Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Biologia Geral
Programa de Pós-Graduação em Genética

Dissertação de Mestrado

DINÂMICA DA MISCIGENAÇÃO EM POPULAÇÕES DA AMÉRICA LATINA

Autora: Fernanda Rodrigues Soares
Orientador: Eduardo Tarazona-Santos

Belo Horizonte
2012

Fernanda Rodrigues Soares

DINÂMICA DA MISCIGENAÇÃO EM POPULAÇÕES DA AMÉRICA LATINA

Dissertação apresentada ao curso de Mestrado do Departamento de Biologia Geral do Instituto de Ciências Biológicas de Universidade Federal de Minas Gerais como pré-requisito parcial para obtenção do título de Mestre em Genética.

Orientador: Eduardo Tarazona-Santos

Belo Horizonte
2012

AGRADECIMENTOS

Agradeço primeiramente à Pós-Graduação em Genética da UFMG, por ter me recebido de portas abertas mesmo vindo de tão longe;

Ao meu orientador, Dr. Eduardo Tarazona-Santos, pelo apoio constante desde meu primeiro momento no Mestrado, por estar sempre presente, paciente e disposto a ensinar e por ter me aceitado em seu laboratório mesmo com meu “aparecimento surpresa”;

Ao prof. Dr. Adrián LLerena, pela confiança e envio dos dados;

À CAPES pelo financiamento com a bolsa durante todo o período de Mestrado;

Aos colegas do LDGH, pelo imenso apoio e ótimo ambiente de trabalho, especialmente ao Dr. Wagner Magalhães, que teve participação crucial em todas as análises deste trabalho;

Às meninas do Genepop, especialmente Jac, Renatinha, Priciane e Michelle pela companhia diária e conversas divertidas, que tornam o ambiente de trabalho bastante agradável;

A todos os professores do departamento, que tiveram participação fundamental em minha formação neste período;

À minha prima Tamy, que me cedeu um cantinho do seu lar durante 10 meses para que eu pudesse me estabelecer em Belo Horizonte;

A toda a minha família e amigos, que me apoiaram muito nesta jornada e tiraram meu stress durante tantos momentos difíceis;

Ao meu namorado Emerson, pelo carinho, ajuda, paciência e amor incondicional oferecido sempre;

E, principalmente, à minha mãe, que é a grande responsável por todas as minhas conquistas profissionais e pessoais, sempre me dando todo apoio possível e impossível para as minhas realizações.

SUMÁRIO

LISTA DE FIGURAS ..... **VI**

LISTA DE TABELAS ..... **VII**

LISTA DE ABREVIATURAS ..... **VIII**

LISTA DE ANEXOS ..... **IX**

RESUMO **1**

ABSTRACT **2**

1. INTRODUÇÃO **3**

1.1. Populações da América Latina *3*

1.2. Miscigenação populacional *6*

1.3. Estimativas de Miscigenação individual e estudo da dinâmica da miscigenação *8*

1.4. Estimativas Moleculares de Miscigenação *12*

2. OBJETIVOS **16**

Objetivo Geral *16*

Objetivos Específicos *16*

3. MATERIAL E MÉTODOS **17**

3.1. Fontes de dados e amostras *17*

3.2. Controle de Qualidade dos dados *18*

3.3. Frequências alélicas *19*

3.4. Heterozigosidade esperada (Hs) *19*

3.5. Análise Molecular de Variância (AMOVA) *20*

3.6. Fst das populações par a par *21*

3.7. Informatividade dos AIMs (In) *21*

3.8. Análise de Componentes Principais (PCA) *22*

3.9. Estimativas de Miscigenação *22*

3.10. Desequilíbrio de Ligação *23*

4. RESULTADOS E DISCUSSÃO **24**

4.1. Controle de qualidade dos dados *24*

4.2. Diferenciação interpopulacional *25*

4.2.1. Frequências alélicas 25

4.2.2. Análise de *Fst* por população 26

4.2.3. Análise de Componentes Principais (PCA) 27

4.3. Diversidade populacional dos AIMs *29*

4.4. Diferenciação intrapopulacional *31*

4.5. Quantificação da miscigenação *34*

4.6. Distribuição da miscigenação individual *35*

4.7. Desequilíbrio de ligação entre loci não ligados gerado pela miscigenação *38*

CONCLUSÕES **40**

REFERÊNCIAS BIBLIOGRÁFICAS **42**

ANEXO 1 ..... **46**

ANEXO 2 ..... **48**

ANEXO 3 ..... **49**

ANEXO 4 ..... **51**

ANEXO 5 ..... **52**

ANEXO 6 ..... **53**

LISTA DE FIGURAS

Figura 1: Modelos hipotéticos de miscigenação com as respectivas relações com valores de r^2 e variância da miscigenação individual.....6

Figura 2: Modelo determinístico de miscigenação de Verdu e Rosenberg (2011).....7

Figura 3: Distribuição da probabilidade da miscigenação de S_1 em cinco cenários diferentes de Verdu e Rosenberg, 2011.....9

Figura 4: Três cenários diferentes de fundação de H com contribuições iguais ao longo das gerações.....10

Figura 5: Três cenários iguais de fundação de H com contribuições diferentes ao longo das gerações, mas com taxas S_1/S_2 iguais.....10

Figura 6: Três cenários diferentes de variância com fundação diferente e miscigenação idêntica ao longo das gerações.....11

Figura 7: Processos opostos de miscigenação: Dois cenários diferentes de variância com mesma fundação e contribuições de miscigenação opostas ao longo das gerações.....11

Figura 8: Fluxograma das análises realizadas neste estudo, com seus respectivos *softwares*.....23

Figura 9: Análise de Componentes Principais a partir dos dados de genótipos dos indivíduos das 13 populações estudadas.....28

Figura 10: Esquema da quantificação da miscigenação individual dos 1957 indivíduos do estudo.....35

Figura 11: Gráficos de distribuição de miscigenação individual nas populações miscigenadas.....35

Figura 12: Gráficos de distribuição dos valores individuais de r^2 para populações miscigenadas e populações parentais.....39

LISTA DE TABELAS

Tabela 1: Exemplo de AIMs para cada uma das populações parentais e para estimar a miscigenação de populações da América Latina, com suas respectivas frequências alélicas.....14

Tabela 2: Tamanho amostral de cada população estudada e número de SNPs genotipados para cada população.....18

Tabela 3: Diferença das frequências alélicas (δ) entre pares de populações parentais para os SNPs do painel de Yeager *et al*. (2008) não incluídos nos dados do CEGEN - Santiago de Compostela.....24

Tabela 4: Frequências alélicas por SNP para cada uma das populações estudadas.....25

Tabela 5: *Fst* par a par para as 13 populações estudadas, com base nos 83 AIMs incluídos no estudo.....27

Tabela 6: Marcadores utilizados no estudo com seus respectivos valores de *In*, *Fct* (F continente/total), *Fst* (F subpopulação/total) e *Fsc* (F subpopulação/continente).....30

Tabela 7: Heterozigosidade esperada calculada para as 13 populações em estudo.....31

Tabela 8: Heterozigosidade esperada por SNP para cada população.....32

Tabela 9: Valores médios de r^2 para as populações parentais e miscigenadas.....38

LISTA DE ABREVIATURAS

AFR – Populações africanas (YRI + MKK + LWK)

AMOVA - Análise Molecular de Variância

CEGEN - Centro Nacional de Genotipado da Espanha – Santiago de Compostela

CEPH - *Centre d'Etude Du Polymorphisme Humain*

CEU - Residentes de Utah com ancestralidade europeia da coleção do CEPH

CQ - Controle de Qualidade

DL – Desequilíbrio de Ligação

EHW - Equilíbrio de Hardy-Weinberg

EM - *Expectation Maximization*

EQU – Indivíduos miscigenados de Quito, Equador

EUR – Populações europeias (CEU + TSI)

EXT – Espanhóis de Extremadura, Espanha

Fct - F variância entre continente em relação ao total

Fsc - F variância entre populações em relação a cada continente

Fst – F variância entre populações em relação a variância total

GLU – *Genotype Library and Utilities*

HP - Ashaninkas do departamento de Junin, Peru

Hs - Heterozigosidade esperada

LIM – Miscigenados de Lima, Peru

LWK - Luhya em Webuye, Quênia

MEX- Descendentes de mexicanos em Los Angeles, Califórnia

MIA – Marcadores Informativos de Ancestralidade

MKK - Maasai em Kinyawa, Quênia

ML - *Maximum Likelihood*

NAT – Populações nativo-americanas (HP + SHI + PUN)

NIC - Indivíduos miscigenados de Managua, Nicarágua

PCA - Análise de Componentes Principais

PUN - Quechuas do departamento de Puno, Peru

RIBEF - Rede IberoAmericana de Farmacogenética

SBE - *Single Base Extension*

SHI - Shimaas do departamento de Cuzco, Peru

SNP – *Single Nucleotide Polymorphism*

STR – *Short Tandem Repeats* (Microssatélites)

TSI - Toscanos da Itália

YRI - Yoruba em Ibadan, Nigéria

LISTA DE ANEXOS

Anexo 1: Informações dos SNPs do painel de ancestralidade utilizados neste estudo.....46

Anexo 2: Linhas de Comandos utilizadas no programa GLU.....47

Anexo 3: Linhas de Comandos utilizadas no pacote Adegenet da plataforma R.....48

Anexo 4: Linhas de Comandos utilizadas no pacote Hierfstat da plataforma R.....50

Anexo 5: Linhas de Comandos utilizadas na plataforma R para a geração dos gráficos de barras.....51

Anexo 6: Linhas de Comandos utilizadas na plataforma R para a geração dos gráficos de distribuição de valores.....52

RESUMO

A miscigenação é uma forma de fluxo gênico entre populações isoladas por um longo período de tempo, que resulta em uma nova população híbrida. O estudo da miscigenação é importante para estudos antropológicos, de genética associativa e epidemiológica para evitar associações estatísticas espúrias e consequentemente falsos positivos. Nosso grupo de pesquisa (LDGH – UFMG) tem como foco estudar aspectos da miscigenação e genética epidemiológica em populações com alta ancestralidade nativo-americana, que são normalmente negligenciadas em estudos populacionais. Este trabalho visa, através de um painel de 83 SNPs marcadores informativos de ancestralidade (AIMs), estimar a distribuição da miscigenação individual e populacional em quatro populações da América Latina: Peru (LIM), Equador (EQU), Nicarágua (NIC) e México (MEX), produto da miscigenação entre europeus, africanos e nativos americanos, além de avaliar a eficiência do painel de SNPs utilizado e inferir qualitativamente aspectos da dinâmica da miscigenação dessas populações no tempo. Empregamos métodos clássicos de genética de populações para calcular a diferenciação inter e intra-populacional – Estatísticas F de Wright, Equilíbrio de Hardy-Weinberg, Heterozigosidade esperada, *Maximum likelihood* para estimar a miscigenação e Desequilíbrio de Ligação (DL). A metodologia de Análise de Componentes Principais (PCA) também foi utilizada para confirmar e ilustrar de forma concisa a diferenciação populacional. Todos os resultados destas análises foram coerentes entre si, evidenciando que os marcadores utilizados têm, em geral, baixa diversidade intracontinental e alta intercontinental, como esperado para AIMs. Nenhuma das populações miscigenadas apresentou alto componente africano, com maior média em 0,12 em NIC. A PCA mostrou um contínuo de miscigenação individual entre europeus e ameríndios nas populações miscigenadas, com os dois primeiros componentes principais explicando 42,7% de toda a variação encontrada nos dados. Os indivíduos de populações africanas foram os mais diferenciados entre todas as populações estudadas. A análise de DL, juntamente com a variância e distribuição da miscigenação individual, nos permitiu inferir qualitativamente a dinâmica da miscigenação das populações EQU, NIC, MEX e LIM, sugerindo que as populações NIC e MEX evidenciam eventos de miscigenação mais recentes, e a ausência de eventos de miscigenação relativamente recentes nas populações LIM e EQU. Em um futuro próximo, usaremos ferramentas conceituais mais sofisticadas para estudar a distribuição da miscigenação individual e DL para aprimorar nossas inferências sobre a dinâmica da miscigenação nessas populações.

Palavras-chave: AIMs, Miscigenação, Genética de Populações.

ABSTRACT

Admixture is the product of gene flow between long-time isolated populations that form a new hybrid population. Admixture studies are important in anthropology, and in genetic epidemiology to avoid spurious statistical associations due to false positive. Our research group (LDGH – UFMG) aims to study admixture aspects and genetic epidemiology of populations with high native American ancestry, that are normally neglected in population studies. This study aims, through a 83-SNP panel of ancestry informative markers (AIMs), estimate the individual and population admixture distribution of four populations of Latin America: Ecuadorian (EQU), Nicaraguan (NIC), Mexican (MEX) and Peruvian (LIM), product of admixture between European, African and Native American, and evaluate the efficiency of the panel used in this investigation, and to perform qualitative inferences of the dynamic of admixture in these populations. We performed classic population genetics analyses in order to assess intra-population and inter-population diversity indexes (F statistics, Hardy-Weinberg Equilibrium, *Maximum likelihood* estimations of admixture and Linkage Disequilibrium (LD) estimators). Principal Component Analysis (PCA) was also used to confirm and illustrate the population differentiation. Overall, these analyses produced consistent results, showing that most of our SNPs have low inter-continental and high inter-continental diversity, as expected for AIMs. None of the admixed populations presented high African ancestry, with the highest average population African contribution of 0,12 in Nicaragua. PCA analysis showed a continuum of individual European and Native American ancestry in the admixed populations, with the two first principal components' statistics explaining 42,7% of the total variance contained in the genetic dataset. Individuals of African populations were the most differentiated between all populations studied. The analysis of LD and of the distribution of individual ancestry allowed us to qualitatively address the temporal dynamics of admixture in EQU, NIC, MEX and LIM. It suggested that NIC and MEX undergone most recent events of admixture, and relatively recent contributions of individuals with high European ancestry were low in EQU and LIM. In the near future, we will use more sophisticated conceptual tools to study the distribution of individual admixture and linkage disequilibrium to improve our inferences about the admixture dynamics in these populations.

Keywords: AIMs, Admixture, Population Genetics.

1. INTRODUÇÃO

1.1. Populações da América Latina

A história evolutiva das populações africanas, europeias e ameríndias pode ser sintetizada da seguinte forma: A nossa espécie se originou na África há cerca de 100 mil anos atrás (Salzano e Bortolini, 2002). A longa permanência desta espécie no continente africano fez com que a mesma adquirisse uma grande variabilidade cultural e genética que são observadas até hoje nos africanos. A migração desses indivíduos ocorreu há cerca de 80 mil anos atrás (segundo a Teoria "*Out of Africa*"), colonizando outros continentes, primeiramente a Ásia e Oceania (aproximadamente 40 mil anos atrás) e então a Europa. Após o estabelecimento da espécie na Ásia Central, houve uma rota migratória para a América, via Estreito de Bering, há cerca de 15 mil anos atrás (Cavalli-Sforza e Feldman, 2003).

As populações atuais de América Latina receberam várias contribuições de populações parentais ao longo do tempo. Os maiores grupos populacionais que contribuíram com esta ancestralidade diversa foram europeus, africanos e os próprios ameríndios. Os dois primeiros grupos deixaram sua contribuição a partir do século XV. Além dessas contribuições, este continente teve uma história de migração muito mais complexa ao longo do tempo, envolvendo outros grupos populacionais com contribuições relativamente menores, como leste-asiáticos, especialmente chineses e japoneses (Salzano e Bortolini, 2002; Galanter *et al*., 2012).

As populações europeias chegaram às Américas a partir de 1492, desde quando migram constantemente para este continente. As populações africanas também começaram a chegar nesta mesma época, através da escravidão imposta pelos europeus que até então colonizavam a América. A escravidão alcançou seus maiores níveis entre os séculos XVI e XIX, com um número estimado de 9 milhões de escravos trazidos para a América entre 1451 e 1870. Este fato é parte da história de muitas populações da época e produziu um trágico impacto nas sociedades do oeste da África (Salzano e Bortolini, 2002; Reader, 1998).

A demografia histórica estuda a variabilidade de uma população do passado, no tempo e espaço. A demografia histórica e a genética se relacionam através de estudos de padrões de acasalamento, que estão relacionados com as migrações, fertilidade e mortalidade de populações do passado, onde podemos, conhecendo os dados destas, fazer inferências sobre como surgiram os indivíduos modernos nestas mesmas populações (Salzano e Bortolini, 2002). No entanto, a genética pode esclarecer a contribuição biológica do isolamento a longo prazo para as populações, enquanto a demografia histórica possui um menor número de ferramentas para isto. A genética pode colaborar produzindo e

analisando dados de DNA para estimar a origem ancestral de determinadas populações, comparando populações miscigenadas com populações parentais, com auxílio de análises estatísticas. Além disso, com cálculos de genética de populações podemos inferir padrões de migração e acasalamento dessas populações.

A genética de populações tem dois objetivos principais: (1) descrever quantitativamente a variabilidade genética presente nas populações e (2), inferir como a combinação de diferentes fatores evolutivos gerou o padrão de variabilidade observado em genes específicos ou no genoma populações. Os fatores evolutivos considerados em genética de populações são a mutação, a deriva genética, o fluxo gênico, a seleção natural, a migração, a recombinação e o padrão de acasalamento (Hartl e Clark, 2010).

A descrição mais simples em genética de populações para um determinado *locus*, consiste em determinar as frequências alélicas e genotípicas. Para descrever quantitativamente a variabilidade genética presente nas populações também se utilizam estatísticas que quantificam a variabilidade intra-populacional e inter-populacional, que são explicadas na seção Materiais e Métodos desta dissertação, e que, a grosso modo, incluem as seguintes estatísticas: F de Wright, Heterozigosidade esperada e Desequilíbrio de Ligação (DL).

O termo estruturação populacional refere-se a como a diversidade genética é distribuída em diferentes níveis hierárquicos, por exemplo: indivíduos, populações, grupos de populações e população humana total. Para quantificar a estrutura genética das populações, foram desenvolvidas metodologias similares à Análise de Variância, como a análise das estatísticas F (Hartl e Clark, 2010), ou a Análise Molecular de Variância (AMOVA, Excoffier, Smouse e Quattro, 1992), utilizado nesta dissertação, cuja descrição quantitativa se encontra na seção Materiais e Métodos.

Atualmente, métodos exploratívos de estatística multivariada, como a Análise de Componentes Principais (PCA, Cavalli-Sforza, 1998), são também utilizados para ilustrar graficamente as relações entre indivíduos e populações a partir de genótipos, e também para quantificar a estrutura populacional, obtida através dos marcadores que estão sendo estudados. Esta análise também é descrita em detalhe na seção Materiais e Métodos desta dissertação.

A genética de populações também tem desenvolvido modelos estatísticos que descrevem estatísticas de variabilidade genética sob determinadas condições definidas pela presença ou ausência de fatores evolutivos específicos. O modelo mais simples é o Equilíbrio de Hardy-Weinberg (EHW), que assume algumas premissas que podem ser ilusórias para conjuntos de dados de populações humanas, como gerações não sobrepostas, cruzamento aleatório, tamanho populacional infinito, migração desprezível, mutações nulas e efeitos de Seleção Natural ínfimos (Hartl e Clark, 2010). No entanto, na

prática, o principal motivo de afastamento do EHW é algum tipo de problema na genotipagem de amostras. Por isso, o EHW é um poderoso método de controle de qualidade dos dados. Tendo como exemplo um *locus* aleatório com os alelos *A* e *a*, o modelo pode ser descrito através da fórmula:

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

Onde:

⇒ *p* = Frequência alélica de *A*;

⇒ p^2 = Frequência genotípica esperada de homozigotos *AA*;

⇒ *2pq* = Frequência de heterozigotos *Aa* esperada na população;

⇒ *q* = Frequência alélica de *a*;

⇒ q^2 = Frequência genotípica esperada de homozigotos *aa*.

Uma característica importante das populações é o desequilíbrio de ligação (DL), que consiste na associação estatística em gametas entre alelos de *loci* diferentes, em contraposição ao conceito de equilíbrio de ligação, que consiste na independência entre alelos de *loci* diferentes. A medida do DL é importante no estudo das populações miscigenadas, porque a miscigenação pode gerar desequilíbrio de ligação, seja entre loci ligados no mesmo cromossomo ou entre loci não ligados, transmitidos pelo mesmo gameta. Espera-se que populações parentais diferenciadas possuam, em seu genoma total, um menor desequilíbrio de ligação (e maior equilíbrio de ligação) que populações miscigenadas, em parte porque são populações mais antigas e possuíram bastante tempo para sofrer recombinação entre seus *loci* e, portanto, possuem menos *loci* ligados. Uma das estimativas para quantificar o desequilíbrio de ligação é o coeficiente de correlação r^2 entre a presença (denotada por 1 ou ausência, denotada por 0) de dois alelos em loci diferentes. Entretanto, quando se trata de dados não ligados, como no nosso caso, a fase gamética dos duplos heterozigotos é desconhecida e difícil de ser inferida utilizando métodos para inferência de haplótipos como o software PHASE (Stephens *et al.* 2001), e nesse caso a proporção de duplos heterozigotos deve ser inferida por outros métodos.

A estimativa mais simples do desequilíbrio de ligação é a do parâmetro *D* (Hartl e Clark, 2010).

$$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$$

Onde:

⇒ *D* = Parâmetro de desequilíbrio de ligação;

⇒ *P* = Frequências dos haplótipos ou gametas subscritos;

⇒ *AB* = Gameta com os alelos *A* e *B*;

⇒ *Ab* = Gameta com os alelos *A* e *b;*

⇒ *aB* = Gameta com os alelos *a* e *B;*

⇒ *ab* = Gameta com os alelos *a* e *b;*

Outras duas estimativas de desequilíbrio de ligação são D' (o D normalizado) e r^2, que consistem, para loci bialélicos, em:

$$D' = \frac{D}{D\ máximo} \qquad r^2 = \frac{D^2}{(p_A q_a p_B q_b)}$$

Onde:

⇒ *p* e *q* = Frequências alélicas dos alelos subscritos.

⇒ *D máximo* = Valor absoluto do máximo valor possível do D dadas a frequência alélica *p*.

1.2. Miscigenação populacional

A miscigenação é uma forma de fluxo gênico onde ocorre uma troca de genes entre duas populações anteriormente isoladas por um longo período de tempo, que resulta em uma nova população híbrida (miscigenada; Tarazona-Santos *et al*., 2007).

Costumava-se trabalhar com dois modelos extremos de miscigenação: (1) modelo de isolamento, onde há apenas um evento migratório e o fluxo gênico cessa, havendo cruzamentos apenas nos indivíduos da população no decorrer das gerações e; (2) modelo de fluxo gênico contínuo, onde há imigrantes inserindo novos alelos na população durante várias gerações seguidas (Figura 1). A maior parte da dinâmica da miscigenação nas populações miscigenadas de América Latina deve se enquadrar em algum lugar do espectro de possibilidades entre estes dois modelos extremos.

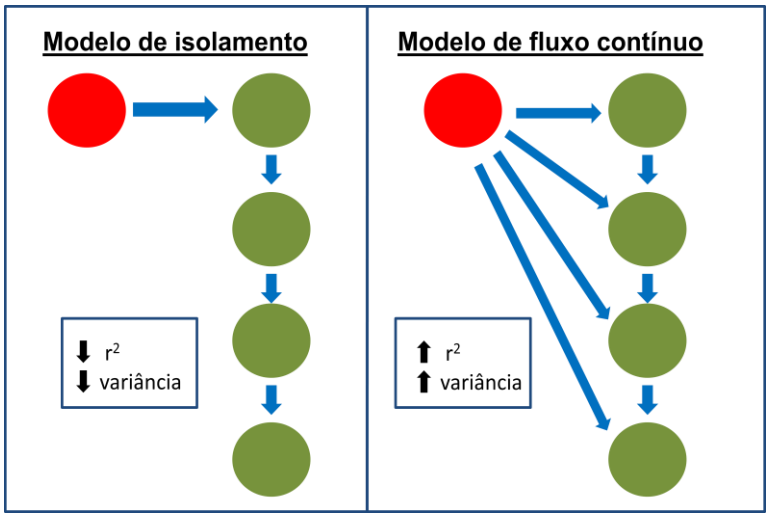

Figura 1: Modelos hipotéticos de miscigenação com as respectivas relações com valores de r^2 e variância da miscigenação individual (Long, 1991).

Recentemente, um modelo que compreende os dois anteriores foi formalizado por Verdu e Rosenberg (2011). Este modelo permite um número de "*n*" populações parentais (S) e uma população híbrida (H) resultante do cruzamento das parentais (Figura 2). Por simplicidade, apresentamos algumas conclusões do estudo de Verdu e Rosenberg (2011) para populações di-híbridas, mas o modelo é geral para *n* populações parentais.

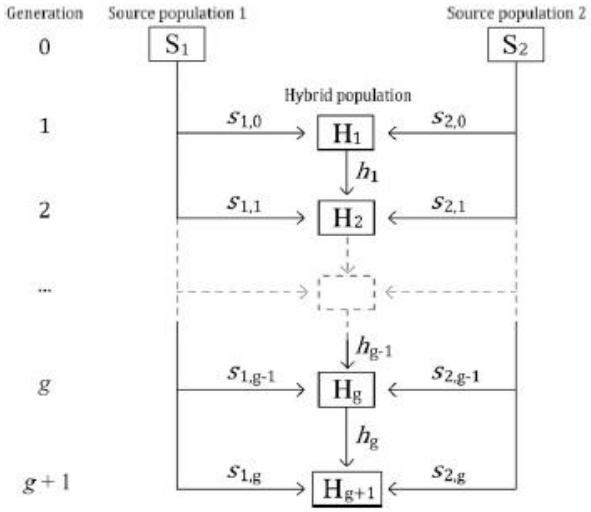

Figura 2: Modelo determinístico de miscigenação com duas populações parentais (S) e uma população híbrida (H). Adaptada de Verdu e Rosenberg (2011).

A Figura 2 esquematiza duas populações parentais (S_1 e S_2) formando uma população híbrida (H_1). As probabilidades de contribuições para a formação de H_2, visto que a primeira híbrida já foi formada, são h_1 (da primeira geração de H), $s_{2,1}$ (da primeira geração de s_2) e $s_{1,1}$ (da primeira geração de s_1). Essa sucessão continua até a geração g + 1, onde a mesma será formada por h_g, $s_{1,g}$ e $s_{2,g}$. As probabilidades de contribuição devem ter valores entre 0 e 1, onde $s_{1,g} + s_{2,g} + h_g = 1$. Na geração 0, onde não há híbridos, $h_g = 0$, então $s_{1,g} + s_{2,g} = 1$.

As populações da América Latina podem ser consideradas como tri-híbridas, com populações africanas, europeias e ameríndias como parentais. Fatores complexos da história de cada uma dessas populações deram origem a diferentes proporções de contribuições ancestrais individuais dentro de uma população e entre populações da América Latina (Galanter *et al*., 2012).

A estimativa de ancestralidade biogeográfica é muito importante em estudos de genética epidemiológica de populações miscigenadas (Tarazona-Santos *et al.,* 2007). Um problema recorrente em estudos de associação são os falsos positivos. Em qualquer estudo caso-controle, o primeiro passo é fazer uma amostragem da população, onde casos e controles devem ser da mesma população, para a amostragem ser etnicamente

homogênea. Se houver uma amostragem etnicamente diferente entre casos e controles, e houver uma maior prevalência da doença na população de casos amostrada, pode haver uma associação estatística espúria, pois indivíduos com a ancestralidade da população onde a doença for mais comum podem estar super-representados na amostragem de casos e qualquer alelo mais comum nesta população pode dar correlação estatística com a doença, mesmo sem ser um alelo de suscetibilidade. Portanto, em todos os estudos de associação, o primeiro passo deve ser estimar a ancestralidade dos indivíduos amostrados, para verificar se casos e controles são etnicamente diferentes e evitar este problema de associação espúria ou controlar o efeito das diferenças de ancestralidade entre casos e controles.

1.3. Estimativas de Miscigenação individual e estudo da dinâmica da miscigenação

Os estudos clássicos costumavam ser focados na miscigenação populacional. Atualmente, a disponibilidade de um maior número de marcadores genotipados com custos razoáveis permitem o estudo da miscigenação de cada indivíduo (individual), cuja distribuição é informativa em relação não unicamente à quantidade de miscigenação, mas à sua dinâmica, ou seja, como ocorreu o processo de miscigenação no tempo.

Verdu e Rosenberg (2011) tem estudado a dinâmica temporal da média e da variância da miscigenação individual sob o modelo determinístico geral de miscigenação mostrado na Figura 2, que incorpora os modelos clássicos de miscigenação e isolamento e de fluxo gênico constante. Eles evidenciam que:

(1) Populações híbridas que possuem 0,5 de contribuição de cada uma das duas parentais na geração 0, mas nas gerações seguintes a contribuição cessa, a distribuição ($P(H_{1,g})$) de um indivíduo exibir uma fração de miscigenação da população 1 ($S_{1,0}$) é sempre em torno de 0,5, por mais que se passem cerca de 6 ou mais gerações (Figura 3A), o que mostra que após um efeito de fundação simétrica sem imigrações adicionais, a distribuição permanece simétrica;

(2) Populações com o mesmo cenário de fundação de (1), mas com contribuições simétricas de S_1 e S_2 (onde Si é a contribuição constante da população i ao longo das gerações das duas populações parentais) após o cenário de fundação, tendem a continuar com uma distribuição simétrica em torno de 0,5 ao longo das gerações pois a contribuição será sempre a mesma (Figura 3B);

(3) Com o cenário de fundação igual ao dos demais modelos apresentados ($S_{1,0}$ = $S_{2,0}$ = 0,5), mas com uma contribuição diferente entre as duas populações parentais ao longo das gerações (S_1 = 0.0001 e S_2 = 0,2), a medida da distribuição de $H_{1,g}$ não é cerca de 0,5, pois os indivíduos vão tender a possuir uma maior ancestralidade da população 2,

produzindo menores valores de contribuição da população 1 (Figura 3C), ou seja, a contribuição posterior ao cenário de fundação tende a predominar;

(4) Uma população contribuindo mais que a outra para o cenário de fundação ($S_{1,0}$ = 0,8 e $S_{2,0}$ = 0,2), com a mesma contribuição nas gerações subsequentes que o modelo anterior (S_1 = 0.0001 e S_2 = 0,2), ou seja, contrárias ao cenário de fundação, vai tender também a ter uma ancestralidade maior para S_2, que contribui mais ao longo das gerações, exceto nas primeiras gerações, onde o cenário de fundação prevalecerá (Figura 3D);

(5) Com cenário de fundação idêntico ao anterior (4) mas com uma contribuição semelhante à da fundação, de S_1 = 0,2 e S_2 = 0,0001, a distribuição de $H_{1,g}$ será acelerada para a direita, com maior contribuição de S_1 ao longo das gerações.

Esses cenários nos demonstram que: (1) se as contribuições de populações parentais ocorrem apenas na primeira geração, os níveis de miscigenação continuam os mesmos ao longo das gerações; (2) as condições iniciais semelhantes podem levar a padrões subsequentes diferentes dependendo das contribuições que essa população receber ao longo do tempo; e (3) com contribuições constantes ao longo do tempo, as condições iniciais influenciam a velocidade que a distribuição da miscigenação adota, sem alterar a forma da distribuição com o passar das gerações.

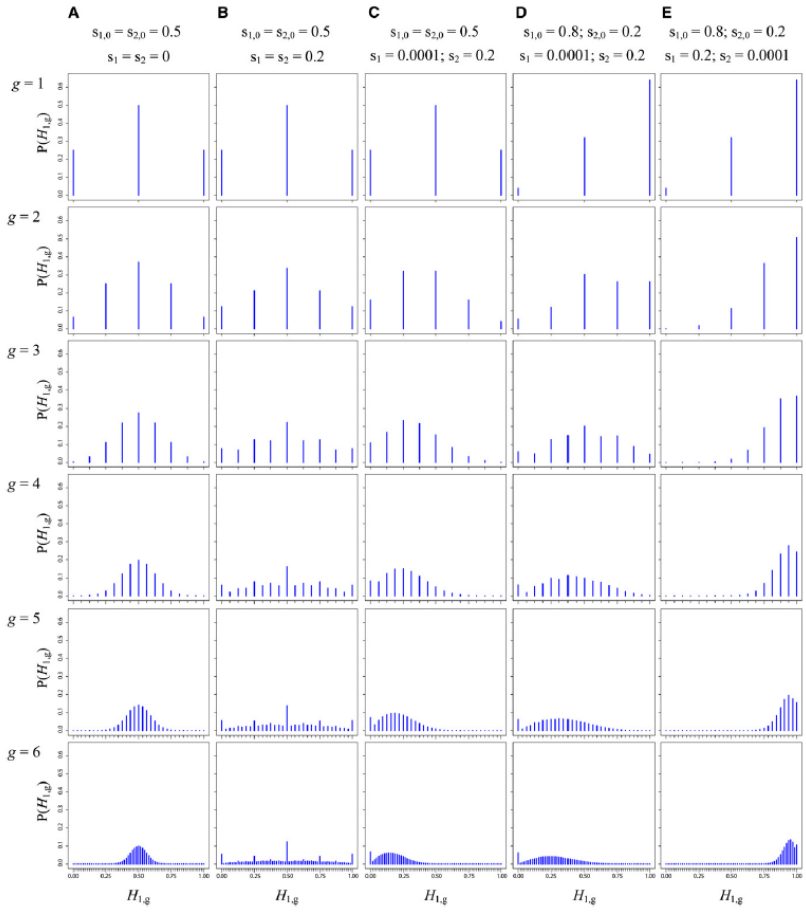

Figura 3: Distribuição da probabilidade da miscigenação de S_1 em cinco cenários diferentes, onde o eixo X representa a fração de miscigenação e o eixo Y representa a probabilidade de um indivíduos exibir uma fração

da miscigenação de S_1. (A) H é fundada com proporções iguais de populações parentais e não recebe contribuições subsequentes, identificando um modelo de isolamento; (B) H é fundada com proporções iguais de parentais e recebe proporções iguais de contribuições subsequentes; (C) H é fundada com proporções iguais de parentais e recebe contribuição de S_2 maior que de S_1 ao longo das gerações; (D) H é fundada com maior contribuição de S_1 mas recebe maior contribuição de S_2 nas gerações subsequentes; (E) H é fundada com maior proporção de S_1 e recebe maior contribuição dessa mesma população parental nas gerações subsequentes. As figuras (B-E) correspondem ao modelo de fluxo contínuo. Adaptada de Verdu e Rosenberg, 2011.

Estes autores também avaliaram o efeito da distribuição da miscigenação individual ao longo do tempo. Se uma população possui cenários de fundação diferentes mas têm os mesmos parâmetros de introgressão (fluxo gênico entre populações isoladas geograficamente por um longo período de tempo, sem isolamento reprodutivo), ou seja, a miscigenação no cenário pós fundacional é constante para diferentes cenários, que tendem a alcançar o mesmo valor esperado de contribuição da população 1 ($E[H_{i,g}]$) ao longo do tempo (Figura 4). De outra forma, se populações possuem 3 cenários iguais de fundação mas possuem diferentes parâmetros de introgressão para S_1 e S_2, mas com a mesma taxa S_1/S_2, o valor esperado da contribuição da população S_1, $H_{1,g}$, também é fixo a longo prazo dependendo de S_1/S_2 (Figura 5).

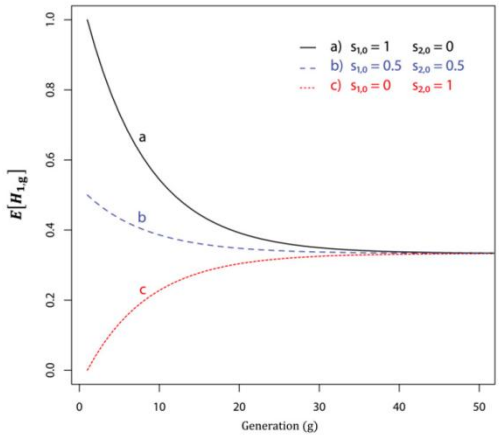

Figura 4: Três cenários diferentes de fundação de H com contribuições iguais ao longo das gerações. Os cenários tendem a alcançar o mesmo valor ao longo do tempo (cerca de 0,35).

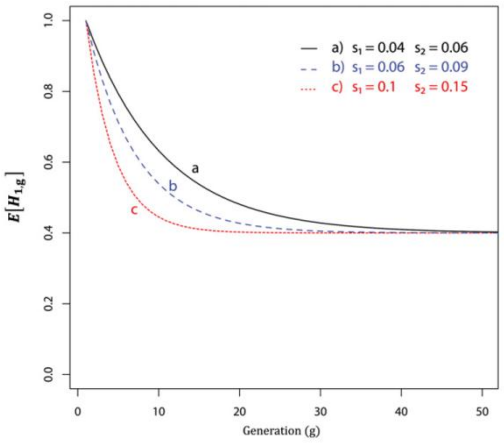

Figura 5: Três cenários iguais de fundação de H com contribuições diferentes ao longo das gerações, mas com taxas S_1/S_2 iguais. Os cenários tendem a alcançar o mesmo valor ao longo do tempo (cerca de 0,4).

A variância da miscigenação individual é também um fator a ser considerado em estudos de miscigenação populacional, pois ela decresce à medida que as populações parentais param de contribuir com a população miscigenada, da mesma forma que a variância é afetada quando a população parental continua contribuindo com fluxo gênico na miscigenada (Verdu e Rosenberg, 2011).

O comportamento da variância interindividual da miscigenação também foi avaliado por Verdu e Rosenberg (2011). Primeiramente três cenários foram testados com diferentes efeitos de fundação, com miscigenação idêntica e constante em todos os cenários. Ao longo das gerações, as variâncias tendem a alcançar um mesmo valor com relação à contribuição da população 1 (Figura 6). Dois cenários foram testados com o mesmo cenário de fundação

mas com efeitos de miscigenação opostos constantes. As variâncias da contribuição da população S_1 também alcançam o mesmo valor ao longo das gerações (Figura 7).

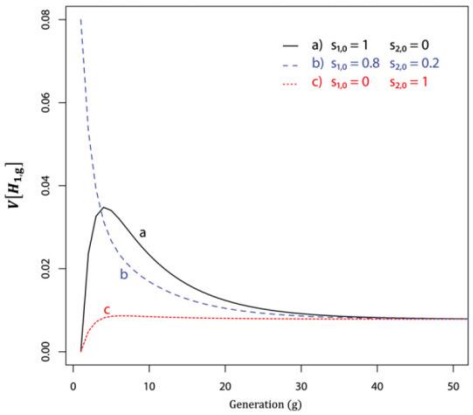

Figura 6: Três cenários com fundação diferente e miscigenação idêntica ao longo das gerações. Os cenários tendem a alcançar o mesmo valor de variância ao longo do tempo (cerca de 0,01).

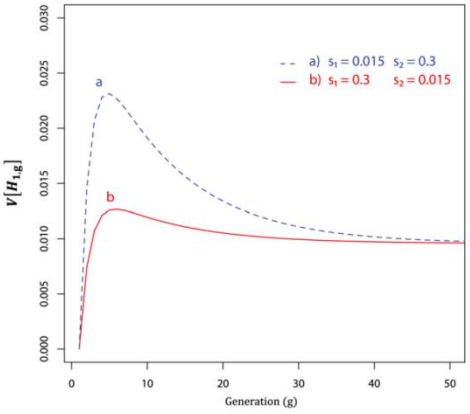

Figura 7: Processos opostos de miscigenação: Dois cenários de mesma fundação e contribuições de miscigenação opostas ao longo das gerações. Os cenários tendem a alcançar o mesmo valor de variância ao longo do tempo (cerca de 0,01).

A miscigenação gera DL na população miscigenada mesmo entre *loci* não ligados. Dados dois *loci* bialélicos não ligados A e B fixados para alelos alternativos diferentes em duas populações parentais 1 e 2, os indivíduos de cada uma das populações apresentarão apenas um tipo de gameta (A1B1 para população 1; A2B2 para população 2). Em uma hibridização panmítica entre as populações 1 e 2 que contribuem cada uma com 50% para a população híbrida panmítica (frequências de A1 = 0,5; B1 = 0,5; A2 = 0,5; B2 =0,5), as frequências genotípicas na primeira geração pós miscigenação, assumindo o EHW nas populações parentais e nas híbridas a partir da primeira geração após a miscigenação, serão (denotando os alelos como A1, A2, B1 e B2):

¼ A1B1/A1B1;

¼ A2B2/A2B2 ;

½ A1A2/B1B2.

Não haverá os genótipos A1A1/B2B2, A2A2/B1B1, A1A1/B1B2, A2A2/B1B2, A1A2/B1B1, A1A2/B1B2, uma vez que existe apenas um tipo de gameta em cada população parental.

Nesta distribuição, é evidente que não existe independência estatística entre alelos, ou seja, existe DL gerado pelo evento de miscigenação, o que depende do fato de que as frequências alélicas nas populações parentais sejam diferentes. Neste caso alelos alternativos estão fixados.

Em ausência de contribuição posterior das populações parentais, as frequências genotípicas tenderão para a independência estatística (equilíbrio de ligação), onde as

frequências serão determinadas pelo produto das frequências genotípicas dos loci A e B (que se encontram sob EHW):

1/16 A1A1/B1B1;

1/16 A2A2/B2B2;

1/16 A1A1/B2B2;

1/16 A2A2/B1B1;

1/8 A1A1/B1B2;

1/8 A1A2/B1B1;

1/8 A1A2/B2B2;

1/8 A2A2/B1B2;

1/4 A1A2/B1B2.

Em geral, quanto maior as diferenças entre as frequências alélicas das populações parentais, maior o DL gerado por miscigenação (neste caso entre *loci* não ligados).

Com o passar das gerações, o DL vai caindo de acordo com o modelo de miscigenação que a população se enquadrar: se a população apresentar o modelo de isolamento, a queda será mais rápida. No caso do modelo de fluxo contínuo, a queda será mais lenta, pois as populações parentais continuarão contribuindo com altas frequências de haplótipos A1B1.

1.4. Estimativas Moleculares de Miscigenação

Para estimar a ancestralidade é possível genotipar um painel de marcadores moleculares que mostram grande diferenciação entre grupos geográficos ancestrais (Yeager *et al*., 2008). Estes marcadores existem porque o isolamento geográfico prolongado de uma população gera diferenças em suas frequências alélicas quando comparada a outras populações. Isso ocorre por deriva genética (flutuação aleatória ao acaso das frequências alélicas) e mutações exclusivas daquela população, que pode gerar alelos específicos na mesma.

Bernstein (1931) e Ottensooser (1944) foram os pioneiros a usar dados de freqüências alélicas em populações parentais e miscigenadas para avaliar as contribuições acumuladas de grupos ancestrais para determinada população miscigenada (Salzano e Bortolini, 2002). Se obtivermos um número suficiente de marcadores genéticos, podemos quantificar a ancestralidade de uma pessoa (Salzano, 1997).

Podemos encontrar estudos com estimativas de miscigenação populacionais, individuais e ao longo dos cromossomos, assim como por continentes, regiões subcontinentais e populações. Todas essas estimativas possuem um erros embutidos, pois é complicado obter amostras das populações parentais genuinamente representativas de uma população, região ou do mundo (Cavalli-Sforza, 1998), além do erro amostral e das

diferenças entre as frequências no momento da amostragem e aquelas no passado nas populações parentais e híbridas devido à deriva genética.

A miscigenação individual pode ser calculada pelo método de *Maximum Likelihood* (ML). Neste estudo utilizamos esse método acrescido do algoritmo *Expectation Maximization* (EM, Dempster, 1977), no caso deste trabalho implementado no programa Admixture (Alexander *et al*., 2009), como é descrito na seção Materiais e Métodos desta dissertação.

Existem vários tipos de marcadores moleculares utilizados para estimar a miscigenação individual. Eles podem ser classificados quanto à sua localização: Cromossomos autossômicos (herança biparental), DNA mitocondrial e Cromossomo Y (herança uniparental); e também quanto à natureza molecular: Microssatélites (repetições em tandem de sequências de nucleotídeos), Indels (polimorfismos de inserção e deleção) e SNPs (*Single Nucleotide Polymorphisms* – variação em um único nucleotídeo de DNA).

Marcadores de Cromossomo Y e DNA mitocondrial são uniparentais (patrilinear e matrilinear, respectivamente) e por isso não estimam ancestralidade individual, mas conseguem traçar a história genética exclusivamente da linhagem paterna ou materna das populações.

Os microssatélites são marcadores multialélicos, e os alelos correspondem ao número de repetições do motivo. O marcador D1S80, por exemplo, possui mais de 30 alelos diferenciados nas populações europeia, africana e ameríndia (Duncan *et al*., 1996).

Os Indels são produtos de uma inserção ou uma deleção num determinado sítio de DNA e estão sendo utilizados também com o intuito de estimar ancestralidade individual, com sucesso (Pena *et al*., 2011).

Os SNPs são os marcadores mais abundantes do genoma (The International HapMap Consotium, 2003). Atualmente eles estão sendo amplamente utilizados em estudos genéticos e vemos um grande número deles já caracterizados e disponíveis em grandes bancos de dados, como o Projeto Internacional Hapmap (The International HapMap Consortium, 2010, www.hapmap.org), Seattle SNPs (http://pga.gs.washington.edu), SNP 500 Cancer (Packer *et al*, 2004) e dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Essa disponibilidade de milhões de marcadores nestes bancos de dados nos levou cada vez mais próximo à ancestralidade real dos indivíduos inclusive no nível de cromossomos (Via *et al*., 2009).

Marcadores com frequências alélicas altamente diferenciadas entre populações parentais são considerados informativos para estimar a ancestralidade de indivíduos miscigenados, e são denominados Marcadores Informativos de Ancestralidade (AIMs). Alguns autores consideram um mínimo de 0,5 de diferenciação em frequências alélicas entre duas populações parentais para considerar um determinado marcador como MIA (Yeager *et al.,* 2008; Galanter *et al*., 2012).

Se um número suficiente de AIMs for caracterizado, podemos inferir a ancestralidade em níveis de resolução refinados, como individual e ao longo de cromossomos, não apenas populacional. O número de AIMs necessários para inferir ancestralidade com um determinado nível de acurácia é menor que o número de marcadores selecionados ao acaso, o que facilita e diminui o custo das análises (Rosenberg *et al*., 2003; Galanter *et al*., 2012).

O número de marcadores utilizados para determinar a ancestralidade de cada população depende da informatividade dos marcadores selecionados. Vários painéis de AIMs para populações humanas são desenvolvidos constantemente visando selecionar os de maior informatividade para estimativas mais robustas de ancestralidade (Yeager *et al.,* 2008; Galanter *et al*., 2012; Santos *et al.*, 2010; Lins *et al*., 2009, da Silva et al., 2010).

Com base em comparações entre estimativas de miscigenação obtidas com 600 mil SNPs e painéis de dezenas-centenas de AIMs em diferentes populações latino-americanas, Galanter *et al* (2012) sugerem que um painel com 88 AIMs possui informatividade similar a painéis de 194 ou 314 AIMs, com os três painéis fornecendo boas estimativas de ancestralidade individual, com erros de cerca de 5%. Yeager *et al* (2008) sugerem ainda que os marcadores selecionados para um painel devem ter uma distância física suficiente uns dos outros (média de 2,4 x 10^7 pares de bases) para fornecer informações independentes. AIMs utilizados em populações latino-americanas para estimar as ancestralidades no nível continental, visam minimizar a heterogeneidade intracontinental e aumentar a diferença intercontinental (Galanter *et al*., 2012). Por isso, em estudos com AIMs esperamos sempre que as frequências alélicas dentro de subpopulações de cada grupo parental sejam semelhantes e entre as populações sejam diferentes.

Para populações latino-americanas especificamente, não há um único conjunto de marcadores ótimos, porque a informatividade dos marcadores depende da combinação das frequências alélicas das populações parentais e da proporção de miscigenação nessas populações (Pfaff *et al*., 2004). Marcadores razoavelmente bons para latino-americanos são, então, aqueles que mostram uma frequência alélica bastante divergente entre populações africanas, europeias e ameríndias (Tabela 1).

Tabela 1: Exemplo de AIMs para cada uma das populações parentais e para estimar a miscigenação de populações da América Latina, com suas respectivas frequências alélicas.

	Frequência do AIM nas populações parentais		
	Africanos	**Europeus**	**Ameríndios**
AIM Africano	**1,0**	0,05	0,02
AIM Europeu	0,00	**0,90**	0,05
AIM Nativo-americano	0,02	0,01	**0,89**

O painel de Yeager *et al*, (2008) foi originalmente desenhado para caracterizar populações Afro-Americanas e possui um total de 106 SNPs. Nosso estudo, com base em trabalhos anteriores (Via *et al*., 2011; Avena *et al.,* 2012), utilizou o painel desenvolvido por Yeager *et al* (2008) para estimar a ancestralidade individual e populacional de quatro populações miscigenadas da América Latina, assim como inferir aspectos de sua dinâmica, visto que poucos estudos têm sido realizados com essas populações (por isso há pouca informação disponível), evidenciando que trata-se de um painel que fornece boas estimativas de ancestralidade europeia, africana e ameríndia.

2. OBJETIVOS

Objetivo Geral

Estimar quantitativamente a miscigenação individual de quatro populações da América Latina (EQU, LIM, MEX e NIC) e qualitativamente aspectos da dinâmica da miscigenação dessas populações.

Objetivos Específicos

a) Avaliar a eficiência do painel de 83 marcadores informativos de ancestralidade (AIMs) de Yeager *et al* (2008) para estimar a ancestralidade individual em populações da América Latina.

b) Fazer o controle de qualidade (CQ) e integrar três conjuntos de dados: (1) Minnesota, (2) RIBEF genotipados no CEGEN de Santiago de Compostela e (3) HapMap.

c) Comparar a diversidade intra e interpopulacional dos AIMs em populações europeias, africanas, nativo-americanas e latino-americanas miscigenadas.

d) Estimar a miscigenação populacional e individual em populações miscigenadas do Equador, México, Nicarágua e Peru.

e) Utilizar a distribuição da miscigenação individual e do padrão de desequilíbrio de ligação (DL) entre AIMs para inferir qualitativamente a distribuição dos eventos de miscigenação no tempo.

3. MATERIAL E MÉTODOS

3.1. Fontes de dados e amostras

Dados de 87 SNPs informativos de ancestralidade (AIMs) retirados do painel de Yeager *et al.* (2008) genotipados em indivíduos miscigenados de Managua, Nicarágua (NIC), miscigenados de Quito, Cuenca e Tulcan, Equador (EQU) e espanhóis de Extremadura (EXT) foram fornecidos pela Rede IberoAmericana de Farmacogenética (RIBEF), por intermédio de seu diretor, Dr. Adrián LLerena (Universidade de Extremadura, Badajoz, Espanha). Estas genotipagens foram realizadas no Centro Nacional de Genotipado da Espanha – Santiago de Compostela (CEGEN, www.cegen.org) utilizando a plataforma Sequenom iPLEX. Dados de populações peruanas foram provenientes de trabalhos anteriores do nosso grupo (Pereira *et al*., submetido à PLoS One), cuja genotipagem com a plataforma Sequenom iPLEX foi terceirizada na Universidade de Minnesota, EUA. Neste caso foram genotipados 106 marcadores, incluindo os 87 genotipados no CEGEN. As populações peruanas foram nativos americanos: Ashaninkas do departamento de Junin, Peru (HP); Quechuas do departamento de Puno, Peru (PUN); Shimaas do departamento de Cuzco, Peru (SHI), e a população miscigenada de Lima (LIM), composta por indivíduos controle do trabalho de Miscigenação e Câncer Gástrico do nosso grupo (Pereira *et al.,* submetido à PLoS One).

A genotipagem no CEGEN consistiu na técnica de "*MassArray*" da plataforma Sequenom iPLEX. A reação realizada foi do tipo iPLEX Gold. Cada ensaio deste processo de genotipagem consiste em uma amplificação dos fragmentos de DNA que contêm os SNPs de interesse por uma PCR *multiplex* e uma reação de discriminação alélica, onde todas as reações terminam com uma extensão de *primers* (*Single Base Extension*, SBE). Para a discriminação dos alelos, a reação incorpora *primers* com massa modificada e então é feita uma separação por espectrometria de massa entre produtos de SBE. Esta modificação permite genotipar até 36 SNPs por ensaio, com uma média de 24 SNPs. No caso da genotipagem de Minnesota, foi utilizada a mesma tecnologia de genotipagem. Houve quatro ensaios, três contendo 26 SNPs e um contendo 28 SNPs. Dois SNPs tiveram de ser removidos dos dados porque não possuíam genotipagens suficientes (< 95%) para comparação com as demais: rs30125 e rs888861. O SNP rs2592888 não apresenta dados no projeto Hapmap, por isso foi também excluído deste conjunto de dados, não participando também das análises deste estudo. O SNP rs1990743 se encontrava nos dados do CEGEN mas não se encontrava na genotipagem dos nativos. Portanto, esse SNP foi também excluído da análise, finalizando nossos dados com um total de 83 SNPs dos 87 recebidos.

Além destes dados, inserimos seis populações do Projeto Internacional HapMap (The International HapMap Consortium, 2010, www.hapmap.org) para efeitos de comparação, através de downloads e estudos anteriores do laboratório (Pereira et al., submetido à PLoS One): Residentes de Utah com ancestralidade europeia da coleção do CEPH - *Centre d'Etude Du Polymorphisme Humain* - (CEU); Luhya em Webuye, Quênia (LWK); Descendentes de mexicanos em Los Angeles, Califórnia (MEX); Maasai em Kinyawa, Quênia (MKK); Toscanos da Itália (TSI) e; Yoruba em Ibadan, Nigéria (YRI). Houve uma filtragem de indivíduos das populações CEU e YRI, que são constituídas por trios, visando a permanência apenas de indivíduos parentais na análise. Portanto, indivíduos aparentados foram retirados para evitar a presença de alelos idênticos por descendência recente. Dados dos tamanhos amostrais e número de marcadores analisados para cada população se encontram na Tabela 2, e dados detalhados dos marcadores utilizados são apresentados no Anexo 1.

Tabela 2: Tamanho amostral de todas as populações utilizadas neste estudo e número de SNPs efetivamente genotipados para cada população.

População	**CEU**	**TSI**	**EXT**	**LWK**	**MKK**	**YRI**	**HP**	**SHI**	**PUN**	**MEX**	**EQU**	**NIC**	**LIM**
Número de indivíduos	60	88	322	90	171	209	186	87	23	77	227	123	294
Número de SNPs genotipados	83	53	83	53	52	81	83	83	83	51	83	83	81

3.2. Controle de Qualidade dos dados

A consolidação dos arquivos HapMap, CEGEN e Minnesota foi feita de forma manual, assim como a exclusão dos 16 SNPs que estavam no arquivo de Minnesota e não estavam no arquivo do CEGEN. Após a junção, foi feito um controle de qualidade manual para verificar se os dados de genótipos dos indivíduos correspondiam aos dados da fonte primária de dados. Dez *loci* de seis indivíduos foram selecionados aleatoriamente no arquivo processado e nos arquivos originais, para comparação. Como nenhuma diferença foi observada, constatou-se que os dados eram igualáveis e não interfeririam na qualidade dos resultados.

Alguns SNPs oriundos do HapMap apresentaram os dados em relação à fita de DNA reversa com relação às genotipagens e/ou outras populações do mesmo projeto (por exemplo, TT ao invés de AA). Este é um problema comum quando se integram dados de diferentes fontes. Embora este problema seja facilmente identificável quando aparecem quatro alelos em um loco, é mais difícil de identificar quando se trata de polimorfismos C/G ou A/T. Nestes casos, tivemos que identificar a fita de referência de cada SNP e padronizar os dados para que todas as leituras sejam feitas na mesma fita de referência. Isto ocorreu predominantemente na população YRI.

O arquivo completo gerado em formato SDAT (indivíduos em linhas e marcadores em colunas) teve que ser transformado diversas vezes para os usos em diferentes *softwares*. Para estas transformações utilizamos principalmente o DivergenomeTools (Magalhães *et al*., 2012), o módulo *transform* do programa GLU (http://code.google.com/p/glu-genetics/) e o programa PLINK (Purcell *et al.,* 2007).

Um controle de qualidade (CQ) inicial dos dados foi feito avaliando o Equilíbrio de Hardy-Weinberg (EHW) com o módulo QC do programa GLU (http://code.google.com/p/glu-genetics/), que gera um valor de *p* (a significância sob o modelo de EHW, obtida pelo teste qui-quadrado (X^2), para cada marcador estudado; Anexo 2).

Outro controle de qualidade foi feito com o programa Haploview (Barret *et al.,* 2005), onde observamos a porcentagem de indivíduos genotipados com sucesso para cada SNP do estudo.

Oito indivíduos foram excluídos deste conjunto de dados porque não apresentavam dados para nenhum dos SNPs.

3.3. Frequências alélicas

A plataforma R (http://www.r-project.org) é um ambiente e uma linguagem de programação desenvolvida principalmente para análises estatísticas, incluindo a visualização de gráficos em alta resolução. Todos os gráficos deste trabalho foram gerados com a plataforma R. Outro fato interessante sobre esta plataforma é que ela é amplamente expansível, devido à sua simples e bem desenvolvida linguagem. Os pacotes de expansão do R utilizados neste estudo foram: Adegenet (Jombart & Solymos, 2008), Ade4 (Chessel *et al*. 2004) e Hierfstat (Goudet, 2005). Os scripts utilizados em linguagem R para as análises estatísticos estão disponíveis como anexos, em um formato padrão do LDGH, para garantir a reprodutibilidade das análises.

A tabela de freqüências alélicas por população foi calculada com o pacote Adegenet (Jombart & Solymos, 2008) da plataforma R (Anexo 3) para nossos dados e a para os genótipos do Projeto HapMap. Um controle de qualidade manual foi feito, escolhendo nove SNPs aleatórios e comparando as freqüências alélicas das populações obtidas do HapMap com as estimadas pelo Adegenet, e nenhuma ou mínima diferença foi encontrada.

3.4. Heterozigosidade esperada (*Hs*)

A diversidade intrapopulacional foi medida através de cálculos de heterozigosidade esperada sob EHW por população, realizada com o pacote Adegenet (Jombart & Solymos, 2008, Anexo 3). As heterozigosidades esperadas por cada SNP foram calculadas com o programa Haploview (Barret *et al*., 2005).

A heterozigosidade esperada no EHW é *2pq*, e, portanto, necessita das freqüências alélicas *p* e *q*. No entanto, o cálculo do Adegenet (Jombart & Solymos, 2008) é um pouco mais sofisticado porque considera a possibilidade (não apresentada no nosso estudo) que os *loci* apresentem mais de dois alelos. Para uma determinada população:

$$Hs = \frac{1}{K}\sum_{k=1}^{K}(1 - \sum_{i=1}^{m(k)} f_i^2)$$

Onde:

- ⇒ *K* = Número de *loci*
- ⇒ *k* = Um determinado *locus*
- ⇒ *m(k)* = Número de alelos de um *locus k*
- ⇒ f_i = Frequência alélica do alelo *i* em uma população.

3.5. Análise Molecular de Variância (AMOVA)

A AMOVA (Excoffier *et al.* 1992) estima a repartição da variância genética em vários componentes hierárquicos, a partir de um enfoque similar ao das análises da variância. A partir deste enfoque são calculadas estatísticas F que estimam os componentes de variância nos diferentes níveis hierárquicos, o que foi realizado com o pacote Hierfstat do programa R (Anexo 4, Goudet, 2005; De Meeûs e Goudet, 2007). Os valores foram calculados por *locus*, de acordo com o algoritmo de Yang (1998), onde:

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

Portanto, para as estatísticas F entre dois níveis *j* e *i*, temos:

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

Para a AMOVA deste trabalho, utilizamos três níveis hierárquicos: (1) indivíduos; (2) populações; (3) continentes. Portanto, os valores calculados na análise molecular de variância foram: *Fct* (F variância entre continente em relação ao total), *Fst* (F variância entre populações em relação a variância total) e *Fsc* (F variância entre populações em relação a cada continente), sendo que, seguindo a fórmula anterior:

$$F_{CT} = \frac{\sigma_a^2}{\sigma_T^2} \qquad F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} \qquad F_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}$$

σ_a^2 é a variância das frequências alélicas entre continentes;

σ_b^2 é a variância das frequências alélicas entre populações;

σ_c^2 é a variância das frequências alélicas entre indivíduos de uma população;

σ_T^2 é a variância total das frequências alélicas entre indivíduos de todos os grupos.

3.6. *Fst* das populações par a par

O cálculo de *Fst* como medida de distância entre pares de populações foi realizado no pacote Adegenet (Jombart & Solymos, 2008) da plataforma R (Anexo 3). Ele utiliza como base a *Hs* das populações em seu valor médio conjuntamente para todos os *loci* dos dados:

$$Fst(A,B) = \frac{(Ht - \frac{n_A Hs(A) + n_B Hs(B)}{n_A + n_B})}{Ht}$$

Onde:

- $\Rightarrow$ n_A = Número de indivíduos da população A;
- $\Rightarrow$ n_B = Número de indivíduos da população B;
- $\Rightarrow$ *Hs(A)* = Heterozigosidade esperada da população A;
- $\Rightarrow$ *Hs(B)* = Heterozigosidade esperada da população B;
- $\Rightarrow$ *Ht* = Heterozigosidade das populações A e B.

3.7. Informatividade dos AIMs (*In*)

O *In* é um índice que sugere a informatividade adicionada por cada marcador para a atribuição de um indivíduo a uma determinada população. Este índice auxilia a determinar quanta informação de marcadores é necessária para estimar uma ancestralidade individual. Possui uma extensão *multilocus* e permite a medida de contribuição de alelos específicos ou populações (Rosenberg, 2003). O *In* assume um modelo sem miscigenação, portanto ele foi realizado somente entre as populações parentais. Desta forma, este índice é tido como:

$$I_n(Q;J) = \sum_{j=1}^{N}(-p_j \log p_j + \sum_{i=1}^{K}\frac{p_{ij}}{K}\log p_{ij})$$

Onde:

Q = Indivíduo de uma população aleatória;

J = Genótipo aleatório de um determinado *locus*;

N = Número de alelos;

j = Alelo específico que está sendo calculado;

p = Frequência de *j*;

K = Número de populações;

i = População específica que está sendo testada.

De modo que *In* = 0 significa que todas as populações possuem frequências iguais para todos os alelos testados, é o valor mínimo. O valor máximo de *In* é o número de populações (K), onde todos os alelos terão frequências específicas para cada população, ou seja, cada população terá seu alelo exclusivo.

Os cálculos de *In* para cada SNP estudado foram realizados pelo programa Infocalc (Rosenberg, 2003), diretamente através dos dados de genotipagem das populações parentais (AFR, EUR e NAT).

3.8. Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística de redução da dimensionalidade que transforma as variáveis originais em novas variáveis (componentes principais), fazendo com que a primeira nova variável, ou componente, seja responsável pela maior parte da variação encontrada no conjunto de dados, a segunda variável seja responsável pela segunda maior variação encontrada, e assim sucessivamente, até que toda a variação dos dados seja elucidada com a propriedade que os componentes são independentes. A PCA identifica padrões ocultos nas variáveis originais, neste caso a partir dos genótipos dos indivíduos (Cavalli-Sforza, 1998).

Esta análise nos permite visualizar claramente os componentes de miscigenação para cada população, a nível individual, assim como a distribuição no espaço da variação de ancestralidade para todos indivíduos de cada população.

A PCA foi realizada com os pacotes Adegenet (Jombart & Solymos, 2008) e ade4 (Chessel *et al*., 2004) da plataforma R (Anexo 3).

3.9. Estimativas de Miscigenação

A miscigenação individual foi estimada pelo método de *Maximum Likelihood* (ML), utilizando o algoritmo *Expectation Maximization* (EM), implementado no programa Admixture (Alexander *et al*., 2009), que não considera DL como outros Softwares como Structure (Pritchard *et al.,* 2000), e por isso sua análise computacional demanda menos tempo. A miscigenação populacional foi obtida calculando a média da miscigenação individual para cada população. Os gráficos de barras verticais, no qual cada barra corresponde a um indivíduo com seus componentes de ancestralidade, e gráficos de distribuição da miscigenação, foram gerados na plataforma R (Anexos 5 e 6).

O método de ML é um procedimento estatístico amplamente utilizado em diversas áreas biológicas. Ele consiste em escolher um modelo probabilístico apropriado para ser atribuído aos dados observados, e em estimar os parâmetros desse modelo, o que é feito com a técnica de *Expectation Maximization* (EM, Do e Batzoglou, 2008).

No caso deste estudo, no qual estamos interessados em estimar os principais componentes de ancestralidade continental da população latino-americana, assumimos que as populações miscigenadas são produto de *k = 3 clusters* ancestrais (europeu, ameríndio ou africano).

3.10. Desequilíbrio de Ligação

O DL foi calculado no programa Haploview (Barret *et al*., 2005), com métodos de estimativas (ML + EM) para inferir a fase dos haplótipos, como já detalhado na Introdução.

Em resumo, as análises deste trabalho podem ser conferidas na Figura 8.

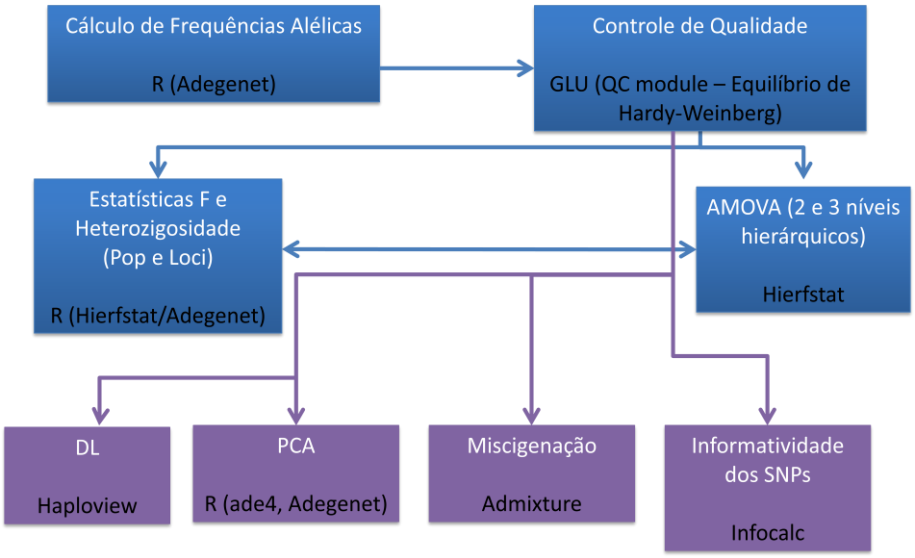

Figura 8: Fluxograma das análises realizadas neste estudo, com seus respectivos *softwares*. Adaptado do Exame de Qualificação do aluno de Doutorado em Bioinformática Giordano Soares-Souza.

4. RESULTADOS E DISCUSSÃO

4.1. Controle de qualidade dos dados

Os resultados da genotipagem dos AIMs se encontravam filtrados pelo controle de qualidade interno do CEGEN - Santiago de Compostela, que obteve resultados para 87 SNPs dos 106 do painel de Yeager *et al*. (2008). Esta perda de 19 SNPs não está associada nem a SNPs genotipados em um ensaio de genotipagem *multiplex* específico, nem à informatividade de uma ancestralidade específica, como mostra a Tabela 3.

Tabela 3: Diferença das frequências alélicas (δ) entre pares de populações parentais para os SNPs do painel de Yeager *et al*. (2008) não incluídos nos dados do CEGEN - Santiago de Compostela .

	AFR-EUR	AFR-NAT	EUR-NAT *
rs10491654	0,245	0,132	0,377
rs10498255	0,498	0,734	0,236
rs10501474	0,575	0,799	0,224
rs10506816	0,691	0,681	0,010
rs1353251	0,196	0,719	0,523
rs1395771	0,611	0,063	0,674
rs1470524	0,589	0,161	0,428
rs1984473	0,449	0,758	0,309
rs2042762	0,033	0,742	0,709
rs257748	0,308	0,310	0,618
rs2595456	0,115	0,314	0,429
rs2817611	0,675	0,705	0,030
rs4852696	0,745	0,671	0,074
rs4934436	0,090	0,366	0,456
rs6883095	0,380	0,819	0,439
rs6911727	0,359	0,709	0,350
rs798887	0,069	0,69	0,759
rs9292118	0,363	0,619	0,256
rs9325872	0,474	0,169	0,643
rs948360	0,527	0,655	0,128
rs304051	0,358	0,691	0,333

* **EUR**: CEU + TSI; **NAT**: HP + SHI + PUN; **AFR**: YRI + MKK + LWK.

Os dados de genotipagem processados em Minnesota, CEGEN e os dados públicos HapMap foram transformados em um arquivo de dados congelados que foi depositado na plataforma bioinformática *DIVERGENOMEdb* (Magalhães *et al*. 2012), desenvolvida pelo nosso grupo. Um controle de qualidade manual foi realizado com estes dados para verificar se os genótipos correspondiam aos da fonte original de dados. Dez *loci* e seis indivíduos aleatórios foram selecionados ao longo do arquivo, e duas tabelas foram montadas, com os dados da fonte primária e com os dados processados. Os dados foram 100% coincidentes, e, portanto, alterações no processamento dos dados que pudessem comprometer a qualidade dos resultados do estudo são muito pouco prováveis.

Entre as populações miscigenadas consideradas neste estudo, os mexicanos residentes na Califórnia (MEX), extraídos do projeto HapMap tiveram 31 (37%) SNPs faltantes em relação às outras populações. Entre as populações parentais, os Maasai (MKK)

do projeto HapMap também apresentaram este problema, com 29 dos 83 SNPs sem genotipagem (35%). De qualquer forma, optamos por manter estas duas populações no estudo pois são importantes na análise final dos dados.

Considerando unicamente os dados genotipados em Minnesota e pelo CEGEN, dois SNPs - rs10510791 e rs1398829 - possuem menos de 95% dos indivíduos genotipados e 81 SNPs (97,5%) obtiveram uma genotipagem para mais de 95% dos indivíduos. Foram obtidas 101.666 (97%) genotipagens das 104.746 (100%) esperadas de acordo com as populações genotipadas.

4.2. Diferenciação interpopulacional

4.2.1. Frequências alélicas

Os valores apresentados na tabela de frequências alélicas (Tabela 4) foram, em geral, concordantes tanto com os valores de *Fct* como com os valores de *In* (Tabela 6), com raras exceções. Dos 83 SNPs do estudo, 22 (realçados em cinza escuro) não apresentaram diferenças contrastantes entre as frequências alélicas das populações de diferentes continentes. Os SNPs rs10515535, rs1498991, rs1934393, rs2840290 e rs6569792 foram concordantes com índices de baixa informatividade como mostrados na Tabela 6 (*In* e *Fct*). Isso significa que estes SNPs são os que apresentam menor contribuição para as análises de miscigenação com o presente conjunto de dados.

A população de PUN apresentou 14 valores de frequências alélicas diferentes das demais populações nativo-americanas, HP e SHI (valores realçados em negrito). Isso pode ter ocorrido devido ao baixo valor amostral para estas populações (n = 22), em comparação a HP e SHI, que possuem valores amostrais maiores (n = 185 e n = 86, respectivamente). Outra explicação possível é um maior nível de miscigenação em PUN.

Tabela 4: Frequências alélicas de cada um dos 83 SNPs para cada uma das populações estudadas.

rs	alelo	Europeus			Africanos			Ameríndios			Miscigenados			
		CEU	TSI	EXT	LWK	MKK	YRI	HP	SHI	PUN	MEX	LIM	EQU	NIC
1004704	A/G	0,225	0,159	0,191	0,139	0,111	0,093	0,742	0,799	0,804	0,584	0,641	0,522	0,407
10131076	G/A	0,125	0,136	0,123	0,722	0,482	0,603	0,003	0,011	0,000	0,136	0,068	0,066	0,142
1013459	A/G	0,942	0,841	0,859	0,294	0,424	0,256	1,000	1,000	1,000	0,877	0,944	0,914	0,866
10214949	A/G	0,858	0,841	0,843	0,417	0,553	0,431	0,973	0,989	0,891	0,942	0,935	0,892	0,854
10248051	C/T	0,717	0,784	0,710	0,144	0,333	0,127	0,610	0,557	0,413	0,591	0,503	0,555	0,480
1036543	C/T	0,058	0,057	0,064	0,678	0,494	0,641	0,718	0,897	0,674	0,429	0,634	0,617	0,419
10484578	G/A	0,317	0,347	0,436	0,350	0,348	0,622	0,094	0,057	0,130	0,351	0,328	0,328	0,394
10486576	G/A	0,308	0,483	0,099	0,483	0,485	0,486	0,796	0,868	0,783	0,487	0,651	0,496	0,366
10488172	G/T	0,183	0,170	0,189	0,006	0,003	0,246	0,935	0,954	0,761	0,448	0,764	0,553	0,398
10491097	A/G	0,333	0,199	0,315	0,978	0,842	0,971	0,836	0,885	0,674	0,532	0,636	0,619	0,630
10492585	A/G	0,083	0,142	0,090	0,144	0,143	0,502	0,040	0,023	0,000	0,143	0,078	0,137	0,142
10497705	C/T	0,450	0,318	0,325	0,233	0,281	0,158	0,987	0,977	0,978	0,500	0,767	0,678	0,577
10498919	G/C	0,333	0,670	0,998	0,672	0,670	0,383	0,522	0,615	0,435	0,669	0,670	0,678	0,760
10500505	A/T	0,308	0,398	0,205	0,400	0,401	0,273	0,548	0,534	0,587	0,396	0,541	0,467	0,354
10507688	A/G	0,767	0,659	0,891	0,661	0,658	0,806	0,516	0,402	0,370	0,656	0,583	0,529	0,626
10508349	A/G	0,142	0,284	NA	0,011	0,009	0,026	0,527	0,649	0,783	0,364	0,585	0,471	0,248
10510791	G/C	0,525	0,426	0,615	0,428	0,427	0,593	0,231	0,420	0,109	0,429	0,262	0,372	0,455
10515535	G/A	0,483	0,392	0,441	0,506	0,819	0,701	0,487	0,506	0,370	0,377	0,422	0,432	0,467
10515919	G/A	0,800	0,869	0,860	0,850	0,865	0,847	0,040	0,017	**0,152**	0,552	0,347	0,456	0,715

10517518	A/G	0,125	0,045	0,048	0,106	0,061	0,144	0,895	0,862	0,870	0,403	0,670	0,577	0,370
10519979	A/G	0,617	0,722	0,506	0,722	0,722	0,481	0,997	1,000	0,935	0,721	0,859	0,773	0,667
10520440	T/G	0,675	0,716	0,705	0,939	0,904	0,947	0,328	0,391	0,348	0,630	0,432	0,502	0,577
10520678	C/T	0,717	0,750	0,741	0,278	0,371	0,182	1,000	0,989	1,000	0,786	0,883	0,892	0,780
1073319	A/G	0,400	0,563	0,205	0,561	0,561	0,352	0,995	0,989	0,870	0,558	0,682	0,604	0,537
12953952	G/A	0,067	0,108	0,076	0,106	0,105	0,457	0,000	0,000	0,000	0,104	0,046	0,040	0,130
138022	G/A	0,275	0,273	0,219	0,800	0,646	0,854	0,005	0,017	0,000	0,182	0,068	0,126	0,175
1397618	A/T	0,467	0,909	0,939	0,911	0,912	0,730	0,992	0,994	1,000	0,909	0,973	0,954	0,947
1398829	A/T	0,958	0,983	0,964	0,422	0,602	0,297	1,000	0,994	1,000	0,929	0,973	0,974	0,890
1451928	G/T	0,233	0,125	0,166	0,083	0,129	0,098	0,637	0,540	0,891	0,351	0,668	0,564	0,459
1477277	C/G	0,525	0,750	0,652	0,750	0,749	0,844	0,901	0,833	0,957	0,747	0,794	0,698	0,593
1498991	G/C	0,075	0,097	0,935	0,089	0,058	0,158	0,269	0,190	**0,043**	0,442	0,393	0,467	0,659
1517634	G/A	0,733	0,733	0,761	0,833	0,749	0,900	0,081	0,132	0,065	0,429	0,265	0,346	0,520
153898	C/T	0,300	0,256	0,311	0,883	0,684	0,911	0,887	0,966	0,957	0,558	0,793	0,689	0,671
1898280	G/A	0,842	0,875	0,752	0,106	0,237	0,151	0,038	0,006	**0,239**	0,584	0,400	0,465	0,492
1919550	T/A	0,742	0,517	0,918	0,517	0,518	0,725	0,078	0,155	0,109	0,519	0,315	0,434	0,638
1934393	G/C	0,800	0,716	0,755	0,233	0,456	0,206	0,465	0,592	**0,130**	0,558	0,372	0,399	0,520
1990745	T/C	0,833	0,813	0,868	0,994	0,959	0,940	0,226	0,305	**0,043**	0,565	0,337	0,471	0,638
2035573	C/T	0,375	0,483	0,248	0,483	0,485	0,301	0,769	0,759	0,804	0,487	0,595	0,524	0,459
2208139	C/T	0,408	0,386	0,376	0,133	0,178	0,165	0,978	0,764	0,891	0,591	0,844	0,729	0,557
2253624	G/T	0,950	0,920	0,964	0,922	0,921	0,921	0,844	0,977	0,978	0,922	0,906	0,923	0,902
2296274	A/G	0,767	0,744	0,781	0,078	0,222	0,033	0,970	0,994	0,935	0,825	0,874	0,861	0,789
249847	C/T	0,442	0,335	0,388	0,067	0,135	0,103	0,992	1,000	0,957	0,649	0,777	0,729	0,593
2569029	T/G	0,683	0,659	0,703	0,856	0,757	0,842	0,126	0,063	**0,261**	0,565	0,277	0,396	0,431
2585901	T/C	0,608	0,409	0,859	0,411	0,412	0,524	0,005	0,006	0,043	0,409	0,231	0,341	0,488
2711070	G/C	0,525	0,500	0,514	0,144	0,249	0,211	0,003	0,000	0,087	0,617	0,298	0,247	0,407
2785279	T/C	0,133	0,205	0,193	0,200	0,202	0,490	0,134	0,287	**0,022**	0,201	0,107	0,137	0,195
2829454	G/A	0,592	0,460	0,696	0,461	0,462	0,670	0,228	0,132	0,174	0,461	0,321	0,372	0,593
2840290	C/T	0,417	0,597	0,335	0,600	0,599	0,701	0,718	0,736	0,674	0,597	0,721	0,637	0,545
304051	T/C	0,317	0,364	0,377	0,717	0,558	0,801	0,005	0,023	0,022	0,123	0,124	0,170	0,264
354747	A/G	0,625	0,528	0,602	0,067	0,170	0,081	0,898	0,868	**0,630**	0,597	0,718	0,652	0,545
3768176	C/G	0,375	0,585	0,890	0,583	0,582	0,416	0,409	0,500	0,478	0,584	0,554	0,553	0,626
3806218	G/A	0,367	0,313	0,380	0,889	0,751	0,914	0,075	0,259	**0,109**	0,260	0,143	0,207	0,341
3828121	T/C	0,133	0,125	0,124	0,017	0,009	0,167	0,591	0,805	0,652	0,253	0,432	0,467	0,276
3860446	T/C	0,717	0,705	0,711	0,022	0,149	0,024	0,019	0,000	0,043	0,331	0,170	0,262	0,370
4013967	C/T	0,800	0,824	0,694	0,822	0,822	0,596	0,992	0,989	1,000	0,825	0,929	0,885	0,768
4034627	C/T	0,900	0,920	0,950	0,472	0,453	0,292	0,944	0,983	0,935	0,890	0,918	0,943	0,886
4076700	T/C	0,192	0,205	0,197	0,733	0,594	0,751	0,005	0,006	0,000	0,136	0,078	0,110	0,187
4130405	C/A	0,850	0,761	0,846	0,606	0,933	0,775	0,406	0,241	0,239	0,617	0,306	0,487	0,537
4130513	C/T	0,900	0,847	0,930	0,844	0,845	0,565	0,970	0,994	**0,804**	0,844	0,810	0,857	0,833
4625554	G/A	0,692	0,733	0,761	0,717	0,576	0,778	0,040	0,057	0,087	0,500	0,226	0,315	0,415
4657449	A/G	0,108	0,114	0,096	0,128	0,164	0,120	0,906	0,902	0,804	0,429	0,714	0,575	0,382
4733652	T/C	0,625	0,500	0,871	0,500	0,500	0,651	0,169	0,236	0,174	0,500	0,332	0,432	0,500
4762106	G/A	0,808	0,841	0,806	0,144	0,254	0,194	0,185	0,006	**0,304**	0,617	0,444	0,436	0,561
5000507	A/T	0,750	0,676	0,685	0,178	0,401	0,127	0,992	0,977	1,000	0,734	0,913	0,848	0,740
567992	A/G	0,133	0,131	0,152	0,006	0,058	0,213	0,927	0,891	0,630	0,435	0,565	0,469	0,301
6569792	A/G	0,800	0,693	0,666	0,139	0,181	0,096	0,495	0,534	0,652	0,649	0,672	0,612	0,553
6684063	G/T	0,200	0,182	0,196	0,789	0,544	0,730	0,962	0,960	**0,783**	0,584	0,714	0,608	0,516
6804094	T/A	0,525	0,523	0,666	0,522	0,520	0,720	0,505	0,305	0,304	0,519	0,379	0,410	0,553
708915	A/T	0,342	0,506	0,203	0,511	0,509	0,615	0,782	0,851	0,652	0,506	0,532	0,513	0,459
719776	C/G	0,200	0,102	0,843	0,961	0,778	0,967	0,634	0,770	0,804	0,208	0,753	0,767	0,683
7463344	G/C	0,025	0,051	0,016	0,050	0,050	0,237	0,000	0,000	0,000	0,052	0,014	0,026	0,077
7535375	C/T	0,550	0,574	0,570	0,167	0,257	0,120	0,801	0,776	0,913	0,708	0,736	0,749	0,720
802524	T/C	0,025	0,057	0,084	0,767	0,637	0,797	0,000	0,006	0,000	0,084	0,037	0,051	0,142
842634	C/T	0,675	0,665	0,801	0,667	0,667	0,801	0,454	0,460	0,717	0,669	0,626	0,639	0,695
868179	G/A	0,100	0,091	0,110	0,794	0,734	0,706	0,000	0,000	0,000	0,091	0,060	0,070	0,163
879780	T/C	0,092	0,074	0,141	0,650	0,614	0,828	0,005	0,011	0,000	0,162	0,054	0,066	0,130
9295316	T/A	0,675	0,489	0,908	0,489	0,488	0,701	0,008	0,006	**0,239**	0,487	0,282	0,427	0,654
9302185	T/C	0,208	0,176	0,185	0,844	0,681	0,856	0,030	0,023	0,022	0,097	0,073	0,104	0,203
9307613	T/A	0,383	0,290	0,408	0,289	0,287	0,533	0,056	0,172	0,065	0,286	0,160	0,251	0,341
9310888	A/G	0,908	0,983	0,935	0,500	0,629	0,471	0,995	1,000	0,978	0,942	0,961	0,952	0,915
9320808	A/G	0,192	0,182	0,158	0,861	0,725	0,873	0,901	0,954	**0,696**	0,532	0,706	0,643	0,569
9323178	A/G	0,558	0,455	0,495	0,117	0,281	0,136	0,989	0,954	0,913	0,656	0,786	0,813	0,707
993314	C/T	0,208	0,227	0,222	0,811	0,787	0,749	0,806	0,897	0,630	0,494	0,628	0,593	0,569

4.2.2. Análise de *Fst* por população

Na Tabela 5, que mostra os *Fst* par a par para as 13 populações estudadas, com base nos 83 AIMs incluídos no estudo, células com cores claras representam variação

intracontinental e células com coloração escura representam variação intercontinental. Os valores foram condizentes com a principal premissa de AIMs, com variação intercontinental maior que a variação intracontinental. Entretanto, chamamos a atenção para o baixo valor de diferenciação entre PUN (n = 22) e EXT (n = 321), realçado em negrito. Este valor, apesar de relativamente baixo, é maior que os valores de *Fst* intracontinental de ambas as populações. Ele pode ser consequência do baixo valor amostral de PUN, além de sugerir que os indivíduos dessa população possuem um componente europeu na sua ancestralidade relativamente maior que as outras populações nativas americanas. Bryc *et al* (2010) sugerem que os tamanhos amostrais diferenciados entre as populações podem gerar um efeito de confusão na análise de *Fst*.

Tabela 5: *Fst* par a par para as 13 populações estudadas, com base nos 83 AIMs incluídos no estudo. Realce em cinza claro indica diferenciação intracontinental e realce em cinza escuro representa variação intercontinental.

	CEU	**TSI**	**EXT**	**LWK**	**MKK**	**YRI**	**HP**	**SHI**	**PUN**	**MEX**	**EQU**	**NIC**
TSI	0,016											
EXT	0,030	0,045										
LWK	0,194	0,191	0,157									
MKK	0,101	0,109	0,139	0,018								
YRI	0,152	0,180	0,206	0,030	0,047							
HP	0,234	0,250	0,293	0,343	0,298	0,366						
SHI	0,283	0,272	0,229	0,355	0,258	0,305	0,014					
PUN	0,187	0,150	**0,085**	0,232	0,122	0,144	0,014	0,033				
MEX	0,044	0,041	0,056	0,174	0,102	0,160	0,127	0,147	0,076			
EQU	0,061	0,071	0,113	0,158	0,138	0,207	0,068	0,057	0,016	0,018		
NIC	0,047	0,054	0,048	0,142	0,096	0,151	0,137	0,133	0,056	0,016	0,015	
LIM	0,079	0,094	0,169	0,176	0,172	0,243	0,041	0,036	0,007	0,029	0,010	0,036

No contexto das populações europeias, é interessante notar que, a partir do observado na tabela de frequências alélicas (Tabela 4), os EXT possuem algumas frequências alélicas diferentes de CEU e TSI. Condizentemente, dentro dos valores de *Fst* da Europa, percebemos que os EXT são mais diferenciados de CEU e TSI do que estas duas últimas populações entre si.

Destacamos os valores altos entre EQU x EXT e LIM x EXT e baixos entre essas populações miscigenadas e as duas outras populações europeias do HapMap (CEU e TSI).

4.2.3. Análise de Componentes Principais (PCA)

A Análise de Componentes Principais revelou dois componentes responsáveis por 42,7% da variabilidade total dos nossos dados (Figura 9).

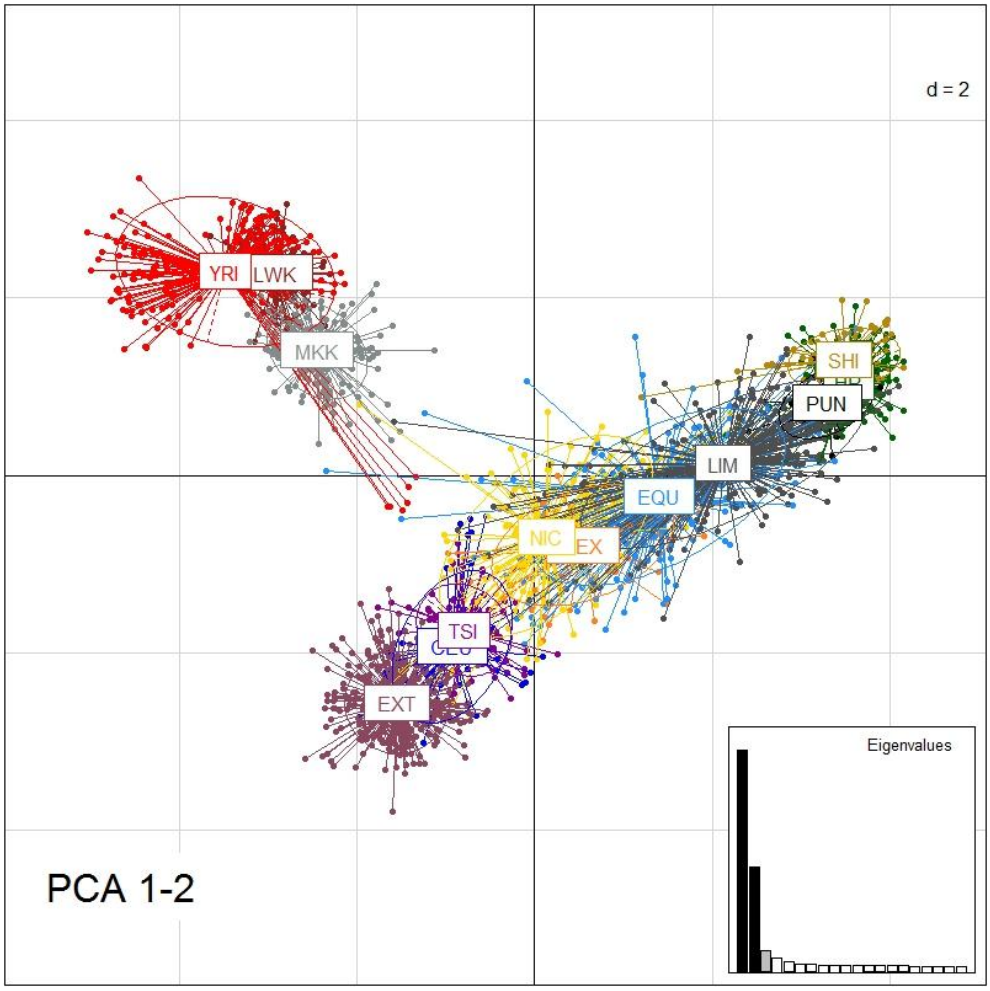

Figura 9: Análise de Componentes Principais a partir dos dados de genótipos dos indivíduos das 13 populações estudadas. O primeiro (abscissa) e o segundo (ordenada) componente principal são mostrados, e representam respectivamente 28,9% e 13,8% da variância total. Siglas: CEU (azul escuro): Residentes de Utah com ancestralidade europeia da coleção do CEPH; LWK (marrom): Luhya em Webuye, Quênia; MEX (laranja): Mexicanos residentes em Los Angeles, Califórnia; MKK (cinza claro): Maasai em Kinyawa, Quênia; TSI (roxo): Toscanos da Itália; YRI (vermelho): Yoruba em Ibadan, Nigéria; EXT (vinho): Espanhóis de Extremadura, Espanha; EQU (azul claro): Miscigenados de Quito, Cuenca e Tulcan, Equador; LIM (cinza escuro): Miscigenados de Lima, Peru; NIC (amarelo): Miscigenados de Manágua, Nicarágua; HP (verde escuro): Ashaninkas do departamento de Junin, Peru; PUN (preto): Quechuas do departamento de Puno, Peru; SHI (verde musgo): Shimaas do departamento de Cuzco, Peru.

Na Figura 9 se percebe dentro das populações miscigenadas (NIC, MEX, EQU, LIM) um amplo gradiente de miscigenação individual entre europeus e nativos americanos, com alguns indivíduos de EQU e NIC com alto componente europeu.

O primeiro componente principal, associado à maior variabilidade, explica 28,9% da variação encontrada nas populações, sendo responsável principalmente pela diferença entre africanos e ameríndios (eixo horizontal). Esta distribuição da variabilidade genética no primeiro componente principal é coerente com os valores de *Fst* (Tabela 5), que mostram a máxima diferenciação ocorrendo entre africanos e ameríndios. O segundo componente

principal explica 13,8% da variância genética total. Este componente está associado à diferenciação entre europeus e africanos (eixo vertical).

Assim como mostraram Silva-Zolezzi *et al.* (2009), a análise de PCA apresentou uma maior distância genética entre os africanos e o restante dos grupos (componente 1 e 2). Yeager *et al.* (2008) mostra que a ancestralidade de diferentes populações pode ser bem evidenciada tanto pelo método de *Maximum Likelihood* (ML) para estimar a ancestralidade, como pelo PCA, o que percebemos também em nosso estudo.

4.3. Diversidade populacional dos AIMs

Como já abordado na introdução desta dissertação, Galanter *et al.* (2012) sugere que um painel com 88 AIMs possui informatividade similar a painéis de 194 ou 314 AIMs, com os três painéis fornecendo boas estimativas de ancestralidade individual. Este resultado sugere que o painel utilizado nesse estudo é suficiente para inferir a ancestralidade individual nas populações latino-americanas amostradas.

AIMs a serem utilizados para estimar a miscigenação africana, europeia e nativo-americana na América Latina devem ter a seguinte propriedade: As populações dentro dos continentes devem ter frequências alélicas semelhantes e baixa heterozigosidade interna, e populações de diferentes continentes devem ter freqüências alélicas bem diferentes. Neste estudo, no contexto da Análise Molecular de Variância genética (AMOVA, Excoffier et al. 1992), quanto maior a variância genética entre continentes, medida pelo *Fct* e menor a variância dentro dos continentes, medida pelo *Fsc*, melhores são os marcadores para diferenciar populações estruturadas. Além desta análise foi calculado o índice *In* de Rosenberg (2003) para verificar a informatividade dos SNPs do painel de ancestralidade. As análises populacionais consideram os 87 AIMs conjuntamente para permitir uma avaliação global do painel de miscigenação utilizado. As análises que consideram separadamente os SNPs, são relevantes para identificar o poder de informatividade da ancestralidade de cada SNP, sugerindo quais SNPs devem ser conservados ou substituídos na preparação de novos painéis de AIMs. Estas análises são relevantes porque o painel utilizado, apresentado por Yeager *et al.* (2008), foi avaliado com base em um menor número de indivíduos e populações. Atualmente, neste trabalho, nós temos uma maior quantidade de dados que nos permite uma melhor análise da informatividade deste painel.

A Tabela 6 mostra o índice *In* para as populações parentais, juntamente com os valores da AMOVA (*Fct, Fst, Fsc*) para melhor comparação. Valores de *In* realçados em cinza indicam os 43 (50%) SNPs com maior informatividade em cada coluna. Os valores de *Fct* destacados são maiores que 0,12, o valor médio observado para populações humanas de diferentes continentes para SNPs escolhidos ao acaso (Barbujani e Colonna, 2009). Os

SNPs realçados em cinza indicam que obtiveram pelo menos dois valores de *In* altos (em cinza) e *Fct* maior que 0,12.

Tabela 6: Os 83 SNPs utilizados no estudo com seus respectivos valores de *In*, *Fct* (F continente/total), *Fst* (F subpopulação/total) e *Fsc* (F subpopulação/continente). Valores de *In* realçados em cinza correspondem aos 43 (50%) maiores valores de cada coluna. Valores de *Fct* realçados são maiores que 0,12. SNPs realçados possuem boa informatividade e alto *Fct*. Dados organizados em ordem decrescente de *Fct*.

	TODAS AS POP	EUR-NAT	EUR-AFR	AFR-NAT	Estatísticas F		
SNP	*In*	*In*	*In*	*In*	*Fct*	*Fst*	*Fsc*
rs9295316	**0,487**	**0,482**	0,016	**0,598**	**0,853**	0,868	0,101
rs1919550	**0,467**	**0,435**	0,027	**0,569**	**0,835**	0,838	0,020
rs12953952	**0,458**	0,025	**0,428**	**0,553**	**0,778**	0,779	0,003
rs2585901	**0,336**	**0,463**	0,023	**0,313**	**0,755**	0,755	0,000
rs10492585	**0,478**	0,006	**0,495**	**0,569**	**0,746**	0,748	0,010
rs1073319	**0,397**	**0,384**	0,017	**0,518**	**0,739**	0,742	0,011
rs10488172	**0,409**	**0,321**	0,061	**0,553**	**0,722**	0,725	0,011
rs10517518	**0,348**	**0,410**	0,005	**0,349**	**0,700**	0,702	0,009
rs567992	**0,358**	**0,320**	0,022	**0,455**	**0,690**	0,701	0,038
rs10486576	**0,463**	**0,289**	**0,347**	**0,347**	**0,684**	0,687	0,009
rs2296274	**0,343**	0,055	**0,254**	**0,473**	**0,659**	0,677	0,055
rs10515919	**0,356**	**0,407**	0,061	**0,408**	**0,651**	0,652	0,003
rs4657449	**0,330**	**0,388**	0,002	**0,349**	**0,646**	0,648	0,005
rs802524	**0,308**	0,025	**0,262**	**0,373**	**0,636**	0,653	0,046
rs868179	**0,294**	0,037	**0,230**	**0,375**	**0,610**	0,610	0,000
rs4130513	**0,316**	0,003	**0,314**	**0,360**	**0,607**	0,625	0,045
rs3860446	**0,275**	**0,323**	**0,249**	0,010	**0,606**	0,615	0,023
rs249847	**0,347**	**0,267**	0,057	**0,503**	**0,587**	0,589	0,004
rs9302185	**0,285**	0,047	**0,204**	**0,389**	**0,579**	0,597	0,044
rs10508349	**0,248**	**0,264**	0,005	0,233	**0,572**	0,593	0,048
rs1898280	**0,264**	**0,345**	**0,216**	0,023	**0,572**	0,589	0,040
rs4733652	0,222	**0,250**	0,000	0,240	**0,571**	0,578	0,017
rs9320808	**0,249**	**0,310**	**0,236**	0,007	**0,565**	0,582	0,040
rs879780	**0,272**	0,039	**0,203**	**0,354**	**0,557**	0,587	0,069
rs1990745	**0,251**	**0,211**	0,018	**0,321**	**0,552**	0,561	0,019
rs7463344	0,194	0,005	**0,177**	0,207	**0,524**	0,521	-0,007
rs138022	**0,279**	0,089	**0,154**	**0,401**	**0,523**	0,548	0,051
rs10519979	**0,298**	**0,199**	0,069	**0,444**	**0,514**	0,512	-0,005
rs1013459	**0,245**	0,048	**0,167**	**0,328**	**0,505**	0,524	0,038
rs10491097	0,194	**0,176**	**0,238**	0,007	**0,500**	0,519	0,039
rs4762106	0,207	**0,253**	**0,204**	0,003	**0,495**	0,511	0,031
rs1517634	**0,248**	**0,249**	0,005	**0,313**	**0,490**	0,502	0,024
rs1398829	0,229	0,011	**0,203**	0,259	**0,487**	0,551	0,125
rs3828121	0,223	**0,163**	0,031	**0,292**	**0,487**	0,507	0,038
rs354747	**0,234**	0,056	**0,142**	**0,344**	**0,486**	0,500	0,027
rs10497705	**0,281**	**0,282**	0,009	**0,375**	**0,485**	0,495	0,020
rs4034627	0,196	0,001	**0,197**	0,213	**0,478**	0,502	0,046
rs6684063	**0,237**	**0,348**	**0,128**	0,069	**0,478**	0,505	0,051
rs10484578	**0,370**	0,082	**0,259**	**0,546**	**0,474**	0,484	0,019
rs5000507	**0,264**	**0,116**	**0,112**	**0,392**	**0,472**	0,514	0,080
rs9323178	**0,261**	**0,173**	0,057	**0,388**	**0,472**	0,487	0,027
rs10520678	**0,249**	0,095	**0,119**	**0,364**	**0,471**	0,490	0,036
rs4076700	**0,235**	0,069	**0,133**	**0,334**	**0,469**	0,483	0,026
rs1036543	0,217	**0,304**	**0,186**	0,019	**0,466**	0,492	0,047
rs708915	0,161	**0,194**	**0,165**	0,001	**0,465**	0,475	0,019
rs2035573	0,216	**0,145**	0,038	**0,303**	**0,449**	0,445	-0,007
rs1004704	0,196	**0,180**	0,007	0,246	**0,446**	0,446	0,001
rs2208139	0,228	**0,171**	0,032	**0,326**	**0,444**	0,457	0,022
rs2829454	0,234	**0,141**	0,056	**0,338**	**0,444**	0,451	0,013
rs3806218	0,203	0,039	**0,134**	**0,294**	**0,443**	0,467	0,044
rs4625554	**0,251**	**0,315**	0,002	0,273	**0,443**	0,463	0,036
rs153898	0,178	**0,227**	**0,154**	0,009	**0,425**	0,458	0,058
rs2569029	0,208	**0,190**	0,010	**0,275**	**0,422**	0,430	0,013
rs993314	0,165	**0,197**	**0,169**	0,001	**0,422**	0,428	0,011
rs10131076	0,197	0,045	**0,124**	0,267	**0,411**	0,433	0,037

rs4013967	0,228	**0,112**	0,085	**0,340**	**0,399**	0,399	0,001
rs4130405	0,178	**0,131**	0,022	0,238	**0,396**	0,417	0,035
rs9307613	**0,243**	0,080	**0,122**	**0,361**	**0,389**	0,403	0,022
rs304051	0,223	**0,140**	0,057	**0,334**	**0,381**	0,414	0,053
rs9310888	0,153	0,014	**0,120**	0,187	**0,357**	0,379	0,033
rs2785279	**0,238**	0,000	**0,262**	0,272	**0,348**	0,385	0,056
rs10507688	0,173	0,108	0,040	0,233	**0,342**	0,358	0,025
rs1397618	0,151	0,044	0,082	0,212	-0,340	0,551	0,665
rs7535375	0,146	0,033	0,087	0,215	**0,340**	0,350	0,015
rs6569792	0,124	0,015	**0,177**	0,093	**0,338**	0,350	0,018
rs10520440	0,144	0,066	0,049	0,213	**0,334**	0,335	0,001
rs1451928	0,135	**0,123**	0,004	0,166	**0,326**	0,342	0,023
rs10214949	0,132	0,026	0,082	0,183	**0,308**	0,319	0,017
rs10248051	0,104	0,012	**0,147**	0,078	**0,287**	0,316	0,041
rs2711070	0,140	**0,208**	0,052	0,066	**0,281**	0,284	0,004
rs2840290	0,105	0,082	**0,143**	0,010	**0,276**	0,276	0,001
rs10500505	0,088	0,065	0,009	0,119	**0,227**	0,222	-0,006
rs842634	0,130	0,058	0,048	0,191	**0,216**	0,251	0,045
rs10510791	0,113	0,061	0,029	0,167	**0,211**	0,254	0,055
rs1934393	0,074	0,042	0,109	0,017	**0,199**	0,257	0,073
rs10515535	0,088	0,001	0,111	0,091	**0,190**	0,199	0,011
rs1498991	0,136	0,104	**0,184**	0,014	**0,187**	0,617	0,529
rs10498919	**0,365**	0,093	**0,542**	0,240	**0,177**	0,652	0,577
rs6804094	0,150	0,027	0,107	0,223	**0,156**	0,219	0,075
rs3768176	0,148	0,085	**0,219**	0,036	**0,146**	0,404	0,303
rs1477277	0,079	0,049	0,109	0,015	0,115	0,206	0,103
rs719776	0,041	0,002	0,056	0,036	-0,076	0,353	0,399
rs2253624	**0,277**	0,010	**0,347**	**0,347**	-0,018	0,068	0,084

A tabela 6 mostra que existem 29 SNPs com valores relativamente baixos e, portanto, os mesmos podem não ter uma informatividade tão relevante para as análises como os outros marcadores do mesmo painel. Ainda nesta análise, 54 SNPs se mostraram como mais informativos em relação aos índices apresentados, o que auxilia o painel a conferir uma boa qualidade dos resultados. Além disso, 79 deles possuem valores de *Fct* claramente maiores que 0,12.

4.4. Diferenciação intrapopulacional

Os valores de diversidade intra-populacional, medidos pela heterozigosidade esperada, podem ser conferidos na Tabela 7. Como esperado para AIMs, os valores menores foram encontrados nas populações parentais (em negrito).

Tabela 7: Heterozigosidade esperada calculada para as 13 populações em estudo. Valores menores que 0,33 realçados em negrito.

Europeus			**Africanos**			**Ameríndios**			**Miscigenados**			
CEU	**TSI**	**EXT**	**LWK**	**MKK**	**YRI**	**HP**	**SHI**	**PUN**	**MEX**	**EQU**	**NIC**	**LIM**
0,346	0,344	**0,303**	**0,311**	0,361	**0,325**	**0,201**	**0,193**	**0,223**	0,398	0,374	0,409	0,337

Os maiores valores de heterozigosidade foram encontrados nas populações miscigenadas MEX, EQU e NIC, o que demonstra uma maior diversidade genética nestas populações. Os miscigenados de Lima apresentam um valor sutilmente mais baixo devido ao alto componente nativo americano desta população (72%).

Em SNPs selecionados ao acaso (que não são AIMs), uma maior heterozigosidade em africanos seria esperada dentro das populações parentais. Estas populações possuem uma ampla variação intracontinental por serem as populações de origem mais remota entre os cinco continentes mundiais, o que indica tempo suficiente para a diversificação intracontinental. Nesse contexto, menores valores de desequilíbrio de ligação seriam também esperados com SNPs selecionados ao acaso para os africanos (Lambert e Tishkoff, 2010; Campbell e Tishkoff, 2010).

A relação matemática entre *Fst* e heterozigosidade esperada prevê que baixos valores de heterozigosidade produzirão altos valores de *Fst* (Wang *et al*, 2007). Valores da Tabela 7 foram concordantes com a análise de *Fst* (Tabela 6). Valores maiores de diferenciação estão entre ameríndios e africanos, seguidos por ameríndios e outras populações. Os menores valores de heterozigosidade esperada estão em ameríndios. A heterozigosidade esperada das populações também foi calculada para cada marcador estudado para avaliar a qualidade dos mesmos (Tabela 8). Na Tabela 8, valores realçados em cinza indicam uma heterozigosidade menor que 0,30, o que seria esperado nas populações parentais com AIMs, menor diversidade intracontinental. Células com "-" identificam dados não suficientes para o cálculo. SNPs marcados em cinza escuro possuem menos de duas populações parentais com valores menores que 0,30 (baixa informatividade).

Tabela 8: Heterozigosidade esperada por SNP para cada população. Valores de heterozigosidade abaixo de 0,30 estão marcados em cinza. SNPs marcados em cinza escuro representam marcadores com menos de duas populações parentais com heterozigosidade menor que 0,30 (baixa informatividade).

SNP	Africanos			Europeus			Ameríndios			Miscigenados			
	MKK	LWK	YRI	EXT	CEU	TSI	HP	PUN	SHI	LIM	EQU	NIC	MEX
rs1004704	0,480	0,292	0,190	0,431	0,370	0,438	0,000	0,000	0,012	0,155	0,253	0,385	0,391
rs10131076	0,233	0,124	0,166	0,475	0,492	0,446	0,011	0,083	0,000	0,342	0,395	0,483	0,455
rs1013459	-	-	0,182	0,324	0,365	-	0,495	0,485	0,497	0,496	0,498	0,457	-
rs10214949	0,488	0,415	0,366	0,242	0,087	0,268	0,000	0,000	0,000	0,101	0,158	0,232	0,216
rs10248051	0,079	0,011	0,098	0,229	0,252	0,305	0,348	0,083	0,425	0,444	0,498	0,462	0,492
rs1036543	0,115	0,189	0,234	0,087	0,180	0,087	0,188	0,227	0,234	0,439	0,488	0,466	0,481
rs10484578	0,494	0,486	0,487	0,265	0,236	0,268	0,052	0,194	0,023	0,117	0,191	0,252	0,110
rs10486576	0,433	0,206	0,139	0,426	0,388	0,381	0,197	0,083	0,057	0,327	0,428	0,442	0,493
rs10488172	0,444	0,247	0,205	0,411	0,388	0,339	0,476	0,485	0,492	0,500	0,494	0,499	0,483
rs10491097	-		0,046	0,317	0,435	-	0,494	0,405	0,495	0,468	0,461	0,424	-
rs10492585	0,282	0,124	0,120	0,479	0,463	0,498	0,169	0,466	0,188	0,403	0,454	0,496	0,481
rs10497705	-	-	-	0,194	0,269	-	0,500	0,466	0,480	0,487	0,499	0,469	-
rs10498919	0,224	0,153	0,166	0,276	0,341	0,219	0,459	0,165	0,492	0,440	0,491	0,497	0,455
rs10500505	-	-	0,323	0,192	0,295	-	0,483	0,499	0,500	0,494	0,494	0,468	-
rs10507688	0,121	-	-	0,257	0,222	0,363	0,482	0,364	0,363	0,421	0,500	0,497	0,473
rs10508349	0,496	0,498	0,400	0,092	0,165	0,146	0,107	0,122	0,023	0,147	0,108	0,198	0,196
rs10510791	-	-	-	0,003	-	-	0,499	0,491	0,470	-	0,436	0,365	-
rs10515535	0,291	0,231	0,262	0,469	0,480	0,474	0,032	0,194	0,354	0,257	0,392	0,494	0,483
rs10515919	0,480	0,488	0,404	0,069	0,053	0,034	0,000	0,000	0,000	0,050	0,051	0,195	0,133
rs10517518	-	-	0,278	0,500	0,500	-	0,000	0,122	0,000	0,240	0,349	0,444	-
rs10519979	0,381	0,278	0,189	0,490	0,496	0,489	0,317	0,159	0,351	0,387	0,376	0,404	0,409
rs10520440	0,403	0,206	0,213	0,500	0,494	0,496	0,016	0,159	0,067	0,334	0,302	0,414	0,452
rs10520678	0,017	0,033	-	0,216	0,208	0,219	0,483	0,454	0,310	0,492	0,498	0,400	0,378

rs1073319	0,366	0,247	0,252	0,415	0,420	0,449	0,218	0,386	0,110	0,397	0,478	0,490	0,492
rs12953952	0,376	0,278	0,162	0,362	0,370	0,391	0,144	0,122	0,223	0,388	0,452	0,499	0,490
rs138022	0,404	0,358	0,252	0,436	0,494	0,434	0,021	0,043	0,034	0,354	0,436	0,488	0,500
rs1397618	0,233	0,255	0,248	0,240	0,298	0,227	0,058	0,258	0,023	0,452	0,496	0,407	0,495
rs1398829	0,372	0,247	0,329	0,500	0,496	0,500	0,000	0,159	0,000	-	0,371	0,483	0,473
rs1451928	0,274	0,223	0,197	0,172	0,150	0,201	0,147	0,315	0,127	0,404	0,489	0,472	0,490
rs1477277	0,171	0,107	0,076	0,416	0,434	0,407	0,441	0,454	0,472	0,489	0,500	0,489	-
rs1498991	-	-	0,499	0,031	-	-	0,000	0,000	0,000	0,027	0,051	0,143	-
rs1517634	0,379	0,247	0,302	0,312	0,252	0,262	0,300	0,423	0,012	0,494	0,492	0,492	0,473
rs153898	-	-	0,101	0,422	0,420	-	0,344	0,268	0,186	0,434	0,467	0,483	-
rs1898280	0,434	0,257	0,210	0,301	0,320	0,290	0,043	0,000	0,000	0,129	0,183	0,321	0,167
rs1919550	0,482	0,391	0,358	0,317	0,298	0,325	0,005	0,000	0,000	0,140	0,196	0,304	0,236
rs1934393	-	-	0,143	0,324	0,343	-	0,005	0,227	0,012	0,434	0,479	0,497	-
rs1990745	-	-	0,058	0,453	0,403	-	0,176	0,083	0,273	0,323	0,421	0,483	-
rs2035573	0,017	0,022	0,039	0,000	-	-	0,499	0,340	0,454	0,485	0,498	0,373	0,463
rs2208139	-	-	0,253	0,223	0,385	-	0,279	0,287	0,357	0,443	0,491	0,500	-
rs2253624	0,488	0,406	0,337	0,359	0,413	0,391	0,063	0,159	0,049	0,346	0,430	0,485	0,500
rs2296274	-	-	0,091	0,371	0,403	-	0,351	0,315	0,360	0,481	0,499	0,497	-
rs249847	-	-	0,035	0,164	0,241	-	0,011	0,364	0,000	0,404	0,489	0,452	-
rs2569029	-	-	-	0,177	0,269	-	0,323	0,340	0,223	0,454	0,500	0,464	-
rs2585901	0,499	0,401	0,473	0,216	0,208	0,236	0,000	0,000	0,000	0,125	0,123	0,244	0,236
rs2711070	0,266	0,043	0,039	0,431	0,427	0,319	0,243	0,439	0,170	0,463	0,472	0,466	0,498
rs2785279	-	-	0,418	0,426	0,343	-	0,006	0,000	0,000	0,129	0,203	0,356	-
rs2829454	0,330	0,306	0,369	0,343	0,286	0,351	0,308	0,466	0,178	0,467	0,483	0,490	0,500
rs2840290	0,467	0,500	0,497	0,120	0,165	0,034	0,011	0,043	0,000	0,073	0,092	0,156	0,110
rs304051	-	-	0,035	0,443	0,499	-	0,500	0,416	0,410	0,470	0,483	0,494	-
rs354747	-	-	-	0,149	0,077	-	0,083	0,165	0,199	0,429	0,491	0,462	-
rs3768176	0,391	0,327	0,396	0,195	0,165	0,165	0,000	0,000	0,000	0,110	0,131	0,272	0,165
rs3806218	0,344	0,075	0,048	0,264	0,261	0,184	0,463	0,315	0,353	0,372	0,358	0,433	0,329
rs3828121	-	-	0,153	0,141	0,043	-	0,000	0,000	0,000	0,086	0,076	0,226	-
rs3860446	-	-	0,264	0,444	0,343	-	0,403	0,440	0,386	0,401	0,462	0,496	-
rs4013967	0,457	0,320	0,223	0,342	0,397	0,397	0,000	0,000	0,000	0,120	0,217	0,288	0,292
rs4034627	-	-	0,232	0,484	0,499	-	0,095	0,087	0,281	0,265	0,376	0,450	-
rs4076700	0,496	0,358	0,316	0,369	0,286	0,407	0,498	0,227	0,480	0,468	0,479	0,499	0,493
rs4130405	0,467	0,401	0,276	0,384	0,405	0,375	0,000	0,000	0,012	0,205	0,190	0,343	0,337
rs4130513	0,006	0,011	-	0,306	0,274	0,283	0,117	0,364	0,089	0,355	0,494	0,479	0,495
rs4625554	0,360	0,175	0,243	0,361	0,208	0,219	0,068	0,364	0,000	0,480	0,497	0,500	0,482
rs4657449	-	-	0,375	0,322	0,295	-	0,335	0,449	0,232	0,498	0,500	0,497	-
rs4733652	0,254	0,043	0,034	0,408	0,370	0,416	0,032	0,083	0,000	0,279	0,386	0,466	0,443
rs4762106	-	-	0,046	0,162	-	-	0,077	0,000	0,046	0,143	0,236	0,244	-
rs5000507	0,344	0,135	0,024	0,341	0,351	0,381	0,053	0,122	0,000	0,215	0,237	0,333	0,289
rs567992	0,290	-	0,080	0,493	0,499	0,477	0,500	0,466	0,500	0,487	0,491	0,498	0,470
rs6569792	-	-	-	0,070	0,039	-	0,263	0,043	0,046	0,169	0,143	0,176	-
rs6684063	0,462	0,358	0,299	0,154	0,018	0,107	0,000	0,000	0,000	0,067	0,097	0,244	0,155
rs6804094	0,198	0,233	0,153	0,307	0,331	0,268	0,382	0,315	0,317	0,459	0,499	0,483	0,486
rs708915	0,493	0,406	0,296	0,470	0,434	0,463	0,000	0,043	0,012	0,211	0,282	0,483	0,216
rs719776	0,474	0,451	0,262	0,240	0,150	0,137	0,005	0,000	0,000	0,101	0,124	0,483	0,272
rs7463344	-	-	0,443	0,243	0,343	-	0,000	0,083	0,000	0,350	0,449	0,483	-
rs7535375	0,374	0,198	0,133	0,471	0,463	0,430	0,135	0,194	0,381	0,243	0,325	0,483	0,385
rs802524	-	-	0,281	0,472	0,480	-	0,352	0,194	0,486	0,385	0,467	0,483	-
rs842634	0,399	0,239	0,208	0,267	0,261	0,298	0,179	0,423	0,068	0,414	0,459	0,483	0,498
rs868179	-	-	0,219	0,312	0,219	-	0,233	0,043	0,415	0,188	0,237	0,483	-
rs879780	-	-	0,308	0,130	0,113	-	0,053	0,315	0,012	0,309	0,245	0,483	-
rs9295316	0,110	0,011	-	0,257	0,198	0,227	0,135	0,466	0,189	0,491	0,498	0,483	0,492
rs9302185	0,110	0,162	0,255	0,120	0,087	0,175	0,392	0,083	0,303	0,477	0,498	0,483	0,493
rs9307613	-	-	-	0,492	0,403	-	0,167	0,227	0,099	0,441	0,441	0,483	-
rs9310888	0,496	0,333	0,390	0,312	0,265	0,288	0,063	0,340	0,067	0,405	0,477	0,483	-
rs9320808	0,297	0,227	0,143	0,445	0,286	0,425	0,500	0,454	0,497	0,439	0,475	0,483	0,455
rs9323178	0,500	0,437	0,458	0,117	0,036	0,107	0,402	0,434	0,154	0,463	0,473	0,483	0,489
rs993314	-	-	0,500	0,114	0,039	-	0,011	0,000	0,000	0,050	0,088	0,483	-

É esperado que AIMs apresentem baixa variabilidade (i.e. heterozigosidade) dentro das populações parentais e maior variabilidade entre estas populações. Os rs1013459, rs10488172, rs10491097, rs2296274, rs4076700 e rs567992, rs10500505, rs2253624, rs2829254, embora apresentem altos valores de *Fct* e *In* (Tabela 4), apresentam também valores relativamente altos de heterozigosidade em algumas das populações parentais. Os SNPs rs1477277 e rs68004094 são aparentemente os menos informativos para ancestralidade, apresentando valores baixos de *In* e *Fct* (Tabela 4) e alta heterozigosidade nas populações parentais.

4.5. Quantificação da miscigenação

A miscigenação individual das populações em estudo está ilustrada na Figura 10, obtida com o método de ML implementado no software Admixture (Alexander *et al.,* 2009). Todas as populações miscigenadas (NIC, MEX, EQU e LIM) mostraram um baixo componente africano. A população NIC mostrou o maior componente africano entre as populações miscigenadas (média de 0,12). Bryc *et al.* (2010) também encontraram extensa variação de ancestralidade europeia e ameríndia e baixo componente africano nas populações do Equador e México. Apesar de o componente africano ser baixo, ele está presente em todas as populações miscigenadas.

O componente de ancestralidade africano nas populações miscigenadas variou entre 0,000001 e 0,537; o de ancestralidade ameríndia está entre 0,049 e 0,999; e o de ancestralidade europeia varia entre 0,000001 e 0,909.

A população parental Maasai (MKK) apresentou um certo nível de miscigenação, o que parece ser devido à uma alta porcentagem de *missing data* (29 dos 83 SNPs não obtiveram nenhum indivíduo genotipado). Por este motivo, os MKK foram retirados da análise de miscigenação e a mesma foi realizada novamente. Nenhuma diferença nos demais indivíduos foi observada (dados não mostrados).

Coerentemente com as análises de *Fst* apresentadas na Tabela 5, NIC possuem valores menores de *Fst* com populações da Europa, justificando seu maior componente ancestral europeu. Os EQU, assim como as outras três populações miscigenadas, possuem alto *Fst* com africanos, corroborando as análises de miscigenação que mostram um baixo componente africano em todas as populações miscigenadas (Figura 10).

Galanter *et al* (2012) verificou que necessita-se de mais SNPs de informatividade européia que de africana e ameríndia em um painel de AIMs devido ao fato de a população européia ser geneticamente e geograficamente intermediária às populações africana e ameríndia. Isso explica por que AIMs de informatividade européia possuem menor *In* em relação aos outros. Em seu estudo, dos 446 AIMs selecionados para seu painel, 202 eram de informatividade européia, 115 de africana e 129 de ameríndia. Dessa forma, com quase o

dobro de marcadores de informatividade européia, eles conseguiram um *In* cumulativo semelhante para as três populações. Uma limitação do nosso painel é que não temos um número maior de AIMs europeus, o que pode estar associado a um maior erro na estimativa de miscigenação europeia. O painel utilizado neste estudo (Yeager *et al.,* 2008) possui cinco SNPs em comum com o painel de Galanter *et al.* (2012).

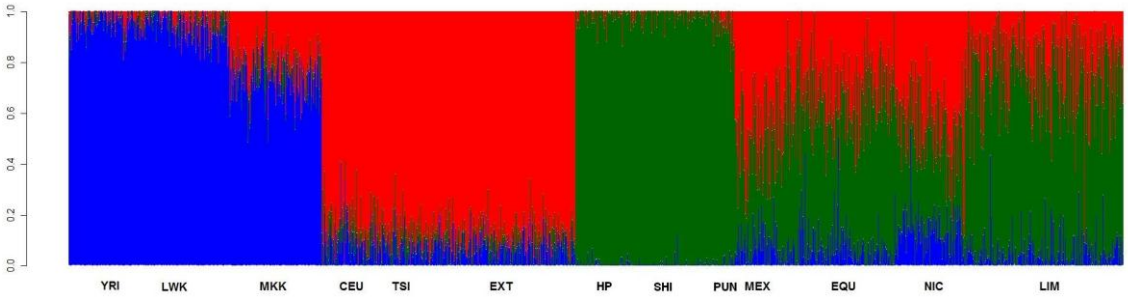

Figura 10: Esquema da quantificação da miscigenação individual dos 1957 indivíduos do estudo. Cada barra representa um indivíduo da amostragem. As cores da barra representam a porcentagem de componente de ancestralidade de cada indivíduo. Azul: Componente Africano; Vermelho: Componente Europeu; Verde: Componente Ameríndio.

4.6. Distribuição da miscigenação individual

Em todas as populações miscigenadas observamos uma ampla distribuição de valores de miscigenação individual, como evidenciado na Figura 11. Como mencionado anteriormente, a população NIC (Figura 11-D) apresentou o maior índice de miscigenação africano (média de 0,12). Entretanto, o maior componente ancestral desta população foi o europeu, assim como verificado nos mexicanos (Figura 11-C).

Os EQU e LIM apresentam os menores índices de miscigenação africana (0,06 e 0,05, respectivamente) e maiores índices de componente nativo americano (0,59 e 0,72, respectivamente), como visualizado na Figura 11.

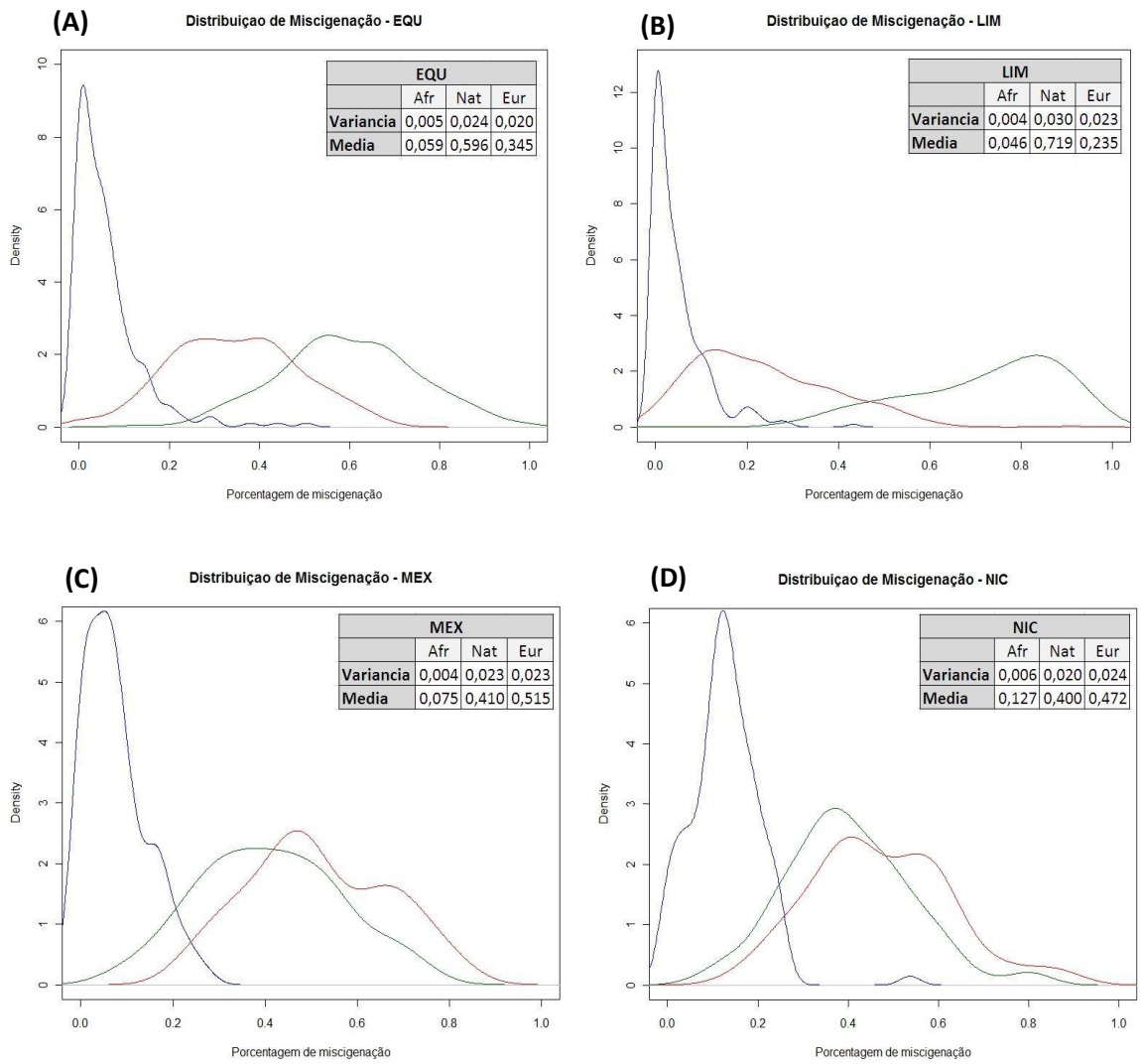

Figura 11: Distribuição de miscigenação individual nas populações: (A) EQU, (B) LIM, (C) MEX e (D) NIC. Azul corresponde à ancestralidade africana, vermelho à ancestralidade europeia e verde à ancestralidade ameríndia. Os valores na tabela em detalhe para cada gráfico indicam a miscigenação populacional das respectivas populações (média da miscigenação individual).

A variância da miscigenação individual é um importante fator para compreendermos a dinâmica da miscigenação de populações (Verdu e Rosenberg, 2011). Sob um modelo de miscigenação e isolamento, a variância de miscigenação individual diminui com o passar do tempo após a miscigenação tendendo a um equilíbrio em que todos os indivíduos tem o mesmo nível de miscigenação. Por outra parte, sob um modelo de fluxo gênico contínuo, a chegada de novos indivíduos das populações parentais ao longo do tempo, mantém a presença de indivíduos com uma alta ancestralidade das populações parentais, mesmo se a miscigenação iniciou há muito tempo. Portanto, de acordo com nossos dados, percebemos um valor baixo de variância africana e mais alto de Ameríndios e Europeus (Figura 11), o que pode significar que o fluxo gênico africano nestas populações cessou ou diminuiu

significativamente após a abolição do tráfico de escravos (metade do século XVI) e o fluxo de europeus continua até os dias atuais.

Nas populações urbanas estudadas de Nicarágua (Manágua), Equador (Quito), Peru (Lima) e Mexicanos (de Los Angeles), a baixa miscigenação média africana, associada com a ausência de indivíduos com ancestralidade africana maior a 20-25% (Figura 11), além de ser coerente com o fim do comércio de escravos africanos para América Latina (1850) é também indicativa de ausência de fluxo gênico a partir de comunidades isoladas de alta ancestralidade de afro-descendentes, equivalentes aos quilombos no Brasil. Este resultado é similar a alguns estudos em algumas populações brasileiras, onde a ancestralidade africana variou de 21-27% em populações do Sudeste e Centro-Oeste (Leite *et al*., 2011; Giolo *et al.,* 2012).

No Brasil, quando populações de diferentes regiões são analisadas separadamente, percebemos um gradiente de ancestralidade africana acentuado, variando de 10% em populações da região Norte a 30% em populações da região Nordeste, onde se instalaram os Quilombos (Pena *et al.,* 2011). Francez *et al.* (2012) mostram que em populações do Norte do Brasil temos indivíduos com variação de componente africano entre 14-21%, o que é indicativo de baixo fluxo gênico muito recente (últimas 2-3 gerações) proveniente de populações de quilombos, mas quando se trata de afro-descendentes, este valor pode chegar a até 69%, o que é indicativo de fluxo gênico recente mais intenso a partir de populações com alta ancestralidade africana. A variação interindividual da ancestralidade do estudo de Santos *et al*. (2010) é também interessante: uma amostra do Rio Grande do Sul evidencia uma alta ancestralidade europeia associada a muitos indivíduos com alta ancestralidade europeia (>70%), indicativa de fluxo gênico continuo recente. Diferentemente, uma amostra de Belém tem maiores níveis de miscigenação africano e nativo, poucos indivíduos com alguma ancestralidade predominante e aparentemente uma menor variância da miscigenação individual, o que é indicativo de eventos de miscigenação mais antigos associados a uma maior tendência à panmixia. No mesmo estudo, a amostra de afrodescendentes da região Amazônica, praticamente não tem indivíduos com ancestralidade europeia ou nativa maior a 20-30%, o que evidencia que o fluxo gênico de origem europeu nas últimas gerações foi relativamente baixo. Estes resultados caracterizam uma grande variação desse componente nas populações brasileiras e latino-americanas.

Em relação ao componente Nativo Americano, as populações de Lima e Quito, além de apresentarem uma maior ancestralidade média ameríndia, apresentam uma maior porcentagem de indivíduos quase completamente nativo-americanos, o que é coerente com o fato que Equador e Peru são países com uma população nativo-americana muito relevante na zona rural que nos últimos decênios têm migrado do campo para as grandes cidades como Lima e Quito. Este fenômeno parece ser menos evidente na população mexicana de

Los Angeles (MEX), e na população de Manágua (NIC). O componente de miscigenação individual europeu evidencia um comportamento complementar ao componente nativo-americano. Mexicanos de Los Angeles e Nicaraguenses de Manágua apresentam maior miscigenação média europeia e mais indivíduos com uma ancestralidade europeia muito alta que Equatorianos e Peruanos, o que sugere um fluxo gênico europeu recente relativamente mais relevante. Porém, nestas considerações, é importante considerar a estratégia de amostragem utilizada. Por exemplo, as amostras de Lima são pacientes de uma unidade de gastrenterologia de um hospital público, o que determina um viés para indivíduos de estrato socioeconômico mais baixo, associado com uma maior ancestralidade nativo-americana (Pereira *et al.* submetido a PLOS One).

4.7. Desequilíbrio de ligação entre *loci* não ligados gerado pela miscigenação

Sabe-se que quanto mais antiga é uma população homogênea, menor desequilíbrio de ligação (DL) ela tem, pois teve tempo o bastante para sofrer uma quantidade significativa de recombinação. Por outro lado, fenômenos de gargalo de garrafa geram DL. No contexto das populações parentais dos latino-americanos, devido a sua história demográfica, Africanos, europeus e nativos americanos apresentam uma ordem crescente de DL (Lambert e Tishkoff , 2010; Campbell e Tishkoff, 2010). Nossos dados evidenciam que a população africana (Figura 12 e Tabela 9) tem menos DL que a europeia, mas a população de Nativos Americanos apresenta, contrariamente ao esperado, um nível de DL similar a aquele observado nos Africanos. Isto pode ser um efeito da escolha dos AIMs do presente painel, para os quais o *Fst* entre Nativos Americanos é menor que entre populações europeias ou entre populações africanas (Tabela 6), pois da mesma forma em que a miscigenação gera DL, o agrupamento artificial de populações diferenciadas também gera DL.

No contexto de populações miscigenadas, o nível de DL depende daquele observado nas populações parentais, em função da sua contribuição, e de fato que, como explicado na introdução, a miscigenação gera DL mesmo entre *loci* não ligados se as populações parentais são suficientemente diferenciadas (Nei e Li, 1973). Este DL gerado pela miscigenação se diluirá com o passar do tempo sob um modelo de miscigenação e isolamento, e cairá mais lentamente sob um modelo de fluxo gênico constante, como explicado na introdução.

Analisando a Tabela 9, observamos os valores de de r^2 para as populações parentais e miscigenadas.

Tabela 9: Média de de r^2 entre os AIMS não ligados do painel estudado para as populações parentais e miscigenadas.

	AFR	EUR	NAT	LIM	EQU	NIC	MEX
r^2	0,022	0,025	0,014	0,028	0,035	0,042	0,043

Os gráficos da Figura 12 corroboram esta afirmação. Na figura 12-B percebemos o baixo DL das populações parentais, com a maior massa da distribuição abaixo de 0,03.

Nas populações miscigenadas (Figura 12-A), observamos um aumento no DL, com a massa da distribuição se estendendo até cerca de 0,09.

Dentro das populações miscigenadas, percebemos que a média de r^2 está sutilmente aumentada em NIC e MEX, indicando uma miscigenação mais recente nestas populações, ou refletindo a maior ancestralidade europeia (a população parental com maior DL) desta população. Metodologias mais sofisticadas serão necessárias para separar estes dois efeitos.

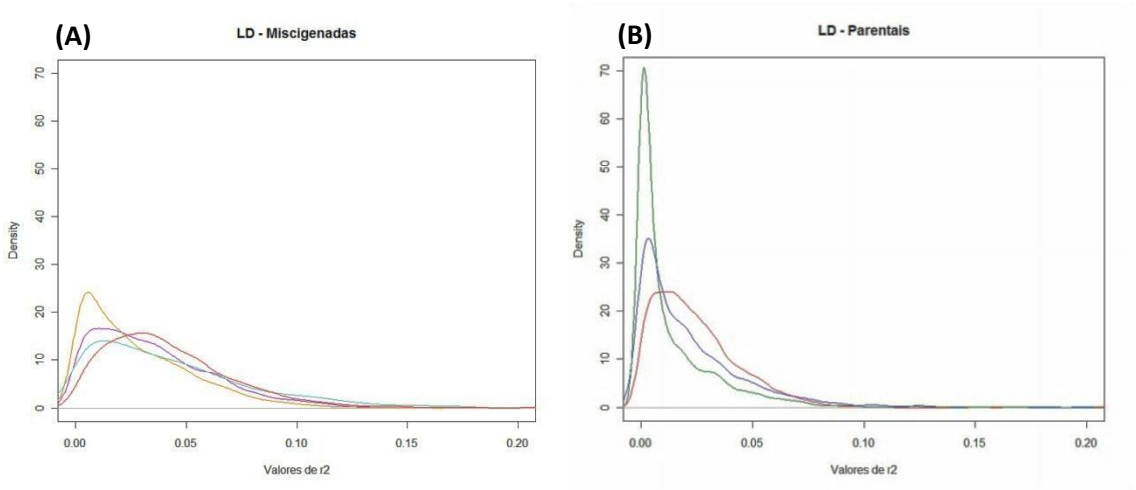

Figura 12: Gráficos de distribuição dos valores individuais de r^2 para (A) populações miscigenadas (amarelo: LIM; violeta: EQU; azul claro: MEX e; vermelho: NIC) e (B) populações parentais (verde: ameríndios; azul: africanos e; vermelho: europeus).

CONCLUSÕES

Na era genômica, a disponibilidade de grandes conjuntos de dados de SNPs está facilitando estudos de miscigenação, permitindo ir além dos estudos clássicos de quantificação da miscigenação populacional, estudando a distribuição da miscigenação individual e o padrão de desequilíbrio de ligação, que são informativos sobre como os processos de miscigenação ocorreram.

Neste estudo, nós integramos com sucesso um conjunto de dados produzidos por nosso grupo com dados públicos do Projeto Internacional HapMap com seus devidos controles de qualidade.

Nossas análises evidenciam que o painel desenvolvido por Yeager *et al*. (2008) com 106 SNPs é suficiente para inferir a miscigenação populacional e individual em populações da América Latina, como evidenciado por diferentes estudos (Via *et al.*, 2011; Avena *et al.*, 2012). Com 83 SNPs deste painel, como sugerido por Galanter *et al*. (2012), conseguimos inferir a miscigenação individual em quatro populações da América Latina, do Equador, Peru, México e Nicarágua. Além disso, com nossas análises de informatividade e diferenciação intra e interpopulacional, podemos ainda sugerir um subconjunto desse se for necessário genotipar menos SNPs em futuros estudos, pois há SNPs pouco informativos neste painel que podem ser removidos.

As medidas de variação genética utilizadas neste estudo: Heterozigosidade esperada, Equilíbrio de Hardy-Weinberg, Frequências alélicas, Estimativas de Miscigenação, AMOVA, *Fst* e Desequilíbrio de Ligação, nos permitiram elucidar fatores quantitativos da miscigenação individual e populacional de todos os indivíduos estudados e de fatores qualitativos na inferência da dinâmica da miscigenação nas quatro populações miscigenadas estudadas. Adicionalmente, a análise multivariada de PCA foi bastante informativa para a elucidação dos dados, visto que os dois primeiros componentes principais conseguiram explicar 42,7% de toda a variação encontrada no conjunto de dados, ilustrando de forma condizente com todas as análises a estimativa de miscigenação individual para todas as populações estudadas, mostrando uma maior diferenciação entre africanos e outros grupos e um baixo componente africano para todas as populações miscigenadas.

As análises de desequilíbrio de ligação, variância e distribuição da miscigenação individual nos auxiliaram a elucidar fatores qualitativos da dinâmica de populações miscigenadas da América Latina (NIC, MEX, EQU e LIM), identificando presença ou ausência de fluxo gênico recente. Percebemos com estas análises que as populações de Nicarágua e México sofreram miscigenação mais recente e/ou provavelmente sofrem maior fluxo gênico de imigrantes principalmente europeus até os dias atuais, diferente das populações miscigenadas da América do Sul: Equador e Peru.

De um ponto de vista metodológico, este trabalho foi realizado no contexto de um esforço para padronizar procedimentos de análises em nosso grupo de pesquisa. Esses esforços incluem o uso do DIVERGENOME (http://pggenetica.icb.ufmg.br/divergenome/pagina/index.php) para seguramente armazenar os dados e fazer procedimentos de controle de qualidade para verificar a integridade dos conjuntos de dados, o uso de arquivos congelados para realizar as análises, o uso de pipelines de análises padronizados e scripts (apresentados nos Anexos) armazenados em um repositório de scripts disponível para os membros do laboratório.

Em um futuro próximo, nós usaremos mais ferramentas conceituais sofisticadas para estudar a distribuição da miscigenação individual e desequilíbrio de ligação para aperfeiçoar nossas inferências sobre a dinâmica da miscigenação nessas populações.

REFERÊNCIAS BIBLIOGRÁFICAS

Alexander, D.H., Novembre,J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research,* 19:1655-1664.

Avena S., Via M., Ziv E., Pérez-Stable E.J. *et al.* (2012). Heterogeneity in Genetic Admixture across Different Regions of Argentina. *PLoS One*. 2012;7(4):e34695.

Barbujani, G. e Colonna, V. (2009). Human genome diversity: frequently asked questions. *Trends in Genetics* 26:285-295.

Barrett JC, Fry B, Maller J, Daly MJ. (2005). Haploview: analysis andvisualization of LD and haplotype maps. *Bioinformatics*.

Bernstein, F. (1931). Die Geographische Verteilung der Blutgruppen und ihre Anthropologische Verteilung. In *Comitato Italiano per lo Studio dei Problemi della Populazione*. Roma: Istituto Poligrafico dello Stato.

Bryc, K. *et al.* (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *PNAS* 107 (2): 786-791.

Campbell, M.C. e Tishkoff, S.A. (2010). The Evolution of Human Genetic and Phenotypic Variation in Africa. *Curr Biol.* 20(4): R166–R173. doi:10.1016/j.cub.2009.11.050.

Cavalli-Sforza, L. L. (1998). The DNA revolution in population genetics. *Trends Genet,* 14**,** 60-5.

Cavalli-Sforza, L. L. e Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet,* 33 Suppl**,** 266-75.

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

Da Silva, M.C., Zuccherato L.W., Lucena F.C., Soares-Souza G.B., et al. (2011). Extensive admixture in Brazilian sickle cell patients: implications for the mapping of genetic modifiers. *Blood*. 118(16):4493-5.

De Meeûs, T. e Goudet, J. (2007). A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetics and Evolution* 7:731–735.

Dempster, A.P., Laird, N.M. e Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39:1–38.

Do, C.B. e Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology* 26:897-899.

Duncan, G., Thomas, E., Gallo, J.C. et al, (1996). Human Phylogenetic Relationships According to the D1S80 locus. *Genetica* 98:227-87.

Excoffier, L., Smouse, P. E. e Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131, 479-91.

Francez, P.A., Ribeiro-Rodrigues, E.M., dos Santos, S.E. (2012). Allelic frequencies and statistical data obtained from 48 AIM INDEL loci in an admixed population from the Brazilian Amazon. *Forensic Sci Int Genet.* Jan;6(1):132-5.

Galanter J.M., Fernandez-Lopez J.C. and the LACE Consortium (2012). Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas. *PLoS Genet*. 8(3):e1002554.

Giolo SR, Soler JM, Greenway SC, Almeida MA et al. (2012). Brazilian urban population genetic structure reveals a high degree of admixture. *Eur J Hum Genet.* 20(1):111-6. doi: 10.1038/ejhg.2011.144.

GLU: *Genotype and Library Utilities. (*http://code.google.com/p/glu-genetics/). Accessed on 24-01-2012.

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

Hartl, D. e Clark, A. (2010) Princípios de Genética de Populações. Editora Artmed, 4ª Ed. Porto Alegre, Brasil.

Jombart T., Solymos P. (2008) Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403-1405.

Lambert e Tishkoff, S.A. (2010). Genetic Structure in African Populations: Implications for Human Demographic History. *Cold Spring Harb Symp Quant Biol.* 74: 395–402. doi:10.1101/sqb.2009.74.053.

Leite T.K., Fonseca R.M., de França N.M., Parra E.J., Pereira R.W. (2011). Genomic ancestry, self-reported "color" and quantitative measures of skin pigmentation in Brazilian admixed siblings. *PLoS One*. 6(11):e27162.

Lins, T. C. *et al.*, (2010). Genetic composition of Brazilian population samples based on a set of twenty eight ancestry informative SNPs. *Am J Hum Biol*. 22: 187-192.

Long, J. (1991). The Genetic Structure of Admixed Populations. *Genetics* 127: 417-428.

Magalhães WCS ; Rodrigues MR ; Silva D ; Soares-Souza GB ; Linannini ML ; Faria-Campos AC ; Tarazona-Santos E. (2012). DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology studies. *Genetic Epidemiology* 36:01-10.

Nei, M. e Li, W.H. (1973). Linkage Disequilibrium In Subdivided Populations. *Genetics* 75: 213-219.

Ottensooser, F. (1944). Cálculo do grau de mistura racial através dos grupos sanguíneos. *Revista Brasileira de Biologia* 4:531-7.

Packer B. R., Yeager M., Staats B. Welch R., *et al.* (2004) SNP500Cancer: a public resource for sequence assay development for genetic variation in candidate genes. *Nucleic Acids Res*. Database issue - 32:D617-D621.

Pena S.D., Di Pietro G., Fuchshuber-Moraes M., Genro J.P. et al. (2011). The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One.* 6;6(2):e17063.

Pereira, L., Zamudio, R., Soares-Souza, G., Herrera, P. *et al.* (submitted to PLoS One.). Socioeconomic and Nutritional Factors Account for the Association of Gastric Cancer with Amerindian Ancestry in a Latin American Admixed Population.

Pfaff, C. L. *et al.* (2004). Information on ancestry from genetic markers. *Genet Epidemiol*. 26: 305-315.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Purcell S., Neale B., Todd-Brown K., Thomas L., *et al.* (2007).PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics,* 81.

Reader, J. (1998). Africa. A Biography of the Continent. New York: Knopf.

Rosenberg, N. A. *et al.* (2003) Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73 (6): 1402-1422.

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

Salzano, F.M., (1997). Human Races: Myth, Invention or Reality? *Interciência* 22:221-7.

Santos NP, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AK, (2010). Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum Mutat.* 31(2):184-90.

SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA *(URL: http://pga.gs.washington.edu)* [02-2012 accessed].

Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC et al. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. *Proc Natl Acad Sci USA*. 26;106(21):8611-6.

Soares-Souza, G. (2012). Integração de Bases de Dados Biológicos em Estudos de Genética de Populações e Epidemiologia Genética: Aplicações em Divergenome e o Projeto Epigen-Brasil. Qualificação de Doutorado em Bioinformática. Universidade Federal de Minas Gerais.

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics,* **68**, 978-989.

Tarazona-Santos, E. *et al.* (2007). Controlling the effects of population stratification by admixture in pharmacogenetics. In: Guilherme Suarez-Kurtz. (Editor): Pharmacogenomics in Admixed populations. Austin: Landes Bioscience.

The dbSNP Database (http://www.ncbi.nlm.nih.gov/SNP/). Accessed on 02-02-2012.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426:789-796.

The International HapMap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58.

The R Project for Statistical Computing (2009) About R. Disponível em http://www.r-project.org/. Acesso: 02/03/2012.

Verdu, P., Rosenberg, N.A. (2011). A General Mechanistic Model for Admixture Histories of Hybrid Populations. *Genetics*, Vol. 189, 1413–1426

Via M., Ziv E., Burchard E.G. (2009). Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. *Clin Genet*. 76(3):225-35.

Via M, Gignoux CR, Roth LA, Fejerman L, Galanter J, et al. (2011) History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico. *PLoS One* 6(1): e16513. doi:10.1371/journal.pone.0016513.

Wang S., Lewis C.M., Jakobsson M., Ramachandran S.,*et al..* (2007) Genetic variation and population structure in Native Americans. *PLoS Genet.* 3:e185.

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

Yaeger,R., Avila-Bront, A., Abdul,K., Nolan, P.C. et al. (2008). Comparing Genetic Ancestry and Self-Described Race in African Americans Born in the United States and in Africa. *Cancer Epidemiol Biomarkers Prev* 17(6):1329-1338.

ANEXO 1

Informação dos SNPs:

SNP	Funcionalidade	Cromossomo	Posição no Cromossomo	Gene	Alelos
rs9310888	NA	3	29286762	NA	A/G
rs1517634	NA	2	224183485	NA	A/G
rs10497705	NA	2	190492014	NA	C/T
rs10498255	intron-variant	2	231612230	CAB39	C/T
rs798887	NA	19	54793188	NA	A/G
rs9325872	NA	8	20480271	NA	A/G
rs10519979	NA	4	149634951	NA	A/G
rs10508349	NA	10	8298964	NA	A/G
rs868179	NA	2	177549497	NA	A/G
rs10488172	intron-variant	7	133335176	EXOC4	G/T
rs10486576	intron-variant	7	28119143	JAZF1	C/T
rs1470524	NA	2	45129515	NA	C/T
rs567992	NA	11	106262397	NA	A/G
rs9320808	intron-variant	6	121654596	C6orf170	A/G
rs249847	NA	12	98867716	NA	C/T
rs2817611	NA	1	11613163	NA	A/G
rs708915	intron-variant	20	8400667	PLCB1	A/T
rs2595456	NA	11	6884763	NA	A/G
rs1498991	NA	3	20900132	NA	C/G
rs10484578	intron-variant	6	35246319	ZNF76	A/G
rs2042762	NA	18	24609876	NA	A/G
rs2785279	NA	10	33709876	NA	C/T
rs10517518	NA	4	61795416	NA	A/G
rs2253624	NA	17	69732081	NA	G/T
rs10500505	NA	16	64943026	NA	A/T
rs879780	intron-variant	11	130008104	APLP2	C/T
rs138022	intron-variant	22	40613036	TNRC6B	A/G
rs1397618	intron-variant	10	120832675	EIF3A	A/T
rs257748	intron-variant	5	15819615	FBXL7	A/T
rs2208139	NA	20	37908954	NA	C/T
rs9292118	NA	5	55900196	NA	C/T
rs4076700	reference	12	117383320	FBXW8	C/T
rs2840290	intron-variant	9	16733957	BNC2	C/T
rs5000507	NA	13	82088954	NA	A/T
rs2569029	NA	5	155240510	NA	G/T
rs12953952	intron-variant	18	67737927	RTTN	A/G
rs7463344	NA	8	33863527	NA	C/G
rs2296274	intron-variant	14	61917178	PRKCH	A/G
rs10520678	NA	15	88937283	NA	C/T
rs10498919	NA	6	78947733	NA	C/G
rs842634	intron-variant	2	61091222	FLJ16341	C/T
rs4013967	NA	9	76897070	NA	C/T
rs10510791	intron-variant	3	57294085	APPL1	C/G
rs10520440	NA	4	180799005	NA	G/T
rs10515919	NA	2	75540596	NA	A/G
rs6684063	NA	1	30699340	NA	G/T
rs354747	NA	20	58912660	NA	A/G
rs9302185	NA	15	54954864	NA	C/T
rs10506816	NA	12	79924857	NA	A/T
rs2585901	intron-variant	13	21420271	XPO4	C/T
rs993314	intron-variant	6	73438572	KCNQ5	C/T
rs1919550	intron-variant	3	121364173	HCLS1	A/T
rs888861	NA	19	35381852	NA	A/G
rs1073319	intron-variant	2	29440454	ALK	A/G
rs1934393	intron-variant	1	49208618	BEND5	C/G
rs2711070	intron-variant	2	159502533	PKP4	C/G
rs1013459	intron-variant	18	11700534	GNAL	A/G
rs9295316	intron-variant	6	158567066	SERAC1	A/T

rs10491654	NA	9	102139527	NA	C/T
rs802524	intron-variant	7	145951642	CNTNAP2	C/T
rs6911727	NA	6	9116398	NA	C/T
rs719776	NA	4	33686660	NA	C/G
rs9323178	NA	14	23113646	NA	A/G
rs10492585	NA	13	105386176	NA	C/T
rs10248051	intron-variant	7	51119353	COBL	C/T
rs1984473	NA	3	155811284	NA	C/T
rs1036543	intron-variant	2	133676214	NCKAP5	C/T
rs3806218	upstream-variant-2KB	1	147011783	BCL9	A/G
rs1898280	NA	8	116074460	NA	A/G
rs3860446	NA	2	104489351	NA	C/T
rs3828121	intron-variant	1	82422200	LPHN2	C/T
rs153898	intron-variant	5	94188622	MCTP1	C/T
rs1353251	intron-variant	5	35857207	IL7R	A/G
rs2829454	intron-variant	21	26273071	LOC339622	A/G
rs4852696	NA	2	83151641	NA	C/G
rs2035573	intron-variant	3	131213506	MRPL3	C/T
rs4657449	intron-variant	1	165465281	LOC400794	A/G
rs1395771	NA	3	96463580	NA	A/G
rs6883095	NA	5	79891047	NA	A/G
rs30125	intron-variant	16	14354661	MKL2	A/G
rs10131076	intron-variant	14	80774385	DIO2-AS1	A/G
rs4762106	intron-variant	12	66018473	LOC100507065	A/G
rs1398829	NA	4	22023275	NA	A/T
rs4130513	intron-variant	16	78458750	WWOX	C/T
rs1990745	NA	5	103381922	NA	C/T
rs10515535	NA	5	143516142	NA	A/G
rs4130405	NA	8	99420775	NA	A/C
rs3768176	intron-variant	1	57568005	DAB1	C/G
rs10507688	NA	13	63406228	NA	A/G
rs2592888	NA	1	159585573	NA	C/T
rs10501474	NA	11	80400647	NA	C/T
rs1451928	NA	14	48340741	NA	G/T
rs10491097	NA	17	19361211	NA	A/G
rs1004704	NA	16	48537421	NA	A/G
rs4733652	NA	8	129844527	NA	C/T
rs10214949	intron-variant	7	79048593	MAGI2	A/G
rs4034627	NA	12	128397472	NA	C/T
rs4934436	NA	10	90783320	NA	C/T
rs9307613	NA	4	130357404	NA	A/T
rs1477277	NA	5	180675022	NA	C/G
rs948360	intron-variant	11	66106725	BRMS1	A/G
rs6569792	intron-variant	6	132694751	MOXD1	A/G
rs6804094	NA	3	187057970	NA	A/T
rs4625554	NA	12	4416304	NA	A/G
rs7535375	intron-variant	1	235913159	LYST	C/T
rs304051	intron-variant	3	4578306	ITPR1	C/T

ANEXO 2

```
/*#=======================================
# DINÂMICA DA MISCIGENAÇÃO DE POPULAÇÕES DA AMÉRICA LATINA
#=======================================
# Controle de qualidade dos dados genéticos
#
# (C) Copyright 2012, by LDGH and Contributors.
#
# /
#/ -----------------
#  GLU - GENETICS
#  -----------------
#
# Original Author: Fernanda Rodrigues-Soares e Wagner Magalhães
# Contributor(s):
# Updated by (and date):
#
# Dependencies:  GLU GENETICS
#
# Command line:
glu qc.summary --hwp input.sdat -o output  # O programa gera um controle de qualidade dos dados,
                identificando os marcadores que estão dentro e fora do EHW

glu transform input.sdat -o arquivo.ldat   # Transforma um arquivo de SDAT para LDAT (inverte linhas
                por colunas)

glu util.join arquivo1.ldat arquivo2.ldat -o doisjuntos.ldat  # Junta linhas de dois arquivos de mesmo
                formato com colunas iguais.

glu transform  Nanda2.sdat --excludesamples=africanos.lst -o Nanda2semafr.sdat  # O programa
                exclui as amostras desejadas, através de um arquivo com os identificadores das
                amostras em uma única coluna.
#
# Sample input files: example.sdat, example.ldat, example2.ldat, example.lst
#
# Arquivo .sdat: ID samples (linhas) x SNPs (colunas).
# Arquivo .ldat: SNPs (linhas) x ID samples (colunas).
# Arquivo .lst: Lista em .txt
#
#######
```

ANEXO 3

```
/*#=======================================
# DINÂMICA DA MISCIGENAÇÃO DE POPULAÇÕES DA AMÉRICA LATINA
#=======================================
#
# (C) Copyright 2012, by LDGH and Contributors.
#
# Cálculos de frequências alélicas, Hs esperada, Fst par a par e PCA
# /
#/ -----------------
#  R - ADEGENET
#  -----------------
#
# Original Author: Giordano Soares-Souza e Fernanda Rodrigues-Soares
# Contributor(s): Wagner Magalhães
# Updated by (and date):
#
# Dependencies:  R, ADEGENET, ADE4
#
# Input file. Example.sdat (arquivo .sdat com uma coluna adicional após a coluna de ID samples com
# o ID das populações de cada amostra)

# Cálculo de frequências alélicas
pop <- read.table(file.choose (), head = TRUE, row.names=1)
pop1 <- data.frame(t(pop))
popt <- data.frame(t(pop1))
pop2 <- df2genind(X = popt[, -1], pop = popt[, 1])
sum(is.na(pop2$tab))
pop3 <- na.replace(pop2, method = "mean")
pop4 <- genind2genpop(pop3)
Xfreq <- makefreq(pop4,quiet=FALSE,missing=NA,truenames=TRUE)

# Cálculo de heterozigosidade esperada
Hs(pop4, truenames=TRUE)

# Cálculo de Fst
pairwise.fst(pop3, res.type=c("dist","matrix"), truenames=TRUE)

# Cálculo de PCA

colorido = c("blue", "darkmagenta", "brown4", "azure4", "chocolate1", "darkgreen", "darkgoldenrod",
"gray0", "red1", "palevioletred4", "dodgerblue", "gold", "gray31", "deeppink")

pop <- read.table(file.choose (), head = TRUE, row.names=1)
pop1 <- data.frame(t(pop))
popt <- data.frame(t(pop1))
obj1 <- df2genind(X = popt[, -1], pop = popt[, 1])
sum(is.na(obj1$tab))
obj2 <- na.replace(obj1, method = "mean")

pca1 <- dudi.pca(obj2$tab, cent = TRUE, scale = FALSE, scannf = FALSE, nf = 3)
barplot(pca1$eig[1:50], main = "Eigenvalues")

s.class(pca1$li, obj2$pop, lab = obj2$pop.names, sub = "PCA 1-2", csub = 2, col = colorido)
add.scatter.eig(pca1$eig[1:20], nf = 3, xax = 1, yax = 2, posi = "bottomright")

s.class(pca1$li, obj2$pop, xax = 1, yax = 3, lab = obj2$pop.names, sub = "PCA 1-3", csub = 2, col =
colorido)
add.scatter.eig(pca1$eig[1:20], nf = 3, xax = 1, yax = 3, posi = "bottomright")
```

```
obj3 <- genind2genpop(obj1, missing = "chi2")

pca2 <- dudi.coa(as.data.frame(obj3$tab), scannf = FALSE, nf = 3)
barplot(pca2$eig, main = "Eigenvalues")

s.label(pca2$li, lab = obj3$pop.names, sub = "PCA 1-2", csub = 2)
add.scatter.eig(pca2$eig, nf = 3, xax = 1, yax = 2, posi = "top")

s.label(pca2$li, xax = 1, yax = 3, lab = obj3$pop.names, sub = "PCA 1-3", csub = 2)
add.scatter.eig(pca2$eig, nf = 3, xax = 2, yax = 3, posi = "bottomright")

# Porcentagem de variancia para cada componente
> pca1$eig
> pca1$eig[1]
> pca1$eig[1]/sum(pca1$eig)
> pca1$eig[2]/sum(pca1$eig)
```

ANEXO 4

```
/*#=======================================
# DINÂMICA DA MISCIGENAÇÃO DE POPULAÇÕES DA AMÉRICA LATINA
#=======================================
#
# (C) Copyright 2012, by LDGH and Contributors.
#
# Cálculo de AMOVA
# /
#/ -----------------
#          HIERFSTAT
#  -----------------
#
# Original Author: Wagner Magalhães
# Contributor(s):
# Updated by (and date):
#
# Dependencies:  R PLATFORM, HIERFSTAT PACKAGE
#
# AMOVA BY LOCUS
   data1 = read.table(as.matrix(file.choose()),head= TRUE, sep = "\t", na.string = "??")
   attach(data1)
   names(data1)
   levels = data1[,c(1)]
            loci1 = data1[,c(2: length(data1))]
   varcomp.glob(levels,loci1)

   data_result1 = matrix(nrow = length(loci1), ncol = 1)
   for (row_data1 in 1: length(loci1))
   {
      fct_values1 = varcomp(data.frame(levels,loci1[row_data1]))
      data_result1[row_data1,1] = fct_values1$F[1,1]
   }
   data_result1
```

ANEXO 5

```
/*#=======================================
# DINÂMICA DA MISCIGENAÇÃO DE POPULAÇÕES DA AMÉRICA LATINA
#=======================================
#
# (C) Copyright 2012, by LDGH and Contributors.
#
# Quantificação da miscigenação individual
# /
#/ -----------------
#  ADMIXTURE
#  -----------------
#
# Original Author: Fernanda Soares
# Contributor(s): Wagner Magalhães
# Updated by (and date):
#
# Dependencies:  GLU GENETICS, PLINK, ADMIXTURE, R
#
# Command line:
glu transform todas.sdat -o todas.ped    # GLU transforma o arquivo SDAT em PED.

plink --noweb --file todas --recode12 --out todaspops      # PLINK recodifica o arquivo PED para
                PED12, que o Admixture reconhece.

admixture arquivo.ped 3            # Linha de comando para rodar o programa admixture, com o nome
                do arquivo PED e o número de k.

# GRÁFICO BARRAS - R
tbl = read.table(file.choose(),sep = ' ')
barplot(t(as.matrix(tbl)),col = c("blue", "red1", "darkgreen"),beside=F,border=NA)
```

ANEXO 6

```
/*#=======================================
# DINÂMICA DA MISCIGENAÇÃO DE POPULAÇÕES DA AMÉRICA LATINA
#=======================================
#
# (C) Copyright 2012, by LDGH and Contributors.
#
#  Gráficos de distribuição de valores
# /
#/ -----------------
#          R
#  -----------------
#
# Original Author: Fernanda Soares
# Contributor(s): Wagner Magalhães
# Updated by (and date):
#
# Dependencies:  R PLATFORM
#
# GRÁFICO DE DISTRIBUIÇÃO DE VALORES
data = read.table(file.choose(),header=T)

data1 = density(data$AFR)
plot(data1, col = "blue", main = "Distribuiçao de Miscigenação - NIC", xlab = "Porcentagem de
miscigenação", xlim = c(0,1), ylim = c(0,6))

data2 = density(data$NAT)
data3 = density(data$EUR)

lines(data2, col = "darkgreen")
lines(data3, col = "red1")

##Arquivos com número de linhas diferentes##
#Arquivos separados

data = read.table(file.choose(),header=T)
data1 = density(data$r2)

plot(data1, col = "darkviolet", main = "LD - Miscigenadas", xlab = "Valores de r2", xlim = c(0,0.2), ylim
= c(0,70))

data2 = read.table(file.choose(),header=T)
data3 = density(data2$r2)
lines(data3, col = "darkorange")

data4 = read.table(file.choose(),header=T)
data5 = density(data4$r2)
lines(data5, col = "goldenrod")

data6 = read.table(file.choose(),header=T)
data7 = density(data6$r2)
lines(data7, col = "red1")
```