que, anteriormente, habitavam regiões próximas ao lago Baikal. O processo de migração foi gradual e resultou na assimilação de outras etnias (Vitebsky, 1990).

Duas rotas poderiam ter sido usadas por populações pleistocênicas para alcançar o continente americano. Uma delas seguiria a partir do Crescente Fértil (Oriente Médio) pela costa do Sudeste Asiático e Mar do Japão (Oceano Pacífico) e a outra rota se daria pela Ásia Central (Zhang *et al*., 2007). Possíveis colonizadores provenientes das regiões meridionais da Ásia seguiriam pela costa sudeste do continente asiático efetivando a partir do sul a colonização do Leste Asiático, região cujos grupos falam predominantemente línguas sino-tibetanas. Assim, o trajeto de colonização se daria pela costa sudeste do continente asiático e seguiria ao norte, em direção ao estreito de Bering, a partir do Mar do Japão. Outros possíveis colonizadores seriam populações da Ásia Central, que nos dias atuais, predominantemente, pertencem ao grupo lingüístico Altaico, originário na atual região da Turquia. Caso estas tenham sido as populações fundadoras do continente americano, o caminho até a América teria se dado predominantemente pelas estepes russas e da Ásia Cental. Zhang e outros (2007) demonstram que a colonização do Leste Asiático

provavelmente se deu a partir do sul do continente e os níveis de miscigenação entre populações da Ásia Central e populações da região Nordeste da Ásia (direção centro-norte) são maiores que as observadas na direção centro-sul.

## 1.3 Medidas de Variabilidade Genética

### *1.3.1 Princípio de Hardy-Weinberg*

O Princípio de Hardy-Weinberg, descrito independentemente por Godfrey Hardy e Wilhelm Weinberg, prediz que uma população comportando-se de maneira mendeliana, mantém as freqüências alélicas inalteradas com o decorrer do tempo. Classifica-se uma população como mendeliana quando os seguintes pressupostos são mantidos: os organismos são diplóides, a população é infinita, de reprodução sexuada e não-preferencial e não há diferenças nas taxas de nascimento e reprodução entre machos e fêmeas. Entretanto, vários fatores são capazes de afastar as freqüências alélicas deste postulado, tais como deriva genética, seleção natural, endocruzamentos, casamentos preferenciais e efeito Wahlund (Li; Graubard, 2009; Lanchace, 2009). Particularmente, os quatro últimos fatores podem ser descritos pelas estatísticas *F* (Lanchace, 2009; Robertson; Hill, 1984). Desvios no Equilíbrio de Hardy-Weinberg relacionam-se ao $F_{IS}$ por indicarem que há alteração nas proporções esperadas de homozigotos e heterozigotos. Ou seja, quando a proporção de heterozigotos ou homozigotos se afasta daquilo que seria esperado sob Equilíbrio em uma subpopulação, obrigatoriamente, houve variação nas freqüências genotípicas.

Além de endogamia, estratificação populacional e seleção, os desvios do Equilíbrio de Hardy-Weinberg também podem ocorrer devido a associações com doenças. Entretanto as implicações não são bem exploradas. Além disso, tais desvios podem ocorrer por dificuldades em genotipar heterozigotos, mutações em sítios de ligação de primers de PCR e devido a polimorfismos comuns do tipo deleção. Dessa maneira, o E-HW é usualmente utilizado como medida de controle de qualidade (Balding, 2006).

*1.3.2 Heterozigosidade Esperada*

A variabilidade genética de uma população é usualmente medida através da heterozigosidade esperada média. E a variância entre os loci que compõe a heterozigosidade média pode ser dividia em dois componentes: a variância interlocus que está relacionada a eventos evolutivos, tais como mutação, seleção e deriva genética; e a variância intralocus que corresponde ao tamanho amostral e freqüências alélicas (Nei; Roychoudhury, 1974). Entre as componentes da variância interlocus, em tese, a ação da deriva poderia ser observada mais facilmente nesse estudo. Entretanto, algumas ressalvas devem ser consideradas. A primeira diz respeito à ação da seleção natural; há intenso debate sobre qual evento evolutivo é proeminente na diferenciação das populações: seleção natural ou deriva (Hurst, 2009). A diferenciação destes eventos é ainda mais complicada nesse estudo, pois os genes do sistema imune são mais propensos a estarem sob seleção (Nielsen *et al*., 2009, Ferrer-Admetlla *et al*., 2008).

Um complicador que poderia emergir nos cálculos de heterozigosidade é o viés de averiguação. O viés de averiguação ocorre devido ao processo de identificação de SNPs, onde alguns poucos indivíduos são seqüenciados para a descoberta de polimorfismos. Como poucos indivíduos são amostrados, a chance de um SNP ser descoberto é proporcional à sua freqüência naquela população e, assim, mais polimorfismos comuns são identificados em relação aos raros. Desse modo, para a população em que os SNPs foram observados a heterozigosidade média através dos sítios polimórficos é maior. Outro complicador é que as freqüências desses SNPs podem estar superestimadas ou subestimadas em relação a outras populações (Clark *et al*., 2005).

As razões acima expostas ajudam a elucidar os possíveis efeitos da utilização de nosso painel de marcadores, uma vez que eles foram identificados em populações européias e, além disso, a escolha para construção do painel priorizou marcadores polimórficos para essa população, o que acentua ainda mais o viés na população européia (Rogers; Jorde, 1996).

*1.3.3 Estatísticas F e AMOVA*

As estatísticas *F* descritas por Wright são utilizadas para estimar o valor de diferenciação genética entre as subpopulações. Essa abordagem estatística consiste na alocação da variabilidade genética em três coeficientes diferentes: (T) nível populacional, (S) subpopulacional, (I) individual. Assim, as estatísticas *F* podem ser descritas desta maneira: $F_{ST}$ é a correlação de alelos escolhidos aleatoriamente dentro da mesma subpopulação em relação a toda população, designando assim a proporção da diversidade genética devida às diferenças nas freqüências alélicas entre populações, quase sempre sendo representada por valores positivos. $F_{IS}$ é a correlação entre alelos interiormente a um indivíduo em relação à subpopulação a qual aquele indivíduo pertence, indicando assim o desvio médio das freqüências genotípicas em relação ao que seria esperado sob o Equilíbrio de Hardy-Weinberg dentro das subpopulações. Os valores positivos de $F_{IS}$ indicam deficiência de heterozigotos enquanto valores negativos indicam deficiência de homozigotos. $F_{IT}$ é a correlação dos alelos interiormente a um indivíduo em relação à população (total) a qual aquele indivíduo pertence, descrevendo assim o afastamento das freqüências genotípicas em relação ao esperado sob E-HW dentro da população (Holsinger; Weir, 2009).

A Análise de Variância Molecular se baseia no cálculo da diferenciação entre haplótipos, podendo, contudo, ser estendida a diferentes tipos de dados moleculares (Excoffier *et al*., 1992). A idéia central de AMOVA é análoga à análise de variância proposta por Weir e Cockerham, onde as variâncias das freqüências alélicas são calculadas e as médias entre dois ou mais grupos são testadas quanto à homogeneidade. Os cálculos de variância em AMOVA são baseados em medidas evolucionárias de distância entre os haplótipos (Holsinger; Weir, 2009). O AMOVA permite a partição da variância genética entre vários loci em dois componentes ou níveis hierárquicos: intrapopulacional e interpopulacional. Há modelos mais gerais que permitem, também, a partição da variância genética em três níveis hierárquicos: Entre grupos, entre populações/dentro dos grupos e dentro das populações. Em ambos os modelos, estatísticas análogas às Estatísticas F são calculadas, sendo nessa abordagem designadas Estatísticas Φ (Excoffier *et al*., 1992).

## 1.4 Identificação de Polimorfismos com Alta Divergência Populacional

A identificação de variantes com grandes diferenças na distribuição de freqüências alélicas entre populações humanas é um dos objetivos primários da pesquisa genética e tem grande interesso médico (Myles *et al.*, 2008). Isso porque doenças que apresentam diferenças de incidência entre populações humanas são tipicamente doenças complexas, cuja etiologia pode estar associada a fatores genéticos e ambientais (Hughes *et al*., 2008). Assim, a identificação de variantes com alta diversidade entre as populações pode estar relacionada às diferenças fenotípicas observadas entre as mesmas (Myles *et al.*, 2008).

### *1.4.1 Implicações em estudos de associação caso-controle*

A identificação de SNPs com alta variabilidade entre grupos continentais é crucial em estudos de epidemiologia genética devido a dois fatores especialmente importantes na população brasileira e de maneira geral em populações miscigenadas: estruturação populacional e desequilíbrio de ligação gerado por miscigenação (Balding, 2006; Smith; O'Brien, 2005).

O objetivo do mapeamento de associação em estudos caso-controle é determinar quais variantes genéticas estão associadas a determinados fenótipos. A idéia consiste em que alelos causais, ou outros próximos a eles, caso exista desequilíbrio de ligação, devem ter freqüências alélicas diferentes nos grupos de casos e controles (Pritchard; Donnelly, 2001). Comumente o desenho de estudos caso-controle prevê que os casos e controles sejam selecionados a partir de amostras populacionais não enviesadas pelo uso de indivíduos relacionados. Além disso, os indivíduos dos dois grupos devem ser pareados de acordo com a idade, sexo, etnicidade, dentre outras possíveis variáveis confundidoras (Thomas; Witte, 2002). Entretanto, a estruturação populacional e a miscigenação podem invalidar resultados obtidos a partir de populações com distribuição heterogênea de freqüências alélicas. (Pritchard; Donnelly, 2001)

A estruturação populacional está relacionada ao risco de associações espúrias em estudos caso-controle devido a diferenças nas freqüências alélicas entre as

subpopulações constituintes do pool gênico de uma população. Três fatores podem levar à sobre-representação de um grupo entre os casos. O primeiro refere-se ao risco inerente de que um alelo pode ser falsamente associado ao fenótipo caso esse último seja mais freqüente numa das populações. Tal fato ocorreria apenas por essa população estar sobre-representada nos casos e o alelo, erroneamente associado, em maior freqüência nessa população em relação às outras constituintes do *pool* gênico. Portanto, para muitos alelos candidatos, a associação se dará apenas pelas diferenças demográficas e não pela causalidade da doença. Marcadores que apresentam grandes diferenças entre grupos populacionais estão mais sujeitos a esse tipo de erro, ou seja, à falsa associação. Outro fator passível de introduzir associações espúrias em estudos de associação é a penetrância diferencial devida a variáveis ambientais. Nesse caso, alguns subgrupos podem ter maior penetrância do genótipo causal devido a pressões ambientais diversas, tais como hábitos alimentares. A terceira variante está relacionada ao viés de averiguação, ou seja, quando há diferenças na amostragem dos indivíduos constituintes do estudo devido a fatores não genéticos tais como, acesso à saúde pública, local de moradia, erros de amostragem (Balding, 2006).

Inadvertidamente poderia concluir-se que apenas a inclusão da etnicidade em estudos casos controles evitaria o problema da estratificação populacional, entretanto, ainda que se leve em conta a etnicidade dos indivíduos amostrados, outros complicadores podem surgir. O cerne dessa questão reside no fato de que a definição de uma população ou etnia é tipicamente arbitrária (Pritchard; Stephens; Donnelly, 2000a). Assim, a estrutura genética populacional pode não refletir as diferenças culturais estabelecidas pelas etnias, ou seja, a estratificação populacional não é bem definida por desígnios como etnia ou nacionalidade, o que corresponde à estruturação populacional críptica. Além disso, indivíduos de populações miscigenadas podem diferir extraordinariamente quanto aos níveis de miscigenação. (Pritchard; Donnelly, 2001)

Haja vista que em determinadas situações é bem interessante a abordagem populacional em estudos de associação caso-controles, opções estão disponíveis para contornar a estratificação. Dois métodos foram publicados entre os anos de 1999 e 2000, cada um apresentando particularidades. Uma abordagem, nomeada Controle

Genômico (GC – *Genomic Control*) consiste na utilização de polimorfismos nulos, que não afetariam a predisposição a doença, para estimar o efeito da estratificação através do genoma. Assim, dois conjuntos de marcadores são analisados, os candidatos e os nulos, e através de um teste qui-quadrado é avaliada a independência entre eles. A magnitude e variância do teste para marcadores nulos estarão infladas caso exista estratificação ou relação críptica entre os indivíduos e conseqüentemente se derivará um multiplicador que permitirá corrigir o valor de significância para os testes de genes candidatos (Devlin; Roeder; Bacanu, 2001). A outra abordagem consiste em um método de associação estrutural (SA – *Structured Association*) descrito em 2000, no qual os indivíduos são designados probabilisticamente às populações estudadas. Em ambas as abordagens, o conhecimento acerca das freqüências de possíveis marcadores informativos de ancestralidade é importante, seja para evitar grandes distâncias genéticas dentro do genoma (GC), seja para designar mais facilmente os indivíduos às subpopulações (SA). Entretanto, se para ambas as abordagens pode-se utilizar populações estruturadas, para populações miscigenadas, apenas a SA pode ser utilizada (Devlin; Roeder; Bacanu, 2001). Atualmente, as análises baseadas em Componentes Principais (ACP) são mais utilizadas para verificar níveis de estruturação genética, isso porque GC tende a ser muito conservativa em várias configurações e SA tende a ser computacionalmente pesada e a definição das subpopulações ainda é controversa (Balding, 2006). Atualmente, as análises de ACP são bastante utilizadas em estudos de associação, principalmente para excluir indivíduos *outliers* das populações (Roeder; Luca, 2009). Novas metodologias têm sido propostas, como utilizar Componentes Principais como covariantes nas regressões (Price *et al*., 2006)

Apesar da ampla discussão sobre a validade de estudos de associação em populações estratificadas e/ou miscigenadas, pouco se sabe sobre a extensão e magnitude desses tipos de obstáculos, uma vez que não há muitos estudos demonstrando claramente resultados falsos positivos (Thomas; Witte, 2002). Um exemplo de caso várias vezes citado (Pritchard; Rosenberg, 1999; Devlin; Roeder; Bacanu, 2001; Pritchard; Donnelly, 2001; Thomas; Witte, 2002) diz respeito ao estudo conduzido por Knowler em 1988 que encontrou falsa associação entre um marcador de ancestralidade européia, $Gm^{3;5,3,14}$, e a presença de diabetes mellitus tipo 2 em populações indígenas, Pima e Papago, do Arizona (Knowler *et al.*, 1988).

Recentemente, Choudhry e colaboradores (2006) demonstraram em um estudo sobre asma que para populações miscigenadas mexicanas e porto-riquenhas, a não correção para ancestralidade individual pode levar a falsos positivos. Entretanto não há muitos mais estudos documentados. Tal fato possivelmente reflete a impossibilidade de se testar os efeitos da estruturação ou miscigenação em estudos mais antigos, de se replicar determinados estudos ou do viés de se publicar artigos com resultados significativos (Pritchard; Donelly, 2001; *cf*. Terwilliger; Weiss, 1998; Cardon; Bell, 2001)

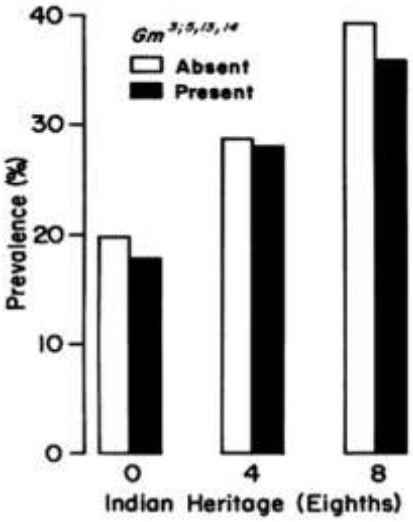



**Figura 2:**

Relação entre a prevalência de diabetes tipo 2 e a presença/ausência do marcador $Gm^{3;5,13,14}$. A relação foi calculada para diferentes níveis de ancestralidade indígena, sendo a miscigenação calculada com base no número de avós com ancestralidade indígena. Estudo conduzido a partir de indivíduos residentes na comunidade indígena de Gila River (EUA).

Em: Knowler *et al.*, 1988

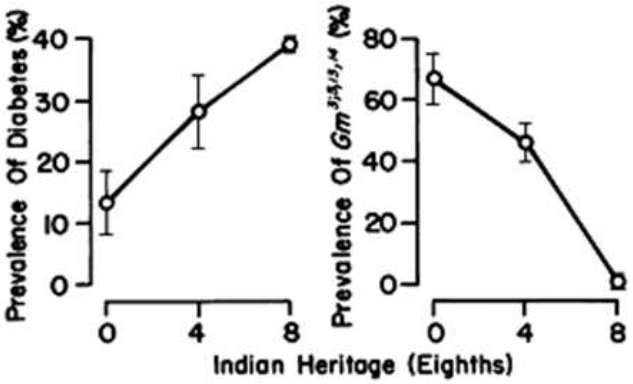



**Figura 3:**

Prevalência de diabetes tipo 2 (esquerda) e do marcador $Gm^{3;5,13,14}$ (direita) entre indivíduos da comunidade de Gila River (EUA), de acordo com a ancestralidade indígena.

Em: Knowler *et al.*, 1988

O segundo fator relevante para estudos de associação em populações como a brasileira, entretanto, é um artifício positivo para os estudos de mapeamento em populações miscigenadas. Nessas populações de miscigenação recente, o desequilíbrio de ligação gerado por miscigenação (DLM) pode ser utilizado para o mapeamento genético de doenças complexas (Smith; O'Brien, 2005). Também nessa

abordagem, a diferença de incidência da doença entre as populações parentais é de grande importância (Balding, 2006). Isso porque as regiões de ancestralidade genômica correspondentes à população em que a doença é mais freqüente serão escaneadas com o intuito de localizar marcadores causais para doença (Seldin, 2007). Essa estratégia é realizada através da localização dessas regiões a partir de Marcadores Informativos de Ancestralidade (MIAs) e assim que os fragmentos cromossômicos são localizados, eles são mapeadas em busca de mutações relacionadas ao fenótipo de interesse (Smith; O'Brien, 2005). Dessa forma, um excesso de casos compartilhando um alelo que é mais comum na população em que a prevalência da doença é maior, pode ser um sinal de que aquele alelo contribui para o risco de desenvolvimento da doença (Balding, 2006).

### *1.4.2 Implicações Evolutivas: Deriva Genética e Seleção Natural*

Diferentes modelos evolucionários têm procurado esclarecer os fatores que levam à diferenciação genética entre as populações. Por exemplo, a Teoria Neutra, que prediz que os polimorfismos evoluem estocasticamente, ou seja, ao acaso. Entretanto, devido a limitações dessa teoria em explicar taxas médias de evolução diferenciais entre taxa e tipos de mutação (por exemplo, sinônimas e não sinônimas), a Teoria Neutra foi sendo substituída pela Teoria Quase-Neutralidade. Sob essa teoria, grande parte da variabilidade entre populações ocorre devido à deriva genética. Entretanto, a adaptação ocorreria devido a pressões seletivas fracas em variantes comuns, ao invés de, se dar através de fortes pressões seletivas em variantes raras. A Teoria Quase-Neutralidade admite três classes de mutações quanto à pressão seletiva: neutras, quase-neutras e não-neutras (deletérias ou vantajosas). E, além disso, assume que as taxas de evolução estão relacionadas ao tamanho efetivo das populações. Ainda hoje, há intenso debate entre defensores da teoria neutra e da seleção adaptativa (Hurst, 2009).

Estudos de genética de populações têm sido utilizados tanto para explicar os padrões de diversidade genética humana em termos de história populacional, quanto para entender as bases genéticas das adaptações fenotípicas. Por trás dessas diferenças há eventos evolutivos moldando a variabilidade genética, seja a deriva, seja a seleção natural. Entretanto, um dos principais obstáculos referentes às

inferências evolutivas repousa justamente na identificação de quais variantes evoluem por deriva genética e quais, devido às pressões seletivas (Balaresque; Ballereau; Jobling, 2007).

O modelo proposto de expansão humana a partir da África (RAO) propõe que as populações humanas sofreram múltiplos eventos fundadores durante a colonização de novas áreas. Dessa forma, a heterozigosidade tende a diminuir de acordo com a distância a partir da África (Novembre; Di Rienzo, 2009). Contudo, *botlenecks* seguidos por expansão espacial podem levar ao aumento da freqüência de um alelo enquanto novas populações colonizam áreas próximas, fenômeno conhecido por *allele surfing* (Hofer *et al*., 2009). Esse fenômeno ocorre devido à ação da deriva genética que advém durante a expansão populacional (Novembre; Di Rienzo, 2009).

A seleção natural pode ser subdividida em três classes: positiva, purificadora e balanceadora. A seleção positiva, ou seleção Darwiniana, está relacionada ao incremento da freqüência de um alelo que aumente o *fitness* do indivíduo (Hurst, 2009). Dessa forma, fenótipos que apresentam grande diferenciação entre populações, possivelmente, estão relacionados a polimorfismos apresentando grandes diferenças nas freqüências alélicas (Myles *et al*., 2008). Regiões sob seleção positiva têm alto desequilíbrio de ligação, isso devido à elevação das freqüências no alelo selecionado ser mais rápida do que a recombinação no local onde ele está situado (Sabeti *et al*., 2002). A seleção purificadora, ou negativa, elimina mutações deletérias. De acordo com a premissa que indivíduos bem adaptados têm maior valor adaptativo, provavelmente esse tipo de seleção é a mais comum. A seleção balanceadora atua no sentido de favorecer a diversidade através da codominância, seleção dependente da freqüência ou coevolução parasita-hospedeiro cíclica. Alelos não são fixados e não podem ser ditos como deletérios ou vantajosos nesse modo de seleção (Hurst, 2009).

A heterozigosidade é capaz de predizer o quão variável são os loci e os afastamentos das expectativas sob neutralidade são capazes de indicar quais loci podem estar sob ação da seleção (Hurst, 2009). A análise dos polimorfismos com maior e menor variabilidade em relação aos demais do genoma pode ser analisada por testes como Mann-Whitney`s U *test*. Polimorfismos com altos valores de

heterozigosidade são indicativos de seleção balanceadora ou de efeito carona (*selective sweep*) (Nielsen *et al*., 2009). Genes do sistema imune, em especial da imunidade inata, são propensos a estarem sob seleção balanceadora devido à pressão seletiva dos patógenos (Ferrer-Admetlla *et al*., 2008), dessa forma, alguns marcadores analisados no presente estudo poderiam estar sob seleção balanceadora.

Os efeitos da seleção positiva e da seleção purificadora na diversidade são parecidos, pois ambas tendem a fixação dos alelos selecionados positivamente ou negativamente (Hurst *et al*., 2009). As variantes positivamente selecionadas podem estar por trás de parte das variações fenotípicas observadas entre humanos e as populações que estão sob seleção positiva, tendem a ter os seus valores de heterozigosidade reduzidos (Myles *et al*., 2008 *apud* Kim; Stephan, 2002). Desse modo, polimorfismos com baixa heterozigosidade tendem a estar sob seleção negativa ou efeito carona recente (Nielsen *et al*., 2009). Entretanto, eventos demográficos podem mimetizar as características encontradas na seleção positiva: efeito carona de alta extensão e diversidade alélica reduzida. Isso porque, efeitos fundadores seguidos por expansões espaciais podem levar à dispersão geográfica de um alelo, ou *Allele Surfing* (Hofer *et al*., 2009).

Alterações não-sinônimas levam à alteração de aminioácidos na seqüência protéica e podem incorrer em alterações conformacionais nas proteínas. Devido ao grande impacto que essas mutações podem causar, geralmente, são selecionadas negativamente. Boyko e colegas (2008) estimam que 27-29% das mutações não-sinônimas são neutrais, 30-42% são moderadamente deletérias e 29-43% são altamente deletérias ou letais. Já as mutações nas regiões promotoras podem influenciar a regulação da tradução das proteínas, positivamente ou negativamente. Taylor e colaboradores sugerem que as regiões promotoras detêm taxas maiores de evolução se comparadas às regiões intrônicas (Taylor *et al*., 2008). Também há vários indícios de seleção positiva nas regiões promotoras de genes relacionados às funções neurais e à nutrição (Haygood *et al*., 2007).

Estudos de variabilidade genética baseados em genes podem ser estendidos a novas perspectivas, tais como a farmacogenômica, que objetiva elucidar os fatores hereditários que influenciam na resposta individual a fármacos e outros estímulos

exógenos (Schork, 2001). O desafio é encontrar essas variantes gênicas para entender como elas interagem entre si e com o ambiente e como ajustar o tratamento de acordo com o indivíduo (Goldstein, 2003). Uma importante ferramenta no desenho de estudos farmacogenômicos baseados em distância genética interpopulacionais é a base de dados PharmGKB: *Pharmacogenetics Knowledge Base* (www.pharmgkb.org). A base de dados PharmGKB contêm informações que incluem genes, proteínas, seqüências referências, regiões de interesse, haplótipos, populações dos indivíduos. Além disso, há informações acerca dos fenótipos celulares, farmacocinética, cinética enzimática, descrição de fármacos, informações sobre estudos clínicos, administração e metabolismo de fármacos (Hewett, 2002).

Outra base de dados relevante na associação entre variantes genéticas e fenótipos é o amiGO (*Gene Ontology Consortium*). O ponto precioso da base Gene Ontology é produzir um vocabulário estruturado, preciso, comum e controlado para descrever diversos genes e os produtos gênicos para diversos organismos. GO permite a integração de diversos banco de dados com vários tipos de informação biológica relevante, incluindo SwiisPROT, Gen-Bank, EMBL, PIR, DDBJ, MPS, YPD & WormPD, Pfam, SCOP e Enzyme (Ashburner *et al*., 2000).

# 2 OBJETIVOS GERAIS E ESPECÍFICOS

## 2.1 Objetivo geral

Identificar SNPs cuja estrutura genética seja muito diferente entre grupos populacionais; em particular entre Nativos Americanos em relação a outros grupos étnicos.

## 2.2 Objetivos específicos

1. Estudar a estrutura genética de 1442 SNPs distribuídos por 411 genes importantes em imunidade, farmacogenética e carcinogênese; a partir de 13 grupos populacionais formados por 56 populações distribuídas pelos 5 continentes.

1.1 Descrever a estrutura genética dos loci estudados através de índices clássicos de variabilidade genética: Equilíbrio de Hardy-Weinberg, Heterozigosidade Esperada e $F_{IS}$ (Coeficiente de Endocruzamento).

1.2 Descrever a estrutura genética dos loci estudados através de níveis hierárquicos de variância molecular, ou seja, através dos coeficientes $F_{CT}$ (variância entre os grupos populacionais), $F_{ST}$ (variância entre as populações a respeito de todo o conjunto populacional) e $F_{SC}$ (variância entre populações dentro de um mesmo grupo).

# 3 METODOLOGIA

## 3.1 Amostragem

Em colaboração com o Dr. Stephen Chanock (NIH – NCI), possuímos dados de genotipagem de 1442 SNPs em 411 genes (Anexo A) (1421 SNPs do *SNPCancerPanel* da Plataforma de Genotipagem *Illumina Golden Gate*® e 21 SNPs adicionais) de 52 populações do CEPH – HGDP (*Centre d'Etude du Polymorphisme Humain - Human Genome Diversity Cell Line Panel*) (Cann, 1998; 2002; Cavalli-Sforza, 2005), 4 populações nativo-americanas do Peru e Equador (Quechua, San Martin, Cayapa e Matsiguenga – dados não publicados) e 4 populações do painel de 102 indivíduos do SNP500Cancer (Afro-americanos, Euro-descendentes residentes em Utah, Hispânicos e Asiáticos) (Packer *et al*., 2004; 2006), perfazendo um total de 1198 indivíduos (Anexo B e E).

## 3.2 Controle de qualidade

A concordância entre as genotipagens de vários SNPs para os indivíduos do painel de SNP500Cancer foi analisada com intuito de encontrar eventuais erros de genotipagem. Foram comparadas as genotipagens realizadas com a plataforma Illumina GoldenGate®, a mesma utilizada nos sujeitos desse estudo, e as tipagens realizadas pelos métodos de seqüenciamento de Sanger e *Real Time PCR*, Taqman (Tabela 5).

A análise de concordância foi realizada através do módulo de controle de qualidade (*QC summary*) do *software* GLU (*Genotyping and Library Utilities*) versão 1.06 para Windows 32 bits (Disponível em: http://code.google.com/p/glu-genetics).

Os SNPs com fortes desvios no Equilíbrio de Hardy-Weinberg foram preditos como sendo erros de genotipagem e por isso, deveriam ser excluídos. Entretanto, para nenhum SNP genotipado a partir da plataforma Illumina® houve desvio significativo do E-HW. Por isso, nenhum SNP foi excluído a partir desses critérios.

Dos 1442 SNPs para os quais dispomos de dados, 1258 SNPs, compreendidos em 396 genes, foram utilizados em todas as análises. Os 184 polimorfismos restantes foram excluídos de algumas análises devido a divergências de *genotyping calling* entre populações de HGDP-CEPH e os dois conjuntos de dados restantes (SNP500Cancer e Nativo-Americanos do LDGH).

Utilizou-se o conjunto de 1442 SNPs para todas as análises baseadas em subpopulações. Entre as análises para as quais utilizamos o conjunto de 1258 SNPs compreendem-se todas aquelas baseadas em grupos populacionais. Adicionalmente, para algumas análises descritas a seguir, também descartamos os polimorfismos presentes no cromossomo X.

**Tabela 5: SNPs com discordância entre seqüenciamento e genotipagem**

| Sequenc. | Illumina | Conc. | Het-Hom | Hom-Het | Hom-Hom | Taxa | Ref E-HW | Comp E-HW |
|---|---|---|---|---|---|---|---|---|
| PARP4-01 | PARP4-01 | 39 | 0 | 0 | 60 | 0,3939 | 0,0443 | 0,0443 |
| IL15RA-02 | IL15RA-02 | 66 | 10 | 11 | 12 | 0,6667 | 0,0693 | 0,1069 |
| TNKS-05 | TNKS-05 | 73 | 0 | 29 | 0 | 0,7157 | $< 10^{-6}$ | 0,4840 |
| PARP4-19 | PARP4-19 | 74 | 0 | 27 | 0 | 0,7327 | $< 10^{-6}$ | 0,1651 |
| NBS1-13 | NBS1-13 | 81 | 1 | 18 | 0 | 0,8100 | $10^{-6}$ | 0,3020 |
| TNKS-33 | TNKS-33 | 83 | 0 | 18 | 0 | 0,8218 | $10^{-5}$ | 0,3676 |
| CYP19A1-09 | CYP19A1-09 | 82 | 7 | 7 | 2 | 0,8367 | 0,2096 | 0,2120 |
| MBL2-03 | MBL2-03 | 87 | 0 | 14 | 0 | 0,8614 | 1 | 0,5997 |
| BIC-34 | BIC-34 | 88 | 0 | 12 | 0 | 0,8800 | $10^{-6}$ | 0,0286 |
| IGF1-22 | IGF1-22 | 91 | 2 | 7 | 1 | 0,9010 | 0,0407 | 0,6010 |
| BAK1-05 | BAK1-05 | 94 | 7 | 0 | 0 | 0,9307 | 0,0003 | 0,8059 |

Sequenc.: SNPs inferidos pelos métodos de seqüenciamento e *Real-Time PCR*. Illumina: SNPs genotipados pelo método de genotipagem Illumina GoldenGate®. Conc.: Concordância entre os genótipos pelas técnicas de seqüenciamento e genotipagem. Het-Hom: Indica quantos genótipos heterozigotos em TaqMan foram identificados como homozigotos em Illumina. Hom-Het: Indica quantos genótipos homozigotos em TaqMan foram identificados como heterozigotos em Illumina. Hom-Hom: Indica quantos genótipos homozigotos não são concordantes entre as duas técnicas. Taxa: Representa o percentual de concordância entre as duas técnicas. Ref E-HW: Valor do Equilíbrio de Hardy-Weinberg para TaqMan. Comp E-HW: Valor do Equilíbrio de Hardy-Weinberg para Illumina. Observação: Apenas os polimorfismos com concordância inferior a 95% estão representados na tabela.

### 3.3 Definição das populações e dos grupos populacionais

Os indivíduos utilizados nesse estudo foram alocados em 56 subpopulações de acordo com a localização geográfica dos mesmos (ver Anexo B, E). Devido ao baixo número de indivíduos presentes nas populações Bantu da África Meridional, agrupamos todos os indivíduos da região sudoeste e sul do continente africano em uma única subpopulação, como realizado em (Rosenberg *et al*., 2002, 2005). Utilizamos as diretrizes do artigo de Bastos-Rodrigues (2006) para agrupar todos os indivíduos da etnia Han (China) em única subpopulação.

As 56 subpopulações foram agrupadas em 9 grupos populacionais, além de mais quatro populações acrescidas posteriormente, miscigenadas e com quatro ascendências diferentes, residentes nos Estados Unidos da América, e pertencentes ao painel de indivíduos do SNP500Cancer (Packer *et al*., 2004, 2006), perfazendo 13 grupos populacionais. Os grupos populacionais foram formados a partir da proximidade geográfica das populações, mas é importante salientar que foram respeitadas as similaridades genéticas entre as populações, de modo a representar com maior fidedignidade possível a real estrutura populacional entre os grupos utilizados nesse estudo. Sendo assim, os grupos populacionais foram estabelecidos não só devido à localização geográfica das populações, mas também de acordo com os resultados anteriores de estudos multilocus de estrutura populacional (Rosenberg *et al*., 2002; 2005). Além das divisões estabelecidas por Rosenberg (2004; 2005) também ampliamos o número de grupos com a subdivisão das populações africanas em dois grupos: Leste e Oeste, (Tishkoff *et al*., 2009); e também a subdivisão das populações nativo-americanas em centro-americanas e sul-americanas (Wang *et al*., 2008). É importante ressaltar que essas subdivisões realizadas em nosso estudo são apenas aproximações dos padrões de estrutura genética observados em populações africanas por Tishkoff e outros (2009) e em populações nativo-americanas por Wang e outros (2008).

### 3.4 Descrição do Banco de dados

Através do projeto de doutorado do aluno Wagner Magalhães, o Laboratório de Diversidade Genética Humana vem desenvolvendo uma plataforma bioinformática, a plataforma DIVERGENOME que permite o armazenamento e a manipulação de dados de diferentes projetos de genética de populações e epidemiologia genética. DIVERGENOME é composto por duas partes: DIVERGENOMEdb, um banco de dados relacional capaz de armazenar dados genotípicos e fenotípicos de diferentes projetos individualmente; e DIVERGENOMEtools, um grupo de ferramentas bioinformáticas que permite a manipulação dos dados armazenados na base de dados do DIVERGENOME, permitindo a criação de arquivos de entrada para diversos programas de genética de populações comumente usados por diversos grupos de pesquisa.

Durante a execução do presente projeto de mestrado, alguns componentes da plataforma DIVERGENOME estavam em fase de desenvolvimento ou em testes. Devido a isso desenvolvemos um banco de dados relacional mais simples, destinado a armazenagem dos dados genotípicos utilizados nesse estudo.

Assim, LDGH_SNPsdb é um banco de dados relacional gerenciado através do SGBD (Sistema de Gerenciamento de Banco de Dados) MySQL. O MySQL utiliza a linguagem SQL (Linguagem de Consulta Estruturada – *Structure Query Language*) como interface para administração dos dados, sendo um dos SGBDs mais utilizados no mundo devido a fatores tais como, código aberto, velocidade de processamento e disponibilidade para vários sistemas operacionais. LDGH_SNPsdb também permite o acesso aos dados através do phpMyAdmin, uma ferramenta em linguagem php, que facilita e permite a interação do usuário com o banco através da *Worl Wide Web* (WWW). O banco é acessível através do servidor http Apache, também um software livre, e está instalado em um sistema operacional Linux Ubuntu.

O *script* de criação do LDGH_SNPsdb foi gerado através do *software* DBDesigner 4.5.6, ferramenta que permite a modelagem visual de um banco dados gerenciado por MySQL.

LDGH_SNPsdb permite o armazenamento de dados provenientes da genotipagem de SNPs, não comportando outros tipos de variações genômicas e nem o armazenamento de características fenotípicas. É constituído por quatro entidades (ou tabelas) assim designados: *SNPs*, *Genotypes*, *Sample*, *Population* (Anexo C).

## 3.5 Obtenção dos arquivos de entrada

### *3.5.1 Criação dos arquivos de entrada para Genepop*

A criação dos arquivos de entrada para Genepop (Rousset, 2008) iniciou-se a partir do resultado da pesquisa em MySQL em formato TRIP – saída padrão de banco de dados – onde os dados estão dispostos linearmente, sendo que para este caso específico cada linha apresenta os seguintes dados dispostos em três colunas: código do indivíduo, código do polimorfismo e genótipo. Esse resultado foi posteriormente transformado em uma matriz SDAT – código do indivíduo x código do polimorfismo – através do algoritmo Polyout, constituinte do DIVERGENOMEtools, desenvolvido pelo aluno de doutorado Wagner Magalhães. Após a obtenção das matrizes, o arquivo de entrada para Genepop foi construído com o auxílio dos programas Excel 2007 e TextPad 5.2.0, sendo que os alelos A, C, G e T foram substituídos por 01, 02, 03 e 04 de acordo com a exigência dos programas Convert (Glaubitiz, 2004) e Genepop (Rousset, 2008), já que esses também trabalham com microssatélites.

### *3.5.2 Criação dos arquivos de entrada para GDA e Arlequin*

Para a obtenção dos arquivos de entrada, requeridos pelos programas de genética de populações, utilizou-se a ferramenta CONVERT versão 1.31 (Glaubitz, 2004). Para tanto, foram criados arquivos no formato Genepop que posteriormente foram transformados em *inputs* para os seguintes programas de análises de dados genéticos: Arlequin (Schneider; Roessli; Excoffier, 2000) e GDA (Lewis; Zaykin, 2001).

## 3.6 Análises Estatísticas

### *3.6.1 Cálculos de freqüências alélicas e genotípicas*

As freqüências alélicas dos 1442 SNPs para todas as populações e grupos populacionais foram obtidas através da funcionalidade "*Produce Table of Allelic Frequencies*" do *software* CONVERT 1.31. Para as freqüências genotípicas, utilizou-se uma *query* em MySQL (Anexo D) que permite a contagem dos genótipos a partir do próprio banco de dados LDGH_SNPsdb.

### *3.6.2 Equilíbrio de Hardy-Weinberg*

Como grande parte das populações desse trabalho apresenta tamanho amostral inferior a 40 indivíduos, utilizou-se um teste análago ao Exato de Fisher para evidenciar afastamentos do Equilíbrio de Hardy-Weinberg, o teste de Guo e Thompson (Guo; Thompson, 1992). Adicionalmente ao teste de Guo e Thompson, utilizou-se o Teste U para avaliar a significância dos testes para déficit de heterozigotos e de homozigotos (Rousset; Raymond, 1995).

Para a análise de significância do E-HW utilizamos a correção de Bonferroni para testes múltiplos, estabelecendo o valor de corte de $10^{-5}$ ($\alpha = 0,05/1442$). O p valor estipulado nesse trabalho é restritivo, uma vez que os 1442 testes não são completamente independentes, há muitos SNPs em desequilíbrio de ligação (DL), principalmente dentro dos genes. Entretanto, também não é adequado assumir que todos os SNPs dentro de um gene estejam em DL, em especial, para genes de longa extensão. Para os cálculos do Equilíbrio de Hardy-Weinberg, todos os loci situados no cromossomo X foram excluídos por apresentarem desvios nas freqüências genotípicas em relação ao que é esperado sob E-HW. Tal fato ocorre devido à dinâmica de segregação dos cromossomos sexuais que difere da dinâmica de segregação dos cromossomos autossômicos.

O Equilíbrio de Hardy-Weinberg foi calculado através do *software* Arlequin, versão 2.000. O teste é realizado a partir de uma versão otimizada do algoritmo de Guo e Thompson. A versão implementada em Arlequin retorna os mesmos resultados,

sendo, porém, mais rápida. Os valores de significância para os testes de déficits de homozigotos ou heterozigotos foram obtidos a partir do Teste U implementado no software Genepop, versão 4.09 (Rousset, 2008).

### *3.6.3 Heterozigosidade Esperada ($H_E$)*

Para estimar a variabilidade dos loci calculamos a Heterozigosidade Esperada a partir do programa GDA versão 1.1 para Windows 32 bits (Lewis; Zaykin, 2001). A heterozigosidade esperada média calculada por GDA consiste em um estimador não enviesado formado pela heterozigosidade esperada ($1-\sum_u p_u^2$) multiplicada pelo fator de correção para tamanho amostral (2n)/(2n-1). Os cálculos foram realizados para estabelecer as heterozigosidades para os loci em cada grupo populacional.

### *3.6.4 Análise de Variância Molecular (AMOVA)*

A Análise de Variância Molecular (AMOVA) foi calculada a partir do *software* Arlequin versão 2.000 (Schneider; Roessli; Excoffier, 2000), para dados genotípicos. As análises envolveram a utilização de dois modelos hierárquicos: de três níveis hierárquicos (entre grupos, interpopulacional e intrapopulacional) e de dois níveis hierárquicos (intra e interpopulacional), em diferentes situações, que serão ilustradas nas seções subseqüentes. Os índices de diversidade utilizados se restrigem àqueles relacionados ao $F_{ST}$. A significância dos valores de AMOVA foi obtida através de 10000 permutações.

Os componentes de AMOVA são assim calculados:

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2}, \quad F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} \quad \text{and} \quad F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}.$$

Onde, $\sigma^2_a$ é o componente da covariância devido a diferenças entre haplótipos entre populações, $\sigma^2_b$ é o componente da covariância devido a diferenças entre haplótipos em diferentes populações dentro do grupo, $\sigma^2_c$ é o componente da covariância devido a diferenças entre haplótipos dentro de uma população e $\sigma^2_T$ é a variância molecular total (Excoffier; Smouse; Quattro, 1992).

### 3.6.4.1 Estruturação de grupos populacionais

Para as Análises de Variabilidade Molecular, analisaram-se todas as 56 subpopulações agrupadas em 9 grupos continentais, excluindo-se, portanto, as populações do SNP500Cancer. Utilizaram-se três configurações gerais de arranjo populacional para descrição da divergência populacional.

O modelo tri-hierárquico foi utilizado para descrever a variância em três componentes: entre grupos populacionais ($F_{CT}$), entre subpopulações/dentro dos grupos ($F_{SC}$) e entre todas as subpopulações ($F_{ST}$).

O modelo de dois componentes foi utilizado em duas situações: na primeira, utilizamos as 56 populações em um mesmo grupo para obter os valores extremos de $F_{ST}$; na segunda, realizaram-se AMOVAs independentes para acessar o valor de $F_{SC}$ para cada grupo populacional e conseqüentemente os níveis de variância inter-populacionais (sub-populações) e intra-populacionais.

### 3.6.4.2 Identificação de SNPs com grande divergência entre populações

Para as análises regionais utilizamos 10 configurações populacionais envolvendo as seguintes entidades: grupos populacionais, regiões geográficas e subpopulações. As configurações populacionais visaram elucidar aspectos histórico-demográficos referentes às populações nativo-americanas e latino-americanas, da qual a subpopulação brasileira faz parte.

Nessas análises buscamos selecionar polimorfismos com grande diferença nas freqüências alélicas entre os conjuntos populacionais. Grande parte dos loci, dentre os triados, provavelmente deve a variabilidade ao "*allele surfing effect*" (Hofer *et al*. 2009), porém todos os marcadores são potenciais alvos da seleção natural, além da importância já explicitada em estudos caso-controle, onde a estruturação populacional pode levar a resultados falso-positivos.

Dessa forma, as análises regionais têm dois objetivos principais: identificar polimorfismos que devido à estruturação populacional possam levar a resultados

falso-positivos e elucidar aspectos histórico-demográficos e de seleção natural envolvendo as populações do Leste Asiático e de Nativo-Americanos. O escopo das quatro primeiras configurações corresponde ao primeiro objetivo, enquanto as seis restantes estão relacionadas aos aspectos de formação do *pool* gênico das Américas e a descrição de similaridade genética entre populações Ameríndias e do Leste Asiático.

Ressalta-se que apesar da divisão artificial entre os dois tipos de configuração, os loci selecionados em ambas configurações podem ter tanto implicações biomédicas, ou seja, associações espúrias em estudos caso-controle, quanto estarem sob seleção natural ou terem sofrido o efeito de *allele surfing.*

A partir das configurações populacionais descritas a seguir selecionamos SNPs com valores de $F_{CT}$ (diversidade entre grupos) maiores que 0,25 e de $F_{SC}$ (diversidade dentro dos grupos) menores que 0,10. Esses valores foram escolhidos de acordo com a média encontrada no genoma humano ($F_{ST}$ de aproximadamente 0,12) e visaram obter uma boa representação daqueles polimorfismos com alto grau de diferenciação entre as populações – grupos populacionais, regiões geográficas ou subpopulações – e com freqüências homogêneas dentro das mesmas, para que possam ser representativos de todo a população. Devido às dificuldades inerentes à utilização de dados provenientes de cromossomos sexuais, todos os SNPs do cromossomo X foram excluídos dos resultados.

3.6.4.2.1 Implicações Biomédicas em estudos caso-controle

Na configuração (EUR-NAT-WAFR) selecionamos loci diferenciados entre as populações parentais da população latino-americana e por extensão, brasileira. Foram formados três conjuntos: África Ocidental, Europa e Nativo-Americanos. O último compreende populações indígenas da América do Sul e Central (Quadro 1).

Adicionalmente, três configurações populacionais com enfoque na estrutura genética das populações parentais foram analisadas. Essas três análises populacionais basearam-se na tomada par-a-par das populações da África Ocidental, Europa e Nativo-Americanos. Assim, selecionamos SNPs com alta variância entre:

(NAT – WAFR); (EUR – NAT) e (EUR – WAFR). Devido aos diferentes níveis e composições de miscigenação na América Latina (Wang *et al.*, 2008), as três últimas configurações são importantes em populações predominantemente di-híbridas.

| **Quadro 1: Grupos utilizados nas configurações referentes às implicações biomédicas** | |
|---|---|
| **Grupos Regionais** | População |
| **África Ocidental (WAFR)** | Mandenka |
| | San |
| | Iorubá |
| **Nativo-Americanos (NAT)** | Maia |
| | Pima |
| | Cayapa |
| | Karitiana |
| | Matsiguenga |
| | Piapoco e Curripaco |
| | Quechua |
| | San Martin |
| | Suruí |
| **Europa (EUR)** | Adygei |
| | Bergamo |
| | Franceses |
| | Bascos Franceses |
| | Orcadianos |
| | Russos |
| | Sardenha |
| | Toscanos |

Os polimorfismos encontrados nas configurações anteriores também poderão ser utilizados com marcadores de ancestralidade (MIAs) em populações latinas miscigenadas de diferentes composições parentais, desde que, obviamente, duas das três populações estudadas estejam envolvidas no processo de formação da população miscigenada de interesse.

3.6.4.2.2 Implicações Evolutivas

As configurações populacionais a seguir têm como objetivo identificar loci cuja alta diferenciação está relacionada a fatores histórico-demográficos envolvidos na formação do *pool* gênico das populações nativo-americanas e seleção natural. Para estas análises utilizou-se o grupo populacional Leste Asiático e as subpopulações mais próximas ao Estreito de Bering.

A configuração regional (EAS – NAT) envolveu a análise de divergência entre os grupos populacionais Leste Asiático e Nativo-americanos (incluí populações indígenas meso e sul-americanas).

Formou-se também um grupo regional, NEAS, constituído por três subpopulações do Nordeste Asiático: Daur, Oroqen e Hezhen. Esse grupo foi confrontado ao dos nativo-americanos, sendo então a sexta configuração analisada (NAT – NEAS).

As configurações restantes envolvem a análise par-a-par entre populações nativo-americanas e subpopulações do Leste Asiático, mais próximas do Estreito de Bering. As configurações restantes são: Nativos e Yakut (NAT – YAK); Nativos e Daur (DAUR – NAT); Nativos e Hezhen (HEZ – NAT) e Nativos e Oroqen (NAT –ORO).

# 4. RESULTADOS

## 4.1 Equilíbrio de Hardy-Weinberg

Oito polimorfismos apresentaram afastamentos significativos para mais de um grupo populacional: CSF3-02 (América do Sul, Centro Sul da Ásia, Leste da Ásia e Oriente Médio); os polimorfismos PIN1-17 e PTGS2-05 apresentam desvios para América do Sul e Centro Sul da Ásia. Já os polimorfismos rs17204605 (gene *GSK3B*) e rs1805087 (gene *MTR*) têm valores significativos de E-HW para Europa e Oriente Médio. HSD3B2-19 apresenta freqüências genotípicas anômalas sob E-HW para América do Sul e Leste da Ásia, enquanto MX1-01 para Leste Asiático e Oceania, e SLC6A18-13 para América Central e Europa (Tabela 6).

**Tabela 6: SNPs com afastamentos significativos de E-HW para Grupos Populacionais**

| GENE | SNP | WAFR | EAFR | CAM | SAM | CSA | EAS | EUR | OCE | ORM |
|---|---|---|---|---|---|---|---|---|---|---|
| *ABCA6* | ABCA6-01 | | | | HOM | | | | | |
| *ABCA6* | ABCA6-05 | | | | HOM | | | | | |
| *ABCG8* | ABCG8-01 | | | | | HOM | | | | |
| *AKR1C3* | AKR1C3-26 | | | | HOM | | | | | |
| *APC* | APC-19 | | | | HOM | | | | | |
| *APEX1* | APEX1-09 | | | | | HOM | | | | |
| *APOB* | APOB-07 | | | | HOM | | | | | |
| *CAV1* | CAV1-02 | | | | HET | | | | | |
| *CCND1* | CCND1-03 | | | | HET | | | | | |
| *CD81* | CD81-06 | | | | | HOM | | | | |
| *CSF3* | CSF3-02 | | | | HET | HET | HET | | | HET |
| *CTH* | CTH-10 | | | | | | | HOM | | |
| *EGF* | EGF-04 | | | | HOM | | | | | |
| *FZD7* | FZD7-16 | | | | | | | HOM | | |
| *FZD7* | FZD7-17 | | | | HOM | | | | | |
| *GPX1* | GPX1-06 | | | | HET | | | | | |
| *GPX3* | GPX3-18 | | | | HOM | | | | | |
| *HAO2* | HAO2-01 | | | | HOM | | | | | |
| *HSD3B2* | HSD3B2-19 | | | | HOM | | HOM | | | |
| *HTR1D* | HTR1D-03 | | | | HOM | | | | | |
| *IGFALS* | IGFALS-91 | | | | | HET | | | | |
| *IL4R* | IL4R-07 | | | | | | | HOM | | |
| *LDLR* | LDLR-08 | | | | | | HET | | | |
| *LIPC* | LIPC-01 | | | | | | | HET | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *LOC646837* | LOC646837-05 | | | | HOM | | | | | |
| *MTRR* | MTRR-07 | | | | HOM | | | | | |
| *MX1* | MX1-01 | | | | | | HOM | | HOM | |
| *PIN1* | PIN1-17 | | | | HET | HOM | | | | |
| *POT1* | POT1-02 | | | | HOM | | | | | |
| *POT1* | POT1-05 | | | | HOM | | | | | |
| *POT1* | POT1-09 | | | | HOM | | | | | |
| *POT1* | POT1-11 | | | | HOM | | | | | |
| *PTGS2* | PTGS2-05 | | | | HET | HOM | | | | |
| *GSKB3* | rs1154597 | | | | | | | HOM | | |
| *GSKB3* | rs17204605 | | | | | | | HOM | | HOM |
| *MTR* | rs1805087 | | | | | | | HOM | | HOM |
| *ARNT* | rs7517566 | | | | HOM | | | | | |
| *SLC23A2* | SLC23A2-25 | | | | | | | HOM | | |
| *SLC39A2* | SLC39A2-05 | | | | HOM | | | | | |
| *SLC4A2* | SLC4A2-04 | | | | | | | HOM | | |
| *SLC6A18* | SLC6A18-13 | | | HOM | | | | HOM | | |
| *TERT* | TERT-15 | | HOM | | | | | | | |
| *TNIP1* | TNIP1-02 | | | | HOM | | | | | |
| *TNKS* | TNKS-23 | | | | HOM | | | | | |
| *TNKS* | TNKS-35 | | | | | HOM | | | | |
| *TNKS* | TNKS-64 | | | | HET | | | | | |
| *VIL2* | VIL2-02 | | | | HOM | | | | | |
| *VIL2* | VIL2-03 | | | | HOM | | | | | |
| **TOTAL** | | 0 | 1 | 1 | 30 | 8 | 4 | 10 | 1 | 3 |

Siglas dos Grupos Populacionais: WAFR: África Ocidental; EAFR: África Oriental; CAM: América Central; SAM: América do Sul; CSA: Centro Sul Asiático; EAS: Leste Asiático; EUR: Europa; OCE: Oceania; ORM: Oriente Médio. HOM: Indica que para aquele SNP há excesso de homozigotos; HET: Indica que para aquele SNP há excesso de heterozigotos.

Nas análises baseadas na distribuição dos indivíduos por subpopulação pode-se observar que 16 SNPs, distribuídos por 15 genes, encontravam-se fora do E-HW. O único gene com dois SNPs fora do equilíbrio de HW foi *GSKB3* representado por: rs1154597 e rs17204605. Três SNPs têm afastamentos para mais de uma população: CYP19A1-16 em Franceses e Sindhi, MX1-01 para Yakut e Papua Nova-Guiné e SLC4A2-04 para Franceses e Russos (Tabela 7).

| Tabela 7: SNPs com afastamentos de E-HW por subpopulação | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GENE | SNP | KLS | FRC | SND | YAK | RUS | QEC | SDN | PIM | PAP | PAL | MAT |
| *CD81* | CD81-06 | | HET | | | | | | | | | |
| *CSF3* | CSF3-02 | HET | | | | | | | | | | |
| *CTH* | CTH-10 | | HOM | | | | | | | | | |
| *CYP19A1* | CYP19A1-16 | | HOM | HOM | | | | | | | | |
| *EGRF* | EGRF-05 | | | | | | | | | | | HOM |
| *ERCC5* | ERCC5-05 | | HOM | | | | | | | | | |
| *FZD7* | FZD7-16 | | HOM | | | | | | | | | |
| *IGFALS* | IGFALS-91 | | HET | | | | | | | | | |
| *MX1* | MX1-01 | | | | HOM | | | | | HOM | | |
| *PIN1* | PIN1-17 | | | | | | HET | | | | | |
| *SLC4A2* | SLC4A2-04 | | HOM | | | HOM | | | | | | |
| *SLC6A18* | SLC6A18-13 | | | | | | | | HOM | | | |
| *GSK3B* | rs1154597 | | | | | HOM | | | | | | |
| *GSK3B* | rs17204605 | | HET | | | | | | | | | |
| *MTR* | rs1805087 | | | | | | | HOM | | | | |
| *MSH2* | rs7602094 | | | | | | | | | | HET | |
| **TOTAL** | | **1** | **8** | **1** | **1** | **2** | **1** | **1** | **1** | **1** | **1** | **1** |

HOM: Indica que para aquele SNP há excesso de homozigotos; HET: Indica que para aquele SNP há excesso de heterozigotos. Sigla das populações: KLS, Kalash; FRC, Franceses; SND, Sindhi; YAK, Yakut; RUS, Russos; QEC, Quechua; SDN, Sardenha; PIM, Pima; PAP, Papua; PAL, Palestinos; MAT, Matsiguenga.

### *4.2 Heterozigosidade Esperada (Loci)*

Apenas os loci polimórficos estão representados nas tabelas abaixo. Sendo assim, os loci com apenas um alelo não foram descriminados entre os menores valores de $H_E$. Ressalta-se também o fato de que alguns SNPs foram ignorados devido aos tamanhos amostrais reduzidos.

| Tabela 8: Valores de Heterozigosidade Esperada para a África Oriental | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| PGR-11* | 0,0158 | NFKB1-09 | 0,5040 |
| PGR-20# | 0,0158 | XRCC4-01* | 0,5040 |
| PGR-28 | 0,0158 | rs334535 | 0,5040 |
| PTH-01* | 0,0158 | rs6770314 | 0,5040 |
| SFTPD-03* | 0,0158 | rs9851174 | 0,5040 |
| TEP1-02* | 0,0158 | SSTR3-03* | 0,5039 |
| TERF2-14 | 0,0158 | BIC-15 | 0,5038 |
| TNKS-46 | 0,0158 | CYP2E1-31 | 0,5038 |
| TSG101-30# | 0,0158 | ERCC5-02* | 0,5038 |
| XBP1-02# | 0,0158 | HSPB8-01 | 0,5038 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Na tabela 8 estão os representados os SNPs com maiores e menores valores de Heterozigosidade para o grupo populacional África Oriental, sendo que 117 marcadores são monomórficos nesse grupo.

| Tabela 9: Valores de Heterozigosidade Esperada para a África Ocidental | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| RERG-33 | 0,0178 | OPRD1-03 | 0,5047 |
| RET-01* | 0,0178 | IL13-03 | 0,5046 |
| SFTPD-03* | 0,0178 | RGS17-01 | 0,5045 |
| SLC23A2-33 | 0,0178 | RGS17-03 | 0,5045 |
| SLC4A2-01 | 0,0178 | CAT-02# | 0,5045 |
| TEP1-02* | 0,0178 | AKR1C3-35 | 0,5044 |
| TGFBR1-03 | 0,0178 | CYP24A1-08 | 0,5044 |
| TNIP1-02# | 0,0178 | IGF1-04 | 0,5044 |
| TP73L-03 | 0,0178 | MTR-05 | 0,5044 |
| rs17388148 | 0,0178 | POLD1-13 | 0,5044 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Na tabela 9 estão representados os loci com maior e menor variabilidade para o grupo populacional do Oeste Africano, sendo que para esse grupo, 136 marcadores são homozigotos para um alelo.

| Tabela 10: Valores de Heterozigosidade Esperada para o Oriente Médio | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| CCNA2-01* | 0,0057 | GPX1-06 | 0,5077 |
| LTA-05 | 0,0057 | GSTP1-01* | 0,5077 |
| PGR-27 | 0,0057 | GPX4-09# | 0,5014 |
| BRCA1-20* | 0,0115 | IL2-01# | 0,5014 |
| MTRR-05* | 0,0115 | CTLA4-17# | 0,5014 |
| MTRR-22* | 0,0115 | FBXW7-01 | 0,5014 |
| INSR-59 | 0,0171 | SCUBE2-02 | 0,5014 |
| SLC6A18-13 | 0,0171 | CCND1-03 | 0,5014 |
| SLC23A1-05* | 0,0228 | CYP24A1-01# | 0,5014 |
| HMGCR-02 | 0,0284 | SLC23A2-03* | 0,5014 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Na tabela 10 estão expostos os dez marcadores com maiores e menores valores de $H_E$ para o grupo Oriente Médio, apenas 6 marcadores não apresentam variabilidade nesse grupo.

| Tabela 11: Valores de Heterozigosidade Esperada para a Europa | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| SLC23A1-05* | 0,0259 | SLC30A1-01 | 0,5027 |
| MTRR-05* | 0,0312 | NFKBIE-02 | 0,5016 |
| MTRR-22* | 0,0312 | MX1-28 | 0,5016 |
| CHEK1-02* | 0,0314 | HSD3B2-25# | 0,5016 |
| CGA-05* | 0,0374 | COASY-01* | 0,5015 |
| BLM-02 | 0,0435 | GATA3-46 | 0,5015 |
| LCAT-03* | 0,0443 | MTHFD2-01 | 0,5015 |
| BRCA1-20* | 0,0495 | SLC19A1-05 | 0,5014 |
| CAT-07# | 0,0495 | CD4-03 | 0,5014 |
| DHFR-07 | 0,0495 | CYP19A1-04 | 0,5014 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Na tabela 11 estão expostos os 10 maiores e 10 menores valores de Heterozigosidade Esperada para o grupo populacional europeu. De acordo com o esperado, devido ao processo de seleção de marcadores, esse grupo populacional foi o que apresentou a menor quantidade de loci monomórficos, apenas 3 não são

variáveis.

| Tabela 12: Valores de Heterozigosidade Esperada para o Centro Sul Asiático | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| AKR1C3-29# | 0,0151 | CYP1B1-27# | 0,5013 |
| ABCB1-12 | 0,0152 | IFNGR2-03 | 0,5013 |
| TNKS-46 | 0,0152 | EFNB3-01 | 0,5013 |
| IGF1R-27 | 0,0200 | ABCA5-01 | 0,5013 |
| SOD2-06# | 0,0200 | BIC-34 | 0,5013 |
| MMP1-01* | 0,0250 | CYP24A1-01# | 0,5013 |
| ATM-02* | 0,0251 | BIC-07 | 0,5013 |
| FANCA-25 | 0,0251 | TERT-14 | 0,5013 |
| PGR-27 | 0,0251 | BIC-32 | 0,5012 |
| IL8-11 | 0,0252 | IL1B-03# | 0,5012 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras.. | | | |

A tabela 12 apresenta os 10 valores máximos e mínimos de Heterozigosidade Esperada para o grupo populacional Centro Sul Asiático. Nesse grupo, 6 marcadores são homozigotos para um mesmo alelo.

| Tabela 13: Valores de Heterozigosidade Esperada para o Leste Asiático | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| CAT-07# | 0,0041 | ENG-06 | 0,5012 |
| CDKN2A-09 | 0,0041 | LRP5-01 | 0,5011 |
| ENPP1-04 | 0,0041 | CDKN2A-16# | 0,5010 |
| ERCC4-01* | 0,0041 | EPHX1-15# | 0,5010 |
| INSR-59 | 0,0041 | FBXW7-44 | 0,5010 |
| NR1H4-18 | 0,0041 | HTR1B-02* | 0,5010 |
| PGR-27 | 0,0041 | ZFPM1-07 | 0,5010 |
| RERG-29 | 0,0041 | NPAT-01 | 0,5010 |
| WDR79-06 | 0,0041 | BCR-02* | 0,5010 |
| rs17810302 | 0,0041 | OCA2-03* | 0,5010 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras.. | | | |

A tabela 13 apresenta o resultado do cálculo de Heterozigosidade Esperada para a população Leste Asiático, apenas os 10 loci mais e menos variáveis foram selecionados. Dentre as populações da Eurásia, esse grupo é o que tem o maior

número de loci monomórficos, 32 no total.

| Tabela 14: Valores de Heterozigosidade Esperada para a Oceania | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| RAB15-04 | 0,0313 | CAT-03* | 0,5084 |
| RET-02* | 0,0313 | CSF3-02 | 0,5082 |
| SLC6A3-10* | 0,0313 | ABCA1-12* | 0,5079 |
| SOAT2-01 | 0,0313 | ABCA1-26* | 0,5079 |
| TERF2-01 | 0,0313 | AKR1C3-21 | 0,5079 |
| TERF2-14 | 0,0313 | ATM-01 | 0,5079 |
| TERT-14 | 0,0313 | ATM-27 | 0,5079 |
| TFF1-01 | 0,0313 | CAV1-09 | 0,5079 |
| TNKS-26 | 0,0313 | HSD17B2-02 | 0,5079 |
| rs7973746 | 0,0313 | IFNGR2-03 | 0,5079 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Os loci com maiores e menores valores de $H_E$ para o grupo populacional Oceania estão representados na tabela 14. Neste grupo, 237 loci são representados por apenas um alelo.

| Tabela 15: Valores de Heterozigosidade Esperada para a América Central | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| TGFBR1-01 | 0,0204 | SOD1-01 | 0,5055 |
| TGFBR1-04 | 0,0204 | rs10842518 | 0,5055 |
| TNKS-22 | 0,0204 | IL1B-12 | 0,5054 |
| TP73L-15 | 0,0204 | SEC14L2-05 | 0,5054 |
| TXNRD2-88 | 0,0204 | CYP19A1-37# | 0,5053 |
| rs1154597 | 0,0204 | CYP2E1-31 | 0,5053 |
| rs1533593 | 0,0204 | HFE-07 | 0,5053 |
| rs17036577 | 0,0204 | IL1RN-05 | 0,5053 |
| rs1719888 | 0,0204 | rs1137196 | 0,5053 |
| rs2345060 | 0,0204 | rs17329025 | 0,5053 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

A tabela 15 apresenta os resultados de Heterozigosidade Esperada para o grupo populacional América Central e aí estão demonstrados os 10 maiores e os 10 menores valores de $H_E$. Nesse grupo, 136 marcadores não apresentam variabilidade.

| Tabela 16: Valores de Heterozigosidade Esperada para a América do Sul | | | |
|---|---|---|---|
| **Menores Valores** | | **Maiores Valores** | |
| PIM1-25 | 0,0081 | SLC23A1-09 | 0,5025 |
| PIN1-01# | 0,0081 | GATA3-46 | 0,5022 |
| RAB15-02 | 0,0081 | IGFBP2-26# | 0,5022 |
| SEPP1-02 | 0,0081 | CTLA4-01* | 0,5021 |
| TP53I3-13# | 0,0081 | RAD23B-05 | 0,5021 |
| WDR79-06 | 0,0081 | SLC23A1-18# | 0,5021 |
| rs16830683 | 0,0081 | MATR3-01 | 0,5021 |
| rs334535 | 0,0081 | APOA4-07 | 0,5021 |
| rs4072520 | 0,0081 | BCL6-11 | 0,5021 |
| rs6770314 | 0,0081 | IGF2R-05* | 0,5020 |
| Os SNPs que apresentaram os 10 menores valores de $H_E$ média por loci estão representados nas colunas à esquerda em ordem crescente, já nas colunas à direita, estão representados os SNPs que apresentaram os 10 maiores valores de diversidade, em ordem decrescente. * SNPs localizados em éxons. # SNPs localizados em regiões promotoras. | | | |

Os resultados de Heterozigosidade Esperada para o grupo populacional América do Sul estão descritos na tabela 16. Nesse grupo, ao contrário do que seria esperado, 83 loci são monomórficos. O padrão usual é de que essa população tenha valores de heterozigosidade esperada inferiores aos grupos Oceania e América Central e, em especial, que o grupo populacional Leste Asiático.

| Tabela 17: Valores de Heterozigosidade nos grupos populacionais para os SNPs representados dentre os maiores valores de $H_E$ em mais de um grupo populacional | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **GENE** | **SNP** | **WAFR** | **EAFR** | **CAM** | **SAM** | **CSA** | **EAS** | **EUR** | **OCE** | **ORM** |
| *CYP24A1* | CYP24A1-01 | 0,4118 | 0,2905 | 0,2619 | 0,3592 | 0,5012 | 0,4859 | 0,4797 | 0,3645 | 0,5014 |
| *CYP2E1* | CYP2E1-31 | 0,4707 | 0,5038 | 0,5052 | 0,4859 | 0,2542 | 0,4559 | 0,1576 | 0,0312 | 0,1995 |
| *GATA3* | GATA3-46 | 0,2960 | 0,2234 | 0,4512 | 0,5022 | 0,4953 | 0,4728 | 0,5015 | 0,0907 | 0,4844 |
| *IFNGR2* | IFNGR2-03 | 0,2207 | 0,1198 | 0,2619 | 0,4716 | 0,5012 | 0,3120 | 0,4810 | 0,5079 | 0,4994 |
| Siglas dos Grupos Populacionais: WAFR: África Ocidental; EAFR: África Oriental; CAM: América Central; SAM: América do Sul; CSA: Centro Sul Asiático; EAS: Leste Asiático; EUR: Europa; OCE: Oceania; ORM: Oriente Médio. | | | | | | | | | | |

Para os três primeiros polimorfismos descritos na tabela 17 parece haver relação entre a localização geográfica e níveis de heterozigosidade. Os maiores valores de heterozigosidade para CYP24A1-01 são encontrados na Eurásia e Oeste da África e os menores valores na América, Oceania e Leste da África. Para CYP2E1-31 os maiores valores de $H_E$ ocorrem na África, Leste Asiático e América, já os menores na Eurásia (à exceção do Leste da Ásia) e Oceania. Os maiores valores de heterozigosidade para GATA3-46 ocorrem na América e Eurásia e os valores mais

baixos na África e Oceania. Em IFGNR2-03 não há um padrão regional claro, os menores valores estão na África, America Central e Leste Asiático, os valores mais altos na Oceania, América do Sul e em parte da Eurásia.

*4.3 Análise de Variância Molecular*

4.3.1 Estruturação dos grupos populacionais

A partição da variabilidade genética se deu através de três componentes hierárquicos: entre grupos populacionais ($F_{CT}$), entre populações dentro dos grupos ($F_{SC}$) e dentro das populações ($F_{ST}$). A porcentagem da variabilidade devido à diferenças dentro das populações é de 86,69%, a correspondente à diferenças entre populações dentro dos grupos é de 3,04% e a entre grupos é de 10,27%. Tais valores correspondem aos seguintes índices de fixação: $F_{ST}$ igual a 0,13, $F_{SC}$ a 0,03 e $F_{CT}$ a 0,10. O valor de $F_{SC}$ descrito acima corresponde, entretanto, a média global dos valores de variabilidade de cada continente. Assim, calculamos os valores de $F_{SC}$ individualmente para cada continente como descrito na tabela 18.

**Tabela 18: Análise de Variabilidade Molecular por Continente**

| VAR | SAM | CAM | WAFR | EAFR | ORM | EUR | CSA | EAS | OCE |
|---|---|---|---|---|---|---|---|---|---|
| **Entre Populações** | 13,07 | 7,78 | 5,27 | 4,95 | 1,84 | 1,17 | 1,95 | 1,99 | 9,49 |
| **Dentro das Populações** | 86,93 | 92,22 | 94,73 | 95,05 | 98,16 | 98,83 | 98,05 | 98,01 | 90,51 |

Os valores descritos nessa tabela correspondem ao cálculo de $F_{ST}$ realizado em cada continente utilizando o modelo de dois componentes de Análise de Variabilidade Molecular (intra e interpopulacional). VAR: Partição da variabilidade. Siglas dos Grupos Populacionais: SAM: América do Sul; CAM: América Central; WAFR: África Ocidental; EAFR: África Oriental; ORM: Oriente Médio; EUR: Europa; CSA: Centro Sul Asiático; EAS: Leste Asiático; OCE: Oceania.

Os resultados encontrados corroboram estudos prévios que utilizaram o painel do CEPH-HGDP ou populações similares (Rosenberg *et al*., 2002; 2005; Bastos-Rodrigues *et al*., 2006; Li *et al*., 2008) e as análises da dissertação elaborada por Chevitarese a partir dos mesmos dados. Os grupos continentais com maior heterogeneidade genética interpopulacional são: América do Sul, Oceania e América Central. Os conjuntos populacionais com menor variabilidade interpopulacional são

Europa, Oriente Médio, Leste Asiático e Centro Sul Asiático, representados por regiões que compõe a Eurásia. As análises de variância molecular demonstram que a maior parte da variância em nosso conjunto de dados se deve a variabilidade dentro das populações.

### *4.3.2 Valores de $F_{ST}$ para todas as subpopulações CEPH-HGDP e Nativo-Americanos*

Na tabela 19 estão representados os dez SNPs com maiores valores de diferenciação entre as subpopulações desse estudo.

**Tabela 19: 10 maiores valores globais de $F_{ST}$**

| Gene | SNP | $F_{ST}$ | p valor |
|---|---|---|---|
| *CYP19A1* | CYP1A1-91 | 0,4248 | 0,0000 |
| *CASP3* | CASP3-08 | 0,4184 | 0,0000 |
| *CASR* | CASR-11 | 0,3769 | 0,0000 |
| *LEPR* | LEPR-01 | 0,3705 | 0,0000 |
| *FANCA* | FANCA-22 | 0,3578 | 0,0000 |
| *FANCA* | FANCA-37 | 0,3576 | 0,0000 |
| *RAD51* | rs2619681 | 0,3516 | 0,0000 |
| *IL4* | IL4-11 | 0,3492 | 0,0000 |
| *HSD17B2* | HSD17B2-01 | 0,3443 | 0,0000 |
| *IL4* | IL4-10 | 0,3438 | 0,0000 |

*4.3.3 Identificação de SNPs com alta divergência populacional*

EUR-NAT-WAFR

A partir dos critérios de configuração populacional e seleção encontrou-se 196 marcadores distribuídos por 111 genes com altos valores de diferenciação interpopulacional (Tabela 20).

Quando se analisa os resultados sob a perspectiva de diferenciação entre populações pode-se selecionar 35 SNPs que têm maior diferenciação entre Europa e África Ocidental e Nativo-Americanos (EUR) – (WAFR – NAT), ou seja, a diferenciação entre africanos e nativos para esses loci não é alta, mas os europeus são distintos das duas outras populações. Para os Nativo-Americanos existem 53 marcadores com freqüências alélicas significativamente diferentes das encontradas em europeus e africanos (NAT) – (EUR-WAFR). E como esperado, a população que apresenta maior diferenciação em respeito às demais é a população da África Ocidental (WAFR) – (EUR-NAT) com 75 polimorfismos. Para 33 polimorfismos as freqüências alélicas encontradas nas três populações apresentam semelhante variância interpopulacional (Tabela 20).

**Tabela 20: Tabela de freqüências alélicas para os grupos populacionais África Ocidental, Europa e Nativo-Americanos**

| Gene | SNP | Gen | EUR | NAT | WAFR | Gene | SNP | Gen | EUR | NAT | WAFR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *ABCA1* | ABCA1-17 | A/T | 0,2089 | 0,1834 | 0,8839 | *IL15* | IL15-02 | T/C | 0,3671 | 0,0263 | 0,1000 |
| *AKR1C3* | AKR1C3-11 | G/A | 0,3822 | 0,8743 | 0,4732 | | IL15-06 | C/T | 0,4808 | 0,0318 | 0,3839 |
| | AKR1C3-36 | G/A | 0,3228 | 0,8634 | 0,3393 | *IL1B* | IL1B-03 | C/T | 0,3000 | 0,8924 | 0,5926 |
| *AMACR* | AMACR-03 | G/T | 0,1424 | 0,0088 | 0,5273 | *IL2* | IL2-03 | T/G | 0,3227 | 0,6994 | 0,0535 |
| | AMACR-17 | G/A | 0,4367 | 0,7485 | 1 | *IL4* | IL4-01 | T/C | 0,1392 | 0,6311 | 0,6574 |
| *ANKK1* | ANKK1-01 | A/G | 0,1783 | 0,6363 | 0,3909 | | IL4-03 | T/C | 0,1338 | 0,6235 | 0,4455 |
| *APC* | APC-09 | T/C | 0,3354 | 0,2000 | 0,8303 | | IL4-10 | A/C | 0,1369 | 0,6324 | 0,3000 |
| *AURKA* | AURKA-16 | C/T | 0,2468 | 0,7882 | 0,0893 | | IL4-11 | C/A | 0,1378 | 0,6308 | 0,2143 |
| *BCL2L1* | BCL2L1-01 | T/G | 0,2134 | 0,0203 | 0,5463 | *IL4R* | IL4R-02 | C/A | 0,0892 | 0,0953 | 0,7091 |
| | BCL2L1-02 | C/T | 0,2908 | 0,0303 | 0,6321 | | IL4R-07 | G/T | 0,0621 | 0,0087 | 0,3704 |
| *BCL6* | BCL6-09 | T/C | 0,3365 | 0,0714 | 0,9018 | *IL6* | IL6-01 | C/G | 0,3662 | 0,0118 | 0 |
| *BIC* | BIC-07 | T/C | 0,2342 | 0,7308 | 0,6786 | | IL6-04 | A/G | 0,3548 | 0,0117 | 0 |
| | BIC-10 | T/G | 0,2389 | 0,7289 | 0,6250 | *IL6R* | IL6R-04 | C/A | 0,3397 | 0,7083 | 0,0363 |
| | BIC-15 | C/T | 0,1709 | 0,6794 | 0,4107 | *IL7R* | IL7R-01 | G/A | 0,2866 | 0,7093 | 0,0982 |
| | BIC-32 | G/A | 0,2342 | 0,7289 | 0,7364 | *INSR* | INSR-13 | G/A | 0,4363 | 0,1192 | 0,8393 |
| | BIC-34 | G/A | 0,2500 | 0,7278 | 0,7411 | *KRT23* | KRT23-03 | T/C | 0,2690 | 0,7169 | 0,2130 |
| *BRIP1* | BRIP1-01 | A/G | 0,4427 | 0,8533 | 0,1339 | *LCAT* | LCAT-05 | G/A | 0,1690 | 0,1804 | 0,7600 |
| | BRIP1-09 | T/C | 0,2057 | 0,7794 | 0,3304 | *LIPC* | LIPC-04 | G/T | 0,1139 | 0,1898 | 0,7411 |
| *CASP3* | CASP3-08 | C/T | 0,2595 | 0,9556 | 0,1727 | | LIPC-37 | T/C | 0,0601 | 0 | 0,3304 |
| *CASP8* | CASP8-07 | G/T | 0,0892 | 0,0116 | 0,5278 | *LRP5* | LRP5-01 | T/C | 0,2633 | 0,7107 | 0,1364 |
| *CASR* | CASR-11 | A/G | 0,2821 | 0,9737 | 0,2232 | *MASP1* | rs696405 | A/C | 0,3548 | 0,1316 | 0,7500 |
| *CAT* | CAT-02 | G/A | 0,3636 | 0,8107 | 0,4907 | *MATR3* | MATR3-01 | T/A | 0,2911 | 0,4881 | 0,9554 |
| *CAV1* | CAV1-19 | T/A | 0,3636 | 0,1893 | 0,4907 | *MBL2* | MBL2-46 | C/T | 0,4272 | 0,0439 | 0,8273 |
| | CAV1-29 | C/T | 0,3291 | 0,0117 | 0,4107 | *MSH3* | MSH3-02 | A/G | 0,0601 | 0,5292 | 0,1785 |
| *CDK5* | CDK5-08 | C/A | 0,2468 | 0,8713 | 0,1909 | | MSH3-07 | G/C | 0,1361 | 0,5407 | 0,1339 |
| | CDK5-16 | A/G | 0,2338 | 0,8713 | 0,2636 | *MTRR* | MTRR-19 | T/A | 0,3892 | 0,8982 | 0,5536 |
| *CDKN2A* | CDKN2A-03 | T/C | 0,0886 | 0,6455 | 0,1875 | *MX1* | MX1-11 | G/A | 0,4363 | 0,8970 | 0,8571 |
| *CGA* | CGA-06 | G/A | 0,2816 | 0,6559 | 0,1161 | | MX1-22 | C/T | 0,3070 | 0,0146 | 0,5893 |
| *CYP19A1* | CYP19A1-01 | G/A | 0,4841 | 0,0617 | 0,1250 | | MX1-28 | G/A | 0,500 | 0,0896 | 0,1607 |
| | CYP19A1-04 | T/G | 0,4905 | 0,9041 | 0,8661 | *MYBL2* | MYBL2-03 | A/G | 0,0854 | 0,0058 | 0,3868 |
| | CYP19A1-06 | T/G | 0,4905 | 0,9077 | 0,8273 | *MYC* | MYC-02 | A/T | 0,1487 | 0,1928 | 0,7545 |
| | CYP19A1-08 | T/G | 0,2722 | 0,8488 | 0,2500 | *MYNN* | MYNN-01 | G/A | 0,2994 | 0,6518 | 0,0182 |
| | CYP19A1-29 | C/T | 0,4744 | 0,0643 | 0,0818 | *NCF2* | NCF2-03 | G/A | 0,4209 | 0,8520 | 0,2856 |
| | CYP19A1-30 | G/T | 0,0443 | 0,2235 | 0,6339 | | NCF2-04 | A/G | 0,4236 | 0,8529 | 0,2818 |
| | CYP19A1-34 | T/C | 0,481 | 0,0636 | 0,1161 | *NCOA3* | NCOA3-04 | G/A | 0,0886 | 0,0232 | 0,4196 |
| | CYP19A1-39 | T/C | 0,4583 | 0,0581 | 0,2054 | *NFKB1* | NFKB1-02 | C/T | 0,3544 | 0,7156 | 0,0089 |
| *CYP1A1* | CYP1A1-14 | G/T | 0,3000 | 0,8876 | 0,9821 | | NFKB1-33 | T/A | 0,3590 | 0,7147 | 0,1518 |
| *CYP1B1* | CYP1B1-27 | C/T | 0,4548 | 0,6717 | 0 | *NFKBIE* | NFKBIE-02 | T/C | 0,5000 | 0,2485 | 0,900 |
| | CYP1B1-31 | C/A | 0,1677 | 0,0268 | 0,7273 | *NR1H4* | NR1H4-05 | C/G | 0,3766 | 0,8601 | 0,6111 |

| Gene | SNP | Gen | EUR | NAT | WAFR | Gene | SNP | Gen | EUR | NAT | WAFR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *CYP2E1* | CYP2E1-02 | G/C | 0,1772 | 0,0446 | 0,7232 | *OCA2* | OCA2-23 | G/A | 0,4525 | 0,9315 | 0,8750 |
| | CYP2E1-31 | G/T | 0,0860 | 0,4357 | 0,6296 | *PAK6* | PAK6-13 | A/C | 0,3874 | 0,7981 | 0,0185 |
| *CYP3A7* | CYP3A7-01 | C/T | 0,0949 | 0,2471 | 0,6607 | *PCNA* | PCNA-10 | C/A | 0,1677 | 0,0462 | 0,7232 |
| *DHDH* | DHDH-02 | G/C | 0,2166 | 0,6627 | 0,8393 | *PCTP* | PCTP-01 | G/A | 0,1146 | 0,3314 | 0,6875 |
| *DRD2* | DRD2-03 | A/G | 0,1442 | 0,6317 | 0,2232 | *PHB* | PHB-02 | G/A | 0,4522 | 0,0088 | 0 |
| *EFNB3* | EFNB3-02 | G/A | 0,3981 | 0,8185 | 0,9364 | *PIM1* | PIM1-03 | G/A | 0,2949 | 0,0202 | 0,4911 |
| *ENPP1* | ENPP1-04 | A/T | 0,0570 | 0,0058 | 0,4091 | *PMS1* | rs1233291 | C/G | 0,2722 | 0,2246 | 0,9732 |
| *EPHX2* | EPHX2-04 | C/A | 0,2612 | 0,1441 | 0,7143 | | rs1233297 | T/C | 0,2816 | 0,2135 | 0,9022 |
| *ERCC1* | ERCC1-05 | C/T | 0,4025 | 0,8274 | 0,9464 | | rs1233299 | C/A | 0,2707 | 0,2130 | 0,8482 |
| | ERCC1-06 | G/C | 0,3790 | 0,8485 | 0,9107 | | rs1233302 | A/C | 0,2727 | 0,2209 | 0,9259 |
| *ERCC5* | ERCC5-01 | T/C | 0,3829 | 0,8706 | 0,2767 | | rs256550 | C/T | 0,1346 | 0,2076 | 0,7232 |
| *ESR1* | ESR1-17 | T/G | 0,0949 | 0,0152 | 0,5833 | | rs256552 | G/A | 0,1424 | 0,2078 | 0,7232 |
| *FANCA* | FANCA-03 | T/C | 0,2660 | 0,8385 | 0,4727 | | rs256563 | G/A | 0,1424 | 0,2118 | 0,7182 |
| | FANCA-16 | G/C | 0,2660 | 0,8363 | 0,6545 | | rs256564 | A/G | 0,1424 | 0,2081 | 0,7232 |
| | FANCA-22 | C/T | 0,3510 | 0,8567 | 0,7321 | | rs256567 | T/C | 0,1424 | 0,2135 | 0,7143 |
| | FANCA-28 | A/G | 0,2677 | 0,8445 | 0,5189 | | rs5742938 | G/A | 0,2707 | 0,2091 | 0,8611 |
| | FANCA-35 | C/G | 0,2742 | 0,8343 | 0,5566 | *POLB* | POLB-05 | C/T | 0,1146 | 0,0149 | 0,8241 |
| | FANCA-37 | A/T | 0,3462 | 0,8410 | 0,8455 | | POLB-08 | G/A | 0,0728 | 0,0266 | 0,6818 |
| *FASLG* | FASLG-01 | A/G | 0,4172 | 0,7719 | 0,0268 | | POLB-16 | G/A | 0,0728 | 0,0203 | 0,6852 |
| *FBXW7* | FBXW7-01 | A/G | 0,2643 | 0,5473 | 0,8750 | *POLD1* | POLD1-13 | C/T | 0,0947 | 0,0507 | 0,4909 |
| | FBXW7-05 | C/T | 0,2866 | 0,0029 | 0 | *RAD51* | rs2412546 | G/A | 0,4490 | 0,1036 | 0,7768 |
| | FBXW7-44 | A/G | 0,2722 | 0,4360 | 0,9732 | | rs2619679 | A/T | 0,4430 | 0,1047 | 0,7857 |
| *FUT2* | FUT2-05 | C/T | 0,4712 | 0,0349 | 0,3929 | | rs2619681 | T/C | 0,1465 | 0,7602 | 0,1827 |
| *GATA3* | GATA3-25 | G/C | 0,2532 | 0,0265 | 0,5446 | | rs4924496 | T/C | 0,3903 | 0,0298 | 0,3909 |
| *GDF15* | GDF15-02 | A/T | 0,1784 | 0,6717 | 0,1696 | *RAD52* | RAD52-07 | C/T | 0,3662 | 0,2006 | 0,7857 |
| *GHR* | GHR-21 | T/C | 0,1582 | 0,0087 | 0,5625 | *RAG1* | RAG1-01 | G/A | 0,1170 | 0,5977 | 0,0714 |
| | GHR-47 | A/C | 0,2373 | 0,0089 | 0,6455 | *RB1CC1* | RB1CC1-10 | C/T | 0,1762 | 0,6598 | 0,0818 |
| *GPX2* | GPX2-17 | A/G | 0,2917 | 0,3825 | 0,9273 | | RB1CC1-24 | C/T | 0,1815 | 0,6479 | 0,0893 |
| | GPX2-21 | T/C | 0,2885 | 0,3899 | 0,9455 | *RERG* | RERG-24 | G/A | 0,2917 | 0,0059 | 0,5268 |
| *GPX3* | GPX3-28 | A/G | 0,1465 | 0,0123 | 0,4519 | | RERG-37 | A/C | 0,1266 | 0,5536 | 0,0268 |
| *GSK3B* | rs16830689 | G/C | 0,2057 | 0,0233 | 0,6250 | | RERG-47 | T/C | 0,2911 | 0,0060 | 0,5636 |
| | rs1719889 | T/A | 0,2184 | 0,0234 | 0,6339 | *RGS5* | RGS5-01 | C/A | 0,0854 | 0,0058 | 0,4821 |
| | rs1732170 | A/G | 0,2848 | 0,3029 | 0,9375 | *RNASEL* | RNASEL-02 | A/G | 0,3846 | 0,0175 | 0,1339 |
| | rs334535 | A/G | 0,2057 | 0,0231 | 0,6250 | *SCARB1* | SCARB1-03 | A/G | 0,3471 | 0,8266 | 0,3482 |
| | rs334559 | T/C | 0,2025 | 0,0291 | 0,6273 | *SEPP1* | SEPP1-01 | A/G | 0,2675 | 0,6235 | 0,0804 |
| | rs4072520 | T/G | 0,2184 | 0,0231 | 0,6091 | *SLAMF1* | SLAMF1-03 | G/A | 0,4684 | 0,0769 | 0,7946 |
| | rs6770314 | T/C | 0,2057 | 0,0203 | 0,6339 | *SLC23A1* | SLC23A1-09 | C/T | 0,3362 | 0,4889 | 0,9898 |
| | rs7617372 | C/T | 0,2184 | 0,0291 | 0,6339 | *SLC4A2* | SLC4A2-02 | G/A | 0,2219 | 0,8727 | 0,4107 |
| | rs7620750 | T/C | 0,2197 | 0,0233 | 0,6339 | *SLC6A3* | SLC6A3-10 | G/A | 0,2019 | 0,0152 | 0,4727 |
| | rs9851174 | T/C | 0,2057 | 0,0234 | 0,6250 | *SOAT2* | SOAT2-01 | G/A | 0,1847 | 0,1765 | 0,7273 |

| | SNP | Gen | EUR | NAT | WAFR | Gene | SNP | Gen | EUR | NAT | WAFR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | rs9878473 | G/A | 0,4304 | 0,3588 | 1 | *SOD1* | SOD1-01 | G/A | 0,0665 | 0,4792 | 0,2091 |
| *GSTM3* | GSTM3-06 | T/G | 0,4051 | 0,8482 | 0,1545 | *SOD3* | SOD3-05 | T/C | 0,2792 | 0,5000 | 0,9643 |
| *HSD17B2* | HSD17B2-01 | G/A | 0,0285 | 0,6441 | 0,1339 | *TCTA* | TCTA-04 | G/A | 0,4262 | 0,9765 | 0,1852 |
| *HSD3B1* | HSD3B1-18 | G/T | 0,3608 | 0,1111 | 0,7857 | *TERT* | TERT-02 | T/C | 0,3390 | 0,9639 | 0,6400 |
| | HSD3B1-22 | G/A | 0,3726 | 0,0116 | 0,0536 | *TLR2* | TLR2-04 | C/T | 0,4588 | 0,1104 | 0,5999 |
| | HSD3B1-24 | G/T | 0,3703 | 0,1140 | 0,7857 | | TLR2-06 | T/A | 0,4708 | 0,9649 | 0,6091 |
| | HSD3B1-25 | A/G | 0,3408 | 0,0088 | 0,0714 | *TP73L* | TP73L-13 | A/G | 0,2184 | 0,0173 | 0,5179 |
| | HSD3B1-26 | A/G | 0,3854 | 0,0145 | 0,2411 | | TP73L-15 | G/A | 0,2532 | 0,0145 | 0,7411 |
| *HSD3B2* | HSD3B2-14 | T/G | 0,1571 | 0,1882 | 0,7182 | | TP73L-16 | A/G | 0,4873 | 0,8462 | 0,2411 |
| *IFNAR2* | IFNAR2-01 | A/G | 0,3248 | 0,7292 | 0,1071 | | TP73L-26 | T/C | 0,1677 | 0,8095 | 0,2143 |
| | IFNAR2-06 | G/T | 0,3653 | 0,7336 | 0,1518 | | TP73L-28 | T/C | 0,4013 | 0,8364 | 0,2500 |
| | IFNAR2-10 | A/G | 0,3280 | 0,7297 | 0,1875 | *VCAM1* | VCAM1-05 | G/A | 0,0285 | 0,0058 | 0,3125 |
| *IGF1R* | IGF1R-05 | A/G | 0,3333 | 0,1813 | 0,8036 | *WDR79* | WDR79-06 | T/A | 0,0316 | 0,0029 | 0,3727 |
| | IGF1R-18 | A/G | 0,3344 | 0,1837 | 0,8036 | | WDR79-11 | G/C | 0,1139 | 0,1294 | 0,9272 |
| *IGF2* | IGF2-16 | G/C | 0,3726 | 0,0340 | 0,6909 | *XRCC4* | XRCC4-01 | A/G | 0,1044 | 0,5500 | 0,5546 |
| *IGFBP5* | IGFBP5-10 | C/T | 0,3070 | 0,7122 | 0,8036 | | XRCC4-05 | T/G | 0,4522 | 0,0434 | 0,6182 |
| *IGFBP6* | IGFBP6-19 | T/C | 0,1154 | 0,0799 | 0,6786 | | XRCC4-10 | C/T | 0,0981 | 0,5497 | 0,5625 |
| *IL13* | IL13-01 | A/G | 0,1419 | 0,7852 | 0,2641 | *XRCC5* | XRCC5-12 | A/G | 0,3942 | 0,8639 | 0,1964 |
| | IL13-06 | T/C | 0,1474 | 0,7929 | 0,8214 | | XRCC5-19 | G/A | 0,4487 | 0,8675 | 0,3019 |

Genótipo: As freqüências alélicas foram escolhidas de acordo com o alelo de menor freqüência na população européia. EUR: Freqüência alélica na população européia. NAT: Freqüência alélica na população Nativo-Americana. WAFR: Freqüência alélica para a população da África Ocidental. Os SNPs marcados em azul estão localizados em éxons. As células marcadas em cinza-escuro indicam que aquela população apresenta freqüências alélicas mais diferenciadas em relação às demais populações.

Os genes com maior quantidade de polimorfismos são: *GSKB3* (11 SNPs), *PMS1* (10), *CYP19A1* (8), *FANCA* (6), *BIC* (5), *HSD3B1* (5), *TP73L* (5), *IL4* (4) e *RAD51* (4). Dos polimorfismos encontrados, 30 estão localizados em éxons e 19 estão localizados na região promotora. Os maiores valores de $F_{CT}$ estão descritos na tabela 21.

**Tabela 21: SNPs com maior divergência entre EUR-NAT-WAFR**

| Gene | SNP | Substituição | Região | $F_{CT}$ | $F_{SC}$ |
|---|---|---|---|---|---|
| *CASP3* | CASP3-08 | | Ex8+567 | 0,6314 | 0,0783 |
| *POLB* | POLB-05 | | IVS1-89 | 0,6295 | 0,0249 |
| *CASR* | CASR-11 | | IVS1+20204 | 0,6221 | 0,0803 |
| *WDR79* | WDR79-11 | R68G | Ex1-230C>G | 0,5563 | 0,0586 |
| *TCTA* | TCTA-04 | | IVS2+321A>G | 0,5531 | 0,0195 |
| *POLB* | POLB-16 | | IVS2-2264 | 0,5417 | 0,0848 |
| *IL13* | IL13-06 | | IVS3-24 | 0,5311 | 0,0243 |
| *POLB* | POLB-08 | | IVS7+171 | 0,5260 | 0,0898 |
| *CYP1A1* | CYP1A1-14 | | IVS1+606 | 0,5238 | 0,0296 |
| *CDK5* | CDK5-08 | | IVS7+11 | 0,5226 | 0,0114 |

NAT-WAFR

Da análise par-a-par entre os grupos África Ocidental e Nativo-Americanos emergiram 235 SNPs, alocados em 128 genes, com elevado grau de diferenciação entre os conjuntos. Os genes com maior representatividade são: *GSKB3* (14 polimorfismos), *PMS1* (13), *MASP1* (9), *MBL2* (8), *AKR1C3* (5), *GATA3* (5), *GPX3* (5), *RAD51* (5) e *TP73L* (5). Obteve-se 34 mutações em éxons distribuídas por 32 genes e 26 polimorfismos encontram-se na região promotora de 15 genes. Ver tabela 22, onde estão os 10 loci com maior distância genética entre Nativo-Americanos e Africanos do Oeste.

**Tabela 22: SNPs com maior divergência entre NAT-WAFR**

| Gene | SNP | Substituição | Região | $F_{CT}$ | $F_{SC}$ |
|---|---|---|---|---|---|
| *POLB* | POLB-05 | | IVS1-89 | 0,8747 | 0,0280 |
| *TCTA* | TCTA-04 | | IVS2+321A>G | 0,8509 | 0,0587 |
| *MBL2* | MBL2-46 | | IVS2-405 | 0,8208 | 0,0063 |
| *TP73L* | TP73L-15 | | | 0,8185 | 0,0166 |
| *CYP1B1* | CYP1B1-31 | | Ex3+939 | 0,7780 | 0,0561 |
| *GHR* | GHR-47 | | IVS2+4144 | 0,7589 | 0,0481 |
| *WDR79* | WDR79-11 | R68G | Ex1-230C>G | 0,7534 | 0,0674 |
| *CYP2E1* | CYP2E1-02 | | IVS7-118 | 0,7364 | -0,0081 |
| *IGF2* | IGF2-16 | | IVS1-285 | 0,7320 | -0,0077 |
| *PCNA* | PCNA-10 | | IVS5+140 | 0,7305 | 0,0725 |

EUR-NAT

A análise de disparidade entre as populações da Europa e Nativo-Americanos apresenta 155 SNPs pertencentes a 90 genes com grande variância interpopulacional. O gene com maior número de marcadores selecionados é *CYP19A1* com 8, seguido por *AKR1C3* e *FANCA* com 6 e *BIC* representado por 5. Tem-se ainda 24 polimorfismos exônicos distribuídos por 21 genes e as regiões promotoras de 16 genes são representadas por 22 loci diferenciados. Os loci com valores de $F_{CT}$ mais altos para a comparação entre Europeus e Nativo-Americanos estão evidenciados na tabela 23.

| Tabela 23: SNPs com maior divergência entre EUR-NAT | | | | | |
|---|---|---|---|---|---|
| **Gene** | **SNP** | **Substituição** | **Região** | **$F_{CT}$** | **$F_{SC}$** |
| *CASR* | CASR-11 | | IVS1+20204 | 0,6826 | 0,0345 |
| *CASP3* | CASP3-08 | | Ex8+567 | 0,6745 | 0,0752 |
| *TERT* | TERT-02 | | IVS10+269 | 0,6101 | 0,0274 |
| *SLC4A2* | SLC4A2-02 | | IVS1-530 | 0,5997 | 0,0061 |
| *IL13* | IL13-06 | | IVS3-24 | 0,5858 | 0,0244 |
| *HSD17B2* | HSD17B2-01 | | IVS4-2328 | 0,5837 | 0,0783 |
| *IL13* | IL13-01 | Q144R | Ex4+98 | 0,5831 | 0,0281 |
| *CDK5* | CDK5-16 | | -903 | 0,5828 | 0,0122 |
| *TP73L* | TP73L-26 | | | 0,5796 | 0,0459 |
| *CDK5* | CDK5-08 | | IVS7+11 | 0,5668 | 0,0139 |

EUR-WAFR

Outra configuração populacional analisada foi formada pelas populações Europa e África Ocidental. Nesta análise obtive-se o maior número de loci com alta variância populacional, foram selecionados 220 polimorfismos em 119 genes. Dois genes tiveram mais de 10 SNPs com valores de $F_{CT}$ superior a 0,25, sendo eles *GSKB3* (19 SNPs) e *PMS1* (11 SNPs), além desses pode-se citar *CYP19A1* (6 SNPs) e *MYBL2* (5 SNPs) como os genes com maior número de marcadores diferenciados. Quando as substituições nucleotídicas localizadas em éxons são consideradas, obtêm-se um total de 36 loci distribuídos por 31 regiões gênicas. A tabela 24 evidencia os SNPs com maiores distâncias genéticas entre Europeus e Africanos Ocidentais.

| Tabela 24: SNPs com maior divergência entre EUR-WAFR | | | | | |
|---|---|---|---|---|---|
| **Gene** | **SNP** | **Substituição** | **Região** | **$F_{CT}$** | **$F_{SC}$** |
| *WDR79* | WDR79-11 | R68G | Ex1-230C>G | 0,7788 | 0,0409 |
| *EGF* | EGF-04 | | IVS22-1443 | 0,6931 | -0,0066 |
| *POLB* | POLB-05 | | IVS1-89 | 0,6877 | 0,0220 |
| *CYP19A1* | CYP19A1-30 | | IVS2+14872 | 0,6508 | 0,0171 |
| *ALDH2* | ALDH2-08 | | IVS1+6933 | 0,6432 | -0,0089 |
| *POLB* | POLB-16 | | IVS2-2264 | 0,6310 | 0,0923 |
| *IL13* | IL13-06 | | IVS3-24 | 0,6295 | 0,0283 |
| *CYBB* | CYBB-27 | | IVS12-350 | 0,6293 | 0,0049 |
| *IL4R* | IL4R-02 | E400A | Ex12+300 | 0,6274 | -0,0123 |
| *POLB* | POLB-08 | | IVS7+171 | 0,6268 | 0,0881 |

EAS-NAT

O exame de diferenciação entre as populações do Leste Asiático e Nativo-Americanos demonstrou a existência de 36 polimorfismos, pertencentes a 26 genes, com freqüências alélicas muito distintas entre os dois conjuntos populacionais. O gene com maior quantidade de marcadores é *IL1RN*, com três. Os marcadores com maior divergência entre Leste Asiático e Nativo-americanos estão relatados na tabela 25.

**Tabela 25: SNPs com maior divergência entre EAS-NAT**

| Gene | SNP | Substituição | Região | $F_{CT}$ | $F_{SC}$ |
|---|---|---|---|---|---|
| *IGF2* | IGF2-16 | | IVS1-285 | 0,5855 | 0,0272 |
| *TSPO* | TSPO-03 | | IVS2-136C>G | 0,5555 | 0,0759 |
| *IGF2* | IGF2-02 | | IVS2+384 | 0,5507 | 0,0351 |
| *BRIP1* | BRIP1-09 | | IVS14+3238 | 0,5356 | 0,0244 |
| *CDKN1B* | CDKN1B-04 | | Ex3-387 | 0,5220 | 0,0599 |
| *ADH1C* | ADH1C-16 | | IVS6-680 | 0,4945 | 0,0855 |
| *OPRM1* | OPRM1-02 | | IVS1+11468 | 0,4697 | 0,0820 |
| *CDKN2A* | CDKN2A-03 | | Ex4+83 | 0,4646 | 0,0205 |
| *CCND1* | CCND1-03 | | Ex5+230 | 0,4562 | 0,0418 |
| *CD86* | CD86-02 | A310T | Ex8+35 | 0,4431 | 0,0338 |

NAT-NEAS

Outra configuração analisada envolveu o conjunto populacional NEAS (Nordeste Asiático), formado pelas subpopulações Daur, Hezhen e Oroqen e o conjunto populacional Nativo-Americanos. A população NEAS foi composta pelas três subpopulações do Leste Asiático mais próximas ao Estreito de Bering, à exceção da subpopulação Yakut. Tal proposta pode elucidar quais populações do Leste Asiático, dentre aquela presentes no painel do CEPH-HGDP, tiveram maior contribuição na composição genética Nativo-Americana.

Forte estrutura genética entre os conjuntos populacionais NEAS e Nativo-Americanos é aparente para 25 polimorfismos circunscritos a 21 genes. Os genes com maior número de marcadores selecionados são *CYP19A1* (3 SNPs), *IGF2* e *OPRM1*

(2 SNPs cada). Dentre os 25 marcadores nenhum se localiza na região promotora e apenas três causam substituições não sinônimas, APOB-01, EPXH2-04 e OPRM1-01. Os SNPs com maior divergência genética entre Nativo-Americanos e Subpopulações do Nordeste Asiático estão representados na tabela 26, abaixo.

**Tabela 26: SNPs com maior divergência entre NAT-NEAS**

| Gene | SNP | Substituição | Região | $F_{CT}$ | $F_{SC}$ |
|---|---|---|---|---|---|
| *IGF2* | IGF2-16 | | IVS1-285 | 0,6931 | 0,0605 |
| *IGF2* | IGF2-02 | | IVS2+384 | 0,6281 | 0,0778 |
| *STAT1* | STAT1-01 | | IVS21-8 | 0,5681 | 0,0750 |
| *GATA3* | GATA3-25 | | IVS4-2162 | 0,5284 | 0,0480 |
| *BRIP1* | BRIP1-09 | | IVS14+3238 | 0,5171 | 0,0486 |
| *HSD17B2* | HSD17B2-01 | | IVS4-2328 | 0,4485 | 0,0850 |
| *OPRM1* | OPRM1-23 | | IVS3-30 | 0,4433 | 0,0993 |
| *TP73L* | TP73L-26 | | | 0,4328 | 0,0415 |
| *CCND1* | CCND1-03 | | Ex5+230 | 0,4073 | 0,0645 |
| *PAK6* | PAK6-13 | | Ex11+696 | 0,4021 | 0,0485 |

NAT-YAK

As dessemelhanças entre a subpopulação Yakut e a população Nativo-Americana podem ser ilustradas pela quantidade de polimorfismos com grande variância interpopulacional, o que corresponde a 98 SNPs alocados em 66 genes. O gene *CYP19A1* apresenta a maior quantidade de marcadores com altos valores de $F_{CT}$, seis marcadores, seguido por *MLB2* com quatro. Apenas *BRIP1* contêm mais de uma substituição não sinônimas, BRIP1-02 e BRIP1-05. Quanto à região promotora, cinco genes têm dois ou mais representantes, *GPX3, MBL2*, *FOXC1*, *NFKBIE* e *TP53I3*. Os maiores valores de distância genética entre Nativo-Americanos e Yakuts estão representados na tabela 27.

| Tabela 27: SNPs com maior divergência entre NAT-YAK | | | | | |
|---|---|---|---|---|---|
| **Gene** | **SNP** | **Substituição** | **Região** | **$F_{CT}$** | **$F_{SC}$** |
| *CASR* | CASR-05 | | IVS1-4243 | 0,6710 | 0,0448 |
| *IGF2* | IGF2-16 | | IVS1-285 | 0,6350 | -0,0042 |
| *MASP1* | rs1533593 | | | 0,6244 | -0,0254 |
| *TLR2* | TLR2-06 | | IVS1+1614 | 0,6146 | -0,0017 |
| *GATA3* | GATA3-25 | | IVS4-2162 | 0,5876 | -0,0018 |
| *MX1* | MX1-22 | R441R | Ex14+50 | 0,5863 | -0,0163 |
| *CD86* | CD86-02 | A310T | Ex8+35 | 0,5847 | 0,0345 |
| *RNASEL* | RNASEL-02 | R462Q | Ex1-96 | 0,5676 | -0,0127 |
| *RERG* | RERG-44 | | IVS2-30357 | 0,5485 | -0,0139 |
| *STAT1* | STAT1-01 | | IVS21-8 | 0,5418 | 0,0358 |

DAUR-NAT

Para a configuração populacional (Daur) – (Nativo-Americanos), identifica-se 44 genes estruturados geneticamente, sendo estes representados por 65 polimorfismos. Os genes com maior estruturação, no que diz respeito ao número de marcadores, são *CYP19A1* (6 SNPs), *TNKS* e *TP73L* (4 SNPs) e *CDKN2A* (3 SNPs). As substituições localizadas em éxons somam oito, distribuídas por sete genes, sendo que *ABCA1* possui duas (ABCA1-12 e ABCA1-16). Os polimorfismos em região promotora totalizam cinco, sendo que nenhum gene tem mais de representante. Abaixo estão os 10 maiores valores de divergência entre grupos para essa configuração populacional.

| Tabela 28: SNPs com maior divergência entre DAUR-NAT | | | | | |
|---|---|---|---|---|---|
| **Gene** | **SNP** | **Substituição** | **Região** | **$F_{CT}$** | **$F_{SC}$** |
| *GATA3* | GATA3-25 | | IVS4-2162 | 0,8115 | 0,0090 |
| *IGF2* | IGF2-02 | | IVS2+384 | 0,7503 | 0,0595 |
| *STAT1* | STAT1-01 | | IVS21-8 | 0,6815 | 0,0642 |
| *CD40* | CD40-01 | | IVS1+1066C>T | 0,6499 | 0,0186 |
| *GATA3* | GATA3-28 | | IVS5+60 | 0,6468 | 0,0086 |
| *CASR* | CASR-05 | | IVS1-4243 | 0,6225 | 0,0613 |
| *CDKN2A* | CDKN2A-19 | | IVS1+7291 | 0,6219 | 0,0014 |
| *CD40* | CD40-03 | | IVS8-114G>A | 0,6201 | 0,0176 |
| *BRIP1* | BRIP1-09 | | IVS14+3238 | 0,6058 | 0,0499 |
| *CDKN2A* | CDKN2A-20 | | IVS1+9477 | 0,6027 | 0,0068 |

HEZ-NAT

A análise de divergência alélica entre Hezhen e Nativo-Americanos indicou a existência de 71 marcadores, em 51 genes, com alta variância entre os dois grupos. Os genes com maior número de SNPs em dissimilaridade são *CYP19A1*, seis marcadores e *CDKN2A*, com quatro. Do total de SNPs selecionados, sete correspondem a substituições exônicas, distribuídos por seis genes: APOB-01, BRIP1-02, CD86-02, CYP19A1-01, RET-01, WDR79-08 e WDR79-11. A tabela 29 relata os polimorfismos com maior diferenciação alélica entre Hezhen e Nativo-Americanos.

**Tabela 29: SNPs com maior divergência entre HEZ-NAT**

| Gene | SNP | Substituição | Região | $F_{CT}$ | $F_{SC}$ |
|---|---|---|---|---|---|
| *STAT1* | STAT1-01 | | IVS21-8 | 0,8708 | 0,0704 |
| *IGF2* | IGF2-02 | | IVS2+384 | 0,8257 | 0,0659 |
| *CCDC97* | CCDC97-03 | | Ex5+1082T>A | 0,7493 | 0,0361 |
| *CASR* | CASR-05 | | IVS1-4243 | 0,6814 | 0,0632 |
| *PARP4* | PARP4-19 | | IVS17-110 | 0,6606 | 0,0172 |
| *ADH1C* | ADH1C-16 | | IVS6-680 | 0,6303 | 0,0932 |
| *CYP19A1* | CYP19A1-39 | | IVS2-14688 | 0,5864 | 0,0765 |
| *ERCC4* | ERCC4-15 | | IVS9-28 | 0,5855 | 0,0183 |
| *CD86* | CD86-02 | A310T | Ex8+35 | 0,5518 | 0,0516 |
| *IL1B* | IL1B-03 | | -580 | 0,5354 | 0,0580 |

NAT-ORO

A última configuração populacional em respeito à formação do *pool* gênico das populações nativo-americanas é (Oroqen) – (Nativos Americanos). Nessa análise identificaram-se 71 SNPs, pertencentes a 49 genes, com alta variância interpopulacional. Os genes com maior número de loci variáveis são: TNKS (5 SNPs), CDKN2  e CYP19A1 (4 SNPs) cada. Dentre os 71, oito correspondem a mutações exônicas e quatro estão localizados na região promotora de seus respectivos genes. Os polimorfismos com valores mais altos de $F_{SC}$ para Oroqen e Nativo-Americanos são mostrados na tabela 30.

| Tabela 30: SNPs com maior divergência entre NAT-ORO | | | | | |
|---|---|---|---|---|---|
| **Gene** | **SNP** | **Substituição** | **Região** | **$F_{CT}$** | **$F_{SC}$** |
| *STAT1* | STAT1-01 | | IVS21-8 | 0,8596 | 0,0662 |
| *IGF2* | IGF2-02 | | IVS2+384 | 0,8084 | 0,0635 |
| *CD40* | CD40-01 | | IVS1+1066C>T | 0,6999 | 0,0185 |
| *CD40* | CD40-03 | | IVS8-114G>A | 0,6766 | 0,0171 |
| *BRIP1* | BRIP1-02 | S919P | Ex19-151 | 0,6439 | 0,0183 |
| *CASR* | CASR-05 | | IVS1-4243 | 0,6225 | 0,0613 |
| *CD86* | CD86-02 | A310T | Ex8+35 | 0,6139 | 0,0500 |
| *BRIP1* | BRIP1-09 | | IVS14+3238 | 0,6058 | 0,0499 |
| *TP73L* | TP73L-13 | | | 0,5851 | 0,0050 |
| *GATA3* | GATA3-28 | | IVS5+60 | 0,5820 | 0,0092 |

Comparações entre as configurações regionais do Leste Asiático e Nativo-Americanos

As comparações entre os 10 maiores valores de $F_{CT}$ para SNPs entre as configurações regionais EAS, NEAS, YAK, DAUR, HEZ e ORO demonstra que os loci com estruturação mais consistente em relação à NAT são: STAT1-01 e CASR-05, em todo o nordeste asiático (NEAS+YAK); IGF2-02 em todo o leste asiático, à exceção dos Yakuts; IGF2-16 em EAS, NEAS e YAK; BRIP1-09 em EAS e NEAS; GATA3-25 em Daur, Yakuts e NEAS e ADH1C-16 em EAS e HEZ (Quadro 2).

| Quadro 2: SNPs entre os maiores valores de $F_{CT}$ para populações do Leste Asiático e Nativo-Americanos | | | | | |
|---|---|---|---|---|---|
| EAS | NEAS | YAK | DAUR | HEZ | ORO |
| IGF2-16 | IGF2-16 | CASR-05 | GATA3-25 | STAT1-01 | STAT1-01 |
| TSPO-03 | IGF2-02 | IGF2-16 | IGF2-02 | IGF2-02 | IGF2-02 |
| IGF2-02 | STAT1-01 | rs1533593 | STAT1-01 | CCDC97-03 | CD40-01 |
| BRIP1-09 | GATA3-25 | TLR2-06 | CD40-01 | CASR-05 | CD40-03 |
| CDKN1B-04 | BRIP1-09 | GATA3-25 | GATA3-28 | PARP4-19 | BRIP1-02 |
| ADH1C-16 | HSD17B2-01 | MX1-22 | CASR-05 | ADH1C-16 | CASR-05 |
| OPRM1-02 | OPRM1-23 | CD86-02 | CDKN2A-19 | CYP19A1-39 | CD86-02 |
| CDKN2A-03 | TP73L-26 | RNASEL-02 | CD40-03 | ERCC4-15 | BRIP1-09 |
| CCND1-03 | CCND1-03 | RERG-44 | BRIP1-09 | CD86-02 | TP73L-13 |
| CD86-02 | PAK6-13 | STAT1-01 | CDKN2A-20 | IL1B-03 | GATA3-28 |

Nomenclatura das populações: EAS (Leste Asiático), NEAS (Nordeste Asiático), YAK (Yakut), DAUR (Daur), HEZ (Hezhen), ORO (Oroqen). SNPs presentes em mais de um conjunto populacional ou subpopulação têm as celas coloridas. Cada cor representa um SNP diferente.

As comparações por gene (onde os SNPs são omitidos em favor dos genes) demonstram que as maiores evidências de diferenciação ocorrem para: *IGF2*, em todas as populações e *GATA3* e *STAT1* em todas as populações do nordeste asiático (Quadro 3).

**Quadro 3: Genes entre os maiores valores de $F_{CT}$ para populações do Leste Asiático e Nativo-Americanos**

| EAS | NEAS | YAK | DAUR | HEZ | ORO |
|---|---|---|---|---|---|
| IGF2 | IGF2 | CASR | GATA3 | STAT1 | STAT1 |
| TSPO-03 | IGF2 | IGF2 | IGF2 | IGF2 | IGF2 |
| IGF2 | STAT1 | MASP1 | STAT1 | CCDC97 | CD40 |
| BRIP1 | GATA3 | TLR2 | CD40 | CASR | CD40 |
| CDKN1B | BRIP1 | GATA3 | GATA3 | PARP4 | BRIP1 |
| ADH1C | HSD17B2 | MX1 | CASR | ADH1C | CASR |
| OPRM1 | OPRM1 | CD86 | CDKN2A | CYP19A1 | CD86 |
| CDKN2A | TP73L | RNASEL | CD40 | ERCC4 | BRIP1 |
| CCND1 | CCND1 | RERG | BRIP1 | CD86 | TP73L |
| CD86 | PAK6 | STAT1 | CDKN2A | IL1B | GATA3 |

Nomenclatura das populações: EAS (Leste Asiático), NEAS (Nordeste Asiático), YAK (Yakut), DAUR (Daur), HEZ (Hezhen), ORO (Oroqen). Os nomes dos SNPs com maior divergência foram omitidos em favor do nome dos genes. Genes presentes em mais de um conjunto populacional ou subpopulação têm as celas coloridas. Cada cor representa um gene distinto.

# 5 DISCUSSÃO

O presente estudo acerca da variabilidade em populações humanas foi desenvolvido em colaboração com a aluna de mestrado Juliana Chevitarese no Laboratório de Diversidade Genética Humana. O enfoque principal da dissertação de mestrado apresentada por Chevitarese é a descrição da estrutura populacional dos grupos e subpopulações, também analisados aqui. É importante ressaltar que diferentes metodologias foram utilizadas nos dois trabalhos, sem que os resultados apresentassem divergências significativas. Desse modo, a discussão acerca da estrutura populacional dos grupos será apresentada resumidamente devido à utilização dos mesmos nas Análises de Variância Molecular, maiores informações sobre grupos populacionais e etnias podem ser vistas em: “Determinação da estrutura genética das populações humanas e inferência dos fatores evolutivos que contribuíram para a sua formação”, dissertação elaborada por Juliana Chevitarese sob orientação do Professor Eduardo Martin Tarazona Santos. Ressalta-se assim que o enfoque principal do presente estudo é a análise de variabilidade dos 1442 loci.

## 5.1 Equilibrio de Hardy Weinberg, Coeficiente de Endocruzamento e Diversidade de Nei (Populacionais)

Desvios a partir do Equilíbrio de Hardy Weinberg e $F_{IS}$ apresentam correlação significativa, Sen e Burmeister (2008) demonstraram que para 17 loci em 325 estudos de associação, essa correlação é de 0,191 (p valor = 0,002). Entretanto, os desvios de Hardy-Weinberg também podem estar relacionados à miscigenação recente, seleção e erros de genotipagem. Para SNPs, o desvio tende a ser devido à deficiência de heterozigotos e indica que quanto maior o afastamento, maior a deficiência de heterozigotos. Além disso, Sen e Burmeister (2008) observaram que os erros de genotipagem se devem principalmente à dificuldade em identificar heterozigotos e que esses erros costumam ser consistentes através dos estudos e ao tamanho amostral.

Pelos motivos acima citados, o Equilíbrio de Hardy Weinberg é essencialmente utilizado como medida de controle de qualidade (Balding, 2006). Inferências a partir dos resultados devem ser cuidadosas, principalmente para desvios em direção à maior proporção de homozigotos. Entretanto, a utilização conjunta de E-HW e $F_{IS}$,

além de testes específicos para identificação de déficits nas proporções de homozigotos e heterozigotos, pode ser de grande valia na descrição genética das populações.

A análise de estrutura genética dentro dos grupos populacionais mostrou resultados similares aos encontrados por Chevitarese (2009). As variações nos valores de $F_{IS}$ (cálculo desse trabalho) e $F_{IT}$ (cálculo do trabalho de Chevitarese) demonstram apenas dois valores de dessemelhança acima de 0,02: Para Oceania, onde a diferença alcançou 0,0424 ($F_{IS}$ = 0,416 nesse trabalho e $F_{IT}$ = 0,0840 no trabalho de Chevitarese); e para América Central, 0,0287 ($F_{IS}$ = 0,0418 nesse, e $F_{IT}$ = 0,0705 no de Chevitarese). Provavelmente, tais diferenças se devem aos métodos de cálculo utilizados nos dois trabalhos. Os cálculos dessa estatística diferem quanto ao método de correção de amostras de tamanho diferente (Rousset, 2007; Goudet, 2005). Assim, a baixa amostragem dos dois grupos com maior diferença provavelmente reflete distinções relativas aos métodos de correção do tamanho amostral. E divergem também em relação à concepção de $F_{IT}$, refletindo a importância da divergência de subpopulações dentro dos grupos nos cálculos de $F_{IS} = F_{IT}$, isso porque as maiores diferenças nos valores de $F_{IS}$ desse trabalho e $F_{IT}$ de Chevitarese correspondem também aos maiores valores de $F_{ST}$ ($F_{SC}$). A comparação dos ranqueamentos entre os valores de $F_{IS}$ e $F_{IT}$ dos dois trabalhos (Chevitarese, Soares-Souza) mostrou que apenas para Oceania e Centro Sul da Ásia houve diferenças superiores a duas posições, respectivamente, oitavo menor valor para quinto; e quinto menor valor para oitavo.

A comparação entre subpopulações foi ainda mais consistente que a por grupos, não houve diferenças superiores a 0,02 nos valores de $F_{IS}$. Todos os valores positivos e negativos coincidiram e as alterações nos ranqueamentos foram pequenas e devido a diferenças, muitas vezes, de poucos milésimos.

As maiores diferenças encontradas no estudo de estruturação subpopulacional ocorreram em relação aos loci com afastamentos do postulado de Hardy-Weinberg. Em geral, o estudo de Chevitarese encontrou maior quantidade de loci com desvios significativos, entretanto esse fato se deve aos diferentes valores de corte para seleção de marcadores, $10^{-4}$ em Chevitarese e $10^{-5}$ em Soares-Souza. Comparações

entre desvios não são possíveis para grupos populacionais e entre os loci para os quais há afastamentos para E-HW, isso porque tais informações não estão disponíveis no estudo de Chevitarese.

A comparação entre os valores de Diversidade de Nei para os dois estudos também foi consistente. Houve pequena variação nos valores encontrados, não mais que 0,02 para cima ou para baixo, tanto para grupos populacionais quanto para subpopulações. Dessa forma, as variações entre os ranqueamentos dos dois estudos são pequenas.

## 5.2 Equilíbrio de Hardy Weinberg e Heterozigosidade Esperada (Loci)

### *5.2.1 Equilíbrio de Hardy Weinberg*

Usualmente o Postulado de Hardy-Weinberg é utilizado como medida de controle de qualidade, entretanto, desvios nas freqüências alélicas podem estar relacionados também a efeitos de estratificação populacional, endogamia e seleção natural (Balding, 2006). Ou seja, afastamentos do E-HW podem invalidar testes de associação caso-controle, principalmente devido às diferenças nas freqüências alélicas entre as populações e a dificuldades na genotipagem de um dos alelos do polimorfismo estudado. A seguir estão relacionados alguns estudos de associação que utilizaram polimorfismos para os quais as freqüências genotípicas distam do que seria esperado pela fórmula $p^2 + 2pq + q^2 = 1$.

Estudo envolvendo o polimorfismo IL4R-07 (rs1805016) em populações brancas não-hispânicas encontrou associação entre o genótipo TT e menor taxa de mortalidade em pacientes com glioma maligno (Scheurer *et al*., 2008). Entretanto, no presente estudo encontramos desvios nas freqüências genotípicas em direção aos homozigotos, especialmente em relação a T. A freqüência genotípica em europeus para TT alcança 0,91, enquanto no estudo de Scheurer e colegas essa freqüência é de 0,89. O estudo requer atenção devido ao excesso de homozigotos para o alelo T em populações européias. O polimorfismo IL4R-07 é não sinônimo e a substituição de T por G leva a alteração de serina por alanina no polipeptídio, entretanto, segundo a tabela de distâncias entre aminoácidos GONNET (Gonnet; Cohen; Benner; 1992) e a

ferramenta *online* Polyphen a alteração S752A provavelmente tem função semelhante e é predita como benigna. Assim, as implicações evolutivas dessa substituição são, provavelmente, pouco significantes.

A associação entre LIPC-01 (rs1800588) e níveis de HDL foi observada por Lu e colegas (2008) em holandeses. Fan e colegas (2009) também encontraram associação em finlandeses entre esse polimorfismo e níveis séricos de colesterol total, HDL, triglicérides e apolipoproteína AI, sendo o genótipo TT relacionado aos maiores níveis. Hamrefors e colaboradores (2009) observaram associação entre rs1800588, em conjunto com outros polimorfismos, e a redução de LDL e aumento de HDL quando há uso concomitante de fluvastatina. Entretanto, Acker e outros (2008) não determinaram associação entre altos níveis de HDL e proteção contra a doença arterial coronária quando há presença de variantes relacionadas à diminuição da concentração de LDL nos genes *CEPT* e *LIPC*. Todos os estudos citados utilizaram populações caucasianas. Porém, as freqüências genotípicas foram discordantes entre o presente estudo e os estudos analisados, HapMap e ALFRED (franceses e espanhóis). Além disso, o desvio de HW indica excesso de heterozigotos, justamente o observado na comparação com as proporções das citadas fontes de freqüências genotípicas. Coincidentemente, apenas a base de dados de SNP500Cancer teve valores semelhantes de freqüências genotípicas, mesmo utilizando o seqüenciamento como método de genotipagem. Tais fatos indicam que possivelmente há erro de genotipagem em nosso conjunto de dados, havendo déficit de alelos G, em especial nos heterozigotos.

O polimorfismo MTR-01 (rs1805087) foi utilizado em estudo de associação entre risco de câncer renal e consumo de vegetais. Polimorfismos em genes relacionados ao metabolismo de folato também foram testados, porém a variante do gene *MTR* não obteve valores significativos de associação (Moore *et al*., 2008). A comparação das freqüências genotípicas obtidas no estudo de Moore e nesta dissertação foi discordante, principalmente em relação às freqüências de AA e GG. A comparação com outras bases de dados não elucidou plenamente a questão, sendo que há inconsistências entre diferentes genotipagens dos mesmos submissores. Em geral, os dados do presente estudo evidenciam um excesso de indivíduos GG em relação a todas as bases de dados. Porém, na comparação entre os dados de CEPH-

HGDP e caucasianos genotipados por Perlegen (CHIP HYB) e seqüenciados através do método de Sanger pelo laboratório de Cold Spring Harbor há déficits de indivíduos AG. Curiosamente, o mesmo laboratório também seqüenciou indivíduos CEPH-CEU e as freqüências foram discrepantes em relação à proporção de indivíduos AA e AG. Para o estudo de Moore e outros, SNP500Cancer (seqüenciamento), Affymetrix (CHIP 250K Sty) e Perlegen (600K) há déficits de homozigotos para A. O estudo de Moore utiliza populações da Europa Central, incluso russos. Impressionantemente, quando apenas a população russa do painel CEPH-HGDP é analisada, a diferença é bastante acentuada, não há homozigotos AA nessa população, sendo que em Moore eles constituem 61% dos indivíduos. A análise de $F_{ST}$ par-a-par indica forte sub-estruturação para esse loci na Europa, onde as populações Adygei e do Cáucaso russo têm altos valores de $F_{ST}$ (> 0,10) em relação às demais. A população da Sardenha também apresenta valores significativos de E-HW para esse SNP em direção aos homozigotos, porém para o alelo A. Possivelmente há mais de um fator levando a essas diferenças nas freqüências genotípicas, talvez a interação entre sub-estruturação populacional e erros de genotipagem.

Talbott e colegas (2008) não encontraram associação entre o risco de câncer de mama pré-menopausa e CYP1A1-16 (rs730154). Nesse estudo foram amostradas mulheres auto-identificadas como brancas (terminologia do autor). Como em uma população caucasiana, a Francesa, esse SNP está fora do equilíbrio de Hardy-Weinberg, poderia haver um erro tipo 2 nesse estudo, erro, entretanto, menos grave que o de falso-positivos. As freqüências alélicas são consistentes entre o observado no presente estudo e o de Talbott e colaboradores, entretanto, as freqüências genotípicas não estão disponíveis para comparação.

### *5.2.2 Heterozigosidade Esperada*

Ao contrário do que seria esperado, não houve diferenças significativas entre os conjuntos de maior e menor variabilidade quanto à distribuição de mutações de ponto em regiões promotoras e codificantes, mesmo entre mutações sinônimas e não-sinônimas (Hughes *et al*., 2003).

Entre os polimorfismos com menor variabilidade, das 18 mutações localizadas

em regiões codificantes, metade é sinônima. De acordo com a ferramenta PolyPhen das nove mutações restantes, sete são preditas como benignas e apenas duas causam alterações estruturais. É importante ressaltar que os termos possivelmente e provavelmente danosa utilizados pela ferramenta PolyPhen se referem à mudanças estruturais e não à morbidade ou letalidade.

MTRR-05 foi classificada como potencialmente danosa e encontra-se em baixa $H_E$ nas populações do Oriente Médio e Europa. Todos os demais grupos populacionais têm valores de $H_E$ superiores aos encontrados nesses dois grupos. Não foram encontradas na literatura evidências de seleção positiva ou associação com doenças para esse polimorfismo. O gene *MTRR* está envolvido na regeneração do cofator de cobalamina, requerido para a manutenção da síntese de metionina (UniProtKB, 2010).

A mutação TEP1-02, inferida como provavelmente danosa, encontra-se em baixa heterozigosidade nos dois grupos populacionais africanos e em valores de heterozigosidade superiores nos demais grupos populacionais. Não há artigos disponíveis demonstrando interação entre esse polimorfismo e doenças ou seleção natural. O gene TEP1 está relacionado à manutenção do telômero via recombinação (Gene Ontology, 2010).

Usualmente, polimorfismos danosos são mantidos em baixas freqüências e eliminados após múltiplos eventos fundadores (Hughes, 2009), tal premissa pode ser observada nas populações América Central e América do Sul, onde não há nenhuma mutação em regiões codificantes entre os 10 menores valores de $H_E$. Entretanto, para as duas substituições não-sinônimas em que pode haver alteração significativa na estrutura protéica, não parece haver seleção purificadora atuando globalmente, pois a heterozigosidade nos demais grupos é maior. Uma possível explicação consiste na análise das funções dos genes envolvidos. *MTRR* está relacionado à síntese de aminoácidos e, por isso, poderia estar associado a adaptações à dieta, como, por exemplo, devido à suplementação alimentar de metionina. *TEP1* atua na preservação do telômero e, por isso, poderia não ter grande impacto no *fitness* do indivíduo, caso levasse à morbidade e mortalidade tardias. Outra explicação poderia residir no fato de *TEP1* estar sob efeito carona ou de *allele surfing* e a pressão seletiva negativa não

ser forte o suficiente para impedir o aumento da variabilidade fora da África.

Entre os SNPs em regiões codificantes com maiores valores de Heterozigosidade Esperada, oito são mutações sinônimas, cinco são alterações não sinônimas benignas e apenas uma é possivelmente danosa, ERCC5-02 (rs17655) no grupo África Oriental. Esse gene está envolvido no reparo do DNA por excisão de nucleotídeos (UniProtKB, 2010). Associações a esse polimorfismo têm sido relatadas em vários tipos de câncer: pulmão em afro-americanos residentes na baía de São Francisco (EUA) (Chang *et al*., 2008); orofaringe, laringe e esôfago em residentes de Los Angeles (EUA) (Cui *et al*., 2006), de mama em técnicas de radiologia nos EUA (Rajaraman *et al*., 2008) e estomacal em chineses (Hussain *et al*., 2009). Usualmente, genes ligados ao reparo de DNA costumam estar sob seleção purificadora (Nielsen *et al*., 2009) e o valor de $H_E$ encontrado é intrigante.

Dos quatro SNPs com maiores valores de $H_E$ em mais de uma população, dois se situam na região promotora (CYP24A1-01 e CYP2E1-31) e dois em regiões não-traduzidas de éxons (GATA3-46 e IFNGR2-03), tal observação pode estar relacionada à importância dessas áreas regulatórias na evolução e adaptação a novas pressões seletivas (Pickering; Willis, 2004; Haygood *et al*., 2007). Isso pode ser explicado pelo fato que alterações não-sinônimas são, em geral, mais propensas à morbidade e letalidade devido às mudanças radicais causadas na estrutura protéica e tal ocorrência pode levar à perda da função original. Entretanto, mutações em regiões regulatórias, em geral, não levam à perda de função e resultam em alterações das taxas de tradução. Em geral, genes relacionados à interação com o ambiente são altamente polimórficos, o que pode ser observado no presente estudo. Os dois polimorfismos em promotores estão em genes relacionados à resposta a xenobióticos, sendo que *CYP24A1* atua na homeostase de cálcio e *CYP2E1* no metabolismo de substâncias exógenas e bioativação de alguns substratos hepatotóxicos e carcinogênicos (UniProtKB, 2010). Coincidentemente, os dois polimorfismos em regiões não traduzidas estão envolvidos na resposta imune, por ativação da transcrição de potenciadores de células T, *GATA3*, e receptores de interferon gama, *IFNGR2* (UniProtKB, 2010).

### 5.3 Implicações biomédicas em estudos de associação caso-controle

Vários artigos sobre estudos de associação caso-controle foram analisados quanto à possibilidade de resultados espúrios devidos à estratificação populacional. Entretanto o levantamento dos artigos apresenta algumas limitações. A primeira se deve à nomenclatura dos loci, muitos estudos, em especial os mais antigos, utilizam nomenclaturas alternativas baseadas na localização ou substituição de aminoácidos e não no código dbSNP. Outra restrição se deve ao acesso aos artigos, apenas estudos publicados em revistas indexadas pelo Pubmed e com acesso liberado pela CAPES foram analisados. Além disso, há o viés de publicação, pois a busca no Pubmed tende a retornar apenas estudos em que a associação entre os loci e os variados fenótipos é significativa. Entretanto, esse viés é positivo, pois permite analisar justamente os estudos sob risco de resultados falso-positivos. Como não é objetivo dessa dissertação analisar todos os estudos de associação e sim, demonstrar possíveis erros devidos à estratificação populacional, as duas primeiras limitações tornam-se desimportantes. Também devido a isso, apenas os 15 loci com maiores diferenças nas freqüências alélicas tiveram o levantamento bibliográfico realizado, independentemente da disponibilidade ou não de artigos para todos eles.

Não compete a esse trabalho avaliar minuciosamente as metodologias utilizadas pelos autores, o principal intuito da discussão a seguir é relatar possíveis erros de associação devidos à estruturação populacional, ponderar sobre as amostragens realizadas e em alguns casos recomendar atenção aos resultados obtidos. Preferencialmente, a discussão deveria se concentrar em estudos conduzidos em populações latino-americanas. Porém, estudos de associação caso-controle são bem menos comuns nessas populações que nos EUA e Europa, seja pela menor quantidade de recursos investidos em pesquisa, seja devido à miscigenação. Devido a tais restrições, também foram selecionados trabalhos realizados em quaisquer populações com índices não desprezíveis (superiores a 5%) de imigrantes africanos ou latino-americanos. Isso porque a estratificação nessas populações poderia mimetizar o que seria encontrado em algumas populações latino-americanas.

*5.3.1 Configuração EUR-NAT-WAFR*

O polimorfismo CASP3-08 (rs1049216) teve a maior divergência entre as populações ameríndias, européias e africanas. Após o levantamento bibliográfico acerca dos estudos de associação que utilizaram esse SNP, encontramos apenas um estudo acerca da susceptibilidade à mielomas múltiplos (Hosgood *et al*., 2009). O SNP foi associado ao menor risco de se desenvolver a doença e apenas mulheres foram amostradas nesse estudo. Porém, não há metodologia clara sobre a correção para ancestralidade e esta foi ajustada apenas para as populações caucasianas e africanas. A análise das freqüências alélicas indica que a maior diferenciação se dá entre as populações nativo-americanas e as demais. Devido ao padrão de miscigenação nas populações americanas indicar grande contribuição de mulheres ameríndias e africanas, há possibilidade de que o estudo esteja enviesado. Além disso, há moderada diferenciação dentro dos grupos populacionais ($F_{SC}$ = 0,0783).

Para o SNP POLB-05 (rs313617) há um estudo de associação, relacionando os heterozigotos ao risco aumentado para câncer de bexiga (Figueroa *et al*., 2007). O estudo foi realizado na Espanha e os casos e controles foram selecionados de acordo com a ancestralidade européia, mas não há indicações de qual foi o critério utilizado para acessar a ancestralidade dos indivíduos. Entretanto, há complicadores nesse estudo, o primeiro é que a maior diferença no AMOVA se dá entre os Africanos e os demais grupos, o segundo fator é que a Espanha foi invadida e ocupada a partir do século VIII por populações muçulmanas do Norte da África (Varela *et al*., 2008). Desse modo, eventos de miscigenação entre as populações Ibéricas e Norte-Africanas ocorreram, mas podem não ser mais reconhecidos pelos atuais descendentes, o que poderia reduzir o valor da auto-identificação. É interessante notar que o próprio autor, através do FDR (*False Discovery Rate test*), reconhece que dentre as associações observadas, algumas podem ser falsos positivos (Figueroa *et al*., 2007).

Oito estudos de associação foram realizados para o SNP CYP1A1-14 (rs2606345). Os estudos foram analisados acerca da possibilidade de falsa associação devida à estruturação populacional. As freqüências alélicas para esse locus são divergentes entre os europeus e as demais populações (nativo-americanos e africanos), porém são baixas as variações dentro das três populações.

Rotunno e colaboradores (2009) encontraram associação entre rs2606345 e o câncer de pulmão. Não fumantes homozigotos para o alelo mais comum (T) apresentam menor risco de desenvolver a doença, enquanto os fumantes com genótipos (TG ou GG) apresentam risco exponencialmente aumentado de acordo com o número de cigarros utilizados por dia, indicando resposta dependente de dosagem. Os sujeitos de pesquisa foram amostrados na Lombardia, Itália, e são preditos como caucasianos, a localização geográfica dos indivíduos foi adicionada ao teste como co-variável (Rotunno *et al*., 2009).

Figueroa e colaboradores (2008) testaram a associação entre esse SNP e cânceres testiculares dos tipos seminoma e não-seminoma. Foram genotipados indivíduos do Exército Norte-Americano, tanto negros quanto brancos (terminologia do autor). Não houve metodologia específica para testar a estruturação da amostra, sendo a única medida adotada, a adição das informações sobre etnicidade à regressão. E o estudo propõe que rs2606345 tem sugestiva associação com o câncer tipo não-seminoma em indivíduos heterozigotos ou homozigotos (TT) (Figueroa *et al*., 2008). A sugestiva associação desse alelo pode configurar um falso-positivo, uma vez que há grande diferença nas freqüências alélicas de afro-descendentes e eurodescentes.

Um estudo multi-étnico foi desenvolvido para acessar o papel de *CYP1A1* (e suas variantes) na ativação ou detoxificação de hidrocarbonetos aromáticos policíclicos (PAHs). Três amostragens foram realizadas, duas nos EUA e uma na Polônia, com os seguintes grupos étnicos: Afro-americanos e Dominicanos, nos EUA e Caucasianos, na Polônia. A ancestralidade foi obtida através de auto-identificação. A formação de adutos – ligação química sem modificação de estrutura – entre PAHs e o DNA foi oberservada em recém-nascidos caucasianos, porém a associação entre rs2606345 e adutos DNA-PAHs não foi confirmada após a correção de significância para comparações múltiplas (Wang *et al*., 2008). Nesse estudo multi-étnico é importante salientar a importância do controle de ancestralidade, isso porque, em populações miscigenadas, partes diferentes do genoma podem ter origens distintas.

A Coorte Norte-Americana “Estudo da Saúde Feminina ao Longo da Nação” (SWAN – *Study of Women’s Health Across The Nation*) foi citada em cinco publicações sobre associação de fenótipos com CYP1A1-14, quatros estudos originais e uma revisão de Sowers e colaboradores sobre os estudos de associação realizados nessa coorte até 2006 (Sowers *et al*., 2006b). Cinco etnias ou nacionalidades estão incluídas nessa coorte: Caucasianas, Hispânicas, Afro-americanas, Japonesas e Chinesas, todas residentes nos EUA. O estudo de Sowers e colaboradores (2006a) buscou relacionar o SNP rs2606345 a diferentes níveis de estradiol, estrogênio e seus metabólitos em quatro diferentes etnias/nacionalidades: Caucasianas, Afro-americanas, Chinesas e Japonesas. O genótipo GG (CC no estudo original) foi associado a baixos níveis de estradiol em mulheres japonesas, a altos níveis de 2-hidroxiestrona em mulheres chinesas e 16-hidroxiestrona em afro-americanas. A pesquisa publicada por Crandall e colegas (2006) utilizou todas as etnias da coorte, à exceção da Hispânica e relacionou o genótipo TG (AC no original) a episódios menos freqüentes de sintomas vasomotores em mulheres chinesas. A etnicidade foi utilizada como covariável (Crandall *et al*., 2006). Kravitz e colaboradores (2006) associaram os genótipos GG e TG (CC e AC no original) à probabilidade duas vezes maior de desenvolver depressão, para mulheres afro-americanas de genótipo GG, o risco é dez vezes maior. Outra pesquisa publicada por Crandall e colegas demonstrou associação entre o locus rs2606345 e maior densidade mamográfica para participantes com Índice de Massa Corporal superior a 30 kg/m$^2$ (Crandall *et al*., 2009).

É interessante notar que nenhum dos estudos envolvendo rs2606345 utilizou qualquer forma de controle genético (ex: MIAs, ACPs) nos casos e controles. O estudo de Rotunno, provavelmente, tem menor risco de estratificação que os demais. Isso devido à homogeneidade dentro dos grupos populacionais, ou seja, o baixo $F_{SC}$ indica que o risco de estruturação para populações não miscigenadas é mínimo. Provavelmente isso também é verdadeiro para a amostragem polonesa de Wang. As averiguações realizadas pelos demais autores envolveram populações norte-americanas e por isso, há possibilidade de que os indivíduos sejam miscigenados e possam levar a resultados espúrios. Aparentemente o estudo mais sujeito à estratificação é o de Figueroa, onde etnias diferentes foram analisadas conjuntamente. Os estudos realizados na coorte americana (SWAN) indicam que as freqüências alélicas também são diferentes entre europeus e populações orientais.

Esse fato é interessante, pois pode indicar a atuação de algum evento evolutivo singular à população européia.

*5.3.2 Configurações NAT-WAFR, EUR-NAT e EUR-WAFR*

Dos SNPs com maior diferenciação entre Nativo-Americanos e Africanos, dois foram anteriormente analisados na configuração EUR-NAT-WAFR, POLB-05 e WDR79-11. Apenas mais dois artigos foram encontrados para os 15 maiores valores de $F_{CT}$, um para CYP2E1-02 e outro para IGF2-16.

Devaney e colegas (2007) realizaram um estudo multiétnico onde demonstraram que IGF2-16 (rs3213221) está significativamente associado em homens a indicadores de dano muscular devido a exercícios. Homens e mulheres amostrados pertenciam aos seguintes grupos étnicos: 73% caucasianos, 13% asiáticos, 7% se classificaram como outras raças (definição do autor), 4% hispânicos e 3% afro-americanos. A baixa amostragem de indivíduos de ancestralidade ameríndia e africana pode não ter contribuído para uma falsa associação, entretanto se as proporções de hispânicos e afro-americanos fossem maiores e o dano muscular fosse mais freqüente em desses grupos, possivelmente o resultado não teria validade. A variabilidade dentro dos grupos populacionais NAT e WAFR para esse SNP é nula ($F_{SC}$ = -0,0081).

Quatro estudos foram discutidos sobre a possibilidade de falsos positivos em amostragens estratificadas de populações Européias e Nativo-Americanas. Haiman e colegas (2008) testaram em um estudo multiétnico a associação entre FANCA-03 (rs1061646) e o desenvolvimento de câncer de mama. Cinco grupos étnicos foram analisados: Afro-americanos, nativos havaianos, nipo-americanos, latinos e euro-americanos, adicionalmente foram analisadas populações caucasianas e asiáticas em estudos de replicata. O SNP rs1061646 foi associado ao risco aumentado para câncer de mama no estudo inicial, e em replicatas envolvendo populações asiáticas (japoneses, chineses e filipinos) e caucasianas. Os autores argumentam que a estratégia utilizada – baseada em tagSNPs para capturar a variabilidade populacional, mesmo com a inclusão de várias etnias, tem maior poder estatístico que análises baseadas em populações únicas, diminuem a possibilidade de erro do tipo I

e ainda incrementam a possibilidade de identificar variantes associadas aos fenótipos em mais de uma população. A etnicidade é estimada através de Componentes Principais (CP) e os resultados são agregados à regressão. A prevalência do câncer de mama é maior em mulheres caucasianas e a mortalidade é maior em mulheres afro-descendentes, porém fatores subjacentes à genética podem estar envolvidos, como exposição a carcinogênicos e acesso à saúde. Os maiores riscos de associação espúria, sob a ótica dessa configuração do AMOVA, seriam para populações latino-americanas, o que não foi confirmado para esse SNP. Além disso, a variação intrapopulacional é muito baixa ($F_{SC}$ = 0,0205), diminuindo a chance de sub-estruturação populacional

Colomer e colaboradores associaram o SNP CYP19A1-08 (rs4646) à eficácia do tratamento baseado em letrazol para pacientes portadoras de carcinoma de mama em estágio avançado. O autor não informou a composição étnica dos sujeitos de pesquisa e não há qualquer controle populacional (Colomer *et al*., 2008). O autor admite que os resultados possam ser falso-positivos, devido ao tamanho amostral ou estratificação populacional. Porém, sem a etnicidade dos participantes, não há como inferir se a estruturação entre europeus e nativo-americanos pode contribuir para tal, embora uma precaução adicional seja importante, uma vez que o valor de divergência dentro dos grupos populacionais não é desprezível ($F_{SC}$ = 0,0603).

Sadeghnejad e colegas desenvolveram um estudo em indivíduos brancos (terminologia do autor) buscando relacionar polimorfismos em IL13-01 e o fumo às funções pulmonares, mas não encontraram associação (Sadeghnejad *et al*., 2007). A única medida de controle para estratificação populacional foi a auto-identificação. O estudo mais interessante tendo em vista a quarta configuração foi o conduzido por Wang e colaboradores em indivíduos jamaicanos. Os pesquisadores evidenciaram efeito protetivo da variante A de rs20541 em relação ao Linfoma Não Hodgkin (Wang *et al*., 2009). Os autores desse estudo utilizaram indivíduos negros e outros – outras raças – (terminologias dos autores) na mesma amostragem, ajustando a regressão também para raça. Segundo Wang, essa metodologia foi aplicada devido ao fato de que as análises estratificadas por raça não eram significativamente diferentes.

Os estudos de Sadeghnejad e Wang não se adéquam perfeitamente à

configuração proposta, entretanto, os dois estudos apresentam possibilidade, ainda que baixa, de estruturação populacional. O termo “branco” se caracteriza mais por uma designação fenotípica do que étnica, e indivíduos brancos podem ter diferentes origens e graus de miscigenação, mesmo nos EUA (*United States Census Bureau*, 2008). Por isso, informações étnicas são mais confiáveis que designações raciais em estudos de associação. O estudo de Wang não apresenta alto risco de resultados espúrios devido à estruturação entre ameríndios e europeus. Isso porque a maior parte dos indivíduos possui ancestralidade africana, sendo baixos os níveis de ancestralidade européia e ameríndia (Simms *et al*., 2009). Entretanto, quaisquer estudos realizados em populações sabidamente miscigenadas, devem utilizar métodos rígidos de controle visando evitar a estratificação.

## 5.4 Implicações Evolutivas: Demografia e Seleção

A habilidade em identificar assinaturas moleculares da seleção natural é de considerável ajuda na identificação de loci que contribuem para a adaptação (Pickrell *et al*., 2009). Assim, um dos métodos comumente utilizado para verificar a atuação da seleção positiva é a análise de diferenciação populacional através da seleção de valores extremos de $F_{ST}$. Entretanto, o fenômeno conhecido por *Allele Surfing* produz assinaturas genômicas similares às da seleção positiva (Hofer *et al*., 2009).

### *5.4.1 Valores extremos de diferenciação populacional*

Dois genes têm dois representantes entre os dez polimorfismos com maiores valores de $F_{ST}$, *FANCA* e *IL4*. Consistentemente, o gene Interleukin-4 tem sido constantemente indicado como sob seleção positiva (Rockman *et al*., 2003; Wang *et al*., 2003; Tang *et al*., 2007; Akey *et al*., 2002) e está associado à resposta imune através da ativação de Células-B (Gene Ontology, 2009). Wang e outros (2003) demonstraram que esse gene é um dos que evoluem mais rapidamente em relação aos primatas do velho mundo. O gene *FANCA* atua no reparo a danos no DNA e essa categoria de processo biológico geralmente está sob seleção purificadora. Entretanto, as duas mutações com altos valores de $F_{ST}$ localizam-se em íntrons, região onde a seleção purificadora é mais relaxada.

Entre os demais representantes, apenas a mutação LEPR-01 é uma substituição não sinônima, porém a alteração é predita como benigna pela ferramenta PolyPhen. O gene *LEPR* participa dos processos metabólicos de reserva de energia, em especial, de lipídeos (Gene Ontology, 2009). Possivelmente, genes relacionados ao metabolismo de lipídeos estão relacionados à adaptação à novas dietas.

CYP1A1-19 está localizado na região promotora, região genômica associada a uma das maiores taxas de evolução em regiões funcionais (Taylor *et al*., 2008). O gene *CYP1A1* está relacionado ao metabolismo de drogas e vitamina D (Gene Ontology, 2009).

CASP3-08 localiza-se em região 3’ não traduzida e substituições nessa região podem alterar a estabilidade do mRNA e alterar sítios de ligação de miRNAs. O gene participa do processo de indução da apoptose celular (Gene Ontolgy, 2009).

Os demais polimorfismos se encontram em íntrons dos genes: *CASR*, associado à percepção de concentrações extracelulares de íons de cálcio; *RAD51* que atua no reparo a danos no DNA e *HSD17B2* que integra a via de biossíntese de esteróides.

Dos oito genes entre os maiores valores de diferenciação populacional, quatro podem estar relacionados a adaptações à dieta, dois à resposta a patógenos e um ao reparo de DNA. Respostas adaptativas à dieta (Kelley; Swanson, 2008; Haygood *et al*., 2007), xenobióticos e patógenos (Vallender; Lahn, 2004) não são incomuns, porém a diferenciação de dois SNPs em um gene relacionado ao reparo de DNA pode indicar um efeito carona proveniente de genes próximos a *FANCA*, já que algumas variantes desse gene estão associados à Anemia Fanconi (UniProtKB, 2010).

*5.4.2 Valores Extremos de Diferenciação Regional*

A baixa proporção de SNPs que levam à alteração de aminoácidos (4 mutações) e em regiões promotoras (3 mutações) em relação ao total, quando comparada às configurações populacionais anteriores, demonstra, possivelmente, a menor influência da seleção natural positiva na distinção entre as populações. Tal inferência parte do pressuposto que as regiões codificantes e regulatórias estão menos propensas à ação da deriva gênica, sendo alvo preferencial da seleção purificadora (Hughes *et al*., 2003).

A consistência dos dados entre as diferentes populações e conjuntos populacionais foi observada através da seleção de polimorfismos presentes em ao menos três e através da omissão dos polimorfismos em favor dos genes. Dessa forma, seis genes experimentam alta divergência entre as freqüências alélicas em asiáticos do leste e ameríndios. Desses seis genes, nenhum parece ainda ter sido evidenciado como sob seleção.

*IGF2*, o gene com valores de $F_{CT}$ mais altos e maior número de *hits* na comparação entre as populações do Leste Asiático está associado à diabetes, cujo número de casos em populações americanas é bem próximo ao encontrado no Leste da Ásia, a despeito desta região representar mais de 20% da população mundial (número de casos na América em 2000: 33 milhões; número de casos no Leste Asiático em 2000: 35 milhões) (Organização Mundial da Saúde, 2010). Dos cinco genes restantes, dois estão envolvidos na resposta imune a vírus, *CD86* e *STAT1*; e os demais na regulação da transcrição, *GATA3*; reparo de DNA, *BRIP1*; e percepção das concentrações extracelulares de cálcio, *CASR* (Gene Ontology, 2009). Esse último gene também esteve presente entre os maiores valores de diferenciação populacional baseados em $F_{ST}$, evidenciando que possivelmente a seleção positiva estaria atuando em um dos dois grupos populacionais, EAS ou NAT.

Dois genes entre os maiores valores de $F_{CT}$, mas para menos de três populações têm sido evidenciados como sob seleção positiva: *IL1B* (Tang *et al*., 2007) e *CD40* (Kelley *et al*., 2006), ambos envolvidos na resposta imune.

A análise de variância molecular realizada entre os grupos Leste Asiático e Nativo-Americanos, demonstra que, como esperado, eles estão mais próximos um do outro do que a média global ($F_{CT}$ = 8,93 versus $F_{CT}$ = 10,27). Além disso, apenas 36 polimorfismos têm valores de diferenciação entre grupos superiores à 0,25. Desses, apenas cinco têm valores superiores à 0,50.

Para o conjunto populacional Nordeste da Ásia (NEAS) – Nativos Americanos, a similaridade entre os dois grupos é ainda maior ($F_{CT}$ = 6,93) e apenas 25 polimorfismos com diferenciação entre grupos. Surpreendentemente, há mais polimorfismos com alta divergência ($F_{CT}$ > 0,5) entre NAT-NEAS (8 SNPs) do que entre EAS-NAT (5 SNPs).

A comparação entre as quatro subpopulações mais próximas ao Estreito de Bering e o grupo populacional Nativo-americanos demonstrou que a subpopulação Yakut é a que tem a menor distância genética ($F_{CT}$ = 1,08) em relação às populações americanas, seguida por Oroqen ($F_{CT}$ = 1,56), Hezhen ($F_{CT}$ = 1,77) e Daur ($F_{CT}$ = 2,83). Entretanto, ainda é incerta a participação das atuais populações siberianas na formação do *pool* gênico das populações nativo-americanas da família lingüística Ameríndia (Mulligan *et al*., 2004; Zhang *et al*., 2007). Segundo a própria tradição oral Yakut, durante o processo de migração em direção ao atual hábitat, a região de Yakutia, diversos povos foram assimilados (Vitebsky, 1990). Dessa forma, a proximidade genética entre os atuais Yakuts e as populações nativo-americanas poderia ser possivelmente explicada pela miscigenação entre antigos povos siberianos e as ancestrais das atuais populações da região siberiana. Conseqüentemente, o fluxo gênico entre essas populações elevaria a proximidade genética entre os dois grupos e, por extensão, entre Yakuts e Nativo-Americanos, caso as populações assimiladas tenham realmente contribuído para a colonização da América.

## 6 CONCLUSÃO

A utilização de variadas estatísticas genéticas como Equilíbrio de Hardy-Weinberg, Heterozigosidade Esperada e cálculos de distância genética, como $F_{ST}$ e seus análogos pode contribuir bastante para o desenho de estudos de associação caso-controle realizados em populações estruturadas geneticamente.

A partir do presente estudo foi possível identificar polimorfismos genéticos que podem: estar sob seleção natural positiva; levar à resultados falso-positivos em estudos de associação caso-controle; ou ambos. Os SNPs selecionados a partir dos cálculos de AMOVA podem ainda serem utilizados como MIAs.

Inferências mais fortes sobre seleção natural apenas poderiam ser feitas sob um contexto mais amplo, envolvendo todo um gene ou região e não apenas um único polimorfismo. Ainda assim, provavelmente, metodologias e testes estatísticos mais robustos seriam necessários para confirmar a atuação da seleção natural e diferenciar os padrões observados daqueles observados em eventos demográficos.

**Referências Bibliográficas:**

ACKER BA, BOTMA GJ, ZWINDERMAN AH, *et al*. **High HDL cholesterol does not protect against coronary artery disease when associated with combined cholesteryl ester transfer protein and hepatic lipase gene variants**. *Atherosclerosis* 2008;200(1):161-7.

AKEY JM, ZHANG G, ZHANG K, JIN L, SHRIVER MD. **Interrogating a high-density SNP map for signatures of natural selection**. *Genome Res* 2002;12(12):1805-14.

ASHBURNER M, BALL CA, BLAKE JA, *et al*. **Gene ontology: tool for the unification of biology**. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9.

BALARESQUE PL, BALLEREAU SJ, JOBLING MA. **Challenges in human genetic diversity: demographic history and adaptation**. *Hum Mol Genet* 2007;16 Spec No. 2:R134-9.

BALDING DJ. **A tutorial on statistical methods for population association studies**. *Nat Rev Genet* 2006;7(10):781-91.

BARBUJANI G, MAGAGNI A, MINCH E, CAVALLI-SFORZA LL. **An apportionment of human DNA diversity**. *Proc Natl Acad Sci U S A* 1997;94(9):4516-9.

BASTOS-RODRIGUES L, PIMENTA JR, PENA SD. **The genetic structure of human populations studied through short insertion-deletion polymorphisms**. *Ann Hum Genet* 2006;70(Pt 5):658-65.

BOYKO AR, WILLIAMSON SH, INDAP AR, *et al*. **Assessing the evolutionary impact of amino acid mutations in the human genome**. *PLoS Genet* 2008;4(5):e1000083.

BROOKES AJ. **The essence of SNPs**. *Gene* 1999;234(2):177-86.

CANN HM. **Human genome diversity**. *C R Acad Sci III* 1998;321(6):443-6.

CANN HM, DE TOMA C, CAZES L, *et al*. **A human genome diversity cell line panel**. *Science* 2002;296(5566):261-2.

CARDON LR, BELL JI. **Association study designs for complex diseases**. *Nat Rev Genet* 2001;2(2):91-9.

CAVALLI-SFORZA LL. **The Human Genome Diversity Project: past, present and future**. *Nat Rev Genet* 2005;6(4):333-40.

CHANG JS, WRENSCH MR, HANSEN HM, *et al*. **Nucleotide excision repair genes and risk of lung cancer among San Francisco Bay Area Latinos and African Americans**. *Int J Cancer* 2008;123(9):2095-104.

CHEVITARESE, J. **Determinação da estrutura genética das populações humanas e inferência dos fatores evolutivos que contribuíram para sua formação**. 2009. 99f. Dissertação (Mestrado em Genética) – Universidade Federal de Minas Gerais, Belo Horizonte.

CHOUDHRY S, COYLE NE, TANG H, *et al*. **Population stratification confounds genetic association studies among Latinos**. *Hum Genet* 2006;118(5):652-64.

CLARK AG, HUBISZ MJ, BUSTAMANTE CD, WILLIAMSON SH, NIELSEN R. **Ascertainment bias in studies of human genome-wide polymorphism**. *Genome Res* 2005;15(11):1496-502.

COLOMER R, MONZO M, TUSQUETS I, *et al*. **A single-nucleotide polymorphism in the aromatase gene is associated with the efficacy of the aromatase inhibitor letrozole in advanced breast carcinoma**. *Clin Cancer Res* 2008;14(3):811-6.

CORELLA A, BERT F, PEREZ-PEREZ A, GENE M, TURBON D. **Mitochondrial DNA diversity of the Amerindian populations living in the Andean Piedmont of Bolivia: Chimane, Moseten, Aymara and Quechua**. *Ann Hum Biol* 2007;34(1):34-55.

CRANDALL CJ, CRAWFORD SL, GOLD EB. **Vasomotor symptom prevalence is associated with polymorphisms in sex steroid-metabolizing enzymes and receptors**. *Am J Med* 2006;119(9 Suppl 1):S52-60.

CRANDALL CJ, SEHL ME, CRAWFORD SL, *et al*. **Sex steroid metabolism polymorphisms and mammographic density in pre- and early perimenopausal women**. *Breast Cancer Res* 2009;11(4):R51.

CRAWFORD DC, AKEY DT, NICKERSON DA. **The patterns of natural variation in human genes**. *Annu Rev Genomics Hum Genet* 2005;6:287-312.

CUI Y, MORGENSTERN H, GREENLAND S, *et al*. **Polymorphism of Xeroderma Pigmentosum group G and the risk of lung cancer and squamous cell carcinomas of the oropharynx, larynx and esophagus**. *Int J Cancer* 2006;118(3):714-20.

DEVANEY JM, HOFFMAN EP, GORDISH-DRESSMAN H, KEARNS A, ZAMBRASKI E, CLARKSON PM. **IGF-II gene region polymorphisms related to exertional muscle damage**. *J Appl Physiol* 2007;102(5):1815-23.

DEVLIN B, ROEDER K, BACANU SA. **Unbiased methods for population-based association studies**. *Genet Epidemiol* 2001;21(4):273-84.

ERLICH HA, MACK SJ, BERGSTROM T, GYLLENSTEN UB. **HLA class II alleles in Amerindian populations: implications for the evolution of HLA polymorphism and the colonization of the Americas**. *Hereditas* 1997;127(1-2):19-24.

EXCOFFIER L, LAVAL G, SCHNEIDER S. **Arlequin (version 3.0): An integrated software package for population genetics data analysis**. *Evol Bioinform Online* 2005;1:47-50.

EXCOFFIER L, SMOUSE PE, QUATTRO JM. **Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data**. *Genetics* 1992;131(2):479-91.

FAN YM, RAITAKARI OT, KAHONEN M, *et al*. **Hepatic lipase promoter C-480T polymorphism is associated with serum lipids levels, but not subclinical atherosclerosis: the Cardiovascular Risk in Young Finns Study**. *Clin Genet* 2009;76(1):46-53.

FERRER-ADMETLLA A, BOSCH E, SIKORA M, *et al*. **Balancing selection is the main force shaping the evolution of innate immunity genes**. *J Immunol* 2008;181(2):1315-22.

FIGUEROA JD, MALATS N, REAL FX, *et al*. **Genetic variation in the base excision repair pathway and bladder cancer risk**. *Hum Genet* 2007;121(2):233-42.

FIGUEROA JD, SAKODA LC, GRAUBARD BI, *et al*. **Genetic variation in hormone metabolizing genes and risk of testicular germ cell tumors**. *Cancer Causes Control* 2008;19(9):917-29.

FRAZER KA, BALLINGER DG, COX DR, *et al*. **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007;449(7164):851-61.

GENE ONTOLOGY: **Gene Ontology website**. Disponível em: http://www.geneontology.org Acesso em: 10 dez 2009 a 04 jan 2010.

GLAUBITZ JC. **convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages**. Molecular Ecology Notes. 2004;4:309–310.

GOLDSTEIN DB, TATE SK, SISODIYA SM. **Pharmacogenetics goes genomic**. *Nat Rev Genet* 2003;4(12):937-47.

GONNET GH, COHEN MA, BENNER SA. **Exhaustive matching of the entire protein sequence database**. *Science* 1992;256(5062):1443-5.

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

GUO SW, THOMPSON EA. **Performing the exact test of Hardy-Weinberg proportion for multiple alleles**. *Biometrics* 1992;48(2):361-72.

HAIMAN CA, HSU C, DE BAKKER PI, *et al*. **Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations**. *Hum Mol Genet* 2008;17(6):825-34.

HAMREFORS V, ORHO-MELANDER M, KRAUSS RM, *et al*. **A gene score of nine LDL and HDL regulating genes is associated with fluvastatin induced cholesterol changes in women**. *J Lipid Res* 2009.

HAYGOOD R, FEDRIGO O, HANSON B, YOKOYAMA KD, WRAY GA. **Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution**. *Nat Genet* 2007;39(9):1140-4.

HEWETT M, OLIVER DE, RUBIN DL, *et al*. **PharmGKB: the Pharmacogenetics Knowledge Base**. *Nucleic Acids Res* 2002;30(1):163-5.

HOFER T, RAY N, WEGMANN D, EXCOFFIER L. **Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection**. *Ann Hum Genet* 2009;73(1):95-108.

HOLSINGER KE, WEIR BS. **Genetics in geographically structured populations: defining, estimating and interpreting F(ST)**. *Nat Rev Genet* 2009;10(9):639-50.

HOSGOOD HD, 3rd, Baris D, Zhang Y, et al. **Genetic variation in cell cycle and apoptosis related genes and multiple myeloma risk**. *Leuk Res* 2009;33(12):1609-14.

HUGHES AL. **Evolution in the post-genome era**. *Perspect Biol Med* 2009;52(2):332-7.

HUGHES AL, PACKER B, WELCH R, BERGEN AW, CHANOCK SJ, YEAGER M. **Widespread purifying selection at polymorphic sites in human protein-coding loci**. *Proc Natl Acad Sci U S A* 2003;100(26):15754-7.

HUGHES AL, WELCH R, PURI V, *et al*. **Genome-wide SNP typing reveals signatures of population history**. *Genomics* 2008;92(1):1-8.

HURST LD. **Fundamental concepts in genetics: genetics and the understanding of selection**. *Nat Rev Genet* 2009;10(2):83-93.

HUSSAIN SK, MU LN, CAI L, *et al*. **Genetic variation in immune regulation and DNA repair pathways and stomach cancer in China**. *Cancer Epidemiol Biomarkers Prev* 2009;18(8):2304-9.

INTERNATIONAL HAPMAP CONSORTIUM, Nature. 2003. 426, 789-796

INTERNATIONAL HAPMAP CONSORTIUM, Nature. 2005. 437, 1299-1320

KELLEY JL, MADEOY J, CALHOUN JC, SWANSON W, AKEY JM. **Genomic signatures of positive selection in humans and the limits of outlier approaches**. *Genome Res* 2006;16(8):980-9.

KELLEY JL, SWANSON WJ. **Dietary change and adaptive evolution of enamelin in humans and among primates**. *Genetics* 2008;178(3):1595-603.
KIM Y, STEPHAN W. **Detecting a local signature of genetic hitchhiking along a recombining chromosome**. *Genetics* 2002;160(2):765-77.

KNOWLER WC, WILLIAMS RC, PETTITT DJ, STEINBERG AG. **Gm$^{3;5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture**. *Am J Hum Genet* 1988;43(4):520-6.

LACHANCE J. **Detecting selection-induced departures from Hardy-Weinberg proportions**. *Genet Sel Evol* 2009;41:15.

LEWIS PO, ZAYKIN D. **GDA (Genetic Data Analysis): Computer program for the analysis of allelic data. Version 1.0 d16c ed**. Storrs University of Connecticut. 2001

LEWONTIN RC. **The apportionment of human diversity**. *Evol. Biol.* 1972;6:381-98

LI Y, GRAUBARD BI. **Testing Hardy-Weinberg equilibrium and homogeneity of Hardy-Weinberg disequilibrium using complex survey data**. *Biometrics* 2009;65(4):1096-104.

LI JZ, ABSHER DM, TANG H, *et al*. **Worldwide human relationships inferred from genome-wide patterns of variation**. *Science* 2008;319(5866):1100-4.

LONG JC, WILLIAMS RC, MCAULEY JE, *et al*. **Genetic variation in Arizona**

**Mexican Americans: estimation and interpretation of admixture proportions**. *Am J Phys Anthropol* 1991;84(2):141-57.

LU Y, DOLLE ME, IMHOLZ S, *et al*. **Multiple genetic variants along candidate pathways influence plasma high-density lipoprotein cholesterol concentrations**. *J Lipid Res* 2008;49(12):2582-9.

MOORE LE, HUNG R, KARAMI S, *et al*. **Folate metabolism genes, vegetable intake and renal cancer risk in central Europe**. *Int J Cancer* 2008;122(8):1710-5.

MULLIGAN CJ, HUNLEY K, COLE S, LONG JC. **Population genetics, history, and health patterns in native americans**. *Annu Rev Genomics Hum Genet* 2004;5:295-315.

MYLES S, TANG K, SOMEL M, GREEN RE, KELSO J, STONEKING M. **Identification and analysis of genomic regions with large between-population differentiation in humans**. *Ann Hum Genet* 2008;72(Pt 1):99-110.

NEI M, ROYCHOUDHURY AK. **Sampling variances of heterozygosity and genetic distance**. *Genetics* 1974;76(2):379-90.

NELSON MR, BRYC K, KING KS, *et al*. **The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research**. *Am J Hum Genet* 2008;83(3):347-58.

NIELSEN R, HUBISZ MJ, HELLMANN I, et al. **Darwinian and demographic forces affecting human protein coding genes**. *Genome Res* 2009;19(5):838-49

NOVEMBRE J, DI RIENZO A. **Spatial patterns of variation due to natural selection in humans**. *Nat Rev Genet* 2009;10(11):745-55.

O'ROURKE DH, MOBARRY A, SUAREZ BK. **Patterns of genetic variation in Native America**. *Hum Biol* 1992;64(3):417-34.

PACKER BR, YEAGER M, STAATS B, *et al*. **SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes**. *Nucleic Acids Res* 2004;32(Database issue):D528-32.

PACKER BR, YEAGER M, BURDETT L, *et al*. **SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes**. *Nucleic Acids Res* 2006;34(Database issue):D617-21.

PARHAM P, OHTA T. **Population biology of antigen presentation by MHC class I**

**molecules**. *Science* 1996;272(5258):67-74.

PICKERING BM, WILLIS AE. **The implications of structured 5' untranslated regions on translation and disease**. *Semin Cell Dev Biol* 2005;16(1):39-47.

POLYPHEN: prediction of functional effect of human nsSNPs. Disponível em: <http://genetics.bwh.harvard.edu/pph> Acesso em: 29 dez 2009 a 04 jan 2010.

PRICE AL, PATTERSON NJ, PLENGE RM, WEINBLATT ME, SHADICK NA, REICH D. **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006;38(8):904-9

PRITCHARD JK, ROSENBERG NA. **Use of unlinked genetic markers to detect population stratification in association studies**. *Am J Hum Genet* 1999;65(1):220-8.

PRITCHARD JK, STEPHENS M, DONNELLY P. **Inference of population structure using multilocus genotype data**. *Genetics* 2000;155(2):945-59.

PRITCHARD JK, STEPHENS M, ROSENBERG NA, DONNELLY P. **Association mapping in structured populations**. *Am J Hum Genet* 2000;67(1):170-81.

PRITCHARD JK, DONNELLY P. **Case-control studies of association in structured or admixed populations**. *Theor Popul Biol* 2001;60(3):227-37.

RAJARAMAN P, BHATTI P, DOODY MM, *et al*. **Nucleotide excision repair polymorphisms may modify ionizing radiation-related breast cancer risk in US radiologic technologists**. *Int J Cancer* 2008;123(11):2713-6

RAMACHANDRAN S, DESHPANDE O, ROSEMAN CC, ROSENBERG NA, FELDMAN MW, CAVALLI-SFORZA LL. **Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa**. *Proc Natl Acad Sci U S A* 2005;102(44):15942-7.

RAYMOND M, ROUSSET F. **genepop version 1.2.: population genetics software for exact tests and ecumenicism**. *Journal of Heredity*. 1995;86:248–249.

ROBERTSON A, HILL WG. **Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients**. *Genetics* 1984;107(4):703-18.

ROCKMAN MV, HAHN MW, SORANZO N, GOLDSTEIN DB, WRAY GA. **Positive selection on a human-specific transcription factor binding site regulating IL4 expression**. *Curr Biol* 2003;13(23):2118-23.

RODRIGUEZ H, DERODRIGUEZ E, LORIA A, LISKER R. **Studies on Several Genetic Hematological Traits of the Mexican Population. V. Distribution of Blood Group Antigens in Nahuas, Yaquis, Tarahumaras, Tarascos and Mixtecos**. *Hum Biol* 1963;35:350-60.

ROEDER K, LUCA D. **Searching for disease susceptibility variants in structured populations**. *Genomics* 2009;93(1):1-4

ROGERS AR, JORDE LB. **Ascertainment bias in estimates of average heterozygosity**. *Am J Hum Genet* 1996;58(5):1033-41.

ROSENBERG NA, MAHAJAN S, RAMACHANDRAN S, ZHAO C, PRITCHARD JK, FELDMAN MW. **Clines, clusters, and the effect of study design on the inference of human population structure**. *PLoS Genet* 2005;1(6):e70.

ROSENBERG NA, PRITCHARD JK, WEBER JL, *et al*. **Genetic structure of human populations**. *Science* 2002;298(5602):2381-5.

ROTUNNO M, YU K, LUBIN JH, *et al.* **Phase I metabolic genes and risk of lung cancer: multiple polymorphisms and mRNA expression**. *PLoS One* 2009;4(5):e5652.

ROUSSET F, RAYMOND M. **Testing heterozygote excess and deficiency**. *Genetics* 1995;140(4):1413-9.

ROUSSET F. **genepop'007: a complete re-implementation of the genepop software for Windows and Linux**. Molecular Ecology Resources. 2008;8:103-106

SABETI PC, REICH DE, HIGGINS JM, *et al*. **Detecting recent positive selection in the human genome from haplotype structure**. *Nature* 2002;419(6909):832-7.

SADEGHNEJAD A, MEYERS DA, BOTTAI M, STERLING DA, BLEECKER ER, OHAR JA. **IL13 promoter polymorphism 1112C/T modulates the adverse effect of tobacco smoking on lung function**. *Am J Respir Crit Care Med* 2007;176(8):748-52.

SCHEURER ME, AMIRIAN E, CAO Y, et al. **Polymorphisms in the interleukin-4 receptor gene are associated with better survival in patients with glioblastoma**. *Clin Cancer Res* 2008;14(20):6640-6.

SCHNEIDER S, ROESSLI D, EXCOFFIER L. **Arlequin Ver. 2.0: A Software for Population Genetics Data Analysis**. Genetics and Biometry Laboratory, University of Geneva: Switzerland. (2000)

SEN S, BURMEISTER M. **Hardy-Weinberg analysis of a large set of published association studies reveals genotyping error and a deficit of heterozygotes across multiple loci**. *Hum Genomics* 2008;3(1):36-52.

SIMMS TM, RODRIGUEZ CE, RODRIGUEZ R, HERRERA RJ. **The genetic structure of populations from Haiti and Jamaica reflect divergent demographic histories**. *Am J Phys Anthropol* 2009.

SOWERS MR, SYMONS JP, JANNAUSCH ML, CHU J, KARDIA SR. **Sex steroid hormone polymorphisms, high-density lipoprotein cholesterol, and apolipoprotein A-1 from the Study of Women's Health Across the Nation (SWAN)**. *Am J Med* 2006;119(9 Suppl 1):S61-8.

SOWERS MR, WILSON AL, KARDIA SR, CHU J, MCCONNELL DS. **CYP1A1 and CYP1B1 polymorphisms and their association with estradiol and estrogen metabolites in women who are premenopausal and perimenopausal**. *Am J Med* 2006;119(9 Suppl 1):S44-51.

SCHORK NJ, FALLIN D, THIEL B, *et al*. **The future of genetic case-control studies**. *Adv Genet* 2001;42:191-212.

SELDIN MF. **Admixture mapping as a tool in gene discovery**. *Curr Opin Genet Dev* 2007;17(3):177-81.

SMITH MW, O'BRIEN SJ. **Mapping by admixture linkage disequilibrium: advances, limitations and guidelines**. *Nat Rev Genet* 2005;6(8):623-32.

TALBOTT KE, GAMMON MD, KIBRIYA MG, *et al*. **A CYP19 (aromatase) polymorphism is associated with increased premenopausal breast cancer risk**. *Breast Cancer Res Treat* 2008;111(3):481-7.

TANG K, THORNTON KR, STONEKING M. **A new approach for using genome scans to detect recent positive selection in the human genome**. *PLoS Biol* 2007;5(7):e171.

TAYLOR MS, MASSINGHAM T, HAYASHIZAKI Y, CARNINCI P, GOLDMAN N, SEMPLE CA. **Rapidly evolving human promoter regions**. *Nat Genet* 2008;40(11):1262-3; author reply 1263-4.

TERWILLIGER JD, WEISS KM. **Linkage disequilibrium mapping of complex disease: fantasy or reality?** *Curr Opin Biotechnol* 1998;9(6):578-94

TISHKOFF SA, REED FA, FRIEDLAENDER FR, *et al*. **The genetic structure and history of Africans and African Americans**. *Science* 2009;324(5930):1035-44.

THOMAS DC, WITTE JS. **Point: population stratification: a problem for case-control studies of candidate-gene associations?** *Cancer Epidemiol Biomarkers Prev* 2002;11(6):505-12.

UNIPROTKB: UniProtKB: Protein Knowledgebase. Disponível em: <http://www.uniprot.org/uniprot/> Acesso em: 29 dez 2009 a 04 jan 2010.

UNITED STATES. US Census Bureau. **American FactFinder Help**. Disponível em: <http://factfinder.census.gov/home/en/epss/glossary> Acesso em: 20 dez 2009.

VARELA TA, FARINA J, DIEGUEZ LP, LODEIRO R. **Gene flow and genetic structure in the Galician population (NW Spain) according to Alu insertions**. *BMC Genet* 2008;9:79.

VITEBSKY, P. Yakuts. In: Smith, G. (ed.) *The nationalities question in the Soviet Union*. London, New York: Longman. 1990:304-5

ZHANG F, SU B, ZHANG YP, JIN L. **Genetic studies of human diversity in East Asia**. *Philos Trans R Soc Lond B Biol Sci* 2007;362(1482):987-95.

WANG HY, TANG H, SHEN CK, WU CI. **Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites**. *Mol Biol Evol* 2003;20(11):1795-804.

WANG S, LEWIS CM, JAKOBSSON M, *et al*. **Genetic variation and population structure in native Americans**. *PLoS Genet* 2007;3(11):e185.

WANG S, RAY N, ROJAS W, *et al*. **Geographic patterns of genome admixture in Latin American Mestizos**. *PLoS Genet* 2008;4(3):e1000037.

WANG S, CHANOCK S, TANG D, LI Z, JEDRYCHOWSKI W, PERERA FP. **Assessment of interactions between PAH exposure and genetic polymorphisms on PAH-DNA adducts in African American, Dominican, and Caucasian mothers and newborns**. *Cancer Epidemiol Biomarkers Prev* 2008;17(2):405-13.

WANG SS, CARREON JD, HANCHARD B, CHANOCK S, HISADA M. **Common genetic variants and risk for non-Hodgkin lymphoma and adult T-cell lymphoma/leukemia in Jamaica**. *Int J Cancer* 2009;125(6):1479-82.

WHO: World Health Organization: **Diabetes Programme**, Disponível em: </http://www.who.int/diabetes/facts/world_figures/en/> Acesso em 04 jan 2010.

## ANEXO A – Número de SNPs por gene

| Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs |
|---|---|---|---|---|---|---|---|
| *GSK3B* | 37 | *HSD3B2* | 4 | *ABCB11* | 2 | *BIRC2* | 1 |
| *KRAS* | 22 | *ICAM1* | 4 | *ABCC4* | 2 | *BPI* | 1 |
| *GHR* | 21 | *IL13* | 4 | *AHRR* | 2 | *C11orf65* | 1 |
| *TNKS* | 20 | *IL15RA* | 4 | *ALOX15* | 2 | *CASP10* | 1 |
| *PMS1* | 19 | *IL1B* | 4 | *APOA4* | 2 | *CBR3* | 1 |
| *CYP19A1* | 18 | *LEPR* | 4 | *BAX* | 2 | *CCDC97* | 1 |
| *PGR* | 17 | *MBD2* | 4 | *BCR* | 2 | *CD14* | 1 |
| *MASP1* | 16 | *MET* | 4 | *BIRC3* | 2 | *CD4* | 1 |
| *AKR1C3* | 15 | *MGMT* | 4 | *CALCR* | 2 | *CDC25A* | 1 |
| *MSH2* | 15 | *MMP1* | 4 | *CCL5* | 2 | *CDC25B* | 1 |
| *CTNNB1* | 14 | *MTHFR* | 4 | *CCND3* | 2 | *CDC25C* | 1 |
| *INSR* | 12 | *MTR* | 4 | *CCNH* | 2 | *CDK4* | 1 |
| *RERG* | 12 | *NBN* | 4 | *CCR3* | 2 | *CDK7* | 1 |
| *FANCA* | 11 | *NFKBIE* | 4 | *CD40* | 2 | *CDKN1B* | 1 |
| *MBL2* | 11 | *OPRM1* | 4 | *CD81* | 2 | *CDKN1C* | 1 |
| *BIC* | 10 | *PIM1* | 4 | *CD86* | 2 | *CG018* | 1 |
| *CDKN2A* | 10 | *PLA2G6* | 4 | *CDH1* | 2 | *COASY* | 1 |
| *GATA3* | 10 | *RAD23B* | 4 | *CDK5* | 2 | *CSF2* | 1 |
| *GPX2* | 10 | *RB1CC1* | 4 | *CRP* | 2 | *CTSB* | 1 |
| *HSD17B4* | 10 | *SHBG* | 4 | *CSF3* | 2 | *CTSH* | 1 |
| *LIPC* | 10 | *SLC23A1* | 4 | *CX3CR1* | 2 | *CYP2D6* | 1 |
| *RAD51* | 10 | *SLC6A3* | 4 | *CYP2C19* | 2 | *CYP3A4* | 1 |
| *TP73L* | 10 | *TNF* | 4 | *CYP2E1* | 2 | *CYP3A7* | 1 |
| *ARNT* | 9 | *WDR79* | 4 | *DHDH* | 2 | *DNAJC18* | 1 |
| *ESR1* | 9 | *XBP1* | 4 | *DIO1* | 2 | *DRD1* | 1 |
| *IGF1* | 9 | *ABCB1* | 3 | *DRD2* | 2 | *ENG* | 1 |
| *MX1* | 9 | *ABCG8* | 3 | *DRD4* | 2 | *ENPP1* | 1 |
| *POT1* | 9 | *AHR* | 3 | *EDN1* | 2 | *EPHX2* | 1 |
| *ATM* | 8 | *ALAD* | 3 | *EFNB3* | 2 | *ERBB2* | 1 |
| *AXIN2* | 8 | *ALDH1L1* | 3 | *ERCC2* | 2 | *FASLG* | 1 |
| *BRCA1* | 8 | *APEX1* | 3 | *ERCC3* | 2 | *FOXA1* | 1 |
| *EPHX1* | 8 | *APOA2* | 3 | *ERCC4* | 2 | *FUT2* | 1 |
| *MYBL2* | 8 | *ARHGDIB* | 3 | *ERCC6* | 2 | *GC* | 1 |
| *SLC23A2* | 8 | *ATP1B2* | 3 | *ESR2* | 2 | *HAO2* | 1 |
| *ALOX5* | 7 | *BAK1* | 3 | *EXO1* | 2 | *HIF1AN* | 1 |
| *AMACR* | 7 | *BCL2L1* | 3 | *FLJ45983* | 2 | *HSPB8* | 1 |
| *AURKA* | 7 | *BHMT* | 3 | *GDF15* | 2 | *IFNG* | 1 |
| *BLM* | 7 | *CASP8* | 3 | *GGH* | 2 | *IFNGR2* | 1 |
| *BRCA2* | 7 | *CASP9* | 3 | *GPX1* | 2 | *IGFBP1* | 1 |
| *CASR* | 7 | *CBR1* | 3 | *GRPR* | 2 | *IGFBP3* | 1 |

| Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs |
|---|---|---|---|---|---|---|---|
| *CAV1* | 7 | *CBS* | 3 | *GSTP1* | 2 | *IL12A* | 1 |
| *CTLA4* | 7 | *CCNA2* | 3 | *GSTZ1* | 2 | *IL3* | 1 |
| *IGF1R* | 7 | *CCND1* | 3 | *HMGCR* | 2 | *IL6R* | 1 |
| *IL10* | 7 | *CCR2* | 3 | *HSD17B1* | 2 | *IL8RA* | 1 |
| *IL4R* | 7 | *CD80* | 3 | *HSD17B2* | 2 | *JTV1* | 1 |
| *LPL* | 7 | *CETP* | 3 | *HTR1B* | 2 | *KRT23* | 1 |
| *ROS1* | 7 | *CHEK1* | 3 | *HUS1* | 2 | *LEP* | 1 |
| *TERT* | 7 | *COL18A1* | 3 | *IFNGR1* | 2 | *LIG3* | 1 |
| *ABCA1* | 6 | *CSF1R* | 3 | *IGFALS* | 2 | *LIG4* | 1 |
| *BRIP1* | 6 | *DHFR* | 3 | *IGFBP5* | 2 | *LMOD1* | 1 |
| *CAT* | 6 | *EGF* | 3 | *IL10RA* | 2 | *LOC389143* | 1 |
| *CTH* | 6 | *EGFR* | 3 | *IL12B* | 2 | *LOC391073* | 1 |
| *CYP17A1* | 6 | *ERCC1* | 3 | *IL1A* | 2 | *LOC646837* | 1 |
| *CYP1B1* | 6 | *ERCC5* | 3 | *IL2* | 2 | *MAOA* | 1 |
| *FOXC1* | 6 | *FAM82A* | 3 | *IL6* | 2 | *MATR3* | 1 |
| *FZD7* | 6 | *FAS* | 3 | *IL7R* | 2 | *MBD4* | 1 |
| *GPX3* | 6 | *FOS* | 3 | *IRF1* | 2 | *MDM2* | 1 |
| *HSD3B1* | 6 | *GSTM3* | 3 | *LCAT* | 2 | *MEST* | 1 |
| *IGF2R* | 6 | *HADHA* | 3 | *LITAF* | 2 | *METTL1* | 1 |
| *MSH3* | 6 | *HFE* | 3 | *LRP6* | 2 | *MPDU1* | 1 |
| *MTRR* | 6 | *HTR1D* | 3 | *LTA* | 2 | *MPO* | 1 |
| *NFKB1* | 6 | *IFNAR2* | 3 | *MLH1* | 2 | *MTHFD2* | 1 |
| *PAK6* | 6 | *IGF2AS* | 3 | *MSH6* | 2 | *MYC* | 1 |
| *PARP1* | 6 | *IGFBP2* | 3 | *MSR1* | 2 | *MYNN* | 1 |
| *TEP1* | 6 | *IGFBP6* | 3 | *NINJ1* | 2 | *NICN1* | 1 |
| *TSG101* | 6 | *IL1RN* | 3 | *NOS2A* | 2 | *NPAT* | 1 |
| *APC* | 5 | *IL8* | 3 | *NOS3* | 2 | *NUBP2* | 1 |
| *APOB* | 5 | *IRF3* | 3 | *NR1H4* | 2 | *P2RX7* | 1 |
| *BARD1* | 5 | *IRS1* | 3 | *OGG1* | 2 | *PHB* | 1 |
| *BCL6* | 5 | *JAK3* | 3 | *OPRD1* | 2 | *PLA2G2A* | 1 |
| *CARD15* | 5 | *LMO2* | 3 | *PCTP* | 2 | *PLK1* | 1 |
| *CFH* | 5 | *LOC727797* | 3 | *PTEN* | 2 | *POLD1* | 1 |
| *CYBB* | 5 | *MYO5A* | 3 | *RAD52* | 2 | *PPP1R13L* | 1 |
| *CYP1A1* | 5 | *NCF2* | 3 | *RET* | 2 | *PTGS1* | 1 |
| *FBXW7* | 5 | *NCOA3* | 3 | *RGS17* | 2 | *RAC1* | 1 |
| *IGF2* | 5 | *NQO1* | 3 | *RNASEL* | 2 | *RAD54L* | 1 |
| *IL15* | 5 | *OCA2* | 3 | *RXRA* | 2 | *RAG1* | 1 |
| *IL4* | 5 | *PCNA* | 3 | *RXRB* | 2 | *RGS5* | 1 |
| *LDLR* | 5 | *PMS2* | 3 | *SAT2* | 2 | *RPA4* | 1 |
| *LIG1* | 5 | *POLB* | 3 | *SEP15* | 2 | *SELE* | 1 |

| Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs | Gene | N de SNPs |
|---|---|---|---|---|---|---|---|
| *LRP5* | 5 | *PPARG* | 3 | *SEPP1* | 2 | *SEPT2* | 1 |
| *PARP4* | 5 | *PTH* | 3 | *SFTPD* | 2 | *SLC2A1* | 1 |
| *PIN1* | 5 | *RAB15* | 3 | *SLC19A1* | 2 | *SLC2A4* | 1 |
| *PTGS2* | 5 | *RGS6* | 3 | *SOD2* | 2 | *SLC30A1* | 1 |
| *SCARB1* | 5 | *SCUBE2* | 3 | *SSTR3* | 2 | *SLC30A4* | 1 |
| *TERF1* | 5 | *SEC14L2* | 3 | *TCTA* | 2 | *SLC6A18* | 1 |
| *TP53* | 5 | *SLAMF1* | 3 | *TGM1* | 2 | *SOD1* | 1 |
| *TP53I3* | 5 | *SLC39A2* | 3 | *TNFRSF10A* | 2 | *SOD3* | 1 |
| *WRN* | 5 | *SLC4A2* | 3 | *TYR* | 2 | *STAT1* | 1 |
| *XRCC4* | 5 | *SOAT2* | 3 | *UCP3* | 2 | *STK11* | 1 |
| *XRCC5* | 5 | *SRA1* | 3 | *VDR* | 2 | *SULT1A2* | 1 |
| *ABCC2* | 4 | *TERF2* | 3 | *VIL2* | 2 | *TFF1* | 1 |
| *ADH1C* | 4 | *TGFBR1* | 3 | *XRCC3* | 2 | *TFF3* | 1 |
| *APAF1* | 4 | *TLR2* | 3 | *ABCA5* | 1 | *TFRC* | 1 |
| *AR* | 4 | *TSPO* | 3 | *AKR1A1* | 1 | *TGFB1* | 1 |
| *CASP3* | 4 | *TXNRD2* | 3 | *AKR1C4* | 1 | *TNFRSF1A* | 1 |
| *CGA* | 4 | *TYMS* | 3 | *AKT1* | 1 | *TNIP1* | 1 |
| *COMT* | 4 | *VCAM1* | 3 | *ALDH2* | 1 | *UGT1A* | 1 |
| *CSTF1* | 4 | *VEGF* | 3 | *ALOX12* | 1 | *XPA* | 1 |
| *CYP24A1* | 4 | *XPC* | 3 | *ANKK1* | 1 | *XRCC1* | 1 |
| *CYP7B1* | 4 | *ZNF350* | 3 | *APOE* | 1 | *ZFPM1* | 1 |
| *GPX4* | 4 | *ABCA6* | 2 | *ARVCF* | 1 | *ZNF230* | 1 |
| *GSTA4* | 4 | *ABCA7* | 2 | *B9D2* | 1 | | |

**ANEXO B – Distribuição das populações e subpopulações estudadas por: grupo populacional, região geográfica e grupo lingüístico.**

| Grupo Populacional | População | Localização | Grupo Linguístico[a] | N de Indivíduos |
|---|---|---|---|---|
| África Ocidental | Mandenka | Senegal | Nigero-Congolesa | 24 |
| | San | Namíbia | Coisã | 7 |
| | Iorubá | Nigéria | Nigero-Congolesa | 25 |
| África Oriental | Bantu NE. | Kênia | Nigero-Congolesa | 11 |
| | Bantu SE. Pedi | África do Sul | Nigero-Congolesa | 1 |
| | Bantu SE. Sotho | África do Sul | Nigero-Congolesa | 1 |
| | Bantu SE. Tswana | África do Sul | Nigero-Congolesa | 2 |
| | Bantu SE. Zulu | África do Sul | Nigero-Congolesa | 1 |
| | Bantu SO. Herero | África do Sul | Nigero-Congolesa | 2 |
| | Bantu SO. Ovambo | África do Sul | Nigero-Congolesa | 1 |
| | Pigmeus Biaka | África Central | Nigero-Congolesa | 31 |
| | Pigmeus Mbuti | República Dem. Congo | Nilo-Saariana | 13 |
| América Central | Maia | México | Maia | 25 |
| | Pima | México | Azteca-Tanoano | 24 |
| América do Sul | Cayapa | Equador | Barbacoano | 7 |
| | Karitiana | Brasil | Equatorial-Tucano | 23 |
| | Matsiguenga | Peru | Equatorial-Tucano | 21 |
| | Piapoco e Curripaco | Colômbia | Equatorial-Tucano | 13 |
| | Quechua | Andes Centrais | Andino | 22 |
| | San Martin | Peru | Andino | 17 |
| | Suruí | Brasil | Equatorial-Tucano | 21 |
| Centro Sul Asiático | Balochi | Paquistão | Indo-Europeu | 25 |
| | Brahui | Paquistão | Dravídica | 25 |
| | Burusho | Paquistão | Isolado | 25 |
| | Hazara | Paquistão | Indo-Europeu | 25 |
| | Kalash | Paquistão | Indo-Europeu | 25 |
| | Makrani | Paquistão | Indo-Europeu | 25 |
| | Pathan | Paquistão | Indo-Europeu | 24 |
| | Sindhi | Paquistão | Indo-Europeu | 24 |
| Europa | Adygei | Cáucaso Russo | Circassiana | 15 |
| | Bergamo | Itália | Indo-Europeu | 13 |
| | Franceses | França | Indo-Europeu | 29 |
| | Bascos Franceses | França | Basco | 24 |
| | Orcadiana | Ilhas Orkney | Indo-Europeu | 16 |
| | Russos | Rússia | Indo-Europeu | 25 |
| | Sardenha | Itália | Indo-Europeu | 28 |
| | Toscanos | Itália | Indo-Europeu | 8 |
| Leste Asiático | Cambojanos | Camboja | Austro-Asiático | 10 |
| | Dai | China | Tai-Kadai | 10 |

| | | | | |
|---|---|---|---|---|
| | Daur | China | Altaico | 10 |
| | Han | China | Sino-tibetano | 39 |
| | Hezhen | China | Altaico | 9 |
| | Japoneses | Japão | Japônico | 30 |
| | Lahu | China | Sino-tibetano | 10 |
| | Miaozu | China | Austro-Asiático | 10 |
| | Mongóis | China | Altaico | 10 |
| | Naxi | China | Sino-tibetano | 10 |
| | Oroqen | China | Altaico | 10 |
| | She | China | Austro-Asiático | 10 |
| | Tu | China | Altaico | 10 |
| | Tujia | China | Sino-tibetano | 10 |
| | Uigur | China | Altaico | 10 |
| | Xibo | China | Altaico | 9 |
| | Yakut | Sibéria | Altaico (?) | 25 |
| | Yizu | China | Sino-tibetano | 10 |
| Oceania | NAN Melanésia | Bougainville | Proto-Oceanico | 15 |
| | Papua | Nova Guiné | Bougainville Sul | 17 |
| Oriente Médio | Beduínos | Israel (Negev) | Afro-asiático | 48 |
| | Drusos | Israel (Carmel) | Afro-asiático | 47 |
| | Mozabite | Algéria (Mzab) | Afro-asiático | 30 |
| | Palestinos | Israel (Central) | Afro-asiático | 49 |
| SNP500Cancer(*) | Afro-americanos | EUA | Indo-Europeu | 24 |
| | Euro-descendentes | EUA | Indo-Europeu | 31 |
| | Hispânicos | EUA | Indo-Europeu | 23 |
| | Asiáticos | EUA | Indo-Europeu | 24 |

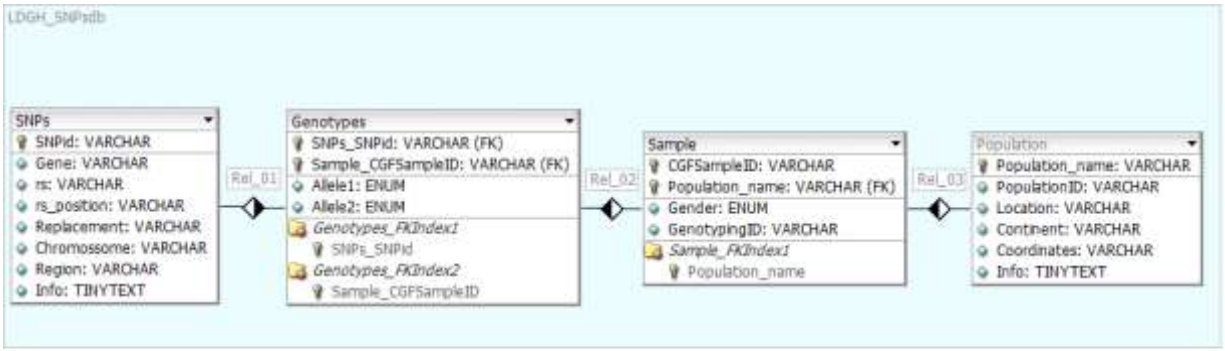
(*) SNP500Cancer não corresponde a um grupo populacional no estudo, entretanto as quatros populações desse painel foram agrupadas sob um único grupo nesta tabela com o intuito de melhor diferenciá-las das demais. Isso porque, para muitas análises, essas populações não foram utilizadas devido à sua origem não autóctone.

[a] Os grupos lingüísticos listados estão representados por filos ou famílias lingüísticas. Foi escolhido o maior nível de classificação visando melhor representar descontinuidades lingüísticas nos grupos populacionais.

(?) Classificação amplamente discutida.

Classificação lingüística de acordo com Ethnologue (www.ethnologue.com)

**ANEXO C – Esquema Entidade-Relacionamento do banco de dados LDGH-SNPsdb**




**Descrição do banco de dados LDGH_SNPs**

A entidade SNPs armazena informações sobre os polimorfismos, sendo constituída pelos seguintes atributos:

SNPid – representa o código de identificação do polimorfismo, no LDGH_SNPs esse identificador corresponde ao código de identificação interno do banco de dados SNP500Cancer.
Gene – determina em qual gene ou região o SNP está presente.
rs – código de identificação para polimorfismos descobertos ou submetidos ao projeto The Single Nucleotide Polymorphism database (dbSNP), um arquivo de domínio público que armazena uma grande quantidade de polimorfismos genéticos simples.
rs_position – posição do SNP no contig de seqüenciamento.
Replacement – determina se há alteração na sequência de aminoácidos devido ao polimorfismo, indicando a posição da troca e quais aminoácidos são alterados.
Chromossome – indica em qual cromossomo e banda citogenética o polimorfismo está presente.
Region – indica a coordenada relativa ao gene do polimorfismo.
Info – campo de texto que permite a adição de informações pertinentes ao SNP, tais como, nomes anteriores, posições diferentes em relação ao anotado no dbSNP e quaisquer outras anotações de importância para projetos desenvolvidos a partir do banco.

A entidade Genotypes armazenas os dados genotípicos provenientes da amostragem

dos indivíduos. É representada pelos atributos Allele1 e Allele2, definindo assim o alelo encontrado em cada uma das fitas genotipadas. É referenciada por duas chaves estrangeiras; uma proveniente da entidade SNPs que permite a identificação do polimorfismo e uma proveniente da entidade Sample que, por sua vez, identifica o sujeito, permitindo assim o relacionamento do genótipo ao polimorfismo e ao indivíduo.

A entidade Sample armazena informações relativas aos sujeitos para os quais dispomos de dados. É constituída pelos seguintes atributos:

CGFSampleId – código de identificação do sujeito genotipado, atualmente contém três identificadores correspondentes aos dados armazendos: HGDP, SNP500Cancer e Nativos-Americanos genotipados pelo SNP500Cancer.
GenotypingId – código de identificação da genotipagem, permite que um mesmo indivíduo tenha mais de uma genotipagem armazenada no banco.
Gender – indica o sexo do sujeito amostrado.

A entidade Population armazena informações relativas às populações de origem dos indivíduos genotipados. Constituída pelos seguintes atributos:

PopulationId – código de identificação da população.
PopulationName – nome da população à qual pertence o indivíduo, é também a chave estrangeira que relaciona a tabela Population à tabela Sample.
Location – localização geográfica da população, podendo ser definida por região ou país.
Continent – indica em qual continente a população se encontra.
Coordinates – indica as coordenadas geográficas dos locais onde os indivíduos foram amostrados.
Info – permite o armazenamento de quaisquer informações relevantes referentes àquela população que não tenham sido anteriormente anotados em algum dos demais campos. Atualmente contêm dados referentes a nomes alternativos de populações.

## **ANEXO D** – Query em MySQL para obtenção de freqüências genotípicas a partir do banco de dados LDGH_SNPsdb

### Exemplos para as populações Adygei e Afro-Americanos:

```
Select genotypes.snpid, population.populationname, concat(genotypes.allele1, genotypes.allele2),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AC')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'CA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AG')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'GA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CG')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'GC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'GG')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'GT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TG')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'TT')/(count(concat(genotypes.allele1, genotypes.allele2)))
From genotypes, sample, population
WHERE population.populationname = 'Adygei'
AND genotypes.cgfsampleid = sample.cgfsampleid
AND sample.populationname = population.populationname
Group by genotypes.snpid
UNION
Select genotypes.snpid, population.populationname, concat(genotypes.allele1, genotypes.allele2),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AC')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'CA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AG')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'GA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'AT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TA')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CG')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'GC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'CT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TC')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'GG')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'GT')/(count(concat(genotypes.allele1, genotypes.allele2))) +
sum(concat(genotypes.allele1, genotypes.allele2) = 'TG')/(count(concat(genotypes.allele1, genotypes.allele2))),
sum(concat(genotypes.allele1, genotypes.allele2) = 'TT')/(count(concat(genotypes.allele1, genotypes.allele2)))
From genotypes, sample, population
WHERE population.populationname = 'Afroamerican'
AND genotypes.cgfsampleid = sample.cgfsampleid
AND sample.populationname = population.populationname
Group by genotypes.snpid
```

**ANEXO E – Distribuição Geográfica das populações de HGDP, SNP500Cancer e LDGH (Nativo-americanos do Peru e Equador)**


**Figura 4: Distribuição geográfica aproximada das populações do HGDP-CEPH, do *SNP500Cancer* e das 4 populações Nativo Americanas do Peru e Equador de nosso laboratório, cujos dados estão disponíveis para esse estudo.**
**1** Bantu NE e SE/SO; **2** Mandenka; **3** Ioruba; **4** San; **5** Pigmeu Mbuti; **6** Pigmeu Biaka; **7** Mozabite; **8** Orcadiana; **9** Adygei; **10** Russa NO; **11** Francesa Basca; **12** Francesa; **13** Bérgamo; **14** Sardenha; **15** Toscana; **16** Beduína; **17** Drusa; **18** Palestina; **19** Balochi; **20** Brahui; **21** Makrani; **22** Sindhi; **23** Pathan; **24** Burusho; **25** Hazara; **26** Uigur; **27** Kalash; **28** Han (China S); **29** Han (China N); **30** Dai. **31** Daur; **32** Hezhen; **33** Lahu; **34** Miaozu (Miao); **35** Oroqen; **36** She; **37** Tujia; **38** Tu; **39** Xibo; **40** Yizu (Yi); **41** Mongólia; **42** Naxi; **43** Camboja; **44** Japonesa; **45** Yakut; **46** Melanésia; **47** Papua; **48** Karitiana; **49** Suruí; **50** Piapoco e Curripaco; **51** Maia; **52** Pima; **53*** Populações Nativas do Peru e Equador (**53-a** Cayapa, **53-b** Quechua, **53-c** San Martín e **53-d** Matsiguenga); **A** Populações do *SNP500Cancer* (Ascendência Caucasiana; Ascendência Asiática; Hispânicos e Afro-americanos).
Em: Chevitarese (2009) - (Figura modificada de Cavalli-Sforza, 2005)

## ANEXO F – Distribuição Geográfica e Definição dos Grupos Populacionais


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

**Grupos Populacionais**: África Ocidental (vermelho); África Oriental (amarelo); Oriente Médio (azul escuro); Europa (rosa); Centro Sul Asiático (azul claro); Leste Asiático (roxo); Oceania (laranja); América Central (preto); América do Sul (marrom). Nativo-americanos correspondem ao conjunto das populações dos grupos América Central (círculos pretos) e América do Sul (círculos marrons). Nordeste Asiático corresponde ao conjunto formado pelas populações Daur (31), Hezhen (32) e Oroqen (35).

Figura modificada de: Cavalli-Sforza, 2005.