Universidade de Brasília
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Biologia Animal

Ancestralidade genética em populações miscigenadas: desafios, aplicações e um olhar sobre a história da formação do Distrito Federal

Luciana Maia Escher dos Santos

Brasília
2023

Universidade de Brasília
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Biologia Animal

Ancestralidade genética em populações miscigenadas: desafios, aplicações e um olhar sobre a história da formação do Distrito Federal

Tese apresentada ao Programa de Pós-Graduação em Biologia Animal do Instituto e Ciências Biológicas da Universidade de Brasília, como requisito parcial para obtenção do título de Doutora em Biologia Animal.

Candidata: Luciana Maia Escher dos Santos

Orientadora: Prof. Dra. Silviene F. de Oliveira

Co-orientadora: Dra. Kelly Nunes

Brasília
2023

Nos dias de muito cansaço eu descansei... eu não desisti.

Por me ensinarem a coragem e abraçarem esse desafio comigo,

incondicionalmente,

aos meus pais, Benevides e Marize,

Dedico

AGRADECIMENTOS

A palavra "agradecer" ***contém a raiz "gratus", do latim, que significa ser acolhido ou acolher com favor, de forma agradável*****.**

O caminho deste doutoramento foi longo, dificil e desafiador, não somente pelas experiências dentro da academia mas pela naturalidade delas sempre estarem vinculadas ao mundo de fora e ele nem sempre é bonito, acolhedor e leve. Chego até aqui para escrever esta parte de um trabalho que nunca foi um sonho, mas um objetivo, um grande, pesado e desafiador objetivo. Honradamente o encerro num concreto pilar de resiliência.

"Sil", minha tão tão querida orientadora. Obrigada por me colocar na tua bagagem, tão rica, tão bonita e tão nobre. Que honra fazer parte da tua "árvore" de orientandos. Uma honra ter tido mais que sua orientação acadêmica, mais que sua energia em dias muito cansativos, mas por ter recebido teu olhar humano em dias muito, muito difíceis. Obrigada sobretudo, por não desistir de mim. A ti, meu respeito, admiração e zelo. Que eu tenha saúde e força para seguir e fazer da tua orientação uma continuidade digna e bonita pelos caminhos da pesquisa.

"Kelinha", minha co-orientadora...você não estava no "Plano Piloto", mas chegou com essa calmaria, com cheiro de terra molhada e derrepente vimos uma tempestade de competência e conhecimento. Acolheu esse projeto com tanta força, nos doando muito da tua energia, que já era dispensada a tantas outras demandas. Particularmente eu te abraço, num abraço cheinho de gratidão. Houveram dias muito difíceis, que foi você, meus braços e meu olhar. Não tenha dúvida do quanto sua colaboração enriqueceu e transformou esse trabalho e minha formação. Á ti querida vai também a minha gratidão por não ter desistido de mim.

“Aos amigos do Laboratório de Genética Humana”, a todos, sem exceção, o meu muitíssimo obrigada pela parceria. Sintam-se abraçados, lembrados e queridos. De uma forma muito especial, um abraço cheinho de afeto “Matheus Castro”.

“Ás amigas Miriam e Marta”... carregadores portáteis de energia durante essa looooonga caminhada, compartilhando e comemorando cada conquista, cada avanço, mas também me percebendo nos dias difíceis, estando presente em diálogo, colo, zelo e força.

“***Sidney”***, meu irmão, meu parceiro para toda hora. 26 anos de amizade e compartilhamento de desafios...Obrigada querido por sempre tentar fazer tudo dar certo, pela torcida, apoio e por sempre me fazer rir !!!! te amo !!

“Á minha família”... parafraseio Isaac Newton “Se eu vi mais longe, foi por estar sobre ombros de gigantes”. São vocês os gigantes desta caminhada. Aos meus pais e irmãos, obrigada por acreditarem em mim até quando eu mesma já não acreditava. Pelo incentivo, amor e amizade. Aos meus filhos, toda gratidão do meu coração por existirem, e me trazer todo dia a responsabilidade de ser melhor que ontem.

“Ás Agências de Financiamento”, CAPES, CNPq, FAPDF, pelo suporte, fazendo sem dúvida, diferença na sobrevivência acadêmica. Levando mais dignidade aos alunos que honrosa e corajosamente se aventuram pelas raízes da acadêmia.

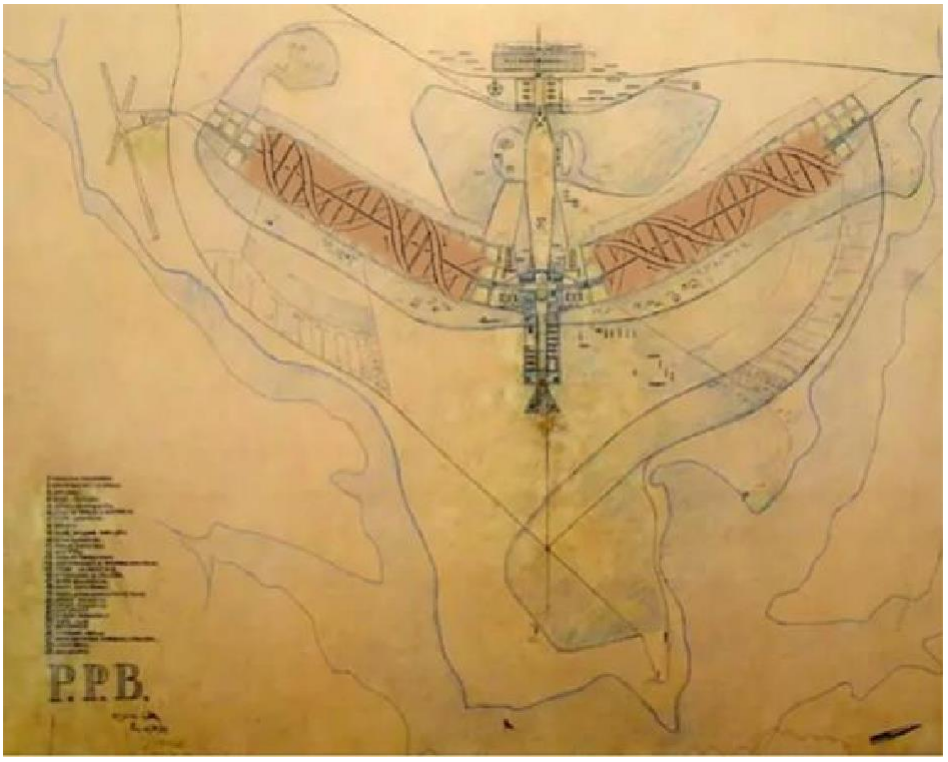


Figura adaptada do Plano Piloto de Brasília, de autoria de Lúcio Costa, vencedor no Concurso Nacional do Plano Piloto da Nova Capital do Brasil, em 1956. Inserida na figura estão moléculas de DNA representando a contribuição de cada imigrante na história de ancestralidade genética desta população. Na década de 50, no Planalto Central Brasileiro, havia toneladas de concreto armado se transformando, mas tomando emprestada as lentes da genética populacional, num olhar atento e sensível vê-se também acontecendo junto ao novo e ousado projeto de Juscelino Kubitschek, **um "balé" chamado fluxo gênico.** ***Genomas de centenas de imigrantes de diversas partes do país se movimentando no palco árido de cerrado, construindo uma população inteira. Nasce um recorte do Brasil em meio ao Planalto Central. Como disse o compositor e poeta Tom Jobim, "seria o "herdeiro" de todas as culturas, de todas as "raças", com um sabor todo próprio".***

Lu Escher

LISTA DE ABREVIAÇÕES

1KGP	Projeto 1000 Genonas (do inglês *1000 Genomes Project*)
ACB	Afro-Caribenho
AIM	Marcadores Informativos de Ancestralidade (do inglês *Ancestry Informative Markers*)
AISNP	SNPs Informativo de Ancestralidade (do inglês *Ancestry Informative SNPs*)
ASW	Afro-Americano
CAAE	Certificado de Apresentação para Apreciação Ética
CLM	Colombiano
CODEPLAN	Companhia de Planejamento do Distrito Federal
CONEP	Comissão Nacional de Ética em Pesquisa
DF	Distrito Federal
DNA	Ácido Desoxirribonucléico (do inglês *Deoxyribonucleic Acid*)
EAS	Leste Asiático
EHW	Equilíbrio de Hardy-Weinberg
FAP	Fundação de Apoio à Pesquisa
HDSNP	Array de Alta Densidade de SNPs (do inglês High-density SNP array)
HGDP	Projeto Genômico da Diversidade Humana (do inglês *Human Genome Diversity Project)*
IBGE	Instituto Brasileiro de Geografia e Estatística
INCOR	Instituto do Coração
LD	Desequilíbrio de Ligação (do inglês *Linkage disequilibrium*)
mtDNA	DNA Mitocondrial
MXL	Mexicano
NAM	Nativo Americano
NGS	Sequenciamento de Nova Geração

	(do inglês *Next Generation Sequencing)*
NRY	Cromossomo Y região não recombinante
PCA	Análise de Componente Principal
	(do inglês *Principal Component Analysis*)
pb	Par de base
PEL	Peruano
PUR	Porto Riquenho
QC	Controle de Qualidade (do inglês *Quality Control*)
SABE	Coorte Brasileira Saúde e Bem Estar
	(do inglês - *Brazilian Cohort of Health, Well-being and Aging*)
STR	Microssatélites (do inglês *Short Tandem Repeat*)
SNP	Polimorfismo de Nucleotídeo Único
	(do inglês *Single Nucleotide Polymorphism*)
TCLE	Termo de Consentimento Livre e Esclarecido
UnB	Universidade de Brasília
USP	Universidade de São Paulo
WGS	Sequenciamento Total do Genoma
	(do inglês *Whole Genome Sequencing)*

LISTA DE TABELAS

CAPÍTULO 1

Tabelas suplementares (arquivos eletrônicos)

Tabelas Notas Suplementares (arquivos eletrônicos)

genomic position in build 38 (Position hg38), reference SNP ID number (SNP ID); AIMs panel in which the SNP is contained (PANEL), reference of the study which describes the AIMS panel (REFERENCE).

Table N2. List of markers with significantly different allelic frequencies between the parental groups of the HGDP and 1KGP dataset.

Table N3. Difference between the maximum expected value of genetic ancestry in each parental group and the median and mean value inferred according to each assessed linkage disquilibrium coefficient.

CAPÍTULO 02

Tabela 01. Distribuição do grupo amostral entre 18 regiões administrativas do DF existentes à época do censo de 1989.

Tabela 02. Autodeclaração, de acordo com as categorias de "cor/raça" utilizadas pelo IBGE, da amostra analisada da população do DF.

Tabela 03. Teste Z para a comparação das proporções de ancestralidade genética entre o DF e as demais regiões geográficas do Brasil.

Tabela 04. Modelo demográfico simplificado de miscigenação do DF.

Tabela 05. Teste do modelo demográfico migratório para o DF.

Tabela 06. Inferência média em porcentagem da ancestralidade genética estimada para cromossomos autossômicos, sexuais (X e Y) e DNA mitocondrial (mtDNA).

Tabela 07. Comparação das proporções de ancestralidade genética inferida para o cromossomo Y entre o DF e as demais regiões geográficas do Brasil utilizando teste Z.

Tabela 08. Haplogrupos identificados na população brasileira.

Tabela 09. Haplogrupos mitocondriais identificados no presente estudo.

Tabela 10. Teste Z para a comparação das proporções de ancestralidade genética inferida para o mtDNA entre o DF e as demais regiões geográficas do Brasil.

LISTA DE FIGURAS

CAPÍTULO 01

CAPÍTULO 2

APÊNDICES

SUMÁRIO

RESUMO GERAL

O perfil de miscigenação da população brasileira está sendo construído ao longo dos últimos cinco séculos, a partir de uma complexa combinação de povos originários nativos americanos, povos europeus, africanos, asiáticos e do Oriente Médio. Os estudos de genética populacional sempre tiveram um papel fundamental para auxiliar a desvendar a história evolutiva das populações e nas últimas décadas com as transformações tecnologias e o acesso à dados genéticos em larga escala novas nuances da história podem ser reveladas. No primeiro capítulo desta tese apresentamos a comparação da inferência de ancestralidade genética a partir de diferentes conjuntos de marcadores genéticos: Marcadores Informativos de Ancestralidade (AIMs), array de alta densidade de SNPs (*High Density SNPs - HDSNPs*) e Sequenciamento do Genoma Completo (*Whole Genome Sequencing* - WGS); e diferentes modelos de miscigenação: tri-híbrido (africano, europeu e nativo americano) e o tetra-híbrido (africano, europeu, nativo americano e leste asiático). Para tanto, analisamos amostras miscigenadas, sendo 1.171 do Brasil e 504 provenientes de populações da América, genotipadas por WGS, com as quais comparamos: (i) 6 painéis AIMS (34AISNP+PIMA; 55AISNP; 128AISNP; 170AISNP; 446AISNP; 665AISNP), (ii); um array de alta densidade de SNPs (HDSNPs - Axiom Human Origins, Thermo Fisher Scientific) e (iii) dados de WGS. Mostramos que tanto a escolha do conjunto de marcadores como do modelo de miscigenação interferem nas inferências de ancestralidade em populações miscigenadas. A menor correlação entre os conjuntos de marcadores e modelos testados foi para o componente ancestral Nativo Americano. As melhores performances, em especial na capacidade de distinção entre os componentes Nativo Americano e Leste Asiático foram observadas para os dados de HDSNPs e WGS. Concluímos que a escolha do conjunto de marcadores e do modelo de miscigenação deve estar de acordo com a história de cada população miscigenada e do propósito de cada estudo. No segundo capítulo, caracterizamos o perfil de ancestralidade genética de uma amostra populacional miscigenada brasileira do Distrito Federal (DF). Situado no Centro-Oeste do Brasil, o DF recebeu um fluxo rápido, amplo e diversificado de individudos de todas as partes do território nacional na década de 60 quando foi construída a capital federal. A fim de reconstruir a história dessa população, realizamos análises do perfil de ancestralidade genética e coletamos informações demográficas numa amostra de 104 indivíduos, nascidos no DF na década de 80. Com base nas informações geradas no capítulo 1, realizamos a genotipagem dessa amostra com HDSNPs e inferimos a ancestralidade genética a partir de um modelo de miscigenação tetra-híbrido. Nosso estudo revelou: (i) as proporções médias de ancestralidade genética observadas a partir das análises dos cromossomos autossómicos foram de 69,95% (±18%) de ancestralidade europeia, 19,8% (±14,4%) africana, 8,57% (±7,2%) nativo americana e 1,68% (±8,6%) leste asiática; essas proporções não apresentam diferenças significativas em relação a média da população brasileira. (ii) As análises dos cromossomos sexuais (X, Y) e do DNA mitocondrial mostraram assinaturas do fenômeno demográfico de acasalamento direcional, com predomínio da contribuição de homens de ascendência europeia e de mulheres com ascendência africana e nativa americana. (iii) As análises demográficas indicam que os ancestrais dos participantes do estudo migraram principalmente das regiões Sudeste (43,64%) e Nordeste (38,22%), seguida pelas regiões Centro-Oeste (9,39%), Norte (5,52%) e Sul (1,65%); (iii) por fim, também foi possível constatar que 51,2% dos matrimônios ocorreram entre indivíduos da mesma região geográfica. Nosso estudo mostra que o perfil de ancestralidade genética observado no DF é uma síntese das contribuições migratórias internas recentes do Brasil e de processos históricos anteriores a essa migração que deixaram fortes assinaturas genéticas no DNA e foram herdadas pela população do DF.

ABSTRACT

The profile of Brazilian miscegenation was built over the last five centuries, from a complex combination of indigenous, European, African, Asian and Middle Eastern peoples. Population genetics studies have played a key role in unraveling the evolutionary history of populations and, in recent decades, with transformed technologies and access to large-scale genetic data, new layers of history could be revealed.In the first chapter of this thesis, we compared the genetic ancestry inference from different sets of genetic markers: Ancestry Informative Markers (AIMs), high-density SNPs array (HDSNPs) and Whole Genome Sequencing (Whole Genome Sequencing - WGS). We also tested different admixture models: the tri-hybrid (African, European and Native American) and tetra-hybrid (African, European, Native American and East Asian). To do so, we analyzed 1,171 unrelated samples from Brazil and 504 unrelated samples from admixed American populations from the 1000 Genomes Project, both genotyped by WGS, with which we compared: (i) 5 AIMs panels (34AISNP+PIMA; 55AISNP; 128AISNP; 170AISNP; 446AISNP), (ii); a high-density array of SNPs (Axiom Human Origins - Thermo Fisher Scientific) and (iii) Whole Genome Sequencing (WGS) data. We show that both the choice of marker set and the admixture model interfere with ancestry inferences in admixed populations. The lowest correlation between the marker sets and models tested was for the inferred Native American ancestral component. The best performances, especially in the ability to distinguish between the Native American and East Asian components, were from the HDSNPs and WGS data. Finally, we conclude and recommend that the choice of the set of markers and the admixture model will depend on the history of each admixed population and the purpose of the study. In the second chapter of the thesis, we characterize the genetic ancestry profile of an admixed Brazilian population sample from the Federal District. Based on the information generated in Chapter 1, we chose to genotype this sample with HDSNPs (Axiom Human Origins - Thermo Fisher Scientific) and infer the genetic ancestry from a tetra-hybrid admixture model. The Federal District (DF) is a region located in the Midwest of Brazil, receiving a fast, wide and diversified flow of individuals from all parts of the national territory from the 60's, when the federal capital was built. To reconstruct the history of this population, we performed analyzes of the genetic ancestry profile and collected demographic information on a sample of 104 individuals, born in the DF in the 1980s. Our study revealed: (i) based on the inference of genetic ancestry from the chromosomes autosomal the average proportions observed were 69.95% (±18%) of European ancestry, 19.8% (±14.4%) African, 8.57% (±7.2%) Native American and 1.68 %(±8.6%) East Asia; these proportions is not significant different from the average of the Brazilian population. (ii) Analysis of sex chromosomes (X, Y) and mitochondrial DNA showed signatures of the demographic phenomenon of directional mating, with a predominance of contributions from men of European descent and women of African and Native American descendants. (iii) Demographic analyzes indicate that the study participants' ancestors migrated mainly from the Southeast (43.64%) and Northeast (38.22%) regions, followed by the Midwest (9.39%), North (5 .52%) and South (1.65%); with 51.2% of marriages occurring between individuals from the same geographic region. Our study shows that the genetic ancestry profile observed in the DF is a synthesis of recent internal migration contributions from Brazil and historical processes prior to this migration that left strong genetic signatures in the DNA and was inherited by the DF population.

1 INTRODUÇÃO GERAL

1.1 ANCESTRALIDADE E ANCESTRALIDADE GENÉTICA

O termo ancestralidade é bastante amplo, remete a algo herdado de nossos ancestrais e indica uma conexão com algum passado, seja ele distante ou recente, cultural, social ou biológico. De acordo com a natureza que o termo ancestralidade se apresenta, este pode ser interpretado e discutido por diferentes lentes com diferentes significâncias: culturais, religiosas, políticas ou biológicas, exercendo influência na formação das identidades pessoais ou coletivas. Quando se fala em ancestralidade genética, estamos nos referindo à similaridade genética entre populações e indivíduos, ou seja, ao grau de compartilhamento do perfil de variantes genéticas e seu padrão de desequilíbrio de ligação no genoma de um indivíduo com o de outro, ou de uma outra população. Quanto maior o grau de similaridade genética, maior é o número de ancestrais recentes que dois indivíduos compartilham, quanto menor for esse grau, menor e mais remotos são os ancestrais comuns entre eles (Novembre & Kang 2015; Mathieson & Scally, 2020).

As ferramentas de busca sobre ancestralidade de um indivíduo podem se valer apenas de dados históricos, ou genealógicos, sendo possível traçar uma árvore e observar como os ancestrais, dos quais se tem conhecimento, estão relacionados entre si. Esse entendimento é mais limitado, uma vez que o conhecimento sobre ancestrais geralmente se restringe a algumas poucas gerações, ou ainda informações unilaterais (matrilinear ou patrilinear) (Mathieson & Scally 2020). A partir destas informações é possível traçar, por exemplo, um heredograma representativo de como os ancestrais genealógicos estão conectados entre si. A riqueza de informações existentes em um heredograma estará necessariamente alinhado à quantidade de informações que se têm sobre gerações passadas, levando para esse gráfico pontos de conexões e caminhos distintos tomados por linhagens distintas, sugerindo tempo e espaço para estes ancestrais. As inferências serão naturalmente mais informativas e precisas se nelas houver a incorporação de dados genéticos (Figura 01).

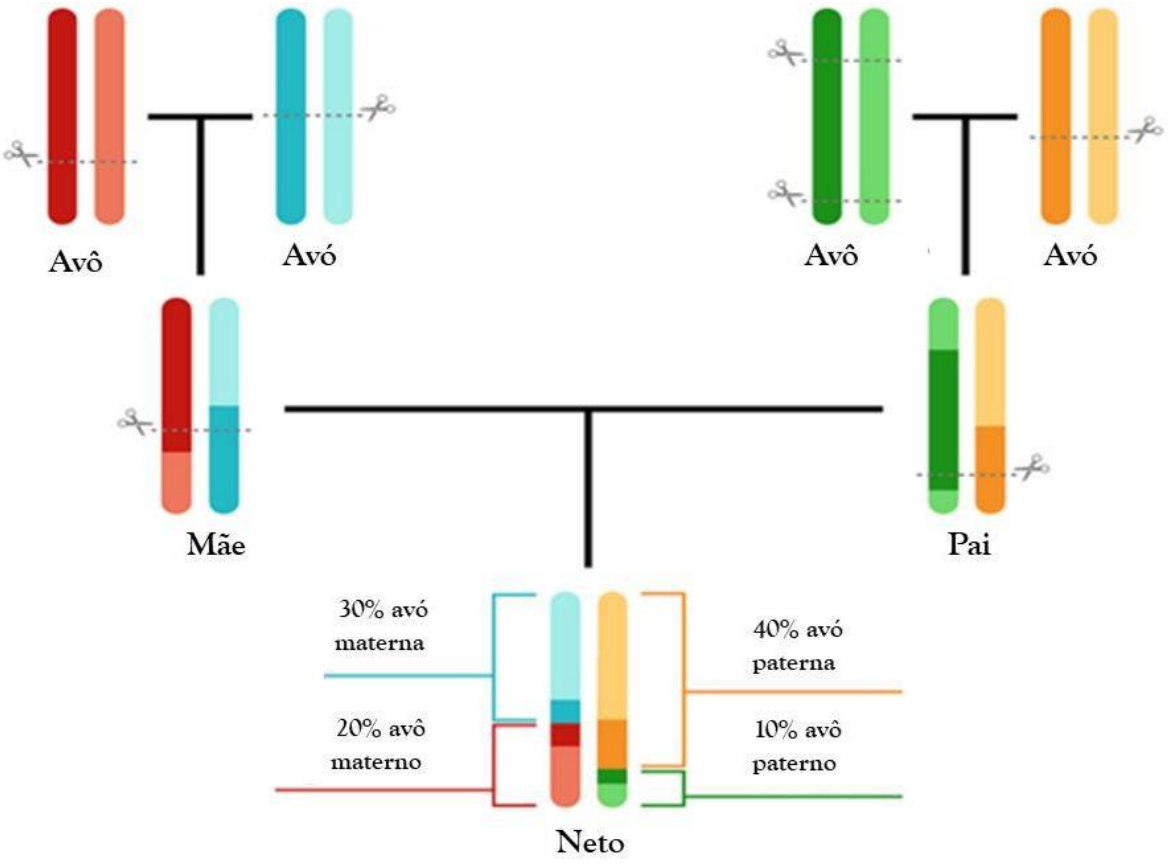


Figura 01. Ascendência genealógica e genética. A figura mostra esquematicamente o compartilhamento do material genético ao longo das gerações em uma genealogia. O DNA contido nos cromossomos autossômicos está representado por barras coloridas, onde cada cor indica uma origem parental distinta. Na parte superior da figura encontra-se representada a geração dos avós, na parte intermediária dos pais e na inferior do indivíduo referência. A tesoura seguida da linha pontilhada indica pontos no cromossomo onde ocorreu o processo de recombinação meiótica. A proporção de ancestralidade herdada de cada um dos avós descrita na figura é ilustrativa e pode variar de individuo para indivíduo. Fonte: https://www.thetech.org/ask-a-geneticist/no-native-american-ancestry-in-results.

A ancestralidade quando inferida a partir de dados genéticos, denominada de "Ancestralidade Genética", segundo Mathieson & Scally (2020), difere da ancestralidade genealógica porque se refere ao subconjunto de caminhos pelos quais o material em seu genoma foi herdado. Assim, por exemplo irmãos, apesar de compartilharem os mesmo ancestrais e a mesma árvore genealógica, herdam material genético distinto, pois o processo de recombinação e o sorteio dos cromossomos do par homólogo que constituirão o conjunto cromossômico dos gametas, diferem em cada meiose. Neste âmbito, a genética de populações é uma disciplina que disponibiliza ferramentas metodológicas que possibilitam o entendimento de como ocorrem as combinações e transformações do material genético ao longo do tempo e do espaço (Novembre & Kang 2015).

A ciência tem possibilitado um olhar transformador sobre o passado humano através do acesso a dados genéticos antigos e modernos reveladores de perfis ancestrais ao longo do tempo. Metodologias das mais variadas têm sido desenvolvidas para inferir e visualizar essas relações num esforço de acessar padrões genéticos cada vez mais informativos sobre a origem geográfica de um indivíduo (revisão em Suarez-Pajes *et al*., 2021). O interesse em quantificar e caracterizar os grupos populacionais humanos começou a quase um século, sendo inicialmente utilizados variantes de grupos sanguíneos e/ou marcadores protéicos. Richard Lewontin (1972),demonstrou a partir da análise de grupos sanguíneos e outros marcadores em populações de diversas partes do mundo, que aproximadamente 90% da variação biológica é encontrada dentro (intra) dos chamados "grupos raciais", e não entre eles. Esses dados mostraram que do ponto de vista biológico os indivíduos são muito mais parecidos do que se imaginava e portanto os "grupos raciais" humanos, definidos por abordagens tipológico-raciais, não encontram suporte biológico. Estudos posteriores confirmaram esses resultados prévios e indicaram que as diferenças observadas entre populações humanas de continentes distintos foi cerca de 7% (Rosemberg *et al.,* 2002). A partir do sequenciamento do genoma humano (*International Human Genome Sequencing Consortium* 2004, Venter *et al.,* 2001) foi possível observar que se tomarmos ao acaso dois indivíduos e compararmos seus genomas, encontraremos 99,9% de semelhanças genéticas entre eles. Ou seja, apenas 0,1% das variações genéticas são responsáveis pelas diferenças fenotípicas (ex. altura, cor cabelo, cor dos olhos), susceptibilidade a doenças, respostas diferencial à medicamentos e diferenças entre indivíduos de populações distintas (Jorde & Wooding 2004). Portanto, é dentro desta pequena proporção de diferença (7% do total de 0,1% de diferenças) que existem variantes genéticas capazes de informar sobre a origem biogeográfica de um indivíduo ou a sua ancestralidade genética (Kosoy *et al.*, 2009); variantes estas que exibem diferencial de frequencia alelica superior a 30% entre quaisquer duas populações parentais (Bonilla *et al,* 2004).

Atribuir de modo preciso a origem biogeográfica de um indivíduo não é uma tarefa simples, pois a variação genética humana ocorre de forma gradual e não abrupta. A distribuição da diversidade e variação genética é reflexo da história evolutiva (seletiva e demográfica) das populações humanas durante todo o percurso de saída do homem moderno do continente Africano até os dias de hoje (Ramachadran *et al.,* 2005; Liu *et al.,* 2006).

Com isso em mente, várias estratégias foram desenvolvidas no intúito de se buscar novas variantes que reflitam de maneira mais acentuada as diferenças entre os grupos populacionais para a inferência da origem biogeográfica de um indivíduo ou de seus ancestrais. Uma das primeiras estratégias, consiste na identificação de "Marcadores Informativos de Ancestralidade" - AIMs (do inglês *Ancestry Informative Markers*), que são conjuntos de marcadores genéticos que apresentam frequências alélicas significativamente diferentes entre populações de diferentes regiões geográficas do mundo (Fondevila *et al*., 2011; Santangelo *et al.,* 2017). Para identificar quais variantes genéticas são mais informativas para as inferências de ancestralidade, várias medidas de diferenciação populacional foram propostas (Rosenberg *et al.,* 2003). A partir dessa identificação são criados "painéis de ancestralidade", que consiste em conjuntos formados por dezenas ou centenas desses AIMs com a finalidade de associar esse conjunto de variantes de um indivíduos a uma ou mais regiões geográficas continentais específicas. Essa estratégia de inferência a partir de painéis de ancestralidade dominou o cenário científico nas últimas três décadas (Phillips *et al.,* 2007; Cheung *et al.,* 2019).

Avanços significativos na área da genética e genômica, com as tecnologias de alto desempenho, tem permitido o acesso a informações genômicas em larga escala (SNPArray e/ou Genomas completos), levando ao desenvolvimento de novas estratégias analíticas. Com acesso a centenas de milhares de variantes genéticas, incluindo aquelas com poder de informatividade geográfica (AIMs) a abordagem agora utilizada parte de um olhar da variante em seu contexto genômico, como por exemplo, partindo do padrão de desequilíbrio de ligação entre a variante e outras adjacentes, visto que esse perfil se mostra peculiar a cada população (Slatkin 2008; Loh *et al.,* 2013). A informação sobre desequilíbrio de ligação, na maioria das vezes é incorporada a partir da definição de haplótipos (combinação de alelos de *loci* adjacentes e que são herdados de forma conjunta) e a comparação desses com aqueles observados nos diferentes grupos populacionais parentais. Diferentes modelos matemáticos foram desenvolvidos usando essa abordagem e buscando maior robustez na inferência de ancestralidade a partir dos dados genômicos em larga escala (Lawson *et al.,* 2012 Mapples *et al.,* 2013; Montserrat *et al.,* 2020). Além disso, ao invés de inferir a ancestralidade a partir da média de AIMs esparsos ao longo do genoma (chamada ancestralidade global), hoje a alta densidade de marcadores permite o desenvolvimento de metodologias próprias e a inferência da ancestralidade a partir de regiões

cromossômicas de interesse (ancestralidade local) (Mapples *et al.,* 2013).

Essa mudança tecnológica será um dos temas explorados na presente tese. No capítulo 1, apresentamos como a escolha de diferentes painéis de ancestralidade: (i) AIMs, (ii) SNPArray ou (iii) dados de Sequenciamento do Genoma Completo, impactam na atribuição biogeográfica e na inferência de ancestralidade genética dos indivíduos. No primeiro capítulo da tese também trazemos um olhar especial às populações miscigenadas e aos desafios de inferir a ancestralidade genética para este grupo populacional.

1.2. ANCESTRALIDADE GENÉTICA EM POPULAÇÕES MISCIGENADAS

Migrações e miscigenação são processos que sempre acompanharam as populações humanas durante sua história evolutiva. Contudo, alguns grupos populacionais continentais divergiram e permaneceram "isolados" ou com baixo fluxo gênico entre eles por milhares de anos. É o caso de africanos e europeus que divergiram há mais de 40.000 anos, e dos povos originário da América (indígenas), que após entrarem no continente Americano pela ponte de gelo no estreito de Bering ficaram isolados dos demais continentes entre 20-15.000 anos. Durante esse período, cada grupo populacional vivenciou seu próprio processo demográfico (com expansões, gargalos populacionais, deriva genética) e seletivo (como, por exemplo, pressões por conjuntos de patógenos, mudanças climáticas e ambientais), acumulando ao longo de seus genomas padrões particulares de variação genética (Balaresque *et al.,* 2007; Lopez *et al.,* 2016).

Com o advento das grandes navegações no século XV, ocorre uma grande diáspora na população humana e, com o retorno do fluxo gênico, indivíduos dessas populações relativamente isoladas voltam a se encontrar. Aqui na América, o encontro desses grupos populacionais propicia o surgimento de um padrão único de miscigenação intercontinental, num primeiro momento, entre os povos originários da América com colonizadores europeus e africanos escravizados (Jobling *et al.,* 2014).

Compreender uma população miscigenada do ponto de vista genético significa entender que neste genoma existem proporções diferentes de populações parentais originais (Figura 02). Ao longo das gerações, o processo de recombinação meiótico gera segmentos cromossômicos derivados dos diferentes conjuntos de populações parentais. A duração, direção e a taxa de fluxo gênico entre as populações parentais

influenciam na proporção de miscigenação e no tamanho dos segmentos cromossômicos. Isso faz com que, com o decorrer das gerações, os cromossomos dos indivíduos miscigenados sejam verdadeiros mosaicos de segmentos com diferentes origens ancestrais (revisão em Soares-Souza *et al.,* 2018).

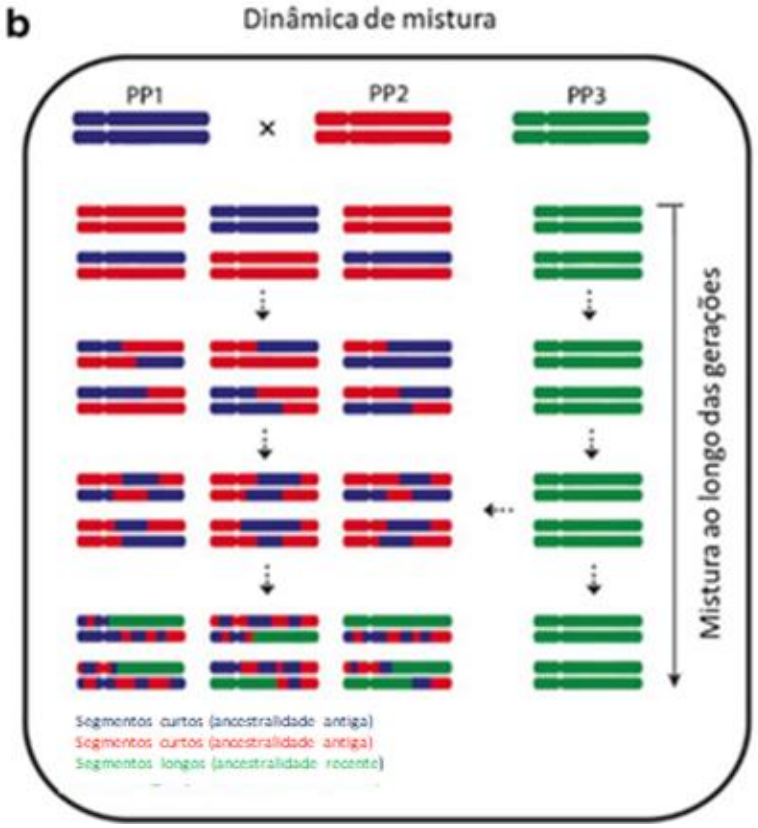


Figura 02. Processo de miscigenação cromossômica: Padrão esquemático de ancestralidade cromossômica após gerações de cruzamento entre populações parentais previamente isoladas (PP1, PP2 e PP3). Assim que uma população miscigenada surge, os cromossomos dos indivíduos miscigenados são inteiramente ou quase inteiramente de um único ancestral. Ao longo das gerações, devido ao processo de recombinação intracromossômica, os segmentos cromossômicos de ancestrais distintos serão emendados em pedaços menores à medida que o número de gerações aumenta desde o evento de miscigenação. Pelo processo de recombinação intracromossômica em cada geração, a ancestralidade recentemente introduzida aparece em segmentos cromossômicos mais longos (ancestralidade PP3, verde) do que ancestrais introduzidos anteriormente (PP1 e PP2, azul e vermelho). Informações sobre o comprimento médio de segmentos cromossômicos de ancestrais distintos podem ser úteis para inferir o número de gerações desde a miscigenação.

Fonte: Soares-Souza *et al.* (2018) (modificada).

Outra peculiaridade das populações miscigenadas é que a ancestralidade genética varia em diferentes níveis (Figura 03): (i) entre população; (ii) entre indivíduos da mesma população; (iii) entre os cromossomos de um mesmo indivíduo (ver exemplo em Gopalan *et al*., 2022). Esse padrão de variação da ancestralidade genética interpopulacional, intra-populacional e intra-individual tem implicação direta nos estudos com populações miscigenadas. Ongaro et al (2019) avaliou os padrões de miscigenação em populações da América do Norte, Caribe e América do Sul. O estudo observa diferenças nas proporções de cada componente ancestral continental entre as populações: tendo as populações afro-americanas e afro-caribenhas altas contribuições do componente africano (74,1% e 87,1% respectivamente), seguidas por Salvador no Brasil (47,8%). As maiores contribuições nativo americanas foram observadas no Peru, México e Equador (59,2%, 41% e 37% respectivamente), enquanto que as maiores contribuições europeias em Bambuí, no Brasil, Porto Rico e Colômbia (82%, 79% e 66% respectivamente). Além do perfil de miscigenação diferir entre as populações da América, são pontuadas duas observações adicionais: (i) a origem parental intracontinental varia entre as populações: por exemplo, populações afro americanas e caribenhas têm maior contribuição de populações originárias do Centro- Oeste (Benim-Nigéria) da África enquanto que Salvador, no Brasil, do Centro-Sul (Angola-Namibia); as populações também diferem quanto a origem do componente europeu, tendo as afro-americanas contribuição britânica, as caribenhas contribuição da França e América do Sul da Península Ibérica; (ii) por fim, observa-se que o número e a data dos eventos migratórios também variam entre as populações.

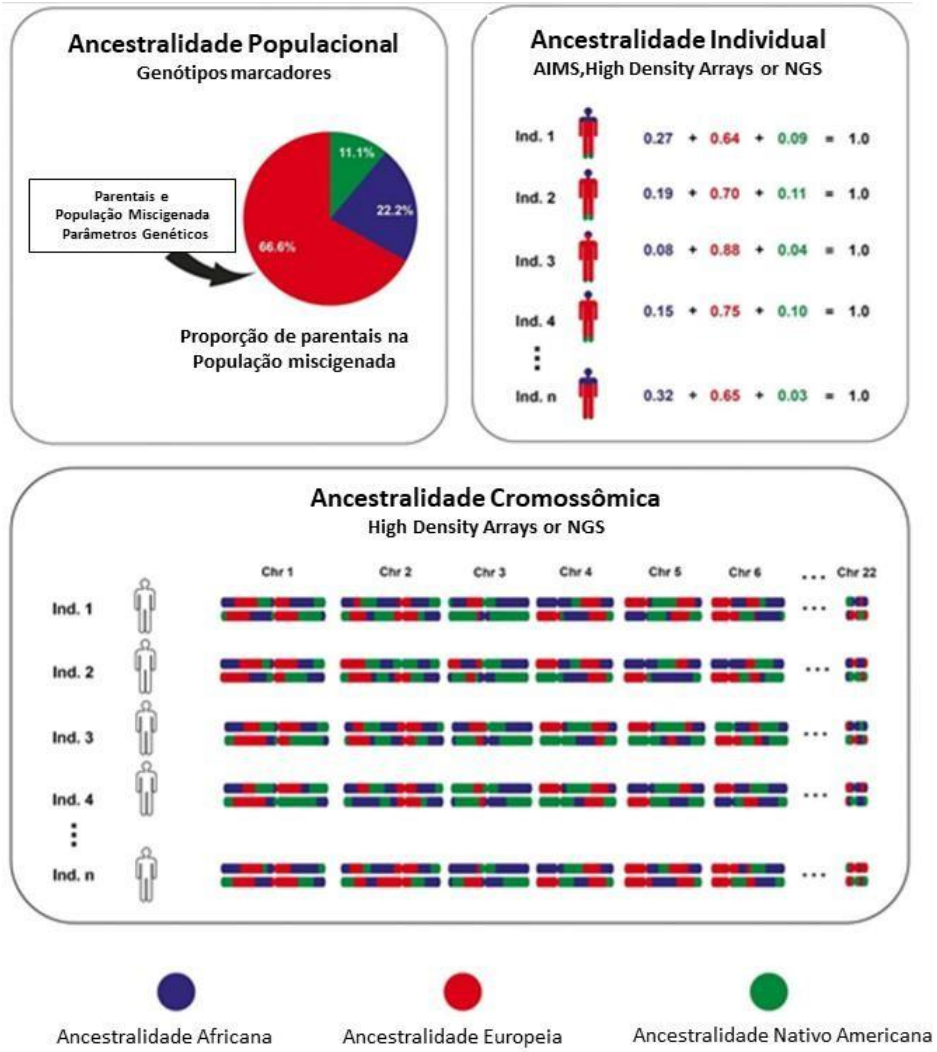


Figure 03. Níveis de variação da ancestralidade genética em indivíduos miscigenados. Em populações miscigenadas a ancestralidade genética varia em múltiplas escalas genéticas: entre populações, entre indivíduos e dentro de genomas individuais. Ancestralidade populacional corresponde às porcentagens do conjunto de genomas de todos os indivíduos da população provenientes de cada população parental, ou seja, é a ancestralidade individual média entre todos os indivíduos da população. Ancestralidade individual é a porcentagem do genoma de cada indivíduo originária de cada população parental. Ancestralidade cromossômica é a origem ancestral de cada fragmento de cada cromossomo de cada indivíduo As cores vermelha, azul e verde representam as diferentes ancestralidades parentais.

Fonte: Soares-Souza *et al.* (2018). (modificada)

Portanto, apesar dos modelos simplificados e das generalizações acerca do processo de miscigenação das Américas, observamos que cada população apresenta um conjunto de particularidades históricas. Por isso, a precisão da inferência de ancestralidade genética para esse tipo de população é tão desafiadora. Em geral, os painéis de ancestralidade e arrays de alta densidade de SNPs são elaborados desconsiderando a heterogeneidade das populações miscigenadas, seja na

composição das populações parentais usadas como referência para o desenho dos painéis/arrays, como na escolha dos modelos usados para inferir o processo de miscigenação (quais populações parentais contribuíram para o processo de formação da população miscigenada, quantas e quando ocorreram as ondas migratórias).

No capítulo 1 da tese exploraramos a acurácia dos painéis de ancestralidade e como a escolha do modelo de miscigenação influenciam na inferência de ancestralidade genética em diferentes populações miscigenadas da América: afro-americanos (EUA), afro-caribenhos (Barbados), mexicanos, colombianos, peruanos, porto riquenhos e brasileiros. As conclusões do primeiro capítulo conduziram as escolhas metodológicas aplicadas no segundo capítulo da tese, o qual consiste num olhar sobre a história de miscigenação e construção da população do Distrito Federal.

A seguir faremos uma breve revisão sobre a história de formação da população brasileira e suas implicações na formação da população do Distrito Federal.

1.3. A POPULAÇÃO BRASILEIRA E SEU PERFIL MULTIÉTNICO

Como comentamos brevemente no tópico anterior, a chegada dos europeus na América no século XV promoveu uma onda migratória intensa principalmente oriundos de dois continentes doadores: Europa e África, sendo a América a receptora de tais etnias (Figura 04). Estima-se que cerca de 15 milhões de europeus tenham cruzado o Atlântico até 1880. A maioria dos migrantes europeus eram jovens, adultos, do sexo masculino em busca de emprego, temporário ou permanente, em território brasileiro. Este último período migratório foi subsidiado pelo governo brasileiro. À época, os europeus viviam em um momento de recessão, onde a dificuldade de acesso à terra e ao alimento eram fatores de repulsão. Assim, sob incentivo e subsídio do governo brasileiro, migraram em massa para América, vislumbrando a possibilidade de acesso à terra e trabalho, uma vez que tinham a garantia de contratação para mão-de-obra em terras americanas (Klein, 2000).

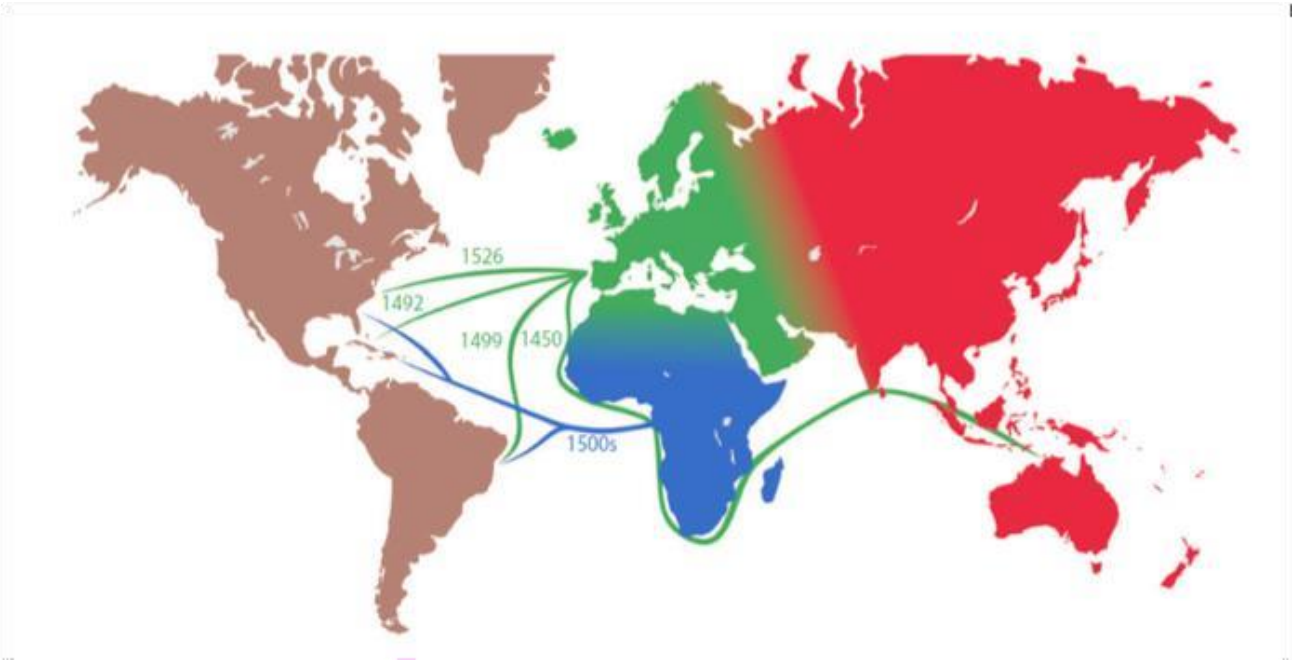


Figura 04. Mapa do processo de miscigenação na América iniciado com as grandes navegações no século XV. As linhas, o fluxo de migração e os números representam a data de início do movimento migratório. Fonte: Winkler et al. (2010).

Os portugueses foram os europeus que mais contribuíram para a colonização do Brasil. Estima-se que em um primeiro momento aproximadamente 700 mil portugueses tenham vindo para cá. A partir de 1700 o fluxo migratório se tornou mais intenso em virtude da descoberta de metais preciosos como o ouro. Entre o período de 1880 e 1960 houve uma migração em massa, com um número de imigrantes europeus estimado em mais de 1,5 milhões de indivíduos. Contribuíram ainda para povoar o Brasil, principalmente a partir de 1890, os italianos, que formaram o segundo maior contingente de migrantes europeus, seguido pelos espanhóis, que iniciaram a migração para o Brasil em épocas mais remotas desde o século XVI, e os alemães a partir de 1884. Outros grupos do Oriente Médio, como Sirío-Libaneses, também contribuíram de modo um pouco menos expressivo (Venâncio, 2000).

Portanto, a primeira colonização neste país envolveu majoritariamente portugueses do sexo masculino. A migração de mulheres europeias, neste primeiro momento, não foi significativa (Ribeiro, 1995), o que desencadeou relacionamentos assimétricos entre homens europeus com mulheres indígenas. Em seguida, com a chegada de africanos escravizados durante o ciclo econômico da cana-de-açúcar, começaram a ocorrer relacionamentos entre europeus e africanas (IBGE, 2000).

Assim, o segundo grupo étnico que veio pra América em massa foi o Africano. No continente americano, o Brasil foi o país que mais importou mão de obra africana (Reis, 2000) trazido pela força, para trabalhar como escravo na agricultura e na extração de metais preciosos. Estima-se que até 1880 tenham sido trazidos para o continente americano pelo menos 10 milhões de africanos. A contribuição africana veio de povos que viviam na África Subsaariana, principalmente da região do Congo e Angola e, em menor número, da Guiné, Gana, Nigéria, Serra Leoa e Moçambique. Esses povos, pertenciam a diferentes grupos linguísticos como bantus, nilo-congo e sudaneses favorecendo ainda mais essa miscigenação complexa das populações na América (Wang *et al.,* 2008).

O Brasil foi a colônia que mais utilizou a mão de obra escrava. Dados estatísticos mostram que cerca de 40% de todos os africanos trazidos para a América vieram para o Brasil (Klein, 1986). O Brasil recebeu africanos de todas as origens e regiões da África subsaariana (IBGE, 2000). Nesse mesmo período, a América espanhola recebeu menos da metade desse contingente (Klein, 2000).

A imigração em massa de europeus e africanos promoveu alterações na composição genética das populações locais brasileiras em decorrência do amplo processo de miscigenação. A medida que os europeus avançaram no processo de conquista, trazendo consigo africanos escravizados, os povos originários americanos, que se encontravam na costa atlântica, foram sendo forçados a migrar para o interior do continente, isto quando conseguiam escapar da morte provocadas pelas guerras ou pelas doenças trazidas pelos europeus (Alencastro, 2000). Estudos estimam uma redução da população originária americana em aproximadamente 90% do que existia à época da colonização (Bedoya *et al.,* 2006; Castro *et al.,* 2021).

Houve ainda, num período mais recente, migração asiática, em especial de japoneses. Esta migração teve como marco inicial a chegada do navio Kasato Maru, em Santos, no dia 18 de junho de 1908. Do porto de Kobe a embarcação trouxe, numa viagem de 52 dias, os 781 primeiros imigrantes vinculados ao acordo imigratório estabelecido entre Brasil e Japão. A maior parte desses imigrantes era formada por camponeses de regiões pobres do norte e sul do Japão, que vieram trabalhar nas prósperas fazendas de café do estado de São Paulo (Daigo, 2008).

Contudo, de acordo com os dados do IBGE (2000, 2002), o destino dos

imigrantes estrangeiros variou entre os estados e as regiões do Brasil. Sendo as regiões Sul e Sudeste do Brasil os principais receptores de imigrantes italianos e alemães. Os estados do Paraná, São Paulo e Pará dos imigrantes japoneses. São Paulo e Paraná dos imigrantes espanhóis. Da mesma forma, umestudo recente mostra que a origem dos Africanos que vieram para o Brasil também é distinta entre as regiões: sendo em Salvador, no Nordeste, o componente de países do Centro-Oeste da África predominante, enquanto em Pelotas, no Sul, observa-se o predomínio da contribuição do componente do Sul-Leste Africano (Gouveia *et al.,* 2020)

Esses dados indicam que o processo de miscigenação ocorre de modo heterogêneo entre as regiões do Brasil. E portanto, cada estado brasileiro pode apresentar perfis de ancestralidade genética distintos, reforçando a necessidade de estudos mais detalhados sobre a composição genética de cada um (Souza *et al.,* 2019).

1.4. ANCESTRALIDADE GENÉTICA NO BRASIL E EM SUAS REGIÕES GEOGRÁFICAS

A heterogeneidade da população brasileira já foi documentada em diversos estudos utilizando tanto marcadores uniparentais quanto autossômicos, onde foi demonstrado um padrão típico multi-étnico, porém não uniforme entre as regiões (Pena *et al.*, 2020; Alves-Silva *et al.,* 2000; Carvalho-Silva *et al.,* 2001; Callegari-Jacques *et al.*, 2003; Godinho *et al.*, 2008; Lins *et al.,* 2010; Manta *et al., 2013a*; De Moura *et al.,* 2015). Souza et al (2019), realizou uma revisão sistemática sobre o perfil da ancestralidade no Brasil, abarcando 81 populações de 19 estados e cinco regiões brasileiras. O objetivo foi apresentar uma visão geral das estimativas de ancestralidade genética para diferentes regiões geográficas brasileiras e analisar os fatores envolvidos nestas estimativas. O estudo mostra que a proporção média de ancestralidade das parentais europeia, africana e nativo americana na população brasileira são respectivamente de 68,1%, 19,6% e 11,6%. Quando analisa a ancestralidade genética por região, infere a mais alta proporção de ancestralidade dos povos originários americanos na região Norte (27,3% ± 7,41%), a africana na região Nordeste (28,8% ± 11,31%) e a europeia na região Sul (80,0% ± 7,60%). Além disso, o estudo também mostra que há variação nas proporções de ancestralidade genética inferidas para cada estado brasileiro (Figura 05).

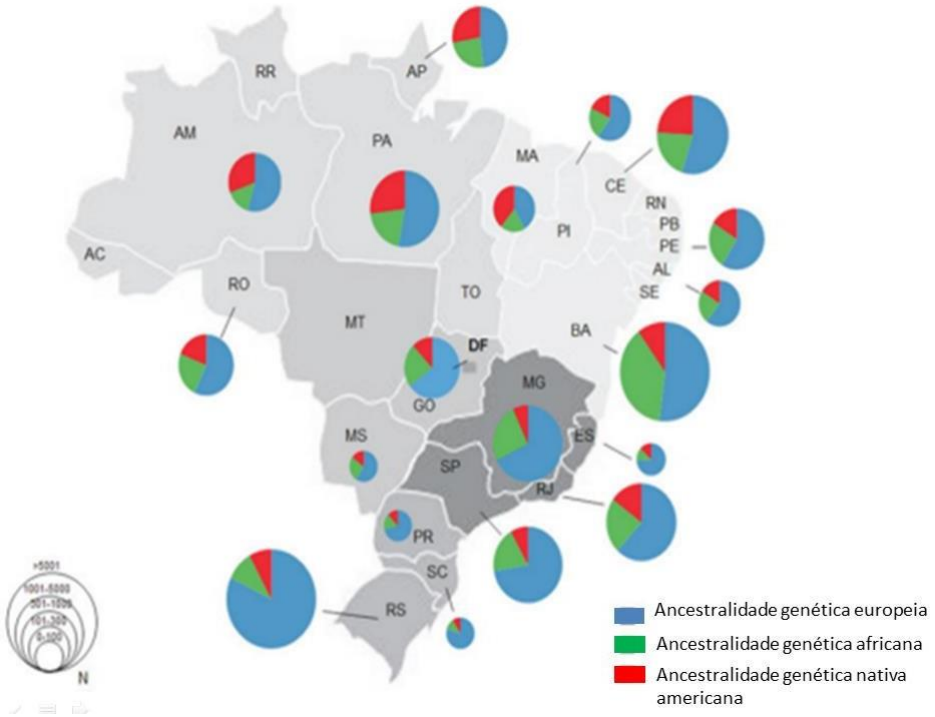


Figura 05. Ancestralidade genética inferida para os estados do Brasil. Cada gráfico circular mostra as proporções médias de ancestralidade inferida para amostras dos estados do Brasil. Os diferentes tamanhos dos gráficos refletem os tamanhos amostrais analisados em cada estado. As cores azul, verde e vermelho, representam as ancestralidades europeia, africana e nativo americana respectivamente.

Fonte: Souza et al (2019) (modificada).

A região Norte, além apresentar um grande número de nativos americanos, contou ao longo de sua história com políticas sociais incentivadoras do casamento entre homens brancos e mulheres nativa americanas, o que corrobora para esse valor expressivo (Benchimol, 1999; Da Cunha, M.C, 2013; Batista dos Santos *et al.*, 1999; Manta *et al.* 2013a). A região Nordeste foi o berço da colonização e, como consequência, recebeu o maior número de indivíduos africanos. Salvador é considerada a cidade mais africana do Brasil, com estimativa de cerca de 80% de afrodescendentes, enquanto que a população brasileira como um todo tem uma estimativa de 56% (Azevedo *et al.*, 1981; IBGE, 2010). A partir do século XVII o componente africano passou a incorporar outras regiões do país, em especial o sudeste devido ao grande fluxo de imigrantes africanos vindos do nordeste devido ao declínio da economia açucareira.

A região Sul recebeu indivíduos africanos em pequenas quantidades (Levy, 1974) e já era escassamente habitada por nativos americanos (Leite, 1996; Ramos, 2006; Ribeiro, 1995). Dessa forma, a única parental com expressão significativa a entrar na

região foi a europeia, sendo o primeiro grupo de colonizadores desta região os portugueses e, mais tarde, de representante de outras partes da Europa por incentivo do governo em troca de trabalho (Salzano & Freire-Maia, 1967; Ribeiro, 1995; Alves-Silva *et al.*, 2000; Marrero *et al.*, 2005; Pena *et al.,* 2011).

A região Sudeste revela um componente de ancestralidade europeia também expressivo (69.1% ± 10.46%) (Souza *et al*., 2019), além de ser geneticamente mais diversificada visto que, por décadas, migrações internas no Brasil tiveram esta região como destino, inserindo na população componentes de ancestralidade dos mais diversos (Giolo *et al.*, 2012). É a região mais densamente povoada do Brasil. Em contrapartida, o Centro-Oeste é a região mais interiorana e uma das últimas a ser densamente povoada. Segundo Cunha 2000/2006 a ocupação dessa região ocorreu basicamente pela expansão agrícola e, mais recentemente, pela construção da cidade de Brasília, sendo que a região recebeu imigrantes de todas as partes do Brasil.

Podemos dizer que no Brasil o padrão migratório num primeiro momento levou a um processo de miscigenação com indivíduos de origem intercontinental (principalmente povos originários Americanos, Europeu e Africanos). E, num segundo momento, ocorreram migrações internas entre as regiões geográficas do Brasil e o processo de miscigenação ocorreu a partir de indivíduos previamente miscigenados. O ápice das migrações internas no Brasil ocorreu entre os anos de 1960-80, com o deslocamento de pessoas que viviam no campo para as grandes cidades (Figura 06). Esse movimento foi especialmente intenso na região Nordeste tendo como principal destino cidades do Sudeste. Nas últimas décadas esse fluxo migratório diminuiu muito, mas o Sudeste ainda é o destino das principais migrações internas no Brasil (IBGE 2000).

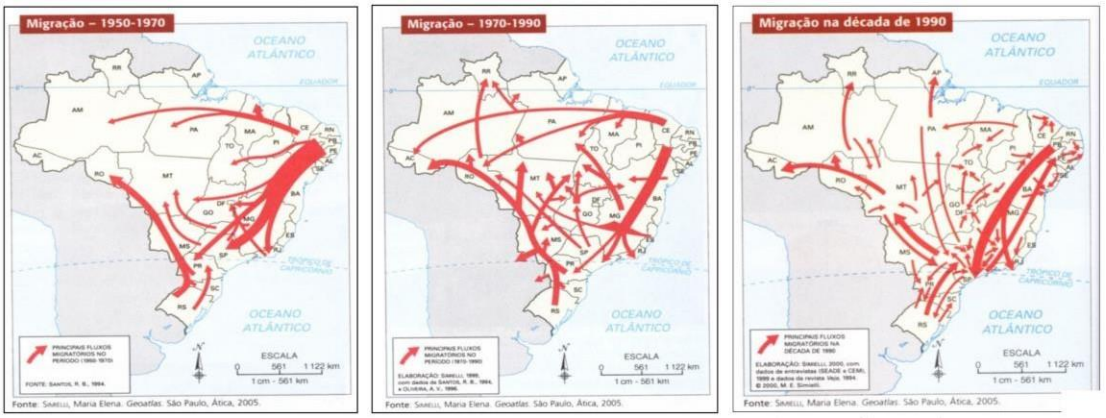


Figura 06. Fluxos migratórios internos no Brasil. As três imagens ilustram as migrações no Brasil, a esquerda o mapa referente ao período de 1950-1970, ao centro de 1970-1990 e a direita na década de 1990 em diante. As setas iniciam no local de origem dos imigrantes e apontam para o destino. A largura das setas indica a proporção de imigrantes.

Fonte: Ilustração GeoAtlas Maria Elena Simielli, editora Ática - 2005.

A construção de Brasília e do Distrito Federal (DF) fazem parte desse segundo processo caracterizado pelas migrações internas no Brasil. A construção de uma cidade planejada no centro do Brasil, erguida em menos de 4 anos (1957-1960), numa região desabitada visando implementar o desenvolvimento socioeconômico da região, contou com a mão de obra de imigrantes vindos de diferentes regiões do Brasil e com histórico prévio de miscigenação. Os "*candangos*" - primeiros trabalhadores de Brasília, se estabeleceram na região do Distrito Federal e uma nova população formada pelo encontro de imigrantes das diversas regiões do Brasil foi estabelecida. Por isso muitos acreditam que o Distrito Federal seja uma síntese da ancestralidade genética observada no Brasil.

No capítulo 2 da presente tese iremos investigar e discorrer sobre a composição genética da população do Distrito Federal, caracterizando esta população, a partir da genotipagem com array de alta densidade de SNPs, e incorporando de forma inédita a investigação sobre o componente ancestral do Leste Asiático nas análises. Ainda, iremos concatenar informações demográficas e sua correspondência com os dados genéticos, para melhor compreensão dos padrões de ancestralidade genética na região do DF.

2. OBJETIVOS

2.1. OBJETIVO GERAL

O presente estudo teve como objetivo contribuir para o entendimento sobre o perfil de ancestralidade genética em populações miscigenadas e discutir os desafios metodológicos para este tipo de inferência, especialmente para a população brasileira, e aplicar esse conhecimento para compreender como ocorreu o processo de formação do Distrito Federal a partir de informações genéticas e demográficas.

2.2. OBJETIVOS ESPECÍFICOS

O primeiro capítulo da tese teve como objetivo avaliar e discutir como a escolha de diferentes conjuntos de marcadores genéticos e modelos de miscigenação influenciam na inferência de ancestralidade genética. Para tanto buscamos:

- Avaliar e comparar a inferência de ancestralidade genética com conjuntos de marcadores genéticos (AIMs, HDSNP e WGS) e com modelos de miscigenação tri- e tetra-híbrido em amostras parentais (Africanas, Europeia, Leste Asiáticas e Nativo Americanas) que contribuíram para a formação da população brasileira.
- Avaliar e comparar a inferência de ancestralidade genética com conjuntos de marcadores genéticos (AIMs, HDSNP e WGS) e com modelos de miscigenação tri- e tetra-híbrido em amostras miscigenadas da América (brasileira, afro-americana, afro-caribenha, colombiana, mexicana, peruana, porto-riquenha).

O segundo capítulo da tese teve como objetivo caracterizar o perfil de ancestralidade e os movimentos migratórios no DF. Para tanto buscamos:

- Analisar a inferência de ancestralidade a partir de marcadores genéticos situados nos cromossomos autossômicos, sexuais e mtDNA.
- Comparar as inferências de ancestralidade obtidas pelo presente estudo com dados da literatura para o Brasil e as cinco regiões geográficas.
- Determinar, a partir das informações demográficas sobre local de nascimento dos genitores dos participantes, a contribuição de cada estado e região geográfica brasileira para a formação da amostra do DF.
- Testar se o perfil de ancestralidade observado é condizente com o perfil migratório dos genitores dos participantes.

CAPÍTULO 1
COMPARAÇÃO DE PAINÉIS DE ANCESTRALIDADE GENÉTICA E MODELOS DE MISCIGENAÇÃO GENÉTICA

Challenges in selecting admixture models and marker sets to infer genetic ancestry in a Brazilian admixed population

Capítulo publicado no formato de artigo no periódico científico internacional Scientific Reports (2022) 12:21240. DOI:10.1038/s41598-022-255217.

1. INTRODUÇÃO AO CAPÍTULO 1

Neste capítulo avaliamos como a escolha de diferentes conjuntos de marcadores genéticos (desde painéis clássicos de AIMs à dados de sequenciamento de genomas completos) e dos modelos de miscigenação (tri-híbrido e tetra-híbrido) influenciam a inferência de ancestralidade genética.

Para tanto, escolhemos seis painéis de AIMs, os quais foram elaborados para distinguir as ancestralidade africanas, europeia, nativo americana e leste asiática, um array de alta densidade de SNPs, o qual foi especialmente elaborado para estudos populacionais, e dados de sequenciamento de genoma completo.

Num primeiro momento, avaliamos a acurácia de cada conjunto de marcadores genéticos em atribuir de forma correta a ancestralidade em amostras de origem biogeográfica conhecida e sem evidências recentes de miscigenação. Para esse teste usamos os dados públicos de 1.511 amostras não miscigenadas do Projeto 1000 Genomas (1KGP) e 543 amostras do Projeto de Diversidade Genômica Humana (HGDP). Num segundo momento, testamos as inferências de ancestralidade pelos diferentes conjuntos de marcadores genéticos em amostras miscigenadas comparando os modelos de miscigenação tri-híbrido (K=3; AFR, EUR e NAM) e tetra-híbrido (K=4, AFR, EUR, NAM e EAS). Para tanto, usamos 1.171 amostras brasileiras e 504 amostras de seis populações miscigenadas do 1KGP (afro-americanos, afro-caribenho, colombianos, mexicanos, peruanos, porto riquenhos).

Este capítulo consiste no manuscrito original do trabalho publicado com o título "*Challenges in selecting admixture models and marker sets to infer genetic ancestry in a Brazilian admixed population*", no periódico científico Scientific Reports - doi: 10.1038/s41598-022-25521-7. PMID: 36481695; PMCID: PMC9731996 (ver ANEXO DIGITAL CAPÍTULO 1).

2. RESUMO

A inferência da ancestralidade genética desempenha um papel cada vez mais proeminente em estudos de genética clínica, populacional e forense. Várias estratégias de genotipagem e metodologias analíticas foram desenvolvidas nas últimas décadas para atribuir indivíduos a regiões biogeográficas específicas. No entanto, apesar desses esforços, a inferência de ancestralidade em populações com história recente de miscigenação, como as do Brasil, continua sendo um desafio. Em populações miscigenadas, a proporção e os componentes da ancestralidade genética variam em diferentes níveis: (i) entre populações; (ii) entre indivíduos da mesma população, e (iii) entre os cromossomos do indivíduo. O presente estudo avaliou 1.171 amostras brasileiras miscigenadas para comparar a ancestralidade genética inferida por modelos de miscigenação tri- / tetra-híbrido e avaliou diferentes conjuntos de marcadores desde aqueles com um pequeno número de marcadores - painéis de marcadores informativos de ancestralidade (AIMs), a arrays de alta densidade de SNPs (HDSNP) e dados de sequência de genoma completo (WGS). As análises revelaram maior variação no coeficiente de correlação dos componentes ancestrais dentro e entre populações miscigenadas, especialmente para componentes ancestrais minoritários. Também observamos correlação positiva entre o maior número de marcadores nos painéis de AIMs e os dados de HDSNP/WGS. Além disso, quanto maior o número de marcadores, mais preciso é o modelo de miscigenação tetra-híbrida.

Challenges in selecting admixture models and marker sets to infer genetic ancestry in a Brazilian admixed population

Luciana Maia Escher[1], Michel S. Naslavsky[2,3], Marília O. Scliar[3], Yeda A. O. Duarte[4,5], Mayana Zatz[2,3], Kelly Nunes[2*+], Silviene F. Oliveira[1*+]

[1] Human Genetics Laboratory, Institute of Biological Sciences, University of Brasilia, DF, Brazil.

[2] Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo, São Paulo, SP, Brazil.

[3] Human Genome and Stem Cell Research Center, University of São Paulo, São Paulo, SP, Brazil.

[5] Medical-Surgical Nursing Department, School of Nursing, University of São Paulo, São Paulo, SP, Brazil.

[6] Epidemiology Department, Public Health School, University of São Paulo, São Paulo, SP, Brazil.

* knunesbio@gmail.com; silviene.oliveira@gmail.com

[+] Authors contributed equally to this work.

3. ABSTRACT

The inference of genetic ancestry plays an increasingly prominent role in clinical, population, and forensic genetics studies. Several genotyping strategies and analytical methodologies have been developed over the last few decades to assign individuals to specific biogeographic regions. However, despite these efforts, ancestry inference in populations with a recent history of admixture, such as those in Brazil, remains a challenge. In admixed populations, proportion and components of genetic ancestry vary on different levels: (i) between populations; (ii) between individuals of the same population, and (iii) throughout the individual's genome. The present study evaluated 1,171 admixed Brazilian samples to compare the genetic ancestry inferred by tri-/tetra-hybrid admixture models and evaluated different marker sets from those with small numbers of ancestry informative markers panels (AIMs), to high-density SNPs (HDSNP) and whole-genome-sequence (WGS) data. Analyses revealed greater variation in the correlation coefficient of ancestry components within and between admixed populations, especially for minority ancestral components. We also observed positive correlation between the number of markers in the AIMs panel and HDSNP/WGS. Furthermore, the greater the number of markers, the more accurate the tetrahybrid admixture model.

4. INTRODUCTION

Understanding how human genetic diversity is distributed and its implications has been a recurrent focus in clinical, population, and forensic genetics studies[1-3]. Since the 1970s, owing to the pioneering work by Richard Lewontin, it has been understood that most human genetic variation occurs between individuals of the same population group, while genetic variation between individuals of distant populations is restricted to a small proportion of the human genome. This study is one of the first to refute the use of social races in biological studies and draw attention to the fact that genetic information is more accurate for biological issues than social groups or ethno-racial self-declaration[4].

Since then, several studies have subsequently confirmed these observations and revealed that the distribution of genetic diversity and population differentiation is a continuous gradient within and between populations across continents[5-6]. Therefore, the categorization of human groups by current geo-political regions are arbitrary choices and not true biological clusters. Thus, these studies clarify that there is only a small set of genetic polymorphisms with a distinct allelic frequency between human populations or continents.

As such, these small genetic differences have been widely studied to address specific biological issues. For example, in clinical genetics, some diseases are recognized as having different incidences among population groups, for example: chronic kidney disease[7], hypertension[8], and inflammatory bowel disease[9]. Identifying associated genetic variants and the correct assignment of individuals in these groups helps in the development of personalized medical care[10]. In forensic genetics, when identifying the individual via CODIS (the Combined DNA Index System), it is often necessary to provide additional information such as phenotypic characteristics (eye, hair, and skin color) and/or the most probable continental origin[11].

Over the last few decades, several methods and strategies have been developed in an attempt to assign an individual´s geographical origin based on DNA variations. This is what geneticists often refer to as genetic ancestry[12]. One of the first strategic approaches applied was to identify sets with a few dozen ancestry informative markers (AIMs - genetic markers which exhibit substantially different frequencies between different populations) in order to compose informative ancestry panels with the purpose

of clustering individuals into continental and subcontinental population groups[13-15]. The development of AIM panels attempts to select a group of genetic markers that compose a small, accurate, low-cost, and highly informative set. In general, AIM panels vary in some characteristics, including: specific loci, number of loci, genotyping strategies, and parental reference populations[16-18]. Studies usually use only one AIM panel or even a complementary set, for example: PIMA+34-plex[19], Pacifiplex and 34-plex[20], and KiddLab+Seldin+34-plex[21].

On the other hand, the emergence of high-throughput genotyping technologies, such as high-density SNPs array (HDSNPs), whole-exome-sequence (WES) and whole-genome-sequence (WGS) enabled high horizontal genome coverage studies for genomic ancestry inference[22,23]. In this scenario, there is not only a significant increase in the number of genetic markers evaluated, the number of genotyped individuals per study grows from hundreds to thousands[24-26]. Consequently, some analytical approaches have been adapted, in addition to the development of new analytical strategies[27-29].

Additionally, in both the AIM and high-throughput genotyping strategies, proper genetic ancestry inference is dependent on the existence of a reference population panel for each ancestral component under study. Genetic ancestry inference remains deficient in some population subgroups due to lack of reference data collection, in addition to some inconclusive or non-validated studies. For example, there is an effort to develop reference panels for Asians and Native Americans, in addition to increasing reference genome data for Latin Americans[30].Admixed populations, such as those in America, pose a peculiar case for genetic ancestry inference as, they originated over the last 500 years through a complex admixture process with population sources of individuals from different continents: Native Americans, Europeans, Africans[31-32] and more recently East Asians[33]. The genomes of admixed individuals are a redistribution of genetic variation observed in parental populations, which produces new genomic combinations of pre-existing genetic variants. This leads to a paradigm shift, in which the geographic origin of the admixed individual becomes a secondary issue, the primary objective being to identify each ancestral component, its distribution and proportion in the individual's admixed genome.

Furthermore, in admixed populations, the proportion and components of genetic

ancestry vary at different levels: (i) between populations; (ii) between individuals of the same population, and (iii) throughout the individual's genome[34]. This has a direct impact on the reproducibility and transposition of study results in these populations. Thus, nowadays, genetic ancestry inference in an admixed population is an essential tool to control the effect of population stratification in association studies[35], for the identification of disease-associated genes[36], precision medical care[11] and to reveal population history[37,38].

Therefore, the correct assignment of each genetic ancestral component is essential for studies with admixed populations. Concerning this issue, some studies compared AIM panels in American admixed populations, observing differences in ancestry proportion inference between panels, which may be related to both the number of markers and parental reference populations of each panel[19, 39, 40].

Of the Latin American countries, Brazil was the only Portuguese colony, which resulted in peculiarities in the admixture process. Brazil received more than 4.5 million African slaves whose origin in the African continent may differ in terms of place and time from those of non-Portuguese colonies in Latin America[37]. Approximately 3 million Native Americans (indigenous people) lived in the current Brazilian territory in the 15th century who, after contact with Europeans, declined in number by at least 90%[41]. Recently, in the 20th century, Brazil continues to receive millions of immigrants, especially from East Asia[33].

Today Brazil is the most populous country in Latin America and the seventh most populated country in the world, with more than 215 million inhabitants. Understanding how different marker sets are assigned to Brazilian genetic ancestry is of extreme relevance to both historical and public health issues. The first comparative studies to address this issue were only performed with AIMs or pharmacogenomic high-density SNP array panels and are based on the trihybrid admixture model (Native American, European and African)[19,39,40]. Herein, we present for the first time a comparison of ancestry estimates between different genetic marks sets - AIMs, high-density array of SNPs and Whole Genome Sequencing - in the Brazilian population using the tri-hybrid (African, European and Native American) and tetra-hybrid admixture models (African, European, Native American and East Asian).

This study analyzed 1,171 admixed samples from Brazil[23], which we compared to (i) 5 AIMs panels: 34AISNP[42]+PIMA[19]; 55AISNP[18]; 128AISNP[43]; 170AISNP[44], and 446AISNP[45], selected by both the tri- and tetra-hybrid admixed models. In addition, we evaluated the combination of the 5 aforementioned panels, referred to herein as 665AISNP; (ii) a high-density SNPs array (HDSNP), specially developed for population genetics studies and therefore without biased markers such as the GWAS SNP array or pharmacogenomics, and (iii) whole genome sequencing (WGS) data.

5. RESULTS

5.1. Ancestry Inference in Parental Population Groups

The proposal of panel sets is to correctly assign individuals to their original continental groups, especially AISNPs. However, factors such as the number of markers and the set of populations used to establish allelic frequencies of the continental group may influence ancestry estimation accuracy. To better assess these issues, we analyzed the influence of marker number by comparing the inference of ancestry in our panel sets for the 4 main parental continental groups (African, European, East Asian, and Native American) which contributed most to the Brazilian population. For the parental populations, we used samples of the Human Genomic Diversity Panel (HGDP)[46] and 1000 Genomes Project phase III (1KGP)[24] (Supplementary Table S1).

First, we evaluated the accuracy of the panel sets in correctly assigning pre-categorized individuals[24,46] into each of the 4 continental groups. Only a small proportion of samples were not assigned correctly (Z-score >|3|; p-value 0.01; ranging from 1.66% to 0.7% and 0.86% to 0.4% in the HGDP and 1KGP, respectively) (Supplementary Tables S2-S4).

Secondly, the distribution of the inferred proportion of individual ancestry for each set of panels and continental group were verified (Figure 1). For African continent samples, all panel sets had high accuracy in inferring African ancestry in both datasets (median >96%, lowest observed dispersion ancestral components inferred). On the other hand, for other continental samples, some panel sets displayed greater dispersion in ancestry inferences, showing median and average values <90%. The 128AISNP and

446AISNP panels had the lowest medians (82-88% and 83-89.5%, respectively) for samples from the continental European, East Asian and Native American groups. Meanwhile, the HDSNP and WGS panels had the lowest dispersions in all continental groups with median values >98%.

Finally, a pairwise comparison of the inferred proportion of individual ancestry was performed between each panel set for each continental group (Supplementary Figure S1A-G). The 8 panel sets evaluated showed high correlation coefficient values (r^2>0.96), ranging from 0.98 to 1 for the African and European samples (Supplementary Figure S1A, B, E and F) and from 0.96 to 1 for East Asians and Native Americans (Supplementary Figure S1C, D and G).

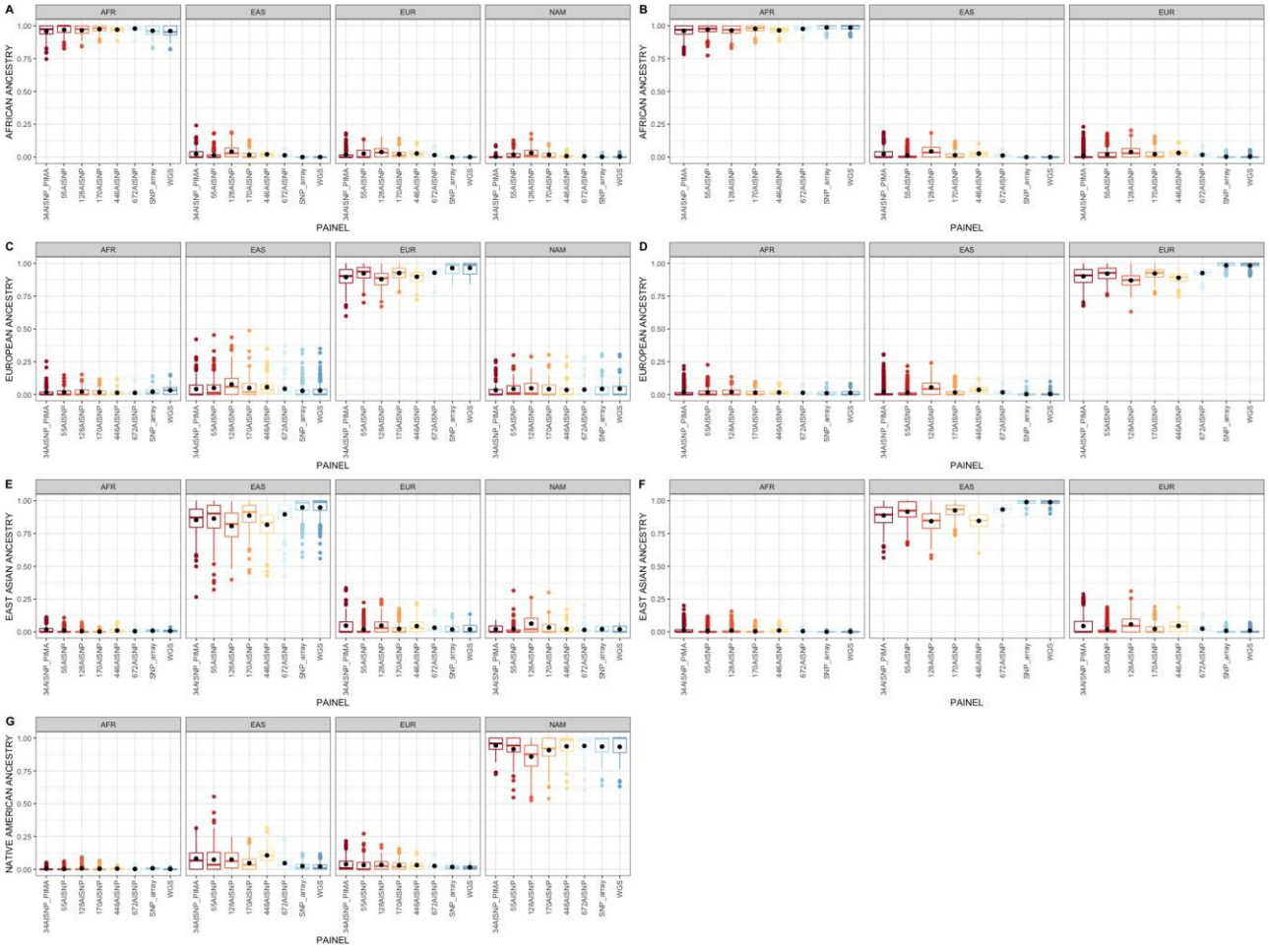


Figure 1. Boxplot with the distribution of ancestry inferences of individuals within each continental group. The boundary of the box closest to zero indicates the 25th percentile, the line within the box represents the median, and the boundary of the box farthest from zero indicates the 75th percentile. Black points within the box mark the average. Whiskers above and below the box indicate the 10th and 90th percentiles. Points above and below the whiskers indicate outliers outside the 10th and 90th percentiles. A, C, E and G refer to the HGDP samples. B, D and F relate to the 1KGP samples.

5.2. Ancestry inference in Brazilian populations

A common question when studying genetic ancestry of the Brazilian population is whether to use a tri- or tetra-hybrid admixture model, and what panel set. In order to explore this issue, we analyzed both the admixed models and the inference of ancestry from different panel sets - AIMs, HDSNP and WGS. For the parental reference populations, we selected Africans (AFR), Europeans (EUR), East Asians (EAS), and Native Americans (NAM) from the HGDP, only including samples with z-score values <|3| (Supplementary Tables: S2-S4).

In general, we observed variations in ancestry inferences according to the admixed model chosen as well as the panel set (Figure 2). By the tri-hybrid model, the average ancestry inferences in the Brazilian sample ranged from 70.02 to 74.16% for EUR ancestral component; 16.91 to 19.58% for AFR, and 8.96 to 10.59% for NAM (Supplementary Table S5). In the tetrahybrid model, the average ancestry inferences were: 66.33 to 73.02% (EUR ancestral component); 16.77 to 18.76% (AFR); 6.46 to 7.26% (NAM), and 2.90 to 8.72% (EAS) (Figure 2; Supplementary Table S6).

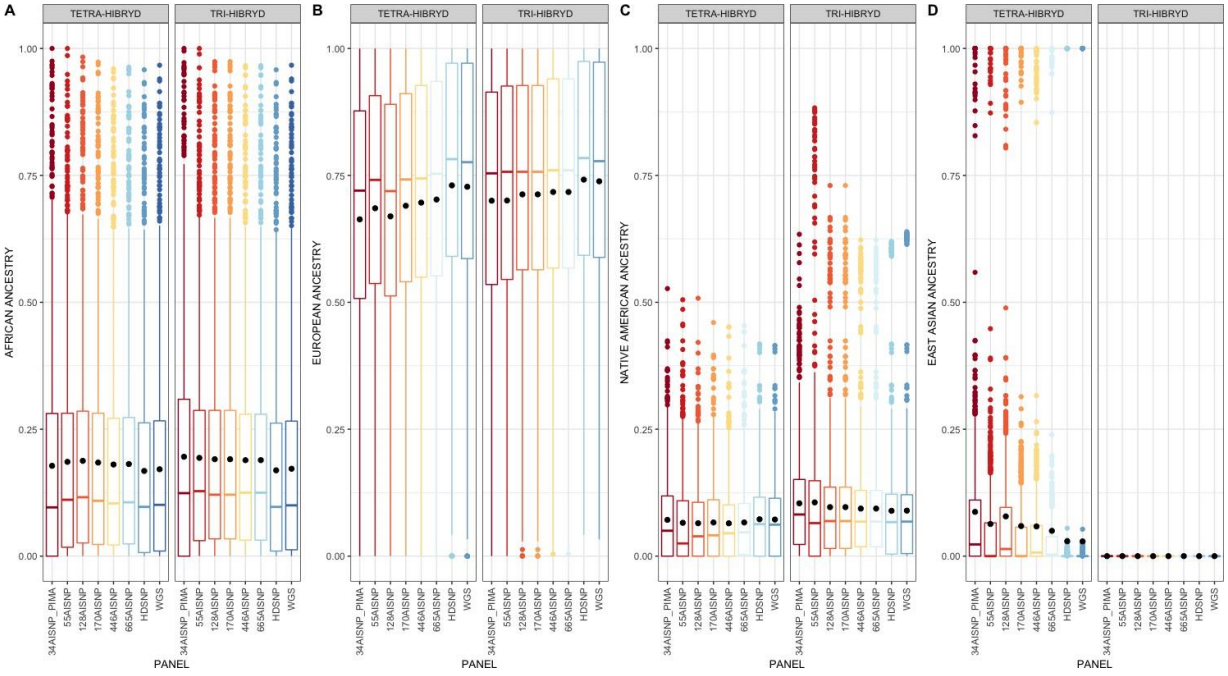


Figure 2. Boxplot with the distribution of ancestry inferences of individuals in the Brazilian population (SABE) for the tri- and tetra-hybrid models. The boundary of the box closest to zero indicates the 25th percentile, the line within the box represents the median, and the boundary of the box farthest from zero indicates the 75th percentile. Black points within the box mark the average. Whiskers above and below the box indicate the 10th and 90th percentiles. Points above and below the whiskers indicate outliers outside the 10th and 90th percentiles.

To determine whether there are significant differences in ancestry inferences according to the tri- or tetra-hybrid models, we adopted two analytical approaches. First, as many studies are interested in the population average of each ancestral component, we performed the paired t-test to compare the average obtained with the same panel from the tri- and tetra-hybrid models (Supplementary Table S7). The averages for the AFR ancestry component inferred did not differ significantly between the models. On the other hand, all panel sets showed significantly different averages for the NAM component (t-test >4; p-value <0.0025), and the 128AISNP for the EUR component (t-test=4.17; p-value=3.13^{-5}). We subsequently performed a pairwise comparison to verify the correlation between the ancestry inferences obtained by the same panel from the tri- and tetra-hybrid models. We observed that the inferred AFR and EUR ancestral component correlation coefficient ranged from 0.97 to 0.99, and the NAM component between 0.46 and 0.67 (Figure 3).

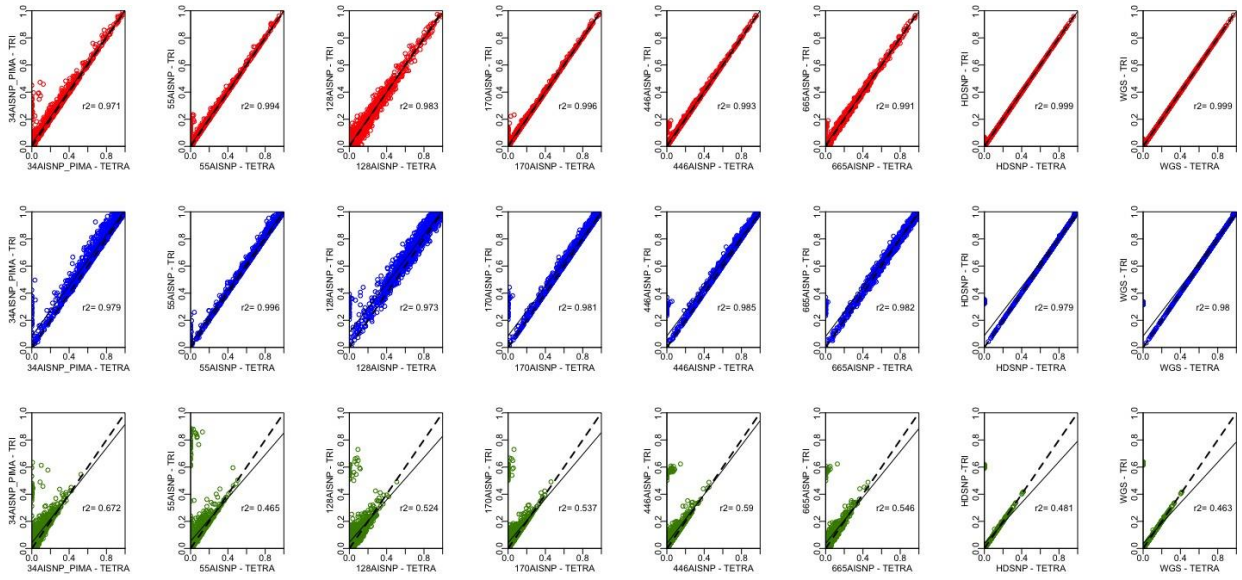


Figure 3. Pairwise comparison of ancestry inferences by tri and tetra-hybrid models for Brazilian samples (SABE) with the 8 panel sets evaluated: 34 AISNP+PIMA; 55 AISNP; 128 AISNP; 170 AISNP; 446 AISNP; 672 AISNP; HDSNP, and WGS. The x-axis corresponds to the ancestry inference by the tetra-hydride model for a given panel. The y-axis corresponds to the inference of genetic ancestry by the tri-hybrid model for a given panel. In the figure, r2 corresponds to the correlation coefficient, the black dashed line represents the trend, and the solid black line corresponds to the perfect agreement between two panels. The red, blue and green colors correspond to the inference for the African, European and Native American ancestral components, respectively.

To better understand how the EAS ancestral component is being detected by the admixture models, we evaluated the assignments of 33 samples in the Brazilian dataset, all of which were self-declared Asian descendants. In the tetra-hybrid model, all panels detected more than 85% EAS ancestral component in the samples analysed. We then analyzed this subset by comparing the inferences of AFR, EUR, and NAM components between the tri- and tetra-hybrid models. For the tetrahybrid model, the inferences of these three ancestral components were close to 0, while between 20 and 66% in the tri-hydride model. In Figure 3, we see that in almost all comparisons, these samples are clustered and offset from the correspondence line. In order to assess how the EAS ancestral component was assigned in the other samples of the dataset, we excluded samples with >85% Asian ancestry. The inferred EAS average for this subset was: 6.17% (s.d.=8.5%) 34AISNP+PIMA; 3.69% (s.d.=6.25%) 55AISNP; 5.31% (s.d.=8.5%) 128AISNP; 3.24% (s.d.=5.17%) 170AISNP; 3.25% (s.d.=4.73) 446AISNP; 2.27% (s.d.=3.55%) 665AISNP; 0.11% (s.d.=0.36%) HDSNP, and 0.09% (s.d.=0.33%) WGS. For HDSNP and WGS, none of the sample had EAS component inference above 5%, while the AIMs panels had samples with maximum observed values ranging from 23.9 to 55.9%.

Finally, we evaluated whether the ancestry inference from the different sets of panels differs from each other (Figure 2). A pairwise comparison of the averages was performed with no significant differences (t-test) observed for the AFR and NAM component in both models (Supplementary Tables S8 and S9). With the tri-hybrid model, the inferences of the EUR ancestral component had significant differences in the comparisons of the 34AISNP+PIMA and 55AISNP panels to HDSNP and WGS with p-values of <0.005 and 0.007, respectively (Supplementary Table S8). In the tetrahybrid model, for inferences of the EUR component, with the exception of the 446AISNP and 665AISNP panels, the others showed significant differences to HDSNP and WGS panels (p-value <0.01). Additionally, for the EAS component, only 665AISNP had no significant difference with HDSNP and WGS (Table S9). We also performed correlation analysis between panels for ancestry inference for the inference of the AFR ancestral component (r^2_{tri}=0.89 to 1; r^2_{tetra}=0.90 to 1); EUR (r^2_{tri}=0.91 to 1; r^2_{tetra}=0.91 to 1); EAS (r^2_{tetra}=0.80 to 1), and NAM component (r^2_{tri}=0.76 to 1; r^2_{tetra}=0.54 to 0.99) (Supplementary Figures: S3 and S4).

5.3. Ancestry inference in admixed American populations

To find out if this pattern observed in the Brazilian sample is similar in other admixed populations in America continent, we performed the same analyzes in the admixed populations of 1KGP: Afro-Caribbean (ACB); Afro-American (ASW); Colombian (CLM); Mexican (MXL); Peruvian (PEL), and Puerto Rican (PUR). The admixed populations evaluated herein have different proportions of parental ancestry, ranging from those with proportions of mostly African ancestry (ACB and AWS), mostly European ancestry (CLM, PUR), and mostly Native American ancestry (MXL and PEL) (Supplementary Tables: S5 and S6).

When comparing ancestry averages inferred by the same panel for each ancestry component with tri- or tetra-hybrid admixture models, nonsignificant differences were observed (except for the 128AISNP panel in the NAM component in the PUR sample; t-test=3.7, p-value=0.029) (Supplementary Table S7). On the other hand, the pattern of correlation coefficients is heterogeneous between the ancestry components and the 1KGP admixed populations (Supplementary Figure S4-F). In the comparisons of ancestry inference averages by the different panels in the same admixture model, we also found heterogeneous results. In the tri-hybrid model, ACB showed differences between the averages inferred for the EUR and NAM componentes, and MXL for the African component (Supplementary Table S8). The pairwise comparison of individual ancestry inferences between the panels shows variation in the correlation coefficients, both between ancestry components in the admixed population, and between the admixed populations. In ACB, the correlation coefficient for African ancestry component ranged from r^2_{TRI}=0.66 to 1 (Supplementary Figure S11A) and in ASW from r^2_{TRI}=0.88 to 1 (Supplementary Figure S12A). In CLM, the correlation coefficient for the EUR ancestry component ranged from r^2_{TRI}=0.78 to 1 (Supplementary Figure S13B), and from from r^2_{TRI}=0.74 to 1 in PUR (Supplementary Figure S16B). Regarding MXL and PER, the correlation coefficient for NAM ancestry ranged from r^2_{TRI}=0.87 to 1 (Supplementary Figures S13C and S15C). A general trend in these comparisons is higher correlation coefficients between panels that share a greater number of markers (e.g. 128AISNP x 170AISNP; 446AISNP x 665AISNP and HDSNP x WGS), in addition to those with the highest number of markers (e.g. 446AINSP, 665AISNP, HDSNP and WGS) (Supplementary Figures: S11 to S16).

6. DISCUSSION

In the present study, we evaluated 8 panel sets: six AISNPs, one HDSNP and one WGS. Using tri- and tetra-hybrid admixture models, we compared ancestry inferences in Brazilian admixed populations and a set of admixed American populations.

To verify the accuracy of the panels, samples from HGDP and 1KGP datasets were used, whose geographic origin is known and without evidence of recent admixture. Despite the low marker overlap observed in the AISNPs panels (see Supplementary Material Notes), all panels showed a high accuracy rate (error rate 0.4-1.66%; Supplementary Table S2) and high degree of correlation in the pairwise comparisons of the panels (r^2>0.96; Supplementary Figures S1A-G). However, it is also possible to observe heterogeneity in the distribution of genetic ancestry inferred within each parental group by the different panel sets (Figure1). The smallest dispersion was observed in AFR (median values >90%), while EUR and EAS presented the largest one (median values <90% in panels such as 128AISNP and 446AISNP).

These results reveal, albeit with varying degrees of accuracy between them, that the available AISNP panels meet the proposed role of correctly attributing ancestry according to the continental group to which the individual belongs. Several studies already compared the accuracy of panels and obtained similar results[18,21,47]. As such, many authors currently argue that there is no necessity for new AIMs panels to assign the 6 biogeographic regions: Sub-Saharan Africa, Europe, Southwest Asia, South Asia, East Asia and the Americas. Instead, efforts should be directed towards building panels for global use, with greater representation of population groups[18].

Most AIMs panels use HGDP and 1KGP data as a reference population for marker selection, including some of those evaluated in the present study[19,42]. These two public databases were essential for understanding the distribution of genetic diversity and affinity among human population groups[24,46,48]. However, they only capture a portion of human population diversity. Therefore, many AIMs panels endeavoured to include more populations from different population groups during their development process, for example: 55 AISNP[18]; 128 AISNP[43]; 446 AISNP[45].

Soundararajan et al (2016)[49] argued that if there is a low representation of data from reference populations, a greater number of markers becomes necessary for the

robustness of allele frequencies for the definition of population groups of interest. Our results converge at this point as we observed greater correspondence in individual ancestry inferences between panels with a greater number of markers, in particular to those of the HDSNP and WGS data.

In the present study, we focused on Brazilian admixed populations. This population emerged in the last half-century, especially from Native American, European, and African sources. More recently, it has also received contributions from other regions, including East Asia and the Middle East.

Admixed populations require a closer look in terms of ancestry inferences as their genomic particularities give rise to several challenges. Each admixed population has a peculiar evolutionary history, differing in parental sources, proportion and time of admixture. Furthermore, the admixing process produces variation at different levels: in ancestry between admixed populations, between individuals in the same admixed population, and throughout the genome of the same admixed individual[34]. For this reason, a method, model or panel that captures the profile in one admixed population or admixed individual well will hardly have the same performance for another.

We know that the EAS contribution is less than 1% for most of the Latin American admixed populations. Therefore, the choice of tri- or tetra-hybrid model will depend on the admixture profile of the population. Our motivation to analyze the tetra-hybrid model lies in the fact that in recent decades there has been a growing migratory flow of East Asian populations to large urban centers in the USA and Brazil. East Asian immigration to Brazil began in 1908 with the Japanese and today, according to the Ministry of Foreign Affairs of Japan, more than 2 million Japanese descendants live in Brazil. São Paulo, the city where the Brazilian samples of the present study were collected, is home to one of the largest Japanese communities outside Japan. The Brazilian cohort has 33 samples with 100% East Asian ancestral component that are direct descendants of the first Japanese immigrants[23]. Data from the last Brazilian census revealed that, in 10 years, there was a 173.7% increase in the number of individuals who declared themselves to be of Asian descent (Japanese, Chinese and Korean)[33].

Based on this scenario, using WGS data from 1,171 Brazilian individuals, we evaluated how different admixed models and sets of panels behave to infer ancestry in

the Brazilian population. First, we checked for differences in the inferences of each ancestral component according to the tri- or tetra-hybrid admixture model. The population average inferred by either admixed model only differed for the NAM ancestral component (Supplementary Table S7). Similarly, the NAM ancestral component is the one with the lowest degree of correlation between the two admixture models (Supplementary Figure S3). These results suggest that the chosen admixture model can influence the inference of the average NAM ancestral component in this Brazilian sample. In Figures 2 and 3, it is also possible to observe a trend of greater proportions in the inference of the NAM ancestral component in the trihybrid model than in the tetrahybrid model, both in terms of the population average and the individual. In order to better understand this trend, it is necessary to evaluate the assignment of the EAS ancestral component in these samples.

Once we had self-declared individuals of Asian descent in this Brazilian cohort, we verified how the tri- and tetra-hybrid models assigned ancestry. For these individuals, the tri-hybrid model, the HDSNP and WGS panels assigned: ~63% to NAM, ~32% to EUR, and ~5% to the AFR ancestral components. There are more ranges of inferred percentage for the AIMs panels (Figure 3). The panels with the highest number of markers (446AISNP and 665AISNP) were closer to the inferences of high-density panels of SNPs, while those with the lowest number of markers (34AISNP+PIMA, 55AISNP, 128AISNP and 170AISNP) had large ranges, in some cases including the assignment of proportions for the African ancestral component >40%. This result shows a redistribution of the EAS component, mostly to the NAM component, followed by the EUR component, and to an even smaller proportion, the AFR component. As the NAM ancestral component is a minority in the Brazilian cohort (<8%), this may constitute to the increase of the NAM ancestral component in the average population discussed in the previous paragraph.

Given the recent migratory flow from East Asia to Brazil and the fact that the samples from the Brazilian cohort were collected in 2010 and had individuals >60 years old (71.86±7.94) at the time of collection (details in [23]), it was unexpected to visualize individuals with this ancestral component as a minority in their genome. Thus, we evaluated the remaining 1,138 samples as probably not possessing the EAS ancestral component. Our results showed that for the tetrahybrid model, especially for the AIM panels, there was more noise in the EAS component inference (Figure 2), while for the

HDSNP and WGS panels, the inferences had less noise (no individual with >5%). These results suggest that the two high-density marker panels are able to better assign ancestral components in the tetrahybrid model. In turn, in the trihybrid model, samples with some proportion of the EAS ancestral component in the tetrahybrid model, showed an increase in the NAM and EUR ancestral components. This observation can be seen as another factor contributing to the differences in the inferred Native American ancestral component between the models.

We also compared the tri- and tetra-hybrid models in other admixed populations in America (1KGP) for which there are no historical records of large migratory flows from EAS (except for Peru). Therefore, it is unusual to analyze the tetra-hybrid model in this dataset and we only performed it in order to better explore the patterns. Unlike the Brazilian cohort, we did not observe significant differences in the population averages of the components between the admixed models. However, Figures S5 to S10 clearly show noise with the inference of the EAS ancestral component in populations for which it is not part of the parental source.

Therefore, choosing an admixed model is not a simple decision, as each model has advantages and disadvantages in each population. The decision of which model to apply will depend on the question the investigator wants to ask and whether there is interest in the population average or ancestry of each individual in the sample. If it is to decipher specific admixture components, for example to learn about an individual's family migratory patterns, then all possible parental populations involved should be included. If they are simply trying to determine the major ancestral component, for example exclusion purposes, then a smaller model with the key continental groups may suffice.

The second objective of our study was to compare ancestry inferences with different sets of panels (AISNP, HDSNP and WGS). In the Brazilian sample, we observed significant differences for the EUR ancestral component averages in the tri-hybrid model (34AISNP+PIMA and 55AISNP x HDSNP and WGS) (Supplementary Table S8) and for the EUR (34AISNP+PIMA, 55AISNP, 128AISNP and 170 AISNP x HDSNP and WGS) and EAS (all panels, except 665AISNP x HDSNP GWS) in the tetrahybrid model (Supplementary Table S9). These results are possibly related to what was observed for the parental populations, where there is greater dispersion in the

distribution of EUR, EAS and NAM ancestry (Figure 1), indicating a variation in the accuracy of correctly assigning this ancestral component. On the other hand, we did not observe differences in the population averages inferred by the sets of panels for the African and NAM ancestral components. Although, in the pairwise analyses, the smallest correlations between panels occurred in inferences from the NAM ancestral component (Supplementary Figures S2C and S3C). This result suggests that although the population average inference of the NAM ancestral component is similar between the panels, there are differences in the inferences on the individual level.

The analysis involving admixed populations of the 1KGP showed heterogeneous results. In paired comparisons of individual ancestry inference between panels (Supplementary Figures S11 to S16), we observed variation correlation coefficients both between ancestry components within the same admixed population and between admixed populations. The inconsistencies observed in the ancestry inferences between the panels were even more evident for the minority ancestry components of the individuals in our results (e.g. Supplementary Figures S14A, S15A and S16C). This probably occurred because the genome of an admixed individual is a mosaic composed of segments from different parental sources. Over generations, due to the process of meiotic recombination, the components of distinct ancestry are shuffled between homologous chromosomes[36,50]. Thus, the greater the number of generations since admixture onset, the smaller the size of the genomic segments of the ancestry will be. In addition, the greater the proportion of an ancestral component, the greater the size of its segments in the genome, while conversely, the smaller the proportion of the ancestral component, the smaller the segments in the genome[50]. In this scenario, due to lower density and genomic coverage, AISNP tends to be less accurate data than higher SNP density and higher genomic coverage.

The NAM component had the lowest correspondence in ancestry inferences between the panels (Supplementary Figures S3C and S16D). It is widely recognized that the Native American populations, due to their recent bottleneck history, are the most differentiated in the world[48] having the lowest number of representatives in the reference panels. Panels were developed with the aim of enriching the NAM component[18,43,45], however, they do not always capture this component well in all admixed populations. Thus, the underrepesentation of Native American sources, in additio to the minority NAM ancestral component in PUR and Brazilian sample, may be

contributing to the observed differences in ancestral inference between the panel sets for this ancestral component.

Through the present study, we verified that there are differences in the inferences of the ancestral components according to the panel chosen. There is greater correspondence of inference between panels that share a greater number of markers (128AISNP and 170AISNP; 446AISNP and 665AISNP; HDSNP and WGS), and among those with the highest number of markers (446AISNP, 665AISNP, HDSNP and WGS). Again, it is important to point out that the choice of panel will depend on the purpose and needs of the study. For example, in forensic genetics, sometimes samples with quantity and quality are not available, which limits the genotyping methodology[51]. Meanwhile, in clinical or genetic association studies, accurate genomic ancestry is essential. Furthermore, it is often necessary to go a step beyond the genomic average and make inferences about ancestry in specific genomic segments[52,53].

The admixed populations of America are being increasingly studied in terms of population history, clinical and forensic studies. Therefore, nowadays, it is essential to discuss and understand how methodological advances, both in genotyping and in analysis, help to improve the inference of genetic ancestry in admixed populations. In the present study, we analysed data from WGS, HDSNP and AIMs in a Brazilian samples, through different admixture models and compared with other admixed populations of the American continent. We showed that heterogeneity within and between admixed populations still poses methodological challenges. Therefore, it is fundamental when defining the research question, to be aware of the advantages and limitations of each mixture model and set of panels for the populations of interest.

7. MATERIALS AND METHODS

7.1. Datasets

Samples from 3 datasets were analyzed: (i) Human Genomic Diversity Panel (HGDP)[46]; (ii) 1000 Genomes Project phase III (1KGP)[24], and (iii) Brazilian Cohort of Health, Well-being and Aging (*Saúde e Bem Estar* – SABE)[23].

Based on the American admixture history, analyses were performed with parental samples from African, European, East Asian and Native American populations of HGDP (543 individuals) and 1KGP (1511 individuals) as described in Table S1. We also analyzed the 504 samples from the 6 admixed populations from the 1KGP, and the 1,171 Brazilian samples from the SABE cohort (Supplementary Table S1).

All samples were genotyped by WGS and are publicly available (https://www.internationalgenome.org/data - HGDP and 1KGP; https://ega-archive.org/studies/EGAS00001005052 – SABE). All individuals enrolled in the SABE cohort signed written consent forms to participate in this study approved by local and national institutional review boards: COEP/FSP/USP OF.COEP/23/10, CONEP 2044/2014, and CEP HIAE 1263-10.

7.2. Ancestry Informative Markers SNPs panel (AISNP panel)

Due to availability in the 3 datasets, we evaluated only SNPs as AIM. Based on this criterion, we selected 5 AIMs panels frequently used in studies with Latin American populations: 34AISNP[42]+PIMA[19]; 55AISNP[18]; 128 AISNP[43]; 170 AISNP[44]; 446 AISNP[45]. In addition, we also evaluated the combination of the 6 panels, which we named 672 AISNP. The SNPs of the AIMs panels used in the present study are described in Supplementary Table S2.

7.3. High-density SNP chip array (HDSNPs panel)

Axiom™ Genome-Wide Human Origins (~600K SNPs – ThermoFisher Scientific) was selected as a representative of high-density SNP arrays. This genotyping panel was optimized for population genetic studies and developed from genomic markers

identified in 11 human populations: France, China, Papua New Guinea, San, Yoruba, Mbuti pygmies, Karitiana, Italy-Sardinia, Melanesia, Cambodia, and Mongolia, avoiding confounding biases introduced using GWAS SNP arrays.

7.4. Merge Datasets

Based on the WGS data from the 3 datasets, the following sets of SNPs were selected: (i) *AISNP panels*: 672 SNPs comprise the 6 AISNP panels selected for the present study. Of these, 5 SNPs (rs12402499; rs17287498; rs1321333; rs10954737; rs10071261) are not detected in all datasets (Supplementary Table S9), of which, 3 SNPs are informative of Native American ancestry and 2 of African ancestry; (ii) *HDSNPs panel*: ~600,000 SNPs that comprise the Axiom Human Origins array. The overlap between the 3 datasets was 555,168 SNPs, and (iii) *WGS data*: the original datasets with more than 60 million variants described. For the present study, we excluded SNPS: (a) MAF<1%; (b) missing data per SNP >1%; (c) Hardy-Weinberg p-value $<1x10^{-8}$, and (d) filter for LD coefficient (r^2=0.1 – see RESULTS section for more details). The final dataset contains 2,018,023 SNPs. For each set of markers, the 3 datasets (HGDP, 1KGP, SABE) were merged using vcftools v.0.1.15 [54] and plink v.1.9[55]. To validate these merge data, a PCA analysis was performed (Supplementary Figure S17). Throughout the text we refer to AISNP, HDSNP and WGS as "panel sets".

7.5. Data analysis

Allelic frequency inferences, Hardy-Weinberg and Fisher's exact test was performed using the PLINK v.1.9 software[55]. Correction for multiple testing was made according to Bonferroni correction.

7.6. Genetic ancestry inference

ADMIXTURE v.1.3[56] was used to perform global ancestry inference. Analyses were performed in an unsupervised manner when considering only parental

populations, and in a supervised manner when considering admixed populations. Parameters used: 4 clusters (K=4) and 2000 bootstrap replicates. Each analysis was repeated 10 runs and the results combined using the CLUMP software (v.1.222)[57]. Information redundancy may occur in high-density SNP data (HDSNP and WGS). Therefore, to minimize and evaluate background linkage disequilibrium in the analyses, we also tested some disequilibrium linkage coefficients (r^2= 0.01, 0.05, 0.1, 0.3 and 0.5), assuming a distance not closer than 200Kb between adjacent markers[55,56].

Comparisons of ancestry inference were performed by correlation analysis and z-score test by R scripts using package stats v.4.1.1[58].

8. REFERENCES

1. Rohlfs, R. V., Fullerton, S. M. & Weir, B. S. Familial Identification: Population Structure and Relationship Distinguishability. PLoS Genet 8(2): e1002469; 10.1371/journal.pgen.1002469 (2012).

2. Garrison, N., Rohlfs, R. & Fullerton, S. Forensic familial searching: scientific and social implications. Nat Rev Genet. 14, 445; 10.1038/nrg3519 (2013).

3. Romero-Hidalgo, S. et al. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. Nat Commun. 8, 1005; 10.1038/s41467-017-01194-z (2017).

4. Lewontin, R. C. The apportionment of human diversity. In Evolutionary biology (eds Steere W.C., Dobzhansky T., Hecht MK), 381-398 (Springer, 1972).

5. Rosenberg, N.A. et al. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1(6):e70. doi: 10.1371/journal.pgen.0010070 (2005).

6. Serre D. & Pääbo S. Evidence for gradients of human genetic diversity within and among continents. Genome Res. 2004 Sep;14(9):1679-85. doi: 10.1101/gr.2529604.

7. Reidy, K. J. et al. Genetic risk of APOL1 and kidney disease in children and young adults of African ancestry. Current opinion in pediatrics. 30, 252-259; 10.1097/MOP.0000000000000603 (2018)

8. Olczak, K. J. et al. Hypertension genetics past, present and future applications. Journal of internal medicine. 290, 1130-1152; 10.1111/joim.13352 (2021).

9. Mak, W. Y. et al. The epidemiology of inflammatory bowel disease: East meets west. Journal of gastroenterology and hepatology. 35, 380-389; 10.1111/jgh.14872 (2020).

10. Bradbury, C. et al. Off-target phenotypes in forensic DNA phenotyping and biogeographic ancestry inference: A resource. Forensic science international. Genetics. 38, 93-104; 10.1016/j.fsigen.2018.10.010 (2019).

11. Tatonetti, N. P. & Noémie E. Fine-scale genetic ancestry as a potential new tool for precision medicine. Nature medicine. 27, 1152-1153; 10.1038/s41591-021-01405-7 (2021).

12. Mathieson I. & Scally A. What is ancestry? PLoS Genet. 9;16(3):e1008624. doi: 10.1371/journal.pgen.1008624. (2020).

13. Rosenberg, N. A. et al. Informativeness of genetic markers for inference of ancestry. American journal of human genetics. 73, 1402-22; 10.1086/380416 (2003).

14. Shriver, M. D. et al. Skin pigmentation, biogeographical ancestry and admixture mapping. Human genetics. 112, 387-99; 10.1007/s00439-002-0896-y (2003).

15. Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. & Lareu, M. V. Inference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets. Methods in molecular biology. 1420, 233–253; 10.1007/978-1-4939-3597-0_18 (2016).

16. Phillips, C. et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. For. Sci. Int. Genet. 1, 273–280; 10.1016/j.fsigen.2007.06.008 (2007).

17. Pereira, R., et al. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PloS one. 7, e29684; 10.1371/journal.pone.0029684 (2012).

18. Kidd, K. K., et al. Progress toward an efficient panel of SNPs for ancestry inference. Forensic science international. Genetics. 10, 23–32; 10.1016/j.fsigen.2014.01.002 (2014).

19. Gontijo, C.C., et al. PIMA: A population informative multiplex for the Americas. Forensic science international. Genetics. 44, 102200; 10.1016/j.fsigen.2019.102200 (2020).

20. Santos, C., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Carracedo, Á., & Lareu, M. V. Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data. Methods in molecular biology. 1420, 255–285; 10.1007/978-1-4939-3597-0_19 (2016).

21. Santos, H. C., et al. A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. European journal of human genetics: EJHG. 24, 725–731; 10.1038/ejhg.2015.187 (2016).

22. Kehdy, F. S., et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences of the United States of America. 112, 8696–8701; 10.1073/pnas.1504447112 (2015).

23. Naslavsky, M. S., et al. Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. Nature communications. 13, 1004; 10.1038/s41467-022-28648-3 (2022).

24. The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature. 526, 68-74; 10.1038/nature15393 (2015).

25. Gurdasani, D., et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 517(7534), 327–332; 10.1038/nature13997 (2015).

26. Kelleher, J., et al. Inferring whole-genome histories in large population datasets. Nature genetics. 51(9), 1330–1338; 10.1038/s41588-019-0483-y (2019).

27. Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. Inference of population structure using dense haplotype data. PLoS genetics. 8(1), e1002453; 10.1371/journal.pgen.1002453 (2012).

28. Hellenthal, G., et al. A genetic atlas of human admixture history. Science. 343(6172), 747–751; 10.1126/science.1243518 (2014).

29. Wu, J., Liu, Y., & Zhao, Y. Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective. Frontiers in genetics, 12, 639877; 10.3389/fgene.2021.639877 (2021).

30. He, G., et al. Massively parallel sequencing of 165 ancestry-informative SNPs and forensic biogeographical ancestry inference in three southern Chinese Sinitic/Tai-Kadai populations. Forensic science international. Genetics. 52, 102475; 10.1016/j.fsigen.2021.102475 (2021).

31. Adhikari, K., Mendoza-Revilla, J., Chacón-Duque, J. C., Fuentes-Guajardo, M., & Ruiz-Linares, A. Admixture in Latin America. Current opinion in genetics & development. 41, 106–114; 10.1016/j.gde.2016.09.003 (2016).

32. Norris, E. T., et al. Genetic ancestry, admixture and health determinants in Latin America. BMC genomics. 19(Suppl 8), 861; 10.1186/s12864-018-5195-7 (2018).

33. Centro de Documentação e Disseminação de Informações (Brazil). Brazil, 500 years of settlement, (ed IBGE- Instituto Brasileiro de Geografia e Estatística) 197-213p. (Rio de Janeiro, 2007).

34. Korunes, K. L., & Goldberg, A. Human genetic admixture. PLoS genetics. 17(3), e1009374; 10.1371/journal.pgen.1009374 (2021).

35. Hellwege, J. N. et al. Population Stratification in Genetic Association Studies. Current protocols in human genetics. 95, 1.22.1-1.22.23; 10.1002/cphg.48 (2017).

36. Suarez-Pajes, E. et al. Genetic Ancestry Inference and Its Application for the Genetic Mapping of Human Diseases. International journal of molecular sciences.

22, 6962; 10.3390/ijms22136962 (2021).

37. Gouveia M.H. et al. Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. Mol Biol Evol. 1,1647-1656; 10.1093/molbev/msaa033 (2020).

38. Ongaro L., et al. Continental-scale genomic analysis suggests shared post-admixture adaptation in the Americas. Hum Mol Genet. 1, 2123-2134; 10.1093/hmg/ddab177 (2021).

39. Pereira, V., et al. Evaluation of the Precision of Ancestry Inferences in South American Admixed Populations. Frontiers in genetics. 11, 966; 10.3389/fgene.2020.00966 (2020).

40. Debortoli, G., de Araujo, G. S., Fortes-Lima, C., Parra, E. J., & Suarez-Kurtz, G. Identification of ancestry proportions in admixed groups across the Americas using clinical pharmacogenomic SNP panels. Scientific reports. 11(1), 1007; 10.1038/s41598-020-80389-9 (2021).

41. Castro E Silva, M.A., et al. Genomic insight into the origins and dispersal of the Brazilian coastal natives. Proc Natl Acad Sci, 117(5):2372-2377; 10.1073/pnas.1909075117 (2020).

42. Phillips, C., Fondevila, M., & Lareau, M. V. A 34-plex autosomal SNP single base extension assay for ancestry investigations. Methods in molecular biology. 830, 109–126; 10.1007/978-1-61779-461-2_8 (2012).

43. Kosoy, R., et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human mutation. 30(1), 69–78; 10.1002/humu.20822 (2009).

Noah A Rosenberg, Jonathan T L Kang, Genetic Diversity and Societally Important Disparities, Genetics, Volume 201, Issue 1, 1 September 2015, Pages 1–12,

44. Pakstis, A. J., et al. Population relationships based on 170 ancestry SNPs from the combined Kidd and Seldin panels. Scientific reports. 9(1), 18874; 10.1038/s41598-

019-55175-x (2019).

45. Galanter, J. M., et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS genetics. 8(3), e1002554; 10.1371/journal.pgen.1002554 (2012).

46. Bergström, A., et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 367(6484), eaay5012; 10.1126/science.aay5012 (2020).

47. Guo, Y. X., Jin, X. Y., Xia, Z. Y., Chen, C., Cui, W., & Zhu, B. F. A small NGS-SNP panel of ancestry inference designed to distinguish African, European, East, and South Asian populations. Electrophoresis, 41(9), 649–656; 10.1002/elps.201900231 (2020).

48. Rosenberg, N. A., et al. Genetic structure of human populations. Science. 298(5602), 2381–2385; 10.1126/science.1078311 (2002).

49. Soundararajan, U., Yun, L., Shi, M., & Kidd, K. K. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. Forensic science international. Genetics. 23, 25–32; 10.1016/j.fsigen.2016.01.013 (2016).

50. Tang, H., et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. American journal of human genetics. 81(3), 626–633; 10.1086/520769 (2007).

51. Jordan, D. & Mills, D. Past, Present, and Future of DNA Typing for Analyzing Human and Non-Human Forensic Samples. Front. Ecol. Evol. 9,646130; 10.3389/fevo.2021.646130 (2021).

52. Blue, E. E., Horimoto, A., Mukherjee, S., Wijsman, E. M., & Thornton, T. A. Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. Alzheimer's & dementia: the journal of the Alzheimer's Association. 15(12), 1524–1532; 10.1016/j.jalz.2019.07.016 (2019).

53. Horimoto, A., et al. Genome-Wide Admixture Mapping of Estimated Glomerular Filtration Rate and Chronic Kidney Disease Identifies European and African

Ancestry-of-Origin Loci in Hispanic and Latino Individuals in the United States. Journal of the American Society of Nephrology: JASN. 33(1), 77–87; 10.1681/ASN.2021050617 (2022).

54. Danecek P.,The variant call format and VCFtools. *Bioinformatics.* 27(15): 2156-2158; 10.1093/bioinformatics/btr330 (2011).

55. Chang, C. C., et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4, 7; 10.1186/s13742-015-0047-8 (2015).

56. Alexander, D. H., Novembre, J., & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 19(9), 1655–1664; 10.1101/gr.094052.109 (2009).

57. Jakobsson, M., and Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 23, 1801–1806; 10.1093/bioinformatics/btm233 (2007).

58. R Core Team (2021). R: language and environment for statistical computing. R Foundation for Statistical Computing. Viena. Austria. URL: https://www.R-project.org/.

ACKNOWLEDGEMENTS

This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES; National Council for Scientific and Technological Development – CNPq (PhD scholarship - L.M.E.S.).

AUTHOR CONTRIBUTIONS

S.O and K.N contributed to study conception and design; M.S.N., M.O.S., Y.A.O.D. and M.Z provided SABE cohort samples and genomic data; L.M.E.S and K.N. performed the analyses; L.M.E.S, K.N and S.O wrote the manuscript and incorporated input from other authors.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The datasets reported in this article are publicly available (https://www.internationalgenome.org/data - HGDP and 1KGP; European Genome-phenome Archive (EGA), under EGA Study accession number EGAS00001005052 – SABE).

MATERIAL SUPLEMENTAR

O Material Suplementar encontra-se no arquivo eletrônico disponibilizado junto com a tese.

CAPÍTULO 2

CARACTERIZAÇÃO GENÉTICA E DEMOGRÁFICA DO DISTRITO FEDERAL

1. INTRODUÇÃO AO CAPÍTULO 2

Neste capítulo concatenamos dados sobre a ancestralidade genética e informações demográficas visando contribuir com uma melhor compreensão da história recente de formação da população do Distrito Federal. Para esse propósito, coletamos material genético e aplicamos um questionário demográfico para 104 participantes em sua maioria nascidos na década de 1980 no Distrito Federal. As amostras foram genotipadas com um array de alta densidade de SNPs, especialmente desenvolvido para estudos populacionais (*Axiom Human Origins* - ThermoFisher™). Esse SNParray apresenta mais de 600.000 marcadores genéticos distribuídos pelos cromossomos autossômicos, sexuais (X e Y) e mtDNA. Com base nos resultados obtidos no capítulo 1 e na informação de haver na amostra descendentes de japoneses, as inferências de ancestralidade genética, a partir dos cromossomos autossômicos e do cromossomo X, foram realizadas usando o modelo de miscigenação tetra-híbrido. A ancestralidade genética foi inferida para os cromossomos autossomos, sexuais (X e Y) e mtDNA e a comparação entre eles permitiu averiguar possíveis sinais de casamentos direcionais com assimetria de ancestralidade genética.

Para responder o quão similar ou distintas as estimativas de ancestralidade genética para a amostra do DF são em relação a população brasileira e as cinco regiões geográficas do país, os resultados gerados no presente estudo foram comparados e testados com dados da literatura. Para melhor compreender como cada região do Brasil contribuiu para o perfil de ancestralidade genética da amostra do DF, usamos a ancestralidade genética e a origem migratória dos genitores dos participantes para testar um modelo simplificado de miscigenação. Além disso, com base nas informações do questionário aplicado, também caracterizamos o perfil de migração do DF com base em cada região e estado brasileiro, gênero, e padrões matrimoniais dos genitores e avós dos participantes do estudo.

2. RESUMO

O Distrito Federal (DF), localizado na região Centro-Oeste do Brasil, possui uma história peculiar de povoamento: sua população começou a se constituir efetivamente na década de 60 com a transferência da capital federal do Rio de Janeiro para o Planalto Central. Desde então, o DF tem recebido imigrantes das diferentes regiões do Brasil. Esse cenário despertou o interesse de nosso grupo de pesquisa em caracterizar a ancestralidade genética e compreender mais profundamente a diversidade genética e demográfica do DF. A ancestralidade genética de 104 brasilienses, nascidos na década de 80, representando 16 diferentes regiões administrativas do DF, foi inferida com base em *array* de alta densidade de SNPs (Axiom Human Origins - Affymetrix / Thermo Fisher). Como população parental, usamos dados de amostras africanas, europeias, nativas americanas e do leste asiático, obtidos do Projeto de Diversidade do Genoma Humano (*Human Genome Diversity Project* - HGDP). Usando a abordagem implementada no software ADMIXTURE e o modelo tetra-híbrido de miscigenação, foi possível inferir a proporção de contribuição das ancestralidades parentais na amostra do DF. Com base na ancestralidade genética inferida a partir dos cromossomos autossômicos, foi observada maior contribuição média do componente ancestral europeu e leste asiático (1,68%±8,6). (70%±18%), seguido pelo africano (19%±14,4), nativo americano (8,5%±7,2) Essas proporções de ancestralidade estão de acordo com a média da proporção de contribuição dos componentes parentais EUR, AFR e NAM da população brasileira (p-valor>0,1) e com o modelo de migração interna do Brasil para o DF (p-valor= 0,22). Com base na ancestralidade genética inferida para o cromossomo X, foi observada contribuição média europeia de 63,1% (±26,1%), africana de 21,7% (±21,6), nativo americana de 13,8% (±13,8) e leste asiática de 2,1% (±12,1). A inferência dos componentes ancestrais europeus e nativo americano apresentam diferenças significativas entre os cromossomos autossômicos e X (p-valor=0,006), indicando casamentos direcionais, refletidos no processo de miscigenação da população brasileira e consequentemente no processo de miscigenação e formação da população do DF. Esse resultado é corroborado pelas análises do cromossomo Y e mtDNA, onde os haplogrupos inferidos para o cromossomo Y revelam que 90,7% são de origem europeia e 9,3% de origem africana e para os haplogrupos

do mtDNA 37,5% são de origem africana, 31,7% europeia, 28,8% nativo americano e 1,9% do leste asiático. Para compreender os processos migratórios, aplicamos um questionário sobre o local de nascimento dos genitores e avós de cada participante. Observamos que a região Sudeste foi responsável por 43,6%e Nordeste por 38,22% dos migrantes, seguida pelo Centro-Oeste (9,3%), Norte (5,5%) e Sul (1,6%). No presente estudo, caracterizamos a população do DF como tendo uma grande contribuição genética europeia, mas também com expressivas contribuições africana e nativo americana. As análises dos cromossomos X, Y e do mtDNA mostraram a predominância do fenômeno demográfico de casamento direcional, com homens de ascendência européia e mulheres com ascendência africana e nativa americana. Esse fenômeno provavelmente é anterior à formação do DF e é característico do processo prévio de formação da população brasileira. Portanto, a resposta das análises dos marcadores autossômicos bem como dos uniparentais estão alinhadas ao contexto histórico de povoamento e miscigenação da população brasileira. De forma geral, este estudo corrobora análises anteriores e ajuda esclarecer com maior precisão como ocorreu e qual o reflexo do processo de migração para o DF.

Palavra-chave: Ancestralidade genética; genética populacional; Distrito Federal; reconstrução histórica.

3. INTRODUÇÃO

Compreender as peculiaridades da história de formação de uma população é o primeiro passo para o enriquecimento do entendimento acerca do padrão genético atual desta. O conhecimento dos eventos demográficos passados até a atualidade se somam às tecnologias direcionadas a estudos genéticos populacionais possibilitando o acesso à informações genômicas cada vez mais refinadas, robustas e, permitindo uma interpretação das forças e eventos que desenharam e desenham ainda a constituição genética de uma população de estudo.

No presente capítulo, iremos explorar a história de formação da população do Distrito Federal a partir de informações genéticas e demográficas. Os dados genéticos permitiram identificar padrões de miscigenação dos componentes genéticos ancestrais de origem continental, enquanto que as informações demográficas ajudaram a contar o processo de migrações internas do Brasil para a formação do Distrito Federal. A seguir, faremos um breve resumo sobre a história do povoamento da região Centro-Oeste do Brasil, como ocorreu o processo de formação e povoamento do Distrito Federal e os principais estudos genéticos nessa população.

3. 1 POVOAMENTO DA REGIÃO CENTRO-OESTE DO BRASIL

Por ser uma região interiorana, a migração de indivíduos com ancestralidade distinta dos povos originários para a região Centro-Oeste foi tardio em relação às outras regiões do Brasil. Há relatos que a partir do século XVII ocorreram os primeiros movimentos exploratórios promovidos pelo governo (“entradas”) ou pelas expedições particulares (“bandeiras”). Estas expedições tinham o objetivo de explorar metais preciosos, promover a colonização de regiões interioranas, capturar indígenas e escravos afrodescendentes fugitivos. Grande parte da região Centro-Oeste pertencia à coroa de Castela e Espanha, mas o avanço das bandeiras para o oeste tornou-a de domínio de Portugal (Bertran, 2000; Palacin *et al*., 1995).

A descoberta de minas de ouro, nos primórdios do século XVIII, propiciou o primeiro evento migratório, trazendo migrantes para esta região, principalmente de Minas Gerais e São Paulo. Esse movimento promoveu, dentre outras modificações na região, a fundação de vilas como Vila Real do Bom Jesus de Cuiabá, atual capital do estado de Mato Grosso, e Vila Boa, hoje a cidade de Goiás (Bertran, 2000; Palacín *et*

al., 1995). Com a decadência da mineração no século XIX, a pecuária, a agricultura (cana-de-açúcar e erva-mate) e a extração da borracha tiveram um avanço. Em decorrência disso, ocorreu o segundo evento migratório. Os migrantes desta segunda fase eram advindos principalmente do atual estado do Maranhão e dos estados de Minas Gerais e São Paulo. Isso viabilizou a construção de estradas e a consolidação do desenvolvimento local, além da construção de novas cidades que atraíram mais migrantes (Palacín & Moraes, 1994).

Já no século XX, ocorreram três principais fluxos de migração para o Centro-Oeste, o primeiro, quando migrantes gaúchos lideraram o desenvolvimento de estradas de ferro e produção de mate, além da agropecuária. O segundo, no governo de Getúlio Vargas (1937-1945), na primeira dita "Marcha para o Oeste", com a intenção de povoar a região central do Brasil. A terceira e mais efetiva, ocorreu durante o governo de Juscelino Kubitschek (1956-1961) com a construção da Capital Federal, Brasília. Nesse momento uma nova "Marcha para o Oeste" foi instaurada, e deslocou principalmente migrantes nordestinos para a região (Cassiano, 2002).

3.2 POVOAMENTO DO DISTRITO FEDERAL (DF)

O recorte do presente estudo está focado no Distrito Federal. Localizado entre os paralelos 15º30’ e 16º03’ de latitude sul e os meridianos 47º25’ e 48º12’ de longitude WGr, na região Centro-Oeste do Brasil, ocupando uma área de 5.789,16 km^2 no Planalto Central do Brasil, centro-leste do Estado de Goiás, equivalente a 0,06% da superfície do país (CODEPLAN, 2012). O Distrito Federal é a menor unidade federativa do Brasil e atualmente compreende 33 regiões administrativas (Figura 01), tendo em seu território a capital federal, Brasília.

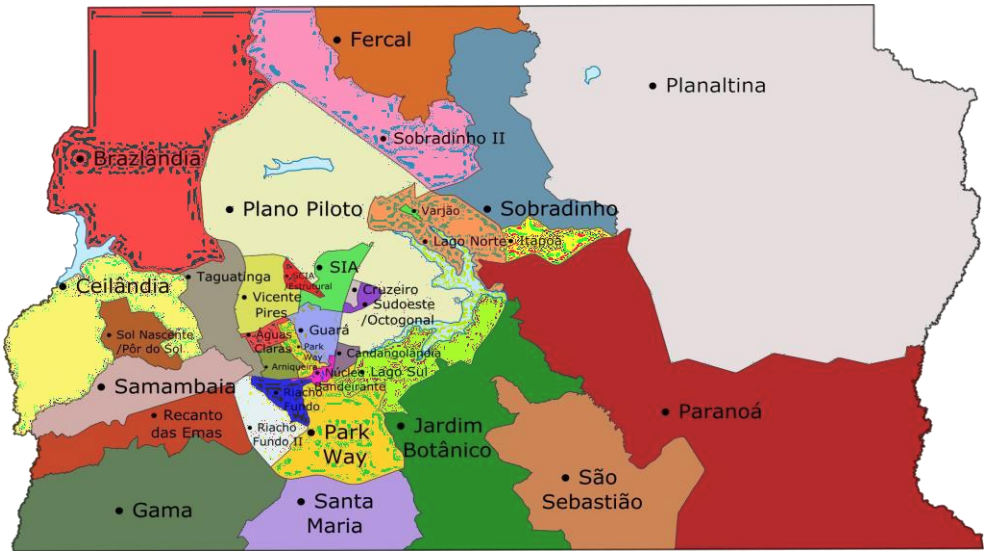


Figura 01. Mapa do Distrito Federal. O mapa mostra as 33 unidades administrativas do DF, de acordo com a Lei Complementar nº 958, de 20 de dezembro de 2019.

Fonte: https://pt.wikipedia.org/wiki/Regi%C3%B5es_administrativas_do_Distrito_Federal_(Brasil).

A população do Distrito Federal começou a ser formada efetivamente na década de 60 com a transferência da capital federal do Rio de Janeiro para o Planalto Central. Diferente das demais regiões brasileiras, seu povoamento se deu pela migração rápida de indivíduos provenientes de todas as regiões do país (em especial Sudeste, Nordeste e Centro-Oeste), onde as populações já vinham de um processo prévio de miscigenação. No ano de 1956, a população da região do Distrito Federal foi estimada em cerca de 6 mil, e o no ano de 1970 havia crescido para mais de 400 mil habitantes. Nesse ano, apenas 22% da população era nativa do DF, sendo 41% provenientes do Nordeste, 37% do Sudeste e 17% do Centro-Oeste (CODEPLAN, 2013).

Em 2010, o Censo apontou que, dos mais de 2,5 milhões de habitantes, cerca de 54% haviam nascido no DF. Em 2021, a população residente estimada foi de 3.094.325 pessoas de acordo com o SIDRA (IBGE – Instituto Brasileiro de Geografia e Estatística. População residente estimada. IBGE, 2010). O retrato que se tem da população do Distrito Federal é de uma série de particularidades, raramente vistas em outro lugar do mundo, com uma formação recente de rápido crescimento, fundada em um território artificial, formada a partir de populações heterogêneas oriundas de migrações internas de outras regiões do Brasil. Portanto, compreender a composição genética e os movimentos migratórios que contribuíram para a formação do DF, são fundamentais para determinar o quanto essa população é representativa ou distinta das demais regiões do Brasil.

3.3 ESTUDOS SOBRE ANCESTRALIDADE GENÉTICA DO DF

Ao longo dos últimos anos, o Laboratório de Genética Humana da UnB vem realizando estudos, com intuito de compreender melhor a dinâmica populacional e a composição genética da população do DF. Os primeiros estudos de linhagens uniparentais foram feitos por Barcelos (2006) que, integrando nosso grupo, analisou a composição genética da população de DF a partir de haplogrupos do cromossomo Y e o mtDNA mitocondrial. Um haplogrupo é um conjunto de variações genéticas específicas que representa uma linhagem de origem paterna (no caso do cromossomo Y) ou materna (no caso do DNA mitocondrial) e, em geral, apresentam distribuição geográfica distinta entre regiões do mundo (Zerjal *et al.,* 2001).

Os resultados obtidos por Barcelos (2006) apontaram que no DF, 90% dos cromossomos Y são de linhagens de origem europeia, 8,5% da africana subsaariana e 1,5% nativo americana. Quanto às linhagens maternas, as análises de mtDNA indicaram que 18% da população do DF têm linhagens europeias, 27% têm linhagens africanas e 44% têm linhagens nativo americanas. Godinho (2008), também integrante do nosso grupo, analisou as mesmas características populacionais, mas a partir de dados de marcadores situados nos cromossomos autossômicos. Utilizando dados de marcadores do tipo microssatélite (*Short Tandem Repeat - STRs*) e marcadores informativos de ancestralidade (*Ancestry informative Markers* - AIMs), mostrou que a contribuição europeia variou ao redor de 60%, a africana entre 25% e a nativo americana, ao redor de 15%.

Esses trabalhos possibilitaram traçar um quadro inicial da composição e estruturação genética do Distrito Federal e ainda tem sido útil à áreas afins, como, por exemplo, a genética forense (Trindade-Filho, 2010; Dalton, 2010), médica (Toledo, 2016; Brandão, 2015; Arcanjo *et al.,* 2013) e antropológica (Arcanjo, 2016; Arcanjo, 2012; Godinho, 2008; Barcelos, 2006).

4. HIPÓTESE

Como o Distrito Federal foi fundado na década de 60 a partir do movimentos migratórios de pessoas oriundas de diferentes regiões do Brasil, temos como hipótese que o perfil genético dessa população é similar a do Brasil como um todo e que o estudo de adultos que nasceram em Brasília pode fornecer informação suficiente sobre a contribuição de cada região do país e seus reflexos na composição genética da população do DF com base em dados genéticos e em informações demográficas.

5. OBJETIVO GERAL

Refinar o entendimento sobre a ancestralidade genética e os movimentos migratórios que resultaram na formação da população do Distrito Federal.

5.1 OBJETIVOS ESPECÍFICOS

- Caracterizar a ancestralidade genética da população do Distrito Federal considerando prioritariamente a geração nascida nos anos 80 com relação aos componentes parentais europeu, africano, asiático e nativo americano;
- Avaliar a intensidade de ocorrência de casamentos direcionais a partir da comparação da ancestralidade genética inferida para os cromossomos autossômicos, sexuais (X e Y) e no DNA mitocondrial.
- Verificar se o perfil de ancestralidade genética no DF é similar ou não a outras regiões do Brasil.
- Verificar se o perfil migratório dos ascendentes dos participantes é condizente com a estrutura genética atual da população do DF.

6. MATERIAL E MÉTODOS

6.1. ASPECTOS ÉTICOS

Este estudo foi aprovado pelo Comitê de Ética e Pesquisa CEP/CONEP da Faculdade de Ciências da Saúde da Universidade de Brasília (UnB), CAAE:72917916.3.0000.0030. O projeto também foi aprovado e contou com o apoio financeiro da Fundação de Apoio à Pesquisa do Distrito Federal (FAP/DF).

Previamente à coleta de dados e material biológico, os participantes da pesquisa receberam esclarecimentos referentes à proposta e então, prontificados a participar, foi solicitada a leitura e assinatura do TCLE deste projeto (Apêndice 1). Em seguida, os mesmos responderam a um questionário que tinha por objetivo acessar dados demográficos incluindo ascendentes e descendentes diretos (Apêndice 2).

Ao participante da pesquisa ficou claro que o mesmo estará contribuindo para um melhor conhecimento do povoamento do Distrito Federal, propiciando a ele um entendimento das suas origens. Foi acordado que o mesmo, receberia como benefício à colaboração, o direito ao acesso a sua ancestralidade genética, entregue sob forma de relatório, permitindo assim, ao participante, o resgate de suas origens ancestrais (Apêndice 3). Os relatórios já foram entregues a cada um dos participantes.

6.2. GRUPO AMOSTRAL

Foram analisados 104 indivíduos, de ambos os sexos (43 homens e 61 mulheres), nascidos no DF entre 1970 e 1989 representantes de 16 das 18 regiões administrativas do Distrito Federal na época (Tabela 01 e Figura 02). O critério de inclusão dos indivíduos na amostra teve como base a região administrativa onde os pais moravam à época do nascimento do participante. Para tanto, foi estabelecido um número máximo de participantes por região administrativa utilizando como base o número de habitantes de cada região de acordo com o censo de 1989. Além disso, os participantes da pesquisa deveriam ser filhos de pessoas com residência fixa no DF à época do nascimento e não ser aparentados com algum outro participante da amostra. A média de idade dos participantes da pesquisa foi de 35 anos, apresentando um desvio padrão de ±7,04 anos. O indivíduo mais velho apresentava 55 anos e o mais jovem 22 anos de idade à época da coleta.

Tabela 01. Distribuição do grupo amostral entre 18 regiões administrativas do DF existentes à época do censo de 1989.

REGIÕES ADMINISTRATIVAS	NÚMERO DE PARTICIPANTES
ASA SUL	16
ASA NORTE	15
CEILÂNDIA	20
LAGO SUL	1
SOBRADINHO	4
CRUZEIRO	1
CANDANGOLÂNDIA	1
SAMAMBAIA	6
GAMA	8
GUARÁ	8
NÚCLEO BANDEIRANTE	3
LAGO NORTE	2
VILA PLANALTO	2
PLANALTINA	1
TAGUATINGA	14
ÁREA OCTOGONAL	1
BRAZILÂNDIA	0
PARANOÁ	0

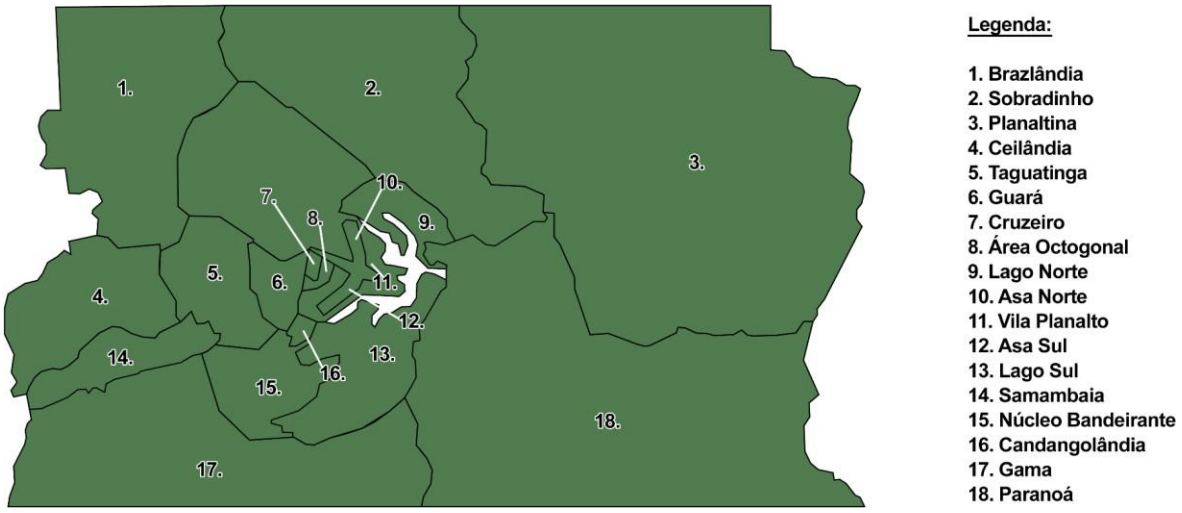


Figura 02. Mapa das regiões administrativas do DF no ano de 1989. O mapa mostra as delimitações de cada região admistrativa no ano de 1989. Os números no mapa representam o nome de cada região, de acordo com a legenda ao lado. No nosso estudo temos participantes de 16 das 18 regiões administrativas, sendo apenas Brazlândia e Paranoá que não tiveram representantes.

Fonte: Codeplan 1990 (modificado).

6.3 TRATAMENTO LABORATORIAL DAS AMOSTRAS

O projeto teve sua parte laboratorial executada no Laboratório de Genética Humana do Departamento de Genética e Morfologia do Instituto de Ciências Biológicas da UnB, onde aconteceram também as análises dos dados. Tivemos o apoio do INCOR-USP (Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo) para o ensaio de genotipagem das placas e SNPArray.

6.3.1 Extração, integridade e quantificação das amostras

Foi coletado de cada participante de pesquisa 8 ml de sangue venoso do antebraço utilizando sistema de coleta a vácuo com agulha estéril e descartável, seguindo o procedimento padrão usado para coleta de sangue venoso no Laboratório de Genética Humana da UnB.

A partir das amostras de sangue, seguiu-se para o processo de extração de DNA, seguindo o protocolo do método Puregene (QUIAGEN). Cada amostra extraída teve a sua integridade verificada em gel de agarose a 1% e imagens obtidas por transiluminação. O processo de quantificação das amostras foi realizado no equipamento Nanodrop ND-1000, da Thermo Scientific. Na etapa seguinte, diluiu-se as

mesmas de forma a ajustar a concentração para a etapa de genotipagem, seguindo o protocolo para SNParray da Affymetrix/Thermo Fisher (10ng/µl), a um volume final de 50µl por amostra. Novamente se quantificou em duplicata todas as amostras para confirmação da concentração padrão necessária à próxima etapa. O DNA extraído e quantificado foi armazenado em microtubos devidamente etiquetados com códigos individuais sem, no entanto, dissociar as amostras dos dados de identificação individual dos participantes.

6.3.2 Do armazenamento

Os materiais biológicos foram mantidos em freezer -20ºC com controle de temperatura e alarme de segurança que monitora oscilações de temperaturas divergentes desse valor, o que garantiu a conservação e estabilidade das amostras para os próximos ensaios.

6.4 GENOTIPAGEM POR SNP ARRAY

As genotipagens foram realizadas com o SNPArray de alta densidade Axiom™ Genome-Wide Human Origins (Thermo Fisher Scientific), o qual apresenta aproximadamente 600.000 marcadores do tipo SNPs. Esse SNPArray foi desenvolvido pela empresa Thermo Fisher Scientific para estudos de genética populacional em humanos e não apresenta viés de representação de SNPs como os observados naqueles de interesse clínicos/farmacogenéticos usados nos estudos de associação com doenças (Lu *et al.,* 2011). Os SNPs foram descobertos comparando dois cromossomos do mesmo indivíduo de ancestralidade conhecida e depois genotipados em um painel maior de amostras da mesma população. A estratégia de escolha dos SNPs que compõe o painel da Axiom Human Origins 1 Array é descrita no trabalho de pesquisa de Keinan A., et al., Nature Genetics (2007). O Axiom Human Origins 1 Array é uma união de 13 painéis diferentes. Os primeiros 12 painéis contêm dezenas de milhares de SNPs por população,sendo elas:1) French, 2) Han Chinese, 3) Papuan, 4) San Bushman, 5) Yoruba, 6) Mbuti Pygmies, 7) Karitiana, 8) Sardinian, 9) Melanesian, 10) Cambodian, 11) Mongolian 12) Papuan2 . O 13º painel é baseado no alinhamento de três genomas diferentes:chipanzé, Chipanzé, Denisova e San Bushman. Detalhes estão descritos em https://assets.thermofisher.com/TFSAssets/LSG/brochures/axiom_hu_origins_datasheet.pdf

6.4.1 Ensaio Simplificado do Sistema de Genotipagem AXIOM

A Figura 3 mostra um esquema simplificado do sistema de genotipagem da tecnologia escolhida para este trabalho. O DNA genômico total (50 ng) é amplificado e aleatoriamente fragmentado em segmentos de 25 a 125 pares de bases (pb). Esses fragmentos são purificados, ressuspensos e hibridados em placas desenhadas especificamente para o Axiom™ Genome-Wide Human Origins Array (Thermo Fisher Scientific). Após a hibridação, o alvo ligado as sondas é lavado sob condições rigorosas para remover inespecificidades, minimizando o ruído de fundo causado por eventos de ligação aleatórios. Após a ligação, as matrizes são coradas e fotografadas no instrumento multicanal GeneTitan™. Ao final, os genótipos são estabelecidos com

auxílio dos softwares Axiom Analysis Suite e Axiom Power Tools.

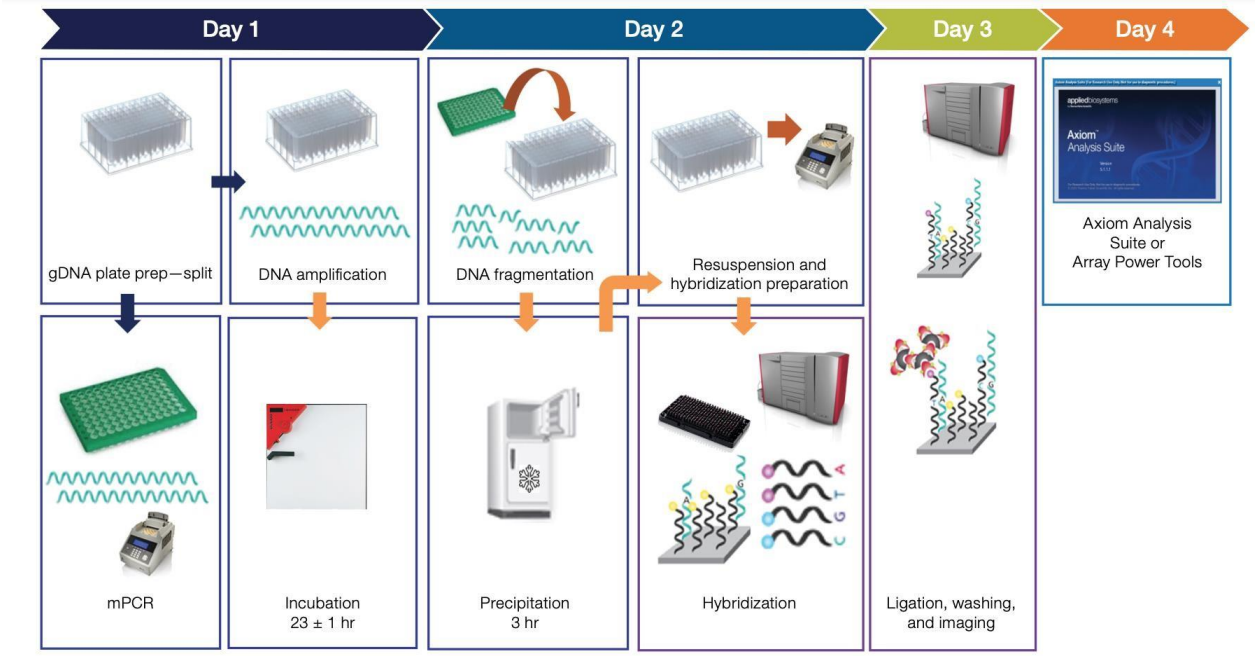


Figura 03. Esquema de genotipagem do sistema AXIOM™ Genome-Wide Human Origins Array. A descrição sobre o ensaio está no texto a cima. Fonte: Thermo Fisher - Analysis Guide Axiom.

6.4.2 Controle de qualidade e filtragem dos dados

O sistema Axiom conta com o software Axiom Analysis Suite para filtragem e limpeza dos dados gerados. A figura 04 demonstra como se dá esse fluxo de trabalho. Primeiro, arquivos do tipo .CEL, os quais armazenam os resultados dos cálculos de intensidade nos valores de pixel fotografados, são selecionados. Em seguida são realizados vários testes de qualidade para identificação e exclusão de marcadores com baixa resolução de genotipagem, amostras com grande número de dados faltantes – *missing data* (*threshold* 5%) - e identificação de eventuais placas problemáticas (placa com padrão de genotipagem muito discrepante em relação às demais - *outliers batches*).

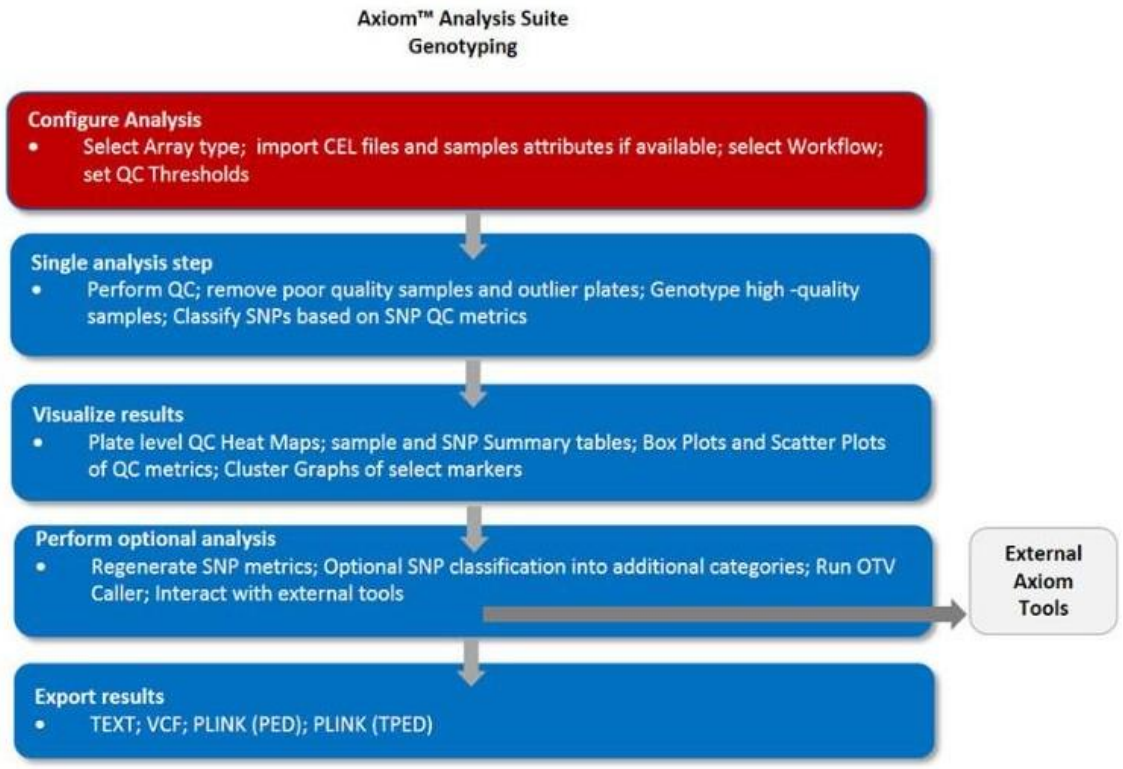


Figura 04. Fluxo de trabalho do controle de qualidade e filtragem pelo Axiom™ Analysis Suite.
Fonte:https://www.thermofisher.com/br/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html

6.5 ANÁLISE DOS DADOS

6.5.1 Grupos Parentais

Levando em consideração o conhecimento histórico do povoamento do Centro-Oeste e estudos genéticos anteriores realizados nessa região, foi decidido que as populações parentais a serem utilizadas nas análises de ancestralidade genética seriam africanos, europeus, leste asiatico e nativos americanos. Os dados para as análises foram obtidos a partir do Projeto de Diversidade do Genoma Humano (*Human Genome Diversity Project - HGDP*), que tem em seu banco público amostras genotipadas para o mesmo array de alta densidade de SNPs usado no presente estudo (https://cephb.fr/en/hgdp_database.php; Dataset 11). Foram utilizados dados de 105 individuos não aparentados da África, 93 da Europa, 139 do Leste Asiático e 36 da América.

6.5.2 Integração dos dados

A integração dos dados públicos com as amostras genotipadas no presente estudo foi realizada através do software Plink v.1.9 (Chang *et al.,* 2015). Para tanto, foi utilizado o comando *--bmerge* para integrar os dados e o comando *--merge-mode* para que a integração fosse realizada apenas para os SNPs comuns aos dois bancos de dados. Uma vez que os dados foram genotipados com o mesmo SNPArray e usaram o mesmo genoma referência (hg19), não foi necessário realizar nenhum ajuste adicional para a integração dos dados.

Em seguida, a função *snpgdsPCA* do pacote em linguagem computacional R, SNPRelate (Zheng *et al.,* 2012) foi usada para realizar uma análise de componente principal (PCA) e verificar se os dados foram integrados de forma correta (figura 5). Para essa análise, usamos a função *snpgdsLDpruning*, também do pacote SNPRelate, para aplicar um filtro de desequilíbrio de ligação (r^2<0.1) visando obter marcadores genômicos (SNPs) independentes.

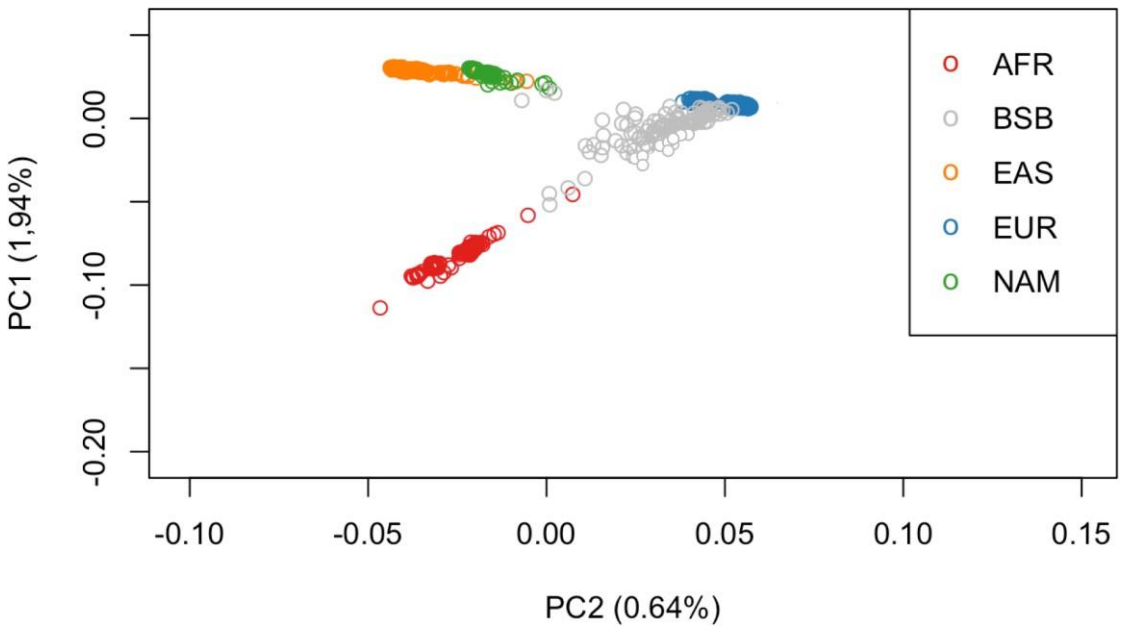


Figura 05. Análise de Componente Principal (PCA) para amostra do DF em relação as populações parentais. Os eixos X e Y representam o segundo e o primeiro componente principal respectivamente (em parênteses a porcentagem da variância explicada por cada componente principal). Cada círculo, representa um indivíduo, as cores vermelha, azul, laranja, verde e cinza representam os grupos populacionais africanos (AFR), europeus (EUR), leste asiáticos (EAS), nativo americanos (NAM) e a amostra do Distrito Federal (BSB) respectivamente.

6.5.3 Inferência da Ancestralidade Genética

A ancestralidade genética foi inferida com o programa ADMIXTURE v.1.3 (Alexander, Novembre & Lange 2009). Esse programa usa uma abordagem de máxima verossimilhança para estimar a ancestralidade individual a partir de genótipos de SNPs multilocus. O método usa um modelo de relaxamento de bloco para que com base nas frequências alélicas as frações de ancestralidade do indivíduo possam ser inferidas. As inferências foram realizadas com SNPs independentes (coeficiente de desequilíbrio de ligação r^2=0.1), usando análises supervisionadas (flag *--supervised*), com quatro clusters ancestrais (K=4), correspondendo às ancestralidades parentais africana, europeia, leste asiática e nativo americana. Para evitar problemas de multimodalidade, foram realizadas 10 corridas, com 2000 réplicas de *bootstrap* cada. Os resultados foram concatenados usando o software CLUMPP v.1.222 (Jakobsson & Rosenberg 2007).

Para as análises envolvendo o cromossomo X, as análises no ADMIXTURE foram realizadas de modo separado para homens e mulheres. Para os homens, neste caso específico, foi usado a flag *--haploid* para indicar que eles apresentam apenas um cromossomo X, como recomendado pelos autores do programa.

6.5.4 Inferência de Haplogrupos para Cromossomo Y

As inferências haplotípicas para o cromossomo Y foram realizadas com o software SNAPPY (Serverson *et al.,* 2018). Esse programa foi desenvolvido para dados de SNPArrays e utiliza como referência a biblioteca do banco de dados ISOGG Y-Tree (*International Society of Genetic Genealogy*; https://isogg.org/tree/). O programa testa os haplótipos do cromossomo Y contidos no banco de dados do ISOGG para encontrar o haplogrupo que é melhor suportado pelos genótipos da nossa amostra. Para isso, é atribuído um *score* para cada haplótipo, o qual utiliza o número de SNPs informativos do haplogrupo, ajustado pelo número de posições informativas que estão omissas na amostra do indivíduo. Ao final, o programa gera dois *scores*, um atribuído aos nós ancestrais na árvore filogenética e outro ao haplogrupo. Desta forma o método considera tanto a quantidade de suporte para o haplogrupo específico, quanto seus haplogrupos ancestrais filogenéticos, fortalecendo a precisão das atribuições do programa. Como parâmetros foram usados *--min_hap_score* 0.8 (*score* mínimo para o haplogrupo) e *--min_deep_score 0.9* (*score* mínimo para os nós da árvore filogenética).

6.5.5 Inferência de Haplogrupos para DNA mitocondrial (mtDNA)

A inferência dos haplogrupos de DNA mitocondrial foram realizadas usando o software HaploGrep3 v3.3.1.0 (Weissensteiner *et al.,* 2016). O software usa informações da árvore filogenética do DNA mitocondrial (http://phylotree.org/) para classificar as variantes genéticas observadas em haplogrupos.

As análises foram realizadas usando arquivos no formato vcf (convertidos no programa Plink, com o argumento *--recode vcf*), contendo apenas as variantes do mtDNA, os quais foram alinhados contra a sequencia de referência do DNA mitocondrial humano (*Revised Cambridge Reference Sequence* - rCRS, hg19/build37, Phylotree v.17). Para indicar que os dados analisados são de SNPArray, a flag *--chip* foi usada, limitando a análise apenas aos SNPs contidos no Axiom Genome-Wide Human Origins Array. Para a definição dos haplogrupos, como critério, foi utilizado o melhor *hit* das variantes observadas com aquelas que definem o haplogrupo (*best hit*) e um *score* de qualidade geral acima de 90%.

6.5.6 Análise de heterogeneidade de ancestralidade entre os cromossomos

Em populações miscigenadas, padrões específicos de casamentos não aleatórios, com contribuição ancestral diferente entre homens e mulheres, podem resultar em diferenças nas proporções de ancestralidade entre os cromossomos autossomos e o cromossomo X. Para identificar essas possíveis diferenças foi aplicado o método CAnD (*Chromosomal Ancestry Difference*), implementado no pacote com mesmo nome, na linguagem computacional R (McHugh *et al.,* 2016). O método utiliza uma abordagem multivariada para testar diferenças sistemáticas nas contribuições genéticas das populações ancestrais entre os cromossomos do próprio indivíduo, bem como entre indivíduos da população. Para tanto, o método verifica se uma ancestralidade específica de um dado cromossomo difere da média dos demais cromossomos no indivíduo e da amostra populacional como um todo. Sob a hipótese nula de não haver diferenças de proporção de uma dada ancestralidade entre os cromossomos, espera-se que o teste CAnD siga a distribuição do x^2 (qui-quadrado) com 1 grau de liberdade. Portanto, a partir das diferenças observadas na proporção de ancestralidade entre os cromossomos foi possível calcular um valor de p, o qual foi corrigido para múltiplos testes (correção de Bonferroni).

7 RESULTADOS E DISCUSSÃO

7.1 Perfil da ancestralidade genética individual estimada a partir de marcadores autossômicos e autodeclaração de "cor/raça"

Como mencionado na introdução do Capítulo 2, estudos prévios sobre a população do Distrito Federal à luz da genética de populações foram iniciados no ano de 2000 por nosso grupo na UnB, no Laboratório de Genética Humana (Barcelos 2006 e Godinho 2008). Quase quinze anos depois, com a possibilidade de novas tecnologias, o presente estudo se propôs novamente a investigar a estrutura genética da população do Distrito Federal a fim de avançar no entendimento fino e concatenar dados genéticos e demográficos com intuito de preencher possíveis lacunas existentes por limitações de tecnologias e meios de análises anteriormente indisponíveis.

Neste trabalho, a ancestralidade genética de 104 indivíduos representantes de 16 diferentes regiões administrativas do DF, foi inferida com base em aproximadamente 600 mil marcadores genéticos do array de alta densidade de SNPs (Axiom Human Origins - Thermo Fisher Scientific).

Num primeiro momento, foi possível observar que a proporção de cada componente ancestral variou entre os indivíduos da amostra do DF (figura 06). Além disso, a figura agrupou os indivíduos pela autodeclaração dos participantes de acordo com as categorias de "cor/raça" pré-estabelecidas pelo IBGE (*"amarelo", "branco", "indigena", "pardo", "preto"*). Nosso objetivo ao usarmos as informações de autodeclaração nas categorias pré-estabelecidas pelo IBGE não foi assumir que elas são grupos biológicos naturais, mas sim reforçar a percepção de que há um ruído em usar essas categorias como equivalente de ancestralidade genética do indivíduo. Esse padrão não é exclusivo da população do DF, mas sim recorrente nos estudos com amostras de diferentes regiões do Brasil (ex. Pena *et al.,* 2011; Nunes *et al.,* 2020, Naslavsky *et al.,* 2022).

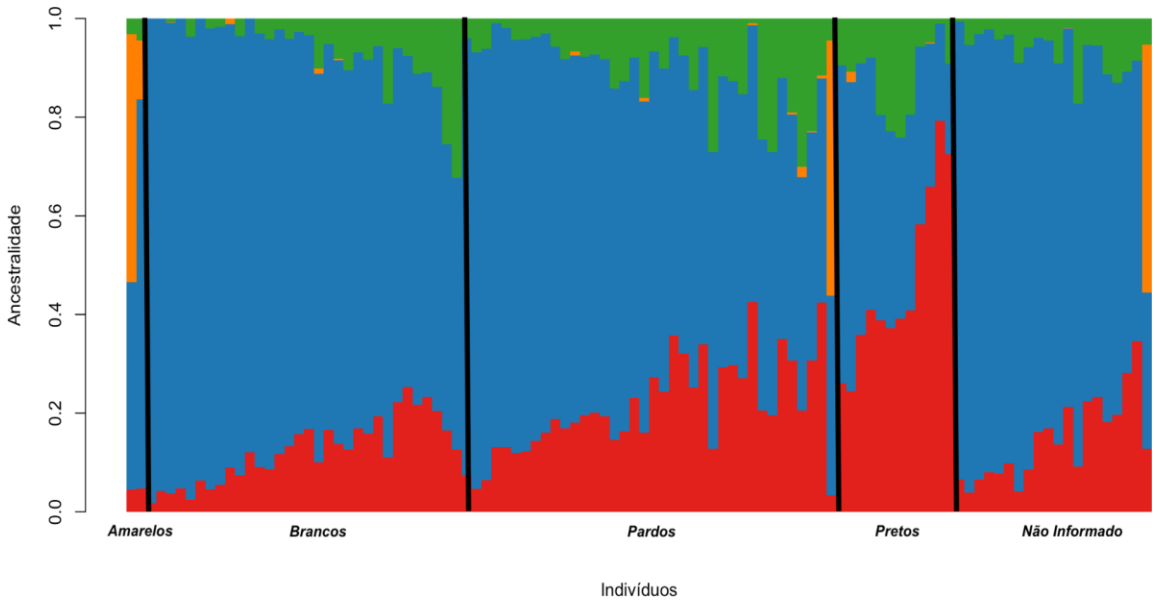


Figura 06. Ancestralidade genética média estimada a partir de marcadores genéticos autossômicos. No eixo X, cada coluna representa um indivíduo da amostra do Distrito Federal. O eixo Y representa a proporção de ancestralidade genética autossômica inferida. As cores vermelha, azul, laranja e verde representam as proporções para ancestralidade africana, europeia, leste asiática e nativo americana respectivamente. As barras pretas verticais separam os grupos de acordo com a autodeclaração dos indivíduos nas categorias do IBGE ("Amarelos", "Brancos", "Pardos", "Pretos" e "Indígenas"). Na amostra analisada nenhum indivíduo se autodeclarou indígena.

Na tabela 02 trazemos o perfil de autodeclaração dos participantes do DF, com base nas categorias nos grupos pré-estabelecidos pelo IBGE e a comparação com os dados do CODEPLAN (2018) e IBGE (2022). É possível observar que quase 10% dos participantes do nosso estudo não se autodeclararam. Nas demais categorias, as proporções são muito parecidas com aquelas reportadas pelo CODEPLAN e IBGE.

Tabela 02. Autodeclaração, de acordo com as categorias de "cor/raça" utilizadas pelo IBGE, da amostra analisada da população do DF. Comparativo da autodeclaração do grupo amostral do presente estudo com dados da CODEPLAN (2018) e IBGE (2022).

Autodeclaração	N	Presente estudo (%)	Presente estudo* (%)	Codeplan 2018 (%)	IBGE DF 2022 (%)
PARDO	37	39,36	44,6	47,4	47,0
BRANCO	32	34,04	38,6	41,1	43,0
PRETO	12	12,75	14,5	10,0	9,1
AMARELO	2	2,12	2,4	1,2	0,7
INDÍGENA	0	0	0	0,3	0,4
Não informado	9	9,57	-	-	-

* dados desconsiderando os 9 individuos que não se autodeclararam

A autodeclaração é como os indivíduos se percebem. Segundo a antropóloga e professora titular do Departamento de Ciências Sociais da Universidade Federal de Santa Maria (UFSM) Maria Catarina Chitolina Zanini, em uma comunicação pessoal, reflete:

> *"a auto identificação é um processo complexo, contínuo, e que, embora esteja centrado no indivíduo, é coletivo. Os pertencimentos étnicos podem estar baseados no compartilhamento da crença, numa origem comum, na autoidentificação e identificação por outros como parte desses coletivos, na escolha de sinais de pertencimento e também nas fronteiras de diferenciação que se estabelecem nos processos interativos"*

A percepção do indivíduo sobre ele mesmo, portanto, não está vinculada ao conhecimento genético da ancestralidade individual ou coletiva, uma vez que o acesso a esse tipo de informação ainda é limitado e distante da população geral. Apesar disso, alguns estudos têm demonstrado que não há uma dissociação completa entre ancestralidade genética e autodeclaração em grupos étnicos. Kedhy et al (2015), ao analisar 5.871 brasileiros, mostrou que quanto menor as proporções de miscigenação no genoma de um indivíduo, maior é a chance de que ele se identifique com a etnia preponderante de sua origem. Contudo, quando o nível de miscigenação é muito alto,

essa relação fica menos definida. Por isso, o uso de autodeclaração como uma aproximação de ancestralidade genética em populações com alto grau de miscigenação, como é o caso da brasileira e do DF, em especial em estudos que envolvem genética médica, não é recomendado (Pena *et al.,* 2011).

7.2 Perfil médio da ancestralidade genética no DF e demais regiões do Brasil

Seguindo para a análise populacional, observa-se que a representatividade do componente ancestral europeu na população do DF revelou uma proporção média de 69,95% (±18%), seguido por um valor médio do componente africano de 19,8% (±14,4%). A contribuição da parental nativo americana está representada por um percentual médio de 8,57% (±7,2%) e em menor expressividade média dos quatro componentes parentais avaliados, está o leste asiático com 1,68%(±8,6%) (Material Suplementar - arquivo digital capítulo 2).

Em estudo anterior do nosso grupo, Godinho (2008), avaliando marcadores do tipo microssatélites e AIMs, observou uma contribuição europeia de cerca de 60%, africana entre 20% e 27% e nativo americana ao redor de 15% numa amostra do DF. Nesse estudo inicial, o componente ancestral do leste asiático não foi avaliado. Em 2019, Souza e colaboradores realizaram uma metanálise, concatenando o resultado de 51 estudos no Brasil sobre ancestralidade genética e com base em 960 amostras de Brasília e Taguatinga eles determinaram a ancestralidade média do DF em 63,0% (±8,67%) de ancestralidade europeia, 24,1% (±6,49%) africana e 12,9% (±2,7%) de nativo americana. Esse estudo também não avaliou a ancestralidade vinda do leste asiático.

Para verificar se as inferências obtidas para o DF diferem ou não das demais regiões do país, realizamos o teste-z para testar a igualdade de proporções. Para tanto, usamos os dados disponibilizados por Souza et al (2019) para as demais regiões do Brasil. Partindo da hipótese nula que duas amostras têm igual proporção de ancestralidade, para cada componente ancestral em separado (Africano, Europeu e Nativo Americano), fizemos comparações par a par entre a amostra do DF com cada uma das regiões do país (tabela 03).

O teste Z revelou diferenças significativas entre a proporção de ancestralidade africana do DF e a região Sul do Brasil (p-valor =0,006), entre a proporção europeia das regiões Norte (p-valor = 0,001), Nordeste (p-valor = 0,004) e Sul (p-valor = 0,015) e

entre a proporção Nativo Americana da região Norte (p-valor = $3,96^{-05}$). Para as regiões Centro-Oeste e Sudeste não foi observada nenhuma diferença significativa nas proporções de ancestralidade genética. É interessante observar também que esses resultados mostram que esta amostra do DF não apresenta diferenças no perfil de ancestralidade genética em relação ao Brasil como um todo (cinco regiões, sem amostras do DF) (tabela 03).

Tabela 03. Teste Z para a comparação das proporções de ancestralidade genética entre o DF e as demais regiões geográficas do Brasil. Os dados do DF foram gerados pelo presente estudo e os demais fazem parte do estudo de Souza et al. 2019.

Região (N)	**Proporção % AFR (z; p valor)**	**Proporção % EUR (z; p valor)**	**Proporção % NAM (z; p valor)**
DF (104)	19,8	69,9	8,6
Centro Oeste (1053)	23,4 (0,50; 0,478)	63,4 (1,46; 0,22)	13,2 (1,40; 0,23)
Norte (3092)	19,9 (0; 1)	**52,9 (11,01; 0,001)**	**27,2 (16,89; $3,96^{-05}$)**
Nordeste (7891)	28,8 (3,63; 0,056)	**55,3 (8,27; 0,004)**	15,8 (3,49; 0,06)
Sudeste (8791)	24,2 (0,85; 0,35)	66,6 (0,36; 0,54)	9,4 (0,01; 0,91)
Sul (11078)	**10,9 (7,46; 0,006)**	**80,0 (5,93; 0,015)**	8,9 (0; 1)
Brasil (31851)	22,6 (0,31; 0,57)	62,4 (2,17; 0,14)	14,7 (2,61; 0,10)

É importante ressaltar que comparamos resultados de estudos cujas inferências de ancestralidade genética foram obtidas a partir de diferentes tipos de marcadores (STRs, AIMs, SNPArray) e diferentes populações parentais de referência, portanto é possível que hajam ruídos introduzidos por esses fatores; Não obstante, o Teste Z comumente é influenciado pelo desvio padrão. No capítulo 1 da tese observamos que há diferenças significativas entre as estimativas de ancestralidade obtidas a partir de alguns painéis de AIMs e de SNPArray. Souza et al (2019) também observam que há discrepâncias significativas nas inferências de ancestralidade genética de acordo com o tipo de marcadores utilizado. Outra diferença entre o nosso estudo com os demais é que estamos usando o modelo de miscigenação tetra-híbrido (AFR, EUR, NAM e EAS), enquanto os demais estudos usaram o modelo tri-híbrido (AFR, EUR, NAM). Nossa escolha pelo modelo tetra-híbrido se deu pelas informações descritas no questionário demográfico sobre a presença de ancestrais de origem

japonesa na nossa amostra do DF. Portanto, os resultados devem ser interpretados com ressalvas.

7.3 Perfil da ancestralidade genética no DF e modelo demográfico de migração

A partir das informações obtidas no questionário demográfico foi possível obter informações a respeito da região do Brasil onde nasceram os pais de cada participante. Com essa informação buscamos verificar como os padrões de migração influenciam na composição da ancestralidade genética na população do DF. Nosso teste consistiu na criação de um modelo demográfico simples no qual usamos as proporções de ancestralidade genética determinadas para cada região geográfica do Brasil, de acordo com Souza et al (2019), e multiplicamos pela proporção de imigrantes de cada região do Brasil informada na nossa amostra (Tabela 04). As estimativas inferidas e observadas foram comparadas a partir de um teste de x^2 (qui-quadrado) (Tabela 05). O teste não revelou diferenças significativas entre o observado e o esperado (p-valor = 0,1353), sugerindo que esse modelo demográfico de migrações é condizente com as proporções de ancestralidade observadas no nosso estudo.

Tabela 04. Modelo demográfico simplificado de miscigenação do DF. A porcentagem de imigrantes foi obtida a partir de informações sobre onde nasceram os pais dos participantes da amostra do DF. A proporção de ancestralidade para cada região do Brasil foi obtida do trabalho de Souza et al (2019) e consiste numa média ponderada pelo tamanho amostral de cada região. A proporção de contribuição para o DF é a multiplicação entre a proporção de imigrantes pela proporção de ancestralidade média em cada região.

Região	**Imigrante (%)**	**Ancestralidade**	**Souza et al (2019) (%)**	**Contribuição para o DF (%)**
Sudeste	43,64	AFR	24,2	10,56
		EUR	66,6	29,06
		NAM	9,4	4,10
Nordeste	38,22	AFR	28,8	11,0
		EUR	55,3	21,13
		NAM	15,8	6,04
Centro-Oeste	9,39	AFR	23,4	2,2
		EUR	63,4	5,95
		NAM	13,2	1,24
Norte	5,52	AFR	19,9	1,98
		EUR	52,9	2,92
		NAM	27,3	1,50
Sul	1,65	AFR	10,9	0,18
		EUR	80,0	1,32
		NAM	8,9	0,15
TOTAL		AFR	22,6	25,02
		EUR	62,4	60,38
		NAM	14,7	13,03

Tabela 05. Teste do modelo demográfico migratório para o DF. A ancestralidade observada refere-se aquela inferida para a amostra do DF e a esperada aquela estimada a partir do modelo demográfico de miscigenação (Tabela 04).

Ancestralidade	Observada	Esperada	x^2
AFR	19,8	25,02	1,08
EUR	66,95	60,38	0,71
NAM	8,57	13,02	1,52
TOTAL			2,23*

* p-valor = 0,1353

7.4 Análise Genética Utilizando marcadores situados no Cromossomo X

O cromossomo X tem um modelo de herança genética distinto entre homens e mulheres. Os homens recebem o cromossomo X de suas mães, enquanto que as mulheres herdam tanto da mãe quanto do pai. Isso faz com que a dinâmica dos processos evolutivos e de recombinação genética seja distinta nesse cromossomo em relação aos autossômicos. Nesse sentido, se pensarmos num modelo demográfico simples onde o tamanho efetivo de homens e mulheres seja igual, seria esperado que o tamanho efetivo para o cromossomo X fosse ¾, por que há três cromossomos X para cada quatro cromossomos autossômicos na população. Desvios dos ¾ no tamanho efetivo populacional pode indicar diferentes histórias na contribuição de homens e mulheres nessas populações, fenômeno esse conhecido como casamento direcional. Num processo de miscigenação com contribuição igual de homens e mulheres da mesma ancestralidade, espera-se que a razão entre a média de ancestralidade no cromossomo X pela média de ancestralidade autossômica seja igual a 1. Um processo de miscigenação com maior contribuição de homens de uma determinada ancestralidade leva a uma redução na média dessa ancestralidade no cromossomo X em relação aos autossomos. Por outro lado, uma maior contribuição de mulheres de uma dada ancestralidade leva a um aumento na média dessa ancestralidade no cromossomo X em relação aos autossomos.

Para verificar possíveis perfil de casamentos direcionais na nossa amostra, inferimos a ancestralidade genética para o cromossomo X e comparamos com a dos demais autossomos (tabela 06). Para averiguar se há diferenças e heterogeneidade na distribuição dos componentes ancestrais entre os cromossomos, aplicamos o método CAnD (figura 07), esse método leva em consideração a diferença entre a média de um dado componente ancestral entre o cromossomo de interesse e a média dos demais cromossomos entre os indivíduos da amostra e entre os cromossomos do indivíduo. Esse método, testa a hipótese nula que a diferença entre as médias das proporções de ancestralidade entre o cromossomo de interesse e os demais é igual a zero.

Observamos diferença significativa na proporção do componente genético Nativo Americano com relação ao cromossomo X nas amostras do Distrito Federal em comparação aos cromossomos autossômicos. Essa diferença corresponde a um aumento médio deste componente ancestral (0,086 nos autossomos para 0,138 no cromossomo X; p-valor = 0,001), indicando maior contribuição de mulher de origem

Nativo Americana. Também observa-se uma redução na média de ancestralidade Europeia no cromossomo X em relação aos autossômicos (63,15% e 69,95%; p-valor = 0,006), indicando maior contribuição de homens de origem europeia. Para os componentes ancestrais africanos e leste asiático, não observa-se diferenças significativas, indicando que a contribuição de homens e mulheres foi equivalente.

Até o momento, poucos estudos averiguaram a ancestralidade genética com base em marcadores genéticos situados no cromossomo X de amostras brasileiras. Contudo, padrões semelhantes ao observado no nosso estudo foram reportados por Ongaro et al (2019) para populações de Salvador (BA), Bambuí (MG) e Pelotas (RS), com desvios significativos da ancestralidade do X em relação aos autossômicos para as proporções de ancestralidade europeia e nativo americana, mas não para a africana.

Tabela 06. Inferência média em porcentagem da ancestralidade genética estimada para cromossomos autossômicos, sexuais (X e Y) e DNA mitocondrial (mtDNA).

Inferência Ancestralidade	**Africana** (%)	**Europeia** (%)	**Nativa Americana** (%)	**Leste Asiática** (%)
Autossômicos	19,2	69,95	8,6	1,7
Cromossomo X	21,9	63,15	13,8	2,1
Cromossomo Y	9,3	90,7	-	-
mtDNA	37,5	31,7	28,8	1,9

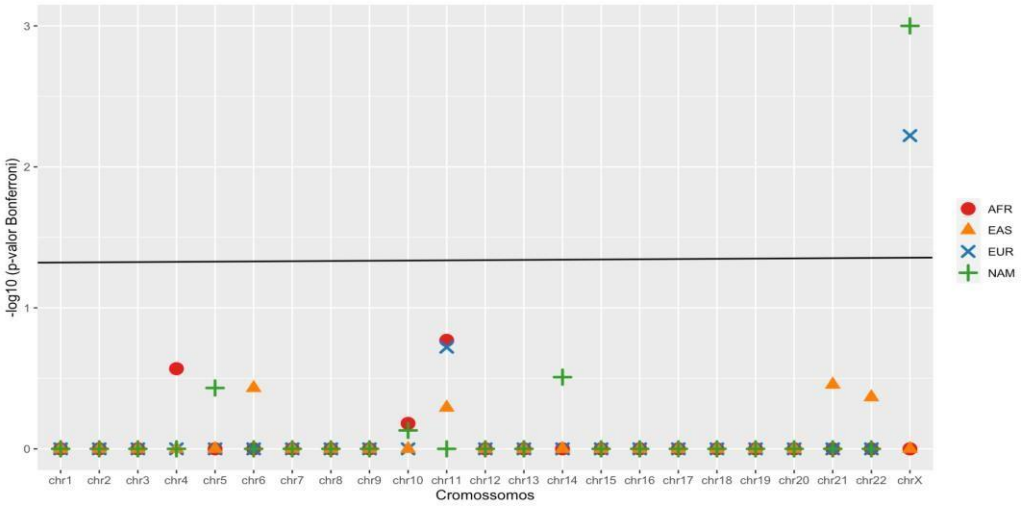


Figura 07. Heterogeneidade na contribuição ancestral entre os cromossomos. O eixo X representa os cromossomos autossômicos e o cromossomo X. O eixo Y representa o p-valor (em escala -log10, com correção de Bonferroni) calculado a partir do método CAnD para a detecção de heterogeneidade nas proporções de cada ancestralidade genética. A linha preta horizontal representa o limiar de significância. As cores vermelha, azul, laranja e verde representam os componentes ancestrais africano (AFR), europeu (EUR), leste asiático (EAS) e nativo americano (NAM) respectivamente.

7.5 Análise Genética Utilizando marcadores situados no cromossomo Y

O cromossomo Y apresenta herança uniparental paterna, ou seja, é transmitido apenas do pai para o filho. Durante décadas, o mapeamento das variações genéticas no cromossomo Y e como elas são distribuídas mundialmente são temas de diversos estudos (Jobling 2001). Hoje é reconhecido que muitas dessas variações genéticas são restritas ou encontradas com maior frequência em determinadas regiões geográficas. Com o uso desta informação, é possível traçar um mapa das migrações humanas e de haplogrupos geográfico-específicos. O *Y Chromosome Consortium* (https://isogg.org/wiki/Y_Chromosome_Consortium) definiu regras para agrupar as mutações de acordo com a relação entre elas, estabelecendo dessa forma a árvore filogenética das variantes genéticas do cromossomo Y. Por convenção, os haplogrupos do cromossomo Y, que corresponde aos principais ramos da árvore filogenética, são agrupados por letras maiúsculas que vão de A até R (YCC 2002). A figura 08 mostra a distribuição global dos principais haplogrupos do cromossomo Y.

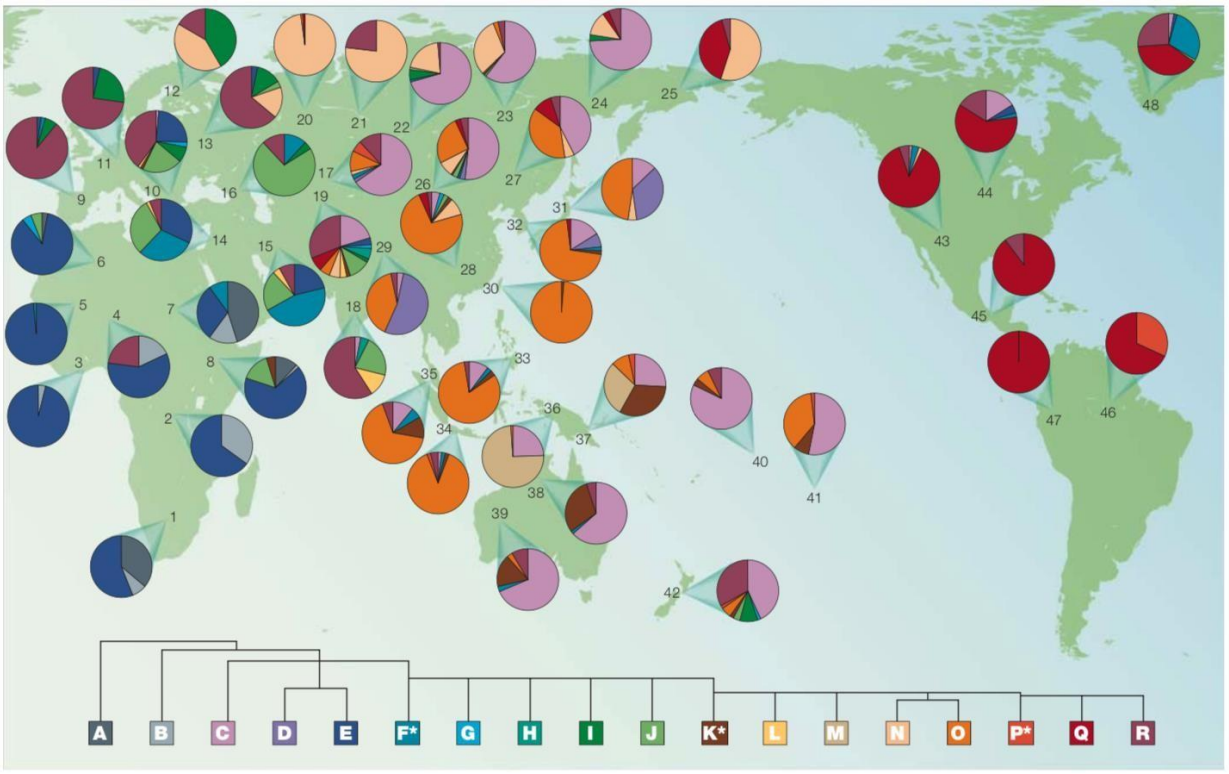


Figura 08 Distribuição global dos haplogrupos do cromossomo Y. Cada círculo representa uma amostra populacional com a frequência dos 18 principais haplogrupos do Y : A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P , Q, R. Os números representam as seguintes populações: 1, !Kung; 2, Pigmeus Biaka; 3, Bamileke; 4, Fali; 5, Senegalês; 6, Berberes; 7, Etíopes; 8, Sudaneses; 9, Bascos; 10, Gregos; 11, Polonês; 12, Saami; 13, Russos; 14, Libanês; 15, Iranianos; 16, Kazbegi (Geórgia); 17, Kazaks; 18, Punjabis; 19, Uzbeques; 20, Floresta Nentsi; 21, Khants; 22, Evenks Orientais; 23, Buryats; 24, Evens; 25, Esquimós; 26, Mongóis; 27, Evenks; 28, Han do Norte; 29, Tibetanos; 30, Taiwanês; 31, Japonês; 32, Coreanos; 33, Filipinos; 34, Javanês; 35, Malaios; 36, Nova Guiné Ocidental (terras altas); 37, Papua Nova Guiné (costa); 38, Australianos (Arnhem); 39, Australianos (deserto arenoso); 40, Cook Islanders; 41, Taitianos; 42, Maori; 43, Navajos; 44 Cheyenne; 45, Mixtec; 46, Makiritare; 47 Cayapa; 48 Inuit Groelândia. Fonte: Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003 Aug;4(8):598-612. doi: 10.1038/nrg1124.

Os haplogrupos do cromossomo Y da nossa população de estudo foram analisados, utilizando o programa SNAPPY (Serverson *et al.,* 2018), com base em 508 SNPs sobrepostos com o ISOGG Y Tree. A partir da distribuição mais frequente dos haplogrupos nas populações parentais, foi estimada sua contribuição na população masculina do Distrito Federal. Observou-se que 90,7% dos haplogrupos são de origem europeia e 9,3% de origem africana. Esses resultados são similares aos reportados no estudo prévio do nosso grupo sobre o DF (90% EUR, 8,5% AFR e 1,5% NAM) (Barcelos, 2006).

Chama a atenção o valor expressivo de haplogrupos com origem europeia, porém cabe revisitar a história de formação da população do DF, que remonta a grande maioria de imigrantes vindo do sudeste e nordeste. Estes, por sua vez, já vinham de regiões cenários de um processo de colonização com relacionamentos completamente assimétricos onde homens africanos e indígenas pouco ou nulamente tiveram a chance de transferir seus alelos para as próximas gerações (Pena *et al.,* 2011).

Resque et al (2016) realizou uma investigação sobre haplogrupos do cromossomo Y no Brasil, no intuito de compreender a estrutura genética de linhagens masculinas no país, analisando um conjunto de 41 Y-SNPs em 1217 homens não aparentados das cinco regiões geopolíticas brasileiras. O estudo revelou que em média a população brasileira apresenta 88% de ancestralidade europeia, 9% africana e 3% nativo americana. Algumas variações destas proporções foram observadas entre as regiões do país. A amostra da região Norte apresentou a maior proporção de ancestralidade nativo americana (8,1%), enquanto a contribuição africana mais acentuada pôde ser observada na população do Nordeste (15,1%). As amostras do Centro-Oeste e do Sul apresentaram as maiores contribuições europeias (95,7% e 93,6%, respectivamente). A região Sudeste apresentou importantes contribuições europeias (86,1%) e africanas (12,0%). A partir da realização do teste z, verificamos que não há diferenças significativas nas proporções de ancestralidade do cromossomo Y entre o DF e as demais regiões do Brasil (tabela 07). Quando comparamos o perfil das linhagens patrilineares do DF com outras regiões brasileiras (Figura 09), o que se percebeu foi o DF seguindo o padrão originário da proporção de ancestralidade das duas regiões que doaram o maior número de imigrantes (Sudeste e Nordeste), como discutido acima.

Tabela 07. Comparação das proporções de ancestralidade genética inferida para o cromossomo Y entre o DF e as demais regiões geográficas do Brasil utilizando teste Z. Os dados do DF foram gerados pelo presente estudo e os demais fazem parte do estudo de Resque et al (2016).

Região (N)	Proporção % AFR (z; p valor)	Proporção % EUR (z; p valor)	Proporção % NAM (z; p valor)
Distrito Federal (43)	9,3	90,7	0
Centro Oeste (135)	3,1 (1,62; 0,20)	95,4 (0,60; 0,44)	1,5 (0; 1)
Norte (272)	6,7 (0,08; 0,76)	85,2 (0,53; 0,46)	8,1 (2,88; 0,09)
Nordeste (243)	14,1 (0,37; 0,54)	84,7 (0,63; 0,42)	1,2 (0; 1)
Sudeste (330)	11,1 (0,01; 0,92)	87,4 (0,14; 0,70)	1,5 (0; 1)
Sul (237)	3,9 (1,31; 0,25)	94 (0,22; 0,63)	2,1 (0,1; 0,74)
Brasil (1217)	8,7 (0; 1)	88,2 (0,07; 0,79)	3,1 (0,51; 0,47)

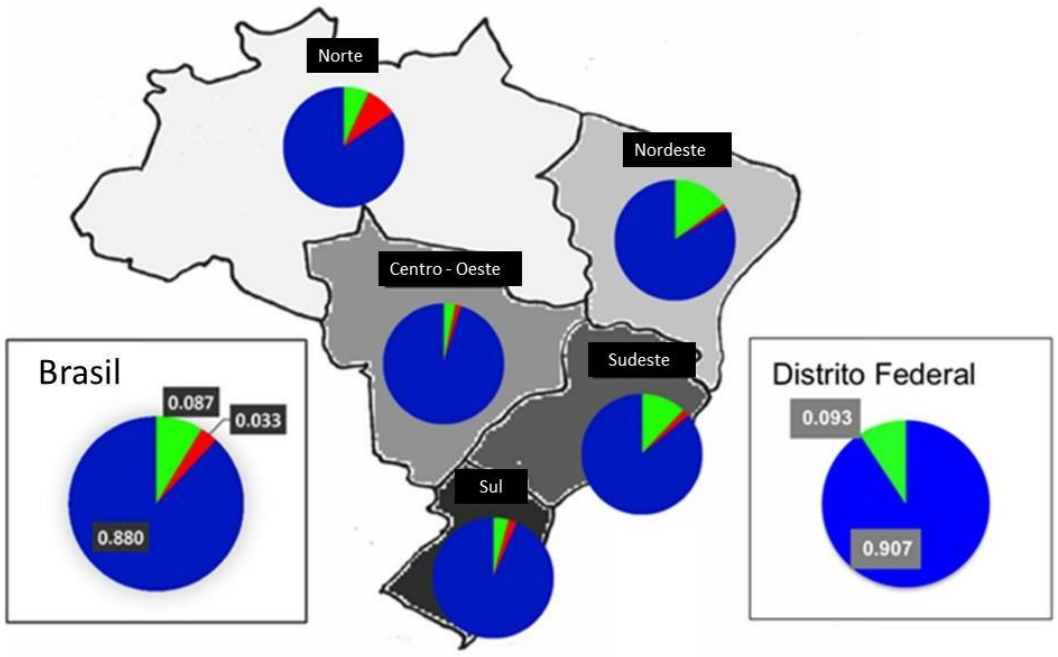


Figura 09. Estimativas de ancestralidade obtidas a partir da análise de marcadores genéticos situados no cromossomo Y. A figura representa as proporções médias de ancestralidade genética inferida para os haplogrupos do cromossomo Y. As inferências para cada uma das cinco regiões geográficas e para o Brasil são do estudo de Resque et al. (2016). A inferência para o Distrito Federal foi realizada no presente estudo. As ancestralidades europeia, africana e nativo americana são representadas pelas cores azul, verde e vermelho respectivamente. Fonte: Resque et al. (2016) (Modificada).

A nível de comparação, vale ressaltar que mesmo nas populações afro-descendentes do nordeste e sudeste, as proporções de linhagens patrilineares de origens europeias estão na orden de 34% e 63%, respectivamente (Abe-Sandes *et al*. 2004; Abe-Sandes *et al*. 2004, Guerreiro-Junior *et al*. 2009, Kimura *et al*., 2017). Isso reforça a força da presença masculina de origem europeia no processo de colonização e seu reflexo nas populações brasileiras atuais.

Na tabela 08, está apresentada a proporção de cada haplogrupo observado na população do DF. O haplogrupo mais representativo foi o R1b1a1b1a de origem europeia, com 55,81% de representatividade. O haplogrupo R1b1a1b é a linhagem europeia mais comum, cerca de 60% dos portugueses e espanhóis apresentam este haplogrupo, ele também é comum em países como Alemanha (50%) (Myres *et al.,* 2011) e Itália (47%) (Grugni *et al.,* 2018). Seguindo com maior valor expressivo na amostra do DF está o G2a (9,30%), também um haplogrupo de origem europeia, mais especificamente oeste e região do caucaso (Broushaki *et al.,* 2016). O haplogrupo G2a foi observado com frequência de 18% na Armênia e 11% na Turquia (Cinnioglu *et al.,* 2004), além de ser comum no oeste da Áustria com frequência variando entre 11% e 40% (Berger *et al.,* 2013). Por fim, observa-se os haplogrupos E1b1a, J2a1 e J2b todos com 6,98% de representatividade. Com relação aos haplogrupos J2a1 e J2b estes ocorrem com maior frequência em países do Oriente Médio e sul da Europa. Esses dois haplogrupos também são frequentes em descendentes de judeus, com frequência de 40% nos grupos sefaraditas (Semino *et al.*, 2004). A literatura diz que a presença do haplogrupo J em nossa população se deu pela migração europeia, m especial pelos portugueses e italianos (Rosser *et al.,* 2000; Semino *et al.,* 2004). E1b1a é um haplogrupo com origem Africana, comum nos grupos subsaarianos. É interessante também reportar a observação do haplogrupo E1a (2,33%), o qual é restrito aos grupos não Bantu, ocorrendo na região de Senegal e Burkina Faso (de Filippo *et al.,* 2011).

Tabela 08. Haplogrupos identificados na população brasileira. Lista de haplogrupos identificados na população brasileira. Ocorrência geográfica mais frequente fora do Brasil (referida como "origem"). Porcentagem observada no Distrito Federal (DF), Brasil (BR), regiões Norte (N), Nordeste (NE), Centro-Oeste (CO), Sudeste (SE) e Sul (S). O dado do DF corresponde ao presente estudo, os dados do Brasil e das cinco regiões geográficas são do estudo de Resque et al (2016). Em parênteses está o tamanho amostral (n).

Haplogrupo	Origem	DF (43)	BR (1217)	N (272)	NE (243)	CO (135)	SE (330)	S (237)
E1a	Africana	2,32%	0,2%	0,6%	0,6%	0	0	0
E1b1a	Africana	6,98%	8,5	6,1%	12,3%	3,1%	10,6%	3,9%
E1b1b-M35	Africana	0	0,4%	0	1,2%	0	0,5	0
E1b1b	Europeia	0	10,5%	7,8%	9,9%	10,3%	12,8 %	10,4 %
G	Europeia	9,3%	5,1%	5,4%	7,4%	3,1%	2,7%	6,9%
I	Europeia	6,98%	8,9%	10%	11,5%	6,1%	8,2%	7,7%
J	Europeia	13,96%	10,1%	9,6%	10,3%	16%	9,4%	8,2%
K	Europeia	2,3%	2,2%	2,6%	1,2%	3,7%	2,1%	2,1%
Q	Americana	0	3,1%	8,1%	1,2%	1,5%	1,8%	2,1%
R1b1a	Europeia	58,1%	51,6%	50%	44,5%	56,7%	51,9%	58,7%

7.6 Análise Genética Utilizando marcadores situados no DNA mitocondrial

A filogenia do mtDNA é classificada em haplogrupos principais caracterizados pelos Polimorfismos de Nucleotídeo Único (SNP - *Single Nucleotide Polymorphism*) formados durante a história evolutiva humana (Stewart & Chinnery, 2015; Gomez *et al.,* 2014). Populações próximas e/ou com origem comuns apresentam haplótipos que, da mesma forma, compartilham polimorfismos comuns. Ao conjunto de haplótipos: nos referimos como haplogrupo (Torroni *et al.,* 1993). A figura 10 apresenta um mapa representativo das migrações populacionais e distribuição geográfica dos principais haplogrupos de mtDNA.

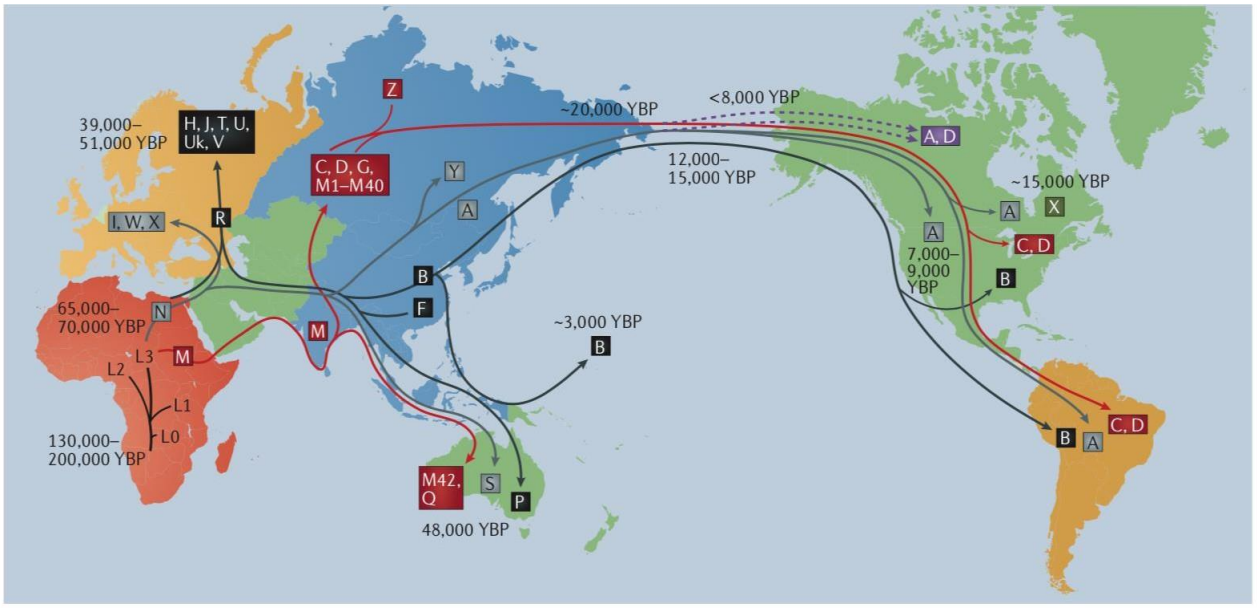


Figura 10. Mapa mundial representativo das migrações populacionais e distribuição geográfica dos principais haplogrupos de mtDNA. Como o DNA mitocondrial (mtDNA) é herdado uniparentalmente, ele sofre recombinação insignificante no nível da população, e as mutações adquiridas ao longo do tempo subdividiram a população humana em vários haplogrupos discretos. Os principais haplogrupos surgiram 40.000–150.000 anos antes do presente (YBP) e definiram diferentes populações humanas à medida que migravam da África e povoavam o globo. A raiz africana foi a fonte de quatro linhagens específicas para a África subsaariana: L0, L1, L2 e L3 (130.000–200.000 YBP). Mais dois haplogrupos, M e N, surgiram do haplogrupo africano L3 65.000–70.000 YBP para povoar o resto do mundo. Com a migração dos humanos, o haplogrupo N foi direcionado para a Eurásia e as linhagens do haplogrupo M se mudaram para a Ásia, dando origem aos haplogrupos A, B, C, D, G e F. Na Europa, o haplogrupo N deu origem ao haplogrupo R, que é a raiz do os haplogrupos europeus H, J, T, U e V, que surgiram 39.000–51.000 YBP126. Os haplogrupos S, P e Q são encontrados na Australásia e foram formados ~48.000 YBP, e os haplogrupos A, B, C e D surgiram <20.000 YBP e povoaram o Leste Asiático e as Américas. Fonte: Stewart JB, Chinnery PF. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet. 2015 Sep;16(9):530-42. doi: 10.1038/nrg3966.

A partir da distribuição dos haplogrupos de acordo com a origem geográfica investigamos também a contribuição parental da população feminina do Distrito Federal. Os haplogrupos do mtDNA foram inferidos com o software HaploGrep3.0. A tabela 09 apresenta a proporção de cada haplogrupo na amostra de Brasília. Em maior representatividade observa-se o haplogrupo H2 (39,42%). O haplogrupo H abrange praticamente toda a variabilidade de mtDNA da Europa (Pereira *et al.,* 2000). Em seguida nota-se o L2 (20,19%), é uma linhagem de DNA mitocondrial característica da África Subsaariana (Salas *et al.,* 2002); os haplogrupo A (10,6%) e B (11,53%) representam matrilinhagens nativo americanas (Gonçalves *et al.,* 2010).

Identificamos cerca de 37,5% dos haplogrupos de origem africana, 31,7% europeia, 28,8% nativo americana e 1,9% do leste asiático (Tabela 09). O reflexo na

população do DF não foi diferente, as análises do cromossomo Y e do mtDNA mostraram a consequência do fenômeno demográfico de acasalamento direcional, de homens com expressiva ascendência européia e mulheres com ascendência africana, nativa americana e européia. Ao comparar nossos resultados com a literatura observamos no Centro-Oeste: 24% de europeus, 31% de africanos e 44% de nativo-americanos (Joerin-Luque *et al.,* 2022). Na região Norte a proporção é de 54% dos haplogrupos de origem nativo americana, 31% europeia e 15% africana. Na região Nordeste: 44% dos haplogrupos de origem africana, 34% origem europeia e 22% nativo americana. No Sudeste 34% africana, 33% nativo americana e 31% europeia. E na região Sul, a matrilinhagem de origem europeia é representada por 66% europeia enquanto a africana e nativo americana estão representadas por 22% e 12% respectivamente (Alves-Silva *et al.,* 2000). Ao compararmos as proporções das ancestralidade observadas no DF com outras regiões do Brasil, observamos diferenças significativas entre o componente africano com as regiões Norte (p-valor=0,009) e Sul (p-valor=0,001), entre o componente europeu com a região Sul e entre o componente nativo americano com as regiões Centro-Oeste (p-valor=0,008) e Norte (p-valor=0,005). Não foi obsevada diferenças significativas entre a média da população brasileira e o DF (Tabela 10).

Vale ressaltar que os dados gerados no presente estudo e os demais estudos apresentam diferenças metodológicas importantes. Enquanto estamos inferindo os haplogrupos a partir de 200 SNPs no mtDNA, o estudo de Alves-Silva et al (2000) determinou os haplogrupos pela avaliação da região hipervariável I do mtDNA e o estudo de Joerin-Luque et al (2022) por captura pelo sequenciamento de exomas. Portanto, os resultados devem ser interpretados considerando essa ressalva.

Tabela 09. Haplogrupos mitocondriais identificados no presente estudo. Lista de haplótipos do mtDNA observados na amostra do DF, proporção (%), ocorrência geográfica mais frequente fora do Brasil (referida como origem geográfica) e a proporção da origem geográfica.

Haplogrupo	Proporção (%)	Ocorrência Geográfica mais frequente	Proporção de origem geográfica (%)
A	12,5		
B	14,4	NAM	28,8
C	1,92	EAS	1,92
D	1,92		
L	37,5	AFR	37,5
H	24,1		
J	4,8		
U	1,9	EUR	31,7
V	0,9		

Tabela 10. Teste Z para a comparação das proporções de ancestralidade genética inferida para o mtDNA entre o DF e as demais regiões geográficas do Brasil. Os dados do DF foram gerados pelo presente estudo e os demais fazem parte do estudo de Alves-Silva et al (2000) e Joerin-Luque et al (2022).

Região (N)	Proporção % AFR (z; p valor)	Proporção % EUR (z; p valor)	Proporção % NAM (z; p valor)
DF (104)	37,5	31,7	28,8
Centro Oeste (323)	31 (1,23; 0,26)	24 (2,05; 0,15)	**44 (6,93; 0,008)**
Norte (48)	**15 (6,83; 0,009)**	31 ($2,26^{-30}$; 1)	**54 (7,92; 0,005)**
Nordeste (50)	44 (0,35; 0,55)	34 (0,01; 0,92)	22 (0,49; 0,48)
Sudeste (99)	34 (0,14; 0,71)	31 ($2,99^{-30}$; 1)	33 (0,24; 0,62)
Sul (60)	**12 (10,99; 0,001)**	**66 (16,78 ; $4,19^{-05}$)**	22 (0,59; 0,44)
Brasil (580)	28 (3,38; 0,07)	39 (1,69; 0,19)	33 (0,53; 0,46)

7.7 Análise de percepção: ancestralidade genética e origem dos avós

7.7.1 Ancestralidade genética no mtDNA e a origem da avó materna

Com base nas respostas do questionário, comparamos a origem informada da avó materna com o resultado de ancestralidade do mtDNA (figura 11 e 12). Foi possível observar: (i) 40,38% dos participantes não informaram ou desconhecem a origem ancestral da avó materna; (ii) dentre os que dizem conhecer (59,62%), houve correspondência entre a origem informada e a inferida pelo mtDNA em 41,93% dos casos. (iii) dentre aqueles com mtDNA de ancestralidade Africana: há correspondência de 12,5% entre a origem informada e a inferida; (iv) dentre aqueles com mtDNA de ancestralidade Europeia há correspondência de 93,3% entre a origem informada e a inferida; (v) dentre aqueles com mtDNA de ancestralidade Asiática: há correspondência de 100% entre a origem informada e a inferida; (vi) dentre aqueles com mtDNA de ancestralidade Nativo Americana: há correspondência de 36,3% entre a origem informada e a inferida (Tabela 11).

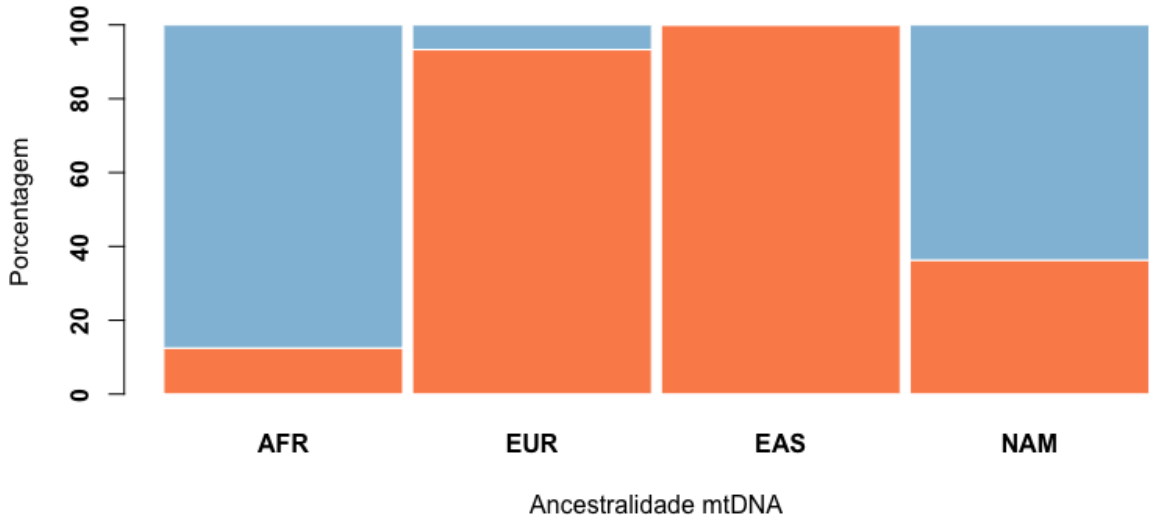


Figura 11. Percepção sobre a origem da Avó materna e concordância com a ancestralidade genética no mtDNA. O eixo X representa a ancestralidade inferida para o mtDNA. O eixo Y representa a porcentagem de concordância ou divergência entre origem das avós maternas informadas pelos participantes e a inferida no mtDNA. A cor laranja representa as concordâncias e a cor azul as divergências observadas.

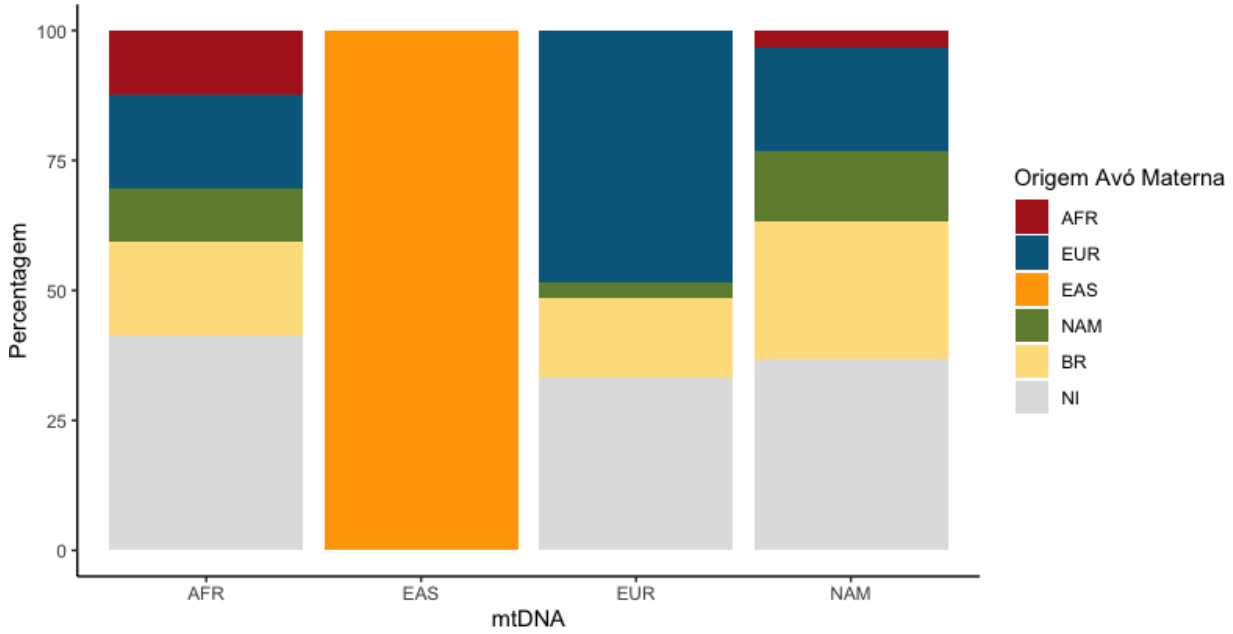


Figura 12. Percepção sobre a origem da Avó materna e a ancestralidade genética no mtDNA. O eixo X representa a ancestralidade inferida para o mtDNA, o eixo Y a porcentagem de origem das avós maternas informadas pelos participantes.

Por fim, aplicamos o teste binomial exato para verificar a hipótese nula de haver correspondência acima de 95% entre a ancestralidade inferida no mtDNA e a origem da avó materna informada (Tabela 11). Observamos que há elevada correspondência para a ancestralidade europeira (93,3%; p-valor=0,54) e leste asiática (100%; p-valor=1). Por outro lado, tanto a ancestralidade africana, como nativo americana tiveram discrepâncias significativas (p-valor<$2{,}15^{-07}$) e correspondêcia inferior à 37% entre a infencia do mtDNA e a origem informada da avó materna.

Tabela 11. Teste binomial exato comparando a ancestralidade encontrada no mtDNA e a correspondência com a origem informada da avó materna.

Ancestralidade mtDNA	Correspondência	Divergência	Total	Probabilidade de sucesso	p-valor
AFR	1	7	8	0,125	$5{,}97^{-09}$
EUR	14	1	15	0,933	0,54
NAM	4	7	11	0,363	$2{,}15^{-07}$
EAS	2	0	2	1	1
TODAS (AFR+EUR+ EAS+NAM)	21	15	36	0,58	$6{,}21^{-11}$

7.7.2 Ancestralidade genética no cromossomo Y e a origem do avô paterno

Comparamos entre os homens, a origem ancestral informada para o avô paterno (figura 13 e 14) com o resultado de ancestralidade do cromossomo Y. Foi possível observar: (i) 44,19% dos participantes não informaram ou desconhecem a origem ancestral do avô paterno; (ii) dentre os que dizem conhecer (55,81%), houve correspondência entre a origem informada e a inferida para o cromossomo Y em 50% dos casos. (iii) dentre aqueles com cromossomo Y de ancestralidade Africana: não houve correspondência entre a origem informada e a inferida; (iv) dentre aqueles com cromossomo Y de ancestralidade Europeia há correspondência de 54,16% entre a origem informada e a inferida.

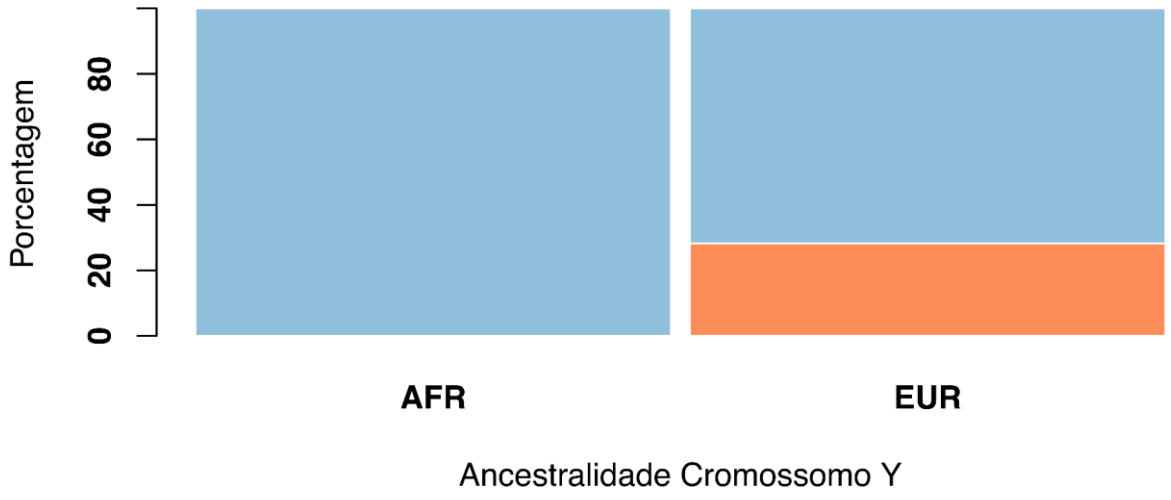


Figura 13. Percepção sobre a origem do Avô paterno e a ancestralidade genética no cromossomo Y. O eixo X representa a ancestralidade inferida para o cromossomo Y. O eixo Y representa a porcentagem de concordância ou divergência entre origem dos avôs paternos informadas pelos participantes e a inferida no cromossomo Y. A cor laranja representa as concordâncias e a cor azul as divergências observadas.

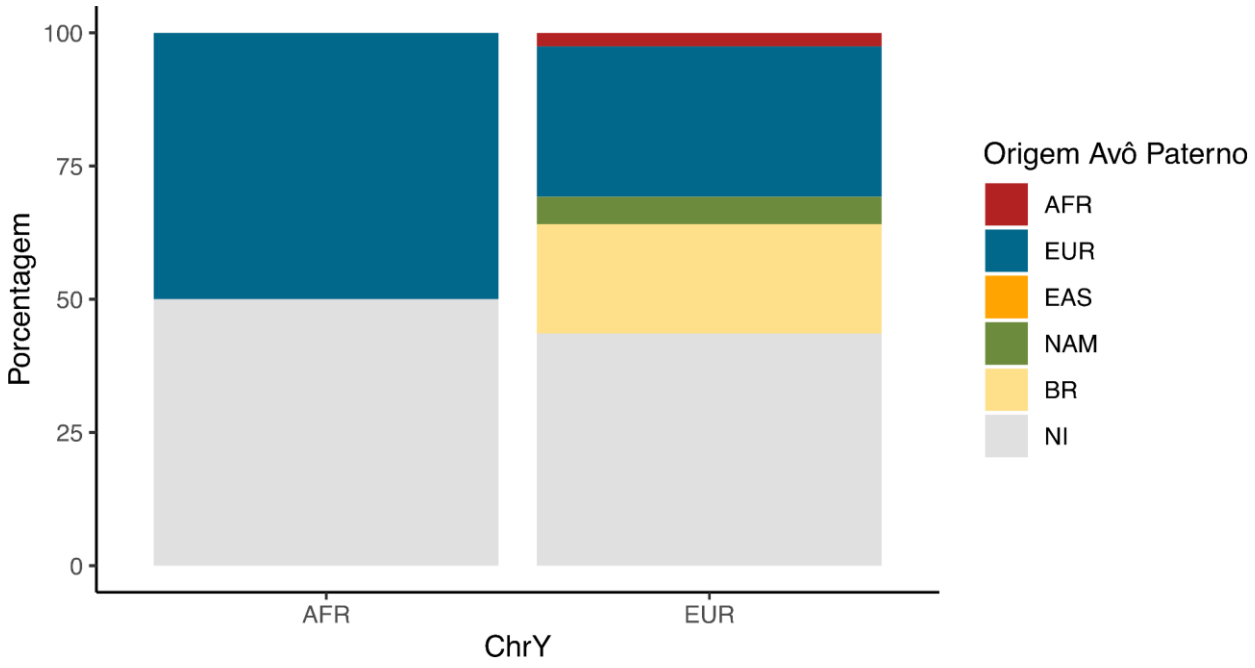


Figura 14. Percepção sobre a origem do Avô paterno e a concordância com a ancestralidade genética no cromossomo Y. O eixo X representa a ancestralidade inferida para o cromossomo Y. O eixo Y representa a porcentagem de origem dos avôs paternos informadas pelos participantes.

Aqui também aplicamos o teste binomial exato para verificar a hipótese nula de haver correspondência mínima de 95% entre a ancestralidade inferida no cromossomo Y e a origem do avô paterno informada (tabela 12). Observamos diferenças significativas para a percepção européia (p-valor=$6{,}63^{-09}$) e europeia+africana ($6{,}85^{-11}$). Nós não testamos o componente africano em separado devido ao baixo número amostral (apenas 4 amostras).

Tabela 12. Teste de x^2 comparando a ancestralidade observada no cromossomo Y e a correspondência com a origem informada do avô paterno.

Ancestralidade Cromossomo Y	Correspondência	Divergência	Total	Probabilidade de sucesso	p-valor
EUR	13	11	24	0,54	$6{,}63^{-09}$
AFR	0	2	2	0	0,0025
TODAS (AFR+EUR)	13	13	26	0,50	$6{,}85^{-11}$

Esse resultado mostra que por um lado a ancestralidade matriarcal europeia e asiática apresenta alta correspondência entre a ancestralidade inferida e a informada (acima de 90%). Por outro, a correspondência para a ancestralidade europeia é muito menor para a ancestralidade patriarcal no cromossomo Y (54%). Uma possível interpretação a essa discrepância é que os eventos de miscigenação mais recentes são mais facilmente distinguíveis na história genealógica dos participantes, enquanto eventos envolvendo miscigenação mais antiga são mais difíceis de acessar.

Ou seja, esse resultado reflete as muitas camadas da história da população brasileira: (i) componentes ancestrais como asiático e mulheres europeias fazem parte de um processo mais recente de miscigenação no Brasil. Dados históricos mostram que a migração nipônica no Brasil teve inicio em 1908 e o maior fluxo migratório de mulheres europeia a partir de 1870 (IBGE 500 anos). Assim, o pequeno número de gerações transcorridas até aqui pode retratar com maior fidelidade a informação genética e origem ancestral das avós; (ii) por outro lado, os componentes ancestrais matriarcais africanos e nativos americanos fazem parte do processo mais antigo de miscigenação. Sabemos que a partir de 1850 o tráfico de africanos escravizados para o Brasil é proíbido, diminuído a entrada do componente africano no país. Além disso, inferências históricas, indicam que grande parte da população nativa americana foi dizimada nos primeiros séculos de colonização europeia no Brasil, sugerindo que a maior parcela da miscigenação com nativo americanas ocorre durante os primeiros séculos da colonização. Deste modo, o maior número de gerações transcorridos e o continuo processo de miscigenação, pode levar a maior discrepância entre a ancestralidade genética e a origem ancestral das avós nesses componentes ancestrais; (iii) por fim, é notório que o componente partriarcal europeu é majoritário desde o início da colonização europeia no Brasil, e os sinais de casamentos direcionais, com assimetria de ancestralidade reforçam o processo de miscigenação antigo de homens de origem europeia com mulheres de origem africana e nativo americana. Para ilustrar como esse processo de antigo miscigenação pode “distorcer” a informação sobre a origem dos avós dos indivíduos, vamos tomar como exemplo alguns quilombos no Brasil, que já entre seus fundadores apresentam indivíduos afro-descendentes com cromossomo Y de origem europeia (ex. Kimura *et al*., 2017).

O Distrito Federal com sua fundação recente (1960), é formado pelo processo secundário de migrações internas do Brasil. Onde indivíduos previamente miscigenados, migram e se estabelecem no DF. Neste sentido, o que observamos é

que apenas 58% dos participantes conseguem dizer com precisão a origem ancestral das avós maternas e 50% dos participantes homens, conseguem precisar a origem ancestral dos avôs paternos. A seguir descreveremos um pouco mais os perfil demográfico e padrões migratórios que deram origem à população do DF.

8 Análise Demográfica

8.1 Perfil de migração dos pais dos participantes para o DF

Com base na análise dos questionários aplicados, foi observada uma maior representatividade de migrantes das regiões Sudeste (43,64%) e Nordeste (38,22%) para o Distrito Federal (Tabela 13). Seguido pelas regiões Centro-Oeste (9,39%), Norte (5,52%) e em menor expressividade está a região Sul com 1,65% do percentual de migração.

Quando observado o perfil de migração destas regiões segundo o gênero percebe-se que são equivalentes e que não há discrepância entre a representatividade de homens e mulheres vindos das regiões Norte e Nordeste. Já com relação à região Centro-Oeste nota-se um maior contingente de mulheres quando comparado aos migrantes do sexo masculino. Observando o padrão de migração da região Sudeste percebe-se uma tendência de migração maior dos homens. Por fim, para os dados dos migrantes da região Sul, vê-se uma representatividade tímida para ambos os sexos.

As tabelas 13.1 e 13.2 demonstram a proporção das regiões de origem dos avós paternos e maternos dos participantes de pesquisa. Nota-se claramente a força da representatividade da origem nas regiões Sudeste e Nordeste. Sobressaindo em segundo lugar na representatividade da origem as regiões Centro-oeste e Norte para avós maternos e paternos respectivamente.

Tabela 13. Perfil de migração dos pais dos participantes para o DF, por região, gênero e a representatividade populacional por região no território brasileiro de acordo com o IBGE (2018).

Região	Migração / Região (%)	Mulheres (%)	Homens (%)	Representatividade Populacional por região brasileira (IBGE,2018)
Sudeste	43,64	40,65	47,19	44,0
Nordeste	38,22	39,56	39,32	30,0
Centro-Oeste	9,39	13,18	5,61	7,25
Norte	5,52	4,39	6,74	8,0
Sul	1,65	2,19	1,12	14,3

Tabela 13.1 Perfil de origem dos avós maternos dos participantes para o DF, por região, gênero.

Região	Migração/Região (%)	Mulheres (%)	Homens (%)
Sudeste	41,60	44,10	38,60
Nordeste	44,09	41,80	46,60
Centro-Oeste	5,59	6,90	4,00
Norte	3,72	2,30	5,30
Sul	3,10	2,30	4,00

Tabela 13.2 Perfil de origem dos avós paternos dos participantes para o DF, por região, gênero.

Região	Migração/Região (%)	Mulheres (%)	Homens (%)
Sudeste	54,34	51,35	49,27
Nordeste	41,25	41,89	40,57
Centro-Oeste	1,43	0	1,44
Norte	5,59	5,40	5,79
Sul	1,39	1,35	1,44

Os três estados mais populosos do Brasil estão na região Sudeste. Esta por sua vez demonstrou contribuir com o maior número de indivíduos imigrantes para o DF. Quando avaliados quais estados teriam sido os maiores "doadores" destes indivíduos pode-se observar que em primeiro lugar está o estado de Minas Gerais perfazendo (26,5%) do total, seguido por São Paulo (7,98%) e Rio de Janeiro (7,40%). Representando a região Nordeste estão os estados do Maranhão (9,52%) e Piauí (8,99%). Goiás (9,57%) como maior representante da região Centro Oeste e o Pará (2,11%) como maior representante da região Norte e (Tabela 14).

O ranqueamento da representatividade de migrantes segundo seus estados de origem para o DF está demonstrado na tabela 14.1. Alguma discrepância na equivalência de representatividade pode-se se dar em função do quantitativo de indivíduos participantes da pesquisa desconhecerem ou não declararem o local de origem dos pais.

Tabela 14. Porcentagem de migrantes (pais do participante da pesquisa) para o DF em relação ao estado de origem e comparativo com dados CODEPLAN e IBGE no ano de 2018.

	Presente estudo		CODEPLAN (2018)	IBGE (2018)
Estado	No. individuos	Migrantes (%)	Migrantes (%)	Habitantes/estado
Goias	18	9,52%	12,2%	6.921.161
Distrito Federal	3	1,58%	-	2.974.703
Minas Gerais	50	26,45%	16,1%	21.040.662
São Paulo	15	7,98%	4,6%	45.538.936
Rio de Janeiro	14	7,40%	5,3%	17.159.960
Espirito Santo	1	0,52%	-	3.972.388
Piauí	17	8,99%	10,6%	3.264.531
Ceará	15	7,98%	7,2%	9.075.649
Maranhão	18	9,52%	10,6%	7.035.055
Pernambuco	6	3%	3,1%	9.496.294
Paraíba	7	3,70%	4,5%	3.996.496
Bahia	6	3,17%	11,1%	14.812.617
Alagoas	2	1,05%	-	3.322.820
Rio Grande do Norte	2	1,05%	-	3.479.010
Acre	1	0,52%	-	869.265
Para	4	2,11%	-	8.513.497

Tabela 14.1. *Ranking* dos estados brasileiros fornecedores de migrantes para o DF

ESTADO	Ranking
MG	1º
MA/GO	2º
PI	3º
RJ	4ª
SP /CE	5º
PB	6º
BA	7º
PE	8º
PA	9º
DF	10º
AL /RN	11º
AC/ ES	12º
RS	13º

Esses resultados estão de acordo com o que se conhece sobre a história de formação do DF. Os fluxos migratórios foram a principal dinâmica formadora nos tempos de construção e consolidação da capital federal no Planalto Central. Num primeiro momento esse fluxo trouxe muitos trabalhadores para trabalhar na área da construção civil, desde especialista no ramo de construção aos peões que trabalhavam nas obras. Em seguida, comerciantes e prestadores de serviços também se estabelecem na nova capital. Devido ao rápido e desordenado crescimento urbano, o governo cria as "cidades-satélites" ou "cidades-dormitórios", ou seja as regiões administrativas, cuja estrutura era inferior a do Plano Piloto e eram destinada aos trabalhadores e à população, em geral, com menor poder aquisitivo que chegavam para trabalhar em Brasília. A contribuição de imigrantes das regiões Nordeste e Sudeste do Brasil na construção e consolidação de Brasília são notórios. Além disso,

há relatos de fluxos migratórios mais intensos até a década de 1980. Após esse período, as migrações diminuem, mas o DF ainda recebe imigrantes de diferentes regiões do país todos os anos (CODEPLAN, 2018).

Também é possível encontrar registros que a migração nordestina para o DF ocorre principalmente em dois momentos. O primeiro, caracterizado pela seca de 1958 no Nordeste, que impulsionou a força de trabalho de centenas de migrantes nordestinos para a região do Distrito Federal. E a segunda, na década de 80, quando houve uma explosão da população urbana no DF, levando à criação de novas regiões administrativas como Samambaia, Riacho Fundo, Quadras Lúcio Costa, Santa Maria, Recanto das Emas. Estas regiões estavam sob “administração” unicamente do governante Joaquim Roriz (1988-1994). O cenário era conhecido como "curral eleitoral", havendo grandiosos assentamentos urbanos, e destes, distribuição de lotes a famílias carentes, o que naturalmente atraiu migrantes de todo Brasil, e mais expressivamente, da região Nordeste (Maniçoba, 2019).

8.2 Perfil de distância marital e a preferência desta união entre regiões

Quando analisada a origem regional dos pais dos participantes de pesquisa, observou-se uma preferência matrimonial de 51,2% entre indivíduos da mesma região geográfica. Com maior expressividade de casos nas regiões Sudeste e Nordeste (Tabela 15), que são os mais representativos em nossa amostra. Fora deste cenário, a preferência matrimonial se mostrou entre indivíduos da região Sudeste por indivíduos da região Nordeste, seguido pela união de indivíduos do Centro-Oeste e Sudeste. Este padrão corrobora com o perfil da amostragem mais expressiva de indivíduos destas três regiões: SE, NE e CO respectivamente na região do DF.

Ainda pôde-se observar que a diversidade de união matrimonial entre estes indivíduos fora da mesma região é maior em indivíduos mais jovens (pais dos participantes de pesquisa) quando se compara ao perfil dos avós maternos e paternos. Segundo as declarações do nosso grupo amostral, a quase totalidade dos matrimônios dos avós foi estabelecida entre indivíduos das regiões Nordeste entre si e Sudeste da mesma forma. Neste contexto (avós do participante), o número pouco expressivo de casos de matrimônio entre indivíduos de regiões distintas sugere que estes, pouco migravam. Esta estimativa é parcial em função dos casos de desconhecimento do participante de pesquisa a respeito desta informação, se fazendo importante considerar

que há lacunas nas informações a respeito das origens dos avós tanto maternos quanto paternos (27,61% dados não informados). Os dados faltantes de avós maternos e paternos corresponderam a 23,33% e 31,90% respectivamente.

Tabela 15. Amostragem do perfil estimado de preferência matrimonial entre indivíduos (pais do participante da pesquisa) de acordo com a região o país (N=Norte, NE=Noreste, CO= Centro-Oeste, SE=Sudeste e S= Sul)

Regiões	Proporção
NE -NE	22 (26,2%)
SE - SE	19 (22,6%)
NE- SE	19 (22.6%)
SE - CO	10 (11,9%)
N - NE	3 (3,6%)
CO - NE	3 (3,6%)
S - SE	3 (3,6%)
N - N	2 (2,4%)
CO -N	1 (1,1%)
N - SE	1 (1,1%)
84 CASAIS	

Analisando o estado de origem de migração do grupo amostral feminino e masculino (Tabela 16) observa-se que a representação feminina da região Sudeste tem forte origem no estado de Minas Gerais, bem como no estado do Piauí na região Nordeste. No que se refere a região Centro-Oeste a grande representatividade está no estado de Goiás. A região Norte está representada pelo estado do Tocantins. Já o grupo amostral masculino da região Sudeste também revelou forte representação do estado de Minas Gerais. A maior representatividade da região Nordeste tem origem no estado do Maranhão, seguido do Estado do Ceará. Goiás representa a região Centro Oeste e a região Norte está representada pelo estado do Tocantins. Na última linha, da mesma tabela, vê-se para cada gênero a descrição dos estados de menor expressividade de migração

Tabela 16. Perfil de migração da região de origem (pais do participante) por gênero.

Região de origem das MULHERES	Região de origem dos HOMENS
GO - 12,37%	GO - 6,52%
PI - 11,34%	PI - 6,52%
MG 26,32%	MG - 28,26%
RJ - 6,18%	RJ - 8,69%
SP - 8,24%	SP - 7,60%
CE - 5,15%	CE- 10,86%
MA - 7,21%	MA - 11,95%
RS - 0	RS - 0
PE - 3,09%	PE - 3,26%
AC - 1,03%	AC - 0
PB - 6,18%	PB - 1,08%
BA - 6,18%	BA - 4,34%
AL - 1,03%	SE - 1,08%
RN - 1,03%	PR - 1,08%
PA - 1,03%	PA- 3,26%
ES - 0	ES - 1,08%

A tendência a maior proporção de casamentos entre pessoas da mesma região geográfica é algo reportado na literatura. Além disso, um estudo revela que os brasileiros tendem a se casar com indivíduos com ancestralidade similar, seja ela caracterizada por característica fenotípica, cultural, social ou geográfica (Kehdy *et al.,* 2015). Os autores do estudo ainda discutem como esse comportamento matrimonial atua na interface entre cultura e biologia e sugerem que após 5 séculos de miscigenação essa pode ser uma das razões por não haver um dissociação total entre a ancestralidade genética e fenótipos como cor da pele e, até mesmo classificação étnico-racial nas populações miscigenadas.

9 CONSIDERAÇÕES FINAIS

Este é o estudo genético e demográfico mais completo realizado na população do Distrito Federal. A partir das inferências de ancestralidade obtidas pelos dos diferentes tipos de cromossomos (autossômicos, sexuais (X e Y) e mtDNA) e associando essas informações com dados demográficos dos mesmos indivíduos, revelamos diferentes nuances da história do DF.

Pela primeira vez, as contribuições genéticas do leste asiático foram avaliadas nessa população. Esse componente ancestral, apesar de minoritário na população do DF (1,96%) revelou estar presente em 50% do genoma de 3 indivíduos. Corroborando e justificando nossa escolha por inferir a ancestralidade genética a partir do modelo tetra-híbrido de miscigenação.

Em linhas gerais, mostramos que apesar do DF apresentar maior contribuição de imigrantes vindos das regiões Sudeste e Nordeste do Brasil, a composição de ancestralidade não difere da média nacional, tanto para os cromossomos autossômicos como para o mtDNA e o cromossomo Y. Por outro lado, mostramos que há diferenças nas proporções de um ou mais componentes ancestrais entre o DF e algumas regiões geográficas do Brasil. Como apresentamos na introdução do presente estudo, os processos migratórios ocorreram de forma heterogênea pelo território brasileiro e portanto as proporções de ancestralidade genética variam entre as regiões do Brasil. O que acaba por levantar a discussão se existe uma única população brasileira representativa do Brasil ou se há várias populações e vários "Brasils". De modo pragmático, podemos dizer que em nossas análises o DF se mostrou como um recorte do perfil médio da ancestralidade genética conhecida para a população brasileira.

A partir das análises de ancestralidade no cromossomo X e do perfil de ancestralidade determinado pelos haplogrupos do cromossomo Y e mtDNA, mostramos que há sinais de casamentos direcionais, com assimetria de ancestralidade na população do DF. Como esse é um padrão geral observado na população brasileira, e remete aos tempos remotos dos primeiros movimentos migratórios intercontinentais no Brasil, provavelmente não é algo que está ocorrendo agora no DF, mas sim um reflexo do efeito fundador da população brasileira.

Por fim, os dados demográficos coletados a partir da nossa amostra sobre os padrões de migração para o DF reproduzem aqueles reportados nos relatórios do CODEPLAN. Portanto, nossa amostragem consegue reproduzir os movimentos

migratórios e a história de formação do DF. Ao associar as informações sobre local de nascimento dos genitores dos participantes com dados sobre a ancestralidade genética média de cada região do Brasil, pudemos construir um modelo demográfico de migração e verificar que o perfil de ancestralidade genética observado no DF é compatível com essas informações migratórias.

A população do DF apresenta uma série de peculiaridades, com uma formação recente, cerca de 60 anos e com poucas gerações nascidas de fato no DF, um rápido crescimento populacional impulsionado um intenso fluxo migratório interno do Brasil, ela revela não apenas história dessa região do Planalto Central, mas também de diferentes partes do Brasil. Portanto, caracterizar a população do DF do ponto de vista genético é compreender como a história recente está sendo registrada no material genético desta população.

CONCLUSÃO GERAL

Com o presente estudo contribuímos para o conhecimento sobre as populações miscigenadas ao testar a influência do conjunto de marcadores e do modelo de miscigenação na inferência de ancestralidade genética. Além disso, aplicamos esse conhecimento e agregamos informações demográficas para compreender melhor o processo recente de formação da população do DF.

No Capítulo 1 mostramos que em populações não miscigenadas os diferentes conjuntos de marcadores testados (painéis de AIMs aos dados de genoma completo) apresentam boa performance: (i) atribuem com acurácia a região biogeográfica dos indivíduos (taxa erro 0,4-1,6%) e (ii) apresentam alta correlação da ancestralidade inferida por eles (r^2>0.96). Por outro lado, mostramos que em populações miscigenadas o padrão é mais complexo. Verificamos que a escolha do modelo tri ou tetra-híbrido de miscigenação: (i) interfere na inferência do componente de ancestralidade nativo americana; (ii) conjuntos de marcadores HDSNP e WGS apresentam melhor performance nos dois modelos e melhor acurácia na atribuição dos componentes ancestrais do leste asiático e nativo americano. Também constatamos que a escolha do conjunto de marcadores pode afetar a inferência de ancestralidades: (i) os painéis com os menores números de marcadores apresentam as menores correlações com os painéis com maior número de marcadores, em especial para as inferências dos componentes ancestrais EUR e EAS; (ii) as inferências de ancestralidade nativo americana apresentam as menores correlações entre os painéis e as populações miscigenadas avaliadas. Em resumo, nosso estudo mostra que a escolha do conjunto de marcadores e do modelo de miscigenação tem impacto na inferência de ancestralidade em populações miscigenadas. Sendo que a escolha do modelo de miscigenação e do conjunto de marcadores para a inferência de ancestralidade dependerá das características históricas de cada população miscigenada e do propósito de cada estudo.

No Capítulo 2, caracterizamos a ancestralidade genética e os padrões de migração em uma amostra de 104 indivíduos do Distrito Federal. Com base na ancestralidade média inferida a partir dos cromossomos autossômicos mostramos: (i) não há diferenças significativas entre o perfil de ancestralidade genética do DF (69,9% EUR, 19,8% AFR, 9,6% NAM e 1,7% EAS) e o perfil conhecido para o Brasil; (ii) há

diferenças significativas entre proporções ancestrais pontuais entre o DF e as regiões Sul, Norte e Nordeste do Brasil; (iii) o perfil de ancestralidade observado no DF é compatível com o modelo de migração construído com base no local de nascimento dos pais dos participantes. Com base na ancestralidade inferida a partir dos cromossomos sexuais (X , Y) e mtDNA observamos; (i) padrões de casamentos direcionais com contribuição assimétrica de homens com origem majoritária EUR e mulheres NAM; (ii) o perfil de ancestralidade do cromossomo Y no DF (90% EUR e 10% AFR) não apresenta diferenças significativas em relação a outras regiões do Brasil e a média brasileira. (iii) para o perfil de ancestralidade do mtDNA (37% AFR, 32% EUR, 29% NAM e 2% EAS) há diferenças significativas entre proporções ancestrais pontuais entre o DF e outras regiões do país, mas não com a média do Brasil. Ao combinarmos informações sobre a ancestralidade genética e informações demográficas verificamos: (i) há diferenças significativas entre a percepção dos participantes sobre a origem dos avós e a ancestralidade observada no cromossomo Y (50% correspondência) e mtDNA (58% correspondência). Por fim, identificamos o perfil de imigrantes para o DF: (i) as regiões Sudeste (43,64%) e Nordeste (38,22%) são as que mais contribuíram com imigrantes; (ii) sendo MG, MA, GO, PI, RJ, SP e CE são os estados que mais contribuíram com imigrantes; (iii) MG, GO, PI, e SP são os estados que mais contribuíram com mulheres; (iv); MG, MA, CE e RJ são os estados que mais contribuíram com homens; (v) 51,2% dos matrimônios ocorrem entre indivíduos oriundos da mesma região geográfica. Em resumo, concluímos que o DF, por apresentar uma formação populacional recente e resultar de um processo secundário de migrações internas do Brasil, é um reflexo de alguns dos padrões históricos de suas populações originais, em especial do Sudeste e Nordeste do Brasil. Assim, o perfil de ancestralidade e os padrões de casamentos direcionais com ancestralidade assimétrica, possivelmente são uma herança ainda da primeira onda migratória formadora da população brasileira, cuja assinatura genética ainda permanece no DNA da nossa população e foi transmitida à população do DF.

Esta tese é um esforço para melhor compreender as populações miscigenadas e as ferramentas para estudá-las. Aqui buscamos explorar diferentes painéis de ancestralidade para trabalhar com populações miscigenadas, diferentes modelos demográficos e integrar diferentes níveis de informações para aplicar esse entendimento na caracterização da população do DF sob o ponto de vista genético e demográfico.

REFERÊNCIAS

Abe-Sandes, K; Silva-Jr, W; Zago, MA. (2004). Heterogeneity of the Y chromosome in Afro-Brazilian populations. Hum Biol 76(1):77-86.

Acuña, MP; Llop, ER; Rothhammer, FE. (2000). Composición genética de la población chilena: las comunidades rurales de los valles de Elqui, Limarí y Choapa. Rev. Méd. Chile 128(6): 593-600.

Agrawal, S; Khan, F. (2005). Reconstructing recent human phylogenies with forensic STR *loci*: A statistical approach. BMC Genetics 6:47-53.

Alencastro, LP. (2000). O trato dos viventes: formação do Brasil no Atlântico Sul. Companhia das Letras, São Paulo/SP, 523p.

Alves, C; Gusmão, L; Damasceno, A; Soares, B; Amorim, A. (2004). Contribution for an African autosomic STR database (AmpF/STR Indentifiler and Powerplex 16 System) and a report on genotypic variations. F Sci Int 139: 201-205.

Alves, C; Gusmão, L; López-Parra, AM; Mesa, MS; Amorim, A; Arroyo-Pardo, E. (2005). STR allelic frequencies for an African population sample (Equatorial Guinea) using AmpFlSTR Identifiler and Powerplex 16 kits. F. Sci. Int. 148: 239-242.

Alves-Silva, J; Santos, MS; Guimarães, PEM; Ferreira, ACS; Bandelt, HJ; Pena, SDJ; Prado, VF. (2000). The Ancestry of Brazilian mtDNA Lineages. Am J Hum Genet 67: 444–461.

Arcot, SS; Adamson, AW; Risch, GW; LaFleur, J; Robichaux, MB; Lamerdin, JE; Carrano, AV; Batzer, MA. (1998). High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. Journal of Molecular Biology 281(5): 843-56.

Arpini-Sampaio, Z; Costa, MC; Melo, AA; Carvalho, MF; Deus, MF; Simões, AL. (1999). Genetic polymorphisms and ethnic admixture in African-derived black communities of northeast Brazil. Human Biology 71(1):69-85.

Astle, W. & Balding, D. J., (2009). Population structure and cryptic relatedness in genetic association studies. Statistical Science, pp. 451-471.

Balaresque PL, Ballereau SJ, Jobling MA (2007). Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet. 15;16 Spec No. 2:R134-9. doi: 10.1093/hmg/ddm242. PMID: 17911157.

Bamshad, M; Wooding, S; Salisbury, BA; Stephens, JC. (2004). Deconstructing the relationship between genetics and race. Nature 5: 598-609.

Bandelt, HJ; Herrnstadt, C; Yao, YG; Kong, QP; Kivisild, T; Rengo, C; Scozzari, R; Richards, M; Villems, R; Macaulay, V; Howell, N; Torroni, A; Zhang, YP. (2003). Identification of Native American Founder mtDNAs Through the Analysis of Complete mtDNA Sequences: Some Caveats. Annals of Human Genetics 67(6):512– 524

Barbujani, G. (2005). Human races: classifying people vs understanding diversity. Current Genomics 6:1-12.

Barcelos, R; Ribeiro, G; Silva Jr, W; Abe-Sandes, K; Godinho, N; Marinho-Neto, F; Gigonzac, M; Klautau-Guimarães, M; Oliveira, S. (2006). Male contribution in the constitution of the Brazilian Centro-Oeste population estimated by Y-chromosome binary markers. International Congress Series 1283:228-230.

Berger B, Niederstätter H, Erhart D, Gassner C, Schennach H, Parson W. High resolution mapping of Y haplogroup G in Tyrol (Austria) (2013). Forensic Sci Int Genet. Sep;7(5):529-36. doi: 10.1016/j.fsigen.2013.05.013.

Bastos-Rodrigues L., Pimenta J. R., Pena S. D. J. (2006) The Genetic Structure of Human Populations Studied Through Short Insertion-Deletion Polymorphisms. Annals of Human Genetics 70:658–665.

Batista Dos Santos, Sidney E., et al. "Differential contribution of indigenous men and women to the formation of an urban population in the Amazon region as revealed by mtDNA and Y-DNA." *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 109.2 (1999): 175-180.

Batzer, M; Stoneking, M; Alegria-Hartman, M; Bazan, H; Kass, DH; Shaik, H; Novick, GH; Ioannou, PA; Scheer, WD; Herrera, RJ; Deininger, PL. (1994). African origin of human-specific polymorphic Alu insertions. Proc Nail Acad Sci 91: 12288-12292.

Batzer, MA; Rubinb, CM; Hellmann-Blumbergc, U; Alegria-Hartmana, M; Leeflangc, EP; Sternb, JD; Bazand, HA; Shaikhd, TH; Deininger, PL; Schmid, CW. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. Journal of Molecular Biology. 247(3):418-27.

Batzer, MA; Knight, A; Stoneking, M; Tiwar, HK; Scheer, WD; Herrera, RJ; Deininger, PL. (1996a). DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. Proc Natl Acad Sci 93: 4360-4364.

Batzer, MA; Arcot, SS; Phinney, JW; Alegria-Hartman, MA; Kass, DH; Miligan, SM; Kimptom, C; Gill, P; Hochmaister, M; Ioannou, PA; Hehhera, RJ; Boudreau, DA; Scheer, WD; Keats, BJ; Deininger, PL; Stoneking, M. (1996b). Genetic variation of recent Alu insertions in human populations. J. Mol. 42: 22-29.

Bedoya, G; Montoya, P; Garcia, J; Soto, I; Bourgeois, S; Carvajal, L; Labuda, D; Alvarez, V; Ospina, J; Hedrick, PW; Ruiz-Linares, A. (2006). Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. PNAS 103(19):7234-7239.

Behar, DM; Villems R; Soodyall, H; Blue-Smith, J; Pereira, L; Metspalu, E; Scozzari, R; Makkan, H; Tzur, S; Comas, D; Bertranpetit, J; Quintana-Murci, L; Tyler-Smith, C; Wells, RS; Rosset, S; Geographic Consortium. (2008). The dawn of human matrilineal diversity. Am J Hum Genet 82:1-11.

Benn-Torres, J; Bonilla, C; Robbins, CM; Waterman, L; Moses, TY; Hernandez, W; Santos, ER; Bennett, F; Aiken, W; Tullock, T; Coard, K; Hennis, A; Wu, S; Nemesure, B; Leske, MC; Freeman, V; Carpten, J; Kittles, RA. (2007). Admixture and population stratification in African Caribbean populations. Ann. Hum. Gen. 71:1- 9.

BENCHIMOL, Jaime Larry. *Dos micróbios aos mosquitos: febre amarela e a revolução pasteuriana no Brasil*. Rio de Janeiro: Editora Fiocruz/Editora UFRJ, 1999.

Bernasconi, A. (2000). Imigrantes italianos na Argentina (1880-1930) – uma aproximação. In

Fausto, B. Fazer a América. Ed. Universidade de São Paulo, São Paulo/SP, p 61-92.

Bertran, P. (2000). História da terra do homem no planalto central: eco-história do Distrito Federal – do indígena ao colonizador. Verano Editora, Brasília/DF, 270p.

Black FL (1992). “Why did they die?”. Science. 5089 (258): 1739-40

Bomfim, TF. (2008). Ancestralidade genômica em uma amostra de portadores do HIV-1 do Estado da Bahia. PPGBSMI – Programa de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa Salvador –Dissertação de Mestrado.

Bonilla, C; Parra, EJ; Plaff, CL; Dios, S; Marshall, JA; Hamman, RF; Ferrell, RE; Hoggart, CL; Mckeigue, PM; Shriver, D. (2004a). Admixture in the hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann. Hum. Gen. 68: 139-153.

Bonilla, C; Shriver, MD; Parra, EJ; Jones, A; Fernández, JR. (2004b). Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York city. Hum. Genet. 115:57-68.

Bonilla, C et al. “Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping.” *Annals of human genetics* vol. 68,Pt 2 (2004): 139-53. doi:10.1046/j.1529-8817.2003.00084.x

Bookstein, R; Lai, CC; To, H; Lee, WH. (1990). PCR-based detection of a polymorphic BamHI site in intron 1 of the human retinoblastoma (RB) gene. Nucleic. Acid. Res. 18(6):1666.

Bowcock, AM; RUiz-Linares, A; Tomfohrde, J; Minch, E; Kidd, JR; Cavalli-Sforza, LL. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 368(6470):455-7.

Bydlowski, SP; Moura-Neto, RS; Soares, RPS; Silva, R; Debes-Bravo, AA; Morganti, L. (2003). Genetic data on 12 STRs (F13A01, F13B, FESFPS, LPL, CSF1PO, TPOX,TH01, vWA, D16S539, D7S820, D13S317, D5S818) from four ethnic groups of São Paulo, Brazil. F Sci Int. 135:67–71.

Caine, LM; Corte-Real, F; Anjos, MJ; Carvalho, M; Serra, A; Antunes, H; Vide, MC; Vieira, DN. (2003). Allele frequencies of 13 *loci* in the Santa Catarina Population of Southern Brazil. J Forensic Sci 48(4): 901-902.

Callegari-Jacques, SM; Grattapaglia, D; Salzano, FM; Salamoni, SP; Crossetti, SC; Ferreira, ME; Hutz, MH. (2003). Historical genetics: spatiotemporal analysis of formation of the brazilian population. Am J Hum Biol 15: 824-834.

Campos-Sánchez R, Barrantes R, Silva S, Escamilla M, Ontiveros A, Nicolini H, Mendoza R, Munoz R, Raventos H. (2006). Genetic structure analysis of three Hispanic populations from Costa Rica, Mexico, and the southwestern United States using Y- chromosome STR markers and mtDNA sequences. Hum. Biol. 78(5):551-63.

Carvalho-Silva, DR; Santos, FR; Jorge Rocha, J; Pena, SDJ. (2001). The Phylogeography of Brazilian Y-Chromosome Lineages. Am J Hum Genet 68: 281–286.

Castro E Silva MA, Ferraz T, Couto-Silva CM, Lemes RB, Nunes K, Comas D, Hünemeier T. (2022). Population Histories and Genomic Diversity of South American Natives. Mol Biol Evol. 7;39(1):msab339.

Cavalli-Sforza, LL; Menozzi, P; Piazza A. (1996). The history and geography of human genes. Princeton University Press, Princeton, New Jersey, 443p.

Cavalli-Sforza, LL. (1997). Genes, peoples, and languages. Proc Natl Acad Sci USA.94(15):7719-24.

Cerda-Flores, RM; Budowle, B; Jin, L; Barton, AS; Keka, R; Chakraborty, R. (2002). Maximum likelihood estimates of admixture in Northeastern Mexico using 13 short tandem repeat *loci*. Am J Hum Biol 14:429-439.

Chakraborty, R. (1985). Gene identity in racial hybrids and estimation of admixture rates. In: Y Ahuja and JV Neel (Eds.): Genetics Microdifferentiation in Human and Other Animal Populations. Delhi, India: Indian Anthropological Association, Delhi University Anthropology Department, pp. 171-180.

Chakraborty, R; Weiss, KM. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between *loci*. Proc Natl Acad Sci USA. 85(23):9119-23.

Cheung EYY, Phillips C, Eduardoff M, Lareu MV, McNevin D. (2019). Performance of ancestry-informative SNP and microhaplotype markers. Forensic Sci Int Genet. Nov;43:102141.

Choudhry, S; Coyle, NE; Tang, H; Salari, K; Lind, D; Clark, SL; Tsai, H; Naqvi, M; Ung, Phong, A; Ung, N; Matallana, H; Avila, PC; Casal, J; Torres, A; Nazario, S; Castro, R; Battle, NC; Perez-Stable, EJ; Kwok, PY; Sheppard, D; Shriver, MD; Rodriguez- Cintron, W; Risch, N; Ziv, E; Burchard, GE. (2006). Population stratification confounds genetic association studies among Latinos. Hum. Genet. 118: 652–664.

Cinnioğlu C, King R, Kivisild T, Kalfoğlu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA. (2004). Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet. 2004 Jan;114(2):127-48.

Cockerham, CC; Weir, BS. (1987). Correlations, descent measures: Drift with migration and mutation. Proc Natl Acad Sci 84: 8512-8514.

Collins-Schramm, HE; Phillips, CM; Operario, DJ; Lee, JS; Weber, JL; Hanson, RL; Knowler, WC; Cooper, R; Li, H; Seldin, MLF. (2002). Ethnic-Difference Markers for Use in Mapping by Admixture Linkage Disequilibrium. Am J Hum Genet 70: 737– 750.

Comas, D; Calafell, F; Benchemsi, N; Helal, A; Lefranc, G; Stoneking, M; Batzer, MA; Bertranpetit, J; Sajantila, A. (2000).Alu insertion polymorphisms in NW Africa and the Iberian Peninsula: evidence for a strong genetic boundary through the Gibraltar Straits. Hum Genet 107:312-319.

Campos-Sánchez, R; Barrantes, R; Silva, S; Escamilla, M; Ontiveros, A; Nicolini, H; Mendoza, R; Munoz, R; Raventos, H. (2006). Genetic analysis of three Hispanic populations from Costa Rica, Mexico, and the Southwestern United States using Y- Chromosome STR markers and mtDNA sequences. Hum Biol 78(5):551-563.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Corte-Real, F; Andrade, L; Anjos, MJ; Carvalho, M; Vieira, DN; Carracedo, A; Vide, MC. (2000). Population genetics of nine STR *loci* in two populations from Brazil. J Forensic Sci. 45(2): 432-435.]

DA CUNHA, Manuela Carneiro. **Índios no Brasil: história, direitos e cidadania**. Editora Companhia das Letras, 2013.

Daigo, Masao (2008). *Pequena história da imigração japonesa no Brasil*. São Paulo, São Paulo: Gráfica Paulos

Dalton, G; Godinho, NMO; Gontijo, CC; Diniz, MECG; Alencar, GF; Amorim, CEG; Barcelos, RSS; Klautau-Guimarães, MN; Oliveira, SF. (2007). Population substructure by social stratification. In: 22nd Congress of the International Society for Forensic Genetics, Denmark. Annals of the 22nd Congress of the International Society for Forensic Genetics.

Dellalibera, E; Havro, MLB; Souza, M; Kajihara, K; Mauricio-da-Silva, L; Silva, RS. (2004). Genetic analysis of 13 STR *loci* in the population from the State of Pernambuco, northeast Brazil. F Sci Int 146:57–59.

Destro-Bisol, G; Boschi, I; Cagliá, A; Tofanelli, S; Pascali, V; Paoli, G; Spedini, G. (2000). Microsatellite variation in Central Africa: an analysis of intrapopulational and interpopulational genetic diversity. Am. J. Phy. Antropology. 112:319-337.

De Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, Nyambe T, Beyer K, Schreiber H, de Knijff P, Luiselli D, Stoneking M, Pakendorf B. (2011). Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol. Mar;28(3):1255-69.

De Moura, R.R., Coelho, A.V.C., Balbino, V.Q., Crovella, S., Brandão, L.A.C. (2015). Meta-Analysis of Brazilian Genetic Admixture and Comparison with Other Latin America Countries. American Journal of Human Biology 27 674–680. doi:10.1002/ajhb.22714

Diamond J . Armas Germes e Aço: Os destinos das sociedades humanas. Tradução: Silvia de Souza Costa. Record.2006.8ed

Didion JP, Yang H, Sheppard K, Fu CP, McMillan L, de Villena FP, Churchill GA. (2012). Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics. Jan 19;13:34.

Dillehay, T. D. (2003). Tracking the first Americans. Nature, pp. 23-24.

Dixon, L; Murray, C; Archer, E; Dobbins, A; Koumi, P; Gill, P. (2005). Validation of a 21- locus autosomal SNP multiplex for forensic identification purposes. Forensic Sci. Int. 154(1):62-77.

Drobnic, K; Budowle, B. (2000). The analysis of three short tandem repeat (STR) *loci* in the Slovene population by multiplex PCR. J. Forensic Sci. 45(4):893-5.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. 16(12):551-8.

Excoffier, L; Smouse, P; Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Fagundes, NJR; Kanitz, R; Eckert, R; Valls, ACS; Bogo, MR; Salzano, FM; Smith, DG; Silva Jr, WA; Zago, MA; Ribeiro-dos-Santos, AKC; Santos, SEB; Petzl-Erler, ML; Bonatto, SL. (2008). Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. American Journal of Human Genetics, v. 82, p. 1-10.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package. Cladistics 5:164–166.

Fernandes, AT; Brehm, A. (2002). Population data of five STRs in three regions from Portugal. F Sci Int 129:72–74.

Ferreira, FL; Leal-Mesquita, ER; Santos, SEB; Ribeiro-dos-Santos, AKC. (2005). Genetic characterization of the population of São Luís, MA, Brazil. Genet Mol Biol 28(1):22- 31.

Ferreira, LB; Mendes-Júnior, CT; Wiezel, CEV; Luizon, MR; Simões, AL. (2006). Genomic ancestry of a sample population from the State of São Paulo, Brazil. Am J Hum Biol 18:702-705.

Gadelha, SR; Alcantara, LC; Costa, GC; Rios, DL; Galvão-Castro, B. (2005). Ethnic differences in the distribution of interleukin-6 polymorphisms among three Brazilian ethnic groups.Hum Biol 77(4):509-14.

Gelernter, J; Kranzler, H; Cubells, JF; Ichinose, H; Nagatsu, T. (1998). DRD2 allele frequencies and linkage disequilibria, including the -141Cins/Del promoter polymorphism, in European-American, African-American, and Japanese subjects. Genomics 51:21-26.

Gene, M; Moreno, P; Borrego, N; Pique, E; Brandt, C; Mas, J; Luna, M; Corbella, J; Huguet, E. (2001). The Bubi population of Equatorial Guinea characterized by HUMTH01, HUMVWA31A, HUMCSF1PO, HUMTPOX, D3S1358, D8S1179,D18S51 and D19S253 STR polymorphisms. Int J Legal Med. 114:298–300.

Gigonzac, MAD. (2002). Análise de marcadores de DNA D7S460, TPA25, ACE, PV92 na população do estado de Goiás. Dissertação de mestrado, Universidade Federal de Goiás.

Giolo SR, Soler JM, Greenway SC, Almeida MA, de Andrade M, Seidman JG, Seidman CE, Krieger JE, Pereira AC. Brazilian urban population genetic structure reveals a high degree of admixture. Eur J Hum Genet. 2012 Jan;20(1):111-6. doi: 10.1038/ejhg.2011.144. Epub 2011 Aug 24. PMID: 21863058; PMCID: PMC3234512.

GODINHO, NMO. (2008). O impacto das migrações na constituição genética de populações latinoamericanas.. Tese (Biologia Animal) - Universidade de Brasília

Góes, ACS; Silva, DA; Gila, EHF; Silva, MTD; Pereira, RW; Carvalho, EF. (2004). Allele frequencies data and statistic parameters for 16 STR *loci*—D19S433, D2S1338, CSF1PO, D16S539, D7S820, D21S11, D18S51, D13S317, D5S818, FGA, Penta E, TH01, vWA, D8S1179, TPOX, D3S1358—in the Rio de Janeiro population, Brazil. F Sci Int. 140:131–132.

Goldberg, A.; Mychajliw, A. M.; Hadly, E. A (2016). Post-invasion demography of prehistoric humans in South America. Nature, v. 532, n. 7598, p. 232-235.

Gomes, AC. (2000). Imigrantes italianos: entre a italianità e a brasilidade. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 159-177.

Gomes, AV; Maurício-da-Silva, L; Raposo, G; Vieira, JRC; Silva, RS. (2007). 13 STR *loci* frequencies in the population from Paraíba, Northeast Brazil. Forensic Science International. 173:231-234.

Gonçalves, R; Jesus, J; Fernandes, AT; Brehm, A. (2002). Genetic profile of a multi-ethnic population from Guiné-Bissau (West African coast) using the new PowerPlex1 16 System kit. F Sci Int. 129:78–80.

González-Andrade, F; Sánchez, K; González-Solórzano, J; Gascón, S; Martínez-Jarreta, B. (2007). Sex-specific genetic admixture of Mestizos, Amerindian Kichwas, and Afro-Ecuadorians from Ecuador. Hum Biol 79(1):51-77.

González-Andrade, F; Sánchez-Q, D; Martínez-Jarreta, B. (2006). Genetic analysis of the Amerindian Kichwas and Afro American descendents populations from Ecuador characterised by 15 STR-PCR polymorphisms. Forensic Sci Int 160(2-3):231-5.

Gopalan S, Smith SP, Korunes K, Hamid I, Ramachandran S, Goldberg A. (2022) Human genetic admixture through the lens of population genomics. Philos Trans R Soc Lond B Biol Sci. Jun 6;377(1852):20200410.

Gotoda, T; Yamada, N; Murase, T; Miyake, S; Murakami, R; Kawamura, M; Kozaki, K; Mori, N; Shimano, H; Shimada, M. (1992). A newly identified null allelic mutation in the human lipoprotein lipase (LPL) gene of a compound heterozygote with familial LPL deficiency. Biochim Biophys Acta. 1138(4):353-6.

Goudet, J; Raymond, M; Meeiis, T; Roussett, F. (1996). Testing Differentiation in Diploid Populations. Genetics. 144: 1933-1940.

Gouveia MH, Borda V, Leal TP, Moreira RG, Bergen AW, Kehdy FSG, Alvim I, Aquino MM, Araujo GS, Araujo NM, Furlan V, Liboredo R, Machado M, Magalhaes WCS, Michelin LA, Rodrigues MR, Rodrigues-Soares F, Sant Anna HP, Santolalla ML, Scliar MO, Soares-Souza G, Zamudio R, Zolini C, Bortolini MC, Dean M, Gilman RH, Guio H, Rocha J, Pereira AC, Barreto ML, Horta BL, Lima-Costa MF, Mbulaiteye SM, Chanock SJ, Tishkoff SA, Yeager M, Tarazona-Santos E. (2020). Origins, Admixture Dynamics, and Homogenization of the African Gene Pool in the Americas. Mol Biol Evol. 1;37(6):1647-1656.

Grandy, DK ; Litt, M; Allen, L; Bunzow, JR; Marchionni, M; Makam, H; Reed, L; Magenis, RE; Civelle, O. (1989).The human dopamine D2 receptor gene is located in chromosome 11 at q22-q23 and identifies a TaqI polymorphism. Am J Hum Genet 45:778-785.

Gratapaglia, D; Schmidt, AB; Costa e Silva, C; Stringher, C; Fernandes, AP; Ferreira, ME. (2001). Brazilian population database for the 13 STR *loci* of AmpFlSTR Profiler Plus TM and Cofiler TM multiplex kits. Forensic Sci. Int. 11:891-

Gregory, V. (2000). Imigração alemã: formação de uma comunidade teuto-brasileira. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 141-157.

Guerreiro, JF; Chautard-Freire-Maia, EA. (1988). ABO and Rh blood groups, migration and estimates of racial admixture for the population of Belém, State of Pará, Brazil. Brazil J Genetics 11:171-186.

Grugni V, Raveane A, Mattioli F, Battaglia V, Sala C, Toniolo D, Ferretti L, Gardella R, Achilli A, Olivieri A, Torroni A, Passarino G, Semino O. (2018). Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective. Ann Hum Biol. Feb;45(1):44-56.

Hammer M. F., Spurdle A. B., Karafet T., Bonner M. R., Wood E. T., Novelleteo A., Malaspina P., MitchellI R. J., Horais S., Jenkins T., Zegira S. L. (1997) The Geographic Distribution of Human Y Chromosome Variation. Genetics 145:787-805.

Hancock, JM. (1996). Microsatellites and other simple sequences in the evolution of the human genome. In Jackson, M.; Strachan, T. & Dover, G. Human Genome Evolution. BIOS. 306 p.

Handley, LJL; Manica, A; Goudet, J; Balloux, F. (2007). Going the distance: human population genetics in a clinal world. Trends in Genetics 23(9):432-439.

Hanski, I. (1998). Metapopulation dynamics. Nature, 396:41–49.Bilton, DT; Freeland, JR; Okamura, B. 2001. Dispersal in freshwater invertebrates. Annual Review of Ecology and Systematic, 32: 159–181.

Hart, D. L.; Clark; A. G. Princípios de genética de populações. 4. ed. Porto Alegre: Editora Artmed, 2010.
Herman, J. (2000). Cenário do encontro de povos: a construção do território. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 17-34.

Hernández-Gutiérrez, S; Hernández-Franco, P; Martínez-Tripp, S; Ramos-Kuria, M; Rangel-Villalobos, H. (2005). STR data for 15 *loci* in a population sample from the central region of Mexico. F Sci Int 151:97–100.

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

Jeunemaitre, X; Soubrier, F; Kotelevtsev, YV; Lifton, RP; Willians, CS; Charru, A. (1992). Molecular basis of human hypertension: role of angiotensinogen. Cell 71(1)169-80.

Jorde, LB; Wooding, SP. (2004). Genetic variation, classification and 'race'. Nat Genet 36(11 Suppl):S28-33.

Joerin-Luque, I.A., Augusto, D.G., Calonga-Solís, V. et al. Uniparental markers reveal new insights on subcontinental ancestry and sex-biased admixture in Brazil. Mol Genet Genomics 297, 419–435 (2022).

Karathanasis, SK. (1985). Apolipoprotein multigene family: tandem organization of human apolipoprotein AI, CIII, and AIV genes. Proc Natl Acad Sci USA 82(19):6374-8.

Keinan A., et.al. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than Europeans. Nature Genetics 39(10):1251-5 (2007).

Kenneth, KK. (2007). The Allele Frequency Database. A resource of gene frequency data on human populations supported by the U. S. National Science Foundation. Disponível em http://alfred.med.yale.edu/alfred/. Yale University.

Keyeux G, Rodas C, Gelvez N and Carter D (2002) Possible migration routes into South America deduced from mitochondrial DNA studies in Colombian Amerindian populations. Hum Biol 74:211-233.

Kitchen A, Miyamoto MM, Mulligan CJ (2008). A Three-Stage Colonization Model for The Peopling of the America

Kidd, KK; Morar, B; Castiglione, CM; Zhao, H; Pakstis, AJ; Speed, WC; Bonne-Tamir, B; Lu, RB; Goldman, D; Lee, C; Nam, YS; Grandy, DK; Jenkins, T; Kidd, J. (1998). A global survey of haplotype frequencies, and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211-227.

Klein, HS. (1986). African Slavery in Latin America and the Caribbean. Oxford University Press, New York, 316p.

Klein, H.S. (2000). Migração internacional na história das Américas. In Fausto, B. Fazer a América. Ed. Universidade de São Paulo, p 13-31.

Klein, HS. (2002). As origens africanas dos escravos brasileiros *in Homo brasilis.*Organizado por Sérgio DJ Pena. Funpec- Editora, São Paulo, p 93-112.

Kimura L, Nunes K, Macedo-Souza LI, Rocha J, Meyer D, Mingroni-Netto RC. (2017) Inferring paternal history of rural African-derived Brazilian populations from Y chromosomes. Am J Hum Biol.29(2).

Kruglyak, S; Durrett, RT; Schug, MD; Aquadro, CF. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95(18):10774-8.

Lawson, D.J.; Hellenthal, G.; Myers, S.; Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. 8, e1002453.

Lee, ST; Nicholls, RD; Jong, MTC; Fukai, K; Spritz, RA. (1995). Organization and sequence of the human P gene and identification of transport proteins. Genomics 26, 354 – 363.

Leite, FPN; Menegassi, FJ; Schwengber, SP; Raimann, PE; Albuquerque, TK. (2003). STR data for 09 autosomal STR markers from Rio Grande do Sul (southern Brazil). F Sci Int 132:223–224.

Lewis, PO & Zaykin, D. (2002). - Genetic Data Analysis: version 1.1 for Windows 95/NT. http://www.lewis.eeb.uconn.edu/lewishome/. 2002. GDA user's manual.

Lima-Costa, M., Rodrigues, L., Barreto, M. *et al.* (2015*)* .Ancestralidade genômica e autoclassificação etnoracial baseada em 5.871 brasileiros residentes na comunidade (The Epigen Initiative). *Sci Rep* 5, 9812.

Lind, JM; Hutcheson-Dilks, HB; Williams, SM; Moore, JH; Essex, M; Ruiz-Pesini, E.; Wallace, DC; Tishkoff, SA; O'Brien, SJ; Smith, MW. (2007).Elevated male European and female African contributions to the genomes of African American individuals. Human Genetics 120:713-722.

Liu, Y; Saha, N; Low, PS; Tay, JS. (1995). Linkage disequilibrium between two *loci* (5'DdeI) of the antithrombin III gene in three ethnic groups in Singapore. Hum. Hered. 45:192-198.

Liu, K; Muse, SZ. (2005). PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics 21:2128-9.

Liu H, Prugnolle F, Manica A, Balloux F. (2006) A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. 79: 230–237.

Loh P.R., Lipson M., Patterson N., Moorjani P., Pickrell J.K., Reich D., Berger B. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193:1233–1254. doi: 10.1534/genetics.112.147330.13

López-Camelo, JS; Cabello, PH; Dutra, MG. (1996). A simple model for the estimation of congenital malformation frequencies in racial mixed populations. Braz. J. Genet. 19:659-663.

López S, van Dorp L, Hellenthal G. (2016). Human Dispersal Out of Africa: A Lasting Debate. Evol Bioinform Online. 21;11(Suppl 2):57-68.

Lovo-Gómez, J; Salas, A; Carracedo, A. (2007). Microsatellite autosomal genotyping data in four indigenous populations from El Salvador. F Sci Int 170:86-91.

Loyo, MA; de Guerra, DC; Izaguirre, MH; Rodriguez-Larralde, A. (2004). Admixture estimates for Churuguara, a Venezuelan town in the State of Falcón. Annals of Hum Biol 31(6):669-80.

Lu Y., et al. Technical design document for a SNP array that is optimized for population genetics. ftp://ftp.cephb.fr/hgdp_supp10/8_12_2011_Technical_Array_Design_Document.pdf

Luis, JR; Terreros, MC; Martinez, L; Rojas, D; Herrera, RJ. (2003). Two problematic human polymorphic Alu insertions. Electrophoresis 24(14):2290-2294.

Luizon, MR. (2007). Dinâmica da mistura étnica em comunidades remanescentes de quilombo brasileiras. Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.

Luizon, MR; Mendes-Junior, CT; de Oliveira, SF; Simões, AL. (2008). Ancestry informative markers in Amerindians from Brazilian Amazon. Am. J. Hum. Biol. 20(1):86-90.

Luna-Vazquez, A; Vilchis-Dorantes, G; Paez-Riberos, LA; Muñoz-Vallec, F; González- Martin, A; Rangel-Villalobos, H. (2003). Population data of nine STRs of Mexican- Mestizos from Mexico City. F Sci Int 136:96–98.

Maniçoba, S, R. CRIAÇÃO DE REGIÕES ADMINISTRATIVAS NO DISTRITO FEDERAL E O HISTÓRICO DA DEFINIÇÃO DE SEUS LIMITES GEOGRÁFICOS Revista Eletrônica: Tempo - Técnica - Território, v.10, n.2 (2019),p.01:30 ISSN: 2177-4366.

Machado, TMBM. (2008). Ancestralidade em Salvador-BA. Dissertação de mestrado em Biotecnologia em Saúde e Medicina Investigativa – Centro de Pesquisa Gonçalo Moniz – FIOCRUZ-BA.

Manica, A; Prugnolle, F; Balloux, F. (2005). Geography is a better determinant of human genetic differentiation than ethnicity. Hum. Genet. 118(3-4):366-71. Revisiting the Genetic Ancestry of Brazilians Using Autosomal AIM-Indels, PLoS ONE 8(9): e75145 (2013). doi:10.1371/journal.pone.0075145

Manta F.S.N., Pereira R., Vianna R., Araújo A.R.B., D.L.G. Gitaí, D.A. Silva, E.V. Wolfgramm, I.M. Pontes, J.I. Aguiar, M.O. Moraes, E.F. Carvalho, L. Gusmão,

Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93:278–88.

Mas Montserrat, D., Bustamante, C. & Ioannidis, A. (2020). LAI-Net: Local-Ancestry Inference With Neural Networks. Proc. IEEE Int. Conf. on Acoust. Speech Signal Process.

Marjanovic, D; Kapur, L; Drobnic, K; Budowle, B; Pojskic, N; Hadziselimovic, R. (2004). Comparative Study of Genetic Variation at 15 STR *Loci* in Three Isolated Populations of the Bosnian Mountain Area. Human Biology 76(1): 15-31.

Mark Jobling, Edward Hollox, Matthew Hurles, Toomas Kivisild, Chris Tyler-SmithHuman Evolutionary Genetics.2nd Edition. 2013. Garland Science: New York. ISBN: (Paperback) 978-0815341482. 443-445p.

Martínez, H; Rodríguez-Larralde, A; Izaguirre, MH; de Guerra, DC. (2007). Admixture estimates for Caracas, Venezuela, based on autosomal, Y-chromosome, and mtDNA markers. Hum Biol. 79(2):201-213.

Martínez-Marignac, VL; Bertoni, B; Parra, EJ; Bianchi, NO. (2004). Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. Hum Biol. 76(4):543-557.

Martinez-Marignac VL, Valladares A, Cameron E, Chan A, Perera A, Globus-Goldberg R, Wacher N, Kumate J, McKeigue P, O'Donnell D, Shriver MD, Cruz M, Parra EJ. (2007). Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. Hum. Genet. 120(6):807-19.

Martinovic, I; Barać, L; Furac, I; Janićijević, B; Kubat, M; Perićić, M; Vidović, B; Rudan,
Mattias Jakobsson, Noah A. Rosenberg. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, *Bioinformatics*, Volume 23, Issue 14, 1801–1806p.

McHugh C, Brown L, Thornton TA. (2016). Detecting Heterogeneity Population Structure Across the Genome in Admixed Populations. Genetics. 2016 Sep;204(1):43-56. doi: 10.1534/genetics.115.184184.

McCombs, JL; Yang, F; Bowman, BH; McGill, JR; Moore, CM. (1986). Chromosomal localization of group-specific component by in situ hybridization. Cytogenet Cell Genet. 42(1-2):62-64.

Mellati, J. C. Índios do Brasil. 9. ed. São Paulo: Editora da Universidade de São Paulo, 2014

Melo, JO. (1996). Historia de Colombia. La dominación Española. Imprenta Nacional de Colombia. Presidencia de la República, 358p.

Mendes-Jr, CT; Simões, AL. (2001). Alu insertions and ethnic composition in a Brazilian population sample. IJHG 1(4):249-254.

Mellati, J. C. Índios do Brasil. 9. ed. São Paulo: Editora da Universidade de São Paulo, 2014.

Miller, SA; Dykes, DD; Polesky, HF. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research. 161, 1215.

Montinaro, F. et al., 2015. Unravelling the hidden ancestry of American admixed populations. Nature Communications, p. 6596.

Morales, JÁ; Monterrosa, JC; Puente, J. (2004). Population genetic data from El Salvador (Central America) using AmpFlSTR Identifiler PCR amplification kit. International Congress Series 1261:223-225.

Morera, B; Barrantes, R; Marin-Rojas, R. (2003). Gene admixture in the Costa Rican population. Annals of Hum Genet. 67:71-80.

Myres NM, Rootsi S, Lin AA, Järve M, King RJ, Kutuev I, Cabrera VM, Khusnutdinova EK, Pshenichnov A, Yunusbayev B, Balanovsky O, Balanovska E, Rudan P, Baldovic M, Herrera RJ, Chiaroni J, Di Cristofaro J, Villems R, Kivisild T, Underhill PA. (2011). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet. 19(1):95-101.

Nalls, MA; Wilson, JG; Patterson, NJ; Tandon, A; Zmuda, JM; Huntsman, S; Garcia, M; Hu, D; Li, R; Beamer, BA; Patel, KV; Akylbekova, EL; Files, JC; Hardy, CL; Buxbaum, SG; Taylor, HA; Reich, D; Harris, TB; Ziv, E. (2008). Admixture mappingof white cell count: genetic locus

responsible for lower white blood cell count in the health ABC and Jackson heart studies. Am J Hum Genet 82:81-87.

Naslavsky MS, Scliar MO, Yamamoto GL, Wang JYT, Zverinova S, Karp T, Nunes K, Ceroni JRM, de Carvalho DL, da Silva Simões CE, Bozoklian D, Nonaka R, Dos Santos Brito Silva N, da Silva Souza A, de Souza Andrade H, Passos MRS, Castro CFB, Mendes-Junior CT, Mercuri RLV, Miller TLA, Buzzo JL, Rego FO, Araújo NM, Magalhães WCS, Mingroni-Netto RC, Borda V, Guio H, Rojas CP, Sanchez C, Caceres O, Dean M, Barreto ML, Lima-Costa MF, Horta BL, Tarazona-Santos E, Meyer D, Galante PAF, Guryev V, Castelli EC, Duarte YAO, Passos-Bueno MR, Zatz M. (2022). Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. Nat Commun. Mar 4;13(1):1004.

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

Nei, M. (1972). Genetic distances between populations. Am. Nat. 106:283-292.

Nell JV, Salzano FM (1967). Further Studies on the Xavante Indians.X. Some hypotheses generalizations resulting from these studies. *Am J Hum Genet* 19:554-74

Noah A Rosenberg, Jonathan T L Kang. (2015). Genetic Diversity and Societally Important Disparities, Genetics, Volume 201, Issue 1, 1 Pages 1–12.

Nunes K, Aguiar VRC, Silva M, Sena AC, de Oliveira DCM, Dinardo CL, Kehdy FSG, Tarazona-Santos E, Rocha VG, Carneiro-Proietti ABF, Loureiro P, Flor-Park MV, Maximo C, Kelly S, Custer B, Weir BS, Sabino EC, Porto LC, Meyer D. (2020). How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation in Admixed Brazilians. Front Immunol. 6;11:584950.

Oliveira, SF; Ferreira, LB; Klautau-Guimarães, MN; Ribeiro, GGBL; Simões, AL. (2006). Reconstrucción historica de poblaciones afro-descendientes aisladas de Brasil: el constraste entre las contribuciones masculina y feminina. Diversidad biológica y salud humana, Murcia – Sociedad Española de Antropologia, pag 203-210.

Oliveira, SF. (1999). Inserções de Alu em populações indígenas da Amazônia brasileira. Tese apresentada para obtenção do título de doutor, Faculdade de Medicina de Ribeirão Preto. Departamento de Genética, Universidade de São Paulo.

Oliveira, SF; Ferreira, LB. (2004). Biological view of the inexistence of human races.Eubios Journal Of Asian And International Bioethics, Ibaraki, 14(2): 60-63.

Olson, S. (2002). Mapping human history: genes, race, and our common origins. Mariner Books, New York/NY, 292p.

Palacín, L; Moraes, MAS. (1994). História de Goiás. Editora UCG, 6a edição, Goiânia/GO, 122p.

Palacín, L; Garciam, LF; Amado, J. (1995). História de Goiás em documentos: Colônia.Editora UFG, Goiânia/GO, 190p.

Paredes, M; Crespillo, M; Luque, JÁ; Valverde, JL. (2003a). STR frequencies for the PowerPlex 16 System Kit in a population from Northeast Spain. F. Sci. Int. 135:75-78.

Paredes, M; Galindo, A; Bernala, M; Ávila, S; Andrade, D; Vergara, C; Rincón, M; Romero, RE; Navarrete, M; Cárdenas, M; Ortega, J; Suarez, D; Cifuentes, A; Salas, A; Carracedo, A. (2003b). Analysis of the CODIS autosomal STR *loci* in four main Colombian regions. F. Sci. Int. 137:67–73.

Palacín, L; Garciam, LF; Amado J. (1995). História de Goiás em documentos: Colônia. Editora UFG. Goiânia- Go, 190p.

Parra, EJ; Kittles, RA; Argyropoulos, G; Pfaff, CL; Hiester, K; Bonilla, C; Sylvester, N; Parrish-Gause, D; Garvey, WT; Jin, L; McKeigue, PM; Kamboh, MI; Ferrell, RE; Pollitzer, WS; Shriver, MD. (2001). Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 114(1):18-29.

Parra, EJ; Marcini, A; Akey, J; Martinson, J; Batzer, MA; Cooper, R; Forrester, T; Allison, DB; Deka, R; Ferrell, RE; Shriver, MD. (1998). Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63(6):1839-51.

Parra, FC; Amado, RC; Lambertucci, JR; Rocha, J; Antunes, CM; Pena, SD. (2003). Color and genomic ancestry in Brazilians. Proc Natl Acad Sci USA 100(1):177-82.

Pena, SDJ. (2002). O retrato molecular do Brasil *in* Homo brasilis: Aspectos Genéticos, Linguísticos, Históricos e Socioantropológicos da Formação do Povo Brasileiro. 1ª ed. Ribeirão Preto: FUNPEC, volume 1, 191 p.

Pena, SDJ, Santos, FR, Tarazona-Santos, E (2020) Genetic admixture in Brazil. American Journal of Medical Genetics 184C:928–938.

Pena SD, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy Fde S, Kohlrausch F, Magno LA, Montenegro RC, Moraes MO, de Moraes ME, de Moraes MR, Ojopi EB, Perini JA, Racciopi C, Ribeiro-Dos-Santos AK, Rios-Santos F, Romano-Silva MA, Sortica VA, Suarez-Kurtz G. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. PLoS One. 2011 Feb 16;6(2):e17063. doi: 10.1371/journal.pone.0017063. PMID: 21359226; PMCID: PMC3040205.

Pfaff, CL; Parra, EJ; Bonilla; C; Hiester, K; McKeigue, PM; Kamboh, MI; Hutchinson, RG; Ferrell, RE; Boerwinkle, E; Shriver, MD. (2001). Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. Am J Hum Genet 68(1):198-207.

Pimentel, BJ; De Azevedo, DA; Silva, LA. (2004). Population genetics of eleven STR *loci* in the State of Sergipe, Northeastern Brazil. J Forensic Sci. 49(2): 49(2):402-3.

Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A; SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007 Dec;1(3-4):273-80. doi: 10.1016/j.fsigen.2007.06.008. Epub 2007 Aug 22. PMID: 19083773.

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

Ramachandran S, Deshpande O, Roseman C C, Rosenberg N A, Feldman M W et al. (2015). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA 102: 15942–15947.

Ramos, JS. (2002). O Brasil sob o paradigma racial: sociologia histórica de uma representação. In Pena, S.D.J. Homo brasilis. Editora Funpec, São Paulo/SP, p 131- 148.

Raymond, M; Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Heredity 86:248-249.

Reis, JJ. (2000). Presença negra: conflitos e encontros. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 79-100.

Reich, D. et al., (2012). Reconstructing Native American population history. Nature, pp. 370-375.

Reynolds, JB; Weir BS; Cockerham, CC. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767-779.

Ribeiro, D. (1996). Os índios e a civilização: a integração das populações indígenas no Brasil moderno. Companhia das Letras, São Paulo/SP, 559p.

Rincon, D (2009). Estudos de DNA mitocondrial em populações remanescentes de quilombo, 2009. Dissertação (Biologia Geral).Universidade de São Paulo

Rock, D. (1985). Argentina: 1516-1982. University of California Press, Los Angeles/California, 478p.

Rodrigues, EMR; Palha, TJBF; Santos, SEB. (2006). Allele frequencies data and statistic parameters for 13 STR *loci* in a population of the Brazilian Amazon Region. Forensic Science International 168(2-3):244-247.

Rodríguez, A; Arrieta, G; Sanou, I; Vargas, MC; García, O; Yurrebaso, I; Pérez, JÁ; Villalta, M; Espinoza, M. (2007). Population genetic data for 18 STR *loci* in Costa Rica. Forensic Science International 168:85-88.

Rohlf, FJ. (1992). NTSYS-pc Numerical taxonomy and multivariate analysis system.Version 1.70, New York: Applied Biostatistics.

Rosenberg N.A., Li L.M., Ward R., Pritchard J.K (2003). Informativeness of Genetic Markers for Inference of Ancestry. Am. J. Hum. Genet. 73:1402–1422.

Rowold, DJ; Herrera, RJ. (2003). Inferring recent human phylogenies using forensic STR technology. Forensic Sci. Int. 133(3):260-5.

Ruitberg, CM; Reeder, DJ; Butler JM. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 29(1): 320– 322.

Saitou, N; Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-25.

Salzano, MF. (2004). Interethnic variability and admixture in Latin América – social implications. Rev Biol Tropical. 52(3): 405-415.

Salzano, MF; Callegari-Jacques, SM. (1988). South americans indians – a case study in evolution. Oxford University Press, New York.

Sans, M. (2000). Admixture studies in Latin America: from the 20th to the 21st century.Hum Biol 72(1):155-77.

Santos, SEB; Guerreiro, JF; Salzano, FM; Weimer, TA; Hutz, MH; Franco, MHLP. (1987). Mobility, blood genetic traits and race mixture in the Amazonian population of Oriximiná. Revista Brasileira de Genética. 4:745-759.

Schröer, K; Schmitt, C; Staak, M. (2000). Analysis of the co-amplified STR *loci* D1S1656, D12S391 and D18S51: population data and validation study for a highly discriminating triplex-PCR. Forensic Sci. Int. 113(1-3):17-20.

Serre, D; Pääbo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. Genome Research 14:1679-1685.

Severson AL; Shortt JA, Mendez FL, Wojcik GL, Bustamante CD, Gignoux CR. (2020) SNAPPY: Single Nucleotide Assignment of Phylogenetic Parameters on the Y chromosome. Gignoux. bioRxiv 454736

Shriver, MD; Parra, EJ; Dios, S; Bonilla, C; Norton, H; Jovel, C; Pfaff, C; Jones, C; Massac, A; Cameron, N; Baron, A; Jackson, T; Argyropoulos, G; Jin, L; Hoggart, CJ; McKeigue, PM; Kittles, RA. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. Hum. Genet. 112(4):387-99.

Shriver, MD; Smith, MW; Jin, L; Marcini, A; Akey, JM; Deka, R; Ferrell, RE. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet. 60(4):957-64.

Silberstein, CF. (2000). A imigração espanhola na Argentina (1880-1930). In Fausto, B. Fazer a América. Ed. Universidade de São Paulo.

Silva, DA; Crouse, CA; Chakraborty, R; Góes, ACS; Carvalho, EF. (2004). Statistical analyses of 14 short tandem repeat *loci* in Brazilian populations from Rio de Janeiro and Mato Grosso do Sul states for forensic and identity testing purposes. Forensic Sci Int. 139:173–176.

Silva, EB; Dellalibera, E; Souza, M; Silva, RS; Maurício-da-Silva, L. (2002). Population genetics of eight STR *loci* - CSF1PO, TPOX, TH01, D16S359, D7S820, D13S317, F13B and LPL in a Brazilian population from the State of Piauí, Northeast Brazil. Forensic Sci. Int. 126:90-92.

Silva, FF; Dellalibera, E; Nigam, P; Mauricio-da-Silva, L; Santos Silva, R. (2003). Microsatellite markers in the population from Rio Grande do Norte, Northeastern Brazil. J Forensic Sci. 48(5):1189-90.

Silva, LAF; Pimentel, BJ; Azevedo, DA; Silva, ENP; Santos, SS. (2002). Allele frequencies of nine STR *loci*—D16S539, D7S820, D13S317, CSF1PO, TPOX, TH01, F13A01, FESFPS and vWA—in the population from Alagoas, northeastern Brazil. Forensic Sci Int. 130:187–188.

Silva, R; Moura-Neto, RS. (2004). Genetic diversity and admixture data on 11 STRs (F13B, TPOX, CSF1PO, F13A01, D7S820, LPL, TH01, vWA, D13S317, FESFPS,and D16S539) in a sample of Rio de Janeiro European-descendants population, Brazil. Forensic Sci. Int. 142:51–53.

Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet.;9(6):477-85.

Simmons, AD; Rodriguez-Arroyo, G; Rodríguez-Larralde, A. (2007). Admixture estimates base don ABO, Rh and nine STRs in two Venezuelan Regions. Annals of Hum Biol 34(1): 56-67.

Steinlechner, M; Schmidt, K; Kraft, HG; Utermann, G; Parson, W. (2002). Gabon black population data on the ten short tandem repeat *loci* D3S1358, VWA, D16S539, D2S1338, D8S1179, D21S11, D18S51, D19S433, TH01 and FGA. Int J Legal Med. 116 :176–178.

Tang, H; Choudhry, S; Mei, R; Morgan, M; Rodrigues-Cintron, W; Burchard, EG; Risch, NJ. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81:626-633.

Tofanelli, S; Boschi, I; Bertoneri, S; Coia, V; Taglioli, L; Franceschi, MG; Destro-Bisol, G; Pascali, V; Paoli, G. (2003). Variation at 16 STR *loci* in Rwandan (Hutu) and implications on profile frequency estimation in Bantu-speakers. Int J Legal Med. 117:121–126.

Tomás, G; Seco, L; Seixas, S; Faustino, P; Lavinha, J; Rocha, J. (2002). The peopling of São Tomé (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. Hum. Biol. 74(3):397-411.

Toscanini, U; Gusmão, L; Berardi, G; Amorim, A; Carracedo, A; Salas, A; Raimondi, E. (2007). Testing for genetic structure in different urban Argentinian populations. Forensic Sci Int. 165:35-40.

Tournamille, C; Le Van Kim, C; Gane, P; Cartron, JP; Colin, Y. (1995). Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism. Hum Genet. 95(4):407-10.

Underhill, PA. (2003) Inferring Human History: Clues from Y-Chromosome Haplotypes. Cold Spring Harbor Symposia on Quantitative Biology, Volume LXVIII. Cold Spring Harbor Laboratory Press 0-87969-709-1/04.

Vainfas, R. (2000). História indígena: 500 anos de despovoamento. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 35-59.

Venâncio, RP. (2000). Presença portuguesa de colonizadores a imigrantes. In Brasil 500 anos de povoamento, IBGE, Rio de Janeiro, p 61-77.

Venter JC, Adams MD, Myers EW et al (2001). The sequence of the human genome. Science.291:1304:1351.

Vieira, TC. (2003). Marcadores Alu em banco de dados da população do estado de Goiás.Dissertação de mestrado, Universidade Federal de Goiás.Wang, J; Caballero, A. (1999). Developments in predicting the effective size of subdivided populations. Heredity 82:212-226.

Wang, S; Lewis Jr, CM; Jakobsson, M; Ramachandran, S; Ray, N; Bedoya, G; Rojas, W; Parra, MV; Molina, JA; Gallo, C; Mazzotti, G; Poletti, G; Hill, K; Hurtado, AM; Salzano, FM. (2007). Genetic variation and population structure in Native Americans. PLOS Genetics, 3:2049-2067.

Wang, S; Ray, N; Rojas, W; Parra, MV; Bedoya, G; Gallo, C; Poletti, G; Mazzoti, G; Hill, K; Hurtado, AM; Camrena, B; Nicolini, H; Klitz, W; Barrantes, R; Molina, JA; Freimer, NB; Bortolini, MC; Salzano, FM; Petzl-Erler, ML; Tsuneto, LT; Dipierri, JE; Alfaro, EL; Bailliet, G; Bianchi, NO; Llop, E; Rothhammer, F; Excoffier, L; Ruiz-Linares, A. (2008). Geographic Patterns of Genome Admixture in Latin American Mestizos. PLoS Genet 4(3):e1000037.

Webb, GC; Coggan, M; Ichinose, A; Board, PG. (1989). Localization of the coagulation factor XIII B subunit gene (F13B) to chromosome bands 1q31-32.1 and restriction fragment length polymorphism at the locus. Human Genetics 81:157-160.

Weir, BS; Cockerham, CC. (1984). Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Whittle, MR; Romano, NL; Negreiros, VAC. (2004). Updated Brazilian genetic data, together with mutation rates, on 19 STR *loci*, including D10S1237. Forensic Sci. Int. 13:9207–10.

Yang-Feng, TL; Opdenak, G; Volckaert, G; Francke, U. (1986). Human tissue-type plasminogen activator gene located near chromosomal breakpoint in myeloproliferative disorder. Am J Hum Genet 39(1):79-87.

Zúñiga, J; Ilzarbe, M; Acunha-Alonzo, V; Rosetti, F; Herbert, Z; Romero, V; Almeciga, I; Clavijo, O; Stern, JNH; Granados, J; Fridkis-Hareli, M; Morrison, P; Azocar, J; Yunis, EJ. (2006). Allele frequencies for 15 autosomal STR *loci* and admixture estimates in Puerto Rican Americans. F Sci Int 164:266–270

Sítios da Internet

CODEPLAN: http:// www.codeplan.df.gov.br - Acessado em janeiro de 2019.

dbSNP NCBI database: http://www.ncbi.nlm.nih.gov/SNP - acessado em 15/02/2019

Funai: http://www.funai.gov.br - Acessado em maio de 2018.

Instituto Brasileiro de Geografia e Estatística (IBGE) PNAD 2007: http://www.ibge.gov.br/home/estatística/população - Acessado em dezembro 2018.

Universidade de Düsseldorf: http://www.uni-duesseldorf.de/WWW/MedFak/Serology/database.html - acessado em 03/02/2018

Governo do Distrito Federal:

www.distritofederal.df.gov.br. http://statgen.ncsu.edu/powermarker/ downloads.htm.

http://www.affymetrix.com/products_services/axiom_custom/axiom_mydesign.affx

http://www.bea.ki.se/documents/axiom_genotyping_solution_analysis_guide.pdf

http://taxonomy.zoology.gla.ac.uk/rod/rod.html.

http://www.spss.com.

https://www.guiaestudo.com.br/regiao-centro-oeste

APÊNDICES

APÊNDICE 1.

Termo de Consentimento Livre e Esclarecido

Universidade de Brasília

Instituto de Ciências Biológicas

Departamento de Genética e Morfologia

Laboratório de Genética Humana

Termo de Consentimento Livre e Esclarecido – TCLE

Convidamos o(a) Senhor(a) a participar voluntariamente projeto de pesquisa **"Reconstrução Histórica do Povoamento do Distrito Federal Utilizando como Ferramenta a Genética de Populações: Será o Distrito Federal a Representação Genética da População Brasileira?"**, sob a responsabilidade da pesquisadora Profa. Dra. Silviene Fabiana de Oliveira. O projeto visa contribuir para o entendimento do povoamento do Distrito Federal e da dinâmica de migração no Brasil. Por causa da história peculiar e recente do DF, a hipótese desse trabalho é de que sua população possa refletir a composição genética da população brasileira como um todo. Para testar essa hipótese, utilizaremos seu perfil genético, aliado ao perfil de cerca de 200 outros participantes de pesquisa do Distrito Federal, para acessarmos a ancestralidade genética da população do Centro-Oeste.

O (a) senhor(a) receberá todos os esclarecimentos necessários antes e no decorrer da pesquisa e lhe asseguramos que seu nome não aparecerá, sendo mantido o mais rigoroso sigilo pela omissão total de quaisquer informações que permitam identificá-lo(a).

A sua participação se dará de duas formas. Primeiro, será preenchido um questionário que nos permita acessar informações sobre o(a) senhor(a) e a origem geográfica de sua família. Em seguida, um pesquisador treinado coletará cerca de 8mL de sangue venoso de um de seus antebraços. O procedimento a ser usado é o padrão para a coleta de sangue venoso. A coleta será feita no Laboratório de Genética da Universidade de Brasília, em data e horário a serem combinados. Todo o procedimento será feito em uma única visita e levará cerca de meia hora (30 minutos).

Os riscos decorrentes de sua participação na pesquisa são dor no local de penetração da agulha, possibilidade de hematoma local, assim como pode haver dano psicológico no caso de desconforto em decorrência de alguma questão do questionário. Se você aceitar participar, estará contribuindo para o melhor conhecimento do povoamento do Distrito Federal e terá como benefício direto o acesso a sua ancestralidade genética.

O(a) senhor(a) pode se recusar a responder (ou participar de qualquer procedimento) qualquer questão que lhe traga constrangimento, podendo desistir de participar da pesquisa em qualquer momento sem nenhum prejuízo. Sua participação é voluntária, isto é, não há pagamento por sua colaboração.

Todas as despesas que o(a) senhor(a) tiver relacionada diretamente ao projeto de pesquisa com alimentação e/ou transporte poderão ser ressarcidas pelo pesquisador.

Caso haja algum dano direto ou indireto decorrente de sua participação na pesquisa, o(a) senhor(a) poderá ser indenizado, obedecendo-se as disposições legais vigentes no Brasil.

Os resultados da pesquisa serão divulgados na Universidade de Brasília e poderão ser publicados posteriormente em periódicos científicos, sempre mantendo a confidencialidade dos participantes. Os dados e materiais serão utilizados somente para esta pesquisa e ficarão sob a guarda do pesquisador por um período de cinco anos, após isso serão destruídos.

Se o(a) senhor(a) tiver qualquer dúvida em relação à pesquisa, por favor telefone para a Profa. Dra. Silviene Fabiana de Oliveira, no Laboratório de Genética da Universidade de Brasília. O número de telefone é (61) 3107 1947, e está disponível em horário comercial, inclusive para ligações a cobrar. O(a) senhor(a) pode também entrar em contato com a pesquisadora pelo e-mail silviene.oliveira@gmail.com.

Este projeto foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Ciências da Saúde (CEP/FS) da Universidade de Brasília. O CEP é composto por profissionais de diferentes áreas cuja função é defender os interesses dos participantes da pesquisa em sua integridade e dignidade e contribuir no desenvolvimento da pesquisa dentro de padrões éticos. As dúvidas com relação à assinatura do TCLE ou aos direitos do participante da pesquisa podem ser esclarecidos pelo telefone (61) 3107-1947 ou pelo e-mail cepfs@unb.br ou cepfsunb@gmail.com. O horário de atendimento é de 10:00hs às 12:00hs e de 13:30hs às 15:30hs, de segunda a sexta-feira. O CEP/FS se localiza na Faculdade de Ciências da Saúde, Campus Universitário Darcy Ribeiro, Universidade de Brasília, Asa Norte.

Caso concorde em participar, pedimos que assine este documento que foi elaborado em duas vias, uma ficará com o pesquisador responsável e a outra com o (a) senhor(a).

_____

nome e assinatura do participante

_____

Profa. Dra. Silviene Fabiana de Oliveira

Brasília,____/___/ _____

APÊNDICE 2

Questionário de acesso a dados demográficos

FICHA CADASTRAL PARA DOADORES DE MATERIAL BIOLÓGICO

PROJETO BRASÍLIA - 2017

DATA DA COLETA:_____/_____/_____

IDENTIFICAÇÃO

Nome:

Endereço:_____CEP:

Telefone:

e-mail:

Documento de identificação

tipo:_____número:_____data de expedição:___/_/_____

Estado civil:_____Escolaridade:

Nascimento:_____/_____/_____Idade na coleta:

Sexo (biológico): [ ]F []M

Local de nascimento no DF:

_ Residência dos pais ao nascimento:_____

FILIAÇÃO

mãe:

local de nascimento:

__tempo no DF:

__motivo da mudança:

ancestralidade:

pai:

local de nascimento:

__tempo no DF:

Motivo da mudança:

Ancestralidade:

Consaguinidade entre os pais:

materna:

local de nascimento:

__ancestralidade:

_ **avô** **materno:**

_ local de nascimento:

_ ancestralidade:

_ **consanguinidade entre os avós:**

_

avó **paterna:**

_ local de nascimento:

__ancestralidade:

avô **paterno:**

__local de nascimento:

__ancestralidade:

consanguinidade entre os avós:

_

CLASSIFICAÇÃO FENOTÍPICA E ANCESTRALIDADE

ancestralidade

Autoatribuída:

_

IBGE:

Preto [ ] Pardo[ ] Branco[ ] Amarela[ ]
Indígena [ ] Não declarada [ ]

Ancestralidade:

Atribuída pelo entrevistador:

_

IBGE:

Preto [ ] Pardo[ ] Branco[ ] Amarela[ ]
Indígena [ ] Não declarada [ ]

Irmãos

mulheres:_____

gêmeos: não[ ]

homens: _____

sim [ ] o doador [ ]
irmãos [ ]

univitelinos [ ]

bivitelinos [ ] sexo:_____

Filhos

número de gestações:

nascidos vivos:

natimortos:

abortos
espontâneos:

abortos espontâneos são recorrentes na família:

DOENÇAS GENÉTICAS NA FAMÍLIA:

doença	afetado
1.	
2.	
3.	
4.	

Eu,
_,acima identificado, autorizo os pesquisadores responsáveis a entrar em contato comigo no futuro para:

[ ] coletar material
[ ] confirmar ou esclarecer informações [ ] coletar informações adicionais

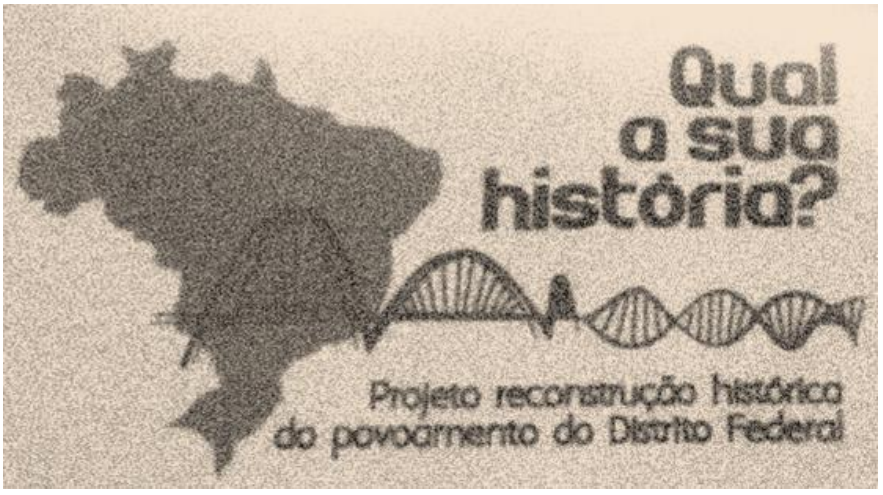
assinatura

Brasília,_____de_____de _____

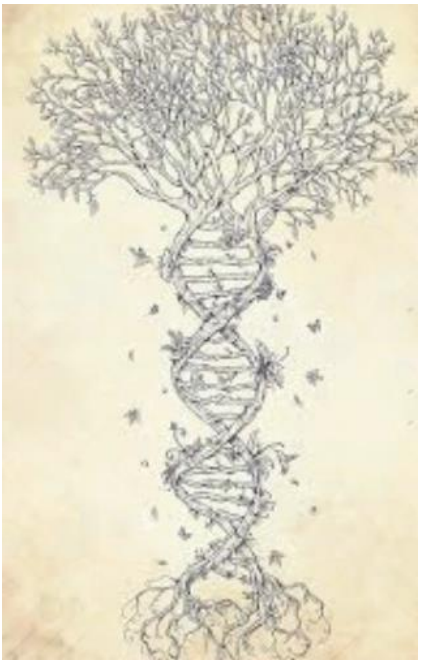
INFORMAÇÕES ADICIONAIS E OBSERVAÇÕES

APÊNDICE 3

Relatório Individual de Ancestralidade



“Sua história não começa por você. O primeiro passo da sua jornada começou na África, onde sua raiz ancestral foi plantada, de acordo com alguns autores, por volta de 300.000 anos; galhos desta árvore se espalharam pelo leste do mediterrâneo em torno de 100.000 anos. Por volta de 80.000 chegaram a China, 40.000 a Austrália e 20.000 na América, tendo se espalhando assim por todos os continentes da terra. Esses galhos deram frutos e sementes gerando novas raízes. As porcentagens de representatividade parental revelados em um teste de ancestralidade são mais que números... são raízes que em algum momento se fixaram em algum lugar, e quando descobertas, desenham um mosaico revelador da sua história genética”



http://www.cabarrusmagazine.com/2018/03/01/167893/dna-the-genetic-lottery

RELATÓRIO INDIVIDUAL DE ANCESTRALIDADE

Prezado/a **xxxxxxxxxxxxxxxxxxxx,**

Este documento é uma devolutiva acordada entre a senhora e o grupo de pesquisa em Ancestralidade e Reconstrução Histórica do Laboratório de Genética Humana da Universidade de Brasília quando da sua participação no projeto: **"RECONSTRUÇÃO HISTÓRICA DO POVOAMENTO DO DISTRITO FEDERAL UTILIZANDO COMO FERRAMENTA A GENÉTICA DE POPULAÇÕES: SERÁ O DF UMA REPRESENTAÇÃO GENÉTICA BRASILEIRA?"**.

É com satisfação que viemos através deste, entregar o resultado dos índices percentuais de representatividade genética das parentais Africana, Europeia, Asiática e Indígena Americana presentes no seu material genético. O conjunto destas porcentagens tem por objetivo proporcionar ao participante da pesquisa, um norteamento de suas origens ancestrais, aproximando o mesmo do conhecimento acerca de suas raízes ascendentes. É particular de cada um a interpretação desta informação. Neto e Santos, 2011, observa o fato de que o indivíduo pode perceber o resultado de um teste de ancestralidade como um agente produtor, redefinidor ou até mesmo desafiador de identidade, onde neste momento o indivíduo pode ser incitado a uma reorientação das perspectivas culturais, psicológicas, físicas e emocionais que o indivíduo tem sobre si mesmo e sobre os continentes de origem de seus antepassados.

Neste primeiro momento, o foco não está em uma análise ao nível subcontinental, com detalhes finos, ou seja, não é possível detalhar ainda as regiões específicas dentro de um continente em que o participante tem ascendência. Esta análise demanda acesso a grandes e específicos bancos de dados. O grupo espera que haja tempo e recursos para realizar tal análise e que esta possa ser acrescentada neste relatório em um segundo momento.

.....

CONTEXTUALIZAÇÃO E RELEVÂNCIA DO PARTICIPANTE DE PESQUISA

O Brasil teve um processo de mistura inusitado na história, gerando o brasileiro atual, o qual PENA (2002) chamou, irreverentemente, de *Homo brasilis*. O perfil genético brasileiro é visto como um dos mais miscigenados do mundo, com contribuições de origem europeia, africana, ameríndia,asiática e sírio libanesa, dentre outros, revelando um padrão de miscigenação pouco visto em outras partes do mundo.

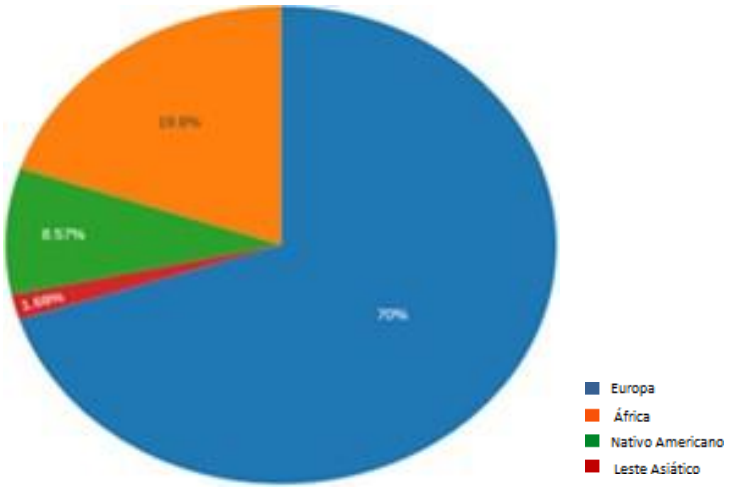
O povoamento do Distrito Federal, diferente das demais regiões brasileiras, se deu pela migração rápida de indivíduos provenientes de todas as regiões do país (em especial Sudeste, Nordeste e Centro-Oeste), onde as populações já vinham de um processo de miscigenação (CODEPLAN, 2013). Como consequência, essa população apresenta uma constituição genética especifica. Para estimar numericamente como essa população esta constituída, cada participante de pesquisa é peça fundamental nas estimativas de ancestralidade biológica de uma população. No caso do Distrito Federal temos uma primeira versão do perfil dessa população.

INFORMAÇÕES

Foi utilizada a técnica de SNPArray, tecnologia Axiom. O Axiom Genome-Wide Human Origins 1 *Array* é um painel exclusivo de genotipagem otimizado para geneticistas de populações que foi projetado em colaboração com os principais pesquisadores da área, exclusivamente para estudar história humana, migração e seleção natural. O *Array* possui marcadores de 11 populações humanas modernas, e conta com aproximadamente 600.000 marcadores todos

com poder de inferência parental. A tecnologia nos permitiu analisar os pontos de equivalência destes marcadores aos existentes no material genético do participante de pesquisa bem como sua porcentagem de correspondência, revelando um índice quantitativo de mistura existente em seu material genético.

RESULTADOS



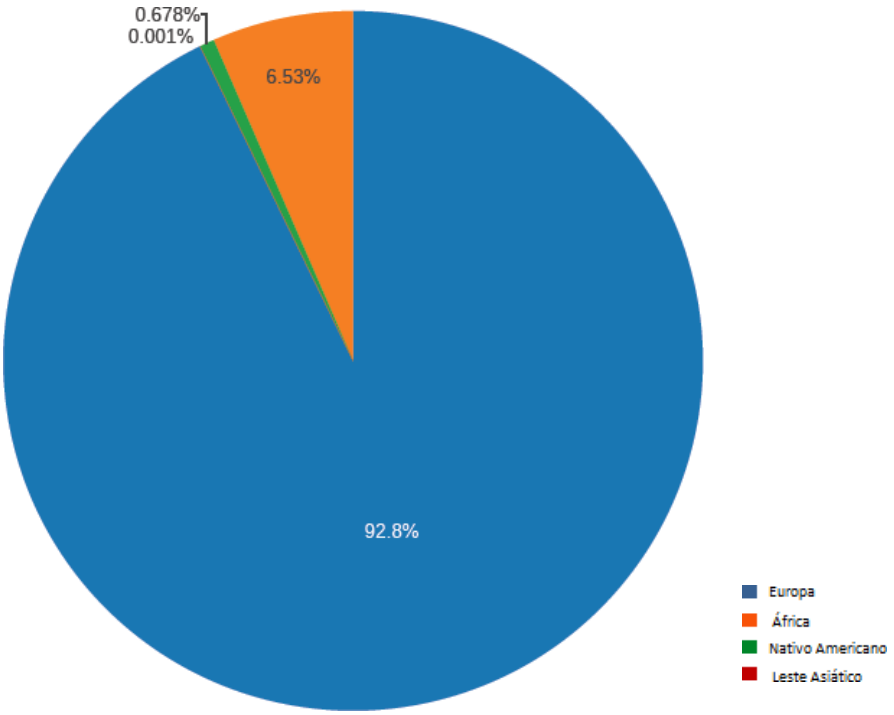
Perfil genético da população do Distrito Federal baseada na análise do *Axiom Genome -Wide Human Origins 1 Array.*

RESULTADOS

O padrão de mistura observado segue o modelo tri-hibrido brasileiro de mistura. Neste estudo, acrescentamos a observação da representatividade do componente asiático na população do DF. Analisada a porcentagem de contribuição dos quatro componentes principais considerados neste estudo, observou-se uma maior contribuição do componente europeu (70%), seguido por um valor mediano do componente africano (19,8%). A contribuição da parental nativo americana está representada por um percentual médio de 8,57% e em

menor expressividade dos quatro componentes parentais avaliados, está o Asiático (1,68%).

RESULTADOS



Seu perfil individual de ancestralidade genética baseada na análise *Axiom Genome-WideHuman Origins 1 Array*

RESULTADOS

Os índices percentuais de representatividade genética das parentais Africana, Europeia, Asiática e Indígena Americana presentes no seu material genético, revelam expressiva contribuição do componente europeu (92,8%), seguido do componente africano (6,53%). Em seguida nota-se uma singela contribuição dos componentes nativo americano (0,678%), e asiático (0,001%). Esse componente asiático pode ser um artefato e eventualmente pode ser somado ao componente indígena americano.

O laboratório de Genética Humana da Universidade de Brasília agradece a sua colaboração.

Dra. Silviene Fabiana de Oliveira dos Santos

MSc. Luciana Maia Escher