Universidade do Estado do Rio de Janeiro
Centro Biomédico
Instituto de Biologia Roberto Alcântara Gomes

Silvia de Oliveira Loiola Martins

Estruturação gênica e ancestralidade da população brasileira: estudo da população nativa da Ilha de Marajó com o emprego de marcadores de DNA

Rio de Janeiro
2022

Silvia de Oliveira Loiola Martins

Estruturação gênica e ancestralidade da população brasileira: estudo da população nativa da Ilha de Marajó com o emprego de marcadores de DNA

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Biociências, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Elizeu Fagundes de Carvalho

Coorientadora: Prof.ª Dra. Leonor Gusmão

Rio de Janeiro

2022

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CB-A

M386 Martins, Silvia de Oliveira Loiola.
Estruturação gênica e ancestralidade da população brasileira: estudo da população nativa da Ilha de Marajó com o emprego de marcadores de DNA / Silvia de Oliveira Loiola Martins. - 2020.
197f.

Orientador: Prof. Dr. Elizeu Fagundes de Carvalho
Coorientadora: Prof.ª Dra. Leonor Gusmão

Doutorado (Tese) - Universidade do Estado do Rio de Janeiro, Instituto de Biologia Roberto Alcântara Gomes. Pós-graduação em Biociências.

1. Hereditariedade – Teses. 2. Marajó, Ilha de (PA) – Teses. 3. Genética de populações – Teses. 4. Cromossomo Y - Teses. 5. DNA mitocondrial – Teses. I. Carvalho, Elizeu Fagundes de. II. Gusmão, Leonor. III. Universidade do Estado do Rio de Janeiro. Instituto de Biologia Roberto Alcântara Gomes. IV. Título.

CDU 575.1

Bibliotecária: Ana Rachel Fonseca de Oliveira
CRB7/6382

_____ _____

Assinatura Data

Silvia de Oliveira Loiola Martins

Estruturação gênica e ancestralidade da população brasileira: estudo da população nativa da Ilha de Marajó com o emprego de marcadores de DNA

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Biociências, da Universidade do Estado do Rio de Janeiro.

Aprovada em 11 de fevereiro de 2022.

Coorientadora: Profª Dra. Leonor Gusmão
Instituto de Biologia Roberto Alcântara Gomes – UERJ

Banca Examinadora: _____
Prof. Dr. Elizeu Fagundes de Carvalho (Orientador)
Instituto de Biologia Roberto Alcântara Gomes – UERJ

_____
Prof.ª Dra. Tatiana de Almeida Simão
Instituto de Biologia Roberto Alcântara Gomes – UERJ

_____
Dra. Fernanda Saloum de Neves Manta
Instituto Oswaldo Cruz

_____
Dra. Verônica Gomes
Universidade do Porto

Rio de Janeiro
2022

DEDICATÓRIA

Dedico este trabalho:

A todos os que se voluntariaram para que ele pudesse ser realizado.

A todos que contribuíram para a minha formação acadêmica.

A Deus, que me capacitou a chegar até aqui.

AGRADECIMENTOS

Agradeço sobretudo a Deus por sempre me sustentar, abençoando todas as áreas da minha vida. Porque dEle, por meio dEle e para Ele são todas as coisas. Dependo sempre dEle.

Aos meus pais, Itamar (*in memorian*) e Sueli, por todo amor, cuidado e incentivo. E por terem me direcionado ao caminho do estudo e do conhecimento.

À minha família como um todo por todo apoio e incentivo e todos os amigos que se alegram e me ajudam com suas orações.

Ao meu marido Joel, por ser um grande companheiro e incentivador, com sua calma e paciência enormes! Com você divido a vida, os momentos alegres que você tem o dom de potencializar e os momentos difíceis, nos quais você me traz calma, demonstrando apoio incondicional e o seu amor.

Ao Prof°. Dr. Elizeu Fagundes de Carvalho, por todo o apoio e oportunidade de aprendizado através das atividades no Laboratório de Diagnósticos por DNA, que me trouxe amadurecimento e crescimento profissional.

À Profª Dra. Leonor Gusmão, por todo o aprendizado, orientação e atenção, além do apoio e paciência sempre presentes durante o desenvolvimento deste projeto; e por ser uma pessoa inspiradora, demonstrando generosidade e desejo de promover o crescimento das pessoas.

À Profª. Drª. Dayse Aparecida da Silva, por todo apoio, ajuda, amizade e conversas sempre incentivadoras, que me trouxeram crescimento profissional e acadêmico.

Ao Professor Dr. Luiz Marcelo de Lima Pinheiro pela parceria neste projeto.

Ao Professor Dr. César Amaral por todo incentivo e aprendizado.

A todos os meus amigos técnicos e da área administrativa do LDD – Jaque, Martinha, Flávio, Paty, Laís, João, Cris, Andréa, Henrique e Ed (*in memorian).* Não tenho palavras para descrever o quanto são importantes no meu dia a dia! Obrigada por tudo...ajuda técnica, parceria, amizade, companhia, risadas e conversas!

À Filipa, por toda a atenção e contribuição no desenvolvimento desse trabalho.

A todos os alunos do LDD com quem aprendi e a quem pude ensinar. Especialmente aos alunos Ju Jannuzzi, July, Masinda, Rodrigo e Allan por compartilharem conhecimentos, por toda a ajuda e pela companhia em tantos momentos!

A todos que, de alguma maneira, contribuíram para a realização deste projeto, o meu muitoobrigada!

RESUMO

MARTINS, Silvia de Oliveira Loiola. **Estruturação gênica e ancestralidade da população brasileira:** estudo da população nativa da Ilha de Marajó com o emprego de marcadores de DNA. 2022. 197 f. Tese (Doutorado em Biociências) – Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

A população brasileira, altamente miscigenada, é resultante de três componentes ancestrais, de origens continentais europeia, africana e nativa americana. Diferentes padrões dessa mistura podem ser vistos ao longo do Brasil e trabalhos realizados com marcadores de linhagem e autossômicos têm demonstrado um predomínio da ancestralidade européia, mais acentuada nas regiões sul e sudeste, seguida pela ancestralidade africana, em maior proporção no nordeste, e pelo componente nativo, mais evidente na região Norte. Além das diferenças observadas entre as regiões geopolíticas do Brasil, existem evidências de subestrutura genética em uma mesma região. Na região Norte já foi visto que há diferenças entre populações urbanas, com maior contribuição europeia, e populações rurais, que têm maior ancestralidade nativa. Este trabalho tem como objetivo o estudo da população da Ilha de Marajó, no estado do Pará, na Região Norte, com o emprego de marcadores de DNA de origem uniparental (cromossomo Y e DNA mitocondrial - mtDNA) e com marcadores autossômicos informativos de ancestralidade (AIMs). Para investigar o *background* genético da população Ilha de Marajó, selecionou-se uma amostra de indivíduos não aparentados, com avós ali nascidos, dos quais 97 homens foram analisados para os 27 *loc*i Y- STRs (*Short Tandem Repeats*) do *kit Yfiler Plus* e 47 Y-SNPs (*Single Nucleotide Polymorphisms)*; a região controle do mtDNA de 101 indivíduos também foi sequenciada e 160 amostras foram genotipadas para o multiplex com 46 marcadores do tipo inserção-deleção (Indels). Resultados dos haplótipos Y-STR indicam uma diversidade haplotípica menor que o observado para populações urbanas de outras regiões (0,9959) e da região Norte, incluindo os Nativo americanos de São Gabriel da Cachoeira. As distâncias genéticas foram baixas entre o Marajó e todas as outras populações do Norte, mas alta com São Gabriel da Cachoeira, indicando uma predominância de linhagens masculinas europeias, o que foi confirmado pela proporções de haplogrupos determinados por Y-SNPs*, de* 71 % de linhagens européias, 17 % de africanas e 12 % de nativas. Esses dados foram corroborados por distâncias genéticas $F_{ST,}$ demonstrando a proximidade das linhagens masculinas brasileiras com o continente Europeu. Na análise do mtDNA foi observada uma diversidade haplotípica de 0,9897, inferior ao valor observado para outras populações urbanas brasileiras. A composição de haplogrupos revelou alta contribuição de linhagens maternas nativo americanas (62,9 %). Haplogrupos africanos representaram 36,1 % das amostras e apenas 1 % de origem europeia. A distribuição das proporções de ancestralidade biparental, com os AIM-Indels, foi de 40,3 % nativa americana, 35,4 % europeia e 24,3 % africana. Esses valores não se distanciam muito dos valores médios para os marcadores uniparentais nessa população, indicando baixo influxo de ancestralidade europeia recente, no âmbito de três gerações. Os resultados obtidos permitirão a criação de uma base de dados para a população miscigenada da Ilha de Marajó, sendo esta a primeira vez que foi avaliado o perfil de ancestralidade genética dessa população.

Palavras-chave: Ancestralidade. Ilha de Marajó. Cromossomo Y. mtDNA. AIMs. Indels.

ABSTRACT

MARTINS, Silvia de Oliveira Loiola. **Genetic structure and ancestry of the Brazilian population:** study of the native population of the Island of Marajó using DNA markers. 2022. 197 f. Tese (Doutorado em Biociências) – Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

Brazilian population is considered highly admixed and different patterns of the mixture between European, African and Native-american ancestry can be observed across the country. Predominance of European ancestry, more pronounced in the South and Southeast regions, followed by African ancestry, in greater proportion in the northeast, and finally the native component, more evident in the North region, have been demonstrated through data from autosomal and lineage markers. The Island of Marajó is in the north of Pará state, Northern Brazil. At the time of Portuguese colonization, it is believed that 30 different indigenous groups inhabited the Island.Currently it is composed by an admixed population, as a result of interethnic mating between European colonizers, Native Americans and African slaves. To investigate the genetic background of the Marajó population, a sample of unrelated individuals was selected, with matrilineal and / or patrilineal local ancestry for at least three generations. A total of 97 males were genotyped for the 27 Y chromosome STR included in the Yfiler Plus kit and 47 Y-SNPs. Additionally, 101 samples were sequenced for the complete mtDNA control region and 160 were genotyped for 46 autosomal insertion-deletion ancestry informative markers in a single multiplex reaction.The haplotype diversities obtained for Y-chromosome and mtDNA markers are lower (0.9959 and 0,9897, respectively) when compared with other Brazilian admixed populations, probably due to the geographic isolation of the island. The mtDNA haplogroup composition revealed a high contribution of Native American maternal ancestry of approximately 63 %, followed by 36 % of African haplogroups and only 1 % where from European origin. Conversely, 71% of male haplogroups, defined by Y-SNPs, were of European origin, 17 % sub-Saharan African and Native American ancestry was responsible for 12 %. In the study of the autosomal ancestry of the individuals, a 40,3 % of Native American ancestry was found, followed by 35.4 % of European and 24,3 % of African contributions. These values are close to the mean values for uniparental markers in this population, indicating a low influx of recent European ancestry within three generations. The results obtained will allow the creation of a database for the admixed population of Ilha de Marajó, this being the first time that the genetic ancestry profile of this population has been evaluated.

Keywords:**:** Ancestry. Marajó Island. Y chromosome. mtDNA. AIMs. Indels.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AMEL Y	gene da amelogenina
DNA	Ácido desoxirribonucléico
AIMs	*Ancestry Informative Markers*
BSA	Albumina Bovina Sérica
CAAE	Certificado de Apresentação de Apreciação Ética
CE	Eletroforese capilar
CEPH	*Center for the Study of Human Polymorphism*
CLUMPP	*CLUster Matching and Permutation Program*
Co.	*Company*
CR	Região controle do mtDNA
CRS	*Cambridge Reference Sequence*
ddNTP	Didesoxinucleotídeo trifosfato
del	Deleção
DIP	*Deletion-Insertion Polymorphism*
d-loop	*Displacement loop*
dNTP	Desoxinucleotídeo trifosfato
DTT	Ditiotreitol
EDTA	Ácido etilenodiaminotetracético
EDTANa2	Ácido etilenodiaminotetracético, sal dissódico
EMPOP	*EDNAP mtDNA population database*
F. A.	Frequência absoluta
F. R.	Frequência relativa
F_{ST}	Índice de Fixação
FTA	*Flinders Technology Associates*
FTDNA	*Family Tree DNA*
HD	Diversidade haplotípica
HGDP	*Human Genome Diversity Panel*
HV-I	Região hipervariável I
HV-II	Região hipervariável II
HV-III	Região hipervariável III

IBGE	Instituto Brasileiro de Geografia e Estatística
Indel	Marcador de inserção-deleção
ISFG	*International Society for Forensic Genetics*
ISOGG	*International Society of Genetic Genealogy*
LDD	Laboratório de Diagnósticos por DNA
LGA	Língua geral da Amazônia
MCMC	Markov Chain Monte Carlo
MDS	*Multidimensional Scaling*
MITOMAP	Base de dados de mtDNA *online*
MPS	*Massive parallel sequencing*
MSY	*Male-specificY chromosome region*
mtDNA	DNA mitocondrial
n	Número
NevGen	*Software* de predição de haplogrupos do cromossomo Y
NGS	*Next-generation Sequencing*
NRY	*Non-recombining Region of the Y chromosome*
PAR1	Região pseudoautossômica 1 do cromossomo Y
PAR2	Região pseudoautossômica 2 do cromossomo Y
PCA	Análise dos Componentes Principais
PCR	*Polimerase Chain Reaction*
PD	Poder de discriminação
pH	Potencial hidrogeniônico
rCRS	*revised Cambridge Reference Sequence*
SBE	*Single Base Extension*
RFU	Unidade relativa de fluorescência
RM	*Rapidly mutating*
rRNA	*RNA ribossomal*
SAP	*Shrimp Alkaline Phosphatase*
SBE	Single Base Extension
SDS	Dodecil-sulfato de sódio
SNP	*Single Nucleotide Polymorfism*
SRY	*Sex-determining region Y chromosome*
SSC	Tampão Citrato de Sódio

STR	*Short Tandem Repeats*
TAE	Tris-acetato-EDTA
TE	TRis-EDTA
Tris	Trihidroximetil aminometano
tRNA	RNA transportador
UERJ	Universidade do Estado do Rio de Janeiro
YHRD	*Y Chromosome Haplotype Reference Database*

LISTA DE SÍMBOLOS

Yp	Braço curto do cromossomo Y
Yq	Braço longo do cromossomo Y
=	Igual
mL	Mililitro
-	Menos
ºC	Grau Celsius
pb	Par de bases
kb	Mil bases
M	Concentração molar
mg	Miligrama
mM	Milimolar
mm2	Milímetro quadrado
µL	Microlitro
ηg	Nanograma
µM	Micromolar
δ	Frequência alélica diferencial
V/cm	Volts por centímetro
min	Minuto
x g	Gravidade
+	Mais
nm	Nanômetro
<	Menor
s	Segundo
cm	Centímetro
km	Quilômetro
h	Hora
ka	Mil anos
p	Nível de significância
π	Diversidade de sítios nucleotídicos
m	Metro

m2	Metro quadrado
A	Adenina
C	Citosina
G	Guanina
T	Timina
In	Índice de informatividade

SUMÁRIO

INTRODUÇÃO

A genética populacional e sua evolução

A genética populacional ou genética de populações é uma área da ciência fundamentada pelos conceitos de genética mendeliana que estuda a distribuição dos alelos e haplótipos nas populações e como suas frequências são mantidas ou alteradas ao longo das gerações, tanto no contexto de uma espécie quanto entre espécies. Engloba o estudo das variações genéticas que ocorrem naturalmente entre os organismos, e visa compreender os processos genético-demográficos atuantes (HARTL, 2008).

Os haplótipos correspondem a uma combinação de grupos de alelos de *loci* adjacentes, podendo ser formado por um ou mais alelos, ou até por um cromossomo inteiro.

A maior mudança no estudo da evolução pela genética nos últimos 50 anos foi a introdução de ferramentas moleculares para estudar as variações nas espécies, em vez de depender de variantes visíveis. A partir daí a genética de populações em nível molecular continua em crescente evolução (CHARLESWORTH; CHARLESWORTH, 2017). Em seus estágios iniciais, estudos pioneiros basearam-se em diferenças detectáveis por eletroforese, de variantes de proteínas solúveis, codificadas por diferentes alelos de um mesmo *locus* - as aloenzimas. Esses estudos fizeram uso da descoberta, feita apenas alguns anos antes, de que os genes codificam polipeptídeos, de modo que a variação nas sequências de proteínas detectadas pela variação em sua mobilidade em um campo elétrico pode ser equiparada à variação no próprio gene. As investigações foram realizadas tanto em humanos como também em outras espécies (HARRIS, 1966; LEWONTIN; HUBBY, 1966).

No entanto, já eram percebidas as limitações da técnica de eletroforese de proteínas para medir variações genéticas, por sua incapacidade de detectar duas categorias de variantes: alterações nas moléculas de proteínas que não afetam sua mobilidade em um gel e variantes na sequência de DNA que deixam a sequência protéica inalterada - os polimorfismos silenciosos ou sinônimos. Tais limitações, assim como o estudo de polimorfismos em regiões não-codificantes, só puderam ser superadas através de estudos da variação na sequência do DNA, que inicialmente foram realizados com enzimas de restrição. O mapeamento de sítios de restrição foi o passo inicial para o desenvolvimento de parâmetros estatísticos relacionados

à diversidade genética nas sequências de DNA, como a diversidade de sítios nucleotídicos (π) (NEI; LI, 1979).

Na década de 1980, diversos estudos forneceram os primeiros *insights* sobre padrões de variação nas sequências de DNA em populações de drosófilas, assim como a correlação positiva entre o nível de variabilidade encontrado e a taxa de recombinação. A recombinação é um recurso muito importante que resulta em maior diversidade, pois em cada geração permite a reorganização da informação genética, criando novas e únicas configurações (AGUADÉ; MIYASHITA**;** LANGLEY, 1989; BEGUN; AQUADRO, 1992). O primeiro estudo de genética populacional com o sequenciamento do DNA foi realizado com técnicas manuais de sequenciamento de Sanger (KREITMAN, 1983; SANGER; NICKLEN; COULSON, 1977), em populações de *D. melanogaster*. Esse estudo foi pioneiro, detectando diferentes tipos de variantes genéticas e ampliando o espectro que se tinha até o momento, revelando polimorfimos de base única ou SNPs (do inglês, *Single Nucleotide Polymorphisms*) silenciosos e fora de regiões codificantes, assim como polimorfismos do tipo inserção/deleção e sequências de homopolinucleotídeos.

Com o surgimento da técnica de PCR (do inglês, *Polymerase Chain Reaction*) e do sequenciamento automático no início da década de 1980, que possibilitaram a amplificação e sequenciamento de regiões específicas do genoma, estudos em larga escala evidenciaram inúmeras variações envolvendo um grande número de genes e espécies; sendo possível estimar novos níveis de diversidade genética. Tais estudos trouxeram ainda muitos outros dados, como diferenças dos níveis de variação entre diferentes *taxa*, além de mostrarem que os SNPs superam em número todos os outros tipos de variantes no DNA (LEFFLER *et al*., 2012; ROMIGUIER *et al*., 2014). A essa altura, devido à necessidade de compilação e de análises em larga escala dos dados obtidos a partir da grande quantidade de sequências geradas ao longo de décadas, já se observava o surgimento de bases de dados públicas, como o GenBank, que engloba sequências de nucleotídeos de mais de 300 mil espécies (BENSON *et al*., 2015); assim como o desenvolvimento de *software* que possibilitam o uso de diversas ferramentas estatísticas já disponíveis (CASILLAS; BARBADILLA, 2017).

Apesar da abundância de dados gerados sobre a diversidade genética, o conhecimento até esse ponto ainda se encontrava restrito a regiões do genoma já estudadas, e não à sua totalidade, o que possibilitaria uma medição mais completa e exata dos parâmetros avaliados. Sendo assim, avanços ainda maiores têm sido feitos visando à caracterização global da variação genética através do sequenciamento de genomas completos, e pode-se dizer que estamos atualmente na era da genômica populacional. Os primeiros trabalhos com

sequenciamento e análise de genomas completos surgiram durante a década passada, com o desenvolvimento das tecnologias de sequenciamento de nova geração ou NGS (do inglês, *Next Generation Sequencing*); também conhecido por MPS (do inglês, *Massive Parallel Sequencing*) (GOODWIN; MCPHERSON; MCCOMBIE, 2016; METZKER, 2010). Tais estudos, antes inviáveis economicamente pelo sequenciamento automatizado de Sanger, iniciaram-se com abordagens em larga escala de polimorfismos em múltiplos *loci* (BLACK *et al*., 2001; LUIKART *et al*., 2003), chegando a genomas completos de milhares de indivíduos de diversas espécies, incluindo a espécie humana (THE 1000 GENOMES PROJECT CONSORTIUM, 2010; THE 1000 GENOMES PROJECT CONSORTIUM , 2015; THE 1001 GENOMES PROJECT CONSORTIUM, 2016; BEGUN *et al*., 2007; FAWCETT *et al*., 2014). Esse alcance se tornou possível devido ao fato de que o MPS fornece dados genômicos populacionais a custos consideravelmente mais baixos, uma vez que permite sequenciar o genoma de grupos de indivíduos (*pool-seq*), e não somente indivíduos separados, como a técnica de Sanger (SCHLÖTTERER *et al*., 2014). Paralelamente à quantidade crescente de dados genômicos populacionais, estimativas cada vez mais precisas das taxas de recombinação ao longo do genoma também estão sendo obtidas, possibilitando inferências mais precisas sobre sua relevância na evolução do genoma (CASILLAS; BARBADILLA, 2017).

As informações provenientes de MPS diferem consideravelmente dos dados de variação obtidos na era das aloenzimas e sequências de Sanger, principalmente no volume desses dados. Sendo assim, nesse novo patamar de quantidade de sequências geradas, a Bioinformática tornou-se essencial para atender às necessidades específicas de todas as etapas do processo, desde a aquisição de dados, verificação de qualidade destes e sua análise e representação. Essa revolução genômica das últimas duas décadas e seu impacto sobre a genética populacional vem enriquecendo essa área com muitos novos conceitos, técnicas moleculares, formatos de dados e métodos estatísticos e computacionais. Sendo assim, modelos matemáticos aplicados na interpretação dos dados também foram evoluindo, primeiramente com estudos que demonstravam que a seleção natural seria uma potente força evolutiva, levando à variação nas espécies através da observação de polimorfismos “visíveis” (ALLISON, 1964). Essa observação passa pela formulação das principais teorias sobre a evolução e a variabilidade das espécies, como o modelo da hipótese “clássica” e o modelo da hipótese “balanceada”, que compõem a teoria selecionista. A diferença fundamental entre esses modelos reside no fato de na via clássica acreditar-se que a seleção agiria sobre genes que apresentavam apenas dois tipos de alelos - alelos funcionais (“selvagens”) e alelos

mutantes, estes considerados deletérios, presentes em frequências muito baixas (MULLER, 1950); enquanto na via balanceada acredita-se que muitos genes poderiam possuir dois ou mais alelos alternativos, com frequências intermediárias mantidas nas populações, por seleção balanceada (DOBZHANSKY, 1955). Alguns anos depois, foi demonstrado que além da seleção natural, e independente desta, a deriva genética aleatória, assim como as mutações neutras, ou seja, mutações cujos efeitos não influenciam a aptidão dos indivíduos, seriam responsáveis por grande parte da variabilidade nas populações, surgindo então a teoria neutralista (KIMURA; CROW, 1964).

A deriva genética pode ser definida como uma amostragem aleatória de gametas em cada geração de uma população finita, o que resulta em uma flutuação das frequências alélicas através das gerações e diminuição de variação genética. Esse fenômeno pode ser observado, por exemplo, quando uma população tem o tamanho drasticamente reduzido por um desastre natural (efeito gargalo) ou quando um pequeno grupo se separa da população principal para fundar uma nova colônia (efeito fundador), podendo resultar em perda de alguns alelos e fixação de outros, independente do caráter adaptativo dos mesmos. Em uma população panmítica idealizada, a força da deriva genética é inversamente proporcional ao tamanho populacional (BAMSHAD *et al*., 2004).

Passado alguns anos, propôs-se um refinamento da teoria neutralista, surgindo então a teoria "quase neutra". Nela se considerou que a maioria das mutações é levemente deletéria ou levemente benéfica, havendo um gradiente nos coeficientes de seleção que atuam sobre as mesmas (OHTA, 1973). Posteriormente, foi desenvolvido, a partir da teoria "quase neutra", um modelo que leva em conta a atuação balanceada tanto do coeficiente de seleção quanto da deriva genética sobre as mutações (OHTA; GILLEPIE, 1996). Esse modelo postula que mutações neutras são afetadas basicamente pela deriva genética, mas não pela seleção natural e que as mutações fortemente deletérias ou benéficas seriam mais afetadas pela seleção natural e não pela deriva. Já as mutações quase neutras seriam fixadas ou não em uma população, dependendo do balanço entre o coeficiente de seleção e o tamanho da população; quanto menor a população, maior o efeito da deriva e quanto maior a população, maior o efeito da seleção natural.

Os primeiros modelos teóricos na genética de populações aqui citados simularam a evolução ao longo do tempo, tentando entender como uma população evoluirá de um tempo passado ou presente para o futuro. Já a teoria da coalescência, surgida posteriormente, segue uma abordagem diferente, centrada na genealogia de uma amostra populacional, utilizando uma modelagem dos processos mutacionais no sentido presente $\rightarrow$ passado, ou seja, eventos

coalescentes são representados como uma genealogia dos genes (KINGMAN, 1982a, 1982b, 2000). Através do rastreamento de uma amostra atual de indivíduos de uma população, a fim de detectar os alelos de seus genes compartilhados por todos os seus membros, chega-se até uma única cópia ancestral, conhecida como o ancestral comum mais recente ou MRCA (do inglês, *Most Recent Common Ancestor*). Após alguns refinamentos, esse modelo retrospectivo simplificou consideravelmente as análises baseadas em modelos neutros, permitindo relacionar a diversidade genética observada em uma amostra com a história demográfica da população da qual foi extraída (CASILLAS; BARBADILLA, 2017).

O principal objetivo da genômica populacional é a descrição e interpretação da variação genética dentro e entre populações, mas podemos dizer que as atuais abordagens tecnológicas revolucionaram esse campo, impulsionando tanto o trabalho empírico quanto suas fundamentações teóricas. Ainda assim, as forças fundamentais da evolução permanecem como os fatores explicativos essenciais (CHARLESWORTH, 2010; LYNCH, 2007). As mutações, consideradas a fonte primária de variabilidade genética nas populações, ocorrem espontaneamente ao longo de todo o genoma. Mas por apresentarem taxas muito baixas, não respondem pelas mudanças rápidas no *genepool* de uma população e pelo nível de variabilidade observado nas populações. Sendo assim, a diversidade genética introduzida por mutações é posteriormente moldada e distribuída por outros processos, tais como deriva gênica, migrações, seleção natural, recombinação, e fluxo gênico. Em conjunto, tais processos, que envolvem a fundação e dispersão de novas populações, e favorecem os cruzamentos entre indivíduos de um mesmo grupo em detrimento a grupos separados geograficamente, levaram à distribuição de diferentes polimorfismos e polimorfismos com diferentes frequências entre populações. Tais diferenças contribuíram para a variabilidade genética que se observa atualmente e na sua distribuição ao nível mundial (AMORIM; BUDOWLE, 2016; BAMSHAD *et al*., 2004; HOLSINGER; WEIR, 2009).

Sabe-se que, em média, dois humanos selecionados ao acaso diferem em aproximadamente três milhões de nucleotídeos, e que a maioria desses polimorfismos é neutra ou quase neutra. A distribuição desses polimorfismos, associada ao conhecimento dos padrões de recombinação e das taxas de mutação, ao longo do genoma, refletem a história demográfica de nossa espécie, permitindo o rastreamento a ancestrais que podem ter vivido em diferentes épocas e partes do mundo, visto que as informações contidas na molécula de DNA são um tipo de registro da história da espécie humana. Desta forma, muitos estudos têm sido realizados com populações atuais, trazendo inferências sobre aspectos histórico-demográficos e evolutivos dos diferentes grupos humanos, corroborando com os registros

históricos já existentes e contribuindo com novas informações. Nomeadamente, locais de origem e rotas de migração ocorridas ao longo do tempo e a compreensão de seu impacto na complexa diversidade mundial contemporânea (CAVALLI-SFORZA, 1998; HARPENDING; ROGERS, 2000; JOBLING; HURLES; TYLER-SMITH, 2003).

As migrações humanas e a origem da população moderna

As migrações consistem no deslocamento de indivíduos de uma área geográfica para outra, que, se já for habitada por indivíduos de um grupo distinto, poderá resultar em um processo de miscigenação e consequente formação de uma população com uma composição genética distinta (dita miscigenada), uma vez que haverá fluxo gênico entre os grupos populacionais envolvidos (JOBLING; HURLES; TYLER-SMITH, 2003).

Atualmente, a teoria mais aceita para a formação da população moderna indica que a origem dos primeiros grupos populacionais humanos se deu na África, há mais de 100 ka (BAMSHAD *et al*., 2004; FORSTER, 2004; PENA *et al*., 2009). Estudos realizados com o DNA mitocondrial (mtDNA) demonstram que as linhagens maternas de todos os humanos originam-se na "Eva mitocondrial", uma mulher nascida há mais de 130 ka no sul ou leste africanos. Evidências genéticas e arqueológicas revelam que há cerca de 100 ka ocorreu inicialmente uma modesta expansão humana em toda a África, seguida de uma grande reexpansão, entre 60-80 ka atrás, na qual os humanos se dispersaram da África para colonizar outras partes do mundo. Provavelmente essa expansão foi possível por ter coincidido com uma diminuição do nível do mar, nunca ocorrida, e que abriu uma rota através do mar vermelho para o Yemen. Acredita-se que o grupo que saiu da África, nessa que ficou estabelecida como a primeira migração moderna bem-sucedida "Out of África", era de tamanho reduzido. Tal conclusão advém do fato de apenas um tipo de linhagem do mtDNA, dentre as linhagens existentes àquela época, ser encontrada fora da África (linhagens L3), da qual todos os não-africanos descendem atualmente. Sendo assim, os primeiros europeus e asiáticos foram originados desse pequeno grupo de migrantes, que se dividiram em duas rotas principais: as rotas costeiras para a Austrália e Papua Nova Guiné, sendo as mais rápidas, em comparação com as rotas com condições climáticas mais severas no Norte, que levavam à Eurásia. Algumas linhagens de mtDNA, que se acredita terem saído da África ou surgido imediatamente após a migração, por exemplo (linhagens M), estão ausentes na população

europeia atual, provavelmente devido a efeitos de deriva mais severos nessas rotas. Com a estabilização das condições climáticas da última Idade do gelo por volta de 30 ka atrás, as populações de Homo sapiens locais cresceram e expandiram intensamente, pela Europa e norte da Ásia, havendo o surgimento de linhagens descendentes fundadoras, conhecidas como haplogrupos, e que atualmente tem distribuição continental ou regional. A partir da Ásia mais uma etapa da expansão humana acontece, levando ao povoamento das Américas, há aproximadamente 25 ka atrás. A teoria mais aceita para o povoamento das Américas relata que o nível do mar era cerca de 120 m inferior ao que é hoje, e a América do Norte e o continente Asiático estavam conectados pela ampla ponte terrestre da Beringia, atualmente submersa. Registros antropológicos e análises genéticas indicam que, através dessa passagem, um pequeno grupo de asiáticos do norte da Sibéria migrou para o continente americano (BONATTO; SALZANO, 1997; FORSTER, 2004; RAY *et al*., 2010). Durante o último período glacial, entretanto, alguns milênios após a chegada desses fundadores, a grande extensão dos mantos de gelo glacial comprometeu a passagem da Beríngia, interrompendo as migrações da Ásia em direção à América do Norte. Embora haja um consenso sobre a origem asiática das linhagens fundadoras na colonização das Américas, existem diferentes hipóteses, baseadas em achados arqueológicos, linguísticos e dados genéticos obtidos de marcadores diversos, sobre o tempo que os fundadores levaram para povoar o continente, assim como sobre o número de ondas migratórias que ocorreram. A maioria dos estudos aponta para uma única migração (BONATTO; SALZANO, 1997; FORSTER *et al*., 1996; MERRIWETHER; ROTHHAMMER; FERREL, 1994), enquanto outros sugerem que uma única e discreta onda de colonização é altamente inconsistente com os níveis observados de diversidade genética, chegando a sugerir três ondas, além da possibilidade de fluxo gênico recorrente entre a Ásia e a América, após a colonização inicial (GREENBERG; TURNER II; ZEGURA, 1986; RAY *et al*., 2010; REICH *et al*., 2012; TAMM *et al*., 2007). Uma das hipóteses sobre o povoamento do continente, chamada de Modelo de Incubação na Beríngia, considera que antes de povoarem o continente como um todo, os fundadores se estabeleceram na Beríngia por aproximadamente 15 ka, isolados por uma barreira geográfica de gelo. Daí surgindo então, uma diversidade que não se observava na população fundadora, antes mesmo de espalharem-se por todo o continente americano. A outra hipótese, chamada Modelo de Colonização Direta, propõe que uma parte dos ancestrais sobreviveu na Beringia, ao ficarem isolados, pelo gelo, da maior parte dos imigrantes, que se moveu mais rapidamente ao longo do continente, do Norte para o Sul. Mais tarde, a população da Beríngia se diferenciaria nos grupos linguísticos “Eskimo-Aleut” e “Na-Dene”, que ficaram restritos à América do Norte; e os que

se dispersaram em direção à América Central e América do Sul originaram a família lingüística dos ameríndios. No decorrer dos milênios seguintes, a costa atlântica do continente americano, assim como o interior foram sendo percorridos e povoados por inúmeros povos ameríndios ou indígenas, que estavam em constante migração e adaptação (BONATTO; SALZANO, 1997; FORSTER, 2004; RAY *et al*., 2010; RIBEIRO, 1995; TAMM *et al*., 2007).

A maioria das populações atuais da América Latina, incluindo a população brasileira, apresenta peculiaridades em relação à sua composição genética, devido ao fato de terem sido alvo de vários fluxos migratórios. Estima-se que cerca de quarenta e cinco milhões de nativos habitavam na América Latina à época do início da vinda de europeus, há cerca de cinco séculos, embora haja projeções que variam de trinta a noventa milhões de ameríndios. Porém houve um rápido decréscimo dessa população e até mesmo a extinção de grupos menores, devido a epidemias e massacres ocorridos na chegada dos invasores (SALZANO; BORTOLINI, 2002; SALZANO; SANS, 2014). As estimativas sobre o número de europeus são muito variáveis, de cerca de quatro a cinquenta milhões de pessoas, em várias ondas, sendo a última delas relativamente recente, após a Segunda Guerra Mundial. Inicialmente vieram principalmente espanhóis e portugueses, entre os séculos XV e XIX, e durante esse período, tais conquistadores europeus trouxeram de diferentes regiões da África subsaariana mais de nove milhões de escravos para trabalhar em suas lavouras, com diversas viagens, sendo cerca de quatro milhões destes apenas para o Brasil. Nos séculos mais recentes alguns países também receberam populações de origem asiática, como os imigrantes japoneses que vieram para São Paulo e Pará, no Brasil, ou as populações de origem chinesa, no Peru. Consequentemente, com tantas contribuições para a diversidade latino americana, pelas diferentes migrações e pelo fluxo gênico envolvendo diferentes ancestralidades, observa-se atualmente uma grande heterogeneidade da ancestralidade entre os países, assim como dentro destes. Porém, pode-se afirmar que essas populações híbridas formaram-se majoritariamente pela contribuição de três componentes genéticos continentais: nativo americano ou ameríndio, europeu e africano. Sendo assim, na América Latina como um todo, e em populações miscigenadas espalhadas pelo mundo, sabe-se que, como reflexo de sua história demográfica, os cromossomos das pessoas que integram essas populações mostram uma complexa mistura de ascendências. Mistura tal que depende tanto do número de gerações de cruzamentos como do padrão de cruzamentos que acontecem, uma vez que as pessoas pertencem a diferentes extratos socioeconômicos (RUIZ-LINARES *et al*., 2014; SALZANO; SANS, 2014).

Marcadores genéticos em estudos forenses e populacionais

O genoma humano, isto é, a informação genética contida no interior das células e transmitida ao longo das gerações, compreende o genoma nuclear, que contém a maior parte da informação genética, e o genoma mitocondrial (mtDNA), cada um dos quais com características distintas. O genoma nuclear, tal como o nome indica, está localizado no interior do núcleo das células dos seres eucariontes. É organizado em cromossomos, sendo que cada célula é composta por 22 pares de autossomos e dois cromossomos sexuais. O cromossomo sexual X está presente em duas cópias nas mulheres (XX) e apenas uma cópia nos homens, que apresentam também um cromossomo Y (XY). Já o mtDNA encontra-se no interior das mitocôndrias, organelas citoplasmáticas responsáveis pela produção de energia, localizadas em células animais eucarióticas. Cada mitocôndria tem de centenas a milhares de cópias do mtDNA, o que representa um pequeno percentual do conteúdo genômico total de uma célula (BUTLER, 2005).

Os polimorfismos são variações na sequência de DNA entre indivíduos, ao nível de um *locus*, ou seja, o local fixo em um cromossomo onde está localizado determinado gene ou marcador genético, Sendo assim, um *locus* é dito polimórfico sempre que nele existirem pelo menos duas variantes (alelos) com frequências superiores a 1 % na população analisada (GELEHRTER; COLLINS, 1990). Os *loci* polimórficos são considerados úteis em diversas aplicações, como estudos populacionais, forenses e genealógicos, visto que possibilitam a diferenciação entre indivíduos e a identificação humana, sendo nomeados de marcadores genéticos (BUTLER, 2005).

Qualquer região do genoma, seja ela codificante ou não-codificante, pode ser rastreada para variantes genéticas, mas na prática, para estudos genéticos populacionais e forenses, marcadores não codificantes são preferíveis, uma vez que não estão diretamente sujeitos aos efeitos da seleção e, portanto, espera-se que reflitam principalmente efeitos neutros em nível populacional, como deriva, expansões, miscigenação e migração (STEIPER, 2010). A escolha de marcadores neutros no campo forense também é recomendada para evitar implicações éticas, uma vez que são menos propensos a divulgar informações associadas a doenças ou suscetibilidade genética (SCHNEIDER, 1997). No entanto, este critério vem sendo reconsiderado com o advento das tecnologias de maior desempenho, como sistemas MPS, que permitem a coleta de informações de grandes porções do genoma. Com o crescente volume de dados genéticos disponíveis, marcadores de ancestralidade, assim como de características

fenotípicas, estão cada vez mais sendo usados em investigações forenses, para obter informações sobre potenciais criminosos sem suspeitos identificados (KAYSER; DE KNIJFF, 2011; PHILLIPS *et al*., 2014a).

Os marcadores genéticos podem apresentar diferenças entre indivíduos tanto na sequência de nucleotídeos, assim como no comprimento de segmentos de DNA. Os polimorfismos de sequência são conhecidos como SNPs, e os polimorfismos de comprimento, que incluem os marcadores de inserção-deleção bialélicos ou Indels, assim como marcadores multialélicos, como os STRs (do inglês, *Short Tandem Repeats*) e outros tipos de variações. Alguns sistemas de classificação consideram os STRs como marcadores indels multialélicos (BUTLER, 2005; WEBER *et al*., 2002).

A seleção de marcadores a serem utilizados, no contexto forense ou da genética populacional, dependerá da demanda requerida, uma vez que diferentes cenários podem ocorrer. Em cada caso, a seleção dos marcadores genéticos apropriados está, em última análise, associado à sua variação intrapopulacional e /ou interpopulacional, assim como ao seu modo de transmissão e também a aspectos técnicos, que são muito relacionados às características das amostras biológicas disponíveis (AMORIM; BUDOWLE, 2016). Marcadores com altas taxas de mutação são muito polimórficos e úteis em genética populacional, para estudar processos recentes; e na genética forense, para identificação humana. Por outro lado, Para o estudo de eventos passados, marcadores de evolução lenta são a escolha mais adequada (SCHLOTTERER, 2000; STEIPER, 2010). Aproveitando sua abundância no genoma, bem como a facilidade de detecção, os marcadores de mutação lenta também podem ser usados em análises forenses para inferir a origem étnica de um indivíduo, por exemplo, uma vez que uma menor recorrência tende a produzir variações de frequência de alelos que refletem uma distribuição geográfica estruturada. Neste trabalho foram usados marcadores genéticos STRs e os Indels, que são polimorfismos de comprimentos; além dos SNPs, que são polimorfismos de variação de sequência.

Marcadores STR - *Short Tandem Repeats*

Dentre os principais marcadores genéticos usados nas últimas décadas estão os marcadores do tipo STR, também conhecidos como microssatélites, que estão entre os polimorfismos mais abundantes do genoma humano, ocorrendo ao longo de todos os

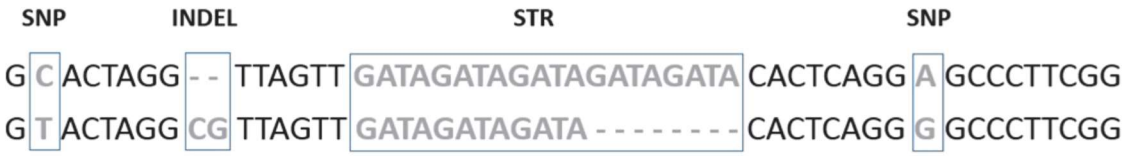
cromossomos a uma razão de 1 a cada 15000 pares de bases (pb), sobretudo em regiões não codificantes. Os STRs são trechos da molécula de DNA compostos de repetições consecutivas (*in tandem*), formadas por unidades de 2 a 6 pb (Figura 1). São caracterizados por variações no comprimento da sequência de DNA devidas a diferenças no número de repetições (ELLEGREN, 2004).

As diferentes variantes alélicas de uma região microssatélite são nomeadas conforme o número de unidades de repetição que formam a sequência. Desta maneira, se um STR é constituído por cinco unidades de repetição, o mesmo apresenta o alelo 5. É importante ressaltar que nem todos os alelos são formados por unidades de repetições completas, sendo nesse caso denominados alelos microvariantes. De acordo com sua estrutura, os STRs podem ser classificados em simples, compostos ou complexos. Os STRs de estrutura simples contêm um único tipo de sequência repetitiva, com sequências e comprimentos idênticos. Os compostos possuem mais de um tipo de sequência repetitiva, enquanto os complexos podem conter diversos blocos repetitivos, de comprimento variável, intermeados por sequências mais ou menos variáveis (URQUHART *et al*, 1994).

Sabe-se que STRs são gerados por mutações aleatórias e que o ganho ou a perda de unidades de repetição ocorrem devido a deslizamentos da enzima polimerase durante a replicação de DNA. As taxas médias de mutação dos STRs são de cerca de 2×10^{-3} (BRINKMANN *et al*., 1998; SCHLOTTERER, 2000), o que faz com que sejam altamente polimórficos quando comparados a outros tipos de polimorfismos, como os Indels e os SNPs. Devido a isso apresentam alelos múltiplos, além de alta heterozigosidade. A complexidade estrutural do STR influencia sua taxa de mutação, sendo os STRs com repetições com 4 nucleotídeos considerados bastante estáveis e por isso, bastante abundantes. Por isso têm sido preferidos ao projetar os *kit*s forenses disponíveis comercialmente, embora repetições maiores apresentem a vantagem de apresentar com menor frequência artefatos de análise do tipo picos *stutter*. Esses artefatos ocorrem *in vitro*, durante a amplificação do fragmento de DNA por PCR, pelo mesmo mecanismo de deslizamento da enzima polimerase. De fato, são produtos da PCR presentes em quantidades menores do que o alelo verdadeiro e diferem em tamanho do produto principal por múltiplos do comprimento da unidade de repetição. Esses artefatos de PCR às vezes podem complicar a análise de STRs, especialmente em situações de pouco DNA molde ou baixa qualidade das amostras, bem como em misturas não balanceadas de dois ou mais contribuintes, mas na maioria das situações eles podem ser facilmente identificados (AMORIM; BUDOWLE, 2016).

Em relação à metodologia de análise, os STRs podem ser analisados a partir de fragmentos de PCR relativamente curtos (100 a 400 pb), sendo possível a amplificação de vários marcadores em uma única reação de PCR multiplex. Os *primers* ou iniciadores referentes a cada região de interesse são marcados com diferentes fluoróforos e os fragmentos gerados são separados de acordo com seu tamanho, durante a eletroforese capilar. Essa simplicidade da técnica de genotipagem, se comparada com outras metodologias, juntamente com a abundância dos STRs no genoma humano e com os altos níveis de diversidade intrapopulacional esperados para esses *loci* de mutação rápida, tornaram esses marcadores amplamente utilizados em laboratórios forenses em todo o mundo. A análise de STRs tornou-se o principal método de escolha para avaliar a diversidade genética em populações ou para realizar a identificação individual, através de uma diversidade de *kit*s comerciais disponíveis para tipagem nos autossomos e nos cromossomos sexuais, e bancos de dados de referência têm sido criados e atualizados, permitindo a comparação entre laboratórios (AMORIM; BUDOWLE, 2016; BUDOWLE; DAAL, 2008).

Figura 1 **-** Exemplos de diferentes tipos de marcadores genéticos: SNP, Indel e STR



Fonte: AMORIM; BUDOWLE, 2016.

Marcadores SNP – *Single Nucleotide Polymorphisms*

Os marcadores SNPs são substituições de um único nucleotídeo que ocorrem em diversas posições no genoma (Figura 1). Alguns autores incluem inserções ou deleções de base única na classe dos SNPs (BUDOWLE; DAAL, 2008), enquanto outros as consideram como polimorfismos do tipo indels (AMORIM; BUDOWLE, 2016).

Os SNPs são considerados os polimorfismos mais abundantes do genoma, representando cerca de 85 % da variação genética humana. O número de SNPs já reportados em humanos é de aproximadamente 38 milhões, o que corresponde a uma densidade de um SNP a cada 80 pb, em média (THE 1000 GENOMES PROJECT CONSORTIUM, 2012).

Teoricamente, qualquer um dos quatro nucleotídeos existentes pode ocorrer em cada posição da sequência de DNA, embora a grande maioria dos SNPs seja um marcador bialélico. Isso pode ser explicado devido à baixa taxa de mutação desses marcadores, que torna a probabilidade de duas ou mais mutações independentes ocorrerem em uma mesma posição ser muito baixa. Ainda assim, SNPs tri-alélicos e tetra-alélicos já foram relatados, e são muito úteis na detecção de misturas em análises forenses (VIGNAL *et al*., 2002; WESTEN *et al*., 2009).

A taxa de mutação dos SNPs é da ordem de 10^{-8} (NACHMAN; CROWELL, 2000), sendo, por isso, considerados bastante estáveis quando comparados aos STRs e muito úteis também em análises de parentesco e na identificação de indivíduos. Mas apesar de serem mais abundantes, o que também seria uma vantagem, os SNPs não são mais utilizados do que os STRs, e isso se deve em parte ao seu baixo poder de discriminação (PD) por *locus*. O fato dos SNPs, em sua grande maioria, serem bialélicos, diminui seu nível de informatividade isoladamente, sendo necessários aproximadamente 50 SNPs para atingir um valor de PD de 12 a 15 STRs (AMORIM; PEREIRA, 2005). Outra limitação para essas aplicações é que não existem bancos de dados populacionais disponíveis para SNPs como vemos para os STRs (KAYSER; DE KNIJFF, 2011).

Dependendo da localização no genoma e das diferenças nas frequências alélicas entre as populações humanas, os SNPs podem estar associados com populações específicas ou correlacionados com características físicas, como cor de olhos, de pele e de cabelo, por exemplo. Daí esses marcadores serem usados também para inferências fenotípicas e de ancestralidade (BUDOWLE; DAAL, 2008). Na área forense, análises de DNA que possam apontar a possível origem ancestral, assim como características físicas visíveis de um criminoso totalmente desconhecido na investigação, podem ser de extrema importância na solução de um caso. Sendo assim, painéis de SNPs selecionados de acordo com finalidades específicas (identificação, inferência fenotípica, inferência de ancestralidade) vêm sendo desenvolvidos e dados populacionais têm sido gerados a partir desses, em diversas populações (AMORIM; BUDOWLE, 2016; GETTINGS *et al*., 2014).

Em relação às técnicas usadas para genotipagem de SNPs, uma vantagem é que podem ser analisados com o mesmo tipo de equipamento comumente usado para os STRs – a eletroforese capilar. Além disso, muitas das dificuldades encontradas no uso de STRs tradicionais não existem para SNPs, como a análise de amostras muito degradadas. Para esse tipo de amostra, o uso de amplicons de PCR mais curtos, de 50 pb ou menos, por exemplo, é essencial para a obtenção de um resultado bem-sucedido. Fragmentos menores, no entanto,

não são possíveis com STRs altamente polimórficos devido à sua sequência repetitiva, mas podem ser empregados com os SNPs, que têm alterações de uma única posição. Nos SNPs também não ocorrem os artefatos de análise do tipo *stutter*, comuns nos STRs (KAYSER; DE KNIJFF, 2011).

Marcadores Indel – Inserção / Deleção

Indels são polimorfismos de comprimento caracterizados pela inserção ou deleção de um ou mais nucleotídeos (Figura 1) e em sua grande maioria são bialélicos, com comprimentos variando entre 3 a 15 pb; embora já tenha se descrito indels de 1 a milhares de pares de bases. Os indels, assim como os STRs e os SNPs, estão entre os polimorfismos mais frequentes do genoma, apresentando uma frequência intermediária entre eles, de 1 a cada 7200 pb e ocorrendo nos 24 cromossomos. Aproximadamente 36 % dos indels estão localizados em regiões promotoras, assim como em introns e exons, podendo causar alguma alteração deletéria na função gênica (MILLS *et al*., 2006; MULLANEY *et al*., 2010; WEBER *et al*., 2002), como por exemplo a alteração encontrada na fibrose cística (COLLINS *et al*., 1987). Sendo assim, os indels, assim como os SNPs, são polimorfismos de potencial interesse na medicina personalizada, uma vez que podem estar associados a alterações fenotípicas (MULLANEY *et al*., 2010). Por apresentarem baixas taxas de mutação, assim como os SNPs, os indels também podem ser úteis para inferências de ancestralidade e estudos de linhagem, dependendo de sua localização no genoma.

Embora apresentem semelhanças com os SNPs, como por exemplo, a possibilidade de serem genotipados com eficácia em amostras degradadas ou antigas, uma vez que necessitam de amplicons menores que os STRs; os indels são analisados por eletroforese capilar com base apenas no tamanho, sem a necessidade de métodos de sequenciamento direto. Ou seja, combinam as vantagens analíticas desses outros tipos de marcadores, o que tem feito com seu uso esteja aumentando em estudos populacionais e forenses (PEREIRA *et al*., 2009; 2012a; 2012b).

Em contraste com os SNPs, que têm sido extensivamente investigados ao longo do genoma humano, os esforços para a descoberta dos indels e de outras formas de variação genética têm sido mais escassos. Embora haja indicações de uma alta densidade de indels no genoma humano, ainda há muito a ser elucidado, uma vez que um número ainda discreto

tenha sido descoberto e validado (MILLS *et al*., 2006). Apesar de serem abundantes, o interesse neste tipo de marcador é muito recente. Um dos estudos pioneiros para a descoberta de indels ao longo do genoma humano foi conduzido por Weber e colaboradores (2002), que identificaram cerca de 2000 polimorfismos. Outros trabalhos relataram a descoberta de indels por análise de dados de sequenciamento de genes alvos relacionados a vias metabólicas, com *software* que identificam indels e SNPs (BHANGALE *et al*., 2005; BHANGALE; STEPHENS; NICKERSON, 2006). Projetos para identificação de indels em indivíduos de diferentes origens geográficas também vem sendo realizados, com abordagens bioinformáticas, em bases de dados de sequências originalmente geradas para a descobertas de SNPs (MILLS *et al*., 2006), assim como em seqüenciamentos de genomas pessoais (LEVY *et al*., 2007; WHEELER *et al*., 2008). Comparando-se os resultados obtidos nesses sequenciamentos, observou-se que tanto o número de indels quanto o tamanho dos mesmos têm apresentado grande variação entre os indivíduos. Isso pode ser explicado pelas diferentes plataformas de sequenciamento utilizadas, assim como as abordagens analíticas. As taxas de validação dos indels por PCR e sequenciamento também variam consideravelmente, sendo necessário o estabelecimento de parâmetros como a faixa de tamanho dos indels para comparações (MULLANEY *et al*., 2010).

Diferentes grupos de trabalho desenvolveram painéis para tipagem de indels já validados em autossomos e no cromossomo X, e estudos colaborativos com o objetivo de relatar frequências alélicas em diferentes populações em todo o mundo vêm sendo realizados. Gradativamente, dados autossômicos têm sido disponibilizados para populações europeias, africanas, asiáticas e nativas americanas (WEBER *et al*., 2002; YANG *et al*., 2005; PEREIRA *et al*., 2009; SANTOS *et al*., 2010).

O genoma humano – recombinação e propriedades de transmissão

Uma das características importantes do genoma a ser levada em consideração no uso de marcadores genéticos é a recombinação, que consiste na troca aleatória de material genético entre os cromossomos, durante a meiose. Através dela ocorre um aumento da variabilidade ao longo das gerações, e a individualização, pois, com exceção de gêmeos idênticos, a probabilidade de duas pessoas compartilharem o mesmo perfil genético é virtualmente nula, para marcadores recombinantes. Dependendo da presença ou ausência de

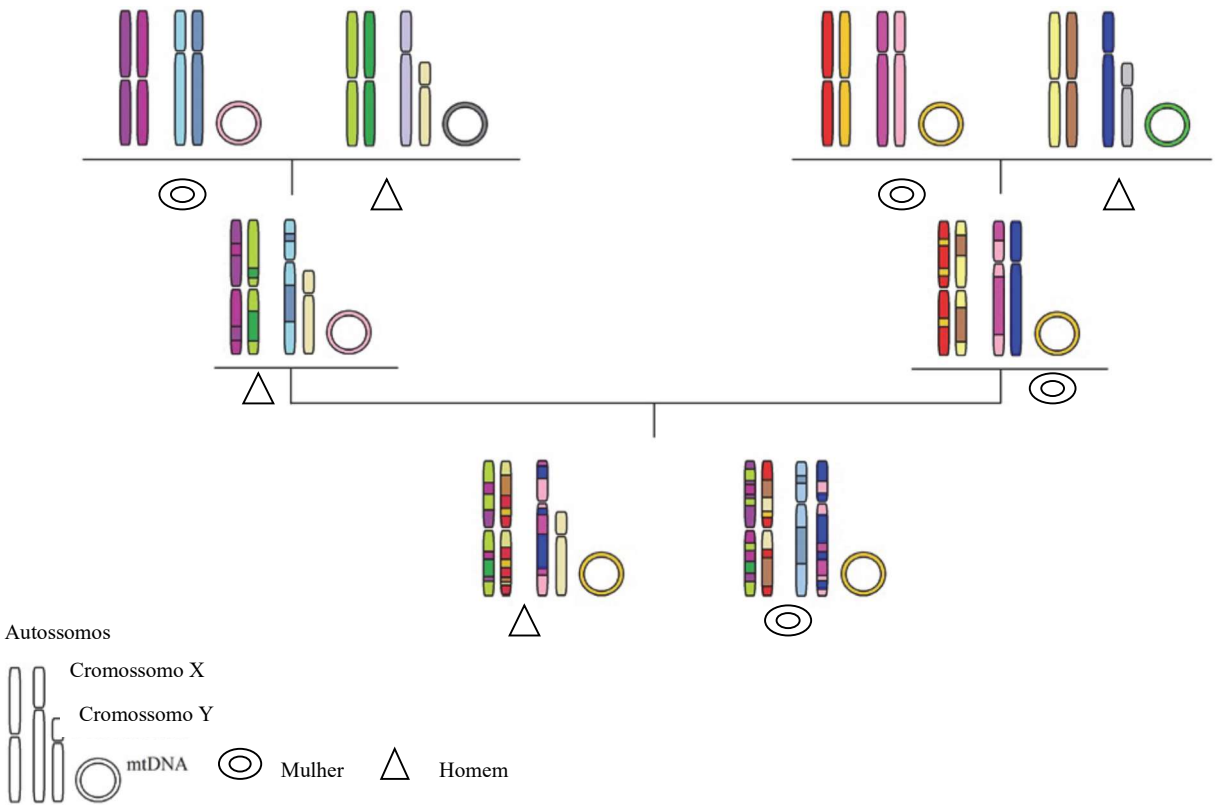
recombinação, podemos classificar as porções genômicas em dois grupos: genomas recombinantes, que permitem a identificação individual (marcadores autossômicos e do cromossomo X) e genomas não recombinantes, que permitem discriminar linhagens parentais e fornecem informações que são compartilhadas por um grupo de pessoas relacionadas (cromossomo Y - na sua maior parte - e mtDNA) (AMORIM; BUDOWLE, 2016). A Figura 2 mostra os diferentes genomas e suas propriedades de transmissão.

Genomas recombinantes: Autossomos e Cromossomo X

Devido à recombinação e ao grande número de cromossomos disponíveis para estudar, os autossomos são a primeira escolha quando se procura variação entre indivíduos. Os marcadores autossômicos ainda são, portanto, a principal fonte de informação para estudos de genética populacional e genética forense, havendo diversos *kit*s comerciais bem padronizados, com alto poder de discriminação entre indivíduos (GILL *et al*., 2012; GJERTSON *et al*., 2007; MORLING *et al*., 2002).

A transmissão biparental dos autossomos é importante para a sua escolha em análises de parentesco, uma vez que tanto o pai quanto a mãe contribuem com metade da informação para seu filho ou filha (Figura 2). Portanto, quaisquer indivíduos do sexo masculino ou feminino com parentesco materno ou paterno compartilharão alelos por descendência com uma probabilidade que depende do coeficiente de co-ancestralidade entre eles (PINTO; SILVA; AMORIM, 2010). Esses marcadores fornecem uma resposta satisfatória na maioria dos casos envolvendo análises de vestígios, identificação individual e na análise de parentesco. Mas em alguns casos mais complexos é necessário complementar a análise com outros tipos de marcadores que geralmente possuem diferentes características, como por exemplo, marcadores do mtDNA ou dos cromossomos X e Y (COBLE *et al*., 2009).

Figura 2 - Diferentes tipos de genomas e suas propriedades de transmissão e de recombinação



Legenda: Os genomas recombinantes, representados pelos autossomos e pelo cromossomo X, com herança biparental. O cromossomo Y, com herança uniparental paterna e o mtDNA com transmissão uniparental materna, ambos sem recombinação.
Fonte: Adaptado de AMORIM; BUDOWLE, 2016.

Marcadores do cromossomo X e autossômicos têm o mesmo PD para perfis femininos em uma população. No entanto, porque os homens têm apenas um cromossomo X, o PD obtido para *loci* específicos deste é menor do que o obtido para *loci* igualmente diversos nos autossomos. Sendo assim, a relevância do cromossomo X se deve principalmente ao seu número diferencial entre homens e mulheres, que tem consequências nos níveis de recombinação: O único cromossomo X masculino é transmitido às filhas sem recombinação, enquanto o cromossomo X herdado da mãe contém a informação recombinada dos dois cromossomos X maternos (Figura 2). Devido a este modo particular de herança, a aplicação desses marcadores depende essencialmente do sexo dos indivíduos envolvidos.

Embora usado em um contexto mais específico, os marcadores do cromossomo X também têm alguma aplicabilidade em ciência forense, como em situações de incesto, onde os dados autossômicos não são suficientes para identificar o parentesco (SZIBOR, 2007;

GOMES *et al*., 2012), ou em investigação de parentesco, nomeadamente em casos de paternidade onde o suposto pai não está disponível e seus parentes mais próximos devem ser analisados em vez dele. Os marcadores do cromossomo X também podem ser de interesse em testes de paternidade, quando dois supostos pais são pai e filho e o filho investigado é do sexo feminino. Nesta situação, ambos os homens não compartilharão entre si alelos do X por descendência e irão transmitir seus haplótipos X intactos para uma filha. Outra utilidade está na resolução de testes de parentesco envolvendo duas irmãs ou meias-irmãs, ajudando a considerar ou excluir a paternidade de forma mais eficiente do que marcadores autossômicos (SZIBOR, 2007).

Genomas não recombinantes: Marcadores de linhagem do Cromossomo Y e do mtDNA

Os marcadores de linhagem, também chamados de uniparentais, encontram-se localizados no genoma mitocondrial e no cromossomo Y e informam quanto à história das linhagens femininas e masculinas, respectivamente. Por essa razão são especialmente úteis no estudo de eventos populacionais que ocorrem diferencialmente no genoma de homens e mulheres (SCHAFFNER, 2004). Além disso, esses marcadores ocorrem como cópia única no genoma e por isso têm um tamanho populacional efetivo reduzido, sendo de 1/4 dos autossomos e 1/3 dos cromossomos X. Essas características, em conjunto, fazem com que tais marcadores apresentem diferenças genéticas acentuadas entre os grupos continentais humanos.

O DNA mitocondrial

O mtDNA é uma molécula circular de fita dupla constituída por duas cadeias polinucleotídicas complementares, de composição nucleotídica distinta (Figura 3). A cadeia pesada (do inglês, *Heavy*, representada por H) tem uma maior percentagem de Guaninas (G) e Adeninas (A), que contribuem para um peso molecular superior, enquanto a cadeia leve (do inglês, *Light*, representada por L) é rica em Citosinas (C) e Timinas (T) apresentando, consequentemente, um menor peso molecular. A molécula tem aproximadamente 16569 pb,

podendo esse número variar devido a inserções e / ou deleções na sequência (ANDERSON *et al*., 1981).

Embora o mtDNA seja considerado monoclonal (o que significa que todas as cópias dentro da célula são iguais), é comum para um indivíduo apresentar heteroplasmia, ou seja, ter mais de um tipo de sequência de mtDNA, variando em tamanho ou em composição nucleotídica (BENDALL *et al*., 1996; GILL *et al*., 1994). Provavelmente as heteroplasmias estão presentes em todos os indivíduos, mas frequentemente em níveis tão baixos que não são detectadas pelos métodos de sequenciamento comumente usados. Geralmente, quando estão em um nível de 20 % ou mais das moléculas de mtDNA já são detectáveis pelo sequenciamento de Sanger. A química usada no sequenciamento, assim como a posição da heteroplasmia também influenciam na detecção (TULLY *et al*., 2001).

O padrão de herança do mtDNA é uniparental por via materna. Isso porque na fecundação, apenas o núcleo do espermatozóide entra no oócito, fundindo-se diretamente com o núcleo materno, e não contribuindo com outros elementos celulares. Além disso, possíveis mitocôndrias paternas presentes são marcadas com ubiquitina e seletivamente degradadas (SUTOVSKY *et al*., 1999). Sendo assim, uma vez que o mtDNA não recombina a cada geração, sua variabilidade deve-se apenas à ocorrência de mutações ao longo das gerações, geralmente envolvendo um único nucleotídeo (SNPs), podendo ocorrer também pequenas inserções ou deleções, especialmente em regiões homopoliméricas ou poli-C. Essas regiões localizam-se entre as posições 302 e 310 e as posições 16183 e 16194 da molécula de mtDNA.

A molécula de mtDNA é composta por duas regiões - a região codificante e a região controle (Figura 3) - classificadas de acordo com a função que desempenham na molécula. A região codificante representa cerca de 90 % do genoma mitocondrial e compreende 37 genes que se traduzem em 22 tRNAs, 2 rRNAs e 13 polipeptídios. Já a porção não codificante do mtDNA, também chamada de região controle (CR, do inglês, *Control Region*) ou *D-loop* (do inglês, Displacement loop), tem uma extensão aproximada de 1120 pb e contém a origem de replicação da cadeia pesada.

A taxa de mutação do genoma mitocondrial é superior à do nuclear, tendo um valor médio na ordem de 10^{-8}. A proximidade da molécula com a cadeia de transferência de elétrons, e a consequente exposição a espécies reativas de oxigênio aumentam a probabilidade de ocorrência de mutação. Além disso, a ausência de histonas protetoras faz com que essa vulnerabilidade aumente. Soma-se também a esses fatores a baixa eficiência do sistema de reparo de erros durante a replicação do mtDNA. A velocidade de mutação ao longo da

molécula é heterogênea, sendo o maior nível de variabilidade genética encontrado na região controle, que se subdivide em três segmentos hipervariáveis: HV-I, HV-II e HV-III (Figura 3). Essa alta variabilidade se explica pelo fato desses segmentos não incluírem informação necessária à codificação de nenhuma substância essencial ao funcionamento celular, sendo as mutações acumuladas nessas regiões fixadas com maior frequência, enquanto as da região codificante são alvo de seleção negativa (ANDERSON *et al*., 1981; SOARES *et al*., 2009).

A sequência completa do mtDNA humano foi descrita pela primeira vez, por Anderson e colaboradores, em 1981, no laboratório de Frederick Sanger, em Cambridge; motivo pelo qual ficou conhecida como Sequência de referência de Cambridge (CRS, do inglês, *Cambridge Reference Sequence*). Cada posição nucleotídica foi numerada, com início na origem de replicação da cadeia pesada (posição 1) e término na posição 16.569 (ANDERSON *et al*., 1981). Quase duas décadas depois, o genoma mitocondrial estudado por Anderson foi reanalisado pelo grupo de Andrews e colaboradores, que identificaram 11 erros na sequência descrita em 1981. As correções foram publicadas e a sequência passou a ser denominada por Sequência de Referência de Cambridge revista (rCRS, do inglês *revised Cambridge Reference Sequence*), sendo essa considerada a sequência padrão de comparação em genética populacional e forense atualmente (ANDREWS *et al*., 1999).

As primeiras investigações genéticas para fins forenses e populacionais basearam-se apenas em HV-I e HV-II, mas atualmente a Sociedade Internacional de Genética Forense – ISFG (do inglês, *International Society for Forensic Genetics*) recomenda que no mínimo seja utilizada a sequência da CR completa para a publicação de dados populacionais de referência (PARSON *et al*., 2014). Em estudos que envolvem a variabilidade do genoma mitocondrial, a separação de duas linhagens maternas pode não ser possível apenas com o estudo da região controle. Na tentativa de ultrapassar esta limitação, começou-se a estudar polimorfismos SNPs localizados na região codificante do mtDNA. Na verdade, com o uso crescente de técnicas de MPS, a expectativa é que cada vez mais dados de sequenciamento da molécula completa de mtDNA, também chamada de mitogenoma, sejam publicados (KING *et al*., 2014; MIKKELSEN *et al*., 2014; STROBL *et al*., 2018).

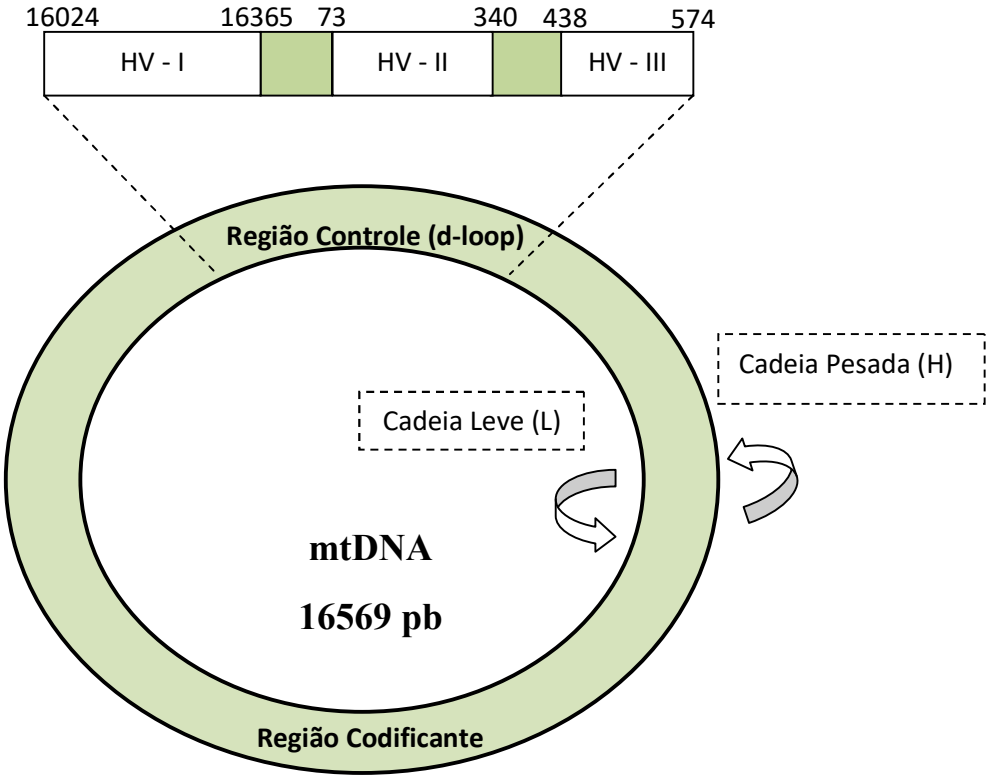
O estudo do mtDNA é amplamente utilizado como ferramenta em pelo menos três áreas distintas: a genética médica, a genética forense e a genética de populações ou antropologia molecular (KAYSER, 2007; TORRONI *et al*., 2006; UNDERHILL ; KIVISILD, 2007). Na genética forense, o uso de marcadores do mtDNA está limitado a alguns casos de investigação forense, pelo seu reduzido PD em comparação aos marcadores autossômicos, uma vez que não permite a identificação a nível individual. O fato de serem transmitidos em

bloco, como se fossem um único marcador, pode ser uma vantagem em situações de identificação de vítimas de desastres em massa ou pessoas desaparecidas, por exemplo, quando os parentes de mesma linhagem materna estão disponíveis, ainda que sejam separados por várias gerações. Mas uma vez que indivíduos não aparentados podem ter sequências semelhantes, sempre que possível, é aconselhável que a análise de mtDNA seja complementada com dados genéticos adicionais.

Quando usado em estudos populacionais, o genoma mitocondrial apenas nos fornece a informação transmitida ao longo das gerações por via materna. No entanto, essa mesma característica, associada ao alto número de cópias, ausência de recombinação e taxa de mutação maior que o DNA nuclear, possibilita o rastreamento de linhagens maternas ao longo do tempo. Supõe-se que todos os tipos de mtDNA no *pool* genético humano podem ser rastreados até um ancestral matrilinear comum que viveu há aproximadamente 200.000 anos na África (MACAULY *et al*., 2005), através de polimorfismos específicos, uma vez que a sequência do mtDNA evoluiu, levando à divergência das linhagens como resultado do acúmulo sequencial de mutações.

Do ponto de vista técnico, a maior vantagem da análise do mtDNA em amostras forenses está no alto número de cópias presentes em cada célula, e também na estrutura circular da molécula, que a torna mais resistente à degradação que o DNA nuclear. Por essas razões, a taxa de sucesso na extração do DNA e subsequente amplificação por PCR é muito maior para o mtDNA do que para o DNA nuclear, o que torna a análise do genoma mitocondrial particularmente útil em amostras antigas ou degradadas, e ainda em amostras muito escassas, em que a quantidade de material biológico disponível para análise é reduzida (LINCH; WHITING; HOLLAND, 2001; PAABO, 1989).

Figura 3 - Representação da estrutura circular do genoma mitocondrial



Legenda: Molécula circular do mtDNA com as cadeias pesada (H) e leve (L). A região não codificante ou região controle em destaque, mostrando as posições dos segmentos hipervariáveis HV-I, HV-II e HV-III.
Fonte: Adaptado de BUTLER, 2005.

Em relação às metodologias de análise para o mtDNA, quando se estuda apenas a região controle ou segmentos desta, comumente se usa a técnica de sequenciamento de Sanger para a detecção dos polimorfismos SNP. Nessa metodologia, o sequenciamento é realizado com *primers* direto e reverso, de forma a obter-se informação complementar das duas fitas, visando um controle de qualidade das sequências (SANGER; NICKLEN; COULSON, 1977). Quando se tem por objetivo o sequenciamento de mitogenomas, as análises realizadas com técnicas de MPS têm apresentado resultados de alta precisão, sensibilidade e eficiência, além de ser de alto rendimento e simples operação. Consistem, resumidamente, na quebra de sequências de DNA em pequenos fragmentos (aproximadamente entre 400-600 pb), o sequenciamento destes em paralelo, seguido da montagem e ordenação num único fragmento contíguo. O produto sequenciado é diretamente detectado sem que haja a necessidade de realizar eletroforese capilar. Essas técnicas têm superado algumas limitações do sequenciamento de Sanger, como a perda de informação em amostras degradadas e a detecção de heteroplasmias e misturas, por exemplo (HOLLAND; MCQUILLAN; O' HANLON, 2011; KING *et al*., 2014; PARSON *et al*., 2015).

Após sequenciamento do genoma mitocondrial, as sequências obtidas são descritas por comparação com a sequência de referência rCRS, sendo representadas pelo conjunto de posições diferentes em relação a ela, ficando subentendido que as posições restantes são idênticas. O conjunto de polimorfismos identificados numa sequência de mtDNA denomina-se haplótipo, o qual caracteriza uma determinada linhagem materna. A nomenclatura dos haplótipos é feita com base em diretrizes detalhadamente descritas pela comunidade científica, de forma a evitar erros de interpretação (AMORIM; FERNANDES; TAVEIRA, 2019; PARSON *et al*., 2014; TULLY *et al*., 2001). A existência de uma nomenclatura única permite o intercâmbio e a comparação de resultados entre laboratórios, a nível internacional, de forma adequada. A partir do agrupamento de haplótipos de mtDNA, definidos por SNPs específicos compartilhados por linhagens próximas entre si que possuem um ancestral comum, formam-se os haplogrupos, que tendem a mostrar alguma especificidade regional, que pode ser mais abrangente ou restringir-se até mesmo a grupos étnicos. Esses haplogrupos são nomeados por letras do alfabeto e podem conter sub-haplogrupos com alterações específicas (AMORIM; FERNANDES; TAVEIRA, 2019; PAKENDORF; STONEKING, 2005).

Visando detalhar as relações filogenéticas de variantes conhecidas do mtDNA, foi desenvolvida uma árvore filogenética – *Phylotree* (VAN OVEN; KAYSER, 2009), englobando os polimorfismos identificados até aquele momento e que definem os haplogrupos e suas ramificações (Figura 4). A *Phylotree* vem sendo atualizada à medida que novos polimorfismos em sequências de mitogenomas completos são descritos e pode ser acessada na página eletrônica http://www.phylotree.org. A detecção da variabilidade da molécula completa de mtDNA garante uma maior precisão na descrição dos haplogrupos do que apenas com a região controle, uma vez que a taxa de mutação nas regiões codificantes é bastante inferior e, por isso, apresentam uma menor probabilidade de recorrência e, consequentemente, uma maior especificidade geográfica.

Os ramos mais profundos e antigos da *Phylotree* incluem haplogrupos do continente africano, mais especificamente da região subsaariana. Em uma resumida descrição da distribuição dos haplogrupos, temos que: Os haplogrupos africanos são L0 a L6 (SALAS *et al*., 2002; TORRONI *et al*., 2006), considerados os de maior diversidade. Os haplogrupos M e N surgem na África oriental, a partir de L3, dispersando-se pela Eurásia e dando origem às linhagens "out-of-Africa" presentes em populações desta região. As linhagens europeias derivam do haplogrupo N, incluindo os haplogrupos H, I, J, K, T, U, V, W e X. Estes, em conjunto, estão presentes em 98 % da população europeia. As linhagens asiáticas derivam

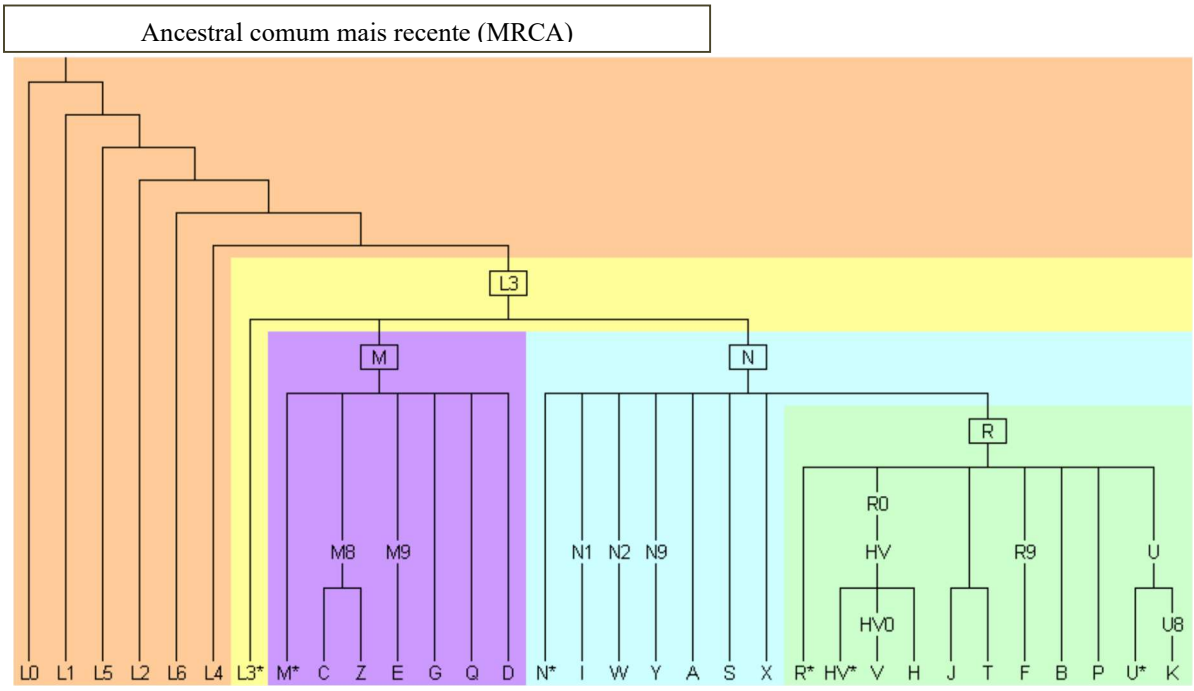
tanto do haplogrupo M como do haplogrupo N. Nas populações nativo-americanas verifica-se a presença dos haplogrupos A, B, C e D, característicos também da Ásia, enquanto os haplogrupos G, Y e Z predominam na Sibéria (MISHMAR *et al*, 2003; QUINTANA-MURCI *et al*., 1999).

A distribuição geográfica e a variabilidade das linhagens do mtDNA podem ser altamente informativas nas inferências de rotas de migração no passado distante. Porém, para estudar a história das populações humanas, além de analisarmos as linhagens existentes em uma população, é crucial o uso de métodos estatísticos que comparem as populações, estimando, por exemplo, distâncias genéticas entre as mesmas, a fim de compreendermos as afinidades populacionais. Além disso, o mtDNA reflete apenas a história via materna de uma população e representa apenas um *locus*, podendo não ser, por isso, tão preciso, devido a efeitos de deriva e da seleção natural sobre esse *locus*. Sendo assim, associar dados do cromossomo Y e de marcadores autossômicos em estudos populacionais pode ser muito relevante (PAKENDORF; STONEKING, 2005; WOODING *et al*., 2004).

Bancos de dados populacionais de mtDNA, assim como bancos de dados de outros tipos de marcadores usados em genética forense, servem de base para estimativas de frequências de haplótipos de mtDNA gerados em laboratório e que sejam pertinentes a casos judiciais. Embora sejam essenciais para os cálculos estatísticos que validam as evidências apresentadas nesse contexto, os dados depositados nessas bases estão sujeitos a erros cometidos por quem os gera. Para estabelecer padrões de qualidade nas etapas de geração, análise, transferência e controle de qualidade de dados do mtDNA, de forma a atenderem os padrões requeridos em análises forenses, foi criada a base de dados EMPOP (*EDNAP mtDNA population database)*, pelo Instituto de Medicina Legal de Innsbruck (PARSON; DÜR, 2007). A base de dados EMPOP é considerada a mais importante para o mtDNA, e inclui desde sequências contendo apenas a região HV-I até mitogenomas completos, sendo a grande maioria sequência da CR completa (https://empop.online). Essa plataforma digital está acessível para consulta pública, com dispositivos de análise disponíveis para a comunidade científica, contribuído tanto como um banco de dados de populações de referência quanto como ferramenta de controle de qualidade para publicações contendo dados de mtDNA (AMORIM; FERNANDES; TAVEIRA, 2019).

Outro importante banco de dados *online* de mtDNA humano é o MITOMAP (http://www.mitomap.org), onde é possível acessar informações sobre as variantes do mtDNA associadas a dados geográficos e a doenças, além de ilustrações, distribuição de haplogrupos e suas frequências (LOTT *et al*., 2013).

Figura 4 - Representação simplificada dos haplogrupos de mtDNA na *Phylotree* e suas ramificações



Legenda: Os haplogrupos africanos L0-L6 são os mais antigos da árvore. O haplogrupo L3 ramifica-se nos haplogrupos euroasiáticos M e N, que por sua vez se subdividem nos haplogrupos euroasiáticos restantes e nos nativo americanos A -D. Os símbolos dos haplogrupos seguidos por um asterisco representam todas as linhagens descendentes não mostradas de um clado particular, para o qual uma letra do alfabeto não foi reservada.
Fonte: VAN OVEN; KAYSER, 2009.

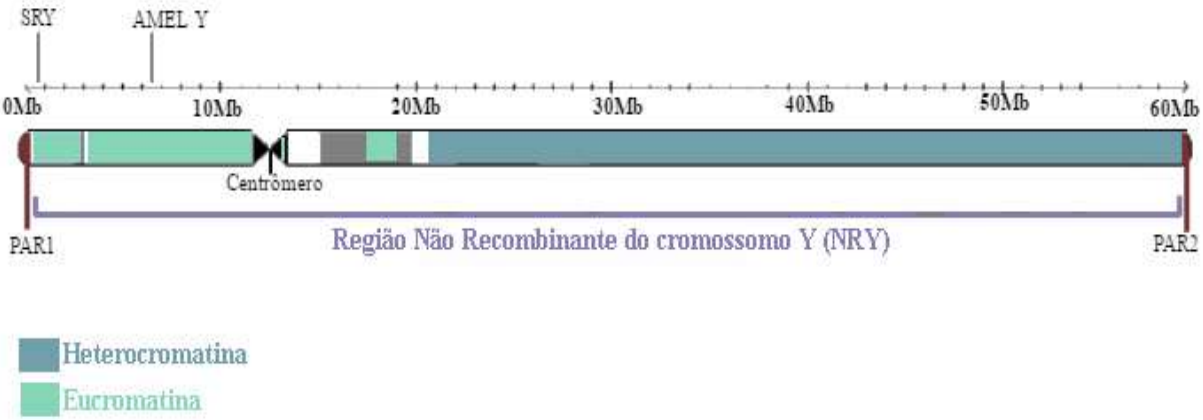
O Cromossomo Y

O cromossomo Y é um dos menores cromossomos do genoma humano, cujo tamanho é de aproximadamente 58 milhões de pares de bases (MORTON, 1991). Ocorre somente em uma cópia por célula e é especifico nos homens, sendo transmitido de pais para filhos (Figura 2). Nesse cromossomo estão presentes dois tipos de cromatina: eucromatina e heterocromatina, localizadas em diferentes porções. A heterocromatina tem tamanho muito variável entre indivíduos, e encontra-se no braço longo. Possui genes em sua extensão, mas em sua maior parte é caracterizada por sequências extremamente repetitivas (também conhecidas como amplicônicas) e não funcionais, que podem se apresentar de maneira consecutiva (*in tandem*) ou como sequências invertidas (palíndromo). A eucromatina está presente no restante do cromossomo e em sua extensão existem genes responsáveis por importantes funções biológicas, como por exemplo, os genes da região de determinação sexual (SRY, do inglês *Sex-determining Region*), e o gene da amelogenina (AMEL Y) (Figura

5). Aproximadamente 95 % da extensão do cromossomo Y, conhecida como região não-recombinante (NRY, do inglês, *Non-recombining Y chromosome region*) ou região masculina específica do cromossomo Y (do inglês, *Male-specificY chromosome region* ou MSY), não está sujeita à recombinação com o cromossomo X, durante a meiose (SKALETSKY *et al*., 2003). Apenas as regiões PAR1 e PAR2, que estão nas extremidades dos braços curto (Yp) e longo (Yq) do cromossomo, recombinam, pois são homólogas a sequências do cromossomo X e são responsáveis pelo pareamento adequado entre os dois cromossomos sexuais durante a meiose. (Figura 5) (BUTLER, 2005).

Algumas características do genoma mitocondrial são comuns ao cromossomo Y, uma vez que, assim como o mtDNA, a região NRY é de natureza haplóide e ausente de recombinação. Por isso o conjunto de alelos dos marcadores do cromossomo Y, também denominado de haplótipo, é integralmente e exclusivamente transmitido a todos os indivíduos do sexo masculino de uma mesma linhagem paterna, e a variabilidade entre eles só vai existir se ocorrerem mutações e arranjos intracromossômicos ao longo do tempo (JOBLING; HURLES; TYLER-SMITH, 2003). Comparado ao mtDNA, porém, o cromossomo Y tem maior diversidade, pois além de marcadores bialélicos SNPs e indels, existem muitos STRs disponíveis para estudos genéticos forenses e populacionais. A escolha adequada dos marcadores do cromossomo Y a serem usados em cada tipo de análise deve levar em consideração suas características singulares, como o nível de diversidade e taxa de mutação.

Figura 5 – O cromossomo Y



Legenda: Cromossomo Y e sua extensão aproximada em megabases. Em destaque estão as regiões PAR1 e PAR 2, que se recombinam com regiões homólogas do cromossomo X, além da região NRY, as porções de heterocromatina e eucromatina e os genes da região de determinação sexual (SRY) e da amelogenina AMEL Y.

Fonte: RÊGO, 2019.

Na genética forense, os marcadores do cromossomo Y encontram algumas limitações, uma vez que são aplicáveis somente em casos que envolvam o sexo masculino, e podem demonstrar vínculos genéticos apenas por via paterna. Por não serem capazes de individualização, mas sim de caracterização de linhagens, os marcadores do Y não podem excluir a paternidade de nenhum outro homem pertencente à mesma linhagem paterna do verdadeiro pai biológico, por exemplo. Sendo assim, devem ser usados em conjunto com marcadores com maior poder de discriminação entre indivíduos, como os autossomos. Apesar dessas limitações, nas análises forenses, os polimorfismos do cromossomo Y são especialmente úteis quando os perfis obtidos a partir de marcadores autossômicos não são suficientemente informativos. Um exemplo são os casos de paternidade na ausência de material biológico pertencente ao suposto pai. Nessas situações (por exemplo, quando o suposto pai já faleceu), é possível acessar o perfil genético completo de seu cromossomo Y usando qualquer parente do sexo masculino na mesma linhagem paterna e comprovar o vínculo (AMORIM; BUDOWLE, 2016; KAYSER, 2017).

Haplótipos do cromossomo Y têm sido amplamente utilizados também em investigações criminais, podendo, por exemplo, revelar se um doador de evidências deixadas em uma cena de crime é do sexo masculino, ou se existe mais de um doador do sexo masculino, no caso de evidências com múltiplas fontes. No caso de evidências onde existe mistura do DNA da vítima com o DNA do agressor, pode ser muito difícil separar os perfis genéticos de ambos usando marcadores autossômicos. Mesmo em amostras que contenham células oriundas de esperma do agressor misturadas às células da vítima, em que existe a possibilidade de uma extração diferencial, separando as frações celulares e obtendo os perfis genéticos separados, a extração diferencial tende a falhar quando a fração de esperma é muito baixa. Nestes casos, quando os marcadores autossômicos são usados, a amplificação preferencial do componente principal presente na mistura (geralmente o DNA da vítima) pode mascarar o perfil genético do estuprador. Porém, em uma mistura de DNA masculino e feminino, mesmo em baixa proporção de material masculino, a genotipagem adicional de marcadores do cromossomo Y pode fornecer um perfil específico do homem (AMORIM; BUDOWLE, 2016; KAYSER, 2017).

Os Y-STRs, por terem taxa de mutação maior que os SNPs e Indels, apresentam maior diversidade alélica; e ainda podem ser encontrados em cópia única ou em múltiplas cópias ao longo da extensão do cromossomo, gerando, assim, uma maior variedade alélica pela combinação dos *loci* existentes. Tal característica torna esses marcadores úteis em diversas aplicações nas quais o principal objetivo é a identificação de linhagens paternas, como:

investigações de parentesco e em investigações criminais, além de estudos de diversidade populacional e identificação de homens vítimas de desastres (BALLANTYNE *et al*., 2010; KAYSER, 2017). Para um maior poder de diferenciação de indivíduos, costuma-se analisar vários *loci* Y-STR simultaneamente. Os marcadores de escolha devem estar bem caracterizados e apresentar alto nível de polimorfismo. Também é necessário conhecer a distribuição das frequências haplotípicas nas populações de interesse, assim como as taxas de mutação desses *loci*.

Apesar de seu amplo uso na ciência forense, os marcadores Y-STR podem apresentar alguns desafios nos cálculos estatísticos envolvendo suas frequências. Isso porque os dados genéticos devem ser tratados como haplótipos, para a determinação dessas frequências, tornando a construção de bancos de dados de Y-STR complexa, uma vez que todo o haplótipo do Y deve ser determinado geneticamente para cada amostra, ao invés de tipagens isoladas de marcadores, como nos autossomos. Como consequência, a maioria dos bancos de dados populacionais é representada principalmente por haplótipos muito raros ou únicos quando um grande número de *loci* está incluído. Sendo assim, para a aplicação prática desses marcadores, bancos de dados adequados devem possuir o maior número de indivíduos da população local com a tipagem do maior número possível de *loci*, a fim de driblar a ocorrência de subestrutura populacional (GUSMÃO; CARRACEDO, 2003). As frequências haplotípicas dos Y-STRs, assim como suas taxas de mutação, podem ser calculadas através de banco de dados próprio, mas distribuições de haplótipos Y-STR em populações de todo o mundo têm sido disponibilizadas por meio de publicações e, mais recentemente, por meio de bases de dados forenses de larga escala. Um dos maiores e mais utilizados bancos de dados de haplótipos de referência *online* para o cromossomo Y é o YHRD (do inglês, *Y- Chromosome Haplotype Reference Database*), cujos objetivos principais consistem na disponibilização de dados de inúmeras populações mundiais, com controle de qualidade, acerca das frequências haplotípicas e das taxas mutacionais de Y-STRs (www.yhrd.org) (KAYSER, 2017; WILLUWEIT; ROWER, 2015). O desenvolvimento desses bancos de dados é importante não apenas para a estimativa de frequência de haplótipos e subsequentes cálculos de probabilidade de correspondência em estudos forenses, mas também para a análise comparativa e populações.

A comunidade forense tem focado alguma atenção em marcadores de mutação rápida localizados no cromossomo Y, os RM-Y-STRs, cujas taxas de mutação são da ordem de 10^{-2}. Por terem taxas de mutação acima da média, estes marcadores irão fornecer uma maior probabilidade de distinguir entre indivíduos intimamente relacionados e podem complementar

análises atualmente feitas com Y-STRs convencionais (BALLANTYNE *et al*., 2012). Alguns *kit*s comerciais mais recentes, como o *PowerPlex Y23 System* (Promega) e *Yfiler Plus* (*Thermo Fisher Scientific*), já incluem alguns desses RM-Y-STRs.

Outra classe de marcadores muito utilizada na análise do cromossomo Y são os SNPs. Por serem bialélicos, oferecem menor informação a nível intrapopulacional em comparação com os marcadores Y-STR. Os marcadores Y-SNP possuem baixas taxas de mutação (na ordem de 10^{-8} por nucleotídeo/por geração) quando comparados com os STRs. A caracterização genética de populações com ambos os tipos de marcadores simultaneamente possibilita uma análise eficiente de eventos relacionados com a evolução e a história das populações humanas. Uma vez que esses marcadores apresentam diferentes taxas de mutação, são capazes de apontar aspectos populacionais a níveis micro e macro-evolutivos, evidenciando dinâmicas ocorridas, como eventos migratórios e padrões de miscigenação.

Assim como no mtDNA, devido às baixas chances de mutação recorrente, os marcadores SNP do cromossomo Y tendem a apresentar eventos moleculares únicos da evolução humana, com uma elevada especificidade geográfica e populacional, uma vez que esses genomas haplóides estão mais sujeitos aos efeitos de deriva gênica, que resultam em frequências diferenciais entre regiões geográficas. Por décadas, esses marcadores foram usados como meio de rastrear a ancestralidade biogeográfica de indivíduos e populações, inicialmente em análises genealógicas, estudos evolutivos e populacionais (JANNUZZI *et al*., 2020; RESQUE *et al*., 2016; UNDERHILL; KIVISILD, 2007), a fim de conhecer rotas de migrações e origens. Tal conhecimento acumulado sobre a distribuição geográfica da diversidade genética do cromossomo Y, atualmente serve como base para as aplicações forenses de inferência de ancestralidade no campo forense, assim como no mtDNA (KAYSER, 2017; UNDERHILL; KIVISILD, 2007). São úteis em situações envolvendo doadores de DNA vestigial em cenas de crimes ou de vítimas de acidentes em massa, onde a identificação não tenha sido possível com marcadores autossômicos. Também podem ser usados em casos de identificação de pessoas desaparecidas e de vítimas de desastres em que não haja outras informações sobre a origem dos restos mortais. Em todas essas situações podem apontar uma linha investigativa a ser seguida.

Por ser um marcador de evolução lenta, a probabilidade de um determinado SNP sofrer a mesma mutação em genomas independentes é muito pequena e, desta maneira, os SNPs são capazes de agrupar linhagens ao associar indivíduos por descendência. Por convenção, um conjunto específico de Y-SNPs, assim como no mtDNA, definem haplogrupos do cromossomo Y. A baixa probabilidade de recorrência dessas mutações

permite assumir que o alelo mutante ou derivado de cada marcador Y-SNP surgiu apenas uma vez na história humana, e todos os homens que possuem tal alelo descendem de um ancestral comum, no qual a mutação ocorreu pela primeira vez (MIZUNO *et al*., 2010; UNDERHILL; KIVISILD, 2007). Os polimorfismos descritos nas árvores filogenéticas do cromossomo Y são capazes de detectar a origem de linhagens com distribuição tanto ao nível inter quanto intracontinental (CHIARONI; UNDERHILL; CAVALLI-SFORZA, 2009; CRUCIANI *et al*., 2011).

O primeiro sistema de filogenia e nomenclatura dos haplogrupos do cromossomo Y foi publicado em 2002, por um grupo de colaboradores, e definiu 18 haplogrupos principais, nomeados pelas letras A – R (Y CHROMOSOME CONSORTIUM, 2002). Desde então, atualizações foram publicadas, incluindo diversos Y-SNPs recentemente descobertos e considerados relevantes para estabelecer as relações evolutivas das linhagens do cromossomo Y, visando aumentar a resolução da árvore filogenética. A versão mais recente da árvore é composta por 20 haplogrupos principais, nomeados pelas letras A – T (JOBLING; TYLER-SMITH, 2003; KARAFET *et al*., 2008; VAN OVEN *et al*., 2014). Algumas fontes de informação *online*, como as páginas eletrônicas mantidas pela Sociedade Internacional de Genealogia Genética (ISOGG; http://www.isogg.org/tree) e por Thomas Krahn (http://ytree.ftdna.com) já mostram um vislumbre da enorme topologia da árvore filogenética 'completa' usando todos os Y-SNPs conhecidos e mapeados filogeneticamente, incluindo muitos que (ainda) não foram publicados na literatura científica.

Nas duas últimas décadas, várias filogenias diferentes do cromossomo Y e nomenclaturas de Y-SNPs e haplogrupos têm sido apresentadas na literatura científica e em conferências que demonstram a presente diversidade no cromossomo Y. Essas configurações, nem sempre consensuais, podem ser atribuídas ao crescimento exponencial do número de Y-SNPs descobertos, devido principalmente ao surgimento das novas tecnologias de MPS, que suportam estudos envolvendo o sequenciamento de genomas completos a um menor custo e maior velocidade que as tecnologias convencionais. Com o volume de informação que se tem gerado a partir de tais estudos a respeito da distribuição de haplogrupos do cromossomo Y em diferentes regiões, novas variantes deverão ser descobertas, havendo uma demanda por remodelações da estrutura da árvore filogenética do cromossomo Y (LARMUSEAU *et al*., 2015; VAN OVEN *et al*., 2014).

A árvore filogenética e a nomenclatura sugeridas por Van Oven e colaboradores, em 2014 (VAN OVEN *et al*., 2014), é uma proposta concisa, que não teve como objetivo usar todos os Y-SNPs conhecidos, mas sim as mutações mais estáveis que dão origem às

ramificações, em busca da melhor resolução em nível mundial. Essa versão mais atual está acessível na página eletrônica www.phylotree.org/Y, e foi baseada em um conjunto de 417 Y-SNPS. Tem sido a mais utilizada na determinação de haplogrupos, nomeando-os juntamente com as mutações que os caracterizam (Figura 6). Sendo assim, como exemplo, a notação A1-M31 refere-se ao haplogrupo A1, que pode ser identificado pela presença da mutação M31.

Resumidamente, na árvore, as duas principais divisões da base são os haplogrupos, A e B, cuja propagação é restrita à África, reforçando as evidências da origem do homem moderno. Ambos haplogrupos primários são geneticamente diversos e ramificam-se em subhaplogrupos geograficamente distintos um do outro, revelando uma possível fragmentação da população, isolamento e subsequente reexpansão no continente africano (Figura 6). O haplogrupo E também é conhecido por ser um haplogrupo africano, porém não se restringe à África, podendo ser encontrado alguns subhaplogrupos no oeste do continente Asiático e mais ao sul da Europa. Acredita-se que estava presente na população africana pela qual se deu a expansão "Out of Africa". Os haplogrupos G, H, I, J, R, T e L são observados em populações da Europa, Oriente Médio e algumas partes da Ásia e do norte da África. Os haplogrupos C e Q podem ser encontrados no continente asiático. Porém, no continente americano, o haplogrupo C está restrito ao norte, enquanto o haplogrupo Q estabeleceu-se mais profundamente (CHIARONI; UNDERHILL; CAVALLI-SFORZA, 2009; UNDERHILL, KIVISILD, 2007).

Uma vez sabidas as mutações que definem os ramos filogenéticos do cromossomo Y, e, portanto, seus haplogrupos; é possível inferir a ancestralidade paterna de um indivíduo do sexo masculino analisando esses Y-SNPs e agrupando os resultados de maneira hierárquica, seguindo os ramos da árvore até o ancestral comum. Essa análise direcionada facilita a genotipagem, visto que não é necessário analisar todos os marcadores, mas somente os de mutações mais recentes, diminuindo o custo e o trabalho envolvido. Ao consultar a árvore (Figura 6), é possível saber, por exemplo, que todo indivíduo pertencente ao haplogrupo A1 possui as mutações M31, V168 e L1085. Porém, na prática, só é necessária a detecção da mutação final – M31.

Devido às diferenças nas taxas mutacionais entre os marcadores Y-STR e Y-SNP, um determinado haplogrupo pode ser composto por difererentes haplótipos Y-STR. Alguns haplogrupos baseados em Y-SNP com diferenças marcantes de frequências entre regiões geográficas exibem uma forte associação com a diversidade haplotípica dos Y-STRs das linhagens que os compõem, de modo que a origem geográfica indicada pelo haplogrupo também pode ser inferida a partir dos haplótipos de Y-STR. A genotipagem de marcadores

STRs é realizada de forma simples, através de PCR e eletroforese capilar somente, enquanto a determinação de polimorfismos SNPs requerem técnicas de sequenciamento ou mini-sequenciamento, que envolvem mais etapas laboratoriais. Pode ser, portanto, vantajoso; utilizar esse tipo de marcador na predição de haplogrupos, principalmente na ausência de dados de marcadores Y-SNP (JOBLING, 2001; KAYSER, 2017). Sendo assim, vários *software* têm sido desenvolvidos, com o objetivo de inferir o haplogrupo correspondente a um haplótipo de Y-STR, a fim de direcionar a escolha dos Y-SNPs a serem analisados.

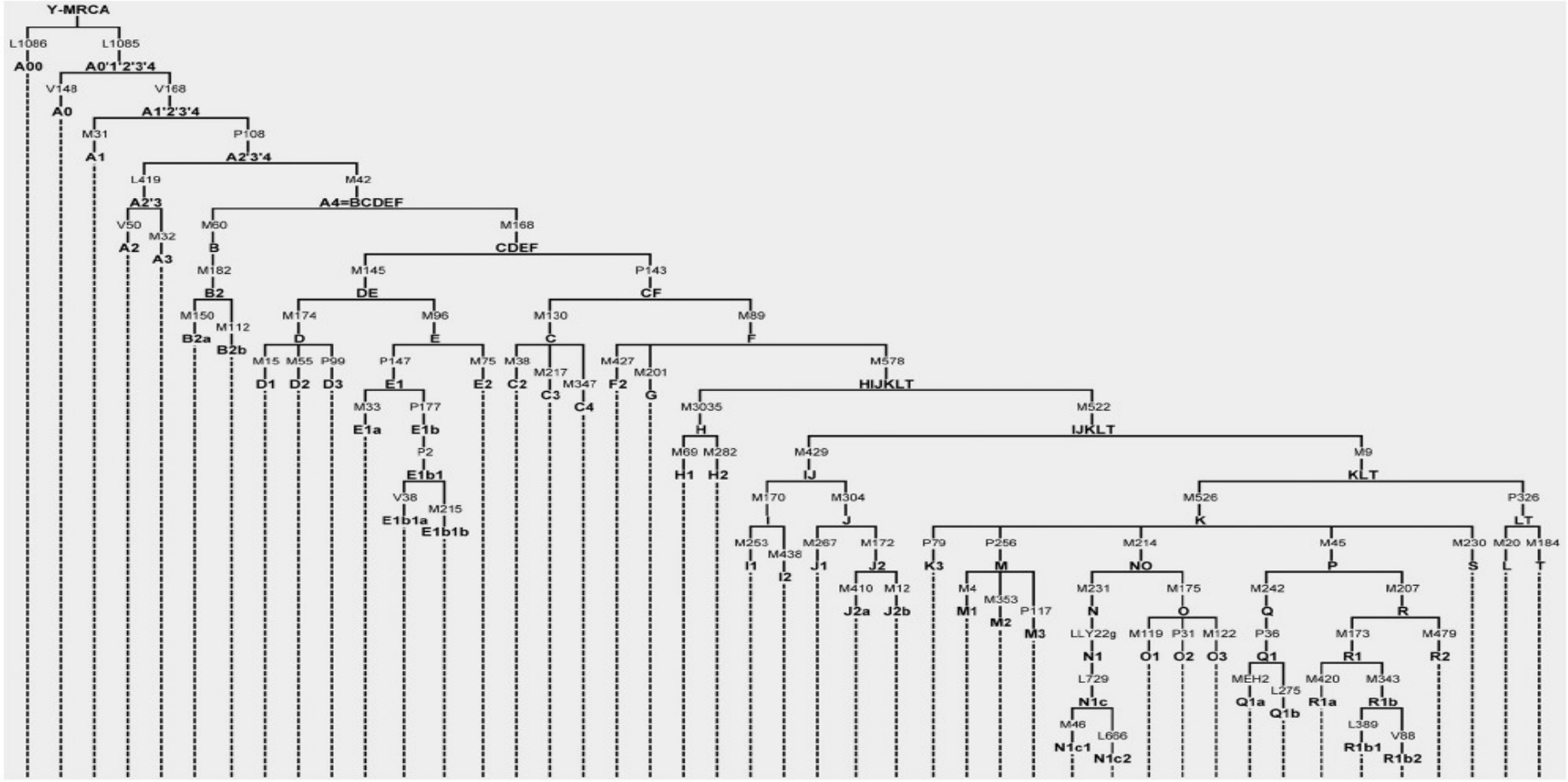
Um *software* gratuito amplamente utilizado pela comunidade científica é o *Haplogroup Predictor* (http://www.hprg.com/hapest5/), que se utiliza de um algoritmo implementado em um programa *online,* com abordagem Bayesiana. O *Haplogroup Predictor* indica a probabilidade de um determinado haplótipo pertencer a um haplogrupo, com base nas frequências alélicas observadas de cada marcador que compõe o haplótipo em diferentes haplogrupos. Essas frequências estão depositadas em base de dados públicas consultadas pelo programa. A abordagem Bayesiana adotada requer que, no momento da análise, seja indicada a possível região geográfica na qual os haplótipos analisados possam ter sido originados, visto que as frequências dos haplótipos podem variar conforme a região considerada. Uma opção é a que considera prioridades iguais para as regiões geográficas, devendo ser usada para indicar outras regiões que o programa não disponibiliza, como por exemplo, regiões do continente americano (ATHEY, 2006; MUZZIO *et al*., 2011).

O *NevGen* Y-DNA *Haplogroup Predictor* é outro *software* gratuito que pode ser usado diretamente na internet ou instalado no computador. O *NevGen* utiliza um algoritimo baseado na abordagem Bayesiana de frequências alélicas, com o diferencial de realizar a correlação (interdependência) de valores de diferentes pares de STRs durante os cálculos de probabilidade de predição de subhaplogrupos. A base de dados consultada pelo *NevGen* é a *Family Tree DNA* (FTDNA - www.familytreedna.com) e somente considera haplótipos que estejam associados com dados de elevadas resoluções de SNPs. Além disso, o *NevGen* não necessita que a possível região de origem do haplótipo seja indicada para o cálculo da estimativa da probabilidade.A fim de evitar a atribuição equivocada de algum haplogrupo, quando o algoritmo utilizado por este programa não é capaz de estimar uma probabilidade de um determinado haplótipo ser associado a algum haplogrupo considerado pelo programa, ele indica como resultado a probabilidade deste pertencer a “Subclados não-suportados” (GENTULA; NEVSKI, 2015).

A abordagem de predição de haplogrupos utilizada por estes *software* tem como limitações principais o tamanho do banco de dados utilizado, visto que se este for de tamanho

reduzido a estimativa das frequências alélicas e a variedade de haplogrupos observados serão precárias; além da falta de representatividade de certos haplogrupos e de grupos populacionais, como por exemplo, populações dos continentes americano e asiático (MUZZIO *et al.*, 2011).

Figura 6 - Árvore filogenética de haplogrupos do cromossomo Y



Legenda: Representação da árvore filogenética do cromossomo Y, com destaque em negrito para a nomenclatura de base dos haplogrupos e os Y-SNPs que definem cada ramificação. No topo, está indicado o ancestral comum mais recente de todos os homens modernos (Y-MRCA, do inglês *Most Recent Common Patrilineal Ancestor of all modern humans*).

Fonte: VAN OVEN et.al., 2014.

Ancestralidade Biogeográfica e Marcadores Informativos de Ancestralidade

Devido ao grande aumento da compreensão da variação populacional humana, pelo detalhamento do genoma e conhecimento de sua diversidade, têm sido identificadas porções variáveis entre indivíduos, com potencial para serem usados como marcadores informativos de ancestralidade – os AIMs (do inglês, *Ancestry Informative Markers*).

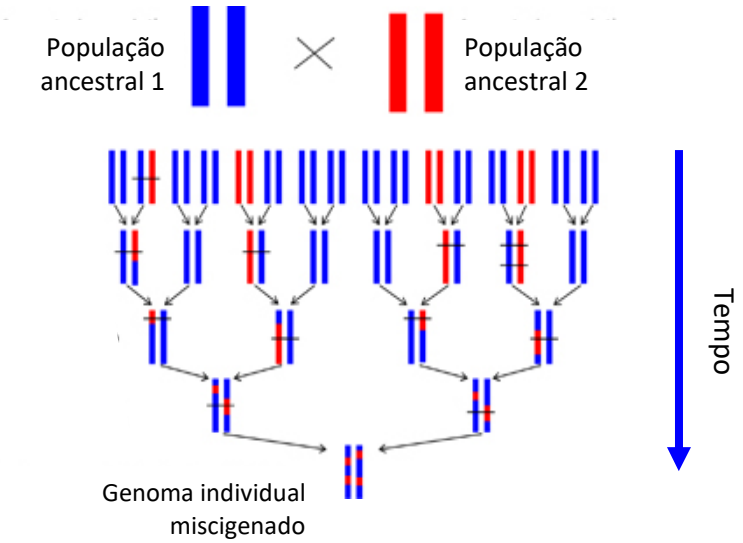
Os AIMs têm como principal característica uma baixa divergência nas frequências alélicas dentro das populações e uma alta divergência destas frequências em populações separadas geograficamente. Estes marcadores são especialmente úteis para inferir a provável origem ancestral de um indivíduo ou estimar as proporções de ancestralidade em indivíduos ou populações miscigenadas (PEREIRA *et al*., 2012a). A miscigenação é uma forma comum de fluxo gênico e refere-se ao processo pelo qual duas ou mais populações genética e fenotipicamente diversas acasalam-se e formam uma nova população, híbrida. Como resultado desse processo, cada cromossomo de um indivíduo miscigenado se apresenta como um mosaico de segmentos cromossômicos derivados de seus ancestrais (Figura 7) (BAYE; WILKE, 2010).

Sendo assim, a ancestralidade pode ser definida como a herança genética que cada indivíduo traz de seus ancestrais diretos, resultante dos acasalamentos entre diversos membros da população em gerações anteriores, ao longo do tempo. A análise da ancestralidade biogeográfica se detém nas variações encontradas em indivíduos e que podem sinalizar a sua origem de uma região geográfica em particular. Não está restrita a avaliações genéticas, podendo ser estimada, por exemplo, através das análises biométricas de esqueletos (PHILLIPS, 2015).

Na genética forense, análises de inferência de ancestralidade biogeográfica encontram diversas aplicações, como no auxílio em investigações, podendo substituir ou colaborar com o relato de testemunhas oculares, uma vez que essas podem cometer enganos. Ou trazendo dados adicionais na ausência de bancos de dados ou coincidências com perfis genéticos depositados nestes. Nesse caso, genótipos de AIMs obtidos de evidências em uma investigação podem indicar a provável ancestralidade de seu doador. São úteis também na identificação de vítimas de desastres ou de pessoas desaparecidas. Na genética médica a confirmação da ancestralidade de doadores que fazem auto-declaração em estudos de associação caso-controle minimizam a chance de resultados espúrios por alteração da frequência de um alelo em um dos grupos do estudo, não por influência desse no fenótipo

estudado, mas por ser mais presente em indivíduos com certa origem ancestral. Dessa forma contribuem para a acurácia de bancos de dados em relação à origem de seus indivíduos. Por fim, vale mencionar a importância dessas análises também do ponto de vista antropológico, em estudos populacionais, visando elucidar rotas de migração e a origem do povoamento de uma região (MORIOT *et al*., 2018).

Figura 7 - Representação esquemática do processo de miscigenação



Legenda: Indivíduos de populações ancestrais diferentes, representadas por cromossomos azuis e vermelhos, acasalam-se ao longo das gerações, e recebem material genético de seus ancestrais, sujeito à recombinação durante a meiose.
Fonte: Adaptado de BAYE; WILKE, 2010.

Embora os marcadores uniparentais do cromossomo Y e do mtDNA apresentem alta diferenciação em termos de origem geográfica continental e sejam informativos de ancestralidade, não são capazes de representar a ancestralidade global do indivíduo como o são os marcadores dos autossomos, por não serem recombinantes ao longo das gerações e devido ao seu padrão de herança. Esses marcadores uniparentais contam histórias complementares e independentes acerca dos ancestrais maternos e paternos dos povos. Tal característica pode resultar em interpretações equivocadas, como no caso de linhagens parentais de origem muito distante da ancestralidade total em um indivíduo pertencente a uma população miscigenada, e que apresente co-ancestralidade, o que é marcadamente comum nas populações urbanas atuais. Outra característica a ser considerada em relação ao uso de marcadores autossômicos, e que pode ser vista como vantajosa, é que necessitam de um

número de indivíduos consideravelmente menor para a elaboração de um banco de dados populacional com o fim de estimar as frequências alélicas para cálculos estatísticos. As variações haplotípicas do mtDNA e do cromossomo Y exigem bancos de dados maiores para estimativas mais precisas, e se tratando de populações provenientes de regiões pouco investigadas isso se torna uma dificuldade considerável. Daí a importância de mais estudos genéticos desses marcadores e o depósito desses dados em plataformas como o YHRD e EMPOP.

Devido a diversos fatores envolvidos em sua formação, a população humana mundial não pode ser vista de maneira uniforme; ao contrário, apresenta uma estrutura complexa. Daí a importância de se estabelecer bem os grupos populacionais, tendo uma base adequada na escolha de conjuntos de marcadores para inferência de ancestralidade, levando-se em consideração tanto o número de marcadores quanto seu nível de resolução geográfica e dados disponíveis de populações de referência. Diferentes tipos de marcadores oferecem diferentes níveis de resolução sobre a estrutura populacional. Evolutivamente, os marcadores com taxas de mutação relativamente baixas (SNPs, Inserções Alu, Indels) são os melhores *loci* para a análise da história humana ao longo do tempo em escalas maiores, ou seja, fornecem uma resolução biogeográfica em nível continental. Já os marcadores de evolução rápida, como os STRs, fornecem maior resolução para escalas de tempo mais curtas (DE KNIJFF, 2000).

Um dos pioneiros em estudos sobre a diversidade genética populacional foi Lewontin (LEWONTIN, 1972) e posteriormente diversos outros estudos vem sendo feitos, que variam nos *loci* e nas populações analisadas, visando avaliar a estrutura dos grupos populacionais humanos modernos. Os resultados encontrados são similares aos de Lewontin, que descreveu que a maior parte das variações em autossomos (aproximadamente 85 %) ocorre dentro das populações. Algumas décadas depois, outro estudo da estrutura populacional foi feito com 1054 indivíduos pertencentes a 52 populações do Projeto de Diversidade do Genoma Humano (HGDP-CEPH) (ROSENBERG *et al*., 2002), utilizando o algoritmo de agrupamentos por similaridade genética do *software* STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000). As análises com o STRUCTURE para o conjunto de 377 STRs autossômicos usados identificaram 5 grupamentos continentais com similaridades genéticas, os quais foram definidos como Eurásia, África Subsaariana, Leste Asiático, América e Oceania. Admitindo-se um número de grupamentos igual a 7 no STRUCTURE, observa-se que a Eurásia se subdivide ainda em Europa, Oriente Médio e Centro-Sul Asiático. Apesar do tamanho reduzido de algumas das populações e lacunas consideráveis do ponto de vista geográfico, esse estudo com 377 STRs autossômicos ainda é considerado o mais informativo envolvendo

a população global. Poucos anos depois 2 outros estudos foram publicados, com um número maior de marcadores STR, sobre diversas populações da América e da Oceania isoladamente (WANG *et al*., 2007; FRIEDLAENDER *et al*., 2008). Esses estudos em conjunto forneceram as primeiras avaliações detalhadas sobre a distribuição da estrutura populacional global. Um novo estudo realizado usou 650000 SNPs para analisar o mesmo grupo de amostras HGDP-CEPH (LI *et al*., 2008) e os resultados encontrados foram muito semelhantes. Rosenberg e colaboradores também reanalisaram as amostras com um número muito maior de STRs (933), obtendo dados concordantes com o estudo anterior (ROSENBERG *et al*., 2005). Pode-se dizer que, apesar do detalhamento alcançado com o maior número de marcadores desses estudos, o padrão global de diversidade e da estrutura populacional humana permanece o mesmo encontrado inicialmente, com discretas, mas consistentes variações, refletindo as diferenças nas frequências alélicas. Essas tendem a aumentar sutilmente com a distância geográfica; mas pequenas descontinuidades existentes levam a separações, daí surgindo os grupamentos populacionais identificados pelo STRUCTURE Foi revelada também, através dos estudos realizados, uma diminuição da variabilidade genética ao longo do eixo África/Ásia – Eurásia – Oceania – América, que pode ser explicada por uma série de eventos fundadores ocorridos nas migrações humanas ao longo do tempo (PHILLIPS, 2015). Os testes de ancestralidade devem levar esses fatos em consideração, uma vez que para um mesmo grupo de marcadores, o nível de heterozigosidade em populações africanas é bem maior quando comparado ao encontrado na América, por exemplo (WANG *et al*., 2007). Pelas mesmas razões, as divergências encontradas entre populações africanas e não africanas usualmente são maiores do que entre outros pares de grupos. Sendo assim, haverá um número inferior de marcadores bem diferenciados entre populações mais próximas, como por exemplo, o par Eurásia – Leste Asiático.

Resumidamente, o que se conclui sobre as características observadas sobre a diversidade genética humana, é que indicam um padrão comum de migrações de grupos reduzidos de pessoas para novas regiões, havendo uma separação do grupo ancestral de origem seguida de rápida expansão. Esses efeitos de deriva genética tem sido uma forte força evolutiva atuante na formação da população humana contemporânea. Forças como a seleção natural, que varia regionalmente de acordo com fatores biogeográficos como clima, dieta e a presença de doenças, e também o fluxo gênico acentuado, presente em processos de miscigenação envolvendo populações diferenciadas, somam entre si na modelagem da diversidade humana. Mais recentemente, após o sequenciamento de genomas arcaicos de origem neandertal e denisovana, estudos realizados relatam uma ancestralidade neandertal

média de aproximadamente 2 % em populações modernas não africanas, e cerca de 7 % de ancestralidade denisovana em populações da Oceania. Esse fluxo gênico envolvendo o homem moderno e esses hominídeos é denominado introgressão arcaica (PHILLIPS, 2015; PICKRELL; REICH, 2014; REICH *et al*., 2011; HUERTA-SANCHEZ *et al*., 2014).

Inferências de ancestralidade geográfica usando marcadores genéticos dependem de três fatores essenciais, que são: um conjunto de AIMs, um banco de dados de amostras de referência genotipadas para os AIMs selecionados e um método estatístico de agrupamento por similaridade (PFAFFELHUBER *et al*., 2020).

A escolha dos marcadores precisa ser pautada em alguns parâmetros inerentes ao marcador em si, mas também ao nível de divergência que esse marcador apresenta entre as populações a serem avaliadas. As medidas de variabilidade de um marcador entre populações podem ser expressas por diferentes sistemas baseados nas distâncias genéticas existentes, sendo esse marcador de natureza bialélica, como os SNPs ou Indels, ou multialélica, como os STRs. As distâncias genéticas, por sua vez, estão correlacionadas com as diferenças das frequências alélicas existentes nas populações para o dado marcador, sendo um dos sistemas mais utilizados o índice F_{ST}, mas também a frequência alélica diferencial (δ) e informatividade (In) (ROSENBERG *et al*., 2003). Os valores dessas medidas variam de 0 a 1, e esses extremos representam, respectivamente, divergência nula e 100 % de divergência para comparações de pares de populações. Uma vez que se conheça esses valores para cada marcador frente às populações de interesse, é possível estabelecer um *ranking* de nível de diferenciação populacional para um conjunto de marcadores (SHRIVER *et al*., 1997; FRUDAKIS *et al*., 2003; GOLDSTEIN *et al*., 1995; ROSENBERG *et al*., 2003) e através de testes compactos obter inferências de ancestralidade. É fundamental, porém, verificar o balanço da divergência que os marcadores selecionados apresentam em conjunto entre os grupos populacionais envolvidos, além da disponibilidade dos dados populacionais disponíveis para esses grupos, necessários para as comparações.

Dado que a distribuição da diversidade humana levou a uma forte divergência entre as populações africanas e não africanas, seguida da divergência entre Eurásia e outras populações, e posteriormente, do Leste asiático com populações da Oceania e da América; sabe-se que a seleção de AIMs tende a encontrar muito mais *loci* informativos africanos do que para comparações de outros grupos. Já os americanos, como um grupo populacional com apenas 15 Ka de separação, têm a menor divergência em relação aos asiáticos do leste, intimamente relacionados (COLONNA *et al*., 2011). Sendo assim, se o objetivo de um teste de ancestralidade for diferenciar a África, a Europa, o Leste Asiático e a América, é mais

difícil encontrar marcadores que distingam europeus e asiáticos do que africanos e americanos. Isso demonstra como os valores de divergência precisam de avaliação criteriosa para as comparações populacionais em um teste, atingindo um balanço adequado no conjunto de marcadores selecionado. Isso é especialmente importante quando se analisa populações miscigenadas, que apresentam co-ancestralidade (TABOADA-ECHALAR *et al*., 2013).

Em relação aos bancos de dados populacionais disponíveis para consulta e seleção de AIMs, pode-se considerar que ainda há pouca informação sobre populações da Oceania, de nativos da América e algumas regiões da Ásia. Portanto, um objetivo válido seria caracterizar uma grande coleção de populações para um pequeno número de painéis de ancestralidade usando tamanhos de amostra razoáveis, em torno de 50 indivíduos ao menos. Porém, com o crescente uso de análises usando plataformas de sequenciamento paralelo massivo, tem sido possível disponibilizar um volume cada vez maior de informações, como os dados de SNPs do estudo com o painel HGDP-CEPH, com 650.000 *loci* (LI *et al*., 2008) e do *1000 Genomes Project Consortium*. Os dados da Fase 3 desse projeto atingiram a marca de aproximadamente 79 milhões de variantes em 2535 indivíduos provenientes de 26 populações (THE 1000 Genomes PROJECT CONSORTIUM, 2012). A partir da disponibilidade desses dados alguns painéis de AIMs com variações no número de SNPs genotipados, que utilizam a química do tipo *Snap-Shot*, de extensão de base única, foram desenvolvidos e posteriormente otimizados (PHILLIPS *et al*., 2007; FONDEVILA *et al*., 2013; KERSBERGEN *et al*., 2009; LAO *et al*., 2006; GETTINGS *et al*., 2014). Subsequentemente, outros painéis contendo maior número de AIMS, com maior abrangência de grupos populacionais e boa capacidade de resolução também foram desenvolvidos e aperfeiçoados. Dois conjuntos de AIMs, compatíveis com plataformas de MPS, com 128 e 55 SNPs, foram combinados (KIDD *et al*., 2011; PHILLIPS *et al*., 2014a), formando o *HID-Ion AmpliSeq Ancestry Panel* , otimizado para o sistema *NGS Ion PGM*, ambos da *Thermo Fisher Scientific*. Além dos dados publicados pela página eletrônica do *1000 Genomes Project Consortium*, frequências alélicas e dados genotípicos de SNPs também podem ser obtidos de outros bancos de dados de consulta pública, como SPSmart, FROGkb e Alfred (AMIGO; SALAS; PHILLIPS, 2011; RAJEEVAN *et al*., 2012a e 2012b).

Em relação aos sistemas estatísticos para inferências de ancestralidade biogeográfica, podem ser citados: Análise Bayesiana, Análise dos Componentes Principais (PCA) e o *software* STRUCTURE, que usa a abordagem Bayesiana. Cada um desses sistemas usa dados de populações de referência e faz inferências através da comparação dos padrões de variação encontrados nessas populações e na população com ancestralidade desconhecida. Por isso, a

relevância, a qualidade e a abrangência dos dados populacionais disponíveis são fatores importantes a serem considerados na análise (PHILLIPS, 2015).

O STRUCTURE é o programa de análise populacional mais amplamente usado atualmente, podendo ser usado com marcadores do tipo SNP e STR simultaneamente (FALUSH; STEPHENS; PRITCHARD, 2003). Nesse *software*, como dito anteriormente, uma vez que a estimativa de agrupamentos ou *clusters* populacionais é feita para as populações de referência, é possível inferir a ancestralidade individual em uma nova amostra, por comparação. A cada "corrida" o STRUCTURE produz uma matriz de coeficientes de associação aos agrupamentos populacionais, permitindo comparações entre coeficientes das amostras de referência e da amostra de ancestralidade desconhecida. A análise de agrupamento normalmente não é supervisionada, portanto as amostras não são rotuladas por região e, consequentemente, os agrupamentos são determinados apenas por padrões de similaridade genética detectada entre as amostras. Este processo é frequentemente usado para testar a eficiência de um conjunto de marcadores para diferenciar grupos específicos, ou seja, se os *clusters* formados corresponderem bem à região geográfica de origem, o painel pode ser considerado informativo para os grupos que foram analisados. Uma variável fundamental na análise pelo STRUCTURE é a estimativa do número de grupos populacionais (K) a serem considerados, uma vez que esse precisa ser coerente com a real estrutura genética das populações analisadas. Nem sempre é simples de se determinar o melhor valor de K, e um *software* que pode auxiliar nessa escolha é o CLUMP (JAKOBSSON; ROSENBERG, 2007; KALINOWSKI, 2011).

A miscigenação é uma característica dominante das populações que habitam as margens dos continentes e uma ocorrência cada vez mais crescente, que se iniciou desde a primeira migração de pequenos grupos humanos. Ao longo de mais de 2 mil anos de história de conquistas, tráfico de escravos e crescente comércio, além das movimentações de massas e a urbanização, dos séculos mais recentes; barreiras culturais e sociais, que antes substituíam a separação geográfica entre os povos, tem sido cada vez menores. Consequentemente, as análises de ancestralidade genética deparam-se atualmente com uma grande variabilidade de padrões nas proporções de mistura entre os indivíduos testados. Na área forense, os investigadores criminais também têm interesse particular nas análises de co-ancestralidade porque esta abriga a possibilidade de combinações incomuns de características físicas em um determinado suspeito. Comparando-se as abordagens de inferência aqui citadas, em relação à capacidade de estimar a ancestralidade em indivíduos miscigenados, pode-se dizer que a de maior limitação é a análise Bayesiana, seguida da PCA, na qual é aconselhavel usar

comparações considerando até 3 populações ancestrais por análises, e finalmente pelo STRUCTURE, mais frequentemente usado (PHILLIPS, 2015). A análise com o STRUCTURE compara os padrões de grupamentos encontrados com os padrões de indivíduos de origem ancestral conhecida (FONDEVILA *et al*., 2013). É importante ressaltar que limitações existem nessa abordagem, como por exemplo o fato de que as populações originais que contribuíram para a miscigenação não podem ser extrapoladas de forma 100 % eficiente a partir de amostras (KIDD *et al*., 2011).

Além dos marcadores AIMS do tipo SNPs, que são os mais clássicos, outros tipos de marcadores também podem ser informativos de ancestralidade, como os Indels, STRs autossômicos e marcadores SNPS multialélicos. Os Indels, que são marcadores bialélicos, assim como os SNPs, analiticamente também poderem ser analisados em fragmentos de PCR curtos e sistemas de multiplexes com *primers* marcados com fluoróforos. Porém são processados diretamente da PCR para a eletroforese capilar (CE), sem etapas intermediárias. Embora existam SNPs mais informativos que os Indels, estes apresentam a vantagem de distinguir melhor picos provenientes de misturas (PEREIRA *et al*., 2009).

Diversos painéis AIM-Indels já foram desenvolvidos, entre eles um painel com 48 marcadores distribuídos em 3 mutiplexes (SANTOS *et al*., 2010), um painel com 46 indels em 1 único multiplex (PEREIRA *et al*., 2012a) e um painel de 21 indels também em 1 multiplex (ZAUMSEGEL; ROTHSCHILD; SCHNEIDER, 2013). O painel de 46 AIM-Indels apresenta divergências entre África, Europa e Leste Asiático comparáveis ao painel de 34 SNPs - 34 plex (FONDEVILA *et al*., 2013; PHILLIPS *et al*., 2007), mas também diferencia os nativo americanos. Por essa razão é uma opção de inferência de ancestralidade genética de simples execução técnica, podendo ser realizada com um único teste de PCR – eletroforese capilar.

Em relação aos marcadores do tipo STR autossômicos, dois tipos de abordagens podem ser usadas: um painel com um grande número de marcadores ou com marcadores de alto poder de diferenciação entre as populações. Quanto à estrutura populacional global, verificou-se que no conjunto de 377 marcadores STRs usados para inferir ancestralidade, marcadores selecionados ao acaso foram mais eficientes quando comparados com marcadores SNPs escolhidos também aleatoreamente, e alguns STRs foram considerados altamente informativos em relação aos SNPs, principalmente os que tem sequência de repetição de dinucleotídeos (ROSENBERG *et al*., 2003).

Por serem marcadores amplamente utilizados na genética forense, utilizar os dados provenientes de STRs para inferir ancestralidade pode ser um campo com grande potencial

de investigação. Alguns estudos já utilizaram, por exemplo, marcadores dos *kit*s comerciais *Identifiler* (PHILLIPS *et al*., 2011), *Identifiler Plus* (LONDIN *et al*., 2010), *Global Filer*, *Fusion*; isoladamente ou combinados entre si ou com outros STRs adicionais e SNPs. A existência de alelos STR específicos de determinadas populações já foi relatada (PHILLIPS *et al*., 2014b), sendo o de maior especificidade o alelo 9 do marcador D9S1120, muito frequente entre os nativos americanos (53 %) (PHILLIPS *et al*., 2008). Um ensaio multiplex com 12 marcadores STR informativos de ancestralidade com repetições de tetranucleotídeos foi desenvolvido, apresentando bom nível de resolução, porém uma taxa maior de erro na detecção da ancestralidade africana (PHILLIPS *et al*., 2013).

Embora o desempenho dos STRs isoladamente ainda não seja considerado promissor como AIMs, esforços crescentes vem sendo feitos, como adaptações de *software* para uso desse tipo de dado e a criação de bancos de dados com perfis e frequências genotípicas (PEREIRA *et al*., 2011), buscando minimizar as distorções nas interpretações dos resultados.

As técnicas de NGS, além de permitirem a ampliação de painéis de AIM-SNPs e AIM-Indels, também podem auxiliar no aumento do nível de informação destes. Isso porque as diversas sequências geradas de um fragmento de centenas de bases permitem genotipar simultaneamente todas as variantes presentes no mesmo, como SNPs ou Indels embutidos em regiões de STRs ou múltiplos SNPs formando haplótipos; muitos dos quais apresentam uma distribuição informativa de ancestralidade. As combinações alélicas ou haplótipos para cada cromossomo, ou seja, sua fase, podem ser determinados, mesmo nos autossomos, o que o sequenciamento de Sanger não possibilitava anteriormente. Esses haplótipos foram classificados, de acordo com o seu tamanho em kb, em minihaplótipos (1–10 kb) e microhaplótipos (200 pb) (PAKSTIS *et al*., 2012; KIDD *et al*., 2014; MORIOT *et al*., 2018). Uma avaliação cuidadosa da taxa de recombinação na região em que ocorrem é necessária para estimar seu grau de informação. Embora taxas de recombinação muito baixas ajudem a preservar os haplótipos em maiores extensões, chegando a quilobases, alguma recombinação é necessária para gerar frequências de haplótipos informativos entre as populações. Da mesma forma, fragmentos muito curtos precisam de maior atividade de recombinação para gerar novos conjuntos de alelos.

Um exemplo desses *loci* haplotípicos são os recentemente descritos marcadores compostos DIP-STR, que apresentam a vantagem de possuir tanto marcadores Indels quanto marcadores STR,combinando os benefícios desses dois tipos de marcadores. A instabilidade do marcador STR faz com que múltiplos alelos surjam, proporcionalmente ao tempo de divergência das populações, enquanto o marcador SNP, por ser mais estável, permite maior

precisão no rastreamento da linhagem de cada haplótipo. Uma outra vantagem é que o número de haplótipos possíveis formado pela combinação desses marcadores é muito maior do que o número de alelos em cada *locus* individual. Essa maior variabilidade aumenta a probabilidade de haplótipos raros, que são facilmente perdidos durante a ocorrência de eventos de deriva populacional, como o efeito fundador e o efeito gargalo (MORIOT *et al*., 2018).

Outra espécie genômica que pode ser informativa para a ancestralidade são os SNPs multialélicos. Inicialmente foram pouco utilizados, não sendo considerados na Fase I do *1000 Genomes Project Consortium*; mas atualmente foram bem caracterizados e constituem um em cada 300 SNPs da Fase III desse projeto. Alguns desses SNPs têm grande variabilidade entre populações, e são também marcadores úteis na deteção de misturas de DNA (WESTEN *et al*., 2009). Possuem maior variabilidade, se comparados aos SNPs bialélicos, devido ao fato de fornecerem um maior número de combinações alélicas possíveis, e a distribuição geográfica desses alelos, influenciada pelos processos de deriva, pode resultar em um alto grau de informação de ancestralidade.

A formação da população brasileira e estudos genéticos de ancestralidade

A população brasileira é considerada uma das mais heterogêneas do mundo, sendo resultado da miscigenação entre as três principais populações ancestrais que também povoaram a América Latina, nomeadamente os povos nativos da América, os europeus e os africanos.

O território brasileiro, que era até então habitado por diferentes grupos nativos (também chamados de indígenas), começou a ser ocupado pelos primeiros grupos europeus que chegaram ao Brasil, de origem portuguesa. Dados históricos relatam que no mês de Abril, em 1500, os portugueses, liderados por Pedro Álvares Cabral, chegaram ao território brasileiro, na região do litoral da Bahia; fato este que constitui um dos episódios da expansão marítima portuguesa, no início do século XV.

Nos primeiros dois séculos de colonização, vieram para o Brasil cerca de 100 mil portugueses, uma média anual de 500 imigrantes. No século seguinte, esse número aumentou: foram registrados 600 mil e uma média anual de 10 mil imigrantes portugueses. O ápice do fluxo migratório ocorreu na primeira metade do século XX, entre 1901 e 1930: a média anual ultrapassou a barreira dos 25 mil. A origem socioeconômica do português imigrante é muito

diversificada: de uma próspera elite nos primeiros séculos de colonização, passou-se a um fluxo crescente de imigrantes pobres a partir da segunda metade do século XIX (IBGE, 2000).

Ao chegarem ao Brasil, os portugueses encontraram populações indígenas que ocupavam o território de forma bem distribuída ao longo da costa litorânea e na região da bacia dos Rios Paraná e Paraguai (abrangendo os estados de São Paulo, Paraná, Mato Grosso do Sul, Minas Gerais, Goiás, Santa Catarina e a região do Distrito Federal). Segundo estudos antropológicos recentes, baseados em afinidades culturais e linguísticas, essas populações compunham dois grandes grupos bem distintos: os tupis-guaranis, que se estendiam por quase toda a costa brasileira, e os tapuias, presentes em alguns pontos do litoral. Esses englobam, por exemplo, os goitacases na foz do Rio Paraíba, os aimorés no sul da Bahia e no norte do Espírito Santo e os tremembés na faixa entre o Ceará e o Maranhão (FAUSTO, 1996). Devido à dificuldade de obtenção de dados sobre os povos indígenas que aqui habitavam na época da chegada dos portugueses, existe uma grande variação nas estimativas quanto ao seu número. Segundo dados do IBGE, estima-se que a população indígena era de, aproximadamente, 2,5 milhões de indivíduos (IBGE, 2000). Praticavam a agricultura, a pesca, a caça e atividades extrativistas, sendo suas relações econômicas basicamente de subsistência e destinada ao consumo próprio.

A chegada dos portugueses iniciou-se com o reconhecimento e a posse da nova terra, através de uma ocupação crescente, em um período colonial que se consolidou ao longo de mais de dois séculos. Essa ocupação, feita também sob o pretexto religioso de catequização dos índios pelos representantes da Igreja Católica, teve como objetivo principal a exploração dos recursos naturais, que começou pelo litoral, com as derrubadas de árvores de pau-brasil pelos próprios indígenas, em troca de objetos de pouco valor. Nesse primeiro momento ocorreu caracteristicamente uma miscigenação entre mulheres indígenas e homens portugueses, visto que a vinda de mulheres europeias ainda era pouco comum (FAUSTO, 1996). Com o objetivo de expandir a colonização da nova terra, a Coroa portuguesa inicia a instalação de engenhos de açúcar e através de incentivos sociais e econômicos promove a formação de latifúndios, concentrados sob o poder de poucos proprietários que tinham ligações com o governo. Diferente das atividades iniciais da colônia, às quais os indígenas estavam familiarizados, como o corte do pau-brasil, as atividades agrícolas às quais os colonizadores queriam impor aos nativos intensificaram conflitos já existentes. O fato é que a conquista do território brasileiro provocou uma drástica diminuição na população indígena, que apesar de ter resistido fortemente aos colonizadores, sobretudo quando se tratou de

escravizá-los, sofreu grande violência cultural, e muitas mortes decorrentes de epidemias e desses conflitos.

De acordo com os resultados do Censo Demográfico 2010, realizado pelo IBGE, 817,9 mil pessoas se declararam indígenas, representando apenas 0,4 % da população total do Brasil. Sabemos que nesse discreto contingente auto declarado indígena inclui-se também uma parte da população mestiça, resultante do contato com os europeus. Esse último censo, em contraste com censos anteriores, que adotavam apenas os quesitos de cor e raça, aprimorou a investigação desse contingente populacional em particular, introduzindo o pertencimento étnico, a língua falada no domicílio e a localização geográfica. Se forem consideradas apenas as pessoas que vivem nas terras indígenas, de uma estimativa de 2,5 milhões para o século XVI, em 1998 chegou-se a um total de apenas 302.888 índios (IBGE, 2000; IBGE, 2012).

Apesar de tantas dificuldades enfrentadas pelos indígenas, desde a época do Brasil colônia, a preservação que se tem hoje, de sua herança biológica, social e cultural, foi possível principalmente pelos deslocamentos contínuos que faziam para regiões cada vez mais distantes, sendo esse isolamento uma forma muito eficaz de resistência.

Nas décadas posteriores à chegada dos primeiros grupos de portugueses, houve a vinda tanto de outros grupos europeus como também de escravos africanos trazidos por frotas portuguesas a partir de diferentes locais da África. A introdução do trabalho escravo a partir de meados do século XVI pelos conquistadores atendia idealmente aos seus interesses, uma vez que nem havia grande oferta de trabalhadores interessados em vir para o Brasil, nem o trabalho assalariado era conveniente para os fins da colonização. Por outro lado, havia uma grande disponibilidade de terras e a necessidade de sua efetiva ocupação produtiva.

Os escravos africanos começaram a ser trazidos para o Brasil para trabalhar principalmente na economia açucareira e nas casas de famílias portuguesas. Posteriormente, trabalharam também nas minas de ouro e diamantes e nas plantações de café. Uma vez que o índio não se adaptou bem ao trabalho compulsório, essa transição da escravidão do indígena para o negro africano foi se dando gradualmente. Estima-se que entre 1550 e 1850 (quando foi criada a lei que abolia o comércio de escravos), aproximadamente, 4 milhões de escravos africanos vindos de diferentes regiões da África, em grande maioria do sexo masculino, entraram pelos portos brasileiros (IBGE, 2000). Os locais com maior influxo de escravos africanos foram Salvador, que recebeu indivíduos em sua maioria da região de Guiné, e o Rio de Janeiro, que recebeu escravos principalmente de Angola (Figura 8). Diferente dos índios, os africanos não conheciam o território, sendo mais difícil fugir, além de ter mostrado uma

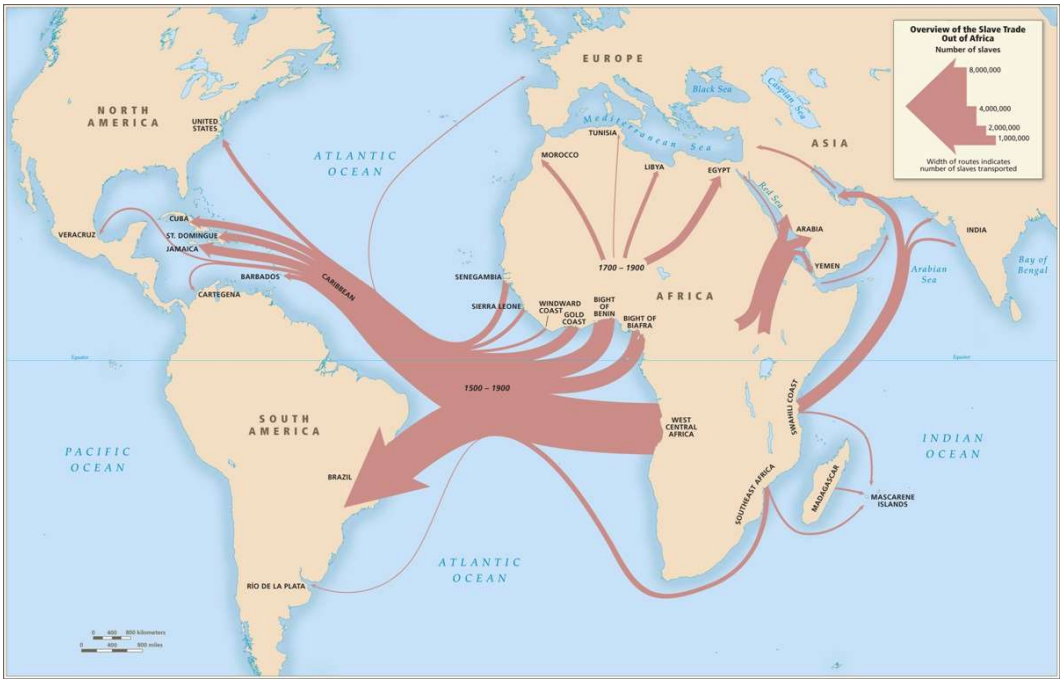
boa capacidade produtiva. Isso fazia deles uma ótima opção à escravização dos índios. Porém, fugas individuais e formações de quilombos, que eram locais de refúgio de escravos que escapavam e se organizavam, eram comuns no Brasil colonial, sendo um dos mais conhecidos o de Palmares, no Estado de Alagoas (FAUSTO, 1996).

As subseqüentes e contínuas migrações de escravos africanos e de europeus foram contribuindo para uma maior mistura na população do Brasil, que ia se formando diferencialmente. Os negros tinham alta representatividade como etnia brasileira em meados do século XIX, mas esse cenário começou a mudar devido a ondas de migrações incentivadas pelas autoridades brasileiras como forma de "branqueamento" da população, transformando o Brasil em um país predominantemente branco, principalmente nas regiões Sudeste e Sul (PENA *et al*., 2011).

Além dos portugueses, houve invasões no nordeste brasileiro, inicialmente de franceses, que se estabeleceram no Maranhão e devido aos riscos de perda territorial atraíram para ali os portugueses, que através do Rio Amazonas, foram avançando para a Região Norte, fundando, em 1616, a cidade de Belém. Posteriormente, os holandeses chegaram ao estado de Pernambuco, em busca do açúcar, influenciando, sobretudo, na formação da população dessa região.

A colonização do norte do Brasil, diferentemente das outras regiões, ocorreu lentamente, com predomínio do trabalho compulsório indígena. Os primeiros negros chegaram à Amazônia não por intermédio dos portugueses, mas graças aos ingleses. Nas feitorias que montaram entre a costa de Macapá e a zona de estreitos, esses ingleses pretenderam realizar um empreendimento agrário de vulto, principalmente, de plantio de cana para a fabricação de açúcar e rum (FAUSTO, 1996; DE CAMPOS; DOLHNIKOFF, 2001). A chegada de escravos africanos em larga escala iniciou-se efetivamente apenas em meados do século XVIII, com estimativas de 53000 indivíduos que chegaram durante o período da escravidão (MIRANDA-NETO, 1976; PALHA *et al*., 2011). Já o Sudeste teve sua colonização inicial através do litoral e a presença maciça de indígenas, que auxiliaram o avanço para o Centro-Sul do país, em direção ao interior. A presença de mestiços resultantes do cruzamento entre brancos e mulheres indígenas (mamelucos) era muito comum no estado de São Paulo, devido ao reduzido número de mulheres brancas.

Figura 8 - Rotas do tráfico de escravos africanos para o Brasil



Legenda: Principais rotas de influxo de escravos africanos para o Brasil. A maioria dos escravos africanos era proveniente das regiões Centro-Oeste e Sudoeste da África, principalmente das regiões da Guiné, Congo e Angola.
Fonte: http://www.slavevoyages.org/static/images/assessment/intro-maps/01.jpg

Na história recente do Brasil, após a abolição da escravidão em 1888, houve a vinda de imigrantes de outras regiões da Europa (Itália, Espanha e Alemanha, principalmente), e da Ásia (Síria, Líbano, China e Japão) (IBGE, 2000), que ocuparam principalmente o sul e sudeste do país; além de intensas migrações internas devido a motivações de ordem econômica.

Em conjunto, essas características da formação da população brasileira resultaram em padrões de miscigenação variáveis ao longo das cinco regiões geopolíticas nas quais o país é atualmente subdividido: Norte, Nordeste, Sudeste, Centro-Oeste e Sul. Ou seja, o território brasileiro apresenta uma distribuição heterogênea das contribuições ancestrais dos nativos americanos, europeus e africanos, uma vez que os encontros e acasalamentos entre esses ocorreram de maneira assimétrica, com europeus e africanos chegando pelo litoral e gradualmente alcançando o interior do país, para onde houve intensa migração de indígenas devido aos conflitos com os europeus (SALOUM DE NEVES MANTA *et al*., 2013).

A heterogeneidade da população brasileira ao longo do seu território tem sido foco de interesse em muitos estudos, que usam diferentes tipos de marcadores genéticos. Em alguns deles, cada região geográfica é tratada como uma única população homogênea, enquanto outros estudos criam diferentes subdivisões: políticas (por exemplo, agrupando populações

por estado), demográficas (por exemplo, populações urbanas e rurais) ou étnicas (por exemplo, cultura, auto-declaração, ou cor da pele).

Os primeiros estudos da diversidade genética da população brasileira analisaram grupos sanguíneos e marcadores protéicos, resultando em uma análise ampla, porém superficial, da heterogeneidade dos brasileiros (SCHNEIDER; SALZANO, 1979). Os marcadores de linhagem também têm sido usados para a compreensão do complexo processo de miscigenação ocorrido no Brasil, sendo capazes de revelar os padrões de cruzamentos interétnicos ocorridos entre africanos, europeus e nativos (PENA *et al.*, 2009; RIBEIRO-DOS-SANTOS *et al*., 2002). Estudos de ancestralidade usando marcadores autossômicos com diferentes tipos de polimorfismos vêm sendo realizados para elucidar padrões de miscigenação em populações brasileiras, uma vez que esses marcadores são informativos da ancestralidade global dos indivíduos (LINS *et al*., 2010; PEREIRA *et al*., 2012a; SANTOS *et al*., 2010).

Salzano e Sans (2014), em um trabalho de revisão da literatura, reuniram dados publicados de diversos trabalhos sobre a população brasileira; sendo 13 deles realizados com marcadores de linhagens associados ou isoladamente, e 27 com marcadores autossômicos. As amostragens e o número de marcadores desses estudos foram variáveis, havendo amostras representativas das 5 regiões geopolíticas brasileiras, assim como de estados e cidades, e até mesmo grupos específicos, como comunidades afro-brasileiras. Nessa compilação de dados foi encontrado o predomínio da proporção de ancestralidade europeia, que é mais acentuada nas regiões sul e sudeste, seguida pela proporção de ancestralidade africana, em maior proporção no nordeste, e por fim o componente nativo, que é mais evidente na região Norte. A região Centro-Oeste, de acordo com estes autores, se assemelha à região Norte do país.

Considerando-se apenas a ancestralidade paterna, análises de marcadores SNP do cromossomo Y apontam na mesma direção, demonstrando que a maioria das linhagens masculinas de populações urbanas do Brasil é de origem europeia, seguida por um menor número de linhagens africanas (provenientes, principalmente, da África subsaariana) ou nativas-americanas, havendo alguma variabilidade entre as cinco regiões geopolíticas do Brasil (JANNUZZI *et al*., 2020; PENA *et al*., 2009; RESQUE *et al*., 2016). Segundo o trabalho de Resque e colaboradores (2016), as regiões Centro-oeste e Sul são as que apresentam os maiores valores de ancestralidades europeias do país, enquanto que as regiões Norte e Nordeste possuem os valores mais altos de ancestralidades nativa americana e africana, respectivamente, observados em populações brasileiras.

As análises realizadas com o mtDNA isoladamente exibem uma maior ancestralidade europeia se considerado o *genepool* da população brasileira como um todo (ALVES-SILVA *et al*., 2000). Porém, diferenças marcantes entre as regiões do Brasil são encontradas, com predomínio das linhagens europeias apenas no sul do país. Já na região Norte o predomínio é nativo, enquanto no Nordeste e maior parte do sudeste as linhagens africanas são maioria (ALVES-SILVA *et al*., 2000; FRIEDMAN *et al*., 2014; SIMÃO *et al*. 2018).

Com marcadores biparentais informativos de ancestralidade, análises realizadas em populações brasileiras mostram que a ancestralidade europeia é a mais prevalente em populações urbanas, com um gradiente crescente do norte para o sul do país, chegando a valores de até 79 %. As ancestralidades africana e nativa estão mais presentes nas regiões Nordeste e Norte, respectivamente (SALOUM DE NEVES MANTA *et al*., 2013; PENA *et al*., 2011).

Considerando-se a demografia do Brasil, além dos diferentes tipos de marcadores e critérios de amostragem utilizados nos estudos já realizados, a ancestralidade dos brasileiros em relação às contribuições europeia, africana e nativa ainda não é totalmente conhecida. Portanto, análises adicionais com AIMs, usando maiores tamanhos amostrais e que incluam populações ainda não caracterizadas, podem revelar novos aspectos da estrutura genética brasileira. No levantamento realizado por Salzano e Sans (2014), observa-se que poucas populações miscigenadas da região Norte do Brasil já foram investigadas quanto à ancestralidade genética com base em marcadores autossômicos, como Macapá e Belém.

A Ilha de Marajó

A Ilha de Marajó está localizada ao norte do estado do Pará, distante aproximadamente 90 km da capital Belém. Com cerca de três mil ilhas e ilhotas, o Marajó é o maior arquipélago flúvio-marítimo do mundo, banhado pelo Oceano Atlântico e pelos rios Amazonas e Tocantins. Sua ilha principal tem uma extensão de 50.000 m², sendo dividida administrativamente em 16 municípios: Afuá, Anajás, Bagre, Breves, Cachoeira do Arari, Chaves, Curralinho, Gurupá, Melgaço, Muaná, Ponta de Pedras, Portel, Santa Cruz do Arari, São Sebastião da Boa Vista, Salvaterra e Soure (Figura 9). A distribuição da população do Marajó apresenta pequeno predomínio da população rural com cerca de 60 %, de acordo com o IBGE, 2010, enquanto a média do país é de 16 %. Dos 16 municípios, apenas três possuem

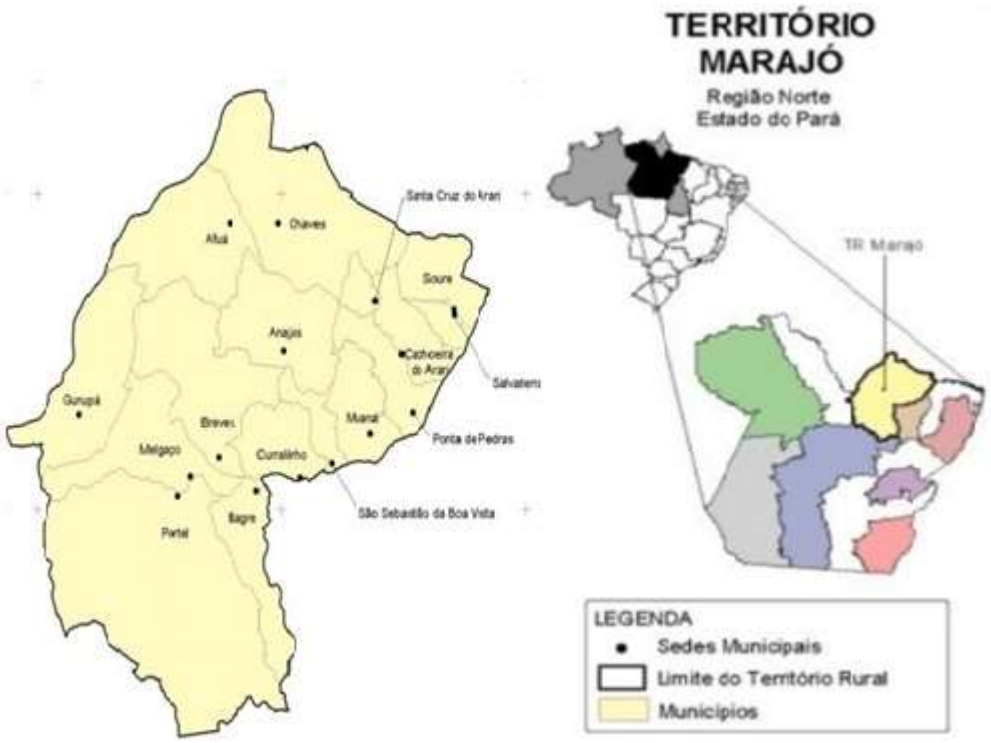
população urbana superior à rural: os municípios de Breves, Soure e Salvaterra (RELATÓRIO ANALÍTICO DO TERRITÓRIO DO MARAJÓ, 2012). Os dois últimos se destacam por serem os mais populosos, com 17000 e 22000 habitantes, respectivamente, sendo a população total da ilha de 250.000 habitantes (www.infoescola.com/geografia/ilha-de-marajo).

A teoria mais aceita sobre a origem do nome Marajó faz menção às observações de índios nativos da ilha, que a denominaram de "Mibaraió", e que em língua tupi significa "anteparo do mar" ou "tapamar". No período colonial era denominada Ilha Grande de Joannes, recebendo o nome atual no século XIX, à época da independência (MORIM, 2014).

O Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) identifica, através de sítios arqueológicos encontrados no Marajó, os sambaquis, indícios da mais antiga ocupação da ilha. Essas formações são enormes montanhas erguidas em baías, praias ou na foz de grandes rios por povos que habitaram o litoral do Brasil na Pré-História e constituem-se principalmente por cascas de moluscos – a própria origem tupi da palavra sambaqui significa "amontoado de conchas". Além desses, os achados arqueológicos apontam para sociedades relativamente autônomas, organizadas em pequenas vilas familiares, vivendo de uma economia de subsistência.

Posteriormente, sociedades complexas que se caracterizaram pelo manejo de terra e de recursos hídricos espalharam-se por toda a ilha, especialmente na área de campos, junto a cabeceiras de rios e igarapés, mas ocupando também a área de floresta. Os sítios arqueológicos coloniais remontam à época do contato com os europeus. São vilas, igrejas, engenhos, fazendas, chalés, com estruturas arquitetônicas e outras evidências materiais datadas do período colonial. Esses remanescentes testemunham um longo processo histórico ocorrido na ilha, e podem auxiliar na compreensão de sua dinâmica cultural, como por exemplo, sobre as formas de contatos interétnicos entre os povos nativos, europeus e africanos (RELATÓRIO ANALÍTICO DO TERRITÓRIO DO MARAJÓ, 2012).

Figura 9 - O Território da Ilha de Marajó



Fonte: Secretaria de Desenvolvimento Territorial/ Ministério de Desenvolvimento Agrário.

Registros da ocupação humana no Arquipélago do Marajó são escassos, havendo relatos da chegada dos primeiros conquistadores, os jesuítas, em 1615, a partir da embocadura do Rio Amazonas. Nas décadas seguintes, a Coroa Portuguesa enviou observadores ao arquipélago, que encontraram ali, surpreendentemente, numerosas fazendas de gado e intensa lavoura em posse dos missionários. A partir daí copiosos relatórios sobre o arquipélago eram enviados ao rei por diversos estudiosos, entre eles o naturalista brasileiro Dr. Alexandre Rodrigues Ferreira (PEREIRA, 1956). Não existem registros escritos pelos jesuítas sobre o comportamento e o modo de vida dos seus habitantes pré-coloniais no trabalho no campo, no lar, isolada ou coletivamente. Assim, a necessidade de se realizar pesquisas sistemáticas ou mesmo o controle e a proteção para a preservação dos sítios arqueológicos existentes na região são de extrema importância (RELATÓRIO ANALÍTICO DO TERRITÓRIO DO MARAJÓ, 2012), a fim de se compreender esses aspectos etno-sociais.

Acredita-se que cerca de 30 diferentes nações indígenas habitavam o Marajó à época da colonização portuguesa. Esses povos, chamados genericamente de “Nheengaíbas” (em Tupi quer dizer "gente de fala incompreensível"), falavam línguas diferentes e localizavam-se

na parte central da ilha. Dentre as nações localizadas no território marajoara, estão: Aruãns, Sacacas, Marauanás, Caiás, Araris, Anajás, Muanás, Mapuás, Pacajás, entre outras (PACHECO, 2010). Apesar de terem grande importância para a formação da população do Marajó, ao final do século XVIII, uma grande parte da população indígena foi dizimada devido a "guerras" com os portugueses, ou havia fugido das expedições portuguesas que escravizavam os índios duramente em lavouras e cidades, tomando para si suas terras (MORIM, 2014). O genocídio dos povos marajoaras, entre eles os Anajás, foi de extrema violência, não deixando, por exemplo, informações sobre as línguas por eles faladas (glotocídio). Os colonizadores, para sanar os problemas relacionados ao multilinguismo amazônico, deram início à imposição do uso da Língua Geral Amazônica (LGA), que era baseada no tupinambá falado pelos índios (BATISTA; NOGUEIRA, 2016; FREIRE, 2003).

Nesse período a região amazônica integrou-se ao mercado mundial através da exploração mercantil e as fazendas e engenhos do século XVIII e XIX continuaram a utilizar-se largamente do trabalho de escravos trazidos da África e homens livres, estes últimos indígenas e mestiços. O Estado do Pará, em 1751, possuía 24 engenhos, classificados de "reais", por serem completos e bem equipados. Em meados do século XIX, Marajó contava com cerca de 2000 escravos negros. Mas a decadência da produção do açúcar pela predileção da produção de aguardente, que era mais lucrativo, levou a região da Amazônia a entrar no Ciclo da Borracha, retrocedendo ao extrativismo predatório. Levas migratórias de colonos lusos se estabelecem na foz do Rio Amazonas, provenientes da África e das ilhas dos Açores. Começa assim a colonização efetiva do arquipélago, onde o gado vai representar papel relevante: os currais de influencia indígena são construídos para o rebanho à medida que os fazendeiros avançam para o interior da ilha (MIRANDA-NETO, 1976).

Dentre alguns estudiosos, o paraense José Veríssimo, publicou relatos sobre as populações indígenas e mestiças da Amazônia, em "Scenas da Vida Amazônica" (VERÍSSIMO, 1899). Nele relatou suas observações sobre os processos de miscigenação que se dera na Ilha Grande de Joannes. Segundo esse crítico "o primeiro colono foi polígamo, as escravas índias faziam um harém aos soldados da conquista. Posteriormente as escravas negras também viveriam esse mesmo tipo de situação. Segundo ele, primeiramente o branco cruzou com o índio, depois o negro com este, e com aquele, e com os resultados destes sucessivos cruzamentos; daí resultou a grande mistura de "sangues" que produziu o curiboca (branco e índio), o mameluco (curiboca e branco), o mulato (branco e preto) e o cafuzo (preto e índio) e ainda outros, de entrelaçamentos destes (PEREIRA, 1956).

A resistência à escravidão mediante fugas deu origem à formação dos quilombos nas várias regiões do arquipélago. Documentos históricos mostram que no decorrer do século XVIII foram muitas as situações e movimentos de fugas da população escravizada, composta tanto por negros quanto por índios. A população de negros, indígenas e mestiços na ilha, a essa altura, correspondia a mais de 80 % da população local. Atualmente, os vaqueiros e capatazes dos latifúndios, que foram passados por herança aos atuais proprietários, descendem, em sua maioria, de antigos escravos que passaram tecnicamente à condição de agregados e dependentes após 1888 (RELATÓRIO ANALÍTICO DO TERRITÓRIO DO MARAJÓ, 2012).

Estudos de genética populacional podem contribuir significativamente para um melhor conhecimento dessa população quanto à contribuição dos diferentes grupos étnicos envolvidos em sua formação. Embora a caracterização genética da população da Ilha de Marajó ofereça interessantes possibilidades ao nível do estudo de sua origem e história, há muito conhecimento ainda por extrair, uma vez que não se encontra na literatura dados consistentes sobre suas origens, o nível de diversidade genética e subestrutura populacional.

Até a presente data apenas um trabalho foi publicado no sentido de avaliar a ancestralidade da população da Ilha de Marajó, com base em marcadores moleculares. Palha e colaboradores (2011) analisaram nove marcadores STR do cromossomo Y de 300 indivíduos de nove comunidades quilombolas da região amazônica (n=300). Dessas comunidades uma era do município de Ponta de Pedras, na Ilha de Marajó (n=58). Para o conjunto das populações quilombolas estudadas, foi observada uma elevada diversidade haplotípica (0,989), com uma maior percentagem de linhagens africanas (53,6 %), seguida pelas europeias (41,4 %) e com uma minoria de linhagens nativo americanas (5 %). Para os 58 indivíduos representantes do Marajó, a diversidade haplotípica foi menor (0,942),tendo sido observadas somente linhagens europeias (61,9 %) e africanas (38,1 %) (PALHA *et al*., 2011).

2 **MATERIAIS E MÉTODOS**

Este estudo foi realizado no Laboratório de Diagnósticos por DNA (LDD), do Departamento de Ecologia, Instituto de Biologia Roberto Alcantara Gomes, Universidade do Estado do Rio de Janeiro (UERJ). A coleta das amostras biológicas foi realizada em colaboração com o pesquisador Dr. Luiz Marcelo de Lima Pinheiro, do Laboratório de Biologia na Saúde da Universidade Federal do Pará – Campus Soure, nos municípios de Soure e Salvaterra.

2.1 **Amostra populacional**

Ao longo deste trabalho, foi analisada uma amostra populacional de indivíduos pertencentes a ambos os sexos, não aparentados, residentes na Ilha de Marajó, no estado do Pará, a fim de contribuir para a caracterização genética da mesma.

Amostras biológicas de 251 participantes foram obtidas através da coleta de manchas de sangue periférico em papel Whatman FTA (Sigma-Aldrich Co.), sendo armazenadas à temperatura ambiente até a extração do DNA. Todos os participantes consentiram na utilização de suas amostras biológicas e o uso de seu material genético, de forma não identificada, em projetos de pesquisa, ao assinar um Termo de consentimento livre e esclarecido (Anexo A) no ato da coleta. O emprego destas amostras em estudos de Genética populacional foi aprovado pelo Comitê de Ética da UERJ, parecer CAAE: 0067.0.228.000-09 (Anexo B). As coletas aconteceram nos municípios de Soure e Salvaterra.

Foram coletados dados sobre o local de nascimento dos doadores, assim como de seus pais e avós maternos e paternos. Essas informações nos serviram para atender aos critérios estabelecidos para seleção da amostra nas análises a serem realizadas (Tabela 1), visando um aprofundamento do conhecimento sobre a história da ancestralidade genética da população da Ilha de Marajó, a qual será referida como “população autóctone”. Por essa razão procurou-se buscar uma ancestralidade de no mínimo três gerações, de forma a eliminar a influência de migrações recentes para a Ilha de Marajó nos resultados a serem obtidos para os marcadores de ancestralidade. Por outro lado, todas as amostras coletadas serão utilizadas para a

caracterização do perfil genético da população atual residente na Ilha de Marajó. Esta informação possibilitará a criação de bancos de dados relevantes para fins forenses ou para estudos clínicos.

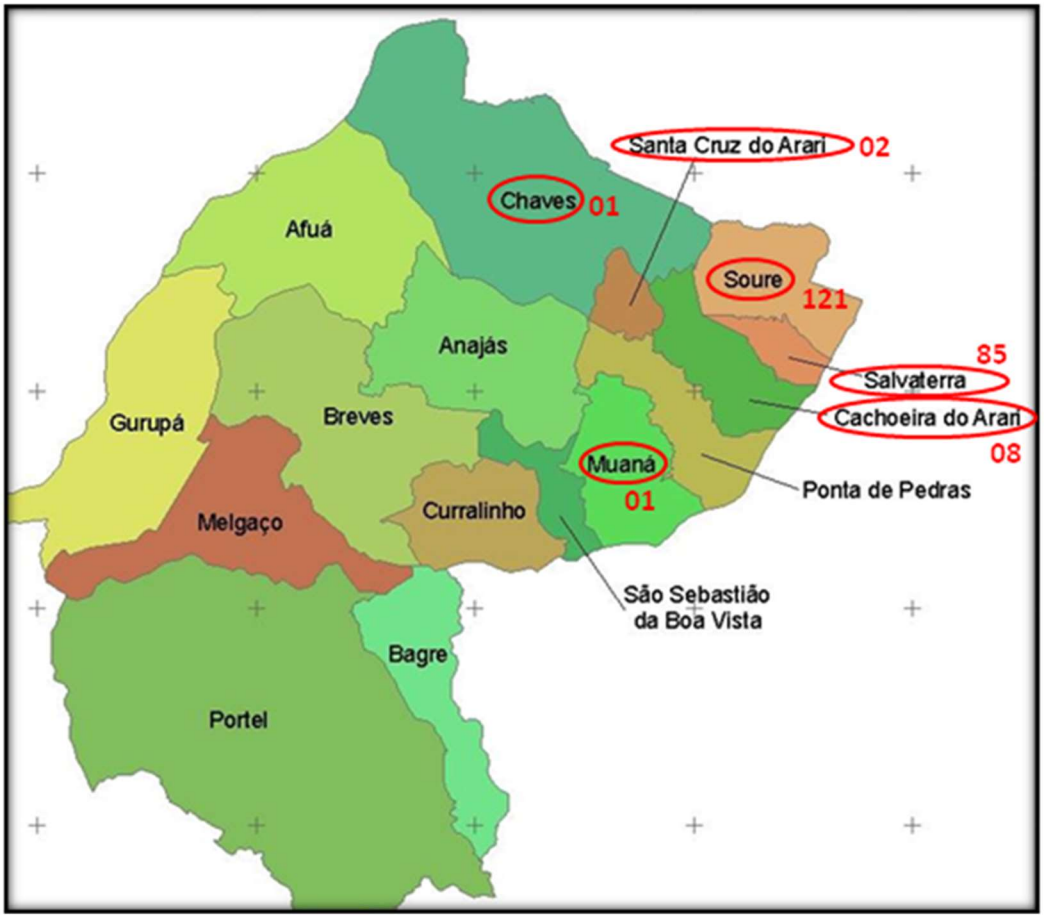
Tabela 1 - Critérios estabelecidos na seleção de indivíduos para o estudo da população autóctone da Ilha de Marajó

Categoria do marcador	Critério	Número de indivíduos selecionados / número total
Cromossomo Y	Indivíduos do sexo masculine com avô paterno nascido na Ilha de Marajó	97/ 251
DNA mitochondrial	Indivíduos de ambos os sexos com avó materna nascido na Ilha de Marajó	241/ 251
AIMs	Indivíduos de ambos os sexos com avós maternos e paternos nascidos na Ilha de Marajó	185/ 251

Legenda: Tipo de marcador a ser estudado, com o critério de seleção de no mínimo 3 gerações e o número de indivíduos que tiveram amostras coletadas e atenderam a esses critérios.

Do total de 251 amostras denominadas de população autóctone, apenas 218 indivíduos nasceram na Ilha de Marajó e 33 nasceram em outros lugares (outros municípios do Estado do Pará - 30, Amapá - 1, Amazonas - 1, Guiana Francesa – 1), apesar de residirem e terem avós naturais do Marajó, atendendo assim aos critérios apresentados na Tabela 1. Os 218 indivíduos nativos da Ilha de Marajó têm como local de nascimento os municípios indicados na Figura 10, todos pertencentes à parte leste da ilha.

Figura 10 - A Ilha de Marajó e seus 16 municípios, com destaque para os municípios de origem de 218 indivíduos da população autóctone



Legenda: Municípios de nascimento de 218 indivíduos autóctones participantes do estudo de ancestralidade da Ilha de Marajó e o número de amostras coletadas de cada município.
Fonte: http://sit.mda.gov.br/download.php?ac=obterDadosBas&m=1501808

2.2 **Extração de DNA genômico**

A extração do DNA das amostras foi realizada no LDD, na UERJ, através de três metodologias distintas: a extração por resina Chelex 100 (BioRad), o método orgânico, com o uso de fenol-clorofórmio/álcool isoamílico; e a extração com o *Kit QIAamp DNA Investigator* (Qiagen).

Embora as eficácias desses métodos de extração possam ser comparáveis, resultando na amplificação do DNA extraído, verificamos que preparações de DNA de amostras extraídas pela resina quelante Chelex apresentaram menor durabilidade, mostrando-se

instáveis em amplificações posteriores. Por essa razão, posteriormente foram realizadas extrações pelo método orgânico, assim como pelo *Kit QIAamp DNA Investigator* (Qiagen).

2.2.1 Extração de DNA - método orgânico - fenol-clorofórmio / álcool isoamílico

Um fragmento de aproximadamente 5 mm^2 de papel Whatman FTA (Sigma-Aldrich Co.), contendo mancha de sangue, foi picotado com o uso de um bisturi e depositado em microtubos de 1,5 mL. Foram adicionados 1 mL de tampão SSC 1x (NaCl 0,15 M e $C_6H_5Na_3O_7.2\ H_2O$ 0,015 M, pH 7) a cada tubo, seguido de agitação em vortex (VX38, Warmnest) e de centrifugação por 5 min a 12000 x g (Centrífuga Centrimicro, modelo 242 – Fanem). O sobrenadante foi descartado com o auxílio de uma pipeta, restando aproximadamente 30 µL no tubo. Foram acrescentados em cada tubo 400 µL de tampão de extração (tris 0,01 M; $EDTANa_2$ 0,01 M; NaCl 0,1 M; SDS 20 %, pH 8), 25 µL de proteinase K 10 mg/mL e 20 µL de DTT 1 M. Em seguida a amostra foi incubada a 37 ºC por 18 a 24 horas em um termobloco. Após a incubação, a extração pelo fenol-clorofórmio e precipitação com etanol seguiram protocolo descrito por Sambrook e colaboradores (SAMBROOK; FRITCSH; MANIATIS, 1989), com adaptações. Foram adicionados 200 µL de fenol/clorofórmio/álcool isoamílico na proporção 25:24:1 (Invitrogen), a preparação foi homogeneizada por agitação em vórtex e centrifugada por 3 min a 12000 x g, havendo a separação da solução em duas fases. A fase aquosa superior foi transferida para um novo tubo de 1,5 mL e foram adicionados 200 µL de clorofórmio/álcool isoamílico (24:1), seguido de homogeneização por agitação em vórtex e centrifugação por 3 min a 12000 x g, separando, novamente, a preparação em duas fases. A fase aquosa (superior) foi transferida para um novo tubo de 1,5 mL e foram adicionados 1 mL de etanol absoluto recém retirado do *freezer* (-20 °C) e 35 µL de acetato de sódio 3 M, seguindo-se agitação no vórtex e incubação a -20 ºC por, pelo menos, 16 horas. Em seguida, a amostra foi centrifugada à temperatura de 4 ºC por 15 min a 10.000 x g (Centrífuga 5418R, Eppendorf). O sobrenadante foi cuidadosamente descartado e foi adicionado 1mL de etanol 70 % , sendo realizada nova centrifugação por 5 min a 12.000 x g. O sobrenadante foi mais uma vez descartado cuidadosamente e as amostras foram deixadas à temperatura ambiente por 18 a 24 horas para evaporação do etanol residual.

Por fim, foram acrescentados 40 µL de TE (tris-HCl 5 mM, EDTA 0,1 mM pH 8), em cada tudo. As amostras foram armazenadas em *freezer* a -20 °C.

2.2.2 Extração de DNA - resina Chelex

Chelex 100 (Bio-Rad) é uma resina que atua como quelante para íons metálicos polivalentes. Durante o processo de extração as células são inicialmente lisadas com água ultrapura. A alcalinidade da solução e a fervura também levam à ruptura das células, permitindo que os grupos quelantes se liguem aos componentes celulares, liberando a molécula do DNA. A extração com Chelex é um bom método, pois é rápido, não requer múltiplas transferências de tubos e não utiliza solventes orgânicos tóxicos como o método com fenol-clorofórmio; porém não consegue remover totalmente inibidores (como a hemoglobina) que podem ser prejudiciais em etapas posteriores (PHILLIPS; MCCALLUM; WELCH, 2012).

A extração de DNA a partir das amostras de sangue colhidas foi realizada de acordo com a técnica desenvolvida por Walsh e colaboradores (WALSH; METZGER; HIGUCHI, 1991). Um fragmento de aproximadamente 5 mm^2 de papel Whatman FTA (Sigma-Aldrich Co.), contendo mancha de sangue, foi picotado com o uso de um bisturi e depositado em microtubos de 1,5 mL. Foi adicionado 1 mL de água ultrapura e as amostras foram deixadas à temperatura ambiente por 30 min. Após este período foram centrifugadas a 12000 x g por 5 min (centrífuga Centrimicro, modelo 242 – Fanem). O sobrenadante foi descartado cuidadosamente com o auxílio de uma pipeta e foram adicionados 200 µL de Chelex 100 a 5 % sob agitação manual, utilizando uma ponteira com a extremidade cortada. Posteriormente, esta preparação foi agitada em vórtex e submetida a duas incubações sucessivas: a primeira a 56 ºC, por 20 min e a segunda a 100 ºC, por 8 min. Após as incubações, as amostras foram novamente centrifugadas a 12000 x g, por 5 min e armazenadas em *freezer* a -20 ºC. As centrifugações foram realizadas na centrífuga Centrimicro, modelo 242 – Fanem.

2.2.3 Extração de DNA - *Kit QIAamp DNA Investigator*

Uma parte das amostras utilizadas teve o DNA extraído com o *kit QIAamp DNA Investigator*, que envolve quatro etapas principais: ruptura das membranas celulares usando uma combinação de atividade enzimática e lise mecânica (aquecimento e agitação); ligação do DNA à membrana de sílica da coluna; lavagem de contaminantes através da membrana usando tampões e a eluição do DNA. Esse método de extração com membrana contendo silica é considerado bastante eficaz, sendo capaz de remover inibidores da amostra enquanto mantém um alto rendimento de DNA de boa qualidade (PHILLIPS; MCCALLUM; WELCH, 2012). Foram seguidas as instruções do protocolo do *kit* para extração do DNA em amostras de sangue em papel FTA e o volume de eluição utilizado foi de 30 µL, em água ultrapura.

2.3 **Quantificação de DNA**

As amostras extraídas com fenol-clorofórmio foram quantificadas utilizando o *kit* comercial Qubit dsDNA HS (Invitrogen), seguindo as instruções do fabricante. A faixa de detecção deste *kit* varia de 0,2-100 ng/µL. Para calibrar o instrumento, dois padrões, *Standard # 1* (0 ng/µL) e *Standard # 2* (10 ng/µL), foram preparados e medidos no instrumento. As amostras a serem quantificadas foram preparadas por combinação de 195 µL de solução reagente e 5 µL de DNA e posteriormente foi feita a conversão do valor de concentração obtido na leitura, levando-se em consideração essa diluição.

2.4 **Análise de marcadores do Cromossomo Y**

A análise molecular de marcadores do cromossomo Y foi realizada em 97 amostras de indivíduos não aparentados pertencentes ao sexo masculino, residentes e que possuem avôs paternos nascidos na Ilha de Marajó.

A determinação dos haplótipos do cromossomo Y foi realizada mediante o emprego dos 25 marcadores STR incluídos no *kit* comercial *Yfiler Plus*, que correspondem a 27 *loci*,

uma vez que os marcadores DYS385 e DYF387S1 correspondem a 2 *loci* cada (*Applied Biosystems*).

A determinação dos haplogrupos presentes nessa amostra foi realizada mediante o emprego de 47 Y-SNPs selecionados de modo a permitirem discriminar os principais haplogrupos que se espera encontrar em populações miscigenadas da América do Sul, como os de origem nativo-americana, europeia e africana. Essa seleção foi feita a partir dos haplótipos Y-STR, através da predição do haplogrupo mais provável, com os *software online Haplogroup Predictor* (http://www.hprg.com/hapest5/, acessado em setembro de 2019) e *NevGen* (http://www.*NevGen*.org, acesso em outubro de 2019). Com base nesse resultado, foram genotipados os Y-SNPs relevantes para a sua confirmação, por reações do tipo SBE (do inglês, *Single Base Extension*) ou extensão de base única, técnica também conhecida por minisequenciamento.

2.4.1 Genotipagem dos *loci* STR do cromossomo Y utilizando o *kit Yfiler Plus*

Para a análise de *loci* Y-STR, o *kit* comercial *Yfiler Plus* (*Applied Biosystems*) foi utilizado, seguindo a metodologia indicada pelo fabricante. Este *kit* consiste em um sistema multiplex de 6 fluorórofos incorporados nos *primer*s por meio de ligantes não-nucleotídicos, que permitem uma amplificação e separação eficientes de 27 *loci* Y-STR em uma única reação de PCR. São utilizados cinco fluoróforos nas amostras (6-FAM, VIC, NED, TAZ e SID), que emitem fluorescências nas cores azul, verde, amarelo, vermelho e roxo, respectivamente, permitindo a análise de diferentes marcadores através da distinção de alelos de *loci* diferentes com o mesmo tamanho. O sexto fluoróforo, que consiste no LIZ, de cor laranja, é utilizado no padrão interno de tamanho, o *GeneScan 600 LIZ* v2.0 (*Thermo Fisher Scientific*). Na Tabela 2 estão indicados os 27 *loci* STR, os fluoróforos utilizados para cada marcador e os alelos incluídos no padrão alélico que compõe o *kit.*

Na reação foram utilizados aproximadamente 1-2 ηg de DNA extraído pelo método orgânico ou 0,5 μL (DNA extraído por Chelex ou *Kit QIAamp DNA Investigator*), 2,0 μL de tampão *Yfiler Plus Master Mix*, 1,0 μL de *Yfiler Plus Primer Set* e água ultrapura autoclavada para completar o volume final de 5,0 μL de reação. Juntamente com as alíquotas de DNA dos indivíduos das amostras populacionais estudadas, também foi amplificado o DNA controle 007 como controle positivo das reações. Como controle negativo foi preparada uma reação

sem o acréscimo de DNA, sendo o volume do mesmo substituído por água ultrapura autoclavada.

O preparo da PCR foi realizado em uma cabine esterilizada por luz ultravioleta de comprimento de onda de 254 nm e as reações foram realizadas em termociclador automático modelo *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*), segundo condições termocíclicas indicadas pelo fabricante do *kit.*

O volume de 1 µL dos produtos da amplificação foi aplicado em placa a ser colocada no analisador genético ABI 3500 (*Applied Biosystems*), juntamente com uma mistura contendo 8,8 µL de Formamida Hi-Di (*Applied Biosystems*) e 0,2 µL de padrão interno de tamanho *GeneScan 600 LIZ dye Size Standard v2.0* (*Applied Biosystems*). Em um dos poços, a essa mistura foi acrescentado 1 µL de *Yfiler Plus Allelic Ladder*. Em seguida, a placa foi levada ao sequenciador para a realização da eletroforese em capilares de 50 cm de comprimento preenchidos com o polímero POP-7 (*Applied Biosystems*).

Por fim, os resultados obtidos são representados graficamente por um eletroferograma no qual se observam picos resultantes do processo de emissão, detecção e tradução do sinal luminoso emitido pelos fluoróforos, cuja intensidade é medida em RFUs, através do programa *GeneMapper v. 4.1* (*Thermo Fisher Scientific*), que também faz a nomeação dos alelos ao comparar os alelos detectados com os alelos do padrão alélico *Yfiler Plus Allelic Ladder.*

2.4.2 Predição de haplogrupos - *software Haplogroup Predictor* e *NevGen*

A fim de evitar genotipagens desnecessárias, a escolha de marcadores Y-SNP utilizados na genotipagem das amostras estudadas foi direcionada por meio dos *software online Haplogroup Predictor* e *NevGen*, que indicam os prováveis haplogrupos de uma determinada amostra com base na análise de seus haplótipos Y-STR.

Tabela 2 - 27 *loci* do *kit Yfiler Plus*

Locus	Alelos incluídos no padrão alélico do *kit Yfiler Plus*	Fluoróforos	DNA controle 007
DYS576	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	6-FAM	19
DYS389I	9, 10, 11, 12, 13, 14, 15, 16, 17		13
DYS635	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30		24
DYS389II	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35		29
DYS627	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27		21
DYS460	7, 8, 9, 10, 11, 12, 13, 14	VIC	11
DYS458	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24		17
DYS19	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19		15
YGATAH4	8, 9, 10, 11, 12, 13, 14, 15		13
DYS448	14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24		19
DYS391	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16		11
DYS456	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	NED	15
DYS390	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29		24
DYS438	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16		12
DYS392	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20		13
DYS518	32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49		37
DYS570	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	TAZ	17
DYS437	10, 11, 12, 13, 14, 15, 16, 17, 18		15
DYS385a/b	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28		11, 14
DYS449	22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40		30
DYS393	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	SID	13
DYS439	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17		12
DYS481	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32		22
DYF387S1 a/b	30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44		35, 37
DYS533	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17		13

Legenda: *Loci Yfiler Plus*, seus respectivos fluoróforos e alelos observados no padrão alélico e no DNA controle 007

Sendo assim, por exemplo, ao colocar no programa o haplótipo de uma determinada amostra e este ter indicado com uma probabilidade elevada (considerada maior que 90 % no presente estudo) que a amostra pode pertencer ao haplogrupo R, a genotipagem desta amostra em particular foi realizada pelo multiplex com *primers* específicos para Y-SNPs

determinantes do haplogrupo R (Multiplex R, segundo a Figura 11). Por outro lado, se os programas não puderam determinar com elevada probabilidade o haplogrupo de uma amostra ou se discordarem entre si; para esta amostra em questão foi feita a genotipagem preliminar com multiplexes que são compostos por *primer*s específicos para Y-SNPs de ramos mais basais da árvore filogenética (Multiplex 1, segundo a Figura 11), a fim de direcionar uma segunda genotipagem com os Y-SNPs determinantes de determinados haplogrupos.

2.4.3 Genotipagem dos marcadores Y-SNP - *kit SNaPshot Multiplex*

Tomando como base os resultados obtidos na predição dos haplogrupos e considerando o conhecimento histórico sobre a origem da população brasileira, assim como os dados obtidos através de estudos genéticos que corroboram para esses relatos (JANNUZZI *et al*., 2020; OLIVEIRA *et al*., 2014; PENA; SANTOS; TARAZONA-SANTOS, 2020; RESQUE *et al*., 2016), foi possível direcionar os marcadores Y-SNP mais interessantes para serem utilizados neste estudo. Foram selecionados 47 marcadores Y-SNP para a análise da ancestralidade das linhagens masculinas do Marajó. Tais marcadores foram agrupados de maneira hierárquica em 5 multiplexes, que caracterizam os principais haplogrupos europeus, africanos e ameríndios relacionados às linhagens masculinas brasileiras (Figura 11). Sendo assim, a genotipagem de marcadores Y-SNP neste estudo foi realizada por PCR multiplex seguida pelo sequenciamento de base única, também conhecido como SBE.

As sequências gênicas que contêm os Y-SNPs de interesse foram amplificadas pela técnica de PCR através da realização dos 5 multiplexes citados anteriormente (Figura 11). Foram utilizados aproximadamente 1-2 ηg de DNA extraído pelo método orgânico ou 0,5-2 µL (DNA extraído por Chelex ou *Kit QIAamp DNA Investigator*), além de 2,5 µL do tampão *QIAGEN Multiplex PCR Master Mix* 2x, proveniente do *QIAGEN Multiplex PCR kit*, 0,5 µL de mix dos *primers* na concentração de 2 µM e água ultrapura autoclavada para completar o volume final de 5,0 µL de reação. Para cada uma das preparações foram feitos um controle negativo (reação sem o acréscimo de DNA) e um controle positivo utilizando DNA de um indivíduo previamente genotipado cujo haplogrupo possuía o estado ancestral (sem mutações) para os Y-SNPs analisados.

Para a amplificação do DNA, o preparo da PCR foi realizado em uma cabine esterilizada por luz ultravioleta de comprimento de onda de 254 nm e as reações foram

realizadas em termociclador automático modelo *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*). As condições termocíclicas foram descritas por Gomes e colaboradores (2010) e são apresentadas na Tabela 3.

Tabela 3 - Condições termocíclicas para a amplificação dos multiplexes para o estudo dos Y-SNPs

Etapa	Ciclos	Temperatura (º C)	Tempo
Desnaturação inicial	1	95	15 min
Desnaturação	35	94	30 s
Anelamento		60	1 min 30 s
Extensão		72	1 min
Extensão final	1	72	10 min

Legenda: Etapas utilizadas na reação de amplificação, com número de ciclos, suas respectivas temperaturas e durações.

Os conjuntos de iniciadores utilizados para cada SNP, assim como os tamanhos de fragmentos obtidos em pares de base e a reação de PCR multiplex a qual pertencem estão descritos na Tabela 4.

Para a avaliação inicial dos produtos de amplificação, foi realizada eletroforese em gel de poliacrilamida (29:1 acrilamida:bis-acrilamida), a uma concentração de 10 %. As amostras foram preparadas adicionando-se 2 µL do produto de amplificação a 1 µL de *Safer Dye* (Kasvi) e posteriormente aplicadas no gel e em uma cuba vertical, a 5,5 V/cm, por 1 h 30 min, em tampão de corrida TBE 1x (tris 89 mM, ácido bórico 89 mM, EDTA 2 mM, pH 8). Também foi aplicado no gel o marcador de tamanho molecular *GeneRuler 50 bp DNA Ladder* (*Thermo Fisher Scientific*), composto de uma mistura de fragmentos de DNA com comprimentos entre 50 a 1000 pb. Após a eletroforese, o gel foi visualizado em um transiluminador de luz azul (modelo K33-333, Kasvi), sendo possível a identificação dos fragmentos de interesse por comparação com o marcador de peso molecular.

A etapa seguinte foi a realização da reação multiplex SBE – de extensão de base única, utilizando-se o *kit SNaPshot Multiplex* (*Applied Biosystems*). Porém, antes da reação SBE, os produtos amplificados foram submetidos à purificação com enzima USB ExoSAP-IT (*Thermo Fisher Scientific*), com o objetivo de remover oligonucleotídeos e dNTPs não incorporados durante a reação de amplificação (1 µL de produto de PCR e 0,5 µL de enzima USB ExoSAP-IT). A purificação foi realizada em termociclador automático modelo *Veriti 96-Well*

Thermal Cycler (*Applied Biosystems*), a 37º C por 30 min e, em seguida, a 80 ºC por 15 min, para inativação da enzima.

Para o preparo das reações SBE foram adicionados diretamente aos produtos de purificação 1 µL de *SNaPshot Multiplex Ready Reaction Mix* (*Thermo Fisher Scientific*) e 2,5 µL de mix de *primer*s SBE (Tabela 5). O marcador 12f2, do Multiplex GIJ, não foi incluído nessa tabela porque seu polimorfismo é detectado pelo tamanho do fragmento gerado na PCR, uma vez que se trata de uma inserção do tipo Alu. As reações SBE foram então submetidas às condições termocíclicas de 25 ciclos a 96 ºC por 10 s, 50 ºC por 5 s e 60 ºC por 30 s em termociclador *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*), segundo as instruções do fabricante. Após a reação de SBE, esta foi purificada adicionando-se 1 µL de enzima *SAP(Thermo Fisher Scientific*), a fim de remover ddNTPs não incorporados. A seguir, essa preparação foi incubada a 37 ºC por 1 h e a 85 ºC por 15 min. O volume de 1 µL dos produtos da purificação foram aplicados em placa a ser colocada no analisador genético ABI 3500 (*Applied Biosystems*) para a eletroforese em capilar de 50 cm preenchido com o polímero POP-7 (*Applied Biosystems*), juntamente com uma mistura contendo 8,8 µL de Formamida Hi-Di (*Applied Biosystems*) e 0,2 µL de padrão interno de tamanho *GeneScan 120 LIZ dye Size Standard v2.0* (*Applied Biosystems*).

Os resultados obtidos do minisequenciamento foram representados graficamente por um eletroferograma no qual se observam picos resultantes do processo de emissão, detecção e tradução do sinal luminoso emitido pelos fluoróforos, cuja intensidade é medida em RFUs, através do programa *GeneMapper v. 4.1* (*Thermo Fisher Scientific*). Uma vez que a cada nucleotídeo que possa ser incorporado na reação SBE está associado um fluorófo específico, informado pelo fabricante do *kit SNaPshot Multiplex* (*Applied Biosystems*), é possível detectar, pela região de emissão de energia do fluoróforo, representada por cor específica informada pelo fabricante do *kit*, qual é a variante alélica para cada SNP selecionado.

Tabela 4 - Conjuntos de *primers* das PCRs multiplexes para o estudo dos Y-SNPs (continua)

Y-SNP	Amplicon (pb)	*Primer* direto (5'→3')	*Primer* reverso (5'→3')	Referência	PCR Multiplex
M9	340	GCAGCATATAAAACTTTCAGG	AAAACCTAACTTTGCTCAAGC	BRION *et al*., 2005	MULTIPLEX 1 EUROPEU (BRION *et al.*, 2005)
M173	172	GCACAGTACTCACTTTAGGTTTGC	GCAGTTTTCCCAGATCCTGA	BRION *et al*., 2005	
SRY1532	167	TCCTTAGCAACCATTAATCTGG	AAATAGCAAAAACTGACACAAGGC	BRION *et al*., 2005	
M213	145	GGCCATATAAAAACGCAGCA	**Primer* SBE reverso	BRION *et al*., 2005	
P25	121	GGACCATCACCTGGGTAAAGT	AGTGCTTGTCCAAGGCAGTA	BRION *et al*., 2005	
Tat	112	GACTCTGAGTGTAGACTTGTGA	GAAGGTGCCGTAAAAGTGTGAA	BRION *et al*., 2005	
M22	106	GCTGATAGTCCTGGTTTCCCTA	TGAGCATGCCTACAGCAGAC	BRION *et al*., 2005	
M70	81	TCATAGCCCACTATACTTTGGAC	CTGAGGGCTGGACTATAGGG	BRION *et al*., 2005	
92R7	55	TGCATGAACACAAAAGACGTA	GCATTGTTAAATATGACCAGC	BRION *et al*., 2005	
M62	309	ACTAAAACACCATTAGAAACAAAGG	CTGAGCAACATAGTGACCCC	BRION *et al*., 2005	MULTIPLEX GIJ (SIMÃO *et al.*, 2021)
M267	256	CGTTGTCCCTGTGTTTCCAT	CTGTTGCCCAGGCTAGTGTC	NOGUEIRO *et al*., 2010	
M172	187	TCCTCATTCACCTGCCTCTC	TCCATGTTGGTTTGGAACAG	BRION *et al*., 2005	
P58	180	ACAGGAGGCCATAATGCAAC	GAGCCTCACACCTTCCTCTG	SIMÃO *et al*., 2021	
M170	158	TGCAGCTCTTATTAAGTTATGTTTTCA	CCAATTACTTTCAACATTTAAGACC	BRION *et al*., 2005	
M201	144	TCAAATTGTGACACTGCAATAGTT	CATCCAACACTAAGTACCTATTACGAA	BRION *et al*., 2005	
12f2	90	CACTGACTGATCAAAATGCTTACAGAT	GGATCCCTTCCTTACACCTTATACA	BRION *et al*., 2005	
P2	180	GCTCCAGCCATCTTTTCCTTA	CTTCTCTCATGAGGGTTTTGGA	GOMES *et al*., 2010	MULTIPLEX E (GOMES *et al*., 2010)
M293	230	AAAGAGATTGATCGGTGCATA	GCTGGCTAATACTTCCACAGAG	GOMES *et al*., 2010	
M154	130	TACTCACACAAACCAAGAAGAAACA	AACCATTGTGTTACATGGCCTA	GOMES *et al*., 2010	
M81	203	GCACTATCATACTCAGCTACACATCTC	AACCATTGTGTTACATGGCCTA	GOMES *et al*., 2010	
M85	283	TGGCATCCAATACTAGCTGATAAAC	AATGCTCACGCTTGTGTTCT	GOMES *et al*., 2010	
M78	235	GGATGGCTGTATGGGTTTCT	ATAGTGTTCCTTCACCTTTCCTT	GOMES *et al*., 2010; BRION *et al*., 2005	
M35	198	GCATGGTCCCTTTCTATGGAT	GAGAATGAATAGGCATGGGTTC	BRION *et al*., 2005	
M96	88	GTGATGTGTAACTTGGAAAACAGG	GGACCATATATTTTGCCATAGGTT	BRION *et al*., 2005	
V6	102	GATGGCACAGTGTTCGACAG	CTTCTCTCCAAATGCCTGCT	GOMES *et al*., 2010	
M2	162	AAGTCCAGACCCAGGAAGGT	ACAGCTCCCCCTTTATCCTC	GOMES *et al*., 2010	
M123	213	TGCTCTCAGGGGAAAATCTG	AGCAAAGTTGAGGTTGCACA	GOMES *et al*., 2010	
M191	122	AAAAATGGAGTGTTTATCAGAGCTT	CCCAGACACACCAAAATATCTC	GOMES *et al*., 2010	
M33	190	CACAACTTCATTGGCTACGG	GTTGAAGCCCCCAAGAGAGAC	GOMES *et al*., 2010	

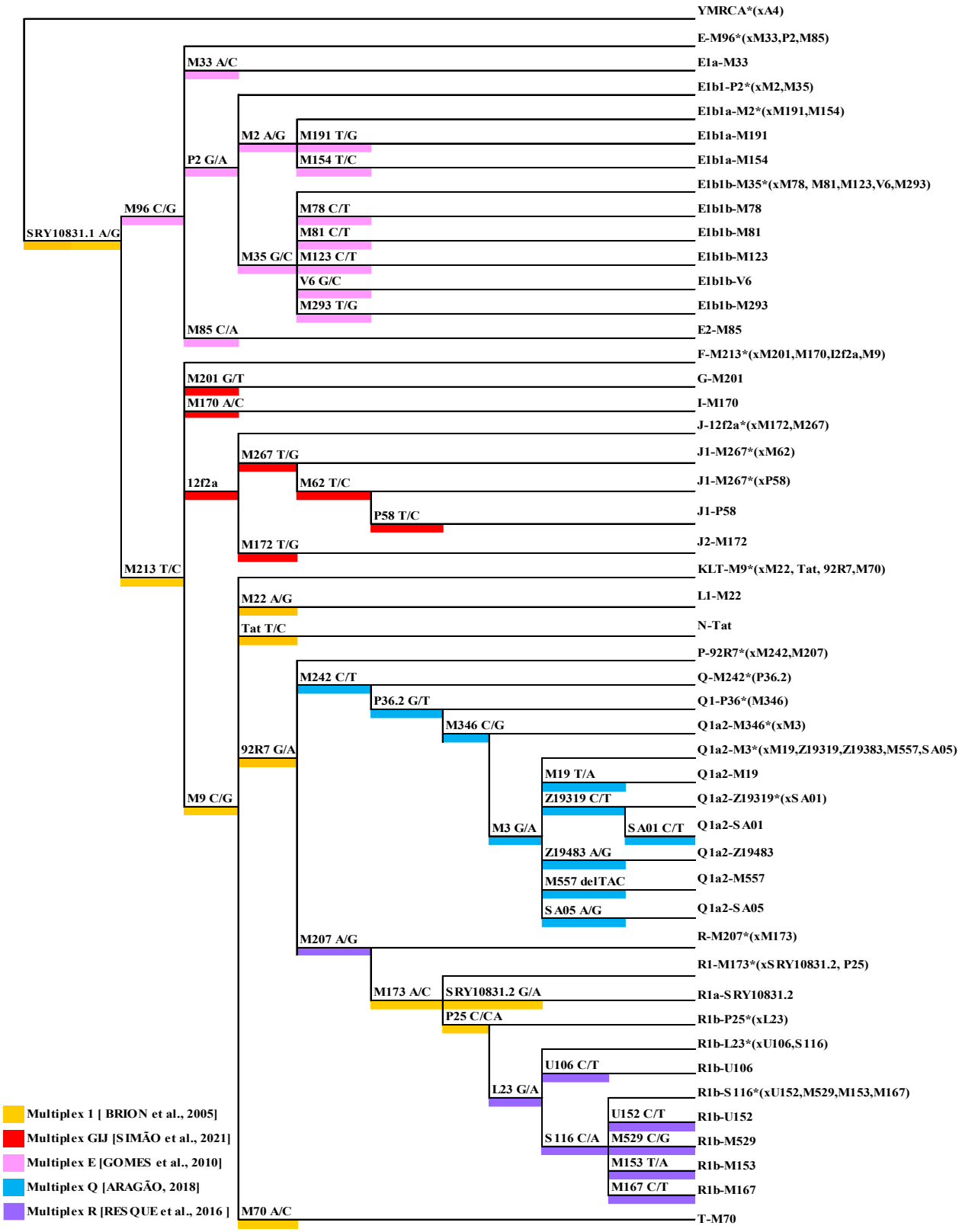
Tabela 4 - Conjuntos de *primers* das PCRs multiplexes para o estudo dos Y-SNPs (conclusão).

Y-SNP	Amplicon(pb)	*Primer* direto (5'→3')	*Primer* reverso (5'→3')	Referência	PCR Multiplex
SA01 M557	370	AAGATCCCACCACTGCACTC	CTCTGGCCCCTAACAAACCT	ARAGÃO, 2018	MULTIPLEX Q (ARAGÃO, 2018)
M3	304	CATTAAAGCCGGTCACAGGT	CTGCCAGGGCTTTCAAATAG	NOGUERA *et al*., 2014	
P36.2	299	GAGGAGGGGGAGAGAGAAAA	TTCAAACAGCCCACCAGATA	NOGUERA *et al*., 2014	
M19	277	TCACCAGAGTTTCAAATAG	ACAGACACAAAGGGCCAACT	NOGUERA *et al*., 2014	
M346	247	GGCCTGAAAATGTGGAAAGA	AGCCTGAAAATGTGGAAAGA	NOGUERA *et al*., 2014	
SA05	236	GAACCAAAGCACAGCACTCA	ATGCTCATGGCCTACACCTC	ARAGÃO, 2018	
Z19483	211	CCATGTAGGAGGAGGCAAAA	CATCACAAAAGCCAAAAGCA	ARAGÃO, 2018	
Z19319	163	TTTGCTGAAGTTGCCTGTCA	AGTTCCAGTCAGGGCAATCA	ARAGÃO, 2018	
M242	155	TTGTGCAAAAAGGTGACCAA	TTTCGCTTTAAGGGCTTTCA	NOGUERA *et al*., 2014	
M153	76	TTCTCAGACACCAATGGTCCTA	TCTGACTTGGAAAGGGGAAA	RESQUE *et al*., 2016	MULTIPLEX R (RESQUE *et al*., 2016)
M167	131	GAGGCTGGGCCAAGTTAAG	CTTCCTCGGAACCACTACCA	RESQUE *et al*., 2016	
U152	163	GAAACATTCCACGCTTGAGG	AGCCTCTTTTTGGCTTCCAT	RESQUE *et al*., 2016	
U106	192	TCCTGAATAGCAAATCCCAAAG	AATGGCAGAGGTAGGAGGAAAT	RESQUE *et al*., 2016	
M529	228	GCCCCCAAAACAACAGAATA	GGAAGCATTCAGAAGCAGGT	RESQUE *et al*., 2016	
L23	229	ACACAGTGAAACCCCGTCTC	AAGATTGTGGGGACAAAGGA	RESQUE *et al*., 2016	
S116	241	TCAGTCAGGGCAAATCTGAA	GGTGGAGTTGGGGCTAAAGT	RESQUE *et al*., 2016	
M207	322	CGTTACAACTATGGGGCAAA	TCCTCTCTGAAATGCCGAAT	RESQUE *et al*., 2016	

Legenda: Nesta tabela, estão listados os marcadores Y-SNP analisados neste estudo, o tamanho dos produtos de amplificação gerados em pares de bases (pb), as sequências dos *primer*s direto e reverso no sentido 5'→3', os artigos de referência para cada conjunto de *primer* e as reações de PCR multiplex às quais os marcadores foram agrupados e amplificados.

Nota: Para o marcador M213, utilizou-se o *primer* SBE reverso na reação de PCR (*cacgtcgtgaaagtctgacaaTCAGAACTTAAAACATCTCGTTAC), pois o *primer* reverso da PCR confeccionado tinha um problema na sequência.

Figura 11 - Marcadores SNP do cromossomo Y usados neste trabalho, em destaque na árvore filogenética do cromossomo Y



Legenda: Conjunto de marcadores Y-SNP utilizados nas amostras deste estudo. As cores indicam cada multiplex e os respectivos marcadores amplificados e ao fim de cada ramo da árvore se encontra o haplogrupo determinado por esses marcadores.

Fonte: A autora.

Tabela 5 - Conjuntos de *primers* das reações SBE multiplexes para o estudo dos Y-SNPs (continua)

Y-SNP	Produto de SBE (pb)	Sequência (5'→3')	Referência	Multiplex SBE
M9 SBE	48	GTGCCACGTCGTGAAAGTCTGACAAGAAACGGCCTAAGATGGTTGAAT	BRION *et al*., 2005	MULTIPLEX 1 EUROPEU (BRION *et al.*, 2005)
M173 SBE	34	GTCTGACAACTTACAATTCAAGGGCATTTAGAAC	BRION *et al*., 2005	
SRY1532	30	TCTGACAATTGTATCTGACTTTTTCACACAGT	BRION *et al*., 2005	
M213 SBE	45	CACGTCGTGAAAGTCTGACAATCAGAACTTAAAACATCTCGTTAC	BRION *et al*., 2005	
P25 SBE	26	CAATCTGCCTGAAACCTGCCTG	BRION *et al*., 2005	
Tat SBE	42	TCGTGAAAGTCTGACAACTCTGAAATATTAAATTAAAACAAC	BRION *et al*., 2005	
M22 SBE	18	GCCATTCCTGGTGGCTCT	BRION *et al*., 2005	
M70 SBE	34	TCTGACAATAGGGATTCTGTTGTGGTAGTCTTAG	BRION *et al*., 2005	
92R7 SBE	28	AAGCATGAACACAAAAGACGTAGAAG	BRION *et al*., 2005	
M62 SBE	27	CTGACAACAATGTTTGTTGGCCATGGA	BRION *et al*., 2005	MULTIPLEX GIJ (SIMÃO *et al.*, 2021)
M267 SBE	37	TGAAAGTCTGACAACTTCCACACAAAATACTGAAMGT	NOGUEIRO *et al*., 2010	
M172 SBE	18	AAACCCATTTTGATGCTT	BRION *et al*., 2005	
P58 SBE	40	CACGTCGTGAAAGTCTGACAATGACATTTGTGTGCTTTGC	SIMÃO *et al*., 2021	
M170 SBE	22	ACACAACCCACACTGAAAAAAA	BRION *et al*., 2005	
M201 SBE	34	CCCCCCCCCCCCCCCGATCTAATAATCCAGTATCAACTGAGG	BRION *et al*., 2005	
P2 SBE	16	GCCCCTAGGAGGAGAA	GOMES *et al*., 2010	MULTIPLEX E (GOMES *et al*., 2010)
M293 SBE	21	AAAGAGATTGATCGGTGCATA	GOMES *et al*., 2010	
M154 SBE	26	AAACATGGCCTATAATATTCAGTACA	GOMES *et al*., 2010	
M81 SBE	27	CCCCCCTAAATTTTGTCCTTTTTTGAA	BRION *et al*., 2005	
M85 SBE	30	CTTGTGTTCTATTAAGTGTAGTTTTGTTAG	GOMES *et al*., 2010	
M78 SBE	34	CCCCCCCCCCACACTTAACAAAGATACTTCTTTC	BRION *et al*., 2005	
M35 SBE	36	CCCCCCCCCCCCCCCCCCCCCAGTCTCTGCCTGTGTC	BRION *et al*., 2005	
M96 SBE	40	CCCCCCCCCCCGTAACTTGGAAAACAGGTCTCTCATAATA	BRION *et al*., 2005	
V6 SBE	42	GCCACGTCGTGAAAGTCTGACAATGCTGTGATTCCTGATGTG	GOMES *et al*., 2010	
M2 SBE	45	GTGCCACGTCGTGAAAGTCTGACAATTTATCCTCCACAGATCTCA	GOMES *et al*., 2010	
M123 SBE	48	TAGGTGCCACGTCGTGAAAGTCTGACAATTCTAGGTATTCAGGCGATG	GOMES *et al*., 2010	
M191 SBE	51	GGTGCCACGTCGTGAAAGTCTGACAACATTTTTTTCTTTACAACTTGACTA	GOMES *et al*., 2010	
M33 SBE	54	GTGCCACGTCGTGAAAGTCTGACAACAGTTACAAAAGTATAATATGTCTGAGAT	GOMES *et al*., 2010	

Tabela 5 - Conjuntos de *primers* das reações SBE multiplexes para o estudo dos Y-SNPs (conclusão)

Y-SNP	Produto de SBE (pb)	*Sequência* (5'→3')	Referência	Multiplex SBE
SA01 SBE	38	gtcgtgaaagtcctgacaaTTTGTCAGTGTAGAGTGG	ARAGÃO, 2018	MULTIPLEX Q (ARAGÃO, 2018)
M557 SBE	45	tgccacgtcgtgaaagtctgacaaGAACAGGGTTGCAAACGGTA	ARAGÃO, 2018	
M3 SBE	17	TCACCTCTGGGACTGA	NOGUERA *et al*., 2014	
P36.2 SBE	35	gtcgtgaaagtctgacaaCATCTATCTATCCATTATTCTCTCT	NOGUERA *et al*., 2014	
M19 SBE	28	tgacaaGTAGAGACATCTGAAACCCAC	NOGUERA *et al*., 2014	
M346 SBE	28	ctgacaaCAGCCAAGAGGACAGTAAGA	NOGUERA *et al*., 2014	
SA05 SBE	48	aggtgccacgtcgtgaaagtctgacaaATGTTTCTAGGGTGAGCCTGT	ARAGÃO, 2018	
Z19483 SBE	42	acgtcgtgaaagtctgacaaATAAGCTGTCTGGCTATTTCA	ARAGÃO, 2018	
Z19319 SBE	40	tcgtgaaagtctgacaaCCATCATCTCAACCTAAAATCC	ARAGÃO, 2018	
M242 SBE	23	aaAAAAAGGTGACCAAGGTGCT	NOGUERA *et al*.,	
M153 SBE	23	AAAGCTCAAAGGGTATGTGAACA	RESQUE *et al.*, 2016	MULTIPLEX R (RESQUE *et al*., 2016)
M167 SBE	16	AAGCCCCACAGGGTGC	RESQUE *et al.*, 2016	
U152 SBE	32	CAAGGATAAGAAAAATGAGTATTGTGAAAATA	RESQUE *et al.*, 2016	
U106 SBE	28	TCTGACAATAGCAAATCCCAAAGCTCCA	RESQUE *et al.*, 2016	
M529 SBE	23	AATAACAACCGCTCTCTCAGACA	RESQUE *et al.*, 2016	
L23 SBE	18	GCGACAGAGCGAGACTCT	RESQUE *et al.*, 2016	
S116 SBE	35	GAAAGTCTGACAAGAGTTGGGGCTAAAGTGAAAG	RESQUE *et al.*, 2016	
M207 SBE	28	AACAAATGTAAGTCAAGCAAGAAATTTA	RESQUE *et al.*, 2016	

Legenda: Nesta tabela, estão listados os *primer*s SBE utilizados neste estudo, suas sequências no sentido 5'→3', os artigos de referência para cada conjunto de *primer*, o tamanho dos produtos de minissequenciamento gerados e as reações de SBE multiplex às quais os marcadores foram agrupados e amplificados.

Nota: Para o marcador M22, o *primer* SBE utilizado foi modificado em duas bases ([CCdel] GCCATTCCTGGTGGCTCT) com relação ao *primer* referência, descrito por Brion *et al*., 2005.

2.4.4 Análise estatística dos resultados do Cromossomo Y

Com base nos dados obtidos pela genotipagem dos 27 *loci* Y-STRs com *Yfiler Plus*, foram realizadas estimativas de parâmetros de diversidade genética intrapopulacional, como as frequências haplotípicas, determinação do número de haplótipos compartilhados e cálculo da diversidade haplotípica (HD), utilizando o *software* Arlequin versão 3.5 (EXCOFFIER; LISCHER, 2010).

O mesmo *software* foi utilizado nos cálculos de distâncias genéticas (F_{ST}) e probabilidades de não-diferenciação entre a amostra estudada e outras 12 populações miscigenadas com dados previamente publicados para *Yfiler Plus*, representativas das regiões geopolíticas brasileiras (Tabela 6). Além dessas, foram comparadas populações de referência de origens africana [Kenya (n=62) - Bantu_Luhya, outro - YA004206] (IACOVACCI *et al*., 2017), europeia [Madri, Espanha (n=126) - YA003147] (MARTÍN *et al*., 2004) e nativa Americana [Oxapampa, Peru (n=58) - Ashaninka - YA004112] (TINEO *et al*., 2015) com dados disponíveis no banco de dados YHRD.

Como não existem dados de *Yfiler Plus* disponíveis para outras populações da região Norte do Brasil, nova análise foi realizada, calculando os valores de diversidade para um conjunto parcial de 18 marcadores para os quais existem dados publicados por Palha e colaboradores (2012) e Purps e colaboradores (2014). Este conjunto incluiu os seguintes Y-STR: DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS570, DYS576, DYS635 e YGATAH4. As distâncias genéticas (F_{ST}) e probabilidades de não-diferenciação entre a amostra estudada e essas populações (Tabela 7) para esse conjunto de marcadores também foram calculadas com o *software* Arlequin v. 3.5.

Nos cálculos das distâncias genéticas foram excluídos os marcadores DYS385 e DYF387S1, e o número de repetições do marcador DYS398I foi subtraído de DYS398II, visto que esses marcadores apresentam mais de um alelo em um haplótipo. Além disso, alelos nulos, microvariantes, deleções, duplicações e triplicações existentes foram codificados como dados em falta e simbolizados com "?".

Para visualização das distâncias genéticas pareadas obtidas, gráficos de escalonamento multidimensional (MDS) foram construídos através do programa *Statistica Software* v. 14.0.0.15 (TIBCO Software Inc.).

As proporções de haplogrupos europeus, africanos e Ameríndios na amostra analisada foram calculadas por contagem direta no Excel (*Microsoft Corporation*).

Tabela 6 - Populações brasileiras utilizadas para comparação com a população da Ilha de Marajó, com base nos dados obtidos pela genotipagem dos 27 Y-STRs - *Yfiler Plus*

População/ Região	Número de indivíduos	Etnia	Referência
Marajó (N)	97	Miscigenada	Este trabalho
Maranhão (NE)	296	Miscigenada	JANNUZZI *et al*., 2020
Piauí (NE)	42	Miscigenada	JANNUZZI *et al*., 2020
Ceará (NE)	38	Miscigenada	JANNUZZI *et al*., 2020
Rio Grande do Norte (NE)	19	Miscigenada	JANNUZZI *et al*., 2020
Pernambuco (NE)	141	Miscigenada	JANNUZZI *et al*., 2020
Mato Grosso (CO)	100	Miscigenada	JANNUZZI *et al*., 2020
Goiás (CO)	204	Miscigenada	JANNUZZI *et al*., 2020
Rio de Janeiro (SE)	258	Miscigenada	JANNUZZI *et al*., 2018
São Paulo (SE)	169	Miscigenada	JANNUZZI *et al*., 2020
Espírito Santo (SE)	409	Miscigenada	JANNUZZI *et al*., 2020; STANGE *et al*., 2019
Espírito Santo-pomeranos (SE)	79	Miscigenada	STANGE *et al*., 2019
Rio Grande do Sul (S)	211	Miscigenada	JANNUZZI *et al*., 2020

Legenda: N - Região Norte, NE - Região Nordeste, CO - Região Centro-Oeste, SE - Região Sudeste, S - Região Sul.

Tabela 7 - Populações da Região Norte do Brasil utilizadas para comparação com a população da Ilha de Marajó, com base nos dados obtidos pela genotipagem de 18 Y-STRs

População	Número de indivíduos	Etnia	Referência
Marajó	97	Miscigenada	Este trabalho
Belém	400	Miscigenada	PALHA *et al*., 2012
Santarém	72	Miscigenada	PALHA *et al*., 2012
Macapá	65	Miscigenada	PALHA *et al*., 2012
Rio Branco	31	Miscigenada	PALHA *et al*., 2012
Porto Velho	135	Miscigenada	PALHA *et al*., 2012
Boa Vista	67	Miscigenada	PALHA *et al*., 2012
Palmas	30	Miscigenada	PALHA *et al*., 2012
Manaus	74	Miscigenada	PALHA *et al*., 2012
São Gabriel da Cachoeira	61	Nativa Americana	PURPS *et al*., 2014

2.5 **Análise de marcadores do DNA mitocondrial**

A determinação dos haplótipos de mtDNA foi realizada através de sequenciamento da Região Controle (RC - entre as posições nucleotídicas 16024 e 576) para 101 indivíduos com avós maternas nascidas na Ilha de Marajó. A determinação dos haplogrupos aos quais pertencem as linhagens maternas dessa população foi feita com base nos haplótipos obtidos por comparação das sequências com a rCRS, através do *software SeqScape v2.7*. Na classificação dos haplótipos seguiram-se as diretrizes propostas pela ISFG (PARSON *et al*., 2014).

2.5.1 Genotipagem da Região Controle do mtDNA

Para a análise da RC completa do mtDNA foi realizado o sequenciamento pelo método de Sanger (SANGER; NICKLEN; COULSON, 1977), com o *kit BigDye Terminator V3.1 Cycle Sequencing* (*Applied Biosystems*).

Foram utilizados os *primers* L15997 e H639 a uma concentração de 0,2 μM, com aproximadamente 1-5 ηg de DNA extraído pelo método orgânico e 2,5 μL de *QIAGEN Multiplex* PCR *Master Mix (QIAGEN), num* volume final de *de* 5,0 μL a ser completado com água ultrapura autoclavada. No caso de amostras de DNA extraído por Chelex ou *Kit QIAamp DNA Investigator*, o volume de amostra utilizado no PCR foi de 0,5 a 2 μL. Juntamente com as alíquotas de DNA das amostras, foi preparada uma reação em que o volume de DNA foi substituído por água ultrapura autoclavada, como controle negativo e uma reação com amostra de DNA previamente utilizada e que tinha bom rendimento de amplificação com os *primers* utilizados.

Para a amplificação do DNA, o preparo da PCR foi realizado em uma cabine esterilizada por luz ultravioleta de comprimento de onda de 254 nm e as reações foram realizadas em termociclador automático modelo *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*). As condições termocíclicas são apresentadas na Tabela 8 e as sequências dos *primers* na Tabela 9.

Algumas amostras de DNA falharam em amplificar a RC com esse par de *primers*, que gera um fragmento de 1211 pb. Por existir a possibilidade dessas amostras estarem

parcialmente degradadas, usamos um protocolo alternativo para a obtenção da sequência completa, que amplifica fragmentos menores. Nesse caso a amplificação foi realizada em 2 fragmentos, denominados RC1 e RC2, definidos pelos pares de *primers* L15967 - H20 (622 pb) e L16475 - H639 (733 pb), respectivamente (Tabela 9).

As condições de amplificação para esse protocolo foram com os *primers* a uma concentração de 0,5 µM, BSA 0,16 µg/µL, com aproximadamente 2-4 ηg de DNA extraído pelo método orgânico ou pelo *Kit QIAamp DNA Investigator* (QIAGEN) e 5,0 µL de *QIAGEN Multiplex* PCR *Master Mix (QIAGEN), num* volume final de *de* 10,0 µL, a ser completado com água ultrapura autoclavada. No caso de amostras de DNA extraído por Chelex ou não quantificadas, o volume de amostra utilizado no PCR foi de 1 a 2 µL. As condições termocíclicas são apresentadas na Tabela 10.

Tabela 8 - Condições termocíclicas para a amplificação da Região Controle do mtDNA com o *QIAGEN Multiplex* PCR *Master Mix*

Etapa	Ciclos	Temperatura (ºC)	Tempo
Desnaturação inicial	1	95	10 min
Desnaturação	35	94	30 s
Anelamento		60	1 min e 30 s
Extensão		72	1 min
Extensão final	1	72	10 min

Legenda: Etapas utilizada na reação de amplificação da região controle do mtDNA, com número de ciclos, suas respectivas temperaturas e durações.

Para a avaliação inicial dos produtos de amplificação, foi realizada eletroforese em gel de agarose a uma concentração de 1,2 % em tampão de corrida TAE 1x (tris 40 mM, ác. acético 20 mM, EDTA 1 mM). As amostras foram preparadas adicionando-se 2 µL do produto de amplificação a 1 µL de *Safer Dye* (Kasvi) e posteriormente aplicadas no gel e em uma cuba horizontal, a 5,5 V/cm, por 30 min. Também foi aplicado no gel o marcador de tamanho molecular *GeneRuler 50 bp DNA Ladder* (*Thermo Fisher Scientific*), composto de uma mistura de fragmentos de DNA com comprimentos entre 50 a 1000 pb.

Após a eletroforese, o gel foi visualizado em um transiluminador de luz azul (modelo K33-333, Kasvi), sendo possível a identificação dos fragmentos de interesse por comparação com o marcador de peso molecular. De acordo com a intensidade da banda obtida em uma avaliação visual, foi possível estipular a quantidade de produto de PCR na reação de sequenciamento. No caso do protocolo alternativo (RC1 e RC2), apenas as amostras com os 2 fragmentos amplificados foram sequenciadas.

Os produtos de PCR foram purificados a partir da adição de 2 µL do reagente *illustra ExoProStar* (*GE Healthcare*) a 5 µL de cada amostra amplificada, deixando, em seguida, a

mistura reagir no termociclador *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*) durante 30 min, sendo a 37 ºC nos primeiros 15 min e 80 ºC nos 15 min finais. Essa etapa, que retira *primers* e nucleotídeos não incorporados na amplificação, antecede a etapa da reação de sequenciamento.

Tabela 9 - *Primers* usados na amplificação e no sequenciamento da Região Controle do mtDNA

Primer	Sentido	Sequência	Referência
L15967	Direto	5'- GTC TTT AAC TCC ACC ATT AGC ACC- 3'	NGUIDI, 2021
L15997	Direto	5'- CAC CAT TAG CAC CCA AAG CT - 3'	WILSON *et al*., 1995
L16475	Direto	5'- TAGCTAAAGTGAACTGTATCC- 3'	NGUIDI, 2021
L16268	Direto	5'- CAC TAG GAT ACC AAC AAA CC - 3'	GABRIEL *et al*., 2001
L16555	Direto	5'- CCC ACA CGT TCC CCT TAA AT - 3'	SIMÃO *et al*., 2018
L314	Direto	5'-CCGCTTCTGGCCACAGCACT-3'	SIMÃO *et al*., 2018
H016	Reverso	5'- CCC GTG AGT GGT TAA TAG GGT - 3'	EDSON *et al*., 2004
H020	Reverso	5'- AGCTCCCGTGAGTGGTTAATA- 3'	NGUIDI, 2021
H159	Reverso	5'- AAATAATAGGATGAGGCAGGAATC - 3'	GABRIEL *et al*., 2001
H388	Reverso	5'- GTTTAAGTGCTGTGGCCAGAAG-3'	SIMÃO *et al*., 2018
H484	Reverso	5'-TGAGATTAGTAGTATGGGAG-3'	EDSON *et al*., 2004
H639	Reverso	5'- GGG TGA TGT GAG CCC GTC TA - 3'	SIMÃO *et al*., 2018

Legenda: Nome dos *primers* diretos e reversos e suas respectivas sequências

Tabela 10 - Condições termocíclicas para a amplificação dos fragmentos RC1 e RC2 do mtDNA com o *QIAGEN Multiplex* PCR *Master Mix*

Etapa	Ciclos	Temperatura (ºC)	Tempo
Desnaturação inicial	1	95	11 min
Desnaturação	36	95	1 min
Anelamento		60-RC1; 55-RC2	1 min
Extensão		72	1 min
Extensão final	1	72	10 min

Legenda: Etapas utilizada na reação de amplificação dos fragmentos RC1 e RC2 da região controle do mtDNA, com número de ciclos, suas respectivas temperaturas e durações.

O sequenciamento foi efetuado através do *kit BigDye Terminator V3.1 Cycle Sequencing* (*Applied Biosystems*), segundo instruções do fabricante. Em casos que não foi atingida dupla cobertura no sequenciamento com os *primers* usados na PCR, os fragmentos foram submetidos a uma ou mais reações com diferentes *primers*, no sentido direto ou reverso, que possibilitaram a cobertura de toda a região controle. Ou seja, dependendo do estado de cada amostra, em relação à qualidade das sequências, com posições duvidosas, ou com presença de heteroplasmias de comprimento e/ou de posição, foram feitos sequenciamentos com outros *primers*, em ambos os sentidos, até que não houvesse dúvida em

qualquer das posições (PARSON; BANDELT, 2007). Nas amostras em que foi observada heteroplasmia de comprimento associada a inserções AC entre as posições 513 e 525 e/ou pela transição na posição 460, o sequenciamento foi também obtido com os *primers* L314, H338 e H484. Os *primers* usados no sequenciamento estão listados na Tabela 9.

As reações de sequenciamento foram realizadas em termociclador automático modelo *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*). Todos os *primers* foram usados a uma concentração de 0,25 µM e a quantidade do produto de PCR purificado a ser utilizado na reação variou de 0,5 a 2,0 µL, dependendo da intensidade da banda correspondente no gel de agarose.

Após a reação de sequenciamento, o produto foi purificado em colunas contendo resina Sephadex G-50 Medium (*GE Healthcare Life Sciences*), que faz a separação das moléculas de ácidos nucléicos de tamanhos diferentes por exclusão. O produto da reação de sequenciamento foi cuidadosamente, adicionado sobre cada coluna, e os tubos foram centrifugados a 3879 x g por 3 min. Ao produto purificado, coletado em microtubo após eluição da coluna, foram adicionados 10 µL de Formamida Hi-Di (*Applied Biosystems*), agente desnaturante. A solução resultante foi aplicada em placa para a eletroforese capilar no Analisador Genético ABI 3500 (*Applied Biosystems*), que realiza a separação dos fragmentos por eletroforese em capilares de 50 cm de comprimento preenchidos com polímero POP-7 (*Applied Biosystems*) e detecta os nucleotídeos incorporados, marcados com um fluoróforo específico.

A determinação dos haplótipos da região controle do mtDNA foi realizada através da comparação das sequências obtidas com a rCRS (*revised Cambridge Reference Sequence*), através do *software SeqScape v2.7* (*Applied Biosystems*). Na determinação dos haplótipos seguiu-se as diretrizes propostas pela ISFG (PARSON *et al*., 2014).

2.5.2 Classificação dos haplogrupos de mtDNA e análise estatística

Os haplogrupos foram classificados com auxílio com a ajuda da base de dados EMPOP database v4/R13 (PARSON; DÜR, 2007) e confirmados manualmente na *Phylotree - mtDNA tree Build* 17, atualizada em Fevereiro de 2016 (VAN OVEN; KAYSER, 2009). Os haplótipos serão submetidas à base de dados EMPOP, para efeitos de controle de qualidade.

As proporções de haplogrupos europeus, africanos e ameríndios na amostra analisada foram calculadas por contagem direta no Excel (*Microsoft Corporation*) e comparadas com outras populações brasileiras anteriormente analisadas incluídas na Tabela 11 e com populações da Região Norte do Brasil (NOGUEIRA *et al*., 2017).

Através da consulta de dados da literatura, baseados nas sequências da RC de algumas populações brasileiras e da América do Sul (Tabela 11), foram obtidos valores referentes ao número de haplótipos na população e sua diversidade haplotípica (HD). Para outras populações, das quais tais dados não estavam disponíveis, esses valores foram calculados juntamente com a população da Ilha de Marajó, através dos haplótipos publicados, utilizando o *software* Arlequin versão 3.5 (EXCOFFIER; LISCHER, 2010). Essas populações foram: Colômbia, Equador, Paraguai, Peru e São Paulo.

O mesmo *software* foi utilizado nos cálculos de distâncias genéticas (F_{ST}) e probabilidades de não-diferenciação entre a amostra estudada e as outras populações com dados previamente publicados para a Região Controle completa (Tabela 11).

O alinhamento das sequências para as análises no *software* Arlequin foi realizado no *software Haplosearch* (FREGEL; DELGADO, 2011) com posterior correção de alinhamento da EMPOP (PARSON; DÜR, 2007). Uma vez que as sequências homopoliméricas que comumente ocorrem nas posições 16183-16194, 302-310 e 568-573 do mtDNA, assim como as repetições diméricas entre 513-525, rotineiramente não são consideradas nas interpretações de perfis na casuística forense e em bases de dados, tais posições foram excluídas das sequências analisadas neste trabalho para fins de comparação de diversidade intra- e interpopulacional (PARSON *et al*., 2014).

Para auxiliar na visualização das distâncias genéticas pareadas obtidas foi construído o gráficos de escalonamento multidimensional (MDS) em duas dimensões através do programa *Statistica Software* v. 14.0.0.15 (TIBCO Software Inc.).

2.6 **Análise dos AIMs**

O sistema utilizado neste trabalho para inferência de ancestralidade reúne 46 AIM-Indels que apresentam diferenças nas frequências alélicas entre os grupos ancestrais africano, europeu, nativo americano e asiático, sendo altamente divergentes entre pelo menos dois destes grupos (PEREIRA *et al*., 2012a)**.** Após a genotipagem das amostras foi feita a análise

de inferência de ancestralidade com o *software* STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000).

Tabela 11 - Populações utilizadas para comparação com a população da Ilha de Marajó, com base nos dados obtidos pelo sequenciamento da Região Controle completa

População	Número de indivíduos	Etnia	Referência
Marajó	101	Miscigenada	Este trabalho
Argentina, Norte	98	Miscigenada	BOBBILO *et al*., 2009
Argentina, Sul	47	Miscigenada	BOBBILO *et al*., 2009
Argentina, Centro	193	Miscigenada	BOBBILO *et al*., 2009
Colômbia, Medelín	94	Miscigenada	AUTON *et al*.,2015
Distrito Federal	306	Miscigenada	FREITAS *et al*., 2019
Equador	156	Miscigenada	BAETA *et al*., 2014; BRANDINI *et al*., 2018
Espírito Santo, Centro	54	Miscigenada	DOS REIS *et al*. (2019)
Espírito Santo, Metropolitano	81	Miscigenada	DOS REIS *et al*. (2019)
Espírito Santo, Norte	17	Miscigenada	DOS REIS *et al*. (2019)
Espírito Santo, Sul	62	Miscigenada	DOS REIS *et al*. (2019)
Paraguai	537	Miscigenada	SIMÃO *et al*., 2021
Paraná	122	Miscigenada	POLETTO *et al*., 2018
Peru, Lima	83	Miscigenada	AUTON *et al*.,2015
Rio de Janeiro	205	Miscigenada	SIMÃO *et al*., 2018
Santa Catarina	80	Miscigenada	PALENCIA *et al*., 2010
São Paulo	142	Miscigenada	PRIETO *et al*., 2011
Venezuela, Caracas	101	Miscigenada	CASTRO DE GUERRA *et al*., 2012

2.6.1 Genotipagem dos 46 AIM-Indels

A genotipagem dos marcadores AIM-Indels autossômicos foi realizada em 160 amostras de indivíduos não aparentados, residentes e que possuem avós e avôs maternos e paternos nascidos na Ilha de Marajó.

Para a análise dos 46 marcadores foi usado um conjunto de 46 pares de *primers* em um único sistema multiplex descrito por Pereira e colaboradores (2012a). O preparo da PCR foi realizado em uma cabine esterilizada por luz ultravioleta de comprimento de onda de 254 nm e as reações foram realizadas em termociclador automático modelo *Veriti 96-Well Thermal Cycler* (*Applied Biosystems*). As condições termocíclicas são apresentadas na Tabela 12. A Tabela 13 contém a informação dos 46 marcadores em relação à sequência polimórfica e ao fluoróforo incorporados ao *primer* por meio de ligantes não-nucleotídicos. Os quatro fluoróforos utilizados (6-FAM, VIC, NED, PET) emitem fluorescência em diferentes comprimentos de onda, permitindo, assim, que os fragmentos dos 46 AIM-Indels

amplificados na mesma reação de PCR sejam separados e detectados eficientemente na eletroforese capilar.

A reação de amplificação dos marcadores AIM-Indels foi adaptada do protocolo de Pereira e colaboradores para um volume de 5 µL. Na PCR, os *primers* foram utilizados a uma concentração de 1 µM, com aproximadamente 1-2 ηg de DNA extraído pelo método orgânico ou 0,5 - 2 µL, no caso de DNA extraído por Chelex ou *Kit QIAamp DNA Investigator*. Além disso, 2,5 µL de *QIAGEN Multiplex* PCR *Master Mix (QIAGEN) e água* ultrapura autoclavada para completar o volume final de reação *de* 5,0 µL. Juntamente com as alíquotas de DNA das amostras, foi preparada uma reação em que o volume de DNA foi substituído por água ultrapura autoclavada, como controle negativo da reação e uma reação com amostra de DNA previamente utilizada e que tinha bom rendimento de amplificação para esse sistema multiplex.

O volume de 1 µL dos produtos da purificação foram aplicados em placa a ser colocada no analisador genético ABI 3500 (*Applied Biosystems*) para a eletroforese em capilar de 50 cm preenchido com o polímero POP-7 (*Applied Biosystems*), juntamente com uma mistura contendo 8,8 µL de Formamida Hi-Di (*Applied Biosystems*) e 0,2 µL de padrão interno de tamanho *GeneScan 500 LIZ dye Size Standard v2.0* (*Applied Biosystems*).

Na análise dos produtos de amplificação e nomeação dos alelos usou-se o *software GeneMapper v. 4.1* (*Thermo Fisher Scientific*).

2.6.2 Análise estatística dos resultados dos AIM-Indels

Para estimar as proporções de cada contribuição continental foi utilizado o *software* STRUCTURE *v2.3.4* (PRITCHARD; STEPHENS; DONNELLY, 2000). Foi realizada uma análise supervisionada, utilizando informações prévias sobre a origem geográfica das amostras de referência da África, da Europa e dos nativos americanos, do painel de diversidade humana HGDP-CEPH (http://www.hagsc.org/hgdp/) [PEREIRA *et al*., 2012a]. As corridas de *STRUCTURE* compreenderam repetições de 100000 etapas seguidas por 100.000 iterações MCMC. Foi considerada uma contribuição tri-híbrida de ameríndios, europeus e africanos (K = 3), por serem sabidamente as populações ancestrais formadoras da população brasileira.

Foi utilizado o modelo *Admixture (Use population Information to test for migrants)*. As frequências de alelos foram correlacionadas e atualizadas utilizando apenas indivíduos com o parâmetro *POPFLAG* = 1 (neste caso, as amostras de HGDP-CEPH utilizadas como referência – africanas, europeias e ameríndias).

Tabela 12 - Condições termocíclicas para a amplificação do multiplex AIM-Indels com o *QIAGEN Multiplex* PCR *Master Mix*

Etapa	Ciclos	Temperatura (ºC)	Tempo
Desnaturação inicial	1	95	15 min
Desnaturação	28	94	30 s
Anelamento		60	1 min e 30 s
Extensão		72	45 s
Extensão final	1	72	10 min

Legenda: Etapas da reação de amplificação dos AIM-Indels, com ciclagem - temperaturas e durações.

Tabela 13 - Sistema multiplex 46 AIM-Indels

Marcador	Polimorfismo Indel / Fluoróforo	Marcador	Polimorfismo Indel/ Fluoróforo
MID-1470	-/GTTAC; 6-FAM	MID-2431	-/ATTG; VIC
MID-777	-/GAA; 6-FAM	MID-2264	-/AAGT; VIC
MID-196	-/CAT; 6-FAM	MID-2256	-/CAT; NED
MID-881	-/ACTT; 6-FAM	MID-128	-/ATT; NED
MID-3122	-/ATCT; 6-FAM	MID-15	-/AAATACACAC; NED
MID-548	-/CT; 6-FAM	MID-2241	-/GTCCAATA; NED
MID-659	-/CT; 6-FAM	MID-419	-/AATGGCA; NED
MID-2011	-/CTAGA; 6-FAM	MID-943	-/TGAT; NED
MID-2929	-/TA; 6-FAM	MID-159	-/CCCCA; NED
MID-593	-/TT; 6-FAM	MID-2005	-/AACAAT; NED
MID-798	-/GGGAAA; 6-FAM	MID-250	-/CA; NED
MID-1193	-/AT; 6-FAM	MID-1802	-/GGA; NED
MID-1871	-/TT; 6-FAM	MID-1607	-/TG; NED
MID-17	-/TAAC; 6-FAM	MID-1734	-/CCAT; PET
MID-2538	-/AACA; 6-FAM	MID-406	-/AG; PET
MID-1644	-/GT; 6-FAM	MID-1386	-/AAACTATTCATTTTTCACCCT; PET
MID-3854	-/TCTA; VIC	MID-1726	-/CAAGAACTATAAT/CACTATCTATTAT; PET
MID-2275	-/TCAGCAG; VIC	MID-3626	-/AATATAATTTCTCCA; PET
MID-94	-/AAC; VIC	MID-360	-/AA; PET
MID-3072	-/GCCCCCA; VIC	MID-1603	-/TTGT; PET
MID-772	-/TAG; VIC	MID-2719	-/AACT; PET
MID-2313	-/ATTATAACT; VIC		
MID-397	-/TTCT; VIC		
MID-1636	-/AA; VIC		
MID-51	-/TTTAT; VIC		

Legenda: O código MID para cada marcador, a sequência Indel polimórfica e o fluoróforo utilizado em cada par de *primer*. O código MID indica a nomeclatura dos marcadores utilizada na base de dados *Marshfield Diallelic Insertion/Deletion Polymorphisms*.

Fonte: Tabela adaptada de PEREIRA *et al*., 2012a.